2014 IEEE International Conference on Systems, Man, and Cybernetics
October 5-8, 2014, San Diego, CA, USA

How to Protect Investors? A GA-based DWD Approach for Financial Statement Fraud Detection

Xinyang Li
School of Information
Renmin University of China
Beijing, China
xyang4173@sina.com

Wei Xu
School of Information
Renmin University of China
Beijing, China
weixu@ruc.edu.cn

Xuesong Tian
School of Information
Renmin University of China
Beijing, China
tianxuesong@outlook.com

Abstract—**As one type of the financial fraud, financial statement fraud has not only led to a huge loss for individual investors and financial institutions, but also impacted the overall stability of the whole industry. This paper used financial and textual features extracted from annually submitted 10-k filings and combined data and text mining techniques for detection of financial statement fraud. When the dimension of samples is larger than the sample size, namely high dimension low sample size (HDLSS), distance weighted discrimination (DWD) model, which has a good generalization performance in HDLSS contexts, is used to detect financial statement fraud. We also adopted genetic algorithm to improve the performance of classifiers, including DWD, Support Vector Machine, Back Propagation Neural Networks and Decision Tree for feature selection and parameter optimization. Compared with other GA-based classification models, the proposed GA-based DWD model achieved relatively high classification accuracy with fewer input features, which proves that this model is a promising tool for detection of fraudulent financial statements.**

Keywords—financial statement fraud; DWD model; Genetic algorithm; Feature selection; Text mining

I. INTRODUCTION

As one of the common financial fraud, financial statement fraud may bring great loss to investors. Compared with traditional audit methods, data mining techniques assist auditors in assessing the fraud risk and providing analytical results for identification of fraudulent financial statement. It helps investors, including individuals, investment companies and financial institutions, with investment decision making and provides quantitative support for audit firms. The early detection of financial statement fraud will prevent fraudulent companies from trapping in financial risk.

In previous studies, data mining and text mining are two major techniques used by researchers for detection of financial statement fraud. The detection of financial statement fraud is usually regards as a classification problem. Michael et.al applied artificial neural network to the detection of management fraud [1]. Kirkos et al. compared the performance of ID3, neural networks and Bayesian belief networks based on the data of 76 Greek manufacturing companies [2]. Ravisankar et al. found that probabilistic neural network outperformed other classifiers on dataset involving Chinese corporations [3]. Kotsiantis et al. used ensemble classifiers, which could provide more accurate results than single classifier, and achieved good performance on the detection of financial statement fraud [4].

Recently, researchers focused on MDA section of Form 10-k, using text mining techniques to identify deception and fraud. Churyk et al. used t-test to confirm the hypotheses of the language characteristics of MDA part of SEC 10-k filings between restated and matched companies [5]. Glancy et al. conducted a cluster analysis based on MDA using singular value decomposition vector (SVD) [6]. Humpherys et al. analyzed the nature of language in deceptive filings with psychological deception theory, and the proposed hypotheses were confirmed with the analysis of independent sample t-test [7].

However, the data mining and text mining techniques were used independently for detection of financial statement fraud in previous researches. Few researchers combined the financial and textual features to detect deception. Financial data can be manipulated easily and only reflect past performance of operation, the contents of 10-k, however, give insight into the results of financial position and future prospect. Therefore, the combination of financial and textual features will improve the ability of detecting financial fraud; on the other hand, will lead to an obvious increment of the number of features. Although it has been proved that Support Vector Machine (SVM) [3, 7] and other classification models have a good performance in detection of financial fraud, there is a strong need for novel HDLSS models. We tested the performance of Distance Weighted Discrimination (DWD) model [8] and used it in detecting financial statement fraud.

To improve the performance of classifiers, redundant and irrelevant features need to be removed in feature space and feature selection is helpful in preprocessing raw data [9]. Feature selection algorithms can be classified into filter approach and wrapper approach. F-score is one of common methods to fulfill filter-based feature selection [10]. Wrapper approach, however, is the feature selection performed dependently of learning algorithms, such as back propagation neural networks (BPNN), and SVM. For example, Guyon et al. proposed an SVM-RFE (Recursive Feature Elimination) feature selection method [11], while Li et al. proposed a statistics-based wrapper of feature selection method and has been applied in financial distress identification research with SVM [12].

978-1-4799-3840-7/14/$31.00 ©2014 IEEE

In addition to feature selection, parameter optimization is usually performed to improve the overall accuracy of classifiers, and genetic algorithm is an efficient tool for feature selection and parameter optimization simultaneously. Huang et al. [13] proposed GA-based SVM for feature selection and parameter optimization, which improve the performance of SVM significantly. Huang [14] later applied the strategy to credit scoring, and the results showed that GA-based SVM outperformed other models with small number of features selected.

However, most researchers encoded solutions as binary string during the process of GA-based feature selection and parameter optimization [13-14] This type of encoding cannot perform neighborhood search in crossover and mutation operation. In this research, we used real number encoding to improve search efficiency. In addition, the number of selected features has an impact on the performance of classifiers, so we introduced a penalty rate during GA process, which increased the probability that individuals with fewer features are selected. We used proposed GA-based DWD to optimize parameters and feature subsets in high dimension low sample size (HDLSS) contexts.

The main contributions of this paper are: (1) Financial and textual features are combined to enhance detection of financial statement fraud; (2) DWD model, which has a good performance in HDLSS settings, is applied to the detection of fraudulent financial statement; (3) An improved GA-based DWD is used to fulfill feature selection and parameter optimization and is proved to achieve relatively high classification accuracy using smallest number of features.

II. THEORITICAL FOUNDATION

In high dimension low sample size settings, where the dimension of features is larger (often much larger) than sample size, traditional multivariate analysis is useless, and classification algorithms, such as SVM [15], are suffered from "data piling". A distance weighted discrimination (DWD), proposed by Marron et al., is regarded as an effective tool to deal with real problems in HDLSS [8].

In the novel view of the performance of classifiers via projecting samples in the direction of normal vector, there is room for improvement in DWD by solving data piling problem. The optimization problem of DWD is to minimize the sum of the reciprocals of the residuals, where samples are distance weighted. In HDLSS contexts, linear classifier is good enough to perform, and the DWD optimization problems can be formulated as below:

$$\min_{r,\mathrm{w},\beta,\xi} \quad \sum_i \frac{1}{r_i} + Ce'\xi \tag{1}$$

$$\begin{aligned} s.t. \quad & r = YX'w + \beta y + \xi \\ & \frac{1}{2} w'w \le \frac{1}{2} \\ & r \ge 0 \\ & w \ge 0 \end{aligned}$$

where y_i denotes the class indicators corresponding with x_i ; $w \in \Re^d$ denotes the normal vector of the separating hyperplane; $\beta \in \Re$ determines the position of hyperplane; ξ is a penalized vector and constant $C \ge 0$ is a penalty parameter. Then the minimum optimization problem is transformed to second-order cone programming (SOCP) problem, and it can be solved by obtaining the dual and using standard optimization program.

III. IMPROVED GA-BASED PARAMETER OPTIMIZATION AND FEATURE SELECTION

A. Overview

Recently, data mining techniques have been successfully applied to fraud detection. In our research, GA-based classification models are used to perform feature selection and parameter optimization. The architectures of the proposed GA-based feature selection and parameters optimization for classification model is shown as Fig. 1.

We conducted 3-fold cross validation. For each of the folds, the average accuracy of the training data via 5-fold cross validation was calculated to evaluate the performance of classifiers. Then, genetic algorithm was used for feature selection and parameter optimization, and the predictive accuracy on the testing set is measured with best parameter and selected features. The raw dataset D is randomly partitioned into 3 equal size subsets, each of which takes turns at being independent testing set T and the remaining 2 subsets serve as training set $k = D - T$. For each of the folds, proceed as follow:

Step1. Using genetic algorithm to find optimal parameters and feature subsets:

Step1.1. Initialization: Generate initial population, each individual i is comprised of the genetic representation of parameters and selected features.

Step1.2. Fitness evaluation: for each individual, do the following:

(i) Decoding chromosomes, and the parameters and selected features are obtained.

(ii) Conduct 5-fold cross validation with the parameters and selected features, obtain the average $Accuracy_i$ and calculate the fitness.

Step1.3. Genetic operation: perform selection, crossover and mutation. Offspring are produced.

Step1.4. Termination conditions: Go to Step1.2 unless the termination condition is reached. Then the optimal parameters and feature subsets are obtained.

Step2. Regenerate the testing set T with the optimal feature subsets, and set the optimal parameters to the classifier. Run the classification model and measure the classification accuracy on the testing set T .

Step3. Overall accuracy is averaged across 3 partitions to evaluate the performance of GA-based models.

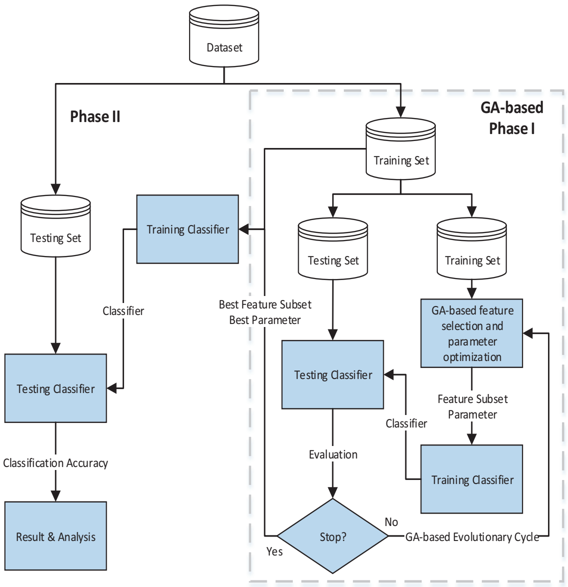

Fig. 1. The framework of GA-based feature selection and parameter optimization

We also conducted the experiments using SVM(Linear), SVM(RBF), NN, DT(C4.5) and improved GA for feature selection and parameter optimization. The parameters to be optimized for these models are summarized in TABLE I.

We applied genetic algorithms to the classification models mentioned above and compared detection performance of them. DWD and SVM(Linear) are instances of linear classifiers. SVM(RBF), NN and C4.5 are nonlinear classification models. Especially, C4.5 results in further dimension reduction because of its pruning process.

B. The modeling process

Several classification tools are used to model the relationship between the fraud results and input samples. To improve the performance of classifiers, genetic algorithms are implemented to optimize parameters and features subsets.

Each chromosome is comprised of two parts, the parameters to be optimized and selected feature subsets. It will produce random individuals during the crossover and mutation process using binary coding system, which cannot perform neighborhood search. Therefore, we encoded the parameters using real number codes and features with binary strings. The bit with value "1" indicates selected feature, and the "0" represents the feature which is not selected. Fig. 2 illustrates the genetic representation of our design.

TABLE I. PARAMETERS TO BE OPTIMIZED FOR CLASSIFICATION MODELS

Model	Parameters to be optimized
SVM(Linear)	C - penalty parameter
SVM (RBF)	C - penalty parameter γ - the kernel parameter gamma
NN	lr - learning rate N - numbers for nodes in the single hidden layer
C4.5	N - minimum number of instances per leaf C - confidence factor

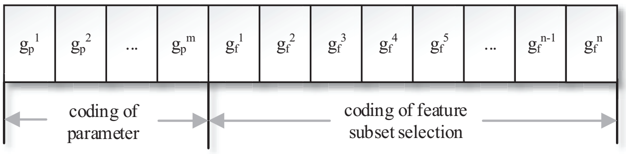

Fig. 2. The chromosome with parameters and input features representation

During the evaluation and selection process, the fitness value of each individual is calculated through fitness function to assess the quality of candidate solutions. Then, crossover and mutation may happen to some individuals with possibilities. The crossover procedure using real number codes is defined as:

$$Offspring_1 = r \times Parent_1 + (1-r) \times Parent_2 \tag{2}$$

$$Offspring_2 = (1-r) \times Parent_1 + r \times Parent_2 \tag{3}$$

where $Offspring_1$, $Offspring_2$ represent the children generated by crossover; $Parent_1$, $Parent_2$ denote the parents selected to breed; $r \in (0,1)$ is a random number between 0 and 1.

Mutation is the process that a bit in a genetic sequence of individual is altered from its original state occasionally, which makes it possible to avoid locally optimal solutions. For the parameters, the real-coded mutation is defined as:

$$Offspring = \begin{cases} Parent + d' & Parent + d > C_{\max} \\ Parent + d & C_{\min} < Parent + d \ < C_{\max} \\ Parent - d'' & Parent + d < C_{\min} \end{cases} \tag{4}$$

where

$$d = r \times amplitude \quad r \in [-1,1]$$
$$d' = r' \times (C_{\max} - Parent) \quad r' \in (0,1)$$
$$d'' = r'' \times (Parent - C_{\min}) \quad r'' \in (0,1)$$

Offspring denotes the child generated by mutation; *Parent* denotes the individual selected to perform mutation;, *amplitude* is the amplitude for mutation; $C_{\min}$, $C_{\max}$ are the minimum and maximum value of parameters, respectively; r is a random number between -1 and 1, both r' and r'' are random numbers between 0 and 1. Selection, crossover and mutation process are repeated until the termination condition is satisfied.

Fitness value depends on the accuracy of classifiers. Classification models, namely DWD, SVM, NN and C4.5, are applied to the training set and testing set using the parameters and selected features generated by individual. When two or more individuals have identical classification accuracy, the one with larger number of features will produce a lower fitness value by introducing a penalty rate δ to reduce the probability of being selected. The penalty of the i th individual is as below:

$$Penalty_i = 1 - \frac{F_i - F_{\min}}{F_{\max} - F_{\min}} \delta \tag{5}$$

where F_i is the number of selected features of *i* th individual; $F_{\min}$, $F_{\max}$ represents the minimum and maximum number of selected features; δ is the penalty rate. Here, we set $\delta = 0.2$ .The fitness of the *i* th individual is defined as follow:

$$Fitness_i = Accuracy_i \times Penalty_i \qquad (6)$$

where $Fitness_i$ is the fitness value of the *i* th individual; $Accuracy_i$ is the average accuracy of the *i* th individual via 5-fold cross validation; $Penalty_i$ is the penalty rate of the *i* th individual.

IV. EXPERIMENTAL RESULTS

A. Company's Form 10-k filing data set

To guarantee available and valid data, we collected annual report 10-k filings of public companies available at Electronic Data-Gathering, Analysis, and Retrieval (EDGAR) database on U.S. Securities and Exchange Commission (SEC) website.

The annual financial statements of 2007 from 111 companies were obtained from the EDGAR database, and 25 10-k filings of companies involved in financial statement fraud were included according to the AAERs. For the companies which are not in the AAERs, we regarded them as nonfraudulent companies. TABLE II showed the selection process of sample companies. 57 companies were remained for our experiment, including 12 fraudulent companies, which account for 21.05%.

According to Standard Industrial Classification (SIC), we further conducted a statistical analysis. TABLE III summarizes composition of sample companies by SIC category. Manufacturing, transportation or communications companies have the largest proportion (33%), and others are widely distributed in other 7 industries, which reveals that sample companies are not tied to any particular industry category.

B. Feature construction

In our research, the financial features were combined with textual features in order to enhance the detection of financial statement fraud using data mining and text mining approaches. The financial data are important indicators to reflect performance and financial position of companies, but they can be easily manipulated and fail to show the future value. Public companies are required to provide an overview of the previous year of operations and future prospects of the upcoming year from management aspects in Management's Discussion and Analysis (MDA) section, which helps readers have a detailed understanding of business and financial status of companies.

1) Financial feature

While constructing the features, we used 35 items proposed by Ravisankar [3] and extracted 30 items from them as listed in TABLE IV for some of the data are unavailable. Some minor adjustment to the features is made according to the latest 10-k form. Financial features include total assets, total liabilities, net profit, accounts receivable and so on, and 23 financial ratios are the relative magnitude of two basic financial values, such as rate of return on total assets, debt to equity ratio and so on.

TABLE II. SAMPLE COMPANIES SELECTION PROCESS

Description	Count of companies
Count of companies of 2007 obtained from SEC website	111
Count of fraudulent companies by searching through AAERs	25
Count of initial sample companies	**136**
Count of duplicated companies	(43)
Count of unique sample companies	**93**
Count of companies with missing values	(36)
Final count of sample companies	**57**

We standardized data using z-score scaling method in data preprocessing step to deal with the financial data of different units and scale.

2) Text feature

For public companies, the Form 10-k is required to file annually by the SEC. In general, audited financial statements are standardized representation of companies' performance without management sayings. The MDA section fills the gap by allowing management to give a narrative explanation about the performance and outline future plan.

High-performance companies make use of Form 10-k to communicate with investors. The fraudulent companies, however, are not willing to explain in detail, so fewer words are expected in MDA section of fraudulent financial statement. Longer sentences and words are served to obfuscate and be less readable. Higher lexical diversity and rich examples may be found in MDA to attempt to be convincing and promising. Furthermore, more present tense verbs are expected to use to reduce the uncertainty of companies and hide negative information. In view of the above, as displayed in TABLE V, textual features, including total words of MDA, indicator of lexical diversity and so on, can be classified into 5 categories.

Both the financial and textual features are indicators to detect fraud, so we combined 30 financial features and 10 textual features proposed above as the input features.

TABLE III. INDUSTRIAL DISTRIBUTION OF SAMPLE COMPANIES

SIC Code	SIC description	Frequency	Percentage
0100 - 0999	Agriculture, Forestry and Fishing,	5	8.77%
1000 - 1499	Mining	2	3.51%
1500 - 1799	Construction	5	8.77%
2000 - 3999	Manufacturing, Transportation, Communications	19	33.33%
4000 - 4999	Electric, Gas and Sanitary service	6	10.53%
5000 - 5199	Wholesale Trade	5	8.77%
5200 - 5999	Retail Trade	7	12.28%
7000 - 8999	Services	8	14.04%
Total		**57**	**100%**

TABLE IV. FINANCIAL FEATURES USED FOR DETECTION OF FINANCIAL STATEMENT FRAUD

No	Financial items	No	Financial items
1	Total Liabilities	16	Current assets/Current liabilities
2	Total Assets	17	Revenue/Property, plant and equipment
3	Gross Profit	18	Cash/Total assets
4	Net Profit	19	Inventory/Current liabilities
5	Revenue	20	Total liabilities/Stockholders' equity
6	Cash and Cash equivalents	21	Long term liabilities/Total assets
7	Accounts Receivable	22	Net profit/Gross profit
8	Inventory/revenue	23	Total liabilities/Total assets
9	Inventory/Total assets	24	Property, plant and equipment/Total liabilities
10	Gross profit/Total assets	25	Cash and Cash equivalents/Current assets
11	Net profit/Total assets	26	Accounts receivable/Total assets
12	Current assets/Total assets	27	Gross profit/Revenue
13	Net profit/ Revenue	28	Revenue /Last year's revenue
14	Accounts receivable/Revenue	29	Account receivable/Accounts receivable of last year
15	Revenue/Total assets	30	Total assets/Total assets of last year

C. Experimental results and comparison

1) Comparison with other feature selection methods

Effective parameter settings improve performance of classification models. Recently, various methods have been proposed for parameter optimization. Grid search [17] is a straightforward method, which tries a group of parameter pairs and the one with the best cross-validation accuracy is chosen as the best parameters.

In DWD model, only penalty parameter C needs to be determined, and we adopted the one-dimensional grid search method. Dataset D is randomly partitioned into 3 subsets, for the k th subset, create training set $T = D - k$ and search as follow:

Step1. Consider a grid search space $C = \{2^{-5}, 2^{-4.5}, 2^{-4}, ..., 2^{8}\}$.

Step2. For each of parameter, conduct 5-fold cross validation on testing set T .

Step3. Choose the parameter with lowest CV error as optimal parameter.

Step4. Use optimal parameter to build model, and conduct classification on testing data.

Overall accuracy is averaged across 3 partitions.

In addition to proper parameters settings, we consider using F-score technique [10] for filter-based feature selection. F-score method is a common method to measure the discrimination of two classes of data, the larger the F-score is, the discriminative the feature is likely to be.

TABLE V. TEXTUAL FEATURE USED FOR DETECTION OF FINANCIAL STATEMENT FRAUD

Category	No.	Text items	Description
Standard Linguistic Variables	1	Total Words	Count of a written character or combination of characters representing a written word
	2	Sentences Length	Total words/total number of sentences
	3	Words Length	Total number of letters/total words
Lexical diversity	4	Lexical diversity	Total number of different words or terms/total number of words or terms * 100, which is the percentage of unique words or terms in all words or terms*100
Content richness	5	Colons	Count of colons/total number of words or terms*100
	6	Semicolons	Count of semicolons/total number of words or terms*100
	7	For example	Count of the term "for example"/total number of words or terms*100
affective tendency	8	Positive emotion [a]	Total number of words or terms indicating positive emotion / total number of words or terms*100;
	9	Negative emotion [a]	Total number of words or terms indicating anxiety/ total number of words or Terms*100
Certainty	10	Present tense	Total number of present tense verbs/total number of words or Terms*100

[a]. Emotional lexicon is defined as Minqing Hu and Bing Liu.[16]

For the given training set $x_k, k = 1, 2, ..., m$, the numbers of positive and negative samples are n_+ and n_- respectively, and the F-score of the i th feature is defined as follow:

$$F(i) \equiv \frac{\left(\bar{x}_i^{(+)} - \bar{x}_i\right)^2 + \left(\bar{x}_i^{(-)} - \bar{x}_i\right)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}\left(x_{k,i}^{(+)} - \bar{x}_i^{(+)}\right)^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_-}\left(x_{k,i}^{(-)} - \bar{x}_i^{(-)}\right)^2} \quad (7)$$

where $\bar{x}_i$, $\bar{x}_i^{(+)}$ and $\bar{x}_i^{(-)}$ are the average of the i th feature of whole, positive and negative data sets respectively; $x_{k,i}^{(+)}$, $x_{k,i}^{(-)}$ are the i th feature of the k th positive and negative samples, respectively.

Dataset D is randomly partitioned into 3 subsets, for the k th subset, create training set $T = D - k$ and the procedure for feature selection and parameter optimization is as follow:

Step1. Calculate F-score of each feature

Step2. Sort F-score, set possible numbers of selected features, namely threshold $f = \left[\frac{n}{2^i}\right], i = 0, 1, ..., m$, where m is the max integer that satisfies $\frac{n}{2^m} \geq 1$.

Step3. For each f , do the following:

(i) Keep the first f features according to F-score.

(ii) For each possible parameter C in grid search, conduct 5-fold cross validation on testing set T.

(iii) Choose the parameter with lowest CV error as optimal parameter.

Step4. Choose the threshold f with lowest CV error, and keep the first f features. Conduct classification on testing data.

Overall accuracy is averaged across 3 partitions.

The two methods for feature selection and parameter optimization stated above are "Grid+DWD" and "Grid+F-score+DWD". They are used to compare with "GA+DWD" proposed in section 3, and the experiment results based on our dataset were summarized in TABLE VI. During the 3-fold cross validation, each fold contains 38 instances and 40 features, which are considered as HDLSS data. The feature selection process may lead to the reduction of dimension; research shows that DWD model still has a good performance in non-HDLSS settings [8].

Since "Grid+DWD" only deals with parameters optimization, all the 40 features are kept to build model with 87.19% classification accuracy. The accuracy of "Grid + F-score + DWD" and "GA+DWD" achieved 89.47% and 91.23%, respectively. "GA+DWD" had higher precision and recall of fraudulent financial statement than other methods, which means 88.9% of the companies classified as fraud are actually fraudulent, and 66.7% fraudulent companies are identified with "GA+DWD".

We performed a Friedman test to find that there is no significant differences among the three methods (p= 0.1407). The average number of selected features of "GA+DWD" is 13.67 and it has a good performance on the detection of fraudulent financial statement. The figure of "Grid+DWD" and "Grid+F-score+DWD" are 40 and 16.67, respectively, which proves "GA+DWD" used fewer features without degrading classification accuracy.

2) *Comparison with other GA-based models*

The parameter setting for genetic algorithm is as follow: population size 20, crossover rate 0.6, mutation rate 0.1, mutation amplitude 5, one-point crossover, roulette wheel selection, and the algorithm stopped when the generation number achieved 1000 or there is no improvement of the fitness value during the last 50 generations.

TABLE VI. RESULTS SUMMARY OF FEATURE SELECTION AND PARAMETER OPTIMIZATION

	Selected feature number	**Overall Accuracy**	**Precision of fraud**	**Recall of fraud**
Grid search+ DWD	40	87.19%	1.000	0.417
Grid search+ F-score+ DWD	16.67	89.47%	0.875	0.583
GA + DWD	13.67	91.23%	0.889	0.667

As the number of fraudulent and nonfraudulent companies is unbalanced, we put more emphasis on the TP rate, namely recall, of the fraudulent companies than overall accuracy. The result with high TP rate in classifying the fraudulent class indicates the companies involved in fraud can be correctly identified among companies, most of which are nonfraudulent ones. In addition, we compared the overall accuracy and precision of DWD, SVM, BPNN and C4.5 via 3-fold cross validation in TABLE VII.

Regarding the precision of fraudulent companies, DWD outperforms other classification models with precision of 88.9%, which means 88.9% of companies classified as fraud are actually fraudulent. Even though BPNN has the lowest accuracy, it is still an effective tool to detect financial statement fraud for its recall is as high as 88.3%. In addition, SVM(Linear) showed good performance of detection. Both SVM(Linear) and DWD are linear classification models, but they are outperformed non-linear classification models. SVM(RBF) and C4.5 has a better performance on the detection of nonfraudulent companies comparing to its recognition ability of the fraudulent companies.

As the sample size of fraudulent companies is larger than that of nonfraudulent companies, BPNN performs best to detect fraudulent financial statement with 83.3% TP rate of fraud, even though DWD achieves the highest overall accuracy. It indicates that 83.3% of fraudulent companies, which takes 21.05% proportion of all sample companies, are identified. The accuracy for SVM(RBF) is 85.97%, it, however, still cannot provide a decision support for detection of companies involved in fraud because the TP rate for fraudulent cases of SVM(RBF) is only 41.7%. Both DWD and SVM(Linear) achieve 66.7% TP rate for fraudulent cases, which outperform SVM(RBF) and C4.5.

The average number of selected features of each GA-based classifier is shown in TABLE VIII. DWD model uses fewest features and also performs well. Although BPNN performs well with 83.33% TP rate, it uses 21.33 features to accomplish detection, which is much larger than 13.67 of DWD. Moreover, DWD and SVM(Linear) have the same TP rate of 66.7%, DWD, however, uses 3 less features than SVM(Linear) with classification accuracy 3.51% higher than SVM(Linear), which shows that DWD achieves relatively high classification accuracy with small number of input features.

Marron [8] indicated that "data piling" caused by the same projections onto the normal direction vector reflects the performance of classifiers. We further studied the numbers of support vectors of SVM. The results showed that the average number of support vectors of SVM(Linear) is 10.67 with better performance, while SVM(RBF) is 25.67. It confirms the conclusion proposed by Marron that "data piling" may lead to the reduction of performance of SVM(RBF).

TABLE VII. RESULTS SUMMARY FOR GA-BASED DWD, SVM, BPNN, C4.5 MODELS

	Overall Accuracy	**Precision**		**Recall**	
		Fraud	***Nonfraud***	***Fraud***	***Nonfraud***
DWD	91.23%	0.889	0.917	0.667	0.978
SVM(Linear)	89.47%	0.800	0.915	0.667	0.956
SVM(RBF)	85.97%	0.833	0.863	0.417	0.978
BPNN	75.44 %	0.455	0.943	0.833	0.733
C4.5	82.46 %	0.600	0.872	0.500	0.911

TABLE VIII. AVERAGE SELECTED FEATURE NUMBERS OF CLASSIFICATION MODELS

	Avg.number of selected features	**Recall(fraud)**	**Overall Accuracy**	**Avg. number of support vectors**
DWD	13.67	0.667	92.98%	-
SVM(Linear)	16.67	0.667	89.47%	10.67
SVM(RBF)	16	0.417	85.97%	25.67
BPNN	21.33	0.833	75.44%	-
C4.5	17.33	0.500	82.46%	-

Regarding the computation time, C4.5 spend less time compared to other classification model. The average time of running one fold is 62.29 seconds, and 98.68 seconds and 107.56 seconds for SVM(Linear) and SVM(RBF), respectively. The running time of BPNN and DWD is much longer than the other 3 models, and BPNN is the most time-consuming, about 1.74 times longer than DWD. In fact, software environment affects the average running time, DWD, SVM and BPNN were performed under MATLAB environment, while SVM was developed by C language by extending the Libsvm[18]. The implementation of C4.5 was carried out with the java package provided by weka software. Basically, it is not comparable for the running time under different software environment, and we are not going to discuss it in further detail.

V. CONCLUSIONS

By combining the data mining and text mining techniques, both financial and textual features are used to detect financial statement fraud. Feature combination increases the interpretability and explanatory power of models, and enhances detection of fraudulent financial statement. Furthermore, we used the DWD model, which has a good performance in HDLSS contexts and applied improved genetic algorithm to performing feature selection and parameter optimization. Experiment results showed that DWD model outperformed other classification models with less number of selected features. Especially in HDLSS settings, nonlinear classification model failed to perform well, sometimes even worse. The DWD model is turned out to be a powerful tool in detecting financial statement fraud, which may help stakeholders reduce unnecessary losses.

In future, semi-supervised learning will be considered to deal with a weakly-labeled dataset of companies, for there are 3.6 to 11.6 years between the release date and the filing date referring to the AAERs issued from 2006 to 2008 [6]. Furthermore, oversampling techniques are to be used to solve the class imbalance problem. Also, new features, such as stock data, can be used to assess the likelihood of fraud risk and provide better detection of financial statement fraud.

ACKNOWLEDGEMENTS

This research work was partly supported by 973Project (Grant No. 2012CB316205), National Natural Science Foundation of China (Grant No.71001103, 91224008, 91324015), Humanities and Social Sciences Foundation of the Ministry of Education (No. 14YJA630075), Beijing Social Science Fund (No. 13JGB035), Beijing Natural Science Foundation (No. 9122013), Beijing Nova Program (No.Z131101000413058), and Program for Excellent Talents in Beijing.

REFERENCES

[1] Fanning, Kurt M., and Kenneth O. Cogger. "Neural network detection of management fraud using published financial data." International Journal of Intelligent Systems in Accounting, Finance & Management 7.1, 1998, pp. 21-41.

[2] Kirkos, Efstathios, Charalambos Spathis, and Yannis Manolopoulos. "Data mining techniques for the detection of fraudulent financial statements." Expert Systems with Applications 32.4, 2007,pp. 995-1003.

[3] P Ravisankar, V Ravi, G Raghava Rao, I Bose. "Detection of financial statement fraud and feature selection using data mining techniques." Decision Support Systems 50.2, 2011, pp. 491-500.

[4] S. Kotsiantis, E. Koumanakos, D. Tzelepis and V. Tampakas. "Forecasting Fraudulent Financial Statements using Data Mining." Enformatika 12, 2006.

[5] Churyk, Natalie Tatiana, Chih-Chen Lee, and B. Douglas Clinton. "Early detection of fraud: Evidence from restatements." Advances in Accounting Behavioral Research 12, 2009, pp. 25-40.

[6] Glancy, Fletche r H., and Surya B. Yadav. "A computational model for financial reporting fraud detection." Decision Support Systems 50.3, 2011, pp. 595-601.

[7] Sean L. Humpherys, Kevin C. Moffitt, Mary B. Burns, Judee K. Burgoon, William F. Felix. "Identification of fraudulent financial statements using linguistic credibility analysis." Decision Support Systems 50.3, 2011, pp. 585-594.

[8] Marron, J. S., Michael J. Todd, and Jeongyoun Ahn. "Distance-weighted discrimination." Journal of the American Statistical Association 102.480, 2007, pp. 1267-1271.

[9] Oreski, Stjepan, and Goran Oreski. "Genetic algorithm-based heuristic for feature selection in credit risk assessment." Expert Systems with Applications 41.4, 2014, pp. 2052-2064.

[10] Chen, Yi-Wei, and Chih-Jen Lin. "Combining SVMs with various feature selection strategies." Feature Extraction. Springer Berlin Heidelberg, 2006, pp. 315-324.

[11] I Guyon, J Weston, S Barnhill, V Vapnik. "Gene selection for cancer classification using support vector machines." Machine learning 46.1-3, 2002, pp. 389-422.

[12] H Li, CJ Li, XJ Wu, J Sun. "Statistics-based wrapper for feature selection: An implementation on financial distress identification with support vector machine." Applied Soft Computing 19, 2014, pp. 57-67.

[13] Huang, Cheng-Lung, and Chieh-Jen Wang. "A GA-based feature selection and parameters optimizationfor support vector machines." Expert Systems with applications 31.2, 2006, pp. 231-240.

[14] Huang, Cheng-Lung, Mu-Chen Chen, and Chieh-Jen Wang. "Credit scoring with a data mining approach based on support vector machines." Expert Systems with Applications 33.4, 2007, pp. 847-856.

[15] Vapnik, Vladimir. The nature of statistical learning theory. springer, 1995.

[16] Hu, Minqing, and Bing Liu. "Mining and summarizing customer reviews." Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2004.

[17] Hsu, Chih-Wei, Chih-Chung Chang, and Chih-Jen Lin. "A practical guide to support vector classification.", 2003.

[18] Chang, C. C., & Lin, C. J. LIBSVM: a library for support vector machines. Available from http://www.csie.ntu.edu.tw/~cjlin/libsvm/. 2001.