The Dissertation Committee for Jing Ai Certifies that this is the approved version of the following dissertation:

Supervised and Unsupervised PRIDIT for Active Insurance Fraud Detection

Committee:

Patrick L. Brockett, Co-Supervisor

Linda L. Golden, Co-Supervisor

William W. Cooper

Bin Gu

Montserrat Guillén

Richard D. MacMinn

Maytal Saar-Tsechansky

Supervised and Unsupervised PRIDIT for Active Insurance Fraud Detection

by

Jing Ai, B.S.; M.S.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

August, 2008

Acknowledgements

I would like to first extend my sincerest appreciation to my advisor Professor Patrick L. Brockett who has guided me through my doctoral study with his insights, diligence, and patience. I am indebted to Professor Brockett for every advancement that I have made on this journey.

I would like to express my great gratitude to my co-supervisor Professor Linda L. Golden, who has supported me throughout this time and has made it easier for me. She has also offered me her great insights and advice on my research and my career.

I am also truly grateful to Professor William W. Cooper, who has continuously provided with me his support and guidance from the very beginning. I would not have had the opportunity to start this endeavor if it was not for Professor Cooper.

I would also like to thank my other committee members Professor Maytal Saar-Tsechansky, Professor Bin Gu, Professor Richard D. MacMinn, and Professor Montserrat Guillen for their valuable comments and advice on my research and my professional development.

Finally, I want to thank my parents, my family, and my friends for their love, understanding, support, and patience.

Supervised and Unsupervised PRIDIT for Active Insurance Fraud Detection

Publication No._____

Jing Ai, Ph.D.

The University of Texas at Austin, 2008

Supervisors: Patrick L. Brockett, Linda L. Golden

This dissertation develops statistical and data mining based methods for insurance fraud detection. Insurance fraud is very costly and has become a world concern in recent years. Great efforts have been made to develop models to identify potentially fraudulent claims for special investigations. In a broader context, insurance fraud detection is a classification task. Both supervised learning methods (where a dependent variable is available for training the model) and unsupervised learning methods (where no prior information of dependent variable is available for use) can be potentially employed to solve this problem.

First, an unsupervised method is developed to improve detection effectiveness. Unsupervised methods are especially pertinent to insurance fraud detection since the nature of insurance claims (i.e., fraud or not) is very costly to obtain, if it can be identified at all. In addition, available unsupervised methods are limited and some of them are computationally intensive and the comprehension of the results may be ambiguous. An empirical demonstration of the proposed method is conducted on a

widely used large dataset where labels are known for the dependent variable. The proposed unsupervised method is also empirically evaluated against prevalent supervised methods as a form of external validation. This method can be used in other applications as well.

Second, another set of learning methods is then developed based on the proposed unsupervised method to further improve performance. These methods are developed in the context of a special class of data mining methods, active learning. The performance of these methods is also empirically evaluated using insurance fraud datasets.

Finally, a method is proposed to estimate the fraud rate (i.e., the percentage of fraudulent claims in the entire claims set). Since the true nature of insurance claims (and any level of fraud) is unknown in most cases, there has not been any consensus on the estimated fraud rate. The proposed estimation method is designed based on the proposed unsupervised method. Implemented using insurance fraud datasets with the known nature of claims (i.e., fraud or not), this estimation method yields accurate estimates which are superior to those generated by a benchmark naïve estimation method.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Insurance Fraud and Detection Methodologies

1.1 Overview of This Dissertation

This dissertation aims at developing statistical and data mining based methodologies for insurance fraud detection. Insurance fraud has become a world concern in recent years. Due to its growth in the past two decades, great efforts have been made to develop models to detect potentially fraudulent claims more effectively. These models can help insurers and regulators to understand insurance fraud better and can potentially reduce the cost incurred due to insurance fraud.

In developing fraud detection models, both supervised learning methods (where a dependent variable is available to train the dataset) and unsupervised learning methods (where no prior information of dependent variable is available for use) are employed by researchers. Applications of the two categories of methods differ in the available data structure. Supervised learning methods make use of prior knowledge of a dependent variable in at least one (training) subset of data. In this situation, one desires to train the model in order to discover patterns in the set of predictor variables, which allows one to predict the value of the dependent variable for incoming datasets. Unsupervised learning methods are applied when the dependent variable is unavailable, and thus these methods are required to extract information directly from predictor variables without access to a true value of the dependent variable even for a subset of data.

In many application areas (e.g., insurance fraud detection) however, labels of the dependent variable are very costly, if at all possible to obtain, making unsupervised techniques either the only option, or the only cost effective option. Despite the rising need, available unsupervised methods are limited and some of them suffer from intensive computation and ambiguous comprehension of the results. As an attempt to attack this problem, I am going to propose in this dissertation new (unsupervised and semi-supervised) methods for the use in insurance fraud detection and related applications. Other important issues of fraud detection are also discussed.

The methods proposed in this dissertation are based on the unsupervised PRIDIT (Principal Component Analysis of RIDIT Scores, where RIDIT refers to "Reference to an Identified Distribution" (Bross, 1958)) method (Brockett et al., 2002). The original PRIDIT method makes use of categorical predictor variables for binary classification task. Chapter 2 presents an overview of the original discrete PRIDIT method. In Chapter 3, I first investigate the development of a general PRIDIT method to include continuous predictor variables. Empirical evaluation of the general PRIDIT method will be conducted and comparisons will be made between the PRIDIT method and a supervised learning method as well as a competitive unsupervised learning method. In Chapter 4, the performance of the PRIDIT method is further assessed by comparing with a set of prevalent supervised learning methods as a form of external validation. The relative strengths and weaknesses of each method are discussed.

In Chapter 5, the information provided by the PRIDIT analyses is used to design a class of new methods in the context of active learning. First, the PRIDIT ranking of claims is used to select an "informative" initial sample for the use of a subsequent

supervised method to build a classification model. Second, a hybrid method is built where the PRIDIT-assessed labels are used in place of true labels for the initial training sample before a supervised method is used to train the model. These two methods are both empirically evaluated using an insurance fraud dataset. Chapter 5 also examines another important issue in fraud detection, namely, how to estimate the fraud rate, i.e. the percentage of fraudulent claims in the claim set. The estimation method proposed is based on the PRIDIT method and is evaluated empirically using insurance fraud datasets.

1.2 INTRODUCTION TO INSURANCE FRAUD DETECTION

Insurance fraud refers to situations when the insureds submit illegitimate claims to the insurance company and attempt to realize financial gains. It has been found in a wide variety of insurance lines (e.g., automobile insurance, health insurance, workers compensation insurance, life insurance, etc.) and can be present at various stages of the insurance product life cycle. Moreover, it can be divided into different types according to various criteria (e.g., whether it is planned or opportunistic, whether it is a first party or a third party insurance claim). However, unlike many other phenomena, there is usually no clear definition of insurance fraud. Litigation is often considered the only means by which one can provide a sure assessment that a claim is fraudulent. In practice, however, suspicious claims are seldom settled by a court but are settled instead through private negotiations between the insurance company and the insured (Brockett et al. 2002). Also for this reason, there is no consensus on how much fraud is inherent in the current insurance market. Most of the available estimates come from reasonable conjectures rather than formal estimation procedures.

Due to the great cost to the insurance industry and the consumers, insurance fraud has become a world concern. The cost was estimated to be $80 billion in the U.S.A. in 2003[1] and has been increasing in recent years. Many efforts have been made by the insurance industry, consumer organizations, and academia to detect insurance fraud. One of the most important approaches is to design statistical or data mining based methodologies for the purpose of effective detection and this is also the topic of this dissertation.

Put in a broader context, insurance fraud detection is a classification task. The insurance companies want to classify the claims into a fraudulent subset and a legitimate subset. One unique feature of insurance fraud is that the dependent variable or the nature of the insurance claim (whether it is fraudulent or not) is usually very costly to detect or even impossible to obtain. And even if one is able to obtain a training sample with dependent variable values, they are subject to errors that will significantly decrease the usefulness of this training sample. This is especially relevant for emerging insurance markets such as China, where there is virtually no insurance fraud data to date. This unique feature makes insurance fraud detection very difficult and has inspired great interest in further developing and improving detection methodologies.

In detection of fraud, a wide variety of statistical based and data mining based approaches have been proposed. These methods can be classified into two broad groups: supervised learning methods and unsupervised learning methods, depending upon the structure of the data available. Supervised learning methods are well developed and

[1] This estimate appears in "Insurance Fraud: the Crime You Pay For" by Coalition Against Insurance Fraud. Available at http://www.insurancefraud.org/fraud_backgrounder.htm.

accessible. However for the purpose of insurance fraud detection, unsupervised data classification methods, which extract information directly from predictor variables without the access to a true value of the dependent variable even for a subset of data, seem to be the most suitable. Therefore, I am going to review literature on both classes of methods in the context of fraud detection, with a particular focus on unsupervised learning methods.

1.3 REVIEW OF SUPERVISED LEARNING METHODS USED IN FRAUD DETECTION

There are a large number of supervised learning methods applied to fraud detection. Among them, standard econometric models are commonly used. A series of studies have been conducted by Derrig, Weisberg and other authors (1991, 1992, 1993, 1994, 1995, and 1998) on automobile fraud detection using statistical based methods. These studies attempt to provide insights into the broad question of handling suspicious claims, from the claims screening process to the identification of potentially fraudulent claims.

In their 1991 baseline study, Weisberg and Derrig established the ground work for insurance fraud study. They collected a representative sample of bodily injury (BI) claims from July 1, 1985 to June 30, 1986. In this baseline study, they use simple descriptive statistics to examine characteristics of the claims from the aspects of the accident, the claimant, the claim, the injury, and the treatment. They attempt to discover the relationships among the claim characteristics and to discuss subjective assessments by adjusters. Although this study is mainly descriptive, it has laid the conceptual background for the future work.

Weisberg and Derrig (1993) continue to explore the insurance fraud patterns by using multiple regressions to analyze bodily injury (BI) and personal injury protection (PIP) claims in 1989. Each claim is coded by four coders including adjusters and claim investigators. Three sets of regressions are run with the suspicion rating given by adjusters, the suspicion rating given by investigators, and the fraud vote as the dependent variable respectively. Twenty five indicators are selected out of a total number of sixty five indicators by correlations with the outcome. Regression models in each set are built using subsets of the twenty five indicators up to the size of ten indicators in each model. The results of the best five models (in terms of R^2) in each set are recorded and compared. Claim handling procedures and techniques are also discussed.

Derrig and Weisberg (1995) further investigate this issue using the 1993 personal injury protection (PIP) claims dataset and apply more sophisticated econometric models. A set of six models are examined and the results are compared. The six models include a "Naïve" model which simply uses the count of indicators, a ten-indicator linear regression, a twenty-indicator linear regression, a linear regression with interactions of the indicators, a CART (Classification And Regression Tree) model, and the linear regression model developed using 1989 claims data as in Weisberg and Derrig (1993). The performances of the models are evaluated using ROC (Receiver Operating Characteristic) curves. As a second step, the claim investigation process is examined and summarized in Derrig and Weisberg (1998). There are two other papers discussing BI tort reform and behavioral factors, which are of less relevance here (Weisberg and Derrig 1992 and Derrig et al. 1994).

The above series were the first studies to formalize the identification of suspicious claims and the improvement of the investigation. General statistical techniques were used to address these issues. As fraud detection attracts more attention and interest, other more sophisticated and more specifically-oriented fraud detection methods are proposed and applied.

Discrete choice models are another class of well-understood econometric models which have received special attention in fraud detection applications. Artis et al. (1999) use a multinomial Logit model to classify three types of claims: legitimate claims, self-benefit fraud and third-party-benefit fraud, after first establishing the function of expected utility from fraud and identifying expected claim cost. A nested multinomial Logit model is also employed to decompose fraud behavior into two steps: (1) the decision to commit fraud and (2) the choice between two types of fraud given fraud. The authors apply their estimation methods to a dataset consisting of insurance claims from the Spanish automobile insurance market. They find that the correct classification percentages obtained by this method differ by types of fraud behavior. Artis et al. (2002) extend their 1999 study by modifying the Logit model to allow for omission errors in the training dataset (previously classified claims), which are the part of the fraudulent claims that are not identified as fraud. Although both omission errors and commission errors (when the designated fraudulent claims are indeed honest claims) may exist, only omission errors are considered since they are more costly to the insurer. They use the same Spanish auto insurance dataset and find that the estimated percentage of misclassification is around 5%. They also find that when omission error is significant, the model is not well specified.

Caudill et al. (2005) further develop Artis et al. (2002) and estimate a multinomial Logit model with missing information in the dependent variable (when some of the labeled "honest" claims set may contain both truly honest claims and truly fraudulent claims). In this case, further classification based on the inaccurate labels in the training datasets is mostly likely to be unreliable. To identify misclassified claims, Caudill et al. apply an Expectation-Maximization (EM) algorithm to estimate the probability of a claim being a fraud case and the conditional probability of a claim being a fraud case given that it is labeled as honest,. The method is tested on a Spanish automobile insurance dataset.

Alternatively, Belhadji et al. (2000) adopt a Probit model in their study. They first use the conditional probability of fraud given each predictor variable to screen the set of predictor variables and then use a Probit model for estimation.

Although the Logit model and the Probit model are generally considered to be interchangeable in the literature, researchers have started to notice the importance of making the choice between these two classes of models, especially in applications where the "events" that the model is trying to capture (e.g., fraudulent cases in the context of insurance fraud detection) are rare (Jin et al., 2005). It is argued that probability estimations and thus classifications can be vastly different under these two models in this case. To select the appropriate model, a test by Silva (2001) is proposed, which is based on the combined likelihood function of the two competing models.

In general, econometric models are familiar and easy to implement. They also have the advantage of providing concrete, interpretable results. However, they depend on model specifications and rely on the quality of the dependent variable. In the fraud detection context, this might cause a problem because the value of the dependent variable

(fraud or not) is subjective and might not be correct. Thus, the results sometimes are not reliable.

The Expert System approach is also used widely for fraud detection. Expert system incorporates human experts expertise to aid the investigation. Major and Riedinger (2002) propose an Electronic Fraud Detection system. They design a five layer system consisting of measuring the behavior, using information numbers to identify "frontiers", natural language querying, an investigator deciding to initiate investigations, and validating and learning.

Fuzzy techniques are often used in expert systems. Stefano and Gisella (2001) use fuzzy expert system techniques to outputs four indices: suspect index, elements of suspect, competence, body injury. Their system is tested on a small Italian auto claims dataset. Pathak et al. (2003) use a fuzzy- expert system. They use fuzzy logic under which membership in a set can be assessed not simply as a 0-1 dichotomy, but can be associated with any real number between 0 and 1. An eight step system is created. The main parts of the eight-step system include human investigation and claims-pickup for further detection, determining the fuzzy membership of the crisps (real world inputs), using fuzzy "If-Then" logic rules to transform crisp input values to linguistic values of outputs, and outputting a decision of potential fraud or not.

Derrig and Ostaszewski (1995) also employ fuzzy logic. They use clusters sorted by experts. A "fuzzy c-means" iterative semantic algorithm is used to deal with numerical data. First, a distance metric is defined to measure degree of "similarity". Second, the center of each cluster is located. Third, the fuzzy membership function values are normalized. Fourth, a stopping rule is applied to decide if the process should be

terminated. And, lastly, small membership values are discarded. This method is applied to a 1989 bodily injury liability claims dataset in Massachusetts.

Expert systems can incorporate valuable and important human expert knowledge. The fuzzy techniques used in many of the expert systems approaches are able to capture "ambiguity"--the very nature of fraud classification. However, compared with other supervised learning methods, such as discrete choice models, expert systems are generally less straightforward to use. Also, core components of the methodology such as membership functions and the rule base need further investigation. The optimality decision is also critical in the success of expert systems.

There are still other supervised methods that have been applied to fraud detection. Naive Bayes, Adaboosted Naive Bayes, Adaboosted Weight of Evidence Scoring are adopted by Viaene et al. (2004a). A variety of criteria including Classification Accuracy, ROC Curve, and AUROC (Area Under ROC curve) are employed for evaluation. Using a PIP automobile insurance claims dataset, they find that Adaboosted Weight of Evidence is superior to Naïve Bayes methods in calibrating probability estimates. A series of supervised classification methods for binary indicators are examined in Viaene et al. (2002), and their performances are evaluated and compared. We very briefly present the methods as follows. Logistic regression, as a well known regression method, estimates the log odds ratio. As in many other studies, it is adopted as the benchmark to assess performance of our method. K-nearest neighbor method estimates average probabilities in local neighborhoods. As suggested by its name, it uses the K-th nearest point to define the concept of "neighborhood". C4.5 decision tree follows the widely used divide-and-conquer approach. The Gain Ratio, as defined in Quinlan (1993), is used as the rule for

node partitioning. Additional stopping conditions are imposed to avoid trivial partitioning. Least-Squares support vector machine classifier, based on Cristianini and Shawe-Taylor (2000) and Vapnik (1995, 1998), is constructed as essentially an optimizing problem to minimize squares of weight vectors. It is calculated using Mercer's condition (Cristianini and Shawe-Taylor 2000 and Vapnik 1995, 1998) and it allows for misclassification. Bayesian methods such as Naïve Bayes and Bayesian learning neural network are also introduced. Naïve Bayes is the simplest form of Bayesian network classifier which assumes the conditional independence of predictors given the class. Then it calculates the posterior class membership by Bayes' Theorem. The particular Bayesian learning neural network presented is a feed-forward multilayer perceptron neural network. It incorporates the Bayesian learning by updating weights and hyper-parameters according to Bayes' Law after observing the training data. By implementing the methods in a 1993 Personal Injury Protection insurance claims dataset from Massachusetts, Viaene et al. (2002) evaluate the performance of the above algorithms. The classification results are presented using mean PCC (Percentage Correctly Classified) and mean AUROC (Area Under Receiver Operating Characteristics curves). A two-way ANOVA analysis and Duncan's multiple range test (Duncan 1955) are used to compare the measures to assess performance. The authors discover that performance differences across the methods studied are quite small, suggesting that easy-to-implement methods are often the best choice.

A special class of supervised learning methods, active learning, may play a potentially significant role in designing new fraud detection tools. As is well known, supervised learning is likely to produce more accurate results since they are trained on

data with true labels. However, acquiring labels for the training data are often expensive. Particularly in insurance fraud detection, the process of investigating claims to obtain labels (fraudulent or not) requires labor input from adjusters and other resources at a cost. Consequently, active learning methods are proposed to minimize the incurred labeling cost while maintaining a satisfactory accuracy. Active learning methods commonly employ an initial (small) training set to develop a supervised classifier and additional "most informative" examples are chosen to be labeled and added to the training sample.

"Informative" examples are usually those with the most uncertain classifications (or probability estimations) under the current classifier. The rationale is that those examples probably contain classification information that has not been fully extracted by the current classifier but which can contribute to the learning process if the labels for these examples are known, i.e., they are "informative." Therefore, labels should be obtained for these examples to improve performance in the most efficient way. Different definitions of "informativeness" have been proposed and algorithms have been designed to select these examples. "Query by Committee" applies two classifiers to the unlabeled examples and chooses to label the examples only when the two classifiers don't agree (Seung et al. 1992). "Uncertainty Sampling" identifies examples with probability estimations around 0.5 as the next set to be trained (Lewis and Gale 1994). It is shown to perform well especially for a small training sample relative to the number of positive examples (Lewis and Gale 1995). Saar-Tsechansky and Provost (2004a) adopts a weighted sampling approach to include additional examples to the initial training sample. Bootstrapped samples of the initial training sample are used to obtain the variance in probability estimates for the unlabelled examples. The most "informative" examples in

this case have the highest variance. These examples are then selected according to probabilities corresponding to their respective "informativeness" and true labels are obtained for the selected examples. They demonstrate that their approach, Bootstrap-LV, is superior to random sampling for small training data size, and thus permits potential cost savings. They also note the difference between the task of class probability estimation (CPE) studied in their paper and the classification task used elsewhere. An algorithm designed for classification, such as Uncertainty Sampling, may be inferior when used for CPE. Therefore, a careful consideration on the purpose of training seems to be necessary before an appropriate approach can be chosen to build a model.

Besides the general classification and class probability estimation objectives, other goals can be set up for active learning as well. For example, Saar-Tsechansky and Provost (2004b) investigates the use of active learning in the decision making process. Model accuracy in this case is not the primary purpose. Instead, the learning method is designed to improve decision-making most efficiently by acquiring examples that are most likely to affect the decisions.

Another relevant class of learning methods is cost-sensitive learning. In many situations, cost effectiveness is a more realistic goal compared with simply higher classification accuracy. Stratification has been proposed as one possible solution. The general idea is to oversample the high-cost class (minority class) or undersample the low-cost class (majority class). However, various drawbacks exist for the stratification approach (Domingos, 1999). A MetaCost method is presented as another solution (Domingos, 1999). Under this method, the training sample is relabeled with each example's estimated "optimal class" which minimizes the misclassification cost by

applying a meta-learner based on a classifier. Then the classifier is re-applied to the relabeled training sample to obtain the model. This is a general method which might be employed with a large set of classifiers and applied to a variety of problems. Viaene et al. (2004b) design claim screening strategies with available cost information. A classification rule is defined based on the cost matrix and the estimated class probability obtained from a Logit model. Six scenarios differing in levels of available information on individual claim amount and audit cost are analyzed and the total savings (costs) are compared. Incorporating individual claim amount and average audit cost into the screening technology seems to be a practical cost-effective strategy.

One should be cautious when employing active learning or other approaches to detect fraud. The imbalanced dataset problem is especially profound in fraud data sets. Since fraud is rare, the majority of the datasets have a much smaller proportion of "event" (i.e., fraud) compared to "nonevent" (i.e., nonfraud). In fact, ACM SIGKDD Explorations in 2004 has a special issue on learning from imbalanced datasets. In the Editorial of this special issue, three general classes of approaches are discussed to combat the imbalanced dataset problem. First, sampling approaches are most often used to deal with this problem. Undersampling of the majority class or oversampling of the minority class seem to be natural solutions, but can cause potential problems. New sampling methods using multiple classifiers or ensemble learning are also designed to alleviate the impact of imbalanced datasets. For example, a "stacking-bagging" approach is developed to address the classification of skewed class distribution and is applied in insurance fraud detection (Phua et al., 2004). Based on three base classifiers (1) Backpropagation, (2) Naïve Bayes and (3) C4.5 Decision Tree, this approach makes use of a meta-classifier on

top of the base classifiers (stacking) with the final class prediction obtained from all three individual predictions (bagging). The results are shown to result in higher cost savings than other sampling approaches, such as undersampling, oversampling, bagging alone, stacking alone, etc. Second, one-class learning is another approach that might be employed in this case. When the imbalance is serious, only the target class is used for classification with the aid of a similarity measure. Third, feature selection can be used in this context as well. To solve classification problems where a large amount of features are available, those features best distinguishing the classes are considered first.

Another possible concern is on the dynamic nature of fraud. When there is a move, there is a counter-move. With the development of fraud detection technologies, fraudsters are also striving to improve their skills. Thus, a static classifier is not likely to achieve the same satisfactory performance in the future as in the current experiments. To address this issue, adaptability becomes one important criterion to evaluate different learning methods (Fawcett and Provost, 1997).

In conclusion, supervised learning methods generally provide satisfactory results since they make use of the knowledge of the dependent variable to train the model. Some of them are widely used and are easy to implement with readily available software programs. However, supervised methods do exhibit a few major problems especially in the insurance fraud detection application. As I mentioned earlier, the assumptions underlying most supervised learning methods are not usually consistent with the nature of fraud data. Obtaining the value of the dependent variable for a fraud application is costly and time consuming, if it can be obtained at all. The uneven class size (fraud vs. non-fraud) in fraud data may also cause a potential problem for certain methods. Furthermore,

different types of misclassifications may entail different costs. This is not currently taken into account by most available supervised learning methods.

1.4 REVIEW OF UNSUPERVISED LEARNING METHODS USED IN FRAUD DETECTION

Although still limited in the literature, unsupervised learning methods are being employed in many applications, and they are of special interest in fraud detection applications. Cluster Analysis is perhaps one of the most commonly used unsupervised learning methods. Under this class of models, data is partitioned according to certain "similarity" and "dissimilarity" measures. The choice of specific measures of "similarity" (or "dissimilarity") is critical. Commonly used clustering methods include Ward's method (which minimizes the within-cluster variance), K-means method (reassignment to the nearest centroid at each iteration), Average linkage (where distance is defined as the average distance between all pairs of members of the two clusters), etc.

Cluster analysis is a popular and easy to implement method and application softwares (e.g., SAS) have ready-to-use programs to implement it. However, several drawbacks exist for cluster analysis. First, the performance of cluster analysis is largely dependent on the choice of the similarity metric. Second, clustering models can result in an extremely uneven split of the sample when it is not desired, so the classification is hardly effective. Moreover, the correspondence between pre-specified classes (such as fraud, non-fraud) and the identified clusters is hard to determine. Finally, there is usually no way to incorporate expert opinions in cluster analysis, which could lead to a great loss of information in applications such as insurance fraud detection. Probably due to these

limitations, even though cluster analysis is designed for classification purposes, there is little literature using cluster analysis directly to identify fraudulent claims.

Another major unsupervised method is the unsupervised neural network. Brockett et al. (1998) use Kohonen Feature Maps for fraud classification. The key part of the process is to discover the mapping between inputs and outputs using a set of weight vectors. The algorithm was first introduced by Kohonen (1982, 1989 and 1990). Briefly, the algorithm for a two-dimensional square form of output units works as follows. The initiation step is to randomly generate a weight vector for output unit (i, j) at epoch t=0, and set parameters α (t), w (t). Then an unused input vector (claim pattern vector) X_p is selected, and an output unit (i_0, j_0) that minimizes the distance between X_p and the weight vectors. This output unit (i_0, j_0) is thus the best matching unit to the pattern X_p. Next, the weight vectors in the neighborhood of (i_0, j_0) are updated for the next epoch using parameters α (t) and w (t). Those weight vectors not in the neighborhood remain unchanged. Then, the algorithm goes to the next epoch and finds an unused pattern to repeat the iterative process. The process terminates at epoch t = T if the pre-specified stopping rule is satisfied. Otherwise, parameters α (t) and w (t) are decreased and the process is repeated until termination occurs at T.

Brockett et al. (1998) apply Kohonen's Feature Map to detect insurance fraud using a Massachusetts BI claims dataset with all binary indicators. They first show that the method can be applied where no dependent variable is available. With the set of binary BI claim vectors as the training sample, "valid" claims are mostly represented by a weight vector containing many zeros, while "strongly suspicious" claims are more likely represented by a vector with a large number of ones. Therefore, they define and calculate

a quantity S_m which measures the distance between the weight vectors and the null pattern. Then a high S_m will indicate that the output unit corresponds to a highly suspicious claim and vice versa. In this way, they are able to calculate S_m for each output unit and match each claim pattern to the feature map.

They also show that it is possible to incorporate expert opinions to categorize claims in the feature map. According to the algorithm described above, each pattern (claim) is matched with an output unit (totally 20*20 output units in their case) and the map locating these claims is drawn. Similar maps are drawn to display adjuster and investigator assessments. In the adjuster and investigator maps, the area where higher suspicion levels are clustered is considered to be where "strongly suspicious" claims are most likely to find their best-matching output units. In this way, an assessment of the suspicion levels of claims could be derived. The same procedure has been executed for both the training dataset and the holdout dataset. Not surprisingly, the overall pattern for the holdout dataset is not as clear as the one for the training dataset. This may arise from the imbalanced sampling and the inferiority of extrapolation. As a byproduct, the study also provides information toward evaluating the two sets of expert assessments. It demonstrates that this method provides satisfactory results and actually outperforms expert assessments.

Unsupervised neural networks are frequently seen in the literature of fraud detection in other fields as well, such as fraud detection in mobile phone use. For example, Burge and Shawe-Taylor (2001) propose a recurrent unsupervised neural network to form both short and long term statistical behavior profiles of mobile phone users and a fraud engine is then designed based on the behavior profiles. They base their

model on the technique introduced in Grabec (1991). Grabec's technique in its simplest form resembles Kohonen's self organizing map, but his method does not restrict the prototypes to a certain topological form.

As suggested in Brockett et al. (1998), it is worth noting that as an unsupervised learning method, unsupervised neural network can incorporate expert opinions, which are usually considered to be the privilege of some supervised learning methods. However, like many other methods in fraud detection, this method is computationally burdensome. Moreover, it is restricted to the information contained in the selected fraud indicators. While this disadvantage is not unique to this method, some of the other data mining methods are able to overcome it. Lastly, a further investigation of the optimal feature map categorization is needed for optimal results.

Other data mining methods have been employed in fraud detection as well. For example, Williams (1999) introduces a "Hot Spots" data mining method. It is essentially a genetic algorithm for classification. In applying the algorithm, the "hot spots", which are the sets of entities of particular interest to domain users, are identified by a three step process. K-means clustering is used at the first step to cluster a dataset to complete and disjoint clusters. Then the rule sets are created according to the rule induction algorithm. Lastly, the domain-specific mapping function is applied to identify "interesting" nuggets. After identifying "hot spots" in the dataset, the fraud investigators will then be able to focus just on the "interesting" cases and explore them more carefully. An Australian health claim dataset is used to illustrate the methodology. This method has several strengths and disadvantages. The "hot spots" method is actually a user (expert) guided system, which combines the advantages of data mining and the expert knowledge to lead

to potentially better results. Also, as an evolutionary method, it automatically provides a possibility to learn from the process itself. However, as is true for other data mining methods, the formula to capture "interestingness" is essential and needs further investigation for the best possible outcome.

There are still other methods applied in fraud detection (Kou et al., 2004). Outlier detection tries to separate outliers from the baseline distribution. These outliers are the "abnormal" points and are considered to have higher potential to be fraudulent, given that fraudulent cases are rare in a large dataset. Bolton and Hand (2001) apply this method to credit card fraud detection. Garvey and Lunt (1991) employ the model based reasoning method to detect computer intrusion, where the fraud behavior is identified by the accumulation of evidences of attacks. Visualization methods and graph theoretic approaches are applied to telecommunication fraud detection (Cox et al., 1997). For a complete survey of fraud detection methods, see Phua et al. (2005). For more references, also see Kou et al. (2004) and Bolton and Hand (2002).

One common problem associated with the methods mentioned above is that they are designed specifically for certain kinds of fraud and may not be directly usable in the case of insurance fraud. They might be modified to fit the nature of insurance fraud, but the literature has not seen the modifications up to this point.

1.5 HYBRID METHODS, THEORETICAL WORK AND OTHER ISSUES

1.5.1 Hybrid methods

Hybrids of supervised methods and hybrids of supervised and unsupervised methods are also used in fraud detection. Usually the unsupervised methods are used first to segment data into pre-clusters and supervised methods are then employed to refine the initial classification. This kind of hybrid methods is seen in Brockett et al. (1998), Williams and Huang (1997) and Williams (1999).

In general, researchers constantly face the choice of using supervised methods or unsupervised methods to solve specific problems. Although supervised methods are still widely used, unsupervised methods have become more and more popular in the recent decades. Compared with supervised methods, unsupervised ones tend to provide more robust results. Not restricted by model specifications, they give the largest freedom for the data to "speak for itself." More importantly, by employing the class of unsupervised methods, it is possible to discover patterns that were not found previously. Since the fraudsters are refining their fraud techniques very rapidly, this is of special importance in the field of fraud detection. Additionally, unsupervised methods become the only available methods when it is impossible or cost-prohibitive to obtain an initial training sample with known values for the dependent variable (e.g., when there is no training dataset available where the nature of the claims is known). The concerns of using unsupervised methods are that they usually do not perform as well as supervised methods since the models are not built using training samples, and sometimes they can not provide clear interpretation. Therefore, it might be wise to use unsupervised methods only when

the conditions are truly in favor of them. Hybrid methods that incorporate the strengths of both classes of methods have great potential and should be explored for better model learning and predictions.

1.5.2 Theoretical work

A body of literature is devoted to building theoretical models for insurance fraud and its detection. In these models, insurance fraud is usually seen as a special class of asymmetric information problems, i.e., ex-post moral hazard. Thus, this problem is placed in the larger context of contract theory and mechanism design. More specifically, different contract features of insurance policies are investigated to provide optimal incentives for limiting fraudulent claims.

Two theoretical approaches are proposed for insurance fraud detection and other related issues. The first one is an ex-post monitoring approach, namely, the costly state verification framework. Since the seminal work of Townsend (1979), many studies adopt the costly state verification approach to study this problem. In Townsend (1979), insurers can obtain private claim information for the insured at a cost (i.e., costly state verification). Under deterministic auditing strategies, Townsend shows that the optimal contract is a debt contract. Within the audit region, the insurer always audits and a deductible policy is preferable. Based on Townsend (1979), Mookherjee and Png (1989) find that random auditing is superior to deterministic auditing and debt contract is no longer optimal when random auditing is allowed. They show that in the non-audit zone, payments from an agent (e.g., insured) to the principal (e.g., insurer) is not a flat amount, but instead depend on the agent's realized income.

While still under the framework of costly state verification, Picard (2000) turns to examine the relationship between the insurer and its auditor. Deductible policies are found to be optimal with exogenous auditing cost. When insureds are able to manipulate auditing cost through collaboration with third parties, a coinsurance policy will be desired. Also under the assumption of endogenous auditing cost, Bond and Crocker (1997) claim the optimal contract to be one that overcompensates easy-to-monitor losses and undercompensates hard-to-monitor losses.

The above works implicitly or explicitly assume that the insurer is willing and able to commit to the pre-specified auditing strategy. However, other researchers are concerned with the case when there is no such commitment. Picard (1996) attacks this problem directly. Under costly state verification, expected utility of honest policyholders are reduced without insurer's commitment, which may lead to the failure of the system. He also proposes that an outside auditing agency may mitigate this problem.

The role of central agencies, insurance fraud bureaus (IFB), in investigating claims is explored by Boyer (2000a). By assuming that insurers cannot commit to audit ex-ante, Boyer finds that the benefit to use IFB depends on the auditing cost structure and policyholders are always better off if IFB conducts investigations at industry-average audit cost. Further, Dionne and Gagne (2002) were able to separate the ex-post moral hazard (opportunistic fraud) from other asymmetric information problems in insurance, such as ex-ante moral hazard and adverse selection. They show that the probability of fraud increases in automobile insurance with the special replacement cost endorsement.

Although studies differ in their assumptions and proposed optimal contracts, they seem to agree on a flattened payment schedule to alleviate insurance fraud. By contrast, a

counter-intuitive compensation schedule is also proposed, where higher claims are over-compensated and lower claims are under-compensated (Boyer, 2004). Boyer argues that the greater difference between the two states will increase insurers' incentive to audit and thus decrease insureds' incentive to defraud. The resulting reduction in fraud will save deadweight costs from auditing and penalties, and thus increases social welfare. Moreover, different forms of insurance taxes are discussed for their respective effectiveness in fighting fraud. When there is fraud, a benefit tax is preferred to a premium tax because the former imposes proportionally higher burden on fraudsters and thus is more efficient in reducing fraudulent claims (Boyer, 2000b, 2001).

Besides the costly state verification framework, another general approach is the costly state falsification framework introduced by Crocker and Morgan (1998), which is an ex-ante design of indemnification mechanism. They argue that when true losses are private information, falsification serves as a signal for the insured's type. In this case, falsification is necessarily induced by the optimal contract, which over-compensates small claims and under-compensates large claims. Crocker and Tennyson (2002) extend Crocker and Morgan (1998) from a first-party setting to a third-party environment. A similar flattened payment schedule is derived with systematic underpayment of claims, the extent of which varies by the difficulty of falsification. Loughran (2005) follows the model of Crocker and Tennyson (2002) but allows for different indemnification schedule for special damages and general damages. To deter fraud, less general damage is compensated relative to special damage as special damage claim increases. Lastly, a deductible policy is demonstrated to be non-optimal under costly state falsification. In

fact, higher deductibles may lead to more fraud in this situation (Dionne and Gagne, 2000).

Besides the two major approaches presented above, other mechanisms have been designed to alleviate insurance fraud. For example, Moreno et al. (2006) models fraud as ex-post adverse selection. A self selection contract, bonus-malus contract, is used to diminish the incentive for fraud by imposing the risk on risk-averse insureds in the form of premium uncertainty in the next period.

Finally, a first link of optimal auditing theory to practical auditing strategy, i.e. scoring, is provided by Dionne et al. (2005). They design an optimal auditing policy based on a suspicion index derived from "red flag" indicator variables. This policy may be explored further as a future research direction.

1.6 SUMMARY OF PRIOR RESEARCH

Supervised learning methods are abundant in the literature and most of them can be used in insurance fraud detection. However, they all rely on the assumption of an available initial training sample of insurance claims knowing the nature of these claims. This training sample is usually obtained by asking insurance investigators ("experts") to rate and classify these claims and can be very costly in terms of time, effort, expertise, money, and other resources. To make it even worse, this costly training sample can have different reliabilities depending, possibly, on different levels of investment, and even with sufficient investment, the quality of the training sample can still be questionable. This potential concern makes unsupervised learning methods a desirable set of candidates.

However, compared to supervised learning methods, unsupervised learning methods are sparse. Among the few available ones, most of them are unsatisfactory because they are not easily interpreted. The lack of interpretability can impede managerial insights and may cause problems in the presence of regulatory requirements (which is often the case in the insurance industry). This property can be especially harmful when researchers are interested not only in obtaining classification results from the learning methods but also in attempting to derive understanding of the underlying problem where interpretability is the key.

Therefore, this dissertation develops unsupervised (and hybrid) classification methodologies for insurance fraud detection based on the unsupervised PRIDIT method. The developed methods have good performance and provide results that are easily interpreted.

CHAPTER 2

The Discrete PRIDIT Method

2.1 INTRODUCTION TO RIDIT SCORING AND THE PRIDIT METHOD

Brockett et al. (2002) introduced a statistical based unsupervised method "PRIDIT (Principal Component Analysis of RIDITs)" for dichotomous classification tasks. The method was based on the "Relative to an Identified Distribution" (RIDIT) scoring system. Using RIDIT scoring, the qualitative responses of each claim on every variable (e.g., claimant appeared to be "claim-wise") are transformed into numerical variable scores, without altering the ranked nature of the data. Under RIDIT scoring, the variable score for an instance responding i on this variable is calculated as $\sum_{j<i} P_j + \frac{1}{2} P_i$, where $\{P_j\}$ is the probability of response i based on a reference distribution of the responses (Bross, 1958). These numerical variable scores are then used in the PRIDIT analysis to develop a binary classification method based on binary or general (ordinal) categorical predictor variables (i.e., the "discrete" PRIDIT method). The choice of scoring system in the PRIDIT method follows Brockett and Levine (1977) and Brockett (1981), where conditional mean scoring was identified for satisfying a set of desired properties of a scoring system, with RIDIT being a special case. RIDIT scoring is also demonstrated to have empirically superior performance relative to other scoring methods (Golden and Brockett, 1987). The development of the discrete PRIDIT method is reviewed in this chapter.

2.2 DEFINING THE VARIABLE SCORES

In the PRIDIT method, the definition of variable scores is a linear transformation of the original RIDIT score defined in Bross (1958). The empirical distribution of the entire sample is chosen to be the "reference distribution." Accordingly, the variable score for an instance responding i is defined for each predictor variable t as

$$B_i = \sum_{j<i} P_j - \sum_{j>i} P_j \ , \tag{1}$$

where P_i is the proportion of instances in category i. For example, assume that a binary predictor variable has three possible response values: $i = 0$, 1 or 2. Also assume that there are three instances in the dataset: instance A responds 0 on this variable, instance B responds 1, and instance C responds 2. Thus, the proportion of instances responding 0 is $P_0 = 1/3$, the proportion of instances responding 1 is $P_1 = 1/3$, and the proportion of instances responding 2 is $P_2 = 1/3$. According to the definition as in (1), variable score for an instance responding 0 is then $B_0 = \sum_{j<0} P_j - \sum_{j>0} P_j = 0 - (P_1 + P_2) = -2/3$; variable score for an instance responding 1 is then $B_1 = \sum_{j<1} P_j - \sum_{j>1} P_j = P_0 - P_2 = 0$; variable score for an instance responding 2 is then $B_2 = \sum_{j<2} P_j - \sum_{j>2} P_j = (P_0 + P_1) - 0 = 2/3$.

The thus defined variable scores possess a few desirable properties. First, the variable score B_i is bounded in [-1, 1]. Second, B_i is monotonically increasing in the rank of response i. Third, B_i is centered around zero.

It should be noted that before obtaining the scores, variables are first constructed in such a fashion that the rank of response categories bears a monotonic relationship with the likelihood of being in the event class. More specifically in the context of insurance

fraud detection, the predictor variables used in the PRIDIT analysis are assumed to have a monotonic relationship with the suspicion of the claim.

2.3 DEVELOPING THE DISCRIMINATORY POWER MEASURE

The PRIDIT method makes use of a discriminatory power measure A_t, which essentially describes each predictor's ability to distinguish between the fraud and non-fraud class or, in other words, the importance of each predictor variable. This measure is used to determine the weights assigned to each predictor to obtain a one-dimensional summative score for each claim. More specifically, A_t is defined as follows:

$$A_t = \sum_{i=1}^{k_t-1} \sum_{j>i} \left\{ \pi_{ti}^{(1)} \pi_{tj}^{(2)} - \pi_{ti}^{(2)} \pi_{tj}^{(1)} \right\},$$

where for q = 1 and 2, $\pi_t^{(q)} = \left(\pi_{t1}^{(q)}, ..., \pi_{tk_t}^{(q)} \right)$ is the multinomial response probability vector of variable t for class q (e.g., in insurance fraud detection, there are two classes: the fraud class and the non-fraud class), i = 1,…, K_t and j = 1,…, K_t are the indices for response categories of variable t. The quantity inside the brackets can be recognized as a 2 × 2 contingency table measure of association. From this definition, we can see that A_t measures the amount of dispersion between the two classes in the latent dimension. Under the above definition of variable score B_{ti} as shown in equation (1), it has been proved that E[B_t| Class 1] = (θ-1)A_t, where θ is the fraud rate (i.e., the percentage of fraudulent claims in the entire claims set) (Brockett et al. 2002).

2.4 VARIABLE WEIGHTS, SUMMATIVE SCORES AND CLASSIFICATION

The individual variable scores obtained in the manner just described are assigned a set of weights to form one-dimensional summative scores for the claims, which are used for classification. Equal weights are first used to get the initial summative scores, and these weights are then updated to take into account each variable's degree of "importance," measured by its consistency with the overall summative score. This updating process is executed iteratively and the weights are successively refined until convergence takes place.

More specifically, the initial summative score matrix $S^{(0)}$ is obtained by applying equal weights $\hat{W}^{(0)}$ (a vector of 1's) to the score matrix F (i.e., $\{B_{nt}\}$, the (n, t) th element of F is instance n's score on variable t), i.e., $S^{(0)} = F\hat{W}^{(0)}$. The initial weight vector is updated according to the individual variable scores' consistency with the summative scores: e.g., $\hat{W}^{(1)}=F^T S^{(0)} /||F^T S^{(0)}||$, where "|| ||" represents the Euclidian distance and the denominator serves as a normalization factor. In this way, the weights are iteratively updated until convergence is secured. The resulting limiting weights $\hat{W}^{(\infty)}$ converge to the first principal component of the squared score matrix $F^T F$. The variable weight for each variable turns out to be linearly related to each variable's discriminatory power measure.

Each claim's summative score using the limiting weights thus measures its potential of being in the fraud class. After the one-dimensional summative scores are obtained, the claims are classified according to the class distribution. When the class distribution is unknown, the classification can be based on the sign of the summative scores since the variable scores are centered around 0. For example, if the response categories of the variables are ranked so that lower response categories indicate higher

potential to be an "event" (i.e., fraud), instances with negative summative scores will be classified accordingly into the "event" class.

2.5 LIMITATIONS OF THE CURRENT METHOD AND CONCLUSION

Although the current PRIDIT method has been implemented using binary predictor variables, it is not restricted to categorical variables mathematically or computationally. In fact, as some of the predictor variables are by nature continuous variables (e.g. time between an accident and a filing of a claim), having only the current PRIDIT method available will force one to discretize the continuous predictors and may impede optimal learning. Thus, it is important to develop a unified method where all types of predictors can be incorporated into the learning process in a consistent manner. This development and the empirical evaluations are shown in Chapter 3 and Chapter 4. I also demonstrate empirically the advantage of the general PRIDIT method over the original discrete PRIDIT method in Chapter 3.

CHAPTER 3

Developing a Unified General PRIDIT Method

3.1 MOTIVATION FOR THE DEVELOPMENT

In this chapter, a unified PRIDIT method is proposed to incorporate both categorical and continuous predictors to optimize the classification performance. Both types of predictors are present in insurance fraud detection and many other applications. For example, in insurance fraud detection, a categorical predictor can be "whether there is a police report" and a continuous predictor can be "the time between an accident and the claim." In the context of market segmentation, a categorical predictor can be "the occupation of the customer" and a continuous predictor can be "the age of the customer." Continuous predictors as well as categorical predictors are desired to be incorporated in a unified classification method. Without such a method, one must convert these continuous variables to categorical (or binary) variables and then apply the discrete method. Moreover, even when the conversion is possible, valuable information may be lost in this process.

To avoid this possible loss, this chapter will present a unified development of a general PRIDIT method to apply to all types of predictor variables. Although the current PRIDIT method has been previously implemented using only binary predictor variables (Brockett et al. 2002), it need not be restricted to binary variables or to ordinal categorical variables, mathematically or computationally. In fact, despite the original motivation to treat qualitative variables, RIDIT scoring is also desired for continuous variables since

the transformed variables (still continuous) possess many preferred properties, as will be seen in later sections (also see Bross, 1958). The next sections present the theoretical derivation of the general (discrete and continuous) PRIDIT method.

3.2 DEFINING THE VARIABLE SCORE AND STOCHASTIC DOMINANCE ASSUMPTION

3.2.1 Definition of the variable scores

To develop the general PRIDIT method, we must first define the continuous analogue of the RIDIT based variable score B_{ti} from the ordinal categorical case. We follow the same variable construction rule--viz., lower responses on a variable indicates a larger potential of being in the event class than higher responses (note that "lower" and "higher" have meaning here since the variable is ordinal, i.e., it indicates the rank of categories or the numeric value). As will be shown later, the thus obtained variable scores possess desirable properties and the scales are unified among categorical and continuous variables. For a mixed dataset with both types of variables, this unification is important to facilitate the successive empirical studies and to ensure valid results.

To derive the analogue of B_{ti} in the continuous case, note that in the binary case (where there are only two response categories for each variable), variable scores (for any variable t) are defined respectively for response category 1 and response category 2 as $B_1 = -\hat{P}_2$, $B_2 = \hat{P}_1$, where $\hat{P}_1$ is the proportion of instances in response category 1 and $\hat{P}_2$ is the proportion of response 2 in the sample. (The index t for variable score B is omitted here since we are considering the calculation for any variable t. The same applies to the rest of the chapter whenever index t is omitted for B.) In the general ordinal categorical

case (where there are more than two ranked categories), the variable score for response category i is $B_i = \sum_{j<i} \hat{P}_j - \sum_{j>i} \hat{P}_j$, where $\hat{P}_j$ is the proportion of instances in response category j. Thus, B_i is defined as the proportion of instances that fall into response categories ranked lower than i minus the proportion that fall into higher ranked response categories.

In the continuous case, if x is a continuous variable with a monotonically decreasing relationship with the unobserved criteria variable (a higher value of this variables leads to event class membership), then by analogue with the discrete case, we define variable score $B(x)$ as the proportion of instances with respondent value less than x minus the proportion of instances with respondent value larger than x. Thus, if $\hat{F}(x)$ is the empirical distribution function of a continuous predictor variable x, then

$$B(x) = \hat{F}(x^-) - [1 - \hat{F}(x)] = [\hat{F}(x) - \hat{P}(x)] - [1 - \hat{F}(x)],$$

or, more succinctly, we have the following definition.

Definition 1. Variable score $B(x)$ for a continuous variable with response value x is $B(x) = 2\hat{F}(x) - 1 - \hat{P}(x)$, where $\hat{F}(x)$ is the empirical distribution of x and $\hat{P}(x)$ is the sample proportion of response x. Thus, variable score for an instance with response value x is $B(x)$.

As thus defined for the continuous case, variable score $B(x)$ preserves the desirable characteristics in the ordinal categorical case, i.e., $B(x)$ is bounded in [-1, 1] so by this scoring method, all variables (whether continuous or discrete) are measured on the same scale. Moreover, $B(x)$ is monotonically increasing and centered around zero:

$$
\begin{aligned}
&\mathrm{E}_{\hat{p}}\left[B(x)\right] \\
&= \sum_{k=1}^{K} B(x_k)\hat{P}(x_k) = \sum_{k=1}^{K}\left[2\hat{F}(x_k) - 1 - \hat{P}(x_k)\right]\hat{P}(x_k) \\
&= 2\sum_{k=1}^{K}\left\{\left[\sum_{l=1}^{k}\hat{P}(x_l)\right]\hat{P}(x_k)\right\} - 1 - \sum_{k=1}^{K}\left[\hat{P}(x_k)\right]^2 \\
&= 2\left\{\sum_{k=1}^{K}\left[\hat{P}(x_k)\right]^2 + \sum_{k=2}^{K}\left[\hat{P}(x_k)\sum_{l=1}^{k-1}\hat{P}(x_l)\right]\right\} - \sum_{k=1}^{K}\left[\hat{P}(x_k)\right]^2 - 1 \\
&= \left[\sum_{k=1}^{K}\hat{P}(x_k)\right]^2 - 1 = 0\,,
\end{aligned}
$$

where variable x takes on an (increasingly ranked) value $x_1, \ldots, x_K$ in the sample.

The "monotonic relationship" assumption used above for variable construction suggests a (first order) stochastic dominance relationship between the two classes of instances, which we formally justify in the next subsection before moving on to the next step of the development.

3.2.2 Stochastic dominance assumption in variable construction

By choice of variables and by construction, the individual predictor variables exhibit a first order stochastic dominance relationship between the two classes (event class, non-event class). Specifically, let $F_1(\cdot)$ be the distribution function of some variable t for class 1 (event class) and $G_2(\cdot)$ be the distribution function for class 2 (non-event class). Recall that variables are constructed so that smaller values suggest higher potential of an event. This is equivalent to saying that $G_2(\cdot)$ first-order stochastically dominates $F_1(\cdot)$, or $\Delta(x) = F_1(x) - G_2(x) \geq 0$, since for a given x, the probability that a response (for some variable t) randomly selected from class 1 is less than x is always higher than the probability that a response randomly selected from class 2 is less than x.

In fact, our intuitive variable construction assumption that the event class tends to score lower on predictor variables than the non-event class is the consequence of a stochastic dominance assumption, as shown in,

Proposition 1. Suppose for some variable *t*, response X is randomly selected from the set of class 1 instances, i.e., X ~ F_1 (·), and response Y (independent of X) is randomly selected from the set of class 2 instances, i.e., Y ~ G_2 (·). If G_2 (·) first-order stochastically dominates F_1 (·), then $P(X < Y) \geq \frac{1}{2}$.

Proof: We use elementary probability and calculus to prove this proposition. Denote $\Delta(y) = F_1(y) - G_2(y)$ and briefly,

$$\begin{aligned} & P(X < Y) \\ &= \int_{-\infty}^{\infty} P(X < y) dG_2(y) = \int_{-\infty}^{\infty} F_1(y) dG_2(y) \\ &= \int_{-\infty}^{\infty} [G_2(y) + \Delta(y)] dG_2(y) \\ &= \frac{1}{2} + \int_{-\infty}^{\infty} \Delta(y) dG_2(y) \geq \frac{1}{2}. \end{aligned}$$

The inequality is established because $\Delta(y) \geq 0$ due to stochastic dominance.

Proposition 1 formally shows that the stochastic dominance relationship leads to the conclusion that a randomly selected class 1 member is more likely to respond lower than a randomly selected member of class 2. This confirms that the variable construction is equivalent to an underlying stochastic dominance assumption between the two classes.

3.3 DEVELOPING THE DISCRIMINATORY POWER MEASURE AND VARIABLE WEIGHTS

The next step of the development is to define A_t, the discriminatory power measure for variable *t*. As briefly described for the discrete case, A_t essentially measures how well the particular variable is able to differentiate the two classes along the latent dimension. If instances in the event class all tend to respond in lower ranked categories of variable *t* and instances in the non-event class all tend to take values in higher ranked categories of this variable, then variable *t* is considered to discriminate well, as reflected by a high value of A_t. Next we will formally develop an approach to define A_t by calculating the conditional mean of the variable score for the event class (class 1).

Take a (random) sample of size *N* consisting of N_1 instances from class 1 and N_2 instances from class 2. The expected proportion θ of class 1 is θ =E [N_1 / N], where N_1 ~Binomial (N, θ). Then the empirical distribution $\hat{F}(x)$ for the entire sample is based on the theoretical distribution $F(x) = \theta F_1(x) + (1-\theta)G_2(x)$, where $F_1(x)$, $G_2(x)$ are the variable distribution functions of class 1 and class 2 respectively. (Again, we are discussing the general case for any variable *t*. The subscript *t* for the distributions is omitted here and in the remainder of this subsection.) Again denote $\Delta(x) = F_1(x) - G_2(x)$. Rewrite $F(x) = F_1(x) - (1-\theta)\Delta(x)$. Then the expected variable score of variable *t* for a member of class 1 can be calculated as:

$$
\begin{aligned}
&\mathrm{E}[B_t \mid \text{Class 1}] \\
&= \int_{-\infty}^{\infty} E[B(x)]dF_1(x) = \int_{-\infty}^{\infty} E[2\hat{F}(x) - 1 - \hat{P}(x)]dF_1(x) \\
&= \int_{-\infty}^{\infty} [2F(x) - 1]dF_1(x) = 2\int_{-\infty}^{\infty} [F_1(x) - (1-\theta)\Delta(x)]dF_1(x) - 1 \\
&= -2(1-\theta)\int_{-\infty}^{\infty} \Delta(x)dF_1(x) = 2(\theta - 1)\int_{-\infty}^{\infty} \Delta(x)dF_1(x).
\end{aligned} \tag{2}
$$

Definition 2. The discriminatory power measure A_t (for variable t) is defined by $A_t = 2\int_{-\infty}^{\infty} \Delta(x)dF_1(x) = 2\int_{-\infty}^{\infty} \Delta(x)dG_2(x)$.

Thus by (2), we have

$$\mathrm{E}[B_t| \text{ Class 1}] = (\theta\text{-}1)A_t \ . \tag{3}$$

It can be easily shown that $2\int_{-\infty}^{\infty} \Delta(x)dF_1(x) = 2\int_{-\infty}^{\infty} \Delta(x)dG_2(x)$, by applying integration by parts and simply noting that the product of $F_1(\cdot)$ and $G_2(\cdot)$ is still a cumulative distribution function, so

$$
\begin{aligned}
&2\int_{-\infty}^{\infty} \Delta(x)dF_1(x) \\
&= 2\int_{-\infty}^{\infty} [F_1(x) - G_2(x)]dF_1(x) = 1 - 2\int_{-\infty}^{\infty} G_2(x)dF_1(x) \\
&= 1 - 2[F_1(x)G_2(x)]\,|_{-\infty}^{\infty} + 2\int_{-\infty}^{\infty} F_1(x)dG_2(x) = 2\int_{-\infty}^{\infty} F_1(x)dG_2(x) - 1 \\
&= 2\int_{-\infty}^{\infty} F_1(x)dG_2(x) - 2\int_{-\infty}^{\infty} G_2(x)dG_2(x) = 2\int_{-\infty}^{\infty} \Delta(x)dG_2(x).
\end{aligned}
$$

A similar calculation yields

$$\mathrm{E}[B_t| \text{ Class 2}] = 2\theta\int_{-\infty}^{\infty} \Delta(x)dG_2(x) = \theta A_t. \tag{4}$$

Thus, the definition of A_t in the continuous case can be derived from calculating the conditional mean score of either class and is equivalent to that in the discrete case. To interpret A_t as "discriminatory power" in the continuous case, note that the integrand *Δ(x)*

measures how "spread out" the variable distributions $F_1(x)$ and $G_2(x)$ are from each other.

The discriminatory power measure A_t plays an important role in our method. As will be shown below, A_t is linearly related to W_t, the estimate $\hat{W}_t$ of which is the limiting variable weight assigned to each individual variable score B_t to form the one dimensional summative score S_n for the instances. A variable t which effectively discriminates instances from the two different classes, as reflected by a higher A_t, is more useful in classification. Hence intuitively, it should have a higher weight W_t. Bearing this in mind, we show in the next subsection how the weights and the overall scores are determined.

3.4 OBTAINING SUMMATIVE SCORES AND CLASSIFICATION

As described in Chapter 2, starting from equal weights, variable weights are iteratively updated until convergence is achieved to the first principal component of the squared variable score matrix. In order to obtain valid summative scores, we need to show that the above procedure converges and consistent estimates of weights exist. This is essentially to prove Theorem 1 in Brockett et al. (2002) in the general case. We first re-state the Theorem here.

Theorem 1. The sequences of predictor variable weights $\{W^{(n)}\}$ and overall summative instance scores $\{S^{(n)}\}$ converge. Moreover, the limiting predictor variable weight $\hat{W}^{(\infty)}$ is the first principal component of $F'F$, which is a consistent estimate of the principal component $W^{(\infty)}$ of E[$F'F$], the tth component of which is explicitly

$$W_t^{(\infty)} = \frac{A_t}{(\mu_1 - U_{tt})\sqrt{\sum_{s=1}^{m} A_s^2 / (\mu_1 - U_{ss})^2}},$$

where μ_1 is the largest eigenvalue of E[$F'F$] and $U_{tt} = N_1\sigma_{1t}^2 + N_2\sigma_{2t}^2$ is the "uniqueness component of variance" in a single-factor analytic model (for q=1 and 2 class, σ_{qt}^2 =variance(B_{qt}), and B_{qt} is the score of a randomly selected instance from class q on variable t).

The way B_t and A_t are defined for the continuous case ensures that the proof follows the same as in the discrete case. (The detailed proofs in the discrete case can be found in Brockett et al. 2002.) It is clear from Theorem 1 that the predictor variable weights $W_t^{(\infty)}$ is proportional to A_t /(μ_1-U_{tt}), validating the intuition that variables with higher discriminatory power should be assigned higher weights.

We define the one-dimensional summative score S_n as follows,

Definition 3. The one-dimensional summative scores for the instances are defined as $S = F\hat{W}^{(\infty)}$, where S is the vector of summative scores for the n instances, F (i.e., $\{B_{nt}\}$) is the variable score matrix with the (n, t) th element being instance n's score on variable t (i.e., $B(x)$, if instance n responds x on variable t), and $\hat{W}^{(\infty)}$ is the vector of limiting variable weights. Thus for instance n, the summative score $S_n = \sum_{t=1}^{m} \hat{W}_t^{(\infty)} B_{nt}$.

The summative scores as defined in Definition 3 are used to carry out the main task of classification by the PRIDIT method. If θ, the proportion of class 1(event class) instances, is known or assumed a priori, we can rank the instances ascendingly by their summative scores and classify the first $N\theta$ instances to class 1, leaving the rest to class 2. If the proportion θ is unknown, which is the case in many applications, a uniform rule is

adopted to assign instances with negative scores to class 1 and instances with positive scores to class 2.

3.5 CONNECTIONS BETWEEN BINARY, CATEGORICAL AND CONTINUOUS CASES

In this section, we relate the discriminatory power measure A_t in the binary, categorical, and continuous cases to further illustrate this important measure and the connections between different cases. In the binary case where only two possible categories of responses exist, (for any variable t) assume class 1 (event class) has the probability vector (p_1, p_2) for the two responses and class 2 (non-event class) has the probability vector (q_1, q_2) (where for i=1 and 2, p_i and q_i are the probability of responding in category i of some variable t for instances in class 1 and class 2, respectively). Consequently, the empirical distribution ($\hat{P}_1, \hat{P}_2$) (where for i=1 and 2, $\hat{P}_i$ is the sample proportion of responses in category i) for a sample with parameter θ is based on the theoretical distribution ($\theta p_1 + (1- \theta)q_1$, $\theta p_2 + (1- \theta)q_2$). By definition, we can calculate variable scores for responses in category 1 and category 2 as $B_{t1}= -\hat{P}_2$ and $B_{t2} = \hat{P}_1$. Thus, a response in category 1 of variable t has expected variable score $E[B_{t1}] = - [\theta p_2 + (1- \theta)q_2]$ and a response in category 2 of variable t has expected variable score $E[B_{t2}] = \theta p_1 + (1- \theta)q_1$. Then

$$
\begin{aligned}
&E[B_t| \text{ Class 1}] \\
&= p_1 \{- [\theta p_2 + (1- \theta)q_2]\} + p_2 [\theta p_1 + (1- \theta)q_1] \\
&= (\theta\text{-}1) (p_1 q_2 - q_1 p_2) .
\end{aligned}
$$

From the stochastic dominance as given in Proposition 1, we can write $p_1 = q_1 + \delta$, $p_2 = q_2 - \delta$, where $\delta > 0$. Then it follows, E[B_t| Class 1] = (θ-1)A_t, where $A_t = \delta$ and δ measures the dispersion of the response probability vectors between the two classes.

For the ordinal categorical case where more than two categories exist, as established by Brockett et al. (2002),

$$A_t = \sum_{i=1}^{k_t-1} \sum_{j>i} \left\{ \pi_{ti}^{(1)} \pi_{tj}^{(2)} - \pi_{ti}^{(2)} \pi_{tj}^{(1)} \right\}, \tag{5}$$

where for q=1 and 2, $\pi_t^{(q)} = \left(\pi_{t1}^{(q)}, ..., \pi_{tk_t}^{(q)} \right)$ is the multinomial response probability vector of variable t for class q, i=1,…, K_t and j =1,…, K_t are the indices for response categories of variable t. The quantity inside the brackets can be recognized as the 2 × 2 contingency table measure of association. Hence, again, A_t measures the amount of dispersion between the two classes in the latent dimension. Under the definition of variable score B_{ti}, it has been proved that E[B_t| Class 1] = (θ-1)A_t.

In the continuous case, rewrite A_t in Definition 2 as

$$
\begin{aligned}
& A_t \\
& = 2\int_{-\infty}^{\infty} \Delta(x) dF_1(x) \\
& = \int_{-\infty}^{\infty} \left[F_1(x) - G_2(x)\right] dF_1(x) + \int_{-\infty}^{\infty} \left[F_1(x) - G_2(x)\right] dF_1(x) \\
& = 1 - \int_{-\infty}^{\infty} G_2(x) dF_1(x) - \int_{-\infty}^{\infty} G_2(x) dF_1(x) \\
& = \int_{-\infty}^{\infty} 1 dF_1(x) - \int_{-\infty}^{\infty} G_2(x) dF_1(x) - \int_{-\infty}^{\infty} G_2(x) dF_1(x) \qquad (6) \\
& = \int_{-\infty}^{\infty} \left[1 - G_2(x)\right] dF_1(x) - \left[F_1(x) G_2(x) \Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} F_1(x) dG_2(x) \right] \\
& = \int_{-\infty}^{\infty} \left[1 - G_2(x)\right] dF_1(x) - \left[1 - \int_{-\infty}^{\infty} F_1(x) dG_2(x) \right] \\
& = \int_{-\infty}^{\infty} \left[1 - G_2(x)\right] dF_1(x) - \int_{-\infty}^{\infty} \left[1 - F_1(x)\right] dG_2(x) \\
& = \int_{-\infty}^{\infty} \int_{x}^{\infty} f_1(x) g_2(u) du dx - \int_{-\infty}^{\infty} \int_{x}^{\infty} g_2(x) f_1(u) du dx \,,
\end{aligned}
$$

where $f_1(x)$ and $g_2(x)$ are density functions (for any variable *t*) of class 1 and class 2 respectively. Note that the fifth equality in (6) is derived by applying integration by parts and for the sixth equality, note again that the product of $F_1(\cdot)$ and $G_2(\cdot)$ is still a distribution function. By changing integrations to summations and densities to probability vectors in the last equality in (6), it is seen immediately that the continuous case (6) replicates the discrete case (5) in the definition of A_t.

Moreover, as in the discrete case, A_t in the continuous case has the properties stated in Proposition 2 below.

Proposition 2. In the continuous case, the discriminatory power measure A_t is bounded between [0, 1], with the right end boundary point (A_t = 1) suggesting perfect separation of the sample based on variable *t* and the left end boundary point (A_t = 0) suggesting no separation based on variable *t*.

Proof: To see A_t is bounded in [0, 1] , first note that from the third equality in (6), $A_t = 1 - 2\int_{-\infty}^{\infty} G_2(x)dF_1(x)$, where $0 \le F_1(x) \le 1$, $0 \le G_2(x) \le 1$, $f_1(x) \ge 0$, $g_2(x) \ge 0$.

Thus $\int_{-\infty}^{\infty} G_2(x) f_1(x)dx \ge 0$ leads to $A_t \le 1$. Moreover, rewrite A_t in Definition 2,

$$\begin{aligned} A_t &= 2\int_{-\infty}^{\infty} \Delta(x)dG_2(x) \\ &= 2\int_{-\infty}^{\infty} (F_1(x) - G_2(x))dG_2(x) \\ &= 2\int_{-\infty}^{\infty} F_1(x)dG_2(x) - 1. \end{aligned}$$

From the proof of Proposition 1, $\int_{-\infty}^{\infty} F_1(x)dG_2(x) \ge \frac{1}{2}$ due to the stochastic dominance assumption, which leads to $A_t \ge 0$.

To see perfect separation corresponds to the right end point, let x^* be the separation value, i.e., all class 1 members respond less than x^* and all class 2 members respond greater than x^*, or $F_1(x^*) = 1$ and $G_2(x^*) = 0$ .This corresponds to the special case in Proposition 1 where $P(X < Y) = 1$. Thus, we have from the last equality in the proof for Proposition 1, $\int_{-\infty}^{\infty} \Delta(x)dG_2(x) = \frac{1}{2}$, and then $A_t = 2\int_{-\infty}^{\infty} \Delta(x)dG_2(x) = 1$.

Similarly, to see that no separation corresponds to the left end point, note that we now have the special case in Proposition 1 where $P(X < Y) = \frac{1}{2}$. Then from the proof in section 3.2.2 for Proposition 1, we have $\int_{-\infty}^{\infty} \Delta(x)dG_2(x) = 0$, i.e., $A_t = 0$.

Thus far, we have developed the PRIDIT method to incorporate continuous predictor variables by closely following the discrete case. However, the method lends itself to the extension after the groundwork is laid to define the variable score B_t and

establish the A_t measure for the continuous case. As a result, all the desired features are preserved and the method is thus validated.

3.6 INTERPRETATIONS OF THE PRIDIT METHOD (RELATIONS TO WILCOXON RANK SUM STATISTIC)

In this subsection, an interpretation of the PRIDIT method is provided by establishing the connection between the PRIDIT method and the Wilcoxon Rank Sum Statistic. As Bross (1958) first pointed out, the RIDIT score is related to the Wilcoxon rank sum statistic. This is encouraging since the Wilcoxon rank sum statistic is traditionally used as a (nonparametric) statistical measure of the deviation between two classes (on a particular variable), and has established statistical properties. In this section, we will provide the connection between the discriminatory power measure A_t and the expected Wilcoxon rank sum statistic $E[W_t^*]$, which helps justify A_t as a measure of variable t's power to differentiate the two classes and thus provides a rationale for our method. This is done through linking the average PRIDIT variable score for each class to the Wilcoxon rank sum statistic of the same class in both the discrete case (a version of this was first shown by Selvin 1977) and the continuous case.

First consider the case when the predictors are all categorical variables. For an instance responding in category i on variable t, by definition,

$$
\begin{aligned}
& B_{ti} \\
&= \sum_{j<i} \hat{P}_j - \sum_{j>i} \hat{P}_j \\
&= \frac{1}{N}\left[y_1 + \ldots + y_{i-1} - (y_{i+1} + \ldots + y_{k_t})\right] \\
&= \frac{1}{N}\left[y_1 + \ldots + y_{i-1} - (N - y_1 - \ldots - y_{i-1} - y_i)\right] \\
&= \frac{1}{N}\left(2y_1 + \ldots + 2y_{i-1} + y_i - N\right) \\
&= \frac{2}{N}(y_1 + \ldots + y_{i-1} + \frac{y_i}{2}) - 1,
\end{aligned}
$$

where for i =1, …, K_t categories, y_i is the total number of instances in category i of variable t from the entire sample of N instances. The relative rank (with ties) of an instance responding in category i among all instances in the sample is $Y_{ti} = y_1 + \ldots + y_{i-1} + \frac{y_i + 1}{2}$. It follows immediately that $B_{ti} = \frac{2Y_{ti}}{N} - 1 - \frac{1}{N}$, or

$$
Y_{ti} = \frac{N}{2} B_{ti} + \frac{N+1}{2} \ . \tag{7}
$$

Now, we will relate the average score $\overline{B}_t^{(q)}$ (q=1 and 2 class) to the Wilcoxon rank sum statistic W_{tq}^* (q=1 and 2 class). For class 1, the average score for variable t is $\overline{B}_t^{(1)} = \frac{\sum_{i=1}^{K_t} x_{ti}^{(1)} B_{ti}}{N_1}$, where for i =1, …, K_t categories, $x_{ti}^{(1)}$ is the number of class 1 instances in category i of variable t and N_1 is the number of instances in class 1. By definition of the Wilcoxon rank sum statistic (computed from class 1),

$$
\begin{aligned}
& W_{t1}^{*} \\
&= \sum_{i=1}^{k_t} x_{ti}^{(1)} Y_{ti} = \sum_{i=1}^{k_t} x_{ti}^{(1)} \left(\frac{N}{2} B_{ti} + \frac{N+1}{2}\right) \\
&= \frac{N}{2} \sum_{i=1}^{k_t} x_{ti}^{(1)} B_{ti} + \frac{N+1}{2} \sum_{i=1}^{k_t} x_{ti}^{(1)} \\
&= \frac{N}{2} \left(N_1 \overline{B}_t^{(1)}\right) + \frac{N+1}{2} N_1 .
\end{aligned}
$$

which leads to

$$
\overline{B}_t^{(1)} = \frac{2W_{t1}^{*}}{NN_1} - 1 - \frac{1}{N} . \tag{8}
$$

Recall that the expected variable score for an instance in class 1, E[B_t| Class 1] = E[$\overline{B}_t^{(1)}$], is proportional to the discriminatory power A_t up to a factor of (θ – 1). Consequently, we have

Proposition 3. E[W_{t1}^{*}] is linearly related to A_t in the discrete case.

Proof: This linear relationship is established from Definition 2 and (8).

The same relationship can also be established for class 2, i.e., $\overline{B}_t^{(2)} = \frac{2W_{t2}^{*}}{NN_2} - 1 - \frac{1}{N}$.

Corollary 1. E[W_{t2}^{*}] is linearly related to A_t in the discrete case.

Thus, A_t is a measure of the dispersion between the two classes on a particular variable t, as is the Wilcoxon rank sum statistic.

The above connections easily extend to the continuous case. For any variable t, note that an instance with response x has variable score $B(x) = 2\hat{F}(x) - 1 - \hat{P}(x)$. The relative rank of this instance among all N instances can be calculated as

$$
\begin{aligned}
& Y_x \\
&= N\hat{P}(X \le x) - \frac{N\hat{P}(x) - 1}{2} \\
&= N\hat{F}(x) - \frac{N\hat{P}(x)}{2} + \frac{1}{2} \\
&= \frac{N}{2} B(x) + \frac{N+1}{2},
\end{aligned}
$$

which is identical to (7) in the discrete case when we substitute response category *i* with (increasingly ranked) response value x_i. The connections thus follow accordingly:

Corollary 2. E[W_{t1}^*] and E[W_{t2}^*] are linearly related to A_t in the continuous case.

The linear relationship found between the expected Wilcoxon rank sum statistic E[W_t^*] and the discriminatory power A_t also reveals the nonparametric nature of our method. In fact, the "distribution-free" feature of RIDIT analysis was first discussed in Bross (1958). In the setting of the PRIDIT method, clearly no functional form is imposed on the distribution of the underlying data. Instead, we make use of the empirical distribution, which is just the proportional division exhibited by the sample. Therefore, our method works with any distributions and different underlying theoretical distributions will not bias the results, i.e., the PRIDIT method also shares with Wilcoxon rank sum statistic the merit of robustness.

3.7 EMPIRICAL DEMONSTRATIONS

In this section, the general PRIDIT method is implemented and the performance is assessed. Perhaps due to the inability of effectively incorporating continuous predictors by current insurance fraud detection methods, the insurance fraud dataset we obtained

consists of only binary predictors. Thus, we demonstrate the general PRIDIT method in an income classification dataset with a mixture of continuous and categorical predictors. This income dataset contains known classifications allowing us to assess the performance of our method against the true classifications. The income classification dataset is used only for demonstration purposes. The general PRIDIT method can be applied similarly to insurance fraud detection when the dataset becomes available, or to a large range of other classification tasks in business or non-business applications.

3.7.1 Data description

The ADULT database (a census income dataset extracted from Current Population Survey) from UCI Machine Learning Repository is used to illustrate the methodology (for details see http://www.ics.uci.edu/~mlearn/MLSummary.html). This dataset aims at a binary classification task (income classification), which is of potential interest to marketers for market segmentation purposes. It is chosen for our analysis since it has a combination of continuous and categorical variables, and has been previously tested using different algorithms (for details, see ftp://ftp.ics.uci.edu/pub/machine-learning-databases/adult/adult.names). Previous work validates the usability of this database and provides a benchmark for our empirical tests.

This dataset contains 45,222 instances after removing those containing missing predictor values. It has been divided to a training dataset and a test dataset by a 2/3 (30,162 instances), 1/3 (15,060 instances) random split. Fourteen census variables are used as indicator variables, six of which are continuous and eight of which are categorical. The continuous variables range over the entire space permitted by the nature

of each particular variable, and the categorical variables have 2 to more than 30 categories. A description of the indicator variables is presented in Table 1. True income classes of individuals are also available in this dataset.

Table 1 Variable Description

Variable Name	Type	Permitted Value	Include in the Analysis?
Age	Continuous	/	Y
Education (Number of Years)	Continuous	/	N
Capital Gain	Continuous	/	Y
Capital Loss	Continuous	/	Y
Working Hours per Week	Continuous	/	Y
(Sampling Frame) Final Weight[a]	Continuous	/	N
Work Class	Categorical	8 Categories	Y
Education	Categorical	16 Categories	Y
Marital Status	Categorical	7 Categories	Y
Occupation	Categorical	14 Categories	Y
Relationship	Categorical	6 Categories	Y
Race	Categorical	5 Categories	Y
Sex	Categorical	2 Categories	Y
Native Country	Categorical	41 Categories	Y

Note: This is a brief description of the dataset. More detailed variable definitions can be found in the data dictionary for the Current Population Survey from the U.S. Census Bureau. http://www.census.gov/cps/.

[a]This variable is used in the sampling process to construct the dataset and is not related to income classification.

3.7.2 Data preparation for the PRIDIT analysis

In order to perform PRIDIT analysis, response categories for each categorical indicator variable are first rank ordered in decreasing propensities of being in the event class, which is designated to be the higher income class (annual income exceeding $50,000). The relationship between continuous variables and the class membership is also examined. Continuous variables and categorical variables are treated separately as explained below.

The initial analysis includes five out of six continuous variables. The variable *education (Number of Years)* (education categories represented by number of years of education) is excluded to avoid duplication with the categorical variable *education*, which also captures the nature of this characteristic. Univariate relationships between the binary dependent variable (income class membership) and continuous indicator variables are examined using logistic regressions. The parameter estimates and their significance are reported in Table 2. We find that all variables except *final weight* have positive coefficients and are significant at 1%. The variable *final weight* is exogenous to the survey and is insignificant at 10%. Thus, it is excluded from further analysis.

Table 2 Parameter Estimates and Significance for Continuous Indicator Variables

Variable	Parameter Estimates	P-Value
Age***	0.0423	<0.0001
Final Weight	< -0.0001	0.1198
Capital Gain***	0.0003	<0.0001
Capital Loss***	0.0007	<0.0001
Working Hours per Week***	0.0469	<0.0001

***indicates significance at 1%.

The eight categorical variables demand a somewhat more complicated treatment. As continuous variables, univariate relationships with the dependent variable are examined for each of the categorical variable by means of logistic regressions. For each variable, a reference response category is chosen and every other category is compared to this reference category on the likelihood of leading to the membership in the event class by odds ratio estimates (against the reference category), thus ranking all categories (including the reference one) monotonically in decreasing propensity of being assigned to

the event class. All of the categorical variables are found significant at 1% in the univariate regressions. Most of the resulting category rankings are fairly intuitive.

3.7.3 The PRIDIT analysis and results on the training dataset

As explained in section 3.2, an initial step to obtain PRIDIT classification is to calculate all individual predictor variable scores for each observation. More specifically, we first obtain the empirical distribution (the proportions of responses) for each variable from the dataset. After the responses for a particular variable (continuous or categorical) are ordered in decreasing likelihood of event class membership, for every specific instance (person) we calculate the proportion of all lower ranked respondents (more event prone) minus the proportion of all higher ranked respondents (less event prone). Recall Definition 1, if one instance responds as x (for categorical variables, x represents the category rank) on variable t, then calculate the variable score $B(x) = \hat{F}(x^-) - [1 - \hat{F}(x)]$, where $\hat{F}(x^-) = \hat{P}(X < x)$ is the proportion of instances responding lower than x on variable t and $[1 - \hat{F}(x)]$ is the proportion higher than x on variable t (by using proportions, this formula applies to both continuous and categorical variables). Recall that the summative scores are produced by a weighted sum of these individual variable scores, with the weights determined in the fashion described in Section 3. By the robust uniform classification rule based on the sign of the summative scores, individual persons with negative one-dimensional summative scores are assigned into the high income class (event class) and individual persons with positive summative scores are assigned to the low income class (non-event class). Note that the above procedure applies to both

continuous and categorical indicator variables in a unified fashion, after they are properly rank ordered.

As a result of the PRIDIT analysis on the training dataset, 14,667 out of 30,162 individuals (49%) are classified as high-income and 15,495 individuals (51%) as low-income. Detailed classification results are shown in Table 3. When compared with the known true class membership for assessment, we can see that PRIDIT is successful in assigning 20,365 out of 30,162 individuals into their true income class, leading to a hit rate of 68%. More importantly, PRIDIT exhibits superior performance in classifying events, i.e., the high income class membership. Out of 7,508 high income individuals, 88% of them are identified correctly by PRIDIT. It is worth noting that the high income class, which accounts only for 25% of the training sample, is the minority class in this dataset. Therefore, unlike most methods designed for classification, PRIDIT performs better for the minority class, which coincides with the event class in most cases.

Table 3 PRIDIT Classification against True Classification for the Training Dataset

	True Low Income	True High Income	PRIDIT Total
PRIDIT Low Income	13762	905	14667
PRIDIT High Income	8892	6603	15495
True Total	22654	7508	30162

Note: [Hit rate: 67.52%; Event class accuracy: 87.95%; Non-event class accuracy: 60.75%].
True classification is known for this dataset and used to evaluate performance. Also, in this table (and in all following tables of this type), accuracy is used as the main criterion to assess performance for the purpose of demonstrating the methodology. Other criteria, such as a cost-related one, may be more appropriate and can be employed. However, by examining event class and non-event class accuracies respectively, we implicitly account for possibly different costs of misclassification types and evaluation under those costs, i.e., if the relative costs of type I and type II error are known, then a cost based criterion could be calculated from the numbers in Table 3.

This "focus" on the event class is beneficial in the following ways. The identification of the event class is usually the ultimate goal of an analysis (for example, selecting high income individuals for marketing purposes in the adult dataset or identifying fraudulent claims in fraud detection) and thus correctly identifying event class instances is more typically the focus of the analysis. There are also cost-related benefits since misclassification cost is usually higher for the event class than for the non-event class (i.e., type I and type II errors are not equally important). For example, in the case of fraud detection, wrongly classifying a non-fraudulent claim to the fraud class may result in unnecessary auditing cost (and perhaps ill will), while missing a true fraudulent claim could cost a substantial unnecessary claim payment to be made by the insurance company. Also in the context of marketing, a much higher loss of profit can be experienced by misidentifying a high income individual than loss of a small promotional expense.

3.7.4 Evaluating PRIDIT in the test dataset

As an unsupervised method, PRIDIT does not require or utilize knowledge of the dependent variable, so it is possible to assess performance directly on the training dataset by comparisons with the true labels of the dependent variable, if they are available. Thus results such as shown in Table 3 are included to exhibit a form of external validity for the PRIDIT procedure. However, using a test dataset allows us to evaluate the stability of the variable scores and variable weights, which affects the applicability of the method in random additional samples. This subsection introduces the empirical procedure of applying trained PRIDIT on the test dataset.

To accomplish this task, PRIDIT is first executed on the training dataset as described before. The rank orders of different response categories and the resulting variable score for each response on each variable (categorical or continuous) are maintained. These variable scores are then assigned directly to every individual in the test dataset according to its respective response on each variable. The variable scores for each individual in the test dataset are next transformed to summative scores by using the weights obtained in the training step. As for the training data, individuals with negative summative scores are assigned to the high income class, while individuals with positive summative scores are classified into the low income class. In this way, rather than conduct a new set of PRIDIT analysis for the test dataset, we take variable scores and weights already developed in the training step to assess individuals in the test dataset, much in the same way as taking training parameters to score the test dataset in regression analysis. Note that even when test dataset structure (i.e., the sample split parameter θ) is different from that of the training dataset, PRIDIT can still generate reasonably good results due to its robustness (also see Bross 1958 and Golden and Brockett 1987 for more comments). In practice, this method makes it possible to analyze data by only running new PRIDIT procedures from time to time, preferably when structural changes are sensed.

The final results for the test dataset are shown in Table 4. From this table, we can see that PRIDIT did achieve a satisfactory classification in the test dataset. The total hit rate (67%), false positive rate (39%), and false negative rate (12%) all remain about the same as in the training dataset, which demonstrates that PRIDIT is fairly stable and has good out-of-sample properties as desired for effective classification tools.

Table 4 PRIDIT Classification against True Classification for the Test Dataset

	True Low Income	True High Income	PRIDIT Total
PRIDIT Low Income	6887	453	7340
PRIDIT High Income	4473	3247	7720
True Total	11360	3700	15060

Note: [Hit rate: 67.29%; Event class accuracy: 87.76%; Non-event class accuracy: 60.63%].

3.7.5 Comparison with a supervised learning method — logistic regression results

While it may be viewed as "unfair" to compare an unsupervised method to a supervised method, we do so here to exhibit external validity and to assess the classification loss incurred when no dependent variable could reasonably be obtained. Logistic regression is thus used as a benchmark to assess performance of the PRIDIT method and probability of 0.5 is used as the threshold for classification. According to Table 5, as expected for the training dataset, logistic regression, optimizing the parameter selection of the model using a set of known true classifications (i.e., using supervised learning), achieves a better hit rate of around 85% in overall classification and 93% in classification of the non-event class. However, it misses 39% in assigning the events for the training dataset. The performance in the test dataset is similar with an 85% total hit rate, a 93% accuracy in non-event class, and a 60% accuracy in event-class. It is clear that the majority of misses produced by logistic regression occur in the event class, since it tends to "overcompensate" by assigning more observations to the majority class (often the non-event class). This property makes it less favorable both for practical purpose and for the cost-sensitive objective, despite its relatively higher overall accuracy.

Table 5 Logistic Classification against True Classification for the Training and Test Dataset

	True Low Income	True High Income	LOGIT Total
Panel A: Training Dataset			
LOGIT Low Income	21036	2934	23970
LOGIT High Income	1618	4574	6192
True Total	22654	7508	30162
Panel B: Test Dataset			
LOGIT Low Income	10533	1481	12014
LOGIT High Income	827	2219	3046
True Total	11360	3700	15060

Note: [Training dataset: Hit rate: 84.91%; Event class accuracy: 60.92%; Non-event class accuracy: 92.86%].

[Test dataset: Hit rate: 84.67%; Event class accuracy: 59.97%; Non-event class accuracy: 92.72%].

To provide further insights obtained from comparing these two methods, we also present cross-classification results from PRIDIT and logistic regression on the training dataset in Table 6. Remarkably, the two methods agree overall on 67% of the classifications even though PRIDIT is unsupervised and logistic regression is supervised. As we investigate their performance along with true classifications in Table 7, we find consistent evidence for PRIDIT's superior performance in the event class just as before, i.e., for an event class, PRIDIT only makes mistakes mostly on the same instances as logistic regression, indicating that those instances are difficult to learn, even in a supervised learning environment.

Table 6 Cross-Classification of PRIDIT and Logistic Regression for the Training Dataset

	PRIDIT Low Income	PRIDIT High Income	LOGIT Total
LOGIT Low Income	14333	9637	23970
LOGIT High Income	334	5858	6192
PRIDIT Total	14667	15495	30162

Note: [Total Agreement: 66.94%].

Table 7 Comparison of PRIDIT, Logistic Regression, and True Classification for the Training Dataset

	True Low Income			True High Income		
	PRIDIT Low Income	PRIDIT High Income	LOGIT Subtotal	PRIDIT Low Income	PRIDIT High Income	LOGIT Subtotal
LOGIT Low Income	13701	7335	21036	632	2302	2934
LOGIT High Income	61	1557	1618	273	4301	4574
PRIDIT Subtotal	13762	8892	22654*	905	6603	7508*

*is the value for "True Total".

3.7.6 PRIDIT as a completely unsupervised method

In the above subsections, it was shown how to use PRIDIT to carry out a binary classification task on a dataset with a mixture of indicator types. However, if we have a sample of observations for which the value of the dependent variable is known, then we can take advantage of the available dependent variable value to decide the ranks in the presence of non-ordinal categorical predictor variables. The performance of PRIDIT on both of the training and the test datasets turned out to be comparable to our benchmark supervised learning technique logistic regression, with superior accuracy attained in classifying the event class instances. If no external assumption of having an assessed

dependent variable exists, we may restrict attention to ordinal categorical or continuous predictor variables and apply PRIDIT. Alternatively, after some pre-treatments, the unsupervised PRIDIT method can make use of all predictor variables (ordinal or not) for classification without any knowledge of the dependent variable at all, making this technique applicable where logistic regression is impossible to perform. This subsection demonstrates treatment details and results for classification using the mixture of predictor variables under the completely unsupervised framework.

In the case of non-binary, non-ordinal categorical predictor variables, a preliminary assignment of (predicted) class labels is required in order to rank the response categories properly for the use of PRIDIT. This is achieved by performing an initial PRIDIT run using only the ordinal variables in the dataset. In the adult dataset used herein, the ordinal variables include *age* (continuous), *education* (categorical), *capital gain* (continuous), and *working hours per week* (continuous), all of which bear a clear positive relation with membership in the high income class (note that our method only requires a *monotonic* relationship between all the predictors and the underlying latent variable. A positive relationship here will work exactly the same as discussed before for the negative relationship after some simple manipulations, e.g., adding a negative sign to the variable or reverse the category ranking).

The PRIDIT classification based on these four ordinal variables is shown in Table 8. The hit rate is about 59%. It is important to note that these PRIDIT classifications of the individuals based on the initial run (unsupervised) are adopted as the dependent variable value for all further analyses under the unsupervised framework. Using these initial PRIDIT classifications, the same steps as described in subsection 3.7.3 are

executed to rank order all indicator variables. Resulting ranks of the categorical variables remain stable as compared to the original ranks derived using known true classifications, as suggested by Spearman rank correlations of the categorical variables under the original and the unsupervised framework shown in Table 9. We may then obtain class assignment of individuals by performing another run of PRIDIT using all indicator variables after they are all ranked. Note that true class memberships are never employed throughout the classification procedure, exhibiting our method as a completely unsupervised one.

Table 8 PRIDIT Classification with the Four Ordinal Variables against True Classification for the Training and Test Dataset

	True Low Income	True High Income	PRIDIT Total
Panel A: Training Dataset			
PRIDIT Low Income	13019	2677	15696
PRIDIT High Income	9635	4831	14466
True Total	22654	7508	30162
Panel B: Test Dataset			
PRIDIT Low Income	6565	1308	7873
PRIDIT High Income	4795	2392	7187
True Total	11360	3700	15060

Note: [Training Dataset: Hit rate: 59.18%; Event class accuracy: 64.34%; Non-event class accuracy: 57.47%].
[Test Dataset: Hit rate: 59.48%; Event class accuracy: 64.65%; Non-event class accuracy: 57.79%].

Table 9 Spearman Rank Correlations of Categorical Variables under the Original and the Unsupervised Framework

Variable	Spearman Rank Correlation	P-Value
Work Class***	0.96	0.0005
Education***	0.83	0.0001
Marital Status*	0.75	0.0522
Occupation***	0.83	0.0002
Relationship	0.66	0.1562
Race**	0.90	0.0374
Sex	1	/
Native Country***	0.66	< 0.0001

***indicates significance at 1%, **indicates significance at 5%, *indicates significance at 10%

The final classification of the training dataset is evaluated against both the true labels and the PRIDIT labels from the initial run. Results are shown in Table 10. The accuracy statistics against true labels (69% total hit rate, 37% false positive rate, 14% false negative rate) are very similar to those obtained with the help of true labels in ranking in the original framework (as shown previously in section 3.7.3), exhibiting the robustness of the PRIDIT method. The accuracy statistics against PRIDIT labels are generally comparable to those obtained previously but with a higher false negative rate. Classifications for the test dataset against true labels and PRIDIT labels (derived from applying the four-ordinal- variable trained PRIDIT on the test dataset) are also obtained and are presented in Table 11. The classification accuracies are generally similar to those of the training dataset.

Table 10 PRIDIT Classification against True Classification and Initial PRIDIT Classification under the Unsupervised Framework (Training Dataset)

	True Low Income	True High Income	PRIDIT Total
Panel A: Against true labels			
PRIDIT Low Income	14332	1068	15400
PRIDIT High Income	8322	6440	14762
True Total	22654	7508	30162
	PRIDIT Label Low Income	PRIDIT Label High Income	PRIDIT Total
Panel B: Against initial PRIDIT labels			
PRIDIT Low Income	9720	5680	15400
PRIDIT High Income	5976	8786	14762
PRIDIT Label Total	15696	14466	30162

Note: [Training dataset against true label: Hit rate: 68.87%; Event class accuracy: 85.78%; Non-event class accuracy: 63.26%]. [Training dataset against PRIDIT label: Hit rate: 61.36%; Event class accuracy: 60.74%; Non-event class accuracy: 61.93%].

Table 11 PRIDIT Classification against True Classification and Initial PRIDIT Classification under the Unsupervised Framework (Test Dataset)

	True Low Income	True High Income	PRIDIT Total
Panel A: Against true labels			
PRIDIT Low Income	7183	629	7812
PRIDIT High Income	4177	3071	7248
True Total	11360	3700	15060
	PRIDIT Label Low Income	**PRIDIT Label High Income**	**PRIDIT Total**
Panel B: Against initial PRIDIT labels			
PRIDIT Low Income	4578	3234	7812
PRIDIT High Income	3295	3953	7248
PRIDIT Label Total	7873	7187	15060

Note: [Test dataset against true label: Hit rate: 68.09%; Event class accuracy: 83.00%; Non-event class accuracy: 63.23%]. [Test dataset against PRIDIT label: Hit rate: 56.65%; Event class accuracy: 55.00%; Non-event class accuracy: 58.15%].

3.7.7 Logistic regression results in the unsupervised framework — the effect of inaccurate training labels on classification accuracy

As an unsupervised learning method, PRIDIT accuracy is independent of the potential confusion which may result from incorrectly classified initial training samples while supervised methods such as logistic regression can be quite sensitive to the accuracy of the assessment of the dependent variable in the training sample (a similar observation is noted by Hausman et al. 1998 for Probit estimation). This section illustrates this point by investigating performances of logistic regression using the set of initial PRIDIT labels (PRIDIT- assessed class membership) as the dependent variable

under the unsupervised framework. Table 12 and Table 13 contain logistic regression classification results for the training and the test dataset respectively.

Table 12 Logistic Regression Classification against True Classification and Initial PRIDIT Classification under the Unsupervised Framework (Training Dataset)

	True Low Income	True High Income	LOGIT Total
Panel A: Against true label			
LOGIT Low Income	12680	2392	15072
LOGIT High Income	9974	5116	15090
True Total	22654	7508	30162
	PRIDIT Label Low Income	**PRIDIT Label High Income**	**LOGIT Total**
Panel B: Against initial PRIDIT label			
LOGIT Low Income	14412	660	15072
LOGIT High Income	1284	13806	15090
PRIDIT Label Total	15696	14466	30162

Note: [Training dataset against true label: Hit rate: 59.00%; Event class accuracy: 68.14%; Non-event class accuracy: 55.97%]. [Training dataset against PRIDIT label: Hit rate: 93.55%; Event class accuracy: 95.44%; Non-event class accuracy: 91.82%].

Table 13 Logistic Regression Classification against True Classification and Initial PRIDIT Classification under the Unsupervised Framework (Test Dataset)

	True Low Income	True High Income	LOGIT Total
Panel A: Against true label			
LOGIT Low Income	6387	1184	7571
LOGIT High Income	4973	2516	7489
True Total	11360	3700	15060
	PRIDIT Label Low Income	**PRIDIT Label High Income**	**LOGIT Total**
Panel B: Against initial PRIDIT label			
LOGIT Low Income	7214	357	7571
LOGIT High Income	659	6830	7489
PRIDIT Label Total	7873	7187	15060

Note: [Test dataset against true label: Hit rate: 59.12%; Event class accuracy: 68.00%; Non-event class accuracy: 56.22%]. [Test dataset against PRIDIT label: Hit rate: 93.25%; Event class accuracy: 95.03%; Non-event class accuracy: 91.63%].

As we have seen earlier, while PRIDIT is quite accurate in assessing the dependent variable, it does make mistakes. Thus, by using initial PRIDIT labels as the dependent variable value, logistic regression is trained with a misclassified training sample in this unsupervised framework. It is worth noting that logistic regression achieves a surprising performance of a higher than 90% overall hit rate compared to the initial PRIDIT classification on both the training and the test dataset (i.e., logistic regression can accurately assess the initial PRIDIT classification). However, the results against the true labels on both of the training and test datasets are not nearly as good (59%). Compared with PRIDIT's consistent performance as a completely unsupervised method, logistic regression sometimes is of limited value in the sense that its performance

relies heavily on the accuracy of the labels in the training sample, as most other supervised learning methods do. Therefore, these methods not only require the existence of training labels, but are sensitive to the quality of labels as well, a characteristic not desired for many applications where the value of dependent variables can be subject to error even in the training set (e.g., insurance fraud cases). Overall, we view PRIDIT as a more stable, flexible, and robust unsupervised alternative to traditional supervised methods which deliver only comparable or sometimes even less accurate classifications.

3.7.8 Comparison with another unsupervised learning method – cluster analysis

A widely used unsupervised method, cluster analysis, is examined to compare performance with PRIDIT. A two-step cluster analysis is executed on SPSS software using the same set of predictors (both continuous and categorical) as for PRIDIT. While it is possible that more than two clusters could be generated by this program (making the dichotomous classification problematic), for this dataset only two clusters were generated. Cluster analysis classifications are presented in Table 14 and Table 15. Cross-classification results between PRIDIT (under the unsupervised framework) and cluster analysis are shown in Table 16. Since there is no guidance on which cluster produced in cluster analysis corresponds with the event class (or even if one does), our results are based on using external data (knowledge of the true label) to choose a cluster labeling for maximum accuracy. The resulting classification accuracies are similar to those by PRIDIT, with slightly better non-event class performance achieved by cluster analysis.

Table 14 Cluster Analysis Classification against True Classification (Training Dataset)

	True Low Income	True High Income	Cluster Total
Cluster 1	7626	6430	14056
Cluster 2	15028	1078	16106
True Total	22654	7508	30162

Note: If cluster 1 corresponds to high income and cluster 2 corresponds to low income, i.e., under maximum accuracy labeling, then [Hit rate: 71.14%; Event class accuracy: 85.64%; Non-event class accuracy: 66.34%].

Table 15 Cluster Analysis Classification against True Classification (Test Dataset)

	True Low Income	True High Income	Cluster Total
Cluster 1	7526	483	8009
Cluster 2	3834	3217	7051
True Total	11360	3700	15060

Note: If cluster 1 corresponds to low income and cluster 2 corresponds to high income, i.e., under maximum accuracy labeling, then [Hit rate: 71.33%; Event class accuracy: 86.95%; Non-event class accuracy: 66.25%].

Table 16 Cross-Classification of PRIDIT (Completely Unsupervised) and Cluster Analysis (Training and Test Dataset)

	PRIDIT Low Income	PRIDIT High Income	Cluster Total
Panel A: Training Dataset			
Cluster 1	1295	12761	14056
Cluster 2	14105	2001	16106
PRIDIT Total	15400	14762	30162
Panel B: Test Dataset			
Cluster 1	7189	820	8009
Cluster 2	623	6428	7051
PRIDIT Total	7812	7248	15060

Note: If cluster 1 corresponds to high income and cluster 2 corresponds to low income for training dataset, i.e., under maximum accuracy labeling, then [Training Dataset: Agreement: 89.07%]. If cluster 1 corresponds to low income and cluster 2 corresponds to high income for test dataset, i.e., under maximum accuracy labeling, then [Test Dataset: Agreement: 90.42%].

Although the performance is similar, several difficulties hinder the use of cluster analysis in this sort of classification task. First, there is essentially no guaranteed way to obtain a "correct" match between clusters and target classes. In our case, we make use of true classifications and match by maximizing accuracy. Obviously, this is not always implementable in applications where no true classifications are available. Even when there exist a subset of instances with known class membership based on which the match might somehow be assessed, we are never assured of a right link due to the existence of classification errors. Second, in cluster analysis, it is more difficult and less intuitive to determine variable importance and obtain variable weights in numerical form, making it difficult to interpret results and to extend or to incorporate cluster results into more complicated further analysis. By way of contrast, PRIDIT not only assigns specific variable weights but also yields a summative score for each instance on which conclusions are drawn. The numerical score is also capable of being correlated with other exogenous variables or being incorporated itself as a predictor variable in further analysis. Additionally, it is hard to control the size of each cluster, meaning that we cannot make use of the sample proportion parameter even when it is known. For the above reasons, PRIDIT can achieve not only as good classifications, but is more informative, more easily interpreted, and a more "manageable" unsupervised alternative than cluster analysis.

3.7.9 Maintaining the continuous form of continuous predictors

We conclude the empirical part of this chapter by examining the value of keeping continuous variables in the continuous form as opposed to discretizing them. It is argued

at the beginning that transforming continuous variables into categorical ones loses information and thus it is important to extend the method to incorporate both continuous and categorical indicators. We here demonstrate this empirically. To single out the impact of continuous variables, the analysis is restricted to using only the four continuous variables for classification. Table 17 shows that accuracy drops somewhat as compared to using all indicators. Based on intuitive cutoff criteria, these continuous variables are then turned into binary variables (the original monotonic relationship remains through this binary transformation), the specific cutoff points of which are listed in Table 18. Table 19 presents PRIDIT performance using the transformed variables. Both total accuracy and event class accuracy drop dramatically (ten percentage points and twenty percentage points respectively) as compared to keeping the four continuous variables in the continuous form. The performance deterioration is likely due to the loss of information resulting from converting continuous variables to binary variables, since the binary form is coarser then the continuous form in keeping the information content of the indicators. This experiment confirms the importance of a learning method being able to accommodate both types of indicator variables, and thus supports the extension of the PRIDIT method.

Table 17 PRIDIT Classification using Only Continuous Indicators against True Classification (Training and Test Dataset)

	True Low Income	True High Income	PRIDIT Total
Panel A: Training Dataset			
PRIDIT Low Income	14684	2734	17418
PRIDIT High Income	7970	4774	12744
True Total	22654	7508	30162
Panel B: Test Dataset			
PRIDIT Low Income	7283	1339	8622
PRIDIT High Income	4077	2361	6438
True Total	11360	3700	15060

Note: [Training dataset: Hit rate: 64.51%; Event class accuracy: 63.59%; Non-event class accuracy: 64.82%]. [Test dataset: Hit rate: 64.04%; Event class accuracy: 63.81%; Non-event class accuracy: 64.11%].

Table 18 Criteria for Transforming Continuous Variables into Binary Variables

Variable	Age	Capital Gain	Capital Loss	Working Hours per Week
Transformed Variable Value=0 if	<50	=0	=0	<40
Transformed Variable Value=1 if	≥50	>0	>0	≥40

Table 19 PRIDIT Classification using Only Continuous Indicators Transformed into Binary Form against True Classification (Training and Test Dataset)

	True Low Income	True High Income	PRIDIT Total
Panel A: Training Dataset			
PRIDIT Low Income	13437	4194	17631
PRIDIT High Income	9217	3314	12531
True Total	22654	7508	30162
Panel B: Test Dataset			
PRIDIT Low Income	6666	2075	8741
PRIDIT High Income	4694	1625	6319
True Total	11360	3700	15060

Note: [Training dataset: Hit rate: 55.54%; Event class accuracy: 44.14%; Non-event class accuracy: 59.31%].
[Test dataset: Hit rate: 55.05%; Event class accuracy: 43.92%; Non-event class accuracy: 58.68%].

3.8 CONCLUSION

In this chapter, I develop a general PRIDIT method for classification and examine its performance empirically. This method can accommodate both continuous and arbitrary rank-ordered categorical predictor variables, and thus maximize classification performance based on given information. The critical steps for the development are to define the variable score and the discriminatory power measure correctly in the continuous case. The entire procedure is then established by building an analogy between the discrete case and the continuous case. As a validation and interpretation, the connection between this method and the Wilcoxon rank sum statistic is also shown.

Implemented in a widely used large dataset with both categorical and continuous predictor variables for income classification, it is found that PRIDIT, as an unsupervised learning method, exhibits comparable overall performance and superior event class accuracy to that obtained using the supervised learning method logistic regression. Also seen is that unlike supervised learning methods, PRIDIT is capable of producing robust results in the presence of an inaccurate training sample. Moreover, an unsupervised learning method, cluster analysis, serves as a competitor methodology for PRIDIT and is also used for assessment. While the two methods yield similar classifications on the income classification dataset, PRIDIT is more informative and more manageable for further analytic purposes since it produces variable weights and one-dimensional summative scores that can be easily interpreted and applied in subsequent analyses. I also empirically investigate the change in performance caused by transforming continuous predictor variables into binary variables. Significant performance deterioration confirms the importance of maintaining the continuous form and thus demonstrates the value of

this development. In general, this method can be applied to a large variety of classification tasks (e.g., market segmentation, insurance fraud detection, bond ratings, default predictions, etc.), as long as a monotonic relationship exists between the predictors and the latent variable underlying the classification.

CHAPTER 4

The Comparison of PRIDIT and Supervised Learning Methods

4.1 A BRIEF OVERVIEW OF THE COMPARISON

In the above income classification dataset, the performance of PRIDIT was mainly assessed using known values of true income classes as a form of external validation. Logistic regression was also used as another form of external validation to further evaluate performance. In insurance fraud detection however, even "true" classifications may contain a relatively large amount of errors since these "true" classifications are often drawn by insurance adjuster investigations, not by court decisions. Therefore, the evaluation of the PRIDIT method against the "true" classifications may not be very reliable. To hedge the risk of using potentially inaccurate true classifications as the validation criteria, I employ another form of external validation to assess the performance of the PRIDIT method, i.e., the PRIDIT classification of an insurance fraud dataset is assessed relative to that generated by supervised learning methods which could have been adopted were a labeled training sample available. Since it is not easy to obtain a set of "true" classifications which accurately represent reality, this comparison allows us to see how the PRIDIT method will perform compared to supervised learning methods given the available set of "true" classifications. If the PRIDIT method yields classification results similar to the supervised methods, then at least the PRIDIT method is able to capture the same amount of information as the supervised methods without requiring a costly-labeled training sample, despite that we

may not know for sure how accurate the PRIDIT method or the supervised methods are compared with the real classification since the available "true" classifications may not be correct.

It is worth noting that these supervised methods are not competitive methods for the PRIDIT method since they are fundamentally different in terms of prior information required. Rather, the supervised learning methods (logistic regression (LR), Support Vector Machines (SVM, cf., Cristianini and Shawe-Taylor, 2000), Bayesian Additive Regression Trees (BART, cf., Chipman et al., 2006)) are used for external validations. Again, since the insurance fraud dataset available to us does not contain a good selection of continuous variables, the comparison of these methods is executed using the discrete PRIDIT method. However, the derived insights should apply similarly to the general PRIDIT method.

The insurance fraud dataset used for this experiment is a U.S. based personal injury protection (PIP) claims dataset produced by Automobile Insurance Bureau (AIB) in Massachusetts (Viaene et al., 2002). It consists of 1399 claims, each of which was assessed by insurance adjusters and experts in the special investigation unit of AIB. Each claim was given a suspicion score by these experts on a 0-to-10 point scale, and was assigned a code indicating whether the suspicion score is above 1, above 4, or above 7. I use an operational definition of fraud to classify any claim with a suspicion score no less than 4 to be fraudulent according to expert assessments. The entire claim set was divided into a training dataset and a test dataset using a 50-50 random split. The PRIDIT and the supervised learning methods are first run on the training set and are assessed on the test set. The PRIDIT results on the training and test dataset are presented below in Table 20.

The implementation of the supervised learning methods is described in the next section and the generated classifications are also presented.

Table 20 PRIDIT Classification against "True" Fraud Labels (Training and Test Dataset)

	True Non-fraud	True Fraud	PRIDIT Total
Panel A: Training Dataset			
PRIDIT Non-Fraud	297	42	339
PRIDIT Fraud	218	143	361
True Total	515	185	700
Panel B: Test Dataset			
PRIDIT Non-Fraud	294	39	333
PRIDIT Fraud	194	172	366
True Total	488	211	699

Note: [Training dataset: Hit rate: 52.86%; Event class accuracy: 77.30%; Non-event class accuracy: 57.67%]. [Test dataset: Hit rate: 66.67%; Event class accuracy: 81.52%; Non-event class accuracy: 60.25%].

4.2 EXTERNAL VALIDATION THROUGH EMPIRICAL ANALYSIS ON AN INSURANCE FRAUD DATASET

This section presents classification results produced by three prevalent supervised learning methods on the U.S.A. PIP insurance fraud dataset, namely Logistic Regression (LR), Bayesian Additive Regression Trees (BART), and Support Vector Machines (SVM). Each method is first evaluated using the "known" fraud assessment as determined by experts and then compared with PRIDIT classifications where the "known" fraud assessment was never used. Moreover, for LR and BART, Pearson correlations and Spearman rank correlations are also calculated between the predicted scores produced by each method and the summative scores produced by the PRIDIT

method. As explained above, the entire dataset was first split into a training set and a test set, and the classifications of both datasets along with the correlations are shown below.

4.2.1 Logistic regression (LR)

I first use logistic regression (LR) to train the classification model using the training set of data and score the test dataset for predicted classifications. The LR classification results against true classifications are presented in Table 21 and Table 22. Again, probability of 0.5 is used as the threshold for classification. We can observe the same pattern as discussed in Chapter 3. Although the overall accuracy is reasonable, the accuracy on the event class (fraud class) is extremely low while the accuracy on the non-event class (non-fraud class) is very high. In fact, the event class accuracy achieved by LR is much lower than what is achieved by the PRIDIT method as shown in Table 20. This is undesirable in the case of insurance fraud detection, since the fraud class is the event class where the misclassification is most costly to the insurers. The LR performance on the training dataset and the test dataset are similar.

Table 21 LR Classification against "True" Fraud Labels (Expert Assessment) (Training Dataset)

	True Non-fraud	True Fraud	Total
LR Non-fraud	485	131	616
LR Fraud	30	54	84
Total	515	185	700

Note: [Training dataset: Hit rate: 77.00%; Event class accuracy: 29.19%; Non-event class accuracy: 94.17%].

Table 22 LR Classification against "True" Fraud Labels (Expert Assessment) (Test Dataset)

	True Non-fraud	True Fraud	Total
LR Non-fraud	457	143	600
LR Fraud	31	68	99
Total	488	211	699

Note: [Test dataset: Hit rate: 75.11%; Event class accuracy: 32.23%; Non-event class accuracy: 93.65%].

To further compare LR and PRIDIT, I have presented in Table 23 and Table 24 the LR classification against the PRIDIT classification. We can see that although these two methods are entirely different in terms of prior information used (LR is supervised while PRIDIT is unsupervised), they agree on the assessment of the claims about 60% of the time. In fact, PRIDIT catches almost all fraudulent claims predicted by LR and assigns more claims to the fraud class than LR does. This helps PRIDIT to be more accurate in identifying potentially fraudulent claims.

Table 23 LR Classification against PRIDIT Classification (Training Dataset)

	LR Non-fraud	LR Fraud	Total
PRIDIT Non-fraud	337	2	339
PRIDIT Fraud	279	82	361
Total	616	84	700

Note: [Agreement: 59.86%].

Table 24 LR Classification against PRIDIT Classification (Test Dataset)

	LR Non-fraud	LR Fraud	Total
PRIDIT Non-fraud	329	4	333
PRIDIT Fraud	271	95	366
Total	600	99	699

Note: [Agreement: 60.66%].

Besides examining cross-classifications between LR and the PRIDIT method, I also calculate the Pearson correlations and Spearman rank correlations between the LR predicted probabilities and the PRIDIT summative scores of the claims. Both of these scores reflect the consistency of the suspicion level assessment of the claims by each method. The correlations of the scores may reveal more information than the binary classifications and will facilitate the comparison of the two methods. These correlations are presented in Table 25 below.

Table 25 Pearson and Spearman Rank Correlations between LR Predicted Probabilities and PRIDIT Scores

	Correlation between LR Predicted Probabilities and PRIDIT Scores	P Value
Panel A: Training Dataset		
Pearson Correlation	0.7543	<0.0001
Spearman Rank Correlation	0.7912	<0.0001
Panel B: Test Dataset		
Pearson Correlation	0.7819	<0.0001
Spearman Rank Correlation	0.8377	<0.0001

From Table 25, we can see that LR predicted probabilities and PRIDIT summative scores for the claims in both the training dataset and the test dataset are significantly highly correlated. This suggests that although the PRIDIT method does not use any true labels to obtain the classification model, its predictions on the suspicion level of the claims are remarkably similar to those obtained by a supervised method where true labels are required to derive any predictions at all. Moreover, this sheds light on why the binary classifications produced by the PRIDIT method may actually be better than those produced by the supervised learning method LR (as shown in Table 20 to

Table 22): LR's assessment of the suspicion level may be systematically biased down toward the non-fraud class since that is the majority class.

4.2.2 Bayesian Additive Regression Trees (BART)

BART is a recent technique designed for predictions (c.f., Chipman et al., 2006). It has been shown that BART has achieved more accurate prediction results than other prevalent supervised learning methods on a variety of applications (Chipman et al., 2006). Therefore, I also examine BART's performance on the insurance fraud dataset and compare the BART classification with the PRIDIT classification. BART classification results are shown in Table 26 and Table 27 below. From these two tables, we can see that BART exhibits a learning pattern similar to LR. More specifically, BART predicts extremely accurately the non-fraud class but extremely inaccurately the fraud class.

Table 26 BART Classification against "True" Fraud Labels (Training Dataset)

	True Non-fraud	True Fraud	Total
BART Non-fraud	504	148	652
BART Fraud	11	37	48
Total	515	185	700

Note: [Training dataset: Hit rate: 77.29%; Event class accuracy: 20.00%; Non-event class accuracy: 97.86%].

Table 27 BART Classification against "True" Fraud Labels (Test Dataset)

	True Non-fraud	True Fraud	Total
BART Non-fraud	470	170	640
BART Fraud	18	41	59
Total	488	211	699

Note: [Test dataset: Hit rate: 73.10%; Event class accuracy: 19.43%; Non-event class accuracy: 96.31%].

Cross-classifications between BART and PRIDIT are also presented in Table 28 and Table 29 below.

Table 28 BART Classification against PRIDIT Classification (Training Dataset)

	BART Non-fraud	BART Fraud	Total
PRIDIT Non-fraud	339	0	339
PRIDIT Fraud	313	48	361
Total	652	48	700

Note: [Agreement: 55.29%].

Table 29 BART Classification against PRIDIT Classification (Test Dataset)

	BART Nonfraud	BART Fraud	Total
PRIDIT Nonfraud	333	0	333
PRIDIT Fraud	307	59	366
Total	640	59	699

Note: [Agreement: 56.08%].

From Table 28 and Table 29, we can again see a pattern similar to that in LR. PRIDIT agrees with BART on all the predicted fraudulent claims by BART and predicts more fraudulent claims than what BART has predicted, which helps improve the low accuracy BART has in the fraud class.

Lastly, Pearson correlations and Spearman rank correlations between BART scores and PRIDIT summative scores are presented in Table 30 below. We can see that the correlations between BART scores and PRIDIT scores are even higher than those calculated between LR scores and PRIDIT scores.

Table 30 Pearson and Spearman Rank Correlations between BART Scores and PRIDIT Scores

	Correlation between BART Scores and PRIDIT Scores	P Value
Panel A: Training Dataset		
Pearson Correlation	0.8793	<0.0001
Spearman Rank Correlation	0.8815	<0.0001
Panel B: Test Dataset		
Pearson Correlation	0.9040	<0.0001
Spearman Rank Correlation	0.9057	<0.0001

4.2.3 Support Vector Machines (SVM)

SVM has been documented as one of the most accurate classification methods among supervised learning methods. Therefore, I also include SVM as one of the supervised methods examined here. The classification results by SVM on the training and test datasets are presented in Table 31 and Table 32 below. We can see that SVM indeed performs very well on the training dataset in terms of the overall accuracy and the accuracy on the non-fraud class. Although the event class accuracy is still much lower than the PRIDIT method, it has achieved the best performance among all the supervised learning methods. However, when the trained SVM model is applied to the test dataset, the accuracy on the fraud class drops significantly. Since these trained classification models will ultimately be used to classify upcoming test datasets, the performance of SVM in the insurance fraud detection dataset is less than satisfactory.

Table 31 SVM Classification against "True" Fraud Labels (Training Dataset)

	True Non-fraud	True Fraud	Total
SVM Non-fraud	504	69	573
SVM Fraud	11	116	127
Total	515	185	700

Note: Parameters used in the SVM model are: gamma=0.5 and cost=1.

[Training dataset: Hit rate: 88.57%; Event class accuracy: 62.70%; Non-event class accuracy: 97.86%].

Table 32 SVM Classification against "True" Fraud Labels (Test Dataset)

	True Non-fraud	True Fraud	Total
SVM Non-fraud	458	191	649
SVM Fraud	30	20	50
Total	488	211	699

Note: Parameters used in the SVM model are: gamma=0.5 and cost=1.

[Test dataset: Hit rate: 68.38%; Event class accuracy: 9.48%; Non-event class accuracy: 93.85%].

Cross-classifications between SVM and PRIDIT are also shown in Table 33 and Table 34. Again, these two methods agree with each other most of the time.

Table 33 SVM Classification against PRIDIT Classification (Training Dataset)

	SVM Non-fraud	SVM Fraud	Total
PRIDIT Non-fraud	322	17	339
PRIDIT Fraud	251	110	361
Total	573	127	700

Note: Parameters used in the SVM model are: gamma=0.5 and cost=1.

[Agreement: 64.14%].

Table 34 SVM Classification against PRIDIT Classification (Test Dataset)

	SVM Non-fraud	SVM Fraud	Total
PRIDIT Non-fraud	328	5	333
PRIDIT Fraud	321	45	366
Total	649	50	699

Note: Parameters used in the SVM model are: gamma=0.5 and cost=1.

[Agreement: 53.36%].

4.3 STRENGTHS AND WEAKNESSES OF ALTERNATIVE METHODS

From the above results, we can see that these supervised methods all exhibit a similar learning pattern: as expected, the overall performance is better than the unsupervised PRIDIT method on both the training and the test dataset; the performance on the non-fraud (non-event) class is extremely good but the accuracy is very poor on the fraud (event) class. This uneven performance is more pronounced in the test dataset. This is partly because these supervised methods rely on the training set to learn the pattern and use the obtained parameters to predict the test dataset. Since the fraud class is the minority class in the training sample, these supervised methods tend to assign most claims to the majority class which leads to a great amount of "misses" in the minority class (fraud class).

Another possible reason is that the value of the labels (whether the claim is fraudulent or not) may not be accurate in this training sample since they are only the best assessment by the experts. If there are indeed inaccurate labels present in the training set, the learning performance of supervised methods can be greatly impaired, especially when they are used to predict the classifications of the test dataset. By way of contrast, the PRIDIT method does not make use of these possibly inaccurate labels. Instead, the

PRIDIT method relies on the predictor variables and discovers pattern from them directly. In the context of insurance fraud detection, the predictor variables (e.g., age of the driver, whether there is a police report, etc.) are readily available and most likely accurate. Therefore the PRIDIT method is not adversely impacted by the existence of inaccurate labels. Here I only consider learning accuracy. If cost of labeling (i.e., claims auditing cost in the context of insurance fraud detection) and possible difference between the two types of misclassification costs (i.e., type I and type II error) are taken into account, the PRIDIT method may provide even more benefits. This point will be further discussed in Chapter 5.

We can also see that the correlations and the rank correlations between the scores produced by supervised methods and those produced by the PRIDIT method are very high. These scores represent the suspicion level of the claims as assessed by each method. Therefore, the high positive correlations indicate that the PRIDIT method captures the same underlying suspicion patterns as the supervised methods do, but does not require the use of a training dataset with known dependent variable values. Also, the PRIDIT method has the additional merit of being relatively free of the bias created by an unevenly distributed training sample and the inaccurate labels. These correlation results further validate the effectiveness of the PRIDIT method in insurance fraud detection, namely, it captures the relative ranking of the suspicion level of the claims similarly to the supervised methods and provides more accurate classification in the fraud class, but can be used in situations where it is impossible or cost prohibitive to use supervised learning methods.

In conclusion, the learning pattern exhibited by the examined supervised learning methods is undesirable in the context of insurance fraud detection where the very purpose is to detect suspicious claims, or members of the minority class. Moreover, there is no easy remedy for the problem of inaccurate and uneven training samples which may cause difficulties for supervised methods. Therefore, the PRIDIT method seems to be a preferable candidate method to achieve the objective of effectively detecting potential fraudulent claims.

CHAPTER 5

Improving the PRIDIT Method in the Context of Active Learning and Fraud Rate Estimation

5.1 INTRODUCTION TO THE IMPROVEMENT OF THE PRIDIT METHOD

A class of hybrid methods can be created by combining the unsupervised learning method PRIDIT with supervised learning methods, such as logistic regression, neural networks, and support vector machine techniques. As an unsupervised method, PRIDIT can apply to a dataset without prior knowledge of the dependent variable (i.e., fraud or non-fraud). Analyses show that PRIDIT can produce satisfactory classification results, especially for the subset of claims with most extreme (positive or negative) PRIDIT scores, i.e., the extreme subset. The PRIDIT performance for the extreme subsets on a U.S.A. insurance fraud dataset (as described in Chapter 4) and a Spanish insurance fraud dataset are shown in Table 35 and Table 36 below. The Spanish insurance fraud dataset contains 1995 physical damage claims in automobile insurance from 1993 to 1996 in Spain (see Artis et al. 2002). Claims in this dataset have been classified by the insurer as fraudulent or honest, and the classification is coded as a binary variable in the dataset. This binary variable is used as true classifications in all empirical analyses. This dataset also contains a set of predictor variables, which I use to conduct the PRIDIT analyses and to build other supervised learning models.

Each extreme subset in Table 35 and Table 36 is constructed by taking a certain percentage of the highest ranked claims and the same percentage of lowest ranked claims

after running the PRIDIT analysis on the entire dataset. For example, the first extreme subset shown in these two tables contains 10% of the data from the entire dataset and is composed of the top and the bottom ranked 5% of the claims according to the PRIDIT ranking. The classification accuracy on each extreme subset is calculated and presented. These accuracies are also compared with accuracies on randomly selected subsets of the same size. It is clear from the tables that PRIDIT achieves good accuracy in the extreme subsets, and these accuracies are much higher than those in the random subsets.

Table 35 PRIDIT Classifications for the Extreme Subsets (Spanish Dataset)

Size of the Subset as a Percentage of the Entire Dataset (%)	Accuracy of classification on the Extreme Subset	Accuracy of classification on a Random Subset
10	0.9347	0.6350
20	0.8697	0.6667
30	0.8581	0.6578
40	0.7729	0.6579
50	0.7462	0.6483

Table 36 PRIDIT Classifications for the Extreme Subsets (U.S.A. Dataset)

Size of the Subset as a Percentage of the Entire Dataset (%)	Accuracy of classification on the Extreme Subset	Accuracy of classification on a Random Subset
10	0.8058	0.6500
20	0.7921	0.6286
30	0.7566	0.6929
40	0.7299	0.6714
50	0.7253	0.6300

Since PRIDIT classifications are quite accurate in the extreme subsets, we can exploit the information provided by the unsupervised PRIDIT method and use it in a subsequent supervised method. Accordingly, to this end, a class of learning methods is developed based on the PRIDIT method within the framework of active learning. As introduced in Chapter 1, in traditional active learning, an initial costly-labeled training

sample is obtained and is first used to train a selected supervised classifier, and additional training examples are subsequently acquired to refine the learning process. Thus, at the point of each acquisition, the total cost is the sum of labeling cost for the initial sample, plus the labeling cost for additional examples selected, plus the misclassification cost when the model is used in the test dataset (incoming claims in the fraud detection context). The active learning literature generally focuses on the design of acquisition algorithms to obtain additional "informative" examples, but rarely discusses the selection of the initial training sample. As a result, a randomly selected initial training sample is usually employed. Alternatively, I propose to use the costless information provided by PRIDIT analyses to select the initial training sample for a subsequently applied supervised method, leading to a class of new methods.

More specifically, two approaches are adopted to develop the new methods. First, the PRIDIT ranking of the claims can be used to construct the initial training sample. The objective is to select the best initial training sample to start with at the beginning of the learning process (when only a small subset of training examples is used to build the classification model) in order to improve classification accuracy. This approach is named "the informative initial sample approach" and is developed in the next section. Second, the PRIDIT assessment of the labels for the extreme subset of claims can be used in place of true labels to prepare the initial training sample for the subsequent supervised method. By using the PRIDIT assessment of the labels, we may sacrifice classification accuracy but save on the cost of obtaining the labels for the initial training sample. This approach is named "the hybrid method" and will be discussed in section 5.3. All empirical analyses

in the following sections of this chapter are implemented on the Spanish insurance fraud dataset.

5.2 USING PRIDIT TO PREPARE AN INFORMATIVE INITIAL TRAINING SAMPLE

Prediction models based on supervised learning methods usually make use of a random (initial) sample, since little or no knowledge can be extracted by a supervised learning method before a training sample is selected and used to train the model. However, with the use of an unsupervised method, a more "informative" initial sample (than a random sample) may be obtained to boost the performance of the subsequently trained supervised method. Specifically, in the context of the PRIDIT method, we can define the "informativeness" of the examples by the ranking of their summative scores generated from running PRIDIT on the entire dataset. Since the summative scores are monotonically decreasing with the suspicion levels of the insurance claims, claims with highest scores and lowest scores (claims in the extreme subset) are the most "typical" ones as determined by the PRIDIT method.[2] These "typical" examples (claims) can potentially contribute the most to the learning process by a supervised method and thus should be included in the initial training sample.

To examine the effectiveness of the informative initial training sample approach, experiments are run on the Spanish insurance fraud dataset. More specifically, I select logistic regression as the component supervised learning method and run three sets of experiments. The entire sample is divided into a training set and a test set by a 50-50 random split. The resulting training set has 998 claims and the test set has 997 claims.

[2] As we have seen in Table 35 and Table 36, these examples are also most accurately classified.

The size of the initial sample has been set to 100 claims. Given the relatively large number of predictor variables in this insurance fraud detection dataset, this sample size seems to be appropriate as shown by undocumented experiments. Additional training examples are acquired in batches of 10 claims to increase the efficiency of the learning methods. The first set starts with a random initial sample and acquires additional training examples randomly to build a classification model. The second set starts with an "informative" initial sample and acquires additional training examples randomly. The third set starts with an "informative" initial sample and acquires additional training examples "actively." The prevalent "Query by Bagging" algorithm is used to acquire additional examples actively (cf., Abe and Mamitsuka 1998). The experiments are repeated 10 times to eliminate the undue influence of certain random samples. The results are presented in Figure 1 and Figure 2 below.

Figure 1 Performance Comparison on an Insurance Fraud Dataset (Spanish Dataset)

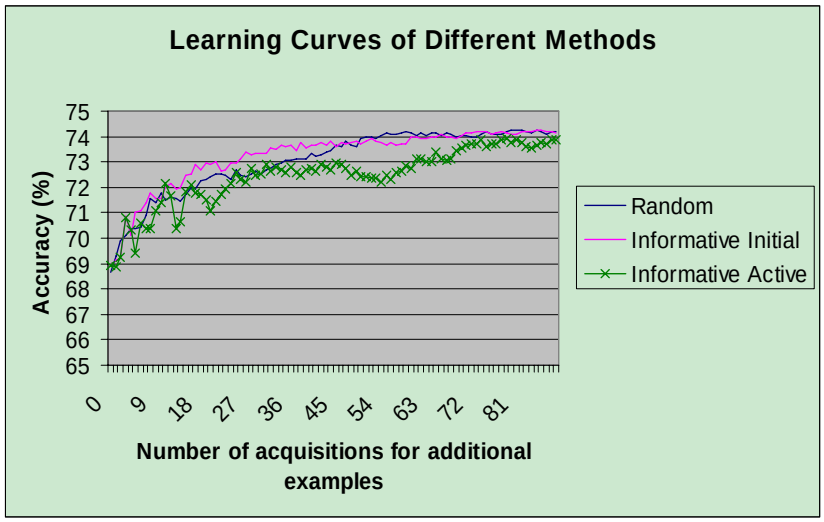
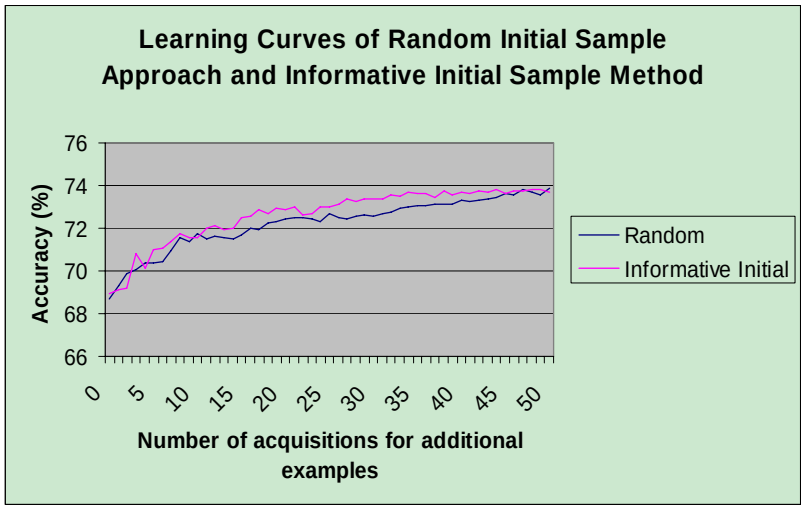


Figure 2 Performance Comparison between the Random Initial Sample Method and the Informative Initial Sample Method on an Insurance Fraud Dataset: First 50 Acquisitions (Spanish Dataset)



From Figure 1, we can see that the learning curve of logistic regression with the informative initial sample lies above that of logistic regression with random sampling at the beginning stage of the learning process. Later on, the random sampling approach begins to dominate for a while and eventually, these two curves converge as all training examples are exhausted. Interestingly, we can see that the active learning method does not perform as well when an "informative" initial sample is used. This may be due to the fact that the "informative" initial sample may have selected the most valuable examples in the dataset and the current active learning algorithms to select the most informative example(s) at each additional acquisition may not be applicable. Whether the traditional active learning method will work with the proposed informative initial training sample method is a potential valuable future research question. Figure 2 presents only the beginning portion of the learning curve where the informative initial sample approach performs better than the random initial sample approach.

The beginning stage is often the most important in the learning process. In insurance fraud detection and most other applications, budgetary and timing concerns usually do not allow one to acquire a large amount of labeled training examples and to continue far along the learning curve. The merit of the idea of active learning (the informative initial training sample or the active acquisition of additional training examples) is to have the benefits front-loaded and put the limited resources to the best use. For example, the insurance company can only spend a limited amount of time, labor, and capital to audit the claims in the initial training sample and the first few additional acquisitions of training claims, in order to obtain an insurance fraud detection model with acceptable detection accuracy. It does not necessarily need to spend a large amount of extra resources to obtain the maximum accuracy. To this end, the informative initial training sample provides more benefits than the random sampling approach in practice since it performs better at the initial training stage. Note also that in an evolving environment wherein the persons committing fraud may be changing tactics (so the model parameters used in the logistic regression are not stationary), the fact that ultimately the random sampling approach exceeds the informative initial sample method in accuracy when more labels are known may be less relevant to practical implementation.

5.3 COST-SENSITIVE PERSPECTIVE OF METHODOLOGY DESIGN

5.3.1 Introduction to the cost-sensitive perspective of the hybrid method

Instead of the costly process of obtaining the true labels for claims with extreme PRIDIT scores, the PRIDIT assessment for these claims could be used as the value of the dependent variable to construct the initial training sample. This PRIDIT-assessed initial sample is then combined with additional examples for which the true labels are obtained in a costly manner to build supervised classification models. Thus, our final model is obtained by applying a supervised learning method to a training sample consisting of both PRIDIT-labeled examples and examples with true labels. This hybrid model should save on total cost since the classification of the initial training sample (via PRIDIT) is costless.

The traditional active learning methods usually pursue an accuracy (classification or class probability estimation accuracy) objective. However, a cost-sensitive objective may be more relevant in insurance fraud detection than classification accuracy. The objective function of using a cost-sensitive classification method acknowledges possible different costs between auditing a non-fraudulent claim and missing a fraudulent claim (i.e., costs of type I and type II error), and also allows one to consider the trade-off between the auditing cost for obtaining "true labels" for use in training samples and the misclassification costs when the final model is applied. After all, the ultimate goal is to reduce the total cost of insurance fraud, be it the auditing cost or the claims cost. The existing literature has seen some development and applications of cost-sensitive learning methods (Domingos, 1999, Viaene et al., 2004b).

To pursue the cost-sensitive objective in this methodology design, I first present a discussion of the traditional cost matrices and the cost matrix appropriate for this application. The performance of the hybrid method with PRIDIT-assessed initial sample is then discussed using a cost-sensitive evaluation criterion according to the cost matrix created. This assessment criterion is also compared with the classification accuracy criterion.

5.3.2 Insurance fraud cost matrix

A cost matrix is usually specified in cost-sensitive learning studies. However, the cost discussed in such studies usually refers only to what is incurred *after* obtaining the classifier, such as claims auditing cost (when the claim is classified as fraudulent) or claims cost (when the claim is classified as non-fraudulent). The cost of learning, which is incurred when the classifier is being trained, is seldom considered in the literature and usually is not included in the cost matrix (e.g., the cost of obtaining a labeled initial training sample for parameter estimation). Table 37 below presents such a cost matrix to be used in insurance fraud detection (Viaene, Derrig and Dedene, 2004b, Viaene et al., 2004). This matrix assumes that the model parameters have already been fit, and that a model prediction of "true legitimate claims" results in no auditing of the claim, but simply a payment of the claim. A model prediction of "true fraudulent claims" results in an audit of the claim and consequently an audit cost. The results of the audit can be savings of the claims cost in the event of a real fraudulent claim, but a "wasted" audited claim if the model falsely predicts "fraud" and the subsequent audit shows the claim to be "legitimate."

Table 37 A Traditional Insurance Fraud Cost Matrix (Viaene, Derrig and Dedene, 2004b, Viaene et al., 2004)

	True Fraudulent Claims	True Legitimate Claims
Predicted Fraudulent Claims	Auditing Cost - Claims Cost	Auditing Cost
Predicted Legitimate Claims	Claims Cost	No Cost

Elkan (2001) discusses some interesting aspects of cost sensitive learning for binary classification. Some "reasonableness" conditions are first given: 1) the cost of labeling an example incorrectly is greater than cost of labeling it correctly, and as a result, 2) neither row in the cost matrix should dominate the other (see Table 37 for the arrangement of rows and columns). Viaene, Derrig and Dedene (2004b) adopt these conditions and further impose three additional common assumptions: 1) the cost matrix is given (or estimated) beforehand, 2) the cost matrix is independent of the predictor vector, 3) the cost matrix is stationary during the learning process. Elkan (2001) also stresses that common baselines should be used when assigning cost and benefit in the cost matrix, especially for opportunity costs, to avoid an inconsistent cost matrix. I have adopted all the above reasonableness conditions when designing the cost matrix for using the hybrid method in insurance fraud detection.

However, in addition to the considerations from the cost matrix literature described above, I propose a cost matrix that also takes into account the cost of learning. This acknowledges that the labels for the training examples are obtained with a cost in most cases. In supervised insurance fraud detection, claims used as training examples for building a fraud detection model need to be audited by human experts before they can be assessed as fraudulent or legitimate. The cost incurred in the auditing process (e.g.,

human time, money, and other resources) thus constitutes the cost of learning in this case. A general version of this type of cost matrix is presented in Table 38.

Table 38 A General Cost Matrix

	True Fraudulent Claims	True Legitimate Claims
No Prediction Model Applied	Claim cost	No cost
Model Predicted Fraudulent Claims	Learning cost + Claim auditing cost - Claim cost	Learning cost + Claim auditing cost + Possible legal expense (due to Bad Faith) + Possible reputational expense (due to Lost Business)
Model Predicted Legitimate Claims	Learning cost + Claim cost	Learning cost

Note: learning cost (labeling cost) equals to claim auditing cost in insurance fraud detection

More specifically in active learning, the cost of learning refers to both the labeling cost for the initial training sample and the cost incurred to acquire additional training examples before the final model is obtained. A cost matrix used for active learning is then described in Table 39. Compared to traditional active learning, the hybrid method with PRIDIT-assessed initial training sample will save on the initial labeling cost and thus on the total cost of learning. Table 40 presents a modified cost matrix that accounts for the different total cost of learning under the hybrid method. Note that although legal expense and reputational expense are valid expenses and could be rather large in practice, they are ignored in the empirical analysis presented in the next subsection since I do not have data on these expenses.

Table 39 A Cost Matrix for Active Learning

	True Fraudulent Claims	True Legitimate Claims
No Model Prediction Applied	Claim cost	No cost
Model Predicted Fraudulent Claims	Initial sample labeling cost + additional sample labeling cost + Claim auditing cost - Claim cost	Initial sample labeling cost + additional sample labeling cost + Claim auditing cost + Possible legal expense (due to Bad Faith) + Possible reputational expense (due to Lost Business)
Model Predicted Legitimate Claims	Initial sample labeling cost + additional sample labeling cost + Claim cost	Initial sample labeling cost + additional sample labeling cost

Note: Labeling cost equals to claim auditing cost in insurance fraud detection

Table 40 A Cost Matrix for the Hybrid Method

	True Fraudulent Claims	True Legitimate Claims
No Model Prediction Applied	Claim cost	No cost
Model Predicted Fraudulent Claims	Additional sample labeling cost + Claim auditing cost - Claim cost	Additional sample labeling cost + Claim auditing cost + Possible legal expense (due to Bad Faith) + Possible reputational expense (due to Lost Business)
Model Predicted Legitimate Claims	Additional sample labeling cost + Claim cost	Additional sample labeling cost

Note: Labeling cost equals to claim auditing cost in insurance fraud detection

Note that in these cost matrices, the labeling cost for the examples are the auditing cost for the claims, which coincides with the misclassification cost when legitimate claims are misclassified as fraudulent. The only difference is that the labeling cost is incurred during the training stage of the learning process, while misclassification cost is incurred during the application stage of the final model. In the current setting, the timing of the cost is not considered, i.e., one dollar of cost incurred in the training stage is treated the same as one dollar of cost incurred in the test stage. Note also that throughout this subsection, I use an implicit assumption that the auditing result is always correct (i.e.,

no further misclassification is considered after auditing). This assumption is commonly adopted in the literature, and is made to facilitate the evaluation of the learning methods.

As shown in Table 39 and Table 40, the hybrid method has the advantage of not requiring labeling of the initial sample. The downside, of course, is that since PRIDIT is not one hundred percent accurate, misclassification may be higher at the beginning of the training process. However, when labeling cost is relatively high, such as in the case of insurance fraud detection, the savings accrued from obtaining fewer labels may more than compensate for the loss in predictive accuracy, resulting in a lower total cost. Subsequently we will empirically examine when the savings obtained by using the PRIDIT-labeled initial training sample exceed the misclassification costs for the insurance fraud dataset.

5.3.3 The hybrid method with the PRIDIT-assessed initial training sample

To implement the hybrid method, PRIDIT analysis is first run on the entire training dataset. Similarly to the informative initial sample approach introduced in section 5.2, the most extreme subset is selected according to the ranking of the PRIDIT summative scores. This subset serves as the initial training sample for a subsequent supervised method. The difference, however, is that true labels of claims in this subset are not acquired. Instead, PRIDIT-assessments of these claims are used as the values of the dependent variable (fraud or not). After the initial sample is prepared, additional examples are selected and true labels of these examples are obtained. These additional training examples are consecutively added to the initial training sample to train a selected supervised classifier until the final model is obtained.

Three sets of experiments are run on the Spanish insurance fraud dataset to evaluate the performance of this hybrid method. Logistic regression is chosen to be the component supervised learning algorithm. The same training and test samples are used as for the informative initial sample method. The training set has 998 claims and the test set has 997 claims. The initial sample size has been set to 100 claims. Additional training examples are acquired in batches of 10 claims to increasing the efficiency of the learning methods. The first set starts with a random initial sample with true labels and acquires additional training examples randomly to build the model. The second set starts with an "informative" initial sample with true labels and acquires additional training examples randomly. The third set starts with an "informative" initial sample with PRIDIT-assessed labels and acquires additional training examples randomly. The experiments are repeated 10 times to eliminate the undue influence of certain random samples. Since in section 5.2., the active learning method "Query for Bagging" for acquiring additional training examples does not perform well when an informative initial sample is used, I do not include the active learning method here for comparisons.

Both the accuracy and the cost matrix method are used as evaluation criteria to assess performance. As explained above, these two evaluation criteria do not necessarily lead to the same conclusion. Learning curves using accuracy as the evaluation criterion are presented below in Figure 3. From Figure 3, we can see that the hybrid method seems to be much less accurate than the random sampling approach and the informative initial sample approach. This is because that the hybrid method starts with a PRIDIT-labeled initial sample which can contain misclassifications. To put these three methods on a "fair" ground for comparisons, we can plot learning curves starting when all methods

have used 100 labeled examples (which is the size of the initial training sample). In this case, the hybrid method would have acquired 10 batches of 10 additional examples with labels while the other two methods started out with an initial sample of 100 labeled examples. The learning curves with a "fair" starting point are plotted in Figure 4.

Figure 3 Performance Comparison Using the Accuracy Criterion (Spanish Dataset)

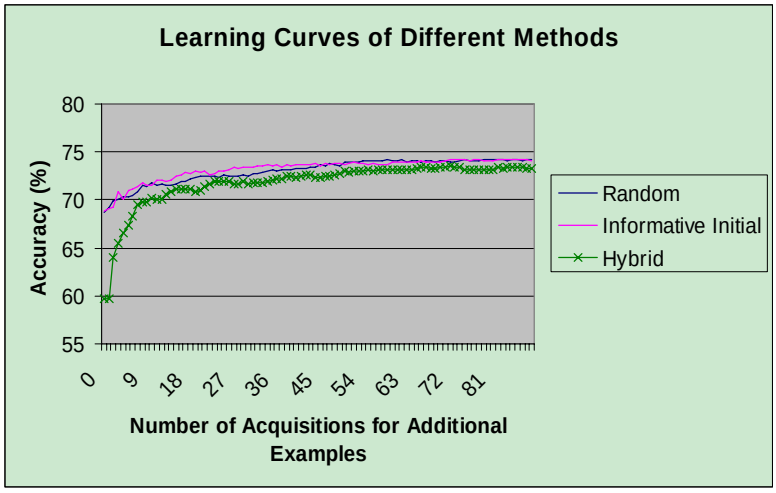


Figure 4 Performance Comparison with a "Fair" Staring Point Using the Accuracy Criterion (Spanish Dataset)

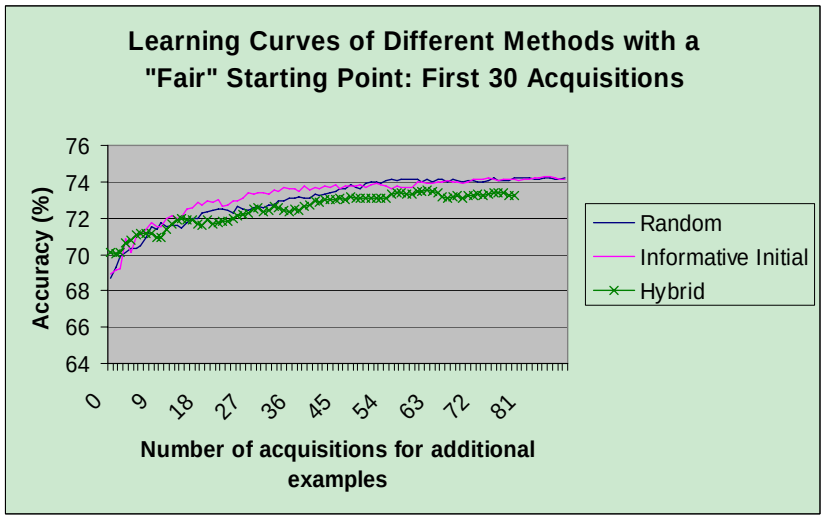
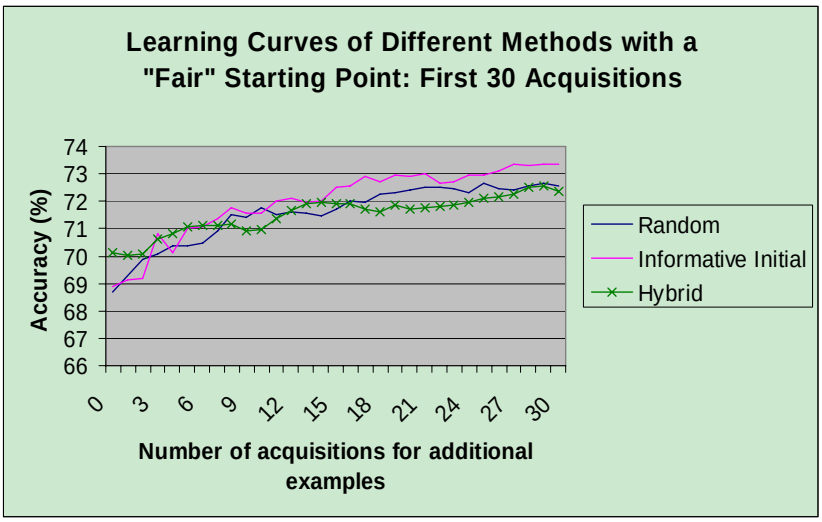


Figure 5 Performance Comparison with a "Fair" Staring Point Using the Accuracy Criterion: First 30 Acquisitions (Spanish Dataset)



From Figure 4 and Figure 5, we can see that the hybrid method performs better at the beginning of the learning process where only a few more batches of additional labeled examples are added to the initial sample to train the model. As more and more true-labeled examples are added for training, the other two methods using a true-labeled initial sample begin to dominate. This is mainly because the inaccurate labels present in the PRIDIT-labeled initial sample hinder the further improvement for the hybrid method. In a sense, as more and more true labels are added in the training sample, these inaccurate labels may "confuse" the supervised learning method and thus slow down the pace of the increase in accuracy. If there is a mechanism to assign weights to reflect the "credibility" of the inaccurate labels and update the weights as learning improves, the adverse impact of these inaccurate labels may be reduced. This point is further discussed in Chapter 6 as a future research direction.

In the above, the accuracy criterion was used to evaluate the performance of the hybrid method. In Figure 4 and 5, the hybrid method has been put on a "fair" ground for comparison with the other methods in terms of the total labeling cost incurred. Alternatively, a cost-sensitive criterion can be used to make this assessment more directly. I use the cost matrix for the hybrid method (as shown in Table 40) to design the cost-sensitive criterion. Recall that legal expenses and reputational expenses are not considered in the empirical analysis since there is no data available on these expenses. This cost matrix contains both the cost of learning incurred at the training stage and misclassification costs incurred at the test stage. More specifically, at each step of the learning process, a certain number of training examples have been selected to train a classification model and the classification model is used to classify the test dataset. Under the hybrid method for insurance fraud detection, the learning cost refers to the labeling cost for those claims acquired with true labels (by claims auditing) which are used for training purposes. The two types of misclassification costs refer to the auditing cost incurred when a predicted fraudulent claim is actually a legitimate claim and the claims cost incurred when a predicted legitimate claim is indeed a fraudulent claim.

In order to use the cost-sensitive criterion to evaluate the hybrid method, I first calculate the cost curves for the hybrid method and the random sampling approach using the current test set which contains 997 claims. I also examine the cost effects in larger test sets by assuming that the same accuracy achieved in the current test set will be achieved in larger test sets. The two larger test sets examined have the size of 2000 claims and 5000 claims respectively.

More specifically, at each step of the learning process, I calculate the sum of the total auditing cost for all the training claims with true labels which are used to build the current classification model (note that the initial training sample is also used to build this model but this sample is costless) and the two types of misclassification costs incurred when the current model is applied to classify claims in the test dataset. For ease of calculation, I assume that the claims cost is four times the auditing cost, which is an estimate of the average costs from the true claims datasets. The calculations for the first six steps of the learning process are demonstrated in Table 41 below.

The first step is to build a classification model with the initial sample (costless for the hybrid method but costly for the random sampling approach). Steps two to five are five acquisitions for additional true-labeled training claims for both methods. Misclassification costs are calculated by multiplying the percentage of misclassifications and the size of the test set (997 claims in Table 41) and the unit cost of misclassification. Recall that the unit misclassification cost for the fraud class (i.e., claims cost) is assumed to be four times that for the nonfraud class (i.e., claims auditing cost). Learning cost (labeling cost) is the unit labeling cost times the number of true-labeled claims used in training the model in each step. Note that the unit labeling cost also equals the claims auditing cost. Finally, the total cost is the sum of misclassification cost for nonfraud class, the misclassification cost for fraud class, and the learning cost. Therefore, all the costs are denominated in the unit claims auditing cost (i.e., the true dollar costs can be calculated by multiplying the costs presented in Table 41 by the dollar amount of claims auditing cost per claim).

Table 41 Total Cost Incurred at Each Step of the Learning Process for the Current Test Set (997 claims): the First Six Steps

Step No.	Percentage of Misclassification (Nonfraud Class) (%)	Misclassification Cost (Nonfraud Class)	Percentage of Misclassification (Fraud Class) (%)	Misclassification Cost (Fraud Class)	Number of True-Labeled Claims Used	Total Cost
Panel A: Cost Calculation for the Random Sampling Approach						
0	12.8485	128.1	18.4754	736.8	100	964.9
1	12.8586	128.2	17.8536	712	110	950.2
2	12.7884	127.5	17.3220	690.8	120	938.3
3	13.0792	130.4	16.8305	671.2	130	931.6
4	12.7482	127.1	16.8907	673.6	140	940.7
5	12.5276	124.9	17.1214	682.8	150	957.7
Panel B: Cost Calculation for the Hybrid Method						
0	19.2076	191.5	21.1133	842	0	1033.5
1	17.5694	175.2	21.8355	870.8	10	1056.0
2	15.9679	159.2	20.0401	799.2	20	978.4
3	15.7172	156.7	18.7663	748.4	30	935.1
4	15.4463	154	17.9840	717.2	40	911.2
5	14.8144	147.7	17.8636	712.4	50	910.1

Note: all costs are denominated in the unit claims auditing cost. The true dollar costs can be calculated by multiplying the costs presented in this table by the dollar amount of claims auditing cost per claim.

The cost curves showing the total costs at each step of the learning process for the two methods using the current test set (calculated as in Table 41) are plotted in Figure 6 below. Figure 7 and Figure 8 plot cost curves for the two methods using a test set of 2000 claims and 5000 claims respectively. The only difference in the calculation processes for these larger test sets is that the misclassification cost equals the same percentage of misclassification times the same unit misclassification cost times the *new* size of the test set (2000 and 5000 respectively).

Figure 6 Cost Curves using the Current Test Dataset (997 Claims) (Spanish Dataset)

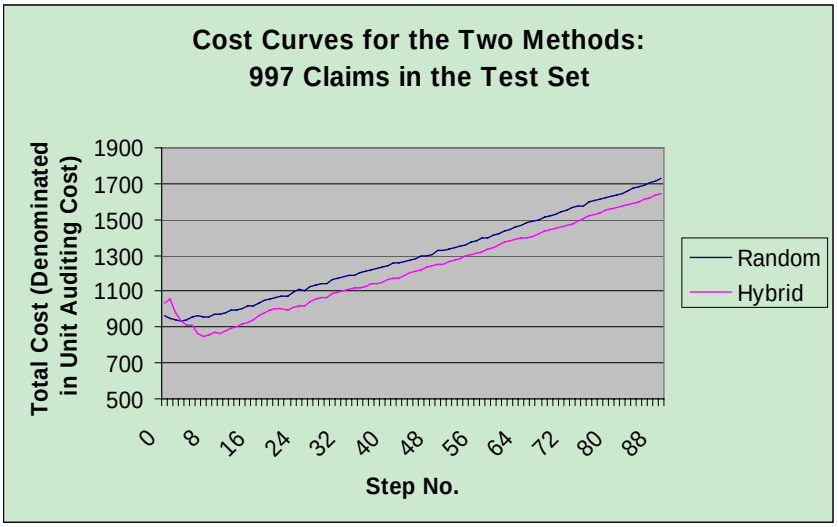


Figure 7 Cost Curves using a Test Dataset Containing 2000 Claims (Based on Spanish Dataset)

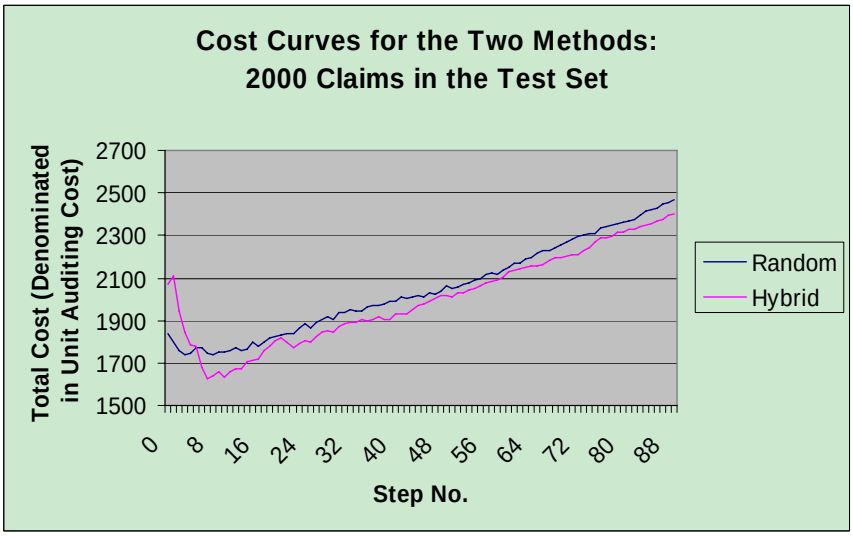
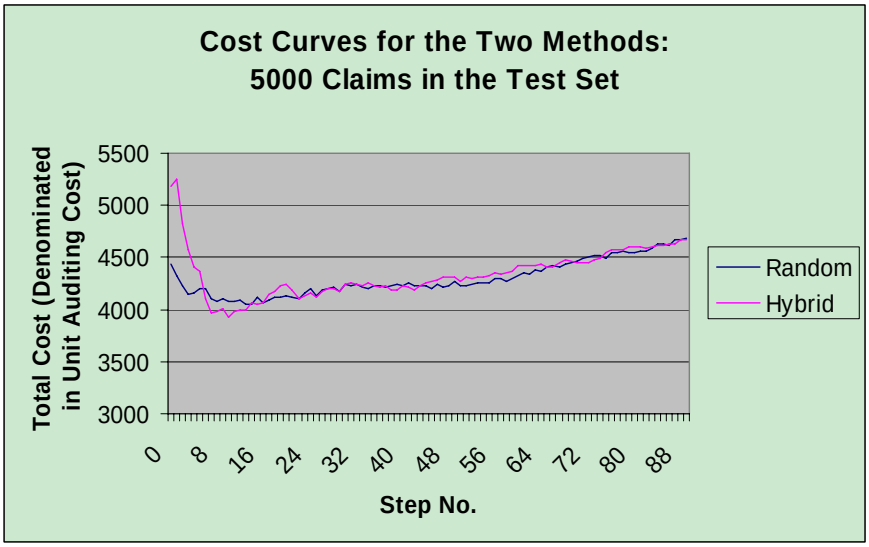


Figure 8 Cost Curves using a Test Dataset Containing 5000 Claims (Based on Spanish Dataset)



From Figure 6 to Figure 8, we can see that the hybrid method is more cost-effective than the random sampling approach except for the very beginning of the learning curve where the hybrid method has a low accuracy in both classes due to the inaccurate initial training sample. As the size of the test set grows bigger, the advantage of the hybrid method naturally diminishes since the total misclassification cost will be much higher than the total labeling cost and the initial savings on the labeling cost from the hybrid method will not make a difference. However, it is worth noting that in practice where there is likely to be a larger test set, the training set is likely to be larger as well. Therefore, the PRIDIT method can be run on the entire larger training set first and the initial PRIDIT-labeled training sample selected can also be larger than what we have now (100 claims). In that case, we will most likely see cost curves as presented in Figure 6 rather than in Figure 7 and Figure 8. Since we do not have a larger insurance fraud dataset, we cannot test this hypothesis empirically here.

From the above figures, we can see that according to the accuracy criterion, the hybrid method performs better than the random sampling approach at the beginning of the training process, but as more correctly classified information is obtained, the correct labels dominate and the performance of the hybrid method is adversely affected by the inaccurate labels in the initial training sample. However, as discussed previously, a cost-sensitive criterion seems to be more appropriate in this context. Traditional active learning usually only considers classification accuracy rather than cost-sensitive criteria mainly due to the follow two reasons: 1) all the training examples are obtained at the same cost and thus only the total number of training examples matter, 2) the two types of misclassification costs are not differentiated and are implicitly assumed to be the same. Since the hybrid method does use training examples obtained at different costs and the two types of misclassifications do need to be differentiated, a cost-sensitive criterion more precisely describes the learning objective. As we can see above, if we instead use the cost-sensitive criterion to assess the different learning methods, it turns out that the hybrid method is more efficient. This is because in this context, the labeling cost (i.e., the auditing cost) is quite high compared with misclassification costs.

5.3.4 Conclusion and discussions

In the above, I have studied the problem of using information provided by an unsupervised method, PRIDIT, to enhance a supervised classification model and improve its performance. I have mainly focused on using the ranking of summative scores and the classification labels provided by the PRIDIT analyses to prepare the initial training sample. The objective of maximizing classification accuracy is used to assess the

supervised model with an "informative" initial training sample. Both the objective of maximizing classification accuracy and the cost-sensitive objective have been employed to evaluate the performance of the hybrid method with a PRIDIT-labeled initial sample.

In the hybrid method discussed in this chapter, the PRIDIT-labeled examples and the examples with true labels are treated in the same manner during the learning process. Other approaches can be employed to combine these examples. For example, different weights can be assigned to these examples to reflect the "credibility" of their labels. In this case, an example with a true label will get a higher weight than an example with a PRIDIT-assessed label. More generally, this falls into the category of training from labeled and unlabeled data. A few studies have examined this problem (Kothari and Jain, 2003, Nigam et al., 2000).

Besides preparing the initial training sample, the information provided by the PRIDIT analysis may also be used to design the algorithm for the acquisition of additional training examples to use in active learning. In order to improve learning efficiency, training examples are to be selected in the sequence of decreasing amount of information. Current active learning methods (Uncertainty sampling, QBC, Bootstrapped-LV, etc.) are based on "ambiguity", i.e. the most informative examples are those that the current model (classifier) is most uncertain about, implying additional information to be extracted. PRIDIT scores and rankings of these training examples may also be used to capture this ambiguity. These issues are further discussed as future research avenues in Chapter 6.

5.4 FRAUD RATE ESTIMATION

5.4.1 Introduction to fraud rate estimation

As we mentioned in Chapter 1, insurance fraud data has the unique characteristic that the nature of the insurance claims (fraud or not) is private knowledge to the insureds which they often try to conceal from the insurance companies. Since most of the cases are not brought to court for settlement, the true value of the fraud variable for these claims may not be obtainable at all. This brought up the controversy in assessing the amount of fraud in the claims data. One current solution is to implement a screening method and pick out the most suspicious claims for more-in-depth auditing in order to construct a sample from which the fraud rate can be estimated. The claims auditing process incurs large expenses (time, human resources, and economic resources) and can only produce a limited amount of data. Moreover, it is difficult to obtain an accurate estimate of the fraud rate directly based on the audited sample of claims since this sample is likely to be biased (claims that are audited and thus included in the sample are more likely to be fraudulent claims than those from a random sample) (Pinquet et al., 2007). In this chapter, I propose a new method for fraud rate estimation by exploiting the PRIDIT method. The resulting estimate can help the insurance industry and the regulatory agencies to understand the magnitude of insurance fraud better and may be used for different analyses in insurance fraud detection (e.g., impact of insurance fraud, comparison among companies and countries, public policy issues, etc.) and in particular, it can be used in claim classification based on the PRIDIT analysis or other methods.

5.4.2 A simple estimation of the fraud rate based on relations among expected class variable scores, discriminatory power measures, and the fraud rate

In this subsection, I develop a simple method to estimate the fraud rate in a claims dataset by exploiting the relationship among different components of the PRIDIT method. Under the PRIDIT method, the relationship between average variable scores, fraud rate θ, and discriminatory power measure A_t can be exploited to estimate θ. More specifically, as in equation (3) and (4) we have E [B_t| Class 1] = (θ -1)A_t and E [B_t| Class 2] = θA_t, where class 1 is the fraud class and class 2 is the non-fraud class. Thus, we have θ = E [B_t| Class 2]/(E [B_t| Class 2] - E [B_t| Class 1]). For each variable t, we can obtain an estimate $\hat{\theta}_t$ of θ from average variable scores: $\hat{\theta}_t = \frac{\overline{B}_{2t}}{\overline{B}_{2t} - \overline{B}_{1t}}$, where $\overline{B}_{1t}$, $\overline{B}_{2t}$ are the average variable t scores for class 1 and class 2 claims respectively. Then, we can obtain a final estimate of θ based on the estimates $\hat{\theta}_t$'s, e.g., taking the average of the $\hat{\theta}_t$'s, or $\hat{\theta} = \frac{1}{T}\sum_{t=1}^{T}\hat{\theta}_t$.

However, in order to implement the above estimation procedure, we need to know the class membership of the claims (fraud or non-fraud) first in order to calculate the mean variable scores for each class. Two approaches can be adopted to tackle this problem. The first method is to rely on a subsample of claims where the nature of the claims (fraudulent or not) are known (e.g., obtained costly by asking insurance adjusters/investigators to label them). Under this method, average variable scores for each class are calculated only from the subset of claims with known class membership,

and an estimate of fraud rate obtained based on the subsample is used as the fraud rate estimate for the entire dataset (or entire population in practice).

The second method is to make use of an initial classification by assigning the claims with extreme suspicion scores (by PRIDIT) to the respective predicted classes (highest scores to the low suspicion class or the non-fraud class and lowest scores to the high suspicion class or the fraud class), which we know is of high accuracy based on empirical experiments (see Table 35 and Table 36). The estimation procedure is then implemented using only these claims with their PRIDIT-assessed class membership. Different proportions of claims with extreme suspicion scores can be selected for estimation and the estimate resulting in the minimum variance can be used as the final estimate of the fraud rate.

In the next subsection, I focus on the first proposed method and present the empirical results on the two insurance fraud datasets. I also discuss the difficulties of implementing the second proposed method and possible solutions to them in subsection 5.4.4.

5.4.3 Empirical evaluation of the fraud rate estimation method

In order to calculate average variable scores for each class, one can select a subsample of claims and obtain the true class memberships for this subsample. Since it is costly to obtain the true classes of the claims, the larger the size of the subsample, the higher is the cost incurred in the fraud rate estimation procedure. However, using a subsample of a larger-size is likely to reduce the estimation variance and thus to increase the estimation accuracy. Therefore, it is useful to compare results from subsamples of

various sizes. The subsamples can be constructed through two different approaches. Under the first approach, claims are ranked by their PRIDIT summative scores and each subsample is composed of an extreme subset of claims (claims with positive extremes scores and those with negative extreme scores). The second approach is to select subsamples of different sizes randomly. These two approaches are respectively described in more detail below and empirical results are also presented.

Specifically, under the first approach, the subsamples are composed of claims with highest (positive) and lowest (negative) extreme PRIDIT summative scores within a certain percentage of the entire dataset. This percentage is chosen to range from 5% to 45% with increments of 5%. Therefore, when the percentage is 5%, 10% of the data is actually used (claims with the highest 5% PRIDIT scores plus claims with the lowest 5% PRIDIT scores). For each subsample, I calculate the estimated theta from each predictor variable and then take the average of these theta estimates to be the final estimate. To avoid undue influence of idiosyncratic variation (resulting from small samples), since theta is a proportion of the sample, I normalize all theta estimates (from each variable and each subsample) to be between 0 and 1 before taking the average. The thus obtained results are presented in Table 42 and Table 43.

Since true class memberships (for the subsamples) are used to derive the above theta estimates, a naïve estimate can be obtained by simply counting the number of fraudulent claims in the labeled subsamples and use the resulting fraud rate for the subsamples as the estimated fraud rate for the entire dataset. The naïve estimates are also included in the tables below.

Table 42 Fraud Rate Estimation using PRIDIT-Ranked Subsamples (Spanish Dataset)

Percentage of the Data Used (%)	Mean	Std Dev	Variance	Median	Range	Interquartile Range	Naïve Estimate
10	0.4481	0.3968	0.1574	0.3135	1.0000	0.7934	0.4724
20	0.6147	0.3612	0.1305	0.6685	1.0000	0.6865	0.5088
30	0.4752	0.2981	0.0889	0.4458	0.9440	0.5413	0.4908
40	0.4510	0.2863	0.0820	0.4078	1.0000	0.3306	0.5157
50	0.5632	0.3121	0.0974	0.5354	1.0000	0.4944	0.5025
60	0.5598	0.2579	0.0665	0.5487	1.0000	0.3200	0.4921
70	0.4297	0.2361	0.0557	0.4561	0.9430	0.2901	0.4946
80	0.4469	0.1841	0.0339	0.4490	0.8104	0.2059	0.5028
90	0.4370	0.1719	0.0296	0.4974	0.6699	0.0896	0.4969

Note: This dataset has 1995 claims and the true theta equals 49.97% for this dataset.

From Table 42 we can see that as the size of the subsample increases, the variance of the theta estimate is continuously reduced, and the estimate becomes closer to the true value. However, it is worth noting that the variance is reduced sharply fairly early on, which indicates that even with a small sample (e.g., less than half of the data), a relatively accurate and reliable theta estimate can be secured.

Table 43 Fraud Rate Estimation using PRIDIT-Ranked Subsamples (U.S.A. Dataset)

Percentage of the Data Used (%)	Mean	Std Dev	Variance	Median	Range	Interquartile Range	Naïve Estimate
10	0.2670	0.2973	0.0884	0.1401	0.8296	0.5036	0.3022
20	0.3929	0.3511	0.1233	0.3716	1.0000	0.4154	0.3262
30	0.2829	0.2760	0.0762	0.1544	0.7932	0.4558	0.3126
40	0.2551	0.2373	0.0563	0.1823	0.6856	0.4376	0.3041
50	0.2336	0.2172	0.0472	0.1546	0.6459	0.3519	0.3104
60	0.2323	0.2222	0.0494	0.1776	0.7742	0.3920	0.3027
70	0.2574	0.2106	0.0444	0.2382	1.0000	0.1946	0.2901
80	0.2734	0.2241	0.0502	0.2032	1.0000	0.2582	0.2895
90	0.2875	0.1945	0.0378	0.2618	1.0000	0.1184	0.2891

Note: This dataset has 1399 claims and the true theta equals 28.31% for this dataset.

Table 43 presents the estimates for the U.S.A. insurance fraud dataset. Unlike the Spanish dataset, this dataset has a fraud rate of 28.31%. We find a pattern similar to the U.S.A. dataset. More specifically, the minimum variance estimate is very close to the true

fraud rate, although the most accurate estimate occurs when using almost the entire set of data (90% of the data). However, the estimate produced using a small amount of data also turns out to be quite accurate. Moreover, the naïve estimates seem to work well in the fraud rate estimation. However, I will later discuss further how naïve estimates become extremely inaccurate when the nature of the fraud datasets is taken into account (see Table 50 to Table 53).

In the above, subsamples are selected by subsequently including claims according to PRIDIT summative scores. Alternatively, subsamples of various sizes can be selected randomly from the entire dataset. I take the same total number of subsamples as before and each subsample is of a size similar to the subsamples taken before as well. The subsamples range from the size of 200 claims (around 10% of the data) to 1800 claims (around 90% of the data) with increments of 200 claims for the Spanish dataset (i.e., totally 9 subsamples) and range from 140 claims (around 10% of the data) to 1260 claims (around 90% of the data) at intervals of 140 claims for the U.S.A. dataset (i.e., totally 9 subsamples). The results from the random subsamples are presented below in Table 44 and Table 45. We can see that the method of randomly selecting subsamples works similarly to the PRIDIT-ranked selection of samples when true class memberships are used in the estimation. However, when subsamples with true class memberships are not available, we may have to rely on PRIDIT-labeled subsamples to implement the estimation, as introduced in section 5.4.2. In this case, we may only use PRIDIT-ranked selection of samples since only these subsamples will guarantee satisfactory accuracy (as shown in Table 35 and Table 36). Random subsamples will no longer work in this case. More discussions regarding this are presented in section 5.4.4.

Table 44 Fraud Rate Estimation using Random Subsamples (Spanish Dataset)

Percentage of the Data Used (%)	Mean	Std Dev	Variance	Median	Range	Interquartile Range	Naïve Estimate
10	0.6388	0.2837	0.0805	0.6499	1	0.4301	0.4750
20	0.4705	0.2744	0.0753	0.4814	1	0.3664	0.5150
30	0.5197	0.1998	0.0399	0.4979	0.7849	0.2327	0.5017
40	0.4417	0.1500	0.0225	0.4736	0.6005	0.1213	0.5100
50	0.5601	0.1762	0.0311	0.5219	0.7289	0.0599	0.4920
60	0.5025	0.0902	0.0081	0.5117	0.3755	0.1376	0.4975
70	0.4988	0.1186	0.0141	0.5271	0.5098	0.0808	0.5193
80	0.4584	0.1172	0.0137	0.4881	0.5275	0.0670	0.5038
90	0.5331	0.1235	0.0153	0.5029	0.5716	0.0419	0.4961

Note: This dataset has 1995 claims and the true theta equals 49.97% for this dataset.

Table 45 Fraud Rate Estimation using Random Subsamples (U.S.A. Dataset)

Percentage of the Data Used (%)	Mean	Std Dev	Variance	Median	Range	Interquartile Range	Naïve Estimate
10	0.4071	0.2797	0.0783	0.3348	1.0000	0.3352	0.2929
20	0.2057	0.1450	0.0210	0.2071	0.5172	0.1656	0.2500
30	0.1919	0.1359	0.0185	0.2011	0.5167	0.2109	0.2690
40	0.3275	0.2625	0.0689	0.2676	1.0000	0.2078	0.2875
50	0.3470	0.2151	0.0463	0.3281	1.0000	0.1505	0.2800
60	0.3280	0.2471	0.0610	0.2914	1.0000	0.1103	0.2893
70	0.2906	0.1788	0.0320	0.2671	1.0000	0.1011	0.2653
80	0.2497	0.1307	0.0171	0.2623	0.4968	0.0478	0.2795
90	0.3273	0.1666	0.0277	0.2916	1.0000	0.0397	0.2841

Note: This dataset has 1399 claims and the true theta equals 28.31% for this dataset.

Most times the true nature of claims (i.e., fraudulent or not) is not known. Therefore, experts' (e.g., insurance adjusters, special investigators, etc.) assessments of the claims are often used in place of the true labels. However, these human assessments are not necessarily consistent from one assessor to another and thus assessments of "true" fraud within a dataset may themselves be ambiguous and lead to inaccurate evaluation criteria for the estimation methods. In fact, the inconsistency in these human assessments (such as the opinions of insurance adjusters and insurance investigators) has been noted in Brockett et al. (1998) and Brockett et al. (2002). The inconsistency is also reflected in

the coding of the "fraud" variable in real world datasets (Viaene et al., 2002). In the U.S.A. insurance claims dataset, suspicion levels of the claims are rated by human experts on a zero to ten point scale with zero representing "clearly non-fraud" and ten representing "highest suspicion of fraud." In the tables above for the U.S.A. dataset, I used a suspicion score of four and above as the cutoff for fraudulent claims. Below, the estimation results are calculated using a more stringent cutoff level of suspicion score of seven and above being classified as fraud, and these results are shown in Table 46 and Table 47.

From Table 46 and Table 47, we can see that it is more difficult for the proposed PRIDIT-based method to estimate the even smaller fraud rate obtained using the more stringent fraud classification criterion. This is mainly because that the PRIDIT method tends to over-predict the fraud class or the minority class and thus decreases the accuracy in the non-fraud class.

Table 46 Fraud Rate Estimation using PRIDIT-Ranked Subsamples (U.S.A. Dataset using Suspicion Score of Seven and Above as the Fraud Indicator)

Percentage of the Data Used (%)	Mean	Std Dev	Variance	Median	Range	Interquartile Range	Naïve Estimate
10	0.1973	0.2996	0.0897	0.0000	1.0000	0.3371	0.1007
20	0.2170	0.3079	0.0948	0.0397	1.0000	0.3162	0.1111
30	0.2580	0.3119	0.0973	0.1733	0.9639	0.4015	0.1146
40	0.1891	0.2241	0.0502	0.0513	0.7038	0.3521	0.1163
50	0.1219	0.1771	0.0314	0.0317	0.6164	0.1743	0.1087
60	0.1015	0.1671	0.0279	0.0072	0.5112	0.0700	0.1001
70	0.0927	0.1253	0.0157	0.0594	0.4725	0.1302	0.0940
80	0.1386	0.2417	0.0584	0.0168	1.0000	0.2273	0.0947
90	0.0779	0.0836	0.0070	0.0560	0.3342	0.0865	0.0921

Note: This dataset has 1399 claims and the true theta equals 8.79% for this dataset.

Table 47 Fraud Rate Estimation using Random Subsamples (U.S.A. Dataset using Suspicion Score of Seven and Above as the Fraud Indicator)

Percentage of the Data Used (%)	Mean	Std Dev	Variance	Median	Range	Interquartile Range	Naïve Estimate
10	0.1063	0.2130	0.0454	0.0295	1.0000	0.1159	0.0714
20	0.1653	0.3083	0.0950	0.0345	1.0000	0.1322	0.1000
30	0.1669	0.2198	0.0483	0.1137	1.0000	0.2271	0.0833
40	0.1930	0.3100	0.0961	0.0993	1.0000	0.1180	0.0875
50	0.1411	0.1477	0.0218	0.1303	0.6783	0.1175	0.0914
60	0.1270	0.2068	0.0428	0.0764	1.0000	0.1228	0.0762
70	0.0901	0.0964	0.0093	0.0762	0.4685	0.0831	0.0745
80	0.0885	0.0867	0.0075	0.0794	0.4600	0.0466	0.0920
90	0.0903	0.0836	0.0070	0.0880	0.3520	0.0723	0.0921

Note: This dataset has 1399 claims and the true theta equals 8.79% for this dataset.

As discussed in detail in the opening chapters, PRIDIT makes use of all predictor variables (or "red flags" for fraud) to calculate a summative score for each claim. Different weights are assigned to these predictor variables to reflect their respective importance. In Table 48 and Table 49 below, theta estimates derived from variables of lower weights (less importance) are deleted in order to further increase the accuracy of the proposed methods. In other words, instead of averaging the theta estimates from all variables, only those from the most important variables are included to obtain the final theta estimate. We can see from Table 48 and Table 49 that theta estimates across subsamples center more tightly around the true value with reduced estimation variances when less important variables are excluded.

Table 48 Fraud Rate Estimation Using PRIDIT-Ranked Subsamples with Variables of High Importance (Spanish Dataset)

Percentage of the Data Used (%)	Mean	Std Dev	Variance	Median	Range	Interquartile Range	Naïve Estimate
10	0.4972	0.3407	0.1161	0.4250	0.8534	0.6521	0.4724
20	0.5179	0.3270	0.1069	0.4174	0.9339	0.5642	0.5088
30	0.5081	0.3235	0.1047	0.4464	0.8915	0.5072	0.4908
40	0.5374	0.3327	0.1107	0.4559	1.0000	0.5640	0.5157
50	0.5499	0.3054	0.0933	0.5070	0.9772	0.4944	0.5025
60	0.5480	0.2444	0.0598	0.4603	0.8336	0.3197	0.4921
70	0.5457	0.1974	0.0390	0.5169	0.6864	0.2030	0.4946
80	0.5464	0.1376	0.0189	0.5401	0.4562	0.1500	0.5028
90	0.5264	0.0617	0.0038	0.5254	0.2242	0.0509	0.4969

Note: This dataset has 1995 claims and the true theta equals 49.97% for this dataset.

Table 49 Fraud Rate Estimation Using PRIDIT-Ranked Subsamples with Variables of High Importance (U.S.A. Dataset)

Percentage of the Data Used (%)	Mean	Std Dev	Variance	Median	Range	Interquartile Range	Naïve Estimate
10	0.2997	0.3141	0.0987	0.1801	0.8296	0.5036	0.3022
20	0.3296	0.3118	0.0972	0.1951	1.0000	0.4293	0.3262
30	0.3070	0.2717	0.0738	0.1685	0.7542	0.4422	0.3126
40	0.2810	0.2409	0.0580	0.1985	0.6856	0.4270	0.3041
50	0.2743	0.2246	0.0504	0.2084	0.6459	0.3784	0.3104
60	0.2585	0.2344	0.0549	0.2000	0.7742	0.3836	0.3027
70	0.2294	0.1594	0.0254	0.1995	0.5223	0.2207	0.2901
80	0.2992	0.2280	0.0520	0.2412	1.0000	0.2793	0.2895
90	0.3067	0.2067	0.0427	0.2734	0.9689	0.1103	0.2891

Note: This dataset has 1399 claims and the true theta equals 28.31% for this dataset.

I now turn to the examination of the applicability of naïve estimates in the context of insurance fraud detection. In the above tables, we see that although the interpretation and reliability of naïve estimates may be questionable, it seems to perform very well. However, it is worth noting that the naïve estimates will only work under the condition that the sample of claims examined has the same expected fraud rate as that of the population of claims (e.g., when the sample is randomly selected). This condition is not always satisfied (Pinquet et al., 2007).

In many cases, the sample of claims selected to obtain true classes (by claims auditing) is not randomly chosen. Rather, it consists of suspicious claims that are chosen according to the outcome of the initial screening procedure. Therefore, we usually end up with a biased sample with more fraudulent claims than there will be in a randomly selected sample or in the original population. This may cause problems for supervised methods where this sample is used as the training sample to estimate and predict the population. Similarly, the naïve estimates derived directly from subsamples of the biased sample will only accurately predict the fraud rate of the *biased sample*, but not the population. Since we are ultimately interested in the fraud rate of the entire population (e.g., the entirety of an insurance company's claim file), the naïve estimates are hardly useful.

By contrast, the PRIDIT-based method using a small amount of true class memberships should still produce fairly accurate theta estimate for the population even when we have a biased sample. This is because, unlike the naïve estimation, the PRIDIT method uses variable scores, not the true class memberships directly, to estimate the fraud rate. The PRIDIT method (with its little cost) can still be run on the entire population to generate variable scores, and a small amount of true class memberships from the biased sample are used only to identify the two classes so that average class variable scores can be calculated.

Below I empirically test the estimation of population fraud rate from a biased sample. Since the entire populations of claims (for both the Spanish and the U.S.A. datasets) are not available for use in my experiments, sample datasets play the role of the populations. As described before, the Spanish dataset has a fraud rate of 49.97% and the

U.S.A. dataset has a fraud rate of 28.31%. I then create two smaller biased samples (i.e., "small samples") from these two datasets respectively. Since the fraud rates of the original datasets are already very large (because they are actually audited samples which I use as populations here), I construct the small samples so that the fraud rates are much lower than the original datasets. This is actually opposite to the real world situation, but the logic should work in the same way as long as the fraud rates of the sample (or the small samples used here) and the population (or the original datasets used here) are very different (i.e., the small samples are biased samples from the original datasets).

The small sample created for the Spanish dataset has 1000 claims and the fraud rate is 25%. The small sample from the U.S.A. dataset has 800 claims and the fraud rate is 12.5%. These numbers are chosen for convenience and they should not affect the empirical results. I implement the same estimation procedure as described in detail previously to estimate the fraud rate of the *original* datasets. More specifically, variable scores are first obtained by running PRIDIT on the *original* datasets. After the two small samples are created, subsamples of different sizes are randomly taken from the *small* samples and true class memberships for these subsamples are then used to calculate the average variable scores for each class. The sizes of subsamples range from 10% to 100% of the small samples. Naïve estimates are also obtained from these subsamples. The results are presented in Table 50 and Table 51, where we can see that the PRIDIT-based estimation method indeed predicts rather accurately the fraud rate of the original datasets, while the naïve estimation, as expected, only predicts the fraud rates of the artificially created small samples which is very different from those of the original datasets.

Table 50 Population Fraud Rate Estimation for a Population Having a True Fraud Rate of 49.97% using Random Subsamples Having a Fraud rate of 25% (Spanish Dataset)

Percentage of the Data Used (%)	Mean	Std Dev	Variance	Median	Range	Interquartile Range	Naïve Estimate
10	0.2709	0.2465	0.0607	0.2928	0.8726	0.4456	0.2100
20	0.4339	0.3668	0.1346	0.4306	1.0000	0.5937	0.3000
30	0.4628	0.3140	0.0986	0.4556	1.0000	0.4198	0.2333
40	0.4972	0.2731	0.0746	0.4775	1.0000	0.1081	0.2575
50	0.4895	0.2750	0.0756	0.4925	1.0000	0.2398	0.2440
60	0.4406	0.2127	0.0452	0.4207	0.8302	0.2850	0.2583
70	0.4178	0.2053	0.0421	0.4564	0.6746	0.2789	0.2557
80	0.5108	0.2107	0.0444	0.4599	0.7141	0.1740	0.2525
90	0.4376	0.1636	0.0268	0.4895	0.6704	0.1939	0.2611
100	0.4824	0.1855	0.0344	0.4945	0.9342	0.1568	0.2500

Note: This dataset has 1995 claims and the true theta equals 49.97% for this dataset. However, subsamples used for estimation are taken from a sample of the original dataset having fraud rate 25%.

Table 51 Population Fraud Rate Estimation for a Population Having a True Fraud Rate of 28.31% using Random Subsamples Having a Fraud rate of 12.5% (U.S.A. Dataset)

Percentage of the Data Used (%)	Mean	Std Dev	Variance	Median	Range	Interquartile Range	Naïve Estimate
10	0.1738	0.2878	0.0828	0.0404	1.0000	0.1774	0.1000
20	0.2495	0.3046	0.0928	0.1906	1.0000	0.3168	0.1375
30	0.2483	0.2445	0.0598	0.2394	1.0000	0.3730	0.1583
40	0.4069	0.3782	0.1430	0.3638	1.0000	0.7265	0.1281
50	0.2789	0.2766	0.0765	0.2358	0.9405	0.3390	0.1100
60	0.3306	0.2672	0.0714	0.2837	1.0000	0.2133	0.1375
70	0.2792	0.2626	0.0689	0.2460	1.0000	0.2348	0.1107
80	0.2934	0.1903	0.0362	0.2957	0.7248	0.2292	0.1250
90	0.2444	0.1862	0.0347	0.2347	0.6152	0.2875	0.1250
100	0.2462	0.1793	0.0322	0.2621	0.6111	0.2422	0.1250

Note: This dataset has 1399 claims and the true theta equals 28.31% for this dataset. However, subsamples used for estimation are taken from a sample of the original dataset having fraud rate 12.5%.

Since it is shown in Table 48 and Table 49 that the PRIDIT-based estimation method tends to perform better when variable scores from predictors with low weights are not used in the final estimation, I repeat the above analyses excluding variable scores

from lower-weighted variables. The results are shown in Table 52 and Table 53. We can see that in general the performance is even better, especially for the Spanish dataset.

Table 52 Population Fraud Rate Estimation for a Population Having a True Fraud Rate of 49.97% using Random Subsamples Having a Fraud rate of 25%, Using Variables of High Importance (Spanish Dataset)

Percentage of the Data Used (%)	Mean	Std Dev	Variance	Median	Range	Interquartile Range	Naïve Estimate
10	0.4023	0.2306	0.0532	0.3483	0.8726	0.1920	0.2100
20	0.4721	0.2692	0.0725	0.4697	1.0000	0.2352	0.3000
30	0.5732	0.2243	0.0503	0.5597	0.7113	0.2367	0.2333
40	0.4928	0.1932	0.0373	0.4694	0.7304	0.0930	0.2575
50	0.4603	0.0991	0.0098	0.4618	0.3219	0.1169	0.2440
60	0.4874	0.1397	0.0195	0.4721	0.4845	0.1595	0.2583
70	0.5326	0.1028	0.0106	0.5175	0.2803	0.1444	0.2557
80	0.4985	0.1125	0.0127	0.5037	0.3537	0.1445	0.2525
90	0.5096	0.0614	0.0038	0.5209	0.1812	0.0863	0.2611
100	0.5151	0.0653	0.0043	0.5276	0.2013	0.0658	0.2500

Note: This dataset has 1995 claims and the true theta equals 49.97% for this dataset. However, subsamples used for estimation are taken from a sample of the original dataset having fraud rate 25%. Only variables of high importance are included.

Table 53 Population Fraud Rate Estimation for a Population Having a True Fraud Rate of 28.31% using Random Subsamples Having a Fraud rate of 12.5%, Using Variables of High Importance (U.S.A. Dataset)

Percentage of the Data Used (%)	Mean	Std Dev	Variance	Median	Range	Interquartile Range	Naïve Estimate
10	0.2367	0.3307	0.1094	0.0886	1.0000	0.3089	0.1000
20	0.2371	0.2224	0.0495	0.2273	0.7803	0.3168	0.1375
30	0.3254	0.2577	0.0664	0.2982	1.0000	0.2110	0.1583
40	0.4860	0.3724	0.1387	0.3705	1.0000	0.6192	0.1281
50	0.3501	0.2250	0.0506	0.3091	0.8632	0.2152	0.1100
60	0.3319	0.1753	0.0307	0.3210	0.5645	0.2111	0.1375
70	0.3261	0.2852	0.0813	0.2671	1.0000	0.1108	0.1107
80	0.3277	0.1395	0.0195	0.3058	0.5972	0.1464	0.1250
90	0.3034	0.1684	0.0284	0.2790	0.6152	0.1277	0.1250
100	0.2896	0.1545	0.0239	0.2965	0.5768	0.0821	0.1250

Note: This dataset has 1399 claims and the true theta equals 28.31% for this dataset. However, subsamples used for estimation are taken from a sample of the original dataset having fraud rate 12.5%. Only variables of high importance are included.

5.4.4 Discussions

In section 5.4, a PRIDIT-based estimation method using subsamples of claims with known class memberships is used for fraud rate estimation. I have shown that the PRIDIT-based estimation method performs well and is especially of value when only a biased sample from the population is available for the estimation.

When it is impossible or cost-prohibitive to obtain a subsample of claims with known class memberships, the PRIDIT classification of the subsamples can be used alternatively. Under this method, extreme subsets of various sizes can be selected to construct subsamples. The claims in these subsamples are classified by their PRIDIT summative scores, and the average variable scores for each such defined class are calculated accordingly. However, since the PRIDIT method is an unsupervised method and tends to over-predict claims in the fraud class as compared to those classified by the insurance experts, theta estimation based on the PRIDIT classifications is adversely affected by the inaccuracy in the initial PRIDIT classification and thus is not very accurate. In order to improve the fraud rate estimation based on this approach, the PRIDIT performance needs to be improved first. One possible solution is to use a small sample of known classification and the hybrid learning method proposed in section 5.3 to improve the initial PRIDIT classification first. Then we may be able to use these classifications to estimate theta and thus reduce the number of labels needed to arrive at a reasonable theta estimate.

Besides using the estimation method introduced above, the estimation of θ may be further improved by exploiting the relationship $E[B_t|\text{ Class 2}] - E[B_t|\text{ Class 1}] = A_t$. Since

A_t is linearly related to the weights, it can be estimated without knowing θ. Thus, a regression framework can be used to obtain the estimate by minimizing the value $(\overline{B}_{2t} - \overline{B}_{1t} - \hat{A}_t)$ under squared loss. Other more sophisticated estimation methods can also be designed in this spirit.

CHAPTER 6

Limitations and Future Research

6.1 LIMITATIONS OF THIS RESEARCH

This dissertation proposes unsupervised and hybrid methodologies to solve important problems in insurance fraud detection, including effective fraud detection from an accuracy stand point, from a cost-sensitive stand point, and the problem of fraud rate estimation. All of the methods developed in this dissertation are empirically demonstrated using real datasets (in insurance fraud detection or other relevant application domains). Insurance fraud detection is a very hard problem by its very nature: it is very costly to obtain a sample of claims where the claims' nature (fraud or not) is known in order to train a detection model, and even if this sample is obtained, these "known" labels are subject to error since they are only the best judgment of experts rather than court decisions. Both of these properties pose difficulties for building an effective model to identify fraudulent claims or to estimate the fraud rate. By proposing new methods, this dissertation aims at advancing the efforts in understanding various problems in insurance fraud detection. However, there are several limitations of this research.

First, the available insurance fraud datasets are limited and the available ones may not be the best for the purpose of assessing the proposed methods. For this reason, I am not able to use an insurance fraud dataset to demonstrate the performance of the general PRIDIT method. Instead, I use an income classification dataset. Moreover, as discussed in various places, the two insurance fraud datasets used in this dissertation are claimed to

have “true” labels. Although they are probably the best datasets available now, these labels are still assigned by experts’ judgment and thus are likely to be subject to errors. In fact, experts tend not to agree with each other on the nature of claims most of the time. Therefore, the assessment of the proposed methods against these “true” labels may not be entirely reliable. All results currently interpreted under the assumption that these “true” labels are accurate need to be interpreted with caution because this assumption may not hold. In addition, the datasets used here are likely to be “biased” samples from the entire claims set, which may also affect the generalizability of the empirical results.

Second, in this dissertation, I have mainly considered the direct economic costs associated with insurance fraud and insurance fraud detection. There are other costs that have not been incorporated in the analyses. For example, misclassifying a nonfraudulent claim to be fraudulent may result in consumer dissatisfaction, loss of reputation and possible “bad faith” lawsuits for the insurance company. These potential costs could also be important considerations in practice.

Third, this dissertation studies the methodology design for insurance fraud detection and fraud rate estimation. There are other interesting problems in insurance fraud detection and fraud detection in general yet to be investigated. For example, a body of theoretical literature has studied the incentives of insurance fraud, its deterrence, and its detection. The insights provided by the theoretical work have not been applied to the methodology design. Also, the inter-adjuster variability in fraud assessment has not been modeled and investigated. Moreover, there are other fraud detection applications to be studied, such as credit card fraud detection, telecommunication fraud detection, etc. These research areas are beyond the scope of this dissertation and are left unexamined.

6.2 FUTURE RESEARCH DIRECTIONS

This dissertation proposes some initial development of unsupervised and hybrid fraud detection methods in the context of active learning and cost-sensitive learning. Both of these areas are emerging research areas so further studies with many different research directions can be conducted to improve the current methods. More sophisticated methods for fraud rate estimation can also be developed. The possible further development of methodology will not only contribute to the insurance fraud detection literature, but could make contributions to the statistical learning literature in general, since some difficulties we see in insurance fraud detection may characterize a set problems and may give rise to a host of new learning methods. We discuss these future research directions in this section.

6.2.1 Improving the general PRIDIT method

The general unsupervised PRIDIT method can be improved along various directions. First, I now only consider a binary classification task, i.e., fraud or not. However, insurance fraud may actually be a continuous variable, e.g., the suspicion level of fraud. Therefore, the current method may be improved for multi-group classification, or even for predicting a continuous dependent variable. Second, I now only assume that there is a single latent dimension underlying the classification. However, there might be multiple latent dimensions to be considered (e.g., for different types of fraud). It will be interesting to see how this method can be extended to capture the multiple latent dimensions. Third, since true labels are some times available, this information may be

used to improve the originally entirely unsupervised PRIDIT method. For example, they may be used to select predictor variables, to modify weights, etc.

6.2.2 Improving the hybrid methods and the fraud rate estimation method

There are several interesting research ideas to improve the hybrid methods developed based on the PRIDIT method. First, the issue of whether there is a best subsample to be used for supervised learning is worth investigating. If we can make use of an unsupervised method to find a best subsample and train a supervised model using only this sample, we may save on the labeling cost and time, and thus may derive a more efficient learning method.

Second, how to most effectively combine the PRIDIT-labeled training examples and the true-labeled examples in training is an important question. We may use an idea similar to the concept of "Credibility" in the actuarial science literature to solve this problem. More specifically, these two types of examples can be weighted to reflect the "credibility" of their labels and these weights can be updated as the learning proceeds.

Third, PRIDIT may be used to develop an "informativeness" measure and to design algorithms to acquire additional training examples, along the lines of traditional active learning literature. This is a challenging problem and deserves further research. It is also interesting to examine why current active learning algorithms do not work well with the proposed informative initial training sample approach.

Fourth, the proposed cost matrices are currently used as assessment criteria for the hybrid methods after the learning is completed. Instead, a learning method taking into account these cost matrices during the learning process can be proposed.

Fifth, in insurance fraud detection, the labeling cost is the claims auditing cost. Therefore, we have the choice of incurring this cost beforehand to train a classification model, or incurring this cost as a type of misclassification cost when the model is applied to the test dataset. It is of interest to see which of these approaches are more efficient and effective. The findings can provide guidelines for insurance companies in their claims auditing and investigation procedure.

Finally, supervised learning methods other than logistic regression can be employed as the component supervised learning methods in developing the hybrid methods. Since these supervised learning methods do not necessarily have the same pattern of performance, it will be interesting to see whether and how the proposed methods will work differently with other supervised learning methods.

Other fraud rate estimation methods can also be developed. Under the proposed estimation method, a small subsample of claims with true classifications is used to implement the method. Later on, we can use initial PRIDIT classifications instead of true classifications, if the accuracy of initial PRIDIT classifications can be further improved. Also, we can develop more sophisticated estimation methods based on the one proposed here (e.g., a regression type of method). Moreover, we may investigate theoretically the properties of the proposed estimation method (e.g., if the estimate is consistent, if it is unbiased, etc.)

This concludes our discussion, although there are without doubt many other avenues of research that remain to be explored.

Bibliography

Abe, N., and Mamitsuka, H. (1998), "Query Learning Strategies Using Boosting and Bagging," *Proceedings of the Fifteenth International Conference on Machine Learning*.

Artís, M., Ayuso, M., and Guillén, M. (1999), "Modeling Different Types of Automobile Insurance Fraud Behavior in the Spanish Market," *Insurance Mathematics and Economics,* 24, 67-81.

Artís, M., Ayuso, M., and Guillén, M. (2002), "Detection of Automobile Insurance Fraud With Discrete Choice Models and Misclassified Claims," *Journal of Risk and Insurance*, 69, 325-340.

Beder, J. H., and Heim, R. C. (1990), "On the Use of RIDIT Analysis," *Psychometrika*, 55, 603-616.

Belhadji, E. B., Dionne, G., and Tarkhani, F. (2000), "A Model for the Detection of Insurance Fraud," *The Geneva Papers on Risk and Insurance,* 25, 517-538.

Bolton, R. J., and Hand, D. J. (2001), "Unsupervised Profiling Methods for Fraud Detection," Conference on Credit Scoring and Credit Control 7, Edinburgh, UK, Sept 5-7.

Bolton, R. J., and Hand, D. J. (2002), "Statistical Fraud Detection: A Review," *Statistical Science*, 17, 235 – 255.

Bond, E. and Crocker, K. (1997), "Hardball and the Soft Touch: The Economics of Optimal Insurance Contracts With Costly State Verification and Endogenous Monitoring Costs," *Journal of Public Economics*, 63, 239-264.

Boyer, M. (2000a), "Centralizing Insurance Fraud Investigation," *Geneva Papers on Risk and Insurance Theory,* 25, 159-178.

Boyer, M. (2000b), "Insurance taxation and insurance fraud," *Journal of Public Economic Theory*, 2, 101-134.

Boyer, M. (2001), "Mitigating Insurance Fraud: Lump-sum Awards, Premium Subsidies, and Indemnity Taxes," *Journal of Risk and Insurance*, 68, 403-436.

Boyer, M. (2004), "Overcompensation as a Partial Solution to Commitment and Renegotiation Problems: the Case of Ex Post Moral Hazard," *Journal of Risk and Insurance*, 71, 559-582.

Brause, R., Langsdorf, T. and Hepp, M. (1999), "Neural Data Mining for Credit Card Fraud Detection," *Proc. of the 11th IEEE International Conference on Tools with Artificial Intelligence*, 103–106.

Brockett, P. (1981), "A Note on Numerical Assignment of Scores to Ranked Categorical Data," *Journal of Mathematical Sociology*, 8, 91-101.

Brockett, P. L., Derrig, R. A., and Xia, X. (1998), "Using Kohonen's Self-Organizing Feature Map to Uncover Automobile Bodily Injury Claims Fraud," *Journal of Risk and Insurance*, 65, 245-274.

Brockett, P. L., Derrig, R. A., Golden, L. L., Levine, A. and Alpert, M. (2002), "Fraud Classification Using Principal Component Analysis of RIDITs," *Journal of Risk and Insurance*, 69, 341-372.

Brockett, P. L., and Golden, L. L. (1992), "A Comment on "Using Rank Values as an Interval Scale" by Dowling and Midgley," *Psychology and Marketing*, 9, 255-261.

Brockett, P. L., and Levine, A. (1977), "On a Characterization of RIDITs," *Annals of Statistics,* 5, 1245-1248.

Bross, I. (1958), "How to Use RIDIT Analysis," *Biometrics*, 14, 18-38.

Burge, P., and Shawe-Taylor, J. (2001), "An Unsupervised Neural Network Approach to Profiling the Behavior of Mobile Phone Users for Use in Fraud Detection," *Journal of Parallel and Distributed Computing*, 61, 915- 925.

Caudill, S., Ayuso, M., and Guillen, M. (2005), "Fraud Detection Using a Multinomial Logit Model with Missing Information," *Journal of Risk and Insurance*, 72, 539-550.

Chawla, N., Japkowicz, N. and Kotcz, A. (2004), "Editorial: special issue on learning from imbalanced data sets," ACM SIGKDD Explorations Newsletter, 6, 1-6.

Chipman, H., George, E. I. and McCulloch, R. E. (2006), Bayesian Ensemble Learning, Neural Information Processing Systems, 2006.

Clarke, M. (1989), "Insurance Fraud," *The British journal of Criminology*, 29, 1-20.

Clarke, M. (1990), "The Control of Insurance Fraud: A Comparative View," *The British Journal of Criminology*, 30, 1-23.

Colquitt, L. and Hoyt, R. (1997), ''An Empirical Analysis of the Nature and Cost of Fraudulent Life Insurance Claims," *Journal of Insurance Regulation*, 15, 451–479.

Cristianini, N. and Shawe-Taylor, J. (2000), *An Introduction to Support Vector Machines*, Cambridge, England: Cambridge University Press.

Crocker, K. and Morgan, J. (1998), "Is Honesty the Best Policy? Curtailing Insurance Fraud through Optimal Incentive Contracts," *Journal of Political Economy*, 106, 355-375.

Crocker, K., and Tennyson, S. (2002), "Insurance Fraud and Optimal Claims Settlement Strategies," *Journal of Law and Economics*, 45, 469-508.

Derrig, R. A. (2002), "Insurance Fraud," *Journal of Risk and Insurance*, 69, 271-288.

Derrig, R. A., and Ostaszewski, K. M. (1995), "Fuzzy Techniques of Pattern Recognition in Risk and Claim Classification," *Journal of Risk and Insurance,* 62, 447-482.

Derrig, R. A., and Weisberg, H. I. (1998), "AIB PIP Claims Screening Experimental Final Report – Understanding and Improving the Claim Investigation Process," AIB Cost Containment/Fraud Filing (DOI Docket R 98-41), Automobile Insurers Bureau of Massachusetts.

Derrig, R., Weisberg, H., and Chen, X. (1994), "Behavioral Factors and Lotteries under No-Fault with a Monetary Threshold: A Study of Massachusetts Automobile Claims," *Journal of Risk and Insurance*, 61, 245–76.

Derrig, R., and Zicko, V. (2002), "Prosecuting insurance fraud-A case study of the Massachusetts experience in the 1990s," *Risk Management and Insurance Review*, 5, 77-104.

Dionne, G. and Gagné, R. (2001), "Deductible contracts against fraudulent claims: evidence from automobile insurance," *The Review of Economics and Statistics*, 83, 290-301.

Dionne, G. and Gagné, R. (2002), "Replacement Cost Endorsement and Opportunistic Fraud in Automobile Insurance," *Journal of Risk and Uncertainty*, 24, 213–30.

Dionne, G., Guiliano, F. and Picard, P. (2003), "Optimal Auditing for Insurance Fraud," CREF 03-03, HEC Montreal.

Domingos, P. (1999), "MetaCost: A General Method for Making Classifiers Cost-sensitive," *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 155–164.

Duncan, D. B. (1955), "Multiple range and multiple F-tests," *Biometrics* 11, 1-42.

Elkan, C. (2001), "The Foundations of Cost-Sensitive Learning," *Proceedings of the IJCAI-01*, 973-978.

Fawcett, T., and Provost, F. (1997), "Adaptive fraud detection," *Data Mining and Knowledge Discovery*, 1, 291-316.

Golden, L. L., and Brockett, P. L. (1987), "The Effects of Alternative Scoring Techniques on the Analysis of Rank Ordered Categorical Data," *Journal of Mathematical Sociology*, 12, 383-414.

Grabec, I. (1991), "Modeling of Chaos by a Self-Organizing Neural Network," in *Artificial Neural Networks*, Elsevier, Amsterdam, 1, 151-156.

Hoyt, R. (1990), "The Effect of Insurance Fraud on the Economic System," *Journal of Insurance Regulation*, 8, 304–315.

Jin, Y., Rejesus, R. and Little, B. (2005), "Binary Choice Models for Rare Events Data: a Crop Insurance Fraud Application," *Applied Economics*, 37, 841-848.

Kohonen, T. (1982). "Self-Organized Formation of Topologically Correct Feature Maps," *Biological Cybernetics*, 43, 59-69.

Kohonen, T. (1989). "Self-Organizing Feature Maps," in Self-Organizing and Associative Memory, New York: Spring-Verlag.

Kohonen, T. (1990). "The Self-Organizing Map," *Proceedings of the IEEE*, 78, 1464-1480.

Kothari, R. and Jain, V. (2003), "Learning from Labeled and Unlabeled Data Using a Minimal Number of Queries," *IEEE Transactions on Neural Networks*, 14, 1496–1505.

Kou, Y., Lu, C., Sirwongwattana, S., and Huang, Y. (2004), "Survey of Fraud Detection Techniques," *Proceedings of the 2004 International Conference on Networking, Sensing, and Control*, 749-754.

Lewis, D. and Gale, W. A. (1994), "A Sequential Algorithm for Training Text Classifiers," *Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 3–12.

Loughran, D. (2005), "Deterring Fraud: The Role of General Damage Awards in Automobile Insurance Settlements," *Journal of Risk and Insurance*, 72, 551-575.

Major, J.A., and Riedinger, D. R. (2002), "EFD: A Hybrid Knowledge/ Statistical-Based System for the Detection of Fruad," *Journal of Insurance Regulation*, 69, 309-324.

Mookherjee, D. and Png, I. (1989), "Optimal Auditing, Insurance and Redistribution," *Quarterly Journal of Economics*, 104, 399-415.

Moreno, I., Vazquez, F., and Watt, R. (2006), "Can bonus-malus allieviate insurance fraud?" *Journal of Risk and Insurance*, 73, 123-151.

Nigam, K., McCallum, Thrun, A. S. and Mitchell, T (2000), "Text classification from labeled and unlabeled documents using EM," *Machine Learning*, 39, 103–134.

Pathak, J., Vidyarthi, N. and Summers, S. (2003), "A Fuzzy-based Algorithm for Auditors to Detect Element of Fraud in Settled Insurance Claims," Odette School of Business Administration.

Phua, C., Alahakoon, D. and Lee, V. (2004), "Minority report in fraud detection: classification of skewed data," *ACM SIGKDD Explorations Newsletter*, 6, 50-59.

Phua, C., Lee, V., Smith, K., and Gayler, R. (2005), "A Comprehensive Survey of Data Mining-based Fraud Detection Research," *Artificial Intelligence Review*.

Picard, P. (1996), "Auditing Claims in the Insurance Market With Fraud: The Credibility Issue," *Journal of Public Economics*, 63, 27-56.

Picard, P. (2000), "Economic Analysis of Insurance Fraud," In G. Dionne (ed.), Handbook of Insurance. Norwell: Kluwer Academic Press, pp. 315–360.

Pinquet, J., Ayuso, M., Guillen, M. (2007), "Selection Bias and Auditing Policies for Insurance Claims," *Journal of Risk and Insurance*, 74, 425-440.

Quinlan, J.R. (1993), *C4.5: Programs for Machine Learning*, San Mateo, California: Morgan Kaufmann.

Saar-Tsechansky, M. and Provost, F. (2004), "Active Sampling for Class Probability Estimation and Ranking," *Machine Learning*, 54, 153-178.

Seung, H.S., Opper, M., and Sompolinsky, H. (1992) "Query by committee," *Proceedings of the 5th Annual Workshop on Computational Learning Theory*, 287-294.

Stefano, B. and Gisella, F. (2001), "Insurance Fraud Evaluation: A Fuzzy Expert System," *Proceedings of IEEE International Fuzzy Systems Conference*, 1491-1494.

Sulzle, K., and Wambach, A. (2005), "Insurance in a Market for Credence Goods," *Journal of Risk and Insurance*, 72, 159-176.

Tennyson, S., and Salsas-Forn, P. (2002), "Claims Auditing in Automobile Insurance: Fraud Detection and Deterrence Objectives," *Journal of Risk and Insurance*, 69, 289-308.

Townsend, R. (1979), "Optimal Contracts and Competitive Markets With Costly State Verification," *Journal of Economic Theory*, 21, 265-293.

Vapnik, V. (1995), *The Nature of Statistical Learning Theory*, New York: Springer.

Vapnik, V. (1998), *Statistical Learning Theory*, New York: Wiley.

Viaene, S. and Dedene, G. (2004), "Insurance Fraud: Issues and Challenges," *Geneva Papers on Risk and Insurance: Issues and Practice*, 29, 313–333.

Viaene, S., Derrig , R. A., Baesens, B., and Dedene, G. (2002), "A Comparison of State-of-the-Art Classification Techniques for Expert Automobile Insurance Claim Fraud Detection," *Journal of Risk and Insurance*, 69, 373-421.

Viaene, S., Derrig, R. and Dedene, G. (2004a), "A Case Study of Applying Boosting Naive Bayes to Claim Fraud Diagnosis," *IEEE Transactions on Knowledge and Data Engineering*, 16, 612- 620.

Viaene, S., Derrig, R. and Dedene, G. (2004b), "Cost-sensitive Learning and Decision Making for Massachusetts PIP Claim Fraud Data," *International Journal of Intelligent Systems*, 19, 1197-1215.

Viaene, S., Gheel, D., Ayuso, M., and Guillen, M. (2004) "Cost-Sensitive Design of Claim Fraud Screens," Lecture Notes in Computer Design, Springer Berlin / Heidelberg.

Weisberg, H. I., and Derrig, R. A. (1991), "Fraud and Automobile Insurance: A Report on Bodily Injury Liability Claims in Massachusetts," *Journal of Insurance Regulation*, 9, 497-541.

Weisberg, H. I. and Derrig, R.A. (1992), "Massachusetts Automobile Bodily Injury Tort Reform," *Journal of Insurance Regulation*, 10, 384-440.

Weisberg, H. I. and Derrig, R.A. (1993), "Quantitative Methods for Detecting Fraudulent Automobile Bodily Injury Claims," AIB Cost Containment/Fraud Filing (DOI Docket R 95-12), Automobile Insurers Bureau of Massachusetts, July, 49-82.

Weisberg, H. I. and Derrig, R.A. (1995), "Identification and Investigation of Suspicious Claims," AIB Cost Containment/Fraud Filing (DOI Docket R 95-12), Automobile Insurers Bureau of Massachusetts, July, 192 – 245.

Weisberg, H. I. and Derrig, R.A. (1998), "Quantitative Methods for Detecting Fraudulent Automobile Bodily Injury Claims," *Risques*, 35, 75 -101.

Weisberg, H. I., Derrig, R. A. and Chen, X. (1994), "Behavioral Factors and Lotteries Under No-Fault With a Monetary Threshold: A Study of Massachusetts Automobile Claims," *Journal of Risk and Insurance*, 61, 245-275.

Williams, G. (1999), "Evolutionary Hot Spots Data Mining: An Architecture for Exploring for Interesting Discoveries," *Proceedings of Pacific Asia Conference on Knowledge Discovery and Data Mining.*

Williams, G. and Huang, Z. (1997), "Mining the Knowledge Mine: The Hot Spots Methodology for Mining Large Real World Databases," *Proceedings of the 10th Australian Joint Conference on Artificial Intelligence.*

Vita

Jing Ai was born in Wuhan, Hubei Province, P.R. China on May 13, 1981, the daughter of Fanglin Ai and Xiaoya Chen. After completing her study at the Experimental High School Attached to Beijing Normal University, Beijing in 1999, she entered Tsinghua University, majoring in Finance. She received the degree of Bachelor of Science in Economics in July 2003. She joined the graduate school of the University of Texas at Austin in August 2003. She received the degree of Master of Science in Information, Risk, and Operations Management in December 2006, and continued pursuing the degree of Doctor of Philosophy in Information, Risk, and Operations Management.

Permanent address: 1630 W. 6TH ST APT P, Austin, TX 78703.

This dissertation was typed by Jing Ai.