Expert Systems with Applications 42 (2015) 2510–2516

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Detecting credit card fraud by Modified Fisher Discriminant Analysis

Nader Mahmoudi, Ekrem Duman *

Özyeğin University Çekmeköy Campus, Industrial Engineering Department, 34794 Istanbul, Turkey

ARTICLE INFO

Article history:
Available online 6 November 2014

Keywords:
Credit card fraud
Linear discriminant
Fisher linear discriminant function
Modified Fisher Discriminant
Profitability

ABSTRACT

In parallel to the increase in the number of credit card transactions, the financial losses due to fraud have also increased. Thus, the popularity of credit card fraud detection has been increased both for academicians and banks. Many supervised learning methods were introduced in credit card fraud literature some of which bears quite complex algorithms. As compared to complex algorithms which somehow over-fit the dataset they are built on, one can expect simpler algorithms may show a more robust performance on a range of datasets. Although, linear discriminant functions are less complex classifiers and can work on high-dimensional problems like credit card fraud detection, they did not receive considerable attention so far. This study investigates a linear discriminant, called Fisher Discriminant Function for the first time in credit card fraud detection problem. On the other hand, in this and some other domains, cost of false negatives is very higher than false positives and is different for each transaction. Thus, it is necessary to develop classification methods which are biased toward the most important instances. To cope for this, a Modified Fisher Discriminant Function is proposed in this study which makes the traditional function more sensitive to the important instances. This way, the profit that can be obtained from a fraud/legitimate classifier is maximized. Experimental results confirm that Modified Fisher Discriminant could eventuate more profit.

1. Introduction

Nowadays, by increasing credit card transactions in not only online purchases but also regular purchases, credit card fraud is becoming rampant. Today, both merchants and clients are affected in terms of financial losses caused by credit card fraud. Some references reported billions of dollars lost annually due to credit card fraud (Chan, Fan, Prodromidis, & Stolfo, 1999; Chen, Chen, & Lin, 2006). CyberSource (2013) reported in 14th annual online fraud that the actual amount of losses will increase by the increasing online sales. It is also reported that the estimated total loss increased up to $3.5 billion in 2012 by 30% increase from 2010. Evidently, with the growth in the number of credit card transactions as a payment system, 70% of consumers in U.S. had concerns about identity fraud significantly (McAlearney & Breach, 2008).

Considering this huge amount of financial loss, prevention of credit card frauds is the most concerning issue for researchers in data mining area. Because of large amount of credit card transactions, detecting about 2.5 percent of frauds leads to save over a million dollar per year (Brause, Langsdorf, & Hepp, 1999). However, along with the development of fraud detection techniques, fraudulent activities done by criminals also have been evolved to avoid detection (Bolton & Hand, 2001). Thus, to perform in the best way, researchers are trying to make modifications in the existing methods or develop new methods to maximize number of frauds detected.

Bolton and Hand (2001) categorized credit card frauds into two groups: application frauds and behavioral frauds. Application frauds occur when fraudsters obtain new cards by presenting false information to issuing companies. On the other hand, behavioral frauds include four types: mail theft, stolen/lost cards, counterfeit cards, and ‘card holder not present’ fraud. In modern banking system, the more the online transactions increase, the more counterfeit and ‘card holder not present’ frauds occur; where in both of these two types of fraud, fraudsters obtain credit card details without the knowledge of card holders. Bolton and Hand (2002) presented a good discussion on the issues and challenges in fraud detection research together with Provost (2002).

In the literature, there are many studies made on credit card fraud detection in some of which methods for learning systems are proposed. If we look at these studies, most of the credit card fraud detection systems are using supervised learning algorithms like neural networks (Aihua, Rencheng, & Yaochen, 2007; Juszczak, Adams, Hand, Whitrow, & Weston, 2008; Quah & Sriganesh, 2007;

* Corresponding author.
E-mail addresses: nader.mahmoudi@ozu.edu.tr (N. Mahmoudi), ekrem.duman@ozyegin.edu.tr (E. Duman).

http://dx.doi.org/10.1016/j.eswa.2014.10.037
0957-4174/© 2014 Elsevier Ltd. All rights reserved.

Schindeler, 2006), decision tree techniques such as ID3, C4.5, and C&RT (Chen, Chiu, Huang, & Chen, 2004; Chen, Luo, Liang, & Lee, 2005; Mena, 2003; Wheeler & Aitken, 2000), and support vector machines (SVMs) (Leonard, 1993).

Sahin and Duman (2011) carried out a study using Artificial Neural Network (ANN) and logistic regression (LR) to score transactions where they are flagged as fraudulent or legitimate transactions. They concluded that ANN outperforms LR based on results. However, as skewness of training set increases, the performance of all models decrease.

Aihua et al. (2007) investigated the efficacy of applying classification models to credit card fraud detection problems. Three different classification methods, i.e. decision tree, neural networks and logistic regression are tested for their applicability in fraud detections. Their paper provides a useful framework to choose the best model to recognize the credit card fraud risk based on different performance measures.

In the most of related studies in literature, the cost of a false negative (labeling a fraudulent transaction as legitimate) and a false positive (labeling a legitimate transaction as fraudulent) are taken as equal to each other. However, in this domain the cost of a false negative is much higher than the cost of a false positive and in fact it varies from transaction to transaction. To cope with the higher cost of a false negative, some researches used adjusted cost matrices during the training phase of their classifiers (Langford & Beygelzimer, 2005; Maloof, 2003; Sheng & Ling, 2006; Zhou & Liu, 2006). However, the variable character of misclassification costs is undertaken in only a few studies so far (Duman & Elikucuk, 2013a; Duman & Ozcelik, 2011; Sahin, Bulkan, & Duman, 2013; Sahin & Duman, 2010; Sahin & Duman, 2011).

Actually the main issue in credit card fraud detection modeling is to get the most possible profit from the use of such a classification model. This study, as a pioneer, tries to implement a linear profit based method to maximize total profit where individual benefits and costs of classifying a transaction are considered during the learning phase. That is, the model which is developed is biased towards the correct classification of beneficial transactions than the others.

This study applied Fisher Linear Discriminant for the first time as a linear discriminant in credit card fraud detection problem. Fisher Linear Discriminant or linear classifier (Christopher, 2006; Fisher, 1936; Fukunaga, 1990; McLachlan, 2004) utilizes dimension reduction method to find the best (D-1)-dimensional hyperplane(s) which can divide a D-dimensional space into two or more subspaces. It is a classic and popular supervised learning method which is commonly used in Face Recognition, Speech/Music Recognition, and Feature Extraction with some modifications (Alexandre-Cortizo, Rosa-Zurera, & Lopez-Ferreras, 2005; Liu & Wechsler, 2002; Witten & Tibshirani, 2011).

The main contributions of this study are introduction of Fisher Discriminant Function for the first time in credit card fraud detection literature and making a simple but effective modification to it to make it an empowered profit-driven classifier in this domain.

The outline of the rest of the paper is as follows: Section 2 reviews related works with detail, Section 3 introduce the methodology of Fisher Discriminant Analysis and improvement carried out in order to make it sensitive to individual profits. Section 4 illustrates the results of implementing the mentioned methods, whereas Section 5 concludes the paper and provides some possible future studies.

2. Related work

Since in this study our problem setup is built as developing a classifier which will help the business users to maximize their profit, here in this section instead of a thorough review of credit card fraud or Fisher Discriminant Analysis publications, we focus on the rather narrow literature on cost sensitive or profit based learning. There is a little number of studies with regard to maximizing total profit (example-dependent) in implementing a classification tool, because as Elkan (2001) mentioned this kind of investigation is in its first steps.

An approach to take cost-sensitivity into account in building up a classifier is to adjust a threshold to make incorrectly classification of instance with higher cost of misclassification harder. In credit card fraud data set, since misclassification cost of fraudulent transactions as legitimate is much higher than misclassification cost of legitimate ones as fraudulent, there should be some modifications in cost matrix to perform better in minimizing total misclassification cost (Sheng & Ling, 2006; Zhou & Liu, 2006; Langford & Beygelzimer, 2005; Maloof, 2003). In real life problems like credit card fraud detection problem misclassification cost of instances may differ based on their classes. So in the mentioned studies, the authors developed a cost matrix showing classification cost of instances from class i as class j as $C(i,j)$. They showed that defining an appropriate cost matrix makes the learning models bias toward the instances with high misclassification cost. Maloof (2003) also indicated that adjusting a cost matrix have as same effect as sampling.

Another way of developing cost sensitive learning method is proposing a new model which is more sensitive to the important instances. Drummond and Holte (2000) developed a new decision tree which applies modified splitting criteria and pruning methods in order to sensitively classify instances with high cost of misclassification. In a similar study, Sahin et al. (2013) proposed a new cost sensitive decision tree which minimizes the misclassification cost while selecting the splitting attribute.

Another method to deal with cost-sensitive problems is using meta-heuristic algorithms with a fitness function taking into account the variable misclassification costs or profits. In a pioneer study, Duman and Ozcelik (2011) combined two well-known meta-heuristic algorithms – Genetic Algorithm (GA) and Scatter Search (SS) – called GASS. The proposed method could improve the performance of classification about 200% in terms of cost. In this study, the authors took the individually variable misclassification costs based on available usable limits.

As a purely relevant study, Duman and Elikucuk (2013a) applied migrating birds optimization (MBO) technique for first time in credit card fraud detection problem with the objective of maximizing total profit obtained by classifying the transactions instead of maximizing classification accuracy. The results show that the MBO algorithm has high performance in classifying most profitable transactions in comparison with the hybrid of Genetic Algorithm and Scatter Search (GASS). The authors on another research (Duman & Elikucuk, 2013b) proposed some modifications on neighborhood sharing function and benefit mechanism by which the total profit obtained could increase up to 94.2%. These results are based on real life data. The authors mentioned MBO as powerful meta-heuristic algorithm in credit card fraud detection problems.

3. Methodology

Below first Fisher Discriminant Analysis (FDA) and then the modification made on it are described.

3.1. Fisher Discriminant Analysis

Linear Discriminant Analysis (LDA) is a kind of supervised learning method by which the input region is divided into decision regions whose boundaries are called decision surfaces or decision boundaries. These decision boundaries are linear function of input vector x which is (D-1)-dimensional hyper-planes in the D-dimensional input vector x. Therefore, linear discriminants also are known as dimension reduction methods by which the learning methods can divide the input region into different convex decision regions. The linear discriminant is a function which takes the input vector x and assign it to one of K classes like C_K using weight vector learned by the training data. A simple way to show the linear discriminant function is

$$y(x) = w^T x + w_0$$

where **w** is weight vector and w_0 is a bias. In the case of 2-class problem, the input vector will assign to class to C_0 if $y(x) > 0$, and to C_1 otherwise. $y(x) = 0$ is corresponds to (D-1)-dimensional hyper-plane as decision boundary. If there was a weight vector by which the samples in both classes can be separated in the perfect way with no misclassification, the dataset is linearly separable.

In the matrix form, let n_0 and n_1 denote the number examples in each class ($n = n_0 + n_1$). For any W, let $z_i = W^T X_i$ which are one-dimensional data that we get after projection. For two class problems, Let M_0 and M_1 be the mean of data in each class which can be calculated as $M_0 = \frac{1}{n_0}\sum_{X_i \in C_0} X_i$ for class 0 and $M_1 = \frac{1}{n_1}\sum_{X_i \in C_1} X_i$ for class 1. Next, mean of projected data can found by $m_0 = W^T M_0$ and $m_1 = W^T M_1$ for class 0 and class 1, respectively.

However, separating two classes by projecting them onto line by which means of the classes join each other works well, there is still a huge amount of overlap in the projected samples as shown in the Fig. 1a. What fisher proposed is minimizing the within-class variance to reduce the overlap while trying to maximize the separation by projecting the samples onto the line joining the means of the classes. The Fisher Criterion (Christopher, 2006) is defined as the ratio of the between-class variance to the within-class variance:

$$J(W) = \frac{(m_0 - m_1)^2}{S_0^2 + S_1^2}$$

where m_0 and m_1 are means of classes while S_0^2 and S_1^2 are variance of the classes. Taking standard deviation of the data in both classes leads to estimate weight vector properly in order to prevent overlap between the projected data in each class (this is shown in the Fig. 1b). To do so, by calculating within-class Scatters (in matrix form), the weight vector onto which the data are projected can split two classes in the best way. Then we might have:

$$\begin{aligned} s_0^2 &= \sum_{X_i \in C_0} (y(x_i) - m_0)^2 = \sum_{X_i \in C_0} \left(W^T X_i - W^T M_0\right)^2 \\ &= \sum_{X_i \in C_0} \left(W^T (X_i - M_0)\right)^2 = \sum_{X_i \in C_0} W^T (X_i - M_0)(X_i - M_0)^T W \Rightarrow s_0^2 \\ &= W^T \left[\sum_{X_i \in C_0} (X_i - M_0)(X_i - M_0)^T\right] W \end{aligned}$$

for class 0, and

$$\begin{aligned} s_1^2 &= \sum_{X_i \in C_1} (y(x_i) - m_1)^2 = \sum_{X_i \in C_0} \left(W^T X_i - W^T M_1\right)^2 \\ &= \sum_{X_i \in C_1} \left(W^T (X_i - M_1)\right)^2 = \sum_{X_i \in C_1} W^T (X_i - M_1)(X_i - M_1)^T W \Rightarrow s_1^2 \\ &= W^T \left[\sum_{X_i \in C_1} (X_i - M_1)(X_i - M_1)^T\right] W \end{aligned}$$

for class 1.

In numerator of the objective function, $(m_0 - m_1)^2$ in the matrix form is:

$$\left(W^T M_0 - W^T M_1\right)^2 = W^T (M_0 - M_1)(M_0 - M_1)^T W.$$

So $(m_0 - m_2)^2 = W^T S_B W$ where $S_B = (M_0 - M_1)(M_0 - M_1)^T$ is $d * d$ matrix and is between-class scatter matrix.

Also for denominator we have $S_0^2 + S_1^2 = W^T S_W W$ where $S_W = \sum_{X_i \in C_0} (X_i - M_0)(X_i - M_0)^T + \sum_{X_i \in C_1} (X_i - M_1)(X_i - M_1)^T$ is $d * d$ matrix which is called within-class matrix.

By replacing $(m_0 - m_1)^2$ with $W^T S_B W$ and $S_0^2 + S_1^2$ with $W_T S_W W$, the objective function which is between-class variance over within-class variance will be:

$$J(W) = \frac{(m_0 - m_1)^2}{S_0^2 + S_1^2} = \frac{W^T S_B W}{W^T S_W W}$$

$J(W) = \frac{W^T S_B W}{W^T S_W W}$ is the objective function in Fisher Discriminant Analysis which maximizing it gives better separated classes with as low overlap as possible within classes.

By differentiating it w.r.t. W and equating to zero we will have:

$$\frac{2 S_B W}{W^T S_W W} - \frac{W^T S_B W}{(W^T S_W W)^2} 2 S_W W = 0$$

which finally gives:

$$S_w^{-1} S_B W - J(W) W = 0$$

solving this generalized problem will give the optimal W weight vector.

Although this criterion is applicable to linearly separable data, one may use it in linearly not separable cases just by accepting some rate of misclassifications in both classes. In the case of linearly not separable case like most of the real-life problems, the weight vector obtained by the equation above will classify the samples with maximum between-class variance and minimum within-class variance.

3.2. Modified Fisher Discriminant Analysis

The main purpose of this study is developing a profit based linear classifier which is more sensitive to more important instances. As (Grela, 2013) shows, weighted average is useful when one

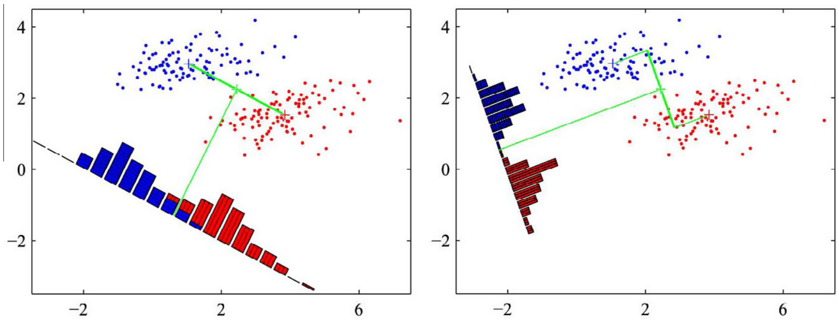

Fig. 1. (a) Considerable overlap between two classes by projecting the samples onto line joining two classes' means. (b) Zero overlap when the within-class variance is taken into account.

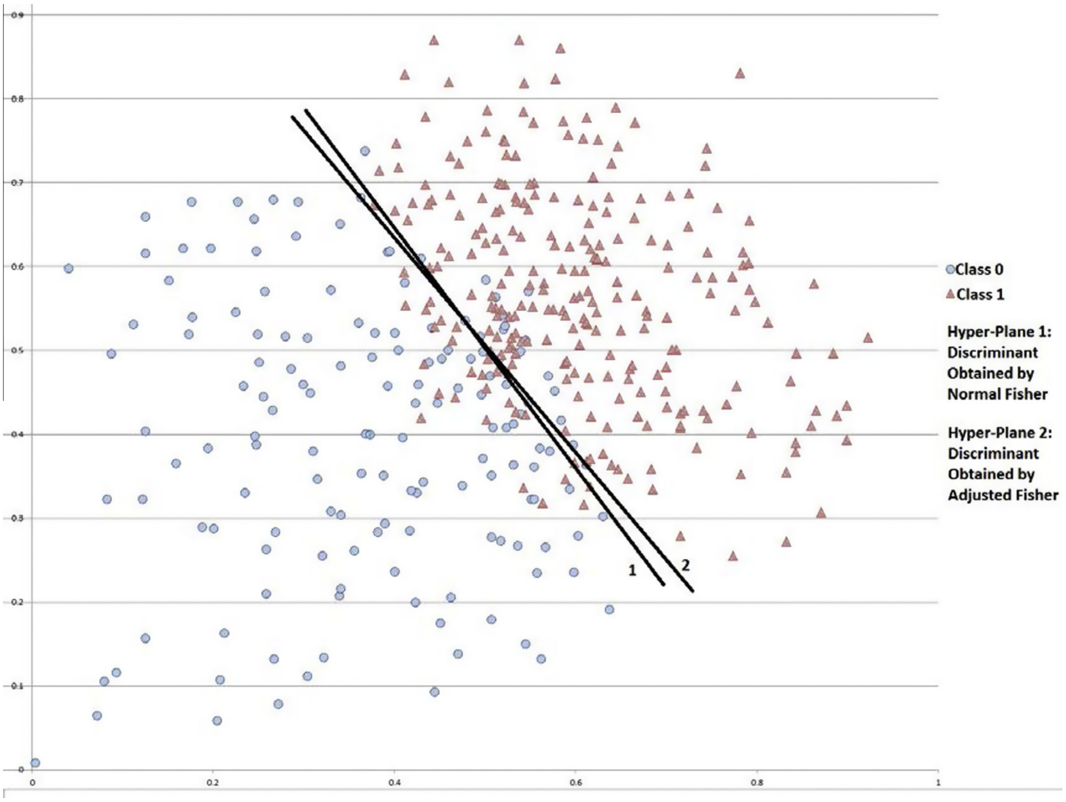

Fig. 2. Change in slope and threshold of the discriminant hyper-plane.

would give importance to foremost instances. However, the weight should be defined carefully with a high sensitivity. Because mentioned study showed that the more similar to each other the weights or values of the attributes are used, the smaller difference between the arithmetic and the weighted mean will be resulted. The modification in the Fisher Criterion which makes it profitable is to apply weighted average for both classes where the weights are defined as total available usable limits on each credit card. This will help us to bias the linear discriminant toward the profitable cases. In other words, when the objective is maximizing correctly classified more important instances; there is a weight vector where a line perpendicular to it separates instances in the optimal way i.e. with more profitable instances. This modification affects both between-class variance and within-class variance so that the mean of classes moves toward the important cases (Fig. 2). Using the Fisher Discriminant Analysis with weighted average will assure that the important instances are classified correctly as much as possible. In this case, each sample has a weight coefficient which shows their importance. Take M'_0 mean for class 0 and M'_1 mean for class 1, which gives the weighted average. By doing so, one can move the mean toward the highly weighted instances. Thus, we will have the means as:

$$M'_0 = \frac{\sum_{x_i \in C_0} PC_{x_i} X_i}{\sum_{x_i \in C_0} PC_{x_i}}$$

for class 0 where PC_{x_i} is weight coefficient of instance x_i, and

$$M'_1 = \frac{\sum_{x_i \in C_1} PC_{x_i} X_i}{\sum_{x_i \in C_1} PC_{x_i}}$$

for class 1.

The most challenging part of this modification is how to define weights for the data. In credit card fraud detection problem, current usable limit on a credit card by which the transaction is made may be good candidate to be the weight of that transaction. However, since this attribute can take very large values, it may cause some instability problems in the calculations. Thus, it should be used in somehow controlled manner. In this regard, we have designed and tested the following five different weight functions to determine the weight of a transaction x_i:

1. $\frac{(\text{Available Usable Limit})_{x_i}}{\text{Average of Available Usable Limit of instances}}$
2. $\left(1 + \frac{(\text{Available Usable Limit})_{x_i}}{\text{Average of Available Usable Limit of instances}}\right)^2$
3. $\left(1 + \frac{(\text{Available Usable Limit})_{x_i}}{\text{Average of Available Usable Limit of instances}}\right)^{\frac{1}{2}}$
4. $\left(\ln\left(e + \frac{(\text{Available Usable Limit})_{x_i}}{\text{Average of Available Usable Limit of instances}}\right)\right)^2$
5. $e^{\left(\frac{(\text{Available Usable Limit})_{x_i}}{\text{Average of Available Usable Limit of instances}}\right)}$

The first function compares the available limit on a card to the average limit on all cards and those cards with higher available limits gets larger weights. Function 2 escalates the ratio considered in Function 1 by taking its square while Function 3 presses it to smaller values by taking the square root. In both functions to start the weights from 1 (that is, to have the minimum weight equal to 1), 1 is added to the ratio. Function 4 is a bit more complicated where we take the square of the ratio and then take its logarithm. Function 5, inflates the ratio by putting it as the power of e. With such different functions we wanted to see whether giving really large or somehow controlled weights to more important cases would result in a better classification model.

Table 1
Saving by top 313 ranked fraudulent predictions of the implemented models and effect of imbalance ratio on the total profit obtained.

		FP (False Positives) Top 313	TP (True Positives) Top 313	Saved Limit Top 313	Saved Limit % Top 313	Total profit Top 313	Total Profit % Top 313	10*FP	100*FP	1000*FP
Decision Tree	1st Fold	85.00	228.00	990419.47	86.63	990262.97	86.61	86.52	85.85	79.16
	2nd Fold	92.00	221.00	1320367.3	91.61	1320210.78	91.60	91.52	90.95	85.20
	3rd Fold	80.00	233.00	1057483.9	88.74	1057327.35	88.73	88.64	88.04	81.99
	Average	**85.67**	**227.33**	**1122756.87**	**88.99**	**1122600.37**	**88.98**	**88.89**	**88.28**	**82.12**
Artificial Neural Network	1st Fold	70.00	243.00	1019186.9	89.14	1019030.37	89.13	89.05	88.50	82.99
	2nd Fold	92.00	221.00	1135966.5	78.82	1135809.97	78.81	78.73	78.15	72.41
	3rd Fold	87.00	226.00	1083842	90.95	1083685.52	90.94	90.85	90.19	83.62
	Average	**83.00**	**230.00**	**1079665.12**	**86.30**	**1079508.62**	**86.29**	**86.21**	**85.61**	**79.67**
Naïve Bayes	1st Fold	102.00	211.00	979931.4	85.71	979774.90	85.70	85.59	84.79	76.76
	2nd Fold	117.00	198.00	1242140.6	86.18	1241983.09	86.17	86.08	85.35	78.04
	3rd Fold	102.00	211.00	1067694.2	89.60	1067537.74	89.58	89.48	88.71	81.01
	Average	**107.00**	**206.67**	**1096588.74**	**87.16**	**1096431.91**	**87.15**	**87.05**	**86.28**	**78.60**
Normal Fisher	1st Fold	70.00	243.00	1019141.2	89.14	1018984.72	89.13	89.04	88.49	82.98
	2nd Fold	79.00	234.00	1213969.7	84.23	1213813.18	84.22	84.15	83.65	78.72
	3rd Fold	71.00	242.00	1049705.4	88.09	1049548.86	88.07	87.99	87.46	82.09
	Average	**73.33***	**239.67***	**1094272.09**	**87.15**	**1094115.59**	**87.14**	**87.06**	**86.54**	**81.27**
Modified Fisher	1st Fold	77.00	236.00	1024499.1	89.61	1024342.62	89.60	89.51	88.90	82.84
	2nd Fold	83.00	230.00	1312708.3	91.08	1312551.79	91.07	91.00	90.48	85.30
	3rd Fold	71.00	242.00	1092816.5	91.70	1092659.96	91.69	91.61	91.07	85.71
	Average	**77.00**	**236.00**	**1143341.29***	**90.80***	**1143184.79***	**90.79***	**90.70***	**90.15***	**84.62***

The model signed by (*) has the best performance in each column.

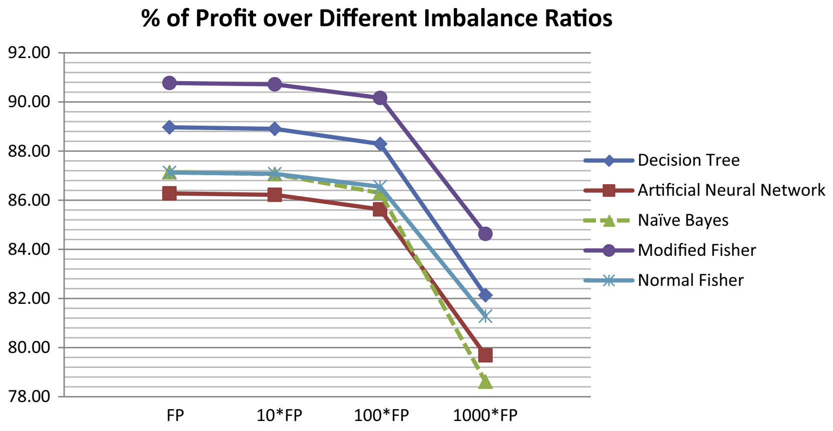

Fig. 3. Changes in % of profit by increasing imbalance ratio.

4. Result and discussion

4.1. Experimental settings

The proposed methods are investigated on a sample of real life dataset taken from an anonymous bank in Turkey where the ratio of fraudulent transactions is about 10 percent. In exact figures, the dataset has 8448 legitimate and 939 fraudulent transactions coupled with 102 attributes. We divided this dataset into 3 different parts (i.e. three different datasets with 2816 legitimate and 313 fraudulent transactions) to apply a three-fold cross validation scheme. The Decision Tree (DT) is used as an attribute selection tool and all algorithms are trained with the attributes that are selected by the DT (the particular DT algorithm we used is Classification & Regression Tree (C&RT) by (Breiman, Friedman, Olshen, & Stone, 1984)).

4.2. Results

In this section, details of experiments on Fisher Discriminant Analysis and its modified version on a real life credit card fraud dataset will be discussed. Also, the model will be evaluated by concentrating on total saving amount as a case-based performance measure. To obtain a comprehensive view of the performance of the proposed methods, three methods are also implemented together with Fisher Discriminant Analysis (FDA) and Modified Fisher Discriminant Analysis (MFDA) on the same datasets which include: Artificial Neural Network (ANN), Decision Tree (DT), and Naïve Bayes (NB).

In this sub-section, first we will discuss the performance of the proposed and implemented methods in terms of confusion matrix entries and classical performance measures. First let's define what performance measure we based our comparisons on:

N = number of legitimate transactions.
P = number of fraudulent transactions.
TP = number of correctly classified fraudulent transactions.
FP = number of false alarms.
Total available limits = summation of available limits in credit cards by which fraudulent transactions (P) are done.
Saved limit[1] = total available limit on fraudulent credit cards which the classifier could detect the transaction (TP).

[1] In real life, our partner bank was provisioning any incoming transaction promptly but then if the card is suspicious they were blocking the card from further transactions. Thus, here for the calculation of available limit and saved limit we are excluding the amount of current transaction.

Table 2
Deviation of the performance of functions in percentage in comparison to function 3.

Functions		FP (false positives) top 313	TP (true positives) top 313	Saved limit top 313	Total profit top 313
Function 1	1st fold	−2.469136	0.862069	−2.87207	−2.87248
	2nd fold	1.4285714	−0.411523	−1.78137	−1.78161
	3rd fold	2.5	−0.858369	−3.78886	−3.78935
	Average	**0.4864785**	**−0.135941**	**−2.8141**	**−2.81448**
Function 2	1st fold	−1.234568	0.431034	−2.93234	−2.93276
	2nd fold	0	0	−0.58925	−0.58933
	3rd fold	7.5	−2.575107	−3.42064	−3.42108
	Average	**2.0884774**	**−0.714691**	**−2.31407**	**−2.31439**
Function 4	1st fold	−1.234568	0.431034	−2.87207	−2.87248
	2nd fold	2.8571429	−0.823045	−0.92441	−0.92453
	3rd fold	2.5	−0.858369	−4.20509	−4.20564
	Average	**1.3741917**	**−0.416793**	**−2.66719**	**−2.66755**
Function 5	1st fold	1.2345679	−0.431034	−6.31097	−6.31188
	2nd fold	5.7142857	−1.646091	−2.08139	−2.08166
	3rd fold	7.5	−2.575107	−3.51328	−3.51374
	Average	***4.8162845***	***−1.550744***	***−3.96855***	***−3.96909***

Total profit = saved limit minus total alert cost for false positives and true positives.

In general terms, classifiers produce scores of being fraudulent for each of the transactions. Then typically, the transactions having a score of 0.5 or greater are predicted as fraud (Positive Class) and confusion matrices are formed using these class assignments. However, as such, the number of positive predictions can be different for each classifier and therefore, confusion (accuracy) matrices based performance measures cannot be compared directly. The Area under ROC Curve (AUC) performance measure (Bradley, 1997) can provide a more robust comparison of classifiers independent of any particular threshold chosen. In our case, we preferred to fix that particular threshold to the top 10% of transactions of test set, since we already know that exactly 10% of our transactions in sample are fraud. By such a fixing, only the best (perfect) classifier will be having 100% true positive rate and the other classifiers' performance can be compared with this perfect classifier. Top 10% corresponds labeling 313 transactions as fraudulent and the remaining 2816 as legitimate. By doing so, the number of false negatives (FN) is equal to number of false positives (FP).

Among the alternative weight functions to be used in modified FDA (MFDA), the one given by Function 3 performed the best. In order to not lose the focus of the paper, we preferred to tabulate its results. First table shows number of FP and TP in top 313 most likely fraudulent transactions. The saved limit and the profit obtained by different algorithms are also displayed. We see that all four algorithms (DT, ANN, NB and FDA) have produced comparative results in terms of the classical measure TP. The performance of FDA as compared to ANN which is a more complex algorithm is really good and this shows that data mining society can pay more attention to it in future.

For the comparison of FDA and modified FDA, as can be seen, while FDA captures more number of positive cases, modified FDA leads more profit by truly classifying the most profitable transactions. As shown in Table 1, total profit obtained by modified FDA is higher than the other algorithms.

At this point, another discussion is worth to be made. If the cut-off point is determined as the fraud score generated for the top 313rd fraudulent transaction we know that 10% of the transactions will be labeled as fraud (because 10% of our training set is fraud). However, when we apply this model in real life to the full set of transactions where the ratio of frauds is about 0.001%, we cannot know exactly how many transactions will have a score above this cut-off. However, we can expect that this number might be quite above 313. Thus, the number of false alerts (FP) can be larger in which case the net profit from our model (savings minus inspection cost of false alerts) can be less. To cope for this, in the last 3 columns of Table 1 we included, what profit we could expect if the number of false alerts will be increased to 10 fold, 100 fold, and 1000 fold, respectively.

Fig. 3 shows the behavior depicted by the last four columns of Table 1. Although, increase in the number of false alerts causes decrease in net profit obtained from a classifier, a change in the relative ordering of different classifiers is not expected. Also, even if the number of alerts can be 1000 times that is observed on the training set, still quite high levels of profit can be attained (about 80 percent).

As an additional note, we want to say a few words on relative performances of the other weight functions as compared to Function 3 whose results are tabulated in Table 1. Table 2 illustrates briefly how the performances of the other weight functions deviate from Function 3. The underlined values in average rows show the best and the *italic* ones show the worst performances when Function 3 is excluded. As we can see, Function 5 has the worst performance both in terms of the TP and the total profit. This can be attributed to its inflating character of the ratio or giving much higher weights to more important instances. We also see that, while Function 1 gives the best result in terms of TP, Function 2 gives the best result in terms of the more important measure total profit. Function 2 is an inflated version of Function 1 from which we can conclude that a little bit inflation of more important cases is useful. However, if we increase the inflation level more, the performance can go worse as in the case of Function 5.

5. Summary and conclusions

In this study, Linear Fisher Discriminant Analysis is implemented for the first time in credit card fraud detection problem. This method tries to find the best (D-1)-dimensional hyperplane by which the within-class variance is minimized to reduce the overlap and between-class variance is maximized to increase the separation. As the main contribution of this study, weighted average is used in calculating these variances which makes the learning sensitive towards the profitable transactions. Linear discriminant is one of the simplest classifiers in terms of complexity whereby the modification proposed in this paper could perform as good as other complex methods in literature.

To sum up as the results obtained show, Fisher Linear Discriminant found to perform better in credit card fraud detection problem based on not only classical performance measures but also saved total available limit. Although, Modified Fisher Discriminant Analysis could capture a bit less number of fraudulent transactions, it has the best performance in terms of the objective of this study which is maximizing total profit.

The modification implemented in this study attempts to consider individually defined weights of transactions for which the linear classifier tries to assign accurate labels to transactions with higher importance. Thus, this method can label transactions with high usable limits on the cards correctly which leads to prevent losing millions of dollars in real life banking systems arisen from fraudulent activities and customers' churn.

For further studies in future, the authors propose to implement other linear discriminant functions like Linear Perceptron Discriminant Function to develop an iterative linear discriminant. Furthermore, as a future study, misclassification cost of negatives could be considered variable as because, sending too many alarms may lead dissatisfaction among customers that may result in variable churn costs.

Acknowledgment

This study was funded by The Scientific and Technological Research Council of Turkey (TÜBİTAK) under Project No. 113M063.

References

Aihua, S., Rencheng, T., & Yaochen, D. (2007). Application of classification models on credit card fraud detection. In *Proceedings – ICSSSM'07: 2007 international conference on service systems and service management*.

Alexandre-Cortizo, E., Rosa-Zurera, M., & Lopez-Ferreras, F. (2005). Application of fisher linear discriminant analysis to speech/music classification. In *EUROCON 2005. The international conference on computer as a tool* (Vol. 2, pp. 1666–1669).

Bolton, R. J., & Hand, D. J. (2001). Unsupervised profiling methods for fraud detection. In *Proceedings of credit scoring and credit control VII* (pp. 5–7).

Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. *Statistical Science, 17*, 235–255.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition, 30*, 1145–1159.

Brause, R., Langsdorf, T., & Hepp, M. (1999). Neural data mining for credit card fraud detection. In *Proceedings 11th international conference on tools with artificial intelligence*.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. *The Wadsworth Statistics Probability Series, 19*, 368.

Chan, P. K., Fan, W., Prodromidis, A. L., & Stolfo, S. J. (1999). Distributed data mining in credit card fraud detection. *IEEE Intelligent Systems*.

Chen, R. C., Chen, T. S., & Lin, C. C. (2006). A new binary support vector system for increasing detection rate of credit card fraud. *International Journal of Pattern Recognition and Artificial Intelligence, 20*, 227–239.

Chen, R. C., Chiu, M. L., Huang, Y. L., & Chen, L. T. (2004). Detecting credit card fraud by using questionnaire-responded transaction model based on support vector machines. In *Intelligent data engineering and automated learning-IDEAL* (pp. 800–806). Springer.

Chen, R. C., Luo, S. T., Liang, X., & Lee, V. (2005). Personalized approach based on SVM and ANN for detecting credit card fraud. In *ICNN&B'05. International conference on neural networks and brain* (Vol. 2, pp. 810–815).

Christopher, M. B. (2006). *Pattern recognition and machine learning*. Cambridge: Springer.

Drummond, C., & Holte, R. C. (2000). Exploiting the cost (in) sensitivity of decision tree splitting criteria. In *ICML* (pp. 239–246).

Duman, E., & Elikucuk, I. (2013a). Applying migrating birds optimization to credit card fraud detection. In *Trends and applications in knowledge discovery and data mining* (pp. 416–427). Springer.

Duman, E., & Elikucuk, I. (2013b). Solving credit card fraud detection problem by the new metaheuristics migrating birds optimization. *Advances in Computational Intelligence*, 62–71.

Duman, E., & Ozcelik, M. H. (2011). Detecting credit card fraud by genetic algorithm and scatter search. *Expert Systems with Applications, 38*(10), 13057–13063.

Elkan, C. (2001). The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence* (Vol. 17, pp. 973–978).

Fisher, R. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics, 7*, 179–188.

Fukunaga, K. (1990). Introduction to statistical pattern recognition. *Pattern Recognition, 22*, 833–834.

Grela, G. (2013). Does weighted average really work? 1219–1227. Retrieved from <http://www.toknowpress.net/ISBN/978-961-6914-02-4/papers/ML13-391.pdf>.

Juszczak, P., Adams, N. M., Hand, D. J., Whitrow, C., & Weston, D. J. (2008). Off-the-peg and bespoke classifiers for fraud detection. *Computational Statistics & Data Analysis, 52*(9), 4521–4532.

Langford, J., & Beygelzimer, A. (2005). Sensitive error correcting output codes. In *Learning theory* (pp. 158–172). Springer.

Leonard, K. J. (1993). Detecting credit card fraud using expert systems. *Computers & Industrial Engineering, 25*(1), 103–106.

Liu, C., & Wechsler, H. (2002). Gabor feature based classification using the enhanced Fisher linear discriminant model for face recognition. *IEEE Transactions on Image Processing, 11*, 467–476.

Maloof, M. A. (2003). Learning when data sets are imbalanced and when costs are unequal and unknown. *Analysis, 21*, 1263–1284.

McAlearney, S., & Breach, T. J. X. D. (2008). Ignore cost lessons and weep. CIO, August 7.

McLachlan, G. J. (2004). *Discriminant analysis and statistical pattern recognition. Wiley series in probability and statistics*. New Jersey: John Wiley & Sons.

Mena, J. (2003). *Investigative data mining for security and criminal detection*. Butterworth-Heinemann.

Provost, F. (2002). [Statistical fraud detection: A review]: Comment. *Statistical Science*, 249–251.

Quah, J. T. S., & Sriganesh, M. (2007). Real time credit card fraud detection using computational intelligence. In *2007 International joint conference on neural networks* (pp. 863–868).

Sahin, Y., & Duman, E. (2010). An overview of business domains where fraud can take place, and a survey of various fraud detection techniques. In *Proceedings of the 1st international symposium on computing in science and engineering*, Aydin, Turkey.

Sahin, Y., & Duman, E. (2011). Detecting credit card fraud by ANN and logistic regression. In *2011 International symposium on innovations in intelligent systems and applications (INISTA)* (pp. 315–319).

Sahin, Y., Bulkan, S., & Duman, E. (2013). A cost-sensitive decision tree approach for fraud detection. *Expert Systems with Applications, 40*(15), 5916–5923.

Schindeler, S. (2006). Fighting card fraud in the USA. *Credit Control, 27*(2), 50.

Sheng, V. S., & Ling, C. X. (2006). Thresholding for making classifiers cost-sensitive. In *Proceedings of the national conference on artificial intelligence* (Vol. 21, pp. 476).

Wheeler, R., & Aitken, S. (2000). Multiple algorithms for fraud detection. *Knowledge-Based Systems, 13*, 93–99.

Witten, D. M., & Tibshirani, R. (2011). Penalized classification using Fisher's linear discriminant. *Journal of the Royal Statistical Society. Series B: Statistical Methodology, 73*, 753–772.

Zhou, Z. H., & Liu, X. Y. (2006). Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on Knowledge and Data Engineering, 18*, 63–77.