A state of the art survey of data mining-based fraud detection and credit scoring

Xun Zhou[*], *Sicong* Cheng, *Meng* Zhu, *Chengkun* Guo, *Sida* Zhou, *Peng* Xu, *Zhenghua* Xue, and *Weishi* Zhang

Technology Centre, HENGCHANG LITONG Investment Management(Beijing) Co. Ltd., Building 5, East District, Yard 10, XIBEIWANG East Road, Beijing, China

Abstract. Credit risk has been a widespread and deep penetrating problem for centuries, but not until various credit derivatives and products were developed and novel technologies began radically changing the human society, have fraud detection, credit scoring and other risk management systems become so important not only to some specific firms, but to industries and governments worldwide. Frauds and unpredictable defaults cost billions of dollars each year, thus, forcing financial institutions to continuously improve their systems for loss reduction. In the past twenty years, amounts of studies have proposed the use of data mining techniques to detect frauds, score credits and manage risks, but issues such as data selection, algorithm design, and hyperparameter optimization affect the perceived ability of the proposed solutions and it is difficult for auditors and researchers to explore and figure out the highest level of general development in this area. In this survey we focus on a state of the art survey of recently developed data mining techniques for fraud detection and credit scoring. Several outstanding experiments are recorded and highlighted, and the corresponding techniques, which are mostly based on supervised learning algorithms, unsupervised learning algorithms, semi-supervised algorithms, ensemble learning, transfer learning, or some hybrid ideas are explained and analysed. The goal of this paper is to provide a dense review of up-to-date techniques for fraud detection and credit scoring, a general analysis on the results achieved and upcoming challenges for further researches.

1 Introduction

Credit risk has been a widespread and deep penetrating problem for centuries, but not until various credit derivatives and products were developed and novel technologies began radically changing the human society, have fraud detection, credit scoring and other risk management systems become so important not only to some specific firms, but to industries and governments worldwide.

Financial fraud, including corporate frauds, money laundering, insurance frauds, credit card fraud, personal loan fraud, peer to peer lending fraud and others, is different from generally acceptable risky credit events such as loan default for the reason that fraud is a

[*] Corresponding author: zhouxun180108@credithc.com

deliberate act, a wrongful or criminal deception that is contrary to law, rule, or policy with intent to abuse a profit organization's system and to obtain unauthorized financial benefit without necessarily leading to direct legal consequences [1-3]. Although there is a universally accepted difference in law between fraud and risky credit events, in credit markets, the boundary between them becomes vague as more credit events are moved to online platforms and more fraudsters are skilled in counterfeiting. Therefore, when managing to reduce loss due to credit risks, a financial institution tends to mix financial detection and credit scoring and apply more characteristics in its decision-making process.

Recently, frauds and unpredictable defaults cost billions of dollars each year, thus, forcing financial institutions to continuously improve their systems for loss reduction and, consequentially, fraud detection and credit scoring became hot spots to explore and, in the past twenty years, a large amount of studies have proposed the use of novel data mining techniques for fraud detection, credit scoring and risk management.

Data mining is a process that uses a variety of data analysis tools to discover hidden patterns and relationships that may support a valid prediction. Phua et al. [2] point out that fraud detection has become one of the best-established applications of data mining in both industry and government. On the other hand, credit scoring is also heavily needed, especially in China's credit market, which has grown to 30 to 40 trillion dollars. The advent of the internet has led to the creation of new business models in China. Ant Financial, for example, began its long term-planned Sesame Credit, the user data-based online credit scoring service in 2015. Sesame Credit generates credit scores based on data from users, partners, public agencies, financial institutions, and various types of merchants. Financial products or services like Sesame Credit have made credit more widely available to customers and have enlarged the China's credit market dramatically by creating various credit related services and serving customers who have limited traditional credit history and who need loans, but bank loans or credit cards are not handy or available. At the same time, China's P2P lending market is explosively growing as more and more Chinese people are connected by internet and accustomed to loans. The explosive growth of China's credit market provides opportunities for related organizations to make profit, for customers to get bland new services and products, for fraudsters to hunt for unauthorized benefits, and thus for researchers to design intelligent fraud detection and credit scoring systems. Nevertheless, issues such as data selection, algorithm design, and hyperparameter optimization affect the perceived ability of the proposed solutions and it is difficult for auditors and researchers to explore and figure out the highest level of general development in this area.

In this survey we will focus on a state of the art survey of recently developed data mining techniques for fraud detection and credit scoring. The types of fraud will be studied here are automobile insurance fraud, financial statement fraud, credit card fraud and peer to peer lending fraud. Credit scoring will be discussed for distinguishing bad credit and good credit. Automobile insurance fraud refers to submitting fake documents regarding causalities in a staged accident or claims for past losses to obtain financial profit [4]. Financial statement fraud can be defined as material omissions or misrepresentations resulting from an intentional failure to report financial information in accordance with accounting standards [5]. Credit card fraud can be divided into two types: behaviour fraud and application fraud. Behaviour fraud refers to theft and fraud committed by using a stolen physical card or card information via internet, phone, shopping, web, or in absence of card holder [6]. Application fraud is a type of identity theft or identity counterfeits that involves opening an account using stolen or fake information.

Several outstanding experiments will be highlighted and the corresponding techniques, which are mostly based on supervised learning algorithms, unsupervised learning algorithms, semi-supervised algorithms, ensemble learning, transfer learning, or some hybrid ideas will be explained and analysed. Going through a number of important researches published within the last few years, this paper aims to provide a dense review of up-to-date techniques for fraud detection and credit scoring, a general analysis on the results achieved and upcoming challenges for further researches.

The rest of the paper is organized as follows. Section 2 contains the related reviews and survey papers. Section 3 presents the classification of data mining techniques and applications. Section 4 detailly explains four highlighted advanced data mining methods for automobile insurance fraud detection, financial statement fraud detection, credit card fraud detection and credit scoring, especially for P2P lending, respectively. Finally, Section 5 analyses the present results, discusses upcoming challenges and concludes the paper.

2 Related works

Over the past few years, a number of review articles have appeared in conference or journal publications. Bolton and Hand [7], for example, reviewed statistical methods for fraud detection, including credit card fraud, money laundering, telecommunications fraud, etc. Zhang and Zhou [8] surveyed financial applications of data mining including stock market and bankruptcy predictions and fraud detection. Phua et al. [2] presented a survey of data mining-based fraud detection research, including credit transaction fraud, telecommunications subscription fraud, automobile insurance fraud and the like. Li et al. [9], Travaille et al. [10] and Liu and Vasarhelyi [11] surveyed and analysed fraud detection statistical methods for health care fraud detection. Richhariya [12], Ngai et al. [13] and Wang [14] provided a comprehensive survey and review for different data mining techniques used to detect financial fraud. Sithic and Balasubramanian [15] presented an extensive survey for fraud types in medical and motor insurance systems and many types of data mining techniques are used to detect fraud in these insurance sectors. Our survey presents herein is an up-to-date, comprehensive and state of the art review of data mining applications in financial fraud detection.

3 Classification of data mining techniques and applications

Figure 1 gives a description of classifier training based on supervised, semi-supervised, and unsupervised learning.

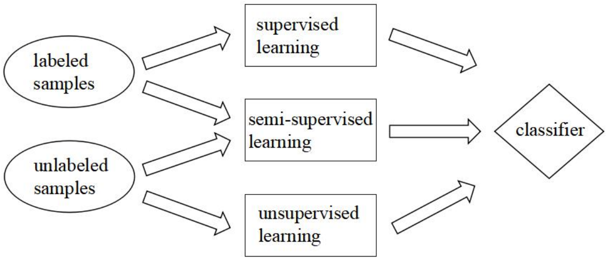

Fig. 1. The relationship between learning style and data set.

3.1 Naive bayes

Naive Bayes method assumes that the presence or absence of any attribute of a class variable is not related to the presence or absence of any other attributes. This technique is named “naive” because it naively assumes independence of the attributes. The classification is done by applying “Bayes” rule to calculate the probability of the correct class.

Viaene et al. [16] applies the weight of evidence formulation of AdaBoost Naive Bayes (boosted fully independent Bayesian network) scoring. This allows the computing of the relative importance (weight) for individual components of suspicion and displaying the aggregation of evidence pro and contra fraud as a balance of evidence which is governed by a simple additivity principle. Panigrahi et al. [17] combined a Dempster-Schaefer adder with a Bayesian learner to solve credit card fraud with their own synthesised data. Hooi et al. [18] developed BIRD, a Bayesian inference approach for ratings fraud detection. The method provides a principled way to combine rating and temporal information to detect rating fraud, and to find a trade-off between users with extreme rating distributions vs. users with larger number of ratings.

Algorithms based on Naive Bayes are easy to implement in engineering, and easy to satisfy the need of fraud detection. However, because Naive Bayes classifier is a log-linear model, it does not provide optimal solutions for non-linear problems with high complexity. To maintain model's interpretability and to improve the ability of solving non-linear problems, we can apply the approaches mentioned above, e.g. AdaBoost, to integrate multiple weak learning machines, and we can also apply decision trees.

3.2 Decision tree

Decision trees have structure of a tree which tries to separate the given records into mutually exclusive subgroups. Decision Tree algorithm recursively partitions a dataset using breadth-first approach or depth-first greedy approach until all the items of data go in a particular class.

Sahin et al. [19] studied the ability of decision trees to identify fraudulent credit card transactions, using a 6-month sample from a major bank. Anis et al. [20] applied random under sampling with feature selection for six decision trees classifier. Results showed that random forest is best classifiers among the other that they have used in this study. Jain et al. [21] presents hybrid approach for credit card fraud detection using rough set and decision tree technique which can be used in credit card fraud detection mechanisms. Save et al. [22] proposed a system which detect fraud in credit card transaction processing using a decision tree with combination of Luhn's algorithm and Hunt's algorithm. Luhn's algorithm is used to validate the card number. More complicated tree structural models, like gradient boosted decision tree (GBDT) and extreme gradient boosted (XGBoost) decision tree models could be used to build fraud detection systems. [23] compared performances of logistic regression, GBDT and deep learning models on credit card fraud detection, and [24] compared performances of random forest (RF) and XGBoost on detecting frauds that conducted through P2P lending platform.

Tree models are interpretable and easy to implement. The more one feature is called, or the more information gain accumulated by the feature, the more important the feature is. This makes tree models powerful not only in classification (or regression) model, but also in feature selecting. However, CART and GBDT share the problem of being over dependent on some features. If creditor forges on these features, or abnormal condition occurs during accessing data of these features, then the engaged model tends to make wrong decisions. As a comparison, random forest based on bagging principle is easy to avoid this problem. In practice, when some features are not reliable enough, then you may consider using RF.

3.3 Logistic regression

Logistic regression is a classification method, which is mainly used for two classification problems. Its main idea is to use the existing data to establish regression equations classification boundaries, so as to classify them.

In 2007, Pinquet et al. [25] and Viaene et al. [26] both studied logistic regression with insurance fraud, concentrating on a database of Spanish automobile insurance claims. Bhattacharyya et al. [27] performed several credit card fraud experiments, comparing two common classification solutions against the well-known logistic regression and observing results across various common metrics. Kibekbaev and Duman [28] proposed a novel profit-based logistic regression which makes the classification considering all individual costs and profits of instances and consequently maximizes the total net profit captured from applying the classification model. Kulkarni and Ade [29] have suggested a framework using logistic regression to tackle the problem of unbalanced data in credit card fraud detection. They have used an incremental learning approach for fraud modelling and detection

Logistic regression is accessible and easy to be parallelly implemented. Considering its interpretability and potential in generalization, logistic regression has been widely used in processing fraud detection and credit scoring. Unlike decision trees, logistic regression is a linear model, hence unable to solve complex non-linear problems. Therefore, single application of LR model onto real operation problem, is usually accompanied by large feature engineering, particularly feature combination. Otherwise, we can use LR model to merge two or more non-linear model to benefit from distinctive advantages of different models.

3.4 Support vector machine

The SVM method finds a special kind of linear model, the maximum margin hyper plane, and it classifies all training instances correctly by separating them into correct classes through a hyperplane.

Patel and Gond [30] proposed the SVM based method with multiple kernel involvement which also includes several fields of user profile instead of only spending profile. Whitrow et al. [31] compared SVM with decision trees in solving credit card fraud, with a focus on aggregating common transactional variables to create new inputs. Maldonado et al. [32] introduced a family of methods based on a backward elimination approach for feature ranking and embedded classification using SVM, which has been adapted to select those attributes that are relevant to discriminate between classes under imbalanced data conditions. Moepya et al. [33] demonstrated weighted Support Vector Machines are superior to the cost-sensitive Naive Bayes and K-Nearest Neighbours classifiers. Mareeswari and Gunasekaran [34] proposed hybrid support vector machine (HSVM) along with communal and spike detection for credit card application fraud detection to overcome the limitation of existing systems.

With kernel techniques, SVM can solve complex non-linear optimization problems, particularly, data structure problems, which makes SVM very suitable for solving complex, changeable, data structuring problems for fraud detection. Notably, the problem how to choose kernel functions has been proposed with a standard solution, and hence, SVM requires continuous trial during operation.

3.5 Artificial neural network

Neural network is a set of connected input/output units and each connection has a weight present with it. During the learning phase, network learns by adjusting weights to guess the correct class labels.

Kolalikhormuji et al. [35] proposed a cascade neural network system with imperialist competitive algorithm for increasing transaction recognition system's and accuracy rate at the same time. Fu et al. [36] proposed a CNN-based framework of mining latent fraud patterns in credit card transactions. Results show its superior performance compared with some state-of-the-art methods. Gulati et al. [37] presented a credit card fraud detection system which works on neural networks seem to detect up to 80% accuracy with sample transaction data. Modi and Dayma [38] indicated that CNN with SMOTE and feature transformation overcome issue of precision and outperforms NN in all terms.

Neural network is accessible for parallel implementation, which will satisfy the demand of large scaled online applications. However, it has high tendency of overfitting if training set is not a good representation of the problem domain, thus requires a large workload of regularization and constant retraining to adapt to novel fraudulent behaviours. Moreover, in recent years, deep learning techniques, such as convolutional neural network (CNN) and recurrent neural network (RNN), are rapidly developed in fields of computer vision, natural language processing, etc. New methods of fraud detection that are based on these techniques, for example, loan customer website analysis, mobile log analysis, micro-expression recognition, will surely be applied more frequently.

3.6 K-means

Unsupervised methods do not need the prior knowledge of fraudulent and non-fraudulent transactions in historical datasets, but instead, detect changes in behaviour or abnormal transactions. One advantage of using unsupervised methods over supervised methods is that previously undetected types of fraud may be detected.

K-means clustering algorithm groups the data based on the similarity of their attribute values. The groups formed by mean clustering algorithm is referred to as cluster. The grouping is formed based on the square of distance and centroid of their data values.

Celebi et al. [39] outlined the K-means initialization methods, focusing on their computational efficiency. Eight commonly used linear time initialization methods are compared on a large and diverse collection of actual and synthetic data sets using various performance criteria. Finally, the experimental results using non-parametric statistical tests are compared. Huang and Su [40] presented a problem based on user behaviour pattern analysis which has the insensitivity of numerical value, strong noise, and uneven spatial and temporal distribution characteristics. The existing clustering methods, trajectory analysis methods, and behaviour pattern analysis methods are analysed, and clustering algorithm is combined into the trajectory analysis. The results show that the improved algorithm has more advantages than the traditional k-means algorithm. Subudhi and Panigrahi [4] used optimized Fuzzy C-Means clustering and supervised classifiers for automobile insurance fraud detection. They combined different supervised classifiers with FCM or GAFCM clustering. The efficacy of the proposed methods is examined by several parallel experiments on a real-world automobile insurance dataset.

The advantage of K-means is that it is simple to implement and understand. However, the drawback is that difficulty classifying noisy data, which many fraud types contain.

3.7 Graph based semi-supervised learning

The method of training the classification model using both unlabelled data and labelled data is called semi-supervised learning in the field of machine learning.

Graph based semi-supervised learning is a semi-supervised learning method studied in recent years, which is based on manifold hypothesis. First, we use graph strategy to create graph models that can reflect all the data relations, and then transfer labels on graphs to get a classification function that satisfies global consistency assumption. Global consistency assumption, namely manifold hypothesis, makes the classification decision surface try to pass the place where labelled and unlabelled data are both sparse, so it is possible to overcome the drawback that the generalization ability limited by the relative lack of the labelled data in the methods such as support vector machine, neural network and so on.

Ramaki [41] proposed a model for fraud detection in credit cards on a semantic connection between data stored for every transaction fulfilled by a user basis and present it by ontology graph and store them then in patterns database. Lebichot et al. [42] proposed several improvements based on an existing Fraud Detection Systems APATE. APATE uses a collective inference algorithm to spread fraudulent influence through a network by using a limited set of confirmed fraudulent transactions. Cao et al. [43] proposed HitFraud, a collective fraud detection algorithm that captures the inter-transaction dependency. Experiments on EA payment transaction data demonstrate that the prediction performance is effectively boosted by HitFraud with different choices of base classifiers.

The advantage of graph based semi-supervised learning is that it is simple to implement and very easy for auditors to understand given visual nature of results. However, the drawback is that it requires high computational power for training and operation, making it unsuitable for real-time function.

4 Selected fraud detection systems

4.1 Automobile insurance fraud detection

Applied on automobile insurance fraud, the traditional statistical machine learning techniques do work, but share limits that must be considered carefully. Firstly, the traditional statistical machine learning techniques depend on manually designed features which are assumed to be exactly specified, complete to describe the problem, but since the standardized information provided in the claims could be counterfeited by the skilled deceivers who might also understand the techniques and those fraudsters keep discovering innovative types of frauds continually, the anti-fraud system must keep up-to-date. Secondly, though the traditional techniques, such as random forests and SVM, can perform linear and nonlinear transformation at a shallow level, they are not good at digging more deeply into the unstructured textual and visual data to find hidden information, while the extra attributes obtained by deep learning could improve the classification, especially considering the extent of mislabelled samples.

As more complex data and hidden information are available, neural network classifiers show their potential to overcome the difficulties of traditional statistical machine learning techniques when they are properly designed and combined with other methods prior to the training process. Considering that the data set is usually extremely skewed and the minority class, though being the very important part which we need to classify, is easily to be ignored, pre-training process would benefit the supervised training process by balancing the data set. Moreover, the validity of textual information shows the possibility to grasp more hidden information.

In Wang and Xu’s research [44], a detection model based on Latent Dirichlet Allocation and deep neural networks is built to handle structural data, consists of numeric data and categorical data, and textual data. In natural language processing, the topic model Latent Dirichlet Allocation is an important generative model based on Gibbs sampling of the Dirichlet-multinomial distribution that can be used to identify hidden topic information in large-scale document collections [45]. In Wang and Xu’s research, the words arising from the text description by insurance experts were viewed as documents for LDA to extract topics regarding automobile insurance behaviours.

The model was tested on real-world data derived from an automobile insurance company in China consisting of 415 fraudulent claims and 36667 non-fraudulent claims. Since the dataset is imbalanced, they applied SMOTE to oversample the fraudulent claims and randomly under-sampled the legitimate claims to get a balanced dataset. Numeric data and one-hot coded categorical data were generated as features directly, while text data was implemented by applying Chinese word segmentation and LDA. LDA and deep neural networks are complementary in the sense that LDA can help explore latent topics while neural networks can dig deeply into the topics for more effective information. Table 1 shows that the performance of LDA and DNN outperforms random forests and SVM applied on the same dataset.

Table 1. The performances of different methods for automobile insurance fraud detection.

Research	Classifier	Dataset	TPR	FPR	Accuracy
Y. Wang and W. Xu [44]	SVM	415/36667	0.853	0.261	0.796
Y. Wang and W. Xu [44]	RF	415/36667	0.802	0.209	0.797
Y. Wang and W. Xu [44]	LDA+DNN	415/36667	0.910	0.082	0.914
M. Vasu et al.	SVM+K-NN+K-means	923/14497	0.792	0.402	0.609
M. Vasu el al.	MLP+K-NN+K-means	923/14497	0.838	0.391	0.623
G. Sundarkumar et al. [46]	SVM+OCSVM	923/14497	0.919	0.416	0.604
G. Sundarkumar et al.	MLP+OCSVM	923/14497	0.646	0.281	0.723
S. Subudhi et al. [4]	SVM+GAFCM	923/14497	0.832	0.115	0.870
S. Subudhi et al.	MLP+GAFCM	923/14497	0.811	0.174	0.824

Compared with the results obtained by M Vasu et al., G Sundarkumar et al. [46], S Subudhi et al. [4], Y Wang and W Xu's result shows a larger improvement from SVM to LDA+DNN, implying that delicately tuning hyperparameters and employing LDA could benefit DNN in financial statement fraud detection.

4.2 Financial statement fraud detection

What differentiate financial statement fraud from other types of fraud are the facts that the fraudsters are usually a group of experts who have in-depth knowledge and are clear attributed of responsibility and that financial statement fraud usually brought a negative impact on capital markets, a loss of shareholder value [47] and may be an effective indicator of substantial financial problems that cause bankruptcy [48]. Amounts of researches has been made on financial statement fraud detection using data mining techniques. Kirkos et al. [49] and Lin et al. [50] applied neural networks, Kotsiantis et al. [51] applied decision trees, and Pai, Huang et al. [52] [53] used support vector machines. Most of previous studies paid more attention on numeric data than on textual data, but since more established natural language processing techniques are acquired, more hidden patterns and information could be uncovered to support fraud detection.

In Hajek and Hentiques' research [54], they examined whether an improved financial fraud detection system could be developed by combining specific features derived from financial information and managerial comments in corporate annual reports.

They identified 311 public U.S. companies involved in alleged instances of fraudulent financial reporting and collected a set of 311 fraudulent annual reports, and also identified 311 firms with the corresponding market capitalization and industry membership and collected a set of U.S. 311 legitimate annual reports. Regarding linguistic variables, they extracted linguistic variables from the management discussion and analysis, representing the most important textual section from the downloaded 10-Ks. By detecting sentiment words, including positive, negative, uncertain, litigious, modal strong, model weak, and constraining words, and calculating the overall tone given by the ratio of the difference of the frequencies of positive and negative words and the total of the frequencies. During the training process, a wide range of data-mining based algorithms are tested, and the results are shown in table 2. We find that BBN significantly outperforms the remaining methods in terms of most classification metrics.

Table 2. The performances of different methods for financial statement fraud detection.

Classifier	Dataset	TPR	FPR	Precision	Accuracy
LR	311/311	0.730	0.229	0.761	0.745
BBN	311/311	0.852	0.046	0.949	0.903
SVM	311/311	0.767	0.206	0.788	0.780
RF	311/311	0.869	0.119	0.880	0.875
MLP	311/311	0.766	0.205	0.789	0.779

However, an ignored problem in the process of training a financial statement fraud detection model is that a report from a specific industry does not include similar information as one from a different industry does. What's more, a report at this stage may not be securely put into a dataset containing previous reports. Considering these problem, we believe a deep understanding of the detected industries and macro-economy is required and a transfer learning model may have a better performance.

4.3 Credit card fraud detection

Using real-life dataset of transactions from an international credit card operation, Bhattacharyya et al. [27] evaluates the performance of support vector machines and random forests, together with the well-known logistic regression models for credit fraud detection. This dataset contains 13-month worth of 50 million credit card transactions on about one million credit cards from a single country, from January 2006 to January 2007. Since fraud transactions in the dataset are very few compared to legitimate transactions, some form of sampling is required to obtain a training dataset containing a sufficient proportion of fraud to non-fraud cases. They use data undersampling, which is a simple method that has been noted to perform well.

Fahmi et al. [55] and West and Bhattacharya [56] both used the same dataset, which is the one used in "UCSD-FICO Data Mining Contest 2009". The competition was organized by University of California, San Diego (UCSD) and FICO a major firm of analytics and decision support in 2009. This dataset is highly imbalanced with a ratio of approximately 97:3 towards legitimate transactions, meaning that 3% of the transactions are fraud while the other 97% are legit. The experimental results are shown in table 3. It could be observed in Fahmi et al.'s experiment that the K-NN based model outperformed other models in all terms. West and Bhattacharya's experimental results show that SVM has the best performance with the highest accuracy and a zero false positive rate.

Table 3. The performances of different methods for credit card fraud detection.

Research	Classifier	Dataset	TPR	FPR	Accuracy
Bhattacharyya et al.	LR	5/45 million	0.654	0.021	0.947
Bhattacharyya et al.	SVM	5/45 million	0.524	0.016	0.938
Bhattacharyya et al.	RF	5/45 million	0.727	0.013	0.962
Fahmi et al.	K-NN	3000/97000	0.738	0.262	0.738
Fahmi et al.	Naive Bayes	3000/97000	0.708	0.292	0.708
Fahmi et al.	SVM	3000/97000	0.692	0.308	0.692
West and Bhattacharya	GA2	3000/97000	0.016	0.000	0.911
West and Bhattacharya	SVM	3000/97000	0.064	0.000	0.915
West and Bhattacharya	GP2	3000/97000	0.025	0.002	0.910

4.4 P2P lending fraud detection and credit scoring

In recent years, P2P lending is explosively growing in China. However, in developing countries like China, where the credit market is not yet well regulated and developing dramatically, most researches are done by scholars or engineers within a relatively narrow environment with only limited, closed data available. [57] used a LR model to test over Lending Club's dataset. Based on Bondora's open-sourced dataset, [58] compared performances of LR classification with ANN classification. [58]'s trial suggests that ANN tends to demonstrate the best performance in classification. [59] applied CART, ANNs and SVM in their experiments, and the results do not differ much. [60] used a P2P company's dataset to set up fraud detection with LR model, while [61] used BP neural network to rate credit of a Chinese P2P lending company's customers. [62] implemented a more complex LSTM model to set up an anti-fraud system on one P2P lending platform based in Jinan, Shandong, China. Moreover, some scholars focus their researches on real problems with Chinese characteristics. For instance, [60] targets on discussion about lending demands fraud of P2P lending services, while [63] targets on construction of laws and regulations of P2P lending industry.

Assuming the homogeneity between training data and test data in P2P lending fraud detection, Xia, Liu, Li and Liu [64] proposed a sequential ensemble credit scoring model based on XGBoost, a variant of gradient boosting machine, and the hyperparameters of XGBoost are adaptively tuned by the tree-structured Parzen estimator, grid search, random search and manual search. A sequential ensemble learning combines a series of weak base learners that process different hypothesizes sequentially to form a better hypothesis, thus making good predictions [65] [66] [67]. To verify the performances of their proposal, five real world credit datasets are utilized, including two datasets from two P2P lending companies. The first dataset from Lending Club [68] contains 1322 good samples and 1320 bad samples, and the second dataset from WE [69] contains 1072 good samples and 349 bad samples. Table 4 and 5 demonstrate that the proposed model outperforms baseline models on average.

However, P2P lending products are transferring huge risks and a relatively larger probability of fraud but, in their experiment, only 10 to 20 features were constructed and thus whether the risks could be detected and controlled is uncertain. When big data is available, some other problems arise: whether the logic behind the data is self-contained and whether characteristics of the same name is independent of time, space, etc. We believe that further studies on transfer learning techniques combined with fundamental analysis could help solve these problems.

Table 4. The performances of different methods for P2P credit scoring 1.

Classifier	Dataset	Type I error	Type II error	Accuracy
XGBoost-MS	1320/1322	0.290	0.376	0.667
XGBoost-GS	1320/1322	0.318	0.356	0.663
XGBoost-RS	1320/1322	0.298	0.361	0.671
XGBoost-TPE	1320/1322	0.298	0.362	0.671
LR	1320/1322	0.414	0.291	0.647
SVM	1320/1322	0.413	0.374	0.607
RF	1320/1322	0.357	0.379	0.632

Table 5. The performances of different methods for P2P credit scoring 2.

Classifier	Dataset	Type I error	Type II error	Accuracy
XGBoost-MS	349/1072	0.521	0.064	0.824
XGBoost-GS	349/1072	0.412	0.080	0.838
XGBoost-RS	349/1072	0.406	0.074	0.845
XGBoost-TPE	349/1072	0.397	0.074	0.847
LR	349/1072	0.923	0.030	0.751
SVM	349/1072	0.972	0.013	0.752
RF	349/1072	0.605	0.060	0.806

5 Conclusion

This paper reviewed the literature describing use of the fraud detection and credit scoring approaches based on supervised, unsupervised, semi-supervised, ensemble and transfer techniques. It is noticed that most fraud detection systems employ at least one supervised learning method. Supervised learning methods we presented here are Naive Bayes, decision tree, logistic regression, support vector machine and neural network, unsupervised learning methods we presented here is k-means and semi-supervised learning methods we presented here is graph-based semi-supervised learning.

These techniques can be used alone or combined with ensemble or meta-learning techniques to build stronger classifiers. Without loss of generality, those approaches are relatively successful in fraud detection and credit scoring, reducing cost, and protecting our economic society. However, there are still challenges in this area. Firstly, data mining-based fraud detection and credit scoring are subject to the same issues as other classification problems, such as feature engineering, parameter selection, and hyperparameter tuning. Secondly, public data is not abundant enough for researchers to train and test their models and it is nearly impossible to represent the complex financial scenarios, particularly those in China. Thirdly, adapting to the development of credit markets, the far-flung risks, and the changes of fraudulent behaviours, fraud detection and credit scoring methods need evolving all the time. Fourthly, since fraud detection and credit scoring are primarily classification problems with vast differences in misclassification costs, cost measurements should be studied in detail with respect to each industry.

References

1. Oxford Concise English Dictionary, Tenth ed, Publisher, 1999.
2. C. Phua, V. Lee, K. Smith and R. Gayler, "A comprehensive survey of data mining-based fraud detection research," *Artificial Intelligence Review,* p. 1–14, 2005.
3. Wang, J; Liao, Y; Tsai, T; Hung, G, "Technology-based financial frauds in Taiwan: issue and approaches," *IEEE Conference on: Systems, Man and Cyberspace,* pp. 1120-1124, 10 2006.
4. Subudhi, S; Panigrahi, S, "Use of optimized Fuzzy C-Means clustering and supervised classifiers for automobile insurance fraud detection," *Journal of King Saud University – Computer and Information Sciences,* 2017.
5. Z. Rezaee, Financial Statement Fraud – Prevention and Detection, John Wiley & Sons, Inc., 2002.
6. K. Chaudhary, J. Yadav and B. Mallick, "A review of fraud detection techniques: credit card," *International Journal of Computer Applications,* 5 2012.
7. Bolton, R J; Hand, D J, "Statistical fraud detection: a review," *Statistical Science,* vol. **17**, no. 3, p. 235–255, 2002.
8. D. Zhang and L. Zhou, "Discovering golden nuggets: data mining in financial application,"*IEEE Transactions on Systems, Man and Cybernetics,* vol. **34**, no. 4, 11 2004.
9. J. Li, K. Huang, J. Jin and J. Shi, "A survey on statistical methods for health care fraud detection," *Health Care Manag. Sci.,* p. 275–287, 2008.
10. P. Travaille, D. Thornton, R. M. Mueller and J. V. Hillegersberg, "Electronic Fraud Detection in the U.S. Medicaid Healthcare Program: Lessons Learned from other Industries," in *Americas Conference on Information Systems*, Detroit, 2011.

11. Q. Liu and V. Miklos, "Healthcare fraud detection: a survey and a clustering model incorporating Geo-location information," in *World Continuous Auditing and Reporting Symposium*, Brisbane, 2013.
12. P. Richhariya and P. K. Singh, "A Survey on Financial Fraud Detection Methodologies," *International Journal of Computer Applications,* vol. **45**, no. 22, pp. 21-24, 2012.
13. Ngai, E W T; Hu, Y; Chen, Y; Sun, X, "The application of data mining techniques in financial fraud detection: a classification framework and an academic review of literature," *Decision Support Systems,* vol. **50**, no. 3, p. 559–569, 2011.
14. S. Wang, "A comprehensive survey of data mining-based accounting-fraud detection research," in *International Conference on Intelligent Computation Technology and Automation*, 2010.
15. L. Sithic and T. Balasubramanian, "Survey of insurance fraud detection using data mining techniques," *International Journal of Emerging Computing Engineering,* vol. **2**, no. 3, pp. 62-65, 2013.
16. S. Viaene, R. Derrig and G. Dedene, "A Case Study of Applying Boosting Naive Bayes to Claim Fraud Diagnosis," *IEEE Transactions on Knowledge and Data Engineering,* vol. **16**, no. 5, pp. 612-620, 2004.
17. S. Panigrahi, A. Kundu, S. Sural and A. K. Majumdar, "Credit card fraud detection: a fusion approach using Dempster–Shafer theory and Bayesian learning," *Information Fusion,* no. 10, pp. 354-363, 2009.
18. B. Hooi, N. Shah, A. Beutel, S. Gunneman, L. Akoglu, M. Kumar, D. Makhija and C. Faloutsos, "BIRDNEST: Bayesian Inference for Ratings-Fraud Detection," in *Siam International Conference on Data Mining*, 2016.
19. Y. Sahin, S. Bulkan and E. Duman, "A cost-sensitive decision tree approach for fraud detection," *Expert Systems with Applications,* no. 40, pp. 5916-5923, 2013.
20. M. Anis, M. Ali and A. Yadav, "A comparative study of decision tree algorithms for class imbalanced learning in credit card fraud detection," *International Journal of Economics, Commerce and Management,* vol. **3**, no. 12, pp. 86-102, 2015.
21. R. Jain, B. Gour and S. Dubey, "A Hybrid Approach for Credit Card Fraud Detection using Rough Set and Decision Tree Technique," *International Journal of Computer Applications,* vol. **139**, no. 10, pp. 1-6, 2016.
22. P. Save, P. Tiwarekar, N. Ketan and N. Mahyavanshi, "A Novel Idea for Credit Card Fraud Detection using Decision Tree," *International Journal of Computer Applications,* vol. **161**, no. 13, pp. 6-9, 2017.
23. G. Rushin, C. Stancil, M. Sun, S. Adams and P. Beling, "Horse race analysis in credit card fraud—deep learning, logistic regression, and Gradient Boosted Tree," in *Systems and Information Engineering Design Symposium*, Charlottesville, 2017.
24. X. Yu, "Machine Learning Application in Online Leading Credit Risk Prediction," 16 7 2017. [Online]. Available: https://arxiv.org/abs/1707.04831. [Accessed 23 2 2018].
25. Pinquet, J; Ayuso, M; Guillen, M;, "Selection bias and auditing policies for insurance claims," *Journal of Risk and Insurance,* no. 74, pp. 425-440, 2007.
26. S. Viaene, M. Ayuso, M. Guillen, D. Van Gheel and G. Dedene, "Strategies for detecting fraudulent claims in the automobile insurance industry," *European Journal of Operational Research,* no. 176, pp. 565-583, 2007.

27. S. Bhattacharyya, S. Jha, K. Tharakunnel and J. C. Westland, "Data mining for credit card fraud: a comparative study," *Decision Support Systems,* no. 50, pp. 602-6613, 2011.
28. A. Kibekbaev and E. Duman, "Profit-based Logistic Regression: A case study in Credit Card Fraud Detection," *The Fourth International Conference on Data Analytics,* pp. 101-105, 2015.
29. P. Kulkarni and R. Ade, "Logistic Regression Learning Model for Handling Concept Drift with Unbalanced Data in Credit Card Fraud Detection System," in *International Conference on Computer and Communication Technologies*, 2016.
30. S. Patel and S. Gond, "Supervised Machine (SVM) Learning for Credit Card Fraud Detection," *International Journal of Engineering Trends & Technology,* vol. **8**, no. 3, pp. 137-139, 2014.
31. C. Whitrow, D. J. Hand, P. Juszczak, D. Weston and N. M. Adams, "Transaction aggregation as a strategy for credit card fraud detection," *Data Mining & Knowledge Discovery,* vol. **18**, no. 1, pp. 30-55, 2009.
32. S. Maldonado, R. Weber and F. Famili, "Feature selection for high-dimensional class-imbalanced data sets using Support Vector Machines," *Information Sciences,* no. 286, 2014.
33. S. O. Moepya, S. S. Akhoury and F. V. Nelwamondo, "Applying Cost-Sensitive Classification for Financial Fraud Detection under High Class-Imbalance," in *2015*, IEEE International Conference on Data Mining Workshop.
34. V. Mareeswari and G. Gunasekaran, "Prevention of credit card fraud detection based on HSVM," in *IEEE International Conference on Information Communication and Embedded Systems*, 2016.
35. M. Kolalikhormuji, M. Bazrafkan, M. Sharifian, S. Javad Mirabedini and A. Harounabadi, "Credit Card Fraud Detection with a Cascade Artificial Neural Network and Imperialist Competitive Algorithm," *International Journal of Computer Applications,* vol. **96**, no. 25, pp. 1-9, 2014.
36. K. Fu, D. Cheng, Y. Tu and L. Zhang, "Credit Card Fraud Detection Using Convolutional Neural Networks," in *International Conference on Neural Information*, 2016.
37. A. Gulati, D. Prakash, MdFuzailC, J. Norman and M. R, "Credit card fraud detection using neural network and geolocation," *IOP Conferece Series: Materials Science and Engineering,* no. 263, 2017.
38. K. Modi and R. Dayma, "Fraud Detection Technique in Credit Card Transactions using Convolutional Neural Network," *International Journal of Advance Research in Engineering, Science & Technology,* vol. **8**, no. 4, pp. 1-7, 2017.
39. M. Celebi, H. Kingravi and P. Vela, "A comparative study of efficient initialization methods for the K-Means clustering algorithm," *Expert Systems with Applications,* vol. **40**, no. 1, pp. 200-210, 2013.
40. X. Huang and W. Su, "An improved K-means clustering algorithm," *Journal of networks,* vol. **9**, no. 1, 2014.
41. A. A. Ramaki, "Credit Card Fraud Detection Based on Ontology Graph," *International Journal of Security, Privacy and Trust Management,* vol. **5**, no. 1, pp. 1-12, 2012.
42. B. Lebichot, F. Braun, O. Caelen and e. al., "A graph-based, semi-supervised, credit card fraud detection system," in *Complex Networks & Their Applications V. Springer International Publishing*, 2016.

43. B. Cao, M. Mao, S. Viidu and Y. S. Philip, "Collective Fraud Detection Capturing Inter- Transaction Dependency," in *Proceedings of Machine Learning Research*, 2017.
44. Y. Wang and W. Xu, "Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud," *Decision Support Systems,* no. 105, pp. 87-95, 2018.
45. D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research,* no. 3, pp. 993-1922, 2003.
46. G. G. Sundarkumar and V. Ravi, "A novel hybrid undersampling method for mining unbalanced datasets in banking and insurance," *Engineering Applications of Artificial Intelligence,* no. 37, pp. 368-377, 2014.
47. M. S. Beasley, J. V. Carcello, D. R. Hermanson and T. I. Neal, "Fraudulent Financial Reporting," in *Committee of Sponsoring Organizations of the Treadway Commission*, Jersey City, 2010.
48. M. Beneish, "The detection of earnings manipulation," *Financial Analysis Journal,* no. 5, pp. 24-36, 1999.
49. E. Kirkos, C. Spathis and Y. Manolopoulos, "Data mining techniques for the detection of fraudulent financial statements," *Expert System Applications,* no. 32, pp. 995-1003, 2007.
50. C. C. Lin, S. Y. Chiu, S. Y. Huang and D. C. Yen, "Detecting the financial statement fraud: the analysis of the differences between data mining techniques and experts' judgements," *Knowledge Based Systems,* no. 89, pp. 459-470, 2015.
51. S. Kotsiantis, E. Koumanakos and D. Tzelepis, "Forecasting fraudulent financial statements using data mining," *International Journal of Computional Intelligence,* no. 3, pp. 104-110, 2006.
52. P. F. Pai, M. F. Hsu and M. C. Wang, "A support vector machine-based model for detecting top management fraud," *Knowledge Based Systems,* no. 24, pp. 314-321, 2011.
53. S. Huang, "Fraud detection model by using support vector machine techniques," *International Journal of Digital Content Technology and Its Applications,* no. 7, pp. 32-42, 2013.
54. P. Hajek and R. Henriques, "Mining corporate annual reprots for intelligent detection of financial statement fraud - A comparative study of machine learning methods," *Knowledge Based Systems,* no. 128, pp. 139-152, 2017.
55. M. Fahmi, A. Hamdy and K. Nagati, "Data Mining Techniques for Credit Card Fraud Detection: Empirical Study," *Sustainable Vital Technologies in Engineering & Informatics,* pp. 1-9, 2016.
56. J. West and M. Bhattacharya, "Some Experimental Issues in Financial Fraud Mining," *Procedia Computer Science,* no. 80, pp. 1734-1744, 2016.
57. R. Emekter, Y. Tu, B. Jirasakuldech and M. Lu, "Evaluating credit risk and loan performance in online Peer-to-Peer (P2P) lending," *Applied Economics,* 2015.
58. A. Byanjankar, M. Heikkila and J. Mezei, "Predicting Credit Risk in Peer-to-Peer Lending: A Neural Network Approach," in *IEEE Symposium Series on Computational Intelligence*, 2015.
59. Y. Jin and Y. Zhu, "A Data-Driven Approach to Predict Default Risk of Loan for Online Peer-to-Peer (P2P) Lending," in *International Conference on Communication Systems and Network Technologies*, Gwalior, 2015.

60. J. J. Xu, Y. Lu and M. Chau, "P2P Lending Fraud Detection: A Big Data Approach," *Intelligence and Security Informatics,* pp. 77-81, 2015.
61. X. Lin, X. Li and Z. Zheng, "Evaluating borrower's default risk in peer-to-peer lending: evidence from a lending platform in China," *Applied Economics,* vol. **49**, no. 35, 2017.
62. Y. Zhang, D. Wang, Y. Chen, H. Shang and Q. Tian, "Credit Risk Assessment Based on Long Short-Term Memory Model".
63. Z. Huang, J. Deng, M. Xiong, Y. Ren and Y. Qiao, "A comparison of US and UK P2P lending regulation systems and study on patterns of P2P lending regulation in China," *Studies on Financial Regulation,* no. 10, pp. 45-58, 2014.
64. Y. Xia, C. Liu, Y. Li and N. Liu, "A boosted decision tree approach using Bayesian hyperparameter optimization for credit scoring," *Expert Systems With Applications,* no. 78, pp. 225-241, 2017.
65. D. S. Nascimento, A. L. Coelho and A. M. Canuto, "Integrating complementary techniques for promoting diversity in classifier ensembles: A systematic study," *Neurocomputing,* no. 138, pp. 347-357, 2014.
66. L. Nanni and A. Lumini, "An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring," *Expert Systems with Applications,* no. 36, pp. 3028-3033, 2009.
67. S. Lessmann, B. Baesens, H. V. Seow and I. C. Thomas, "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research," *European Journal of Operational Research,* no. 247, pp. 124-136, 2015.
68. "Lending Club," [Online]. Available: https://www.lendingclub.com/. [Accessed 23 2 2018].
69. "WE," [Online]. Available: https://www.we.com/. [Accessed 23 2 2018].