Cost Sensitive Credit Card Fraud Detection using Bayes Minimum Risk

Alejandro Correa Bahnsen, Aleksandar Stojanovic, Djamila Aouada and Björn Ottersten
Interdisciplinary Centre for Security, Reliability and Trust
University of Luxembourg, Luxembourg
http://www.uni.lu/snt/
Email: {alejandro.correa, aleksandar.stojanovic, djamila.aouada, bjorn.ottersten}@uni.lu

***Abstract*—Credit card fraud is a growing problem that affects card holders around the world. Fraud detection has been an interesting topic in machine learning. Nevertheless, current state of the art credit card fraud detection algorithms miss to include the real costs of credit card fraud as a measure to evaluate algorithms. In this paper a new comparison measure that realistically represents the monetary gains and losses due to fraud detection is proposed. Moreover, using the proposed cost measure a cost sensitive method based on Bayes minimum risk is presented. This method is compared with state of the art algorithms and shows improvements up to 23% measured by cost. The results of this paper are based on real life transactional data provided by a large European card processing company.**

***Keywords*-Credit card fraud detection; Bayesian decision theory; Cost sensitive classification**

I. Introduction

The use of credit and debit cards has increased significantly in the last years, unfortunately so has the fraud committed with them. According to the European Central Bank [1], during 2010 the total level of fraud reached 1.26 billion in the Single Euro Payments Area. Currently, financial institutions deal with fraud detection with a series of if-then rules created by internal risk teams. If the result of the rule is that a possible fraud is suspected, depending on the rule, the transaction can be denied or an alert is emitted for further investigation. The rules perform well as long as there are no new fraud patterns, as repeated frauds are required for the team to detect new patterns. There is, therefore, a clear need for a better approach to the credit card fraud detection problem.

The use of machine learning in fraud detection has been an interesting topic in recent years. However, due to the confidentiality of financial information and non availability of public databases, few researches have had the opportunity to work on developing methods specific to credit card fraud detection [2]. Nevertheless, the literature on credit card fraud detection is growing and it has been shown that machine learning can be used successfully for this problem, in particular: neural networks [3], artificial immune systems [4], association rules [5], Bayesian learning [3], support vector machines [6], and peer group analysis [7].

The databases used in developing credit card fraud detection systems have a very low ratio of fraudulent transactions, ranging from 0.005% to 0.5%, see [4] and [6]. This generates complications during the training of the different algorithms [8]. Because of this, a common practice in the research community is to carry out an under-sampling procedure [9], consisting in creating a sample of the database with a higher percentage of fraudulent cases.

Most of these studies compare they proposed algorithm with a benchmark logistic regression, and all of them make the comparison using a classical evaluation measure such as misclassification, precision and recall [2]. The particularity of credit card fraud is that wrongly predicting a fraudulent transaction as legitimate carries a significantly different cost than the inverse case. In [10], a method that differentiates between these costs was proposed, but it assumes a constant difference between them, which is a typical assumption is cost sensitive classification [11]. By contrast, we propose an evaluation measure that realistically represents the monetary gains and losses due to fraud and its detection. Moreover, we present a Bayes minimum risk classifier including the real financial costs of credit card fraud detection in order to have a cost sensitive detection system. The proposed cost sensitive method decreased significantly the cost due to fraud as compared with state of the art techniques.

Using a real transactional database with fraudulent and legitimate transactions from a large European card processing company, we compare standard algorithms, using both classical measures and the proposed financial measure. Afterwards, because of the poor performance of the state of the art techniques, a cost sensitive system is developed in order to integrate the real financial costs due to credit card fraud. We first use a thresholding optimization technique and finally a Bayes minimum risk classifier.

The remainder of the paper is organized as follows. In Section II, we present our proposed cost sensitive evaluation measure. Afterwards, we explain the prior work in Section III. In Section IV, we present the cost sensitive credit card fraud detection using Bayes minimum risk. Section V describes the data we use for experiments. Then the results are presented in Section VI. Finally, the conclusions of the paper are given in Section VII.

978-0-7695-5144-9/13 $31.00 © 2013 IEEE
DOI 10.1109/ICMLA.2013.68

Table I
CONFUSION MATRIX OF A BINARY CLASSIFICATION SYSTEM

		True Class (y_i)	
		Fraud	Legitimate
Predicted	Fraud	TP	FP
Class (p)	Legitimate	FN	TN

Table II
COST MATRIX USING FIXED FN COSTS PROPOSED IN [10]

		True Class (y_i)	
		Fraud	Legitimate
Predicted	Fraud	C_a	C_a
Class (p)	Legitimate	$100 \cdot C_a$	0

Table III
COST MATRIX USING REAL FINANCIAL COSTS

		True Class (y_i)	
		Fraud	Legitimate
Predicted	Fraud	C_a	C_a
Class (p_i)	Legitimate	Amt_i	0

II. COST SENSITIVE CREDIT CARD FRAUD DETECTION EVALUATION MEASURE

Once a credit card fraud detection system is developed, it is very important to be able to evaluate and compare it to other state of the art fraud detection systems. In Table I, the classical confusion matrix of a credit card fraud detection system is shown. This matrix is typically used to evaluate binary classification algorithms. The following traditional statistics are extracted from it:

- Misclassification $= 1 - \frac{TP+TN}{TP+TN+FP+FN}$
- Recall $= \frac{TP}{TP+FN}$
- Precision $= \frac{TP}{TP+FP}$
- F_1-Score $= 2\frac{\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}}$,

where TP and FN are the numbers of true positives and false negatives, respectively. We define as positive the case when a fraud is committed, and negative otherwise. These statistics are typically used to evaluate credit card fraud detection systems [2], [10]. However, they assume that FP carries the same cost as FN, but as will be shown later this is not the case in credit card fraud detection. Hand et al. [10] proposed a cost matrix [11] which differentiates the costs of FP and FN, as given in Table II, where in the case of FP the associated cost is the administrative cost C_a related to analysing the transaction and contacting the card holder. This cost is the same assigned to a TP because, in this case, the card holder will have to be contacted. However, in the case of an FN in which a fraud is not detected, the cost is defined to be a hundred times C_a.

Nevertheless, in practice, losses due to a specific fraud range from few to thousands of Euros, which means that assuming constant cost due to an FN is unrealistic. In order to address this limitation, we introduce a new cost matrix in Table III. We defined the cost of an FN to be the amount Amt_i of the transaction i. This cost matrix is a better representation of the actual costs, since when a fraud is not detected, the losses of that particular fraud correspond to the stolen amount.

We define below the cost measure using the cost matrix with real financial costs proposed in Table III:

$$C = \sum_{i=1}^{m} y_i \left(p_i C_a + (1-p_i)Amt_i\right) + (1-y_i)p_i C_a. \quad (1)$$

This measure evaluates the sum of the cost for m transactions, where y_i and p_i are the real and predicted labels, respectively. Moreover, this cost matrix is not only used for evaluation but it is also used to develop a cost sensitive classification algorithm using Bayes minimum risk as described in the following section.

III. BAYES MINIMUM RISK

We use Bayes minimum risk as a method for cost sensitive credit card fraud detection. As defined in [12], the Bayes minimum risk classifier is a decision model based on quantifying tradeoffs between various decisions using probabilities and the costs that accompany such decisions. In the case of credit card fraud detection, there are two decisions, either predict a transaction as fraud p_f or as legitimate p_l. The risk associated with predicting a transaction as fraud is defined as

$$R(p_f|x) = L(p_f|y_f)P(p_f|x) + L(p_f|y_l)P(p_l|x), \quad (2)$$

and when the transaction is predicted as legitimate it is

$$R(p_l|x) = L(p_l|y_l)P(p_l|x) + L(p_l|y_f)P(p_f|x), \quad (3)$$

where y_f and y_l are the real labels for fraudulent and legitimate transactions respectively. $P(p_l|x)$ is the estimated probability of a transaction being legitimate given x, similarly $P(p_f|x)$ is the probability of a transaction being fraud given x. Finally $L(a,b)$ is the loss function when a transaction is predicted as a and the real label is b. Once both risks are calculated, a transaction is classified as fraud if $R(p_f|x) \leq R(p_l|x)$, meaning if the risk associated with that decision is lower than the risk associated with classifying it as legitimate.

Since in the credit card fraud detection case the losses are equal to the cost, first we use the cost matrix with fixed cost for FN as defined in Table II. Then a transaction will be classified as fraud if:

$$C_a P(p_f|x) + C_a P(p_l|x) \leq 100 \cdot C_a P(p_f|x), \quad (4)$$

and as legitimate otherwise.

Lastly, we test while using the proposed cost matrix with real financial costs as in Table III. A transaction will be classified as fraud if the following condition is true:

$$C_a P(p_f|x) + C_a P(p_l|x) \leq Amt_i P(p_f|x), \quad (5)$$

and as legitimate if false.

Table IV
DATABASE ATTRIBUTES

Attribute name	Description
Date	Date and hour of the transaction
Account number	Identification number of the account
Card number	Identification number of the card
Transaction type	Type of transaction (Internet, Card present, ATM)
Amount	Amount of transaction in Euros
Merchant ID	Identification of the merchant
Merchant group	Merchant group identification provided by the card processing company
Country	Country where the transaction took place
Country 2	Country of residence of the card holder
Type of card	Card brand (Visa debit, Visa Classic, Mastercard Gold, ...)
Gender	Gender of the card holder
Age	Card holder age
Bank	Issuer bank of the card
Fraud	Whenever the transaction was or not fraud

IV. DATA

We use a database provided by a large European card processing company. The database consists of fraudulent and legitimate transactions made with credit and debit cards during 2012. The total database contains 80,000,000 individual transactions, each one with 27 attributes. From the original attributes we manually select those that contain useful information with help from the card processing company risk team. Table IV shows the selected attributes.

Using the initial attributes we derived additional 260 attributes using the methodology proposed in [6] and [13]. The idea behind the derived attributes consists in using a transaction aggregation strategy in order to capture consumer spending behavior in the recent past. The derivation of the attributes consists in grouping the transactions made during the last given number of hours, first by card or account number, then by transaction type, merchant group, country or other, followed by calculating the number of transactions or the total amount spent on those transactions. An example of a derived attribute is: number of transactions made during the last 6 hours on the internet by the same individual in the same country.

The database also includes a fraud label indicating whenever a transaction is identified as fraud. This label was created internally in the card processing company, either because the internal risk team detected a fraud or because a client reported a fraudulent transaction on his/her card statement and after an internal investigation the fraud is confirmed. Small amounts that may have slipped the attention of inattentive card holders notwithstanding, this database can be regarded as being extremely accurate. In the database only 20,000 transactions were labelled as fraud, leading to a fraud ratio of 0.025%.

For our experiments, we select a smaller subset of transactions with a higher fraud ratio, corresponding to a specific group of transactions. This database contains 750,000 transactions and a fraud ratio of 0.467%. In this database, the total financial losses due to fraud are 866,410 Euros. We select this database because it is the one where most frauds are being made. Still, in order to capture consumer patterns, we used the full database to calculate the derived attributes. From this database we used the first 10 months (January to October 2012) for training and the last 2 months (November and December 2012) for testing. The motivation behind this approach is the need to train the system in the same way it will be implemented, in which past months are used to predict the current month [6], [13].

Table V
DESCRIPTION OF DATABASES: ONE TRAINING DATABASE AND DIFFERENT UNDER-SAMPLED DATABASES VARYING IN PERCENTAGE OF FRAUDS.

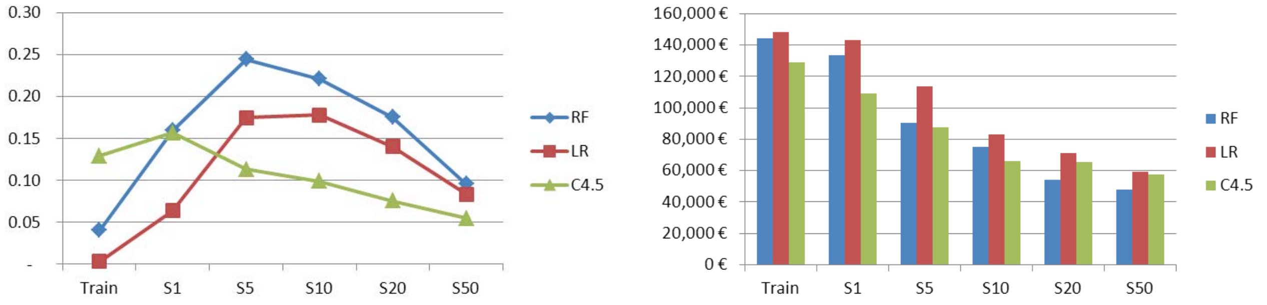
Database	Transactions	Frauds	Fraud Ratio	Fraud Amount
Total	750,000	3,500	0.467%	866,410
Train	625,000	2,900	0.464%	721,349
Test	125,000	600	0.480%	148,562
S1	290,000	2,900	1%	721,349
S5	58,000	2,900	5%	721,349
S10	29,000	2,900	10%	721,349
S20	17,500	2,900	20%	721,349
S50	5,800	2,900	50%	721,349

V. RESULTS

First we test algorithms that have previously been used to solve the credit card fraud detection problem [14], namely, logistic regression (LR), C4.5 and random forest (RF), see [8]. The implementation of these algorithms in Scikit-learn was used [15]. For LR, an ℓ_2 norm regularization was selected. The C4.5 algorithm was trained using the default parameters of the Scikit-learn package. For the RF algorithm, the maximum number of estimators in each split was set to 10 and the Gini criterion for measuring the quality of a split was selected.

Because these algorithms suffer when the label distribution is skewed towards one of the classes [8], we make an under-sampling of the legitimate transactions in order to have a more balanced class distribution. The under-sampling has proved to be a better approach on such problems, see [9]. We create 5 different databases S1, S5, S10, S20 and S50, each one having a different percentage of frauds 1%, 5%, 10%, 20% and 50%, respectively. The motivation to create the different databases is to evaluate how the algorithms perform on different class distributions. Table V summarizes the different databases. It is important to note that the under-sampling procedure was only applied to the training dataset since the test database must reflect the real fraud distribution.

In addition to the traditional aforementioned algorithms we also evaluate a thresholding optimization to make the classifiers cost sensitive, based on the method proposed in [16]. The idea behind this approach is to adaptively modify the probability threshold of an algorithm such that a certain criterion is minimized; in our case the cost due to fraud. By default the probability threshold of an algorithm is 50%,



(a) F_1-Score

(b) Cost

Figure 1. Results using LR, C4.5 and RF algorithms on different under-sampled databases, with $C_a = 2,50$ Euros. RF outperforms the other algorithms measured by cost and by F_1-Score. The best results in terms of cost are found when a higher percentage of frauds is used for training. Nevertheless, the best model using as a reference the F_1-Score is when an under-sampled database with a 5% fraud rate is used, leading to the conclusion that when selecting the algorithms by traditional statistics results are different than when a realistic financial measure is used.

meaning that when the probability of a positive event is greater than 50% that example is classified as positive. This default threshold is not necessarily the one that minimizes the cost due to fraud. So we make an optimization in the training dataset, in order to find the threshold which minimizes the cost measure. Then, this new threshold is applied to the test dataset to obtain the results, and by doing so, we make the algorithm cost sensitive by threshold optimization.

Afterwards, we apply the Bayes minimum risk classifier proposed in Section III, using the cost matrix with fixed FN cost described in Table II and the cost matrix with real financial costs proposed in Table III.

Subsequently, we adjust the estimated probabilities since when applying the under-sampling methodology the estimated probabilities of fraud are overestimated. This may lead to methods that rely on true probabilities to have inconsistencies, which is the case of the Bayes minimum risk classifier [12]. The reason this happens, is because the prior probability of fraud is artificially increased by the under-sampling. In order to solve this we adjust the estimated probabilities with respect to the difference of the fraud distribution in the under-sampled and the training datasets.

For the different algorithms using the test dataset we evaluate the statistics defined in Section II. We also evaluate the cost due to fraud as defined in (1), assuming the C_a parameter is equal to 2.50 Euros. Finally we test the sensibility of the results with respect to the C_a parameter.

A. Traditional algorithms

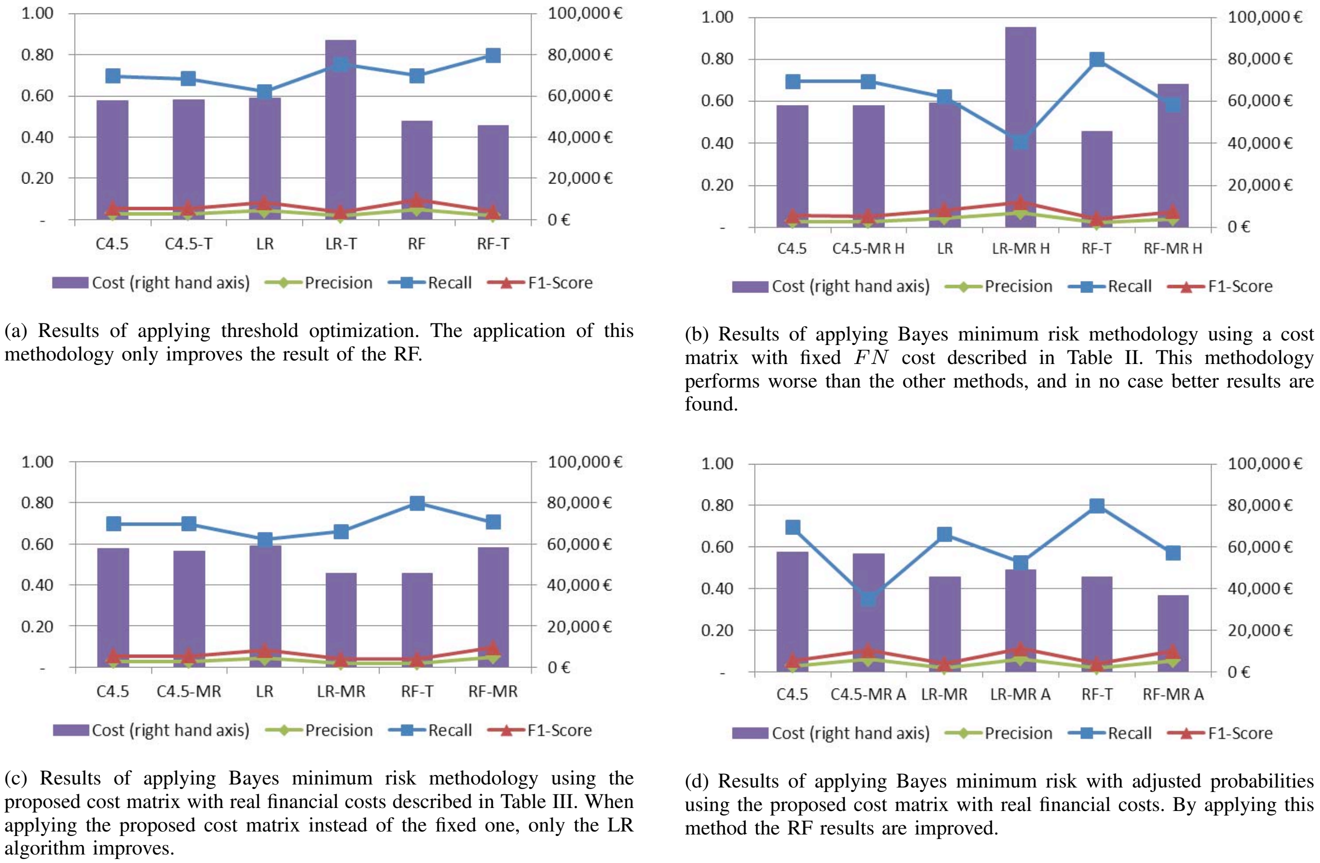
We evaluate the LR, C4.5 and RF algorithms, on the full database and on the different under-sampled databases. Results are shown in Figure 1. It is clear that when applying under-sampling the best results are found when there is a balanced distribution of frauds and not frauds on the training database. In all cases the tree based models outperform the LR. Additionally, when comparing the results based on the traditional comparison measures, the best models measured by F_1-Score are with the database S5 meaning when the frauds are 5% of the training database. The model selected with the traditional F_1-Score performs poorly in terms of cost. There are no significant savings as compared to using no model at all, which corresponds to a cost of 148,562 Euros, equal to the total amount lost due to fraud in the test database. This is why, for the following experiments, we select the models trained in the S50 database, where there are savings in money of up to 76%, despite the low F_1-Score.

B. Thresholding optimization

We apply the threshold optimization technique to the algorithms trained with the S50 database, since the best results using the traditional algorithms and measured by cost, are found on that database. As described before, this methodology attempts to make a classifier cost sensitive by changing the probability threshold. Figure 2a presents the results of applying this technique to LR (LR-T), C4.5 (C4.5-T) and RF (RF-T) on the test dataset. Interestingly, when applying this methodology not all models are improved. The LR-T actually performs the worst, but the RF-T performs much better, and even though in both cases there is an increase in the recall, meaning the number of frauds detected by the algorithms, the precision decreases in both cases. With the C4.5-T algorithm, there is no change in the results when applying the threshold optimization. On the other hand, when the threshold optimization procedure is applied to RF the cost is reduced by 2,165 Euros while the F_1-Score remains the same.

C. Bayes minimum risk classifier

We compare the previous algorithms when applying the Bayes minimum risk classifier. First we evaluate this algorithm using the cost matrix with fixed FN cost described in Table II. As can be seen in Figure 2b, applying these



(a) Results of applying threshold optimization. The application of this methodology only improves the result of the RF.

(b) Results of applying Bayes minimum risk methodology using a cost matrix with fixed FN cost described in Table II. This methodology performs worse than the other methods, and in no case better results are found.

(c) Results of applying Bayes minimum risk methodology using the proposed cost matrix with real financial costs described in Table III. When applying the proposed cost matrix instead of the fixed one, only the LR algorithm improves.

(d) Results of applying Bayes minimum risk with adjusted probabilities using the proposed cost matrix with real financial costs. By applying this method the RF results are improved.

Figure 2. Results of applying Bayes minimum risk methodology using different cost matrices. Using a cost matrix with the real financial costs due to fraud gives the best results both in terms of cost and F_1-Score. The best overall result is found when the methodology Bayes minimum risk with adjusted probabilities using the proposed cost matrix with real financial costs is applied to RF.

algorithms does not give better results. In fact in all cases the model performs worse. The results clearly show that it is unrealistic to assume a constant false negative cost of a hundred times the administrative cost, which motivates the application of Bayes minimum risk using the cost matrix proposed in Table III. In Figure 2c, the results are shown. This methodology when applied to LR (LR-MR) performs very well, increasing savings by 13,685 Euros compared with LR. But then it becomes interesting that the RF algorithm, which has already been improved by using the thresholding methodology, performs very badly with the Bayes minimum risk. This may be because the probability estimates were trained using the under-sampled databases, so the algorithms overestimate the true probabilities of fraud.

D. Bayes minimum risk classifier with adjusted probabilities

We adjust the estimated probabilities as described before. The results of the Bayes minimum risk with adjusted probabilities are shown in Figure 2d. Now it can be seen that the RF-MR A algorithm performs very good, being the overall best model in terms of cost. It is even more interesting that this model saves 8,870 more Euros than the RF-T, but the Recall is down by almost 25%, meaning that this model is detecting the most relevant frauds, that is, fraud with high amounts.

E. Sensibility of C_a

Finally, in order to check that the results were not biased by the selection of the C_a parameter, we compare the results of all the algorithms by varying the parameter from 1.00 to 5.00 Euros. In Table VI the cost of the algorithm when the administrative cost is 1.00, 2.50 and 5.00 Euros are shown. It can be seen that in all cases, the best results are found using RF-MR A.

F. Summary of results

Traditional algorithms applied to detect credit card fraud perform well only when an under-sampling methodology is applied. When applying the thresholding optimization methodology, the RF-T algorithm improves the result in terms of cost without sacrificing the F_1-Score. However, in the case of the LR-T, the new model performs very badly. It

Table VI
COST RESULTS VARYING THE ADMINISTRATIVE COST IN EUROS

Algorithm	C_a=1,00	C_a=2,50	C_a=5,00
C4.5	35,466	57,726	86,215
LR	46,530	59,157	80,202
RF	33,641	47,669	61,969
C4.5-T	35,531	57,888	86,215
LR-T	56,704	87,127	94,977
RF-T	26,598	45,504	66,374
C4.5-MR H	35,531	57,888	79,879
LR-MR H	56,357	95,190	104,027
RF-MR H	37,964	67,977	90,092
C4.5-MR	35,120	56,176	83,117
LR-MR	28,320	45,472	62,570
RF-MR	32,380	58,165	71,694
C4.5-MR A	31,631	56,551	70,127
LR-MR A	27,212	48,915	69,185
RF-MR A	**20,929**	**36,634**	**52,003**

is expected that on the training dataset this methodology performs at least as well as not using it, but seeing the different results found on the test database with the different algorithms leads to the conclusion that this method is overfitting the training data.

More importantly, it turns out that real financial costs need to be used when applying Bayes minimum risk. Using a cost matrix with fixed FN cost as proposed in [10], gives poor results. This is because in practice the cost of for different FN varies significantly. When using our proposed cost matrix with the real financial cost, very good results are found in the case of LR-MR. In this case the cost is reduced by 13,685 Euros. Furthermore, when adjusting the estimated probabilities before applying Bayes minimum risk to RF, the best results in terms of cost are found. In this case the best overall model is found with a cost of 36,634 Euros, meaning savings of 23% as compared to RF.

The above result is verified for different amounts of C_a, that is, when varying C_a between 1.00 and 5.00 Euros. The proposed RF-MR A is consistently the best method.

VI. CONCLUSION

In this paper we have shown the importance of using the real financial costs of credit card fraud when selecting credit card fraud detection algorithms. Also, it is not enough to have a fixed difference between FP and FN but it is important to have the real FN cost of each transaction. Moreover, our evaluations confirmed that including the real cost by creating a cost sensitive system using a Bayes minimum risk classifier, gives rise to much better fraud detection results in the sense of higher savings.

REFERENCES

[1] European Central Bank, "Report on card fraud July 2012," Tech. Rep., 2012.

[2] R. J. Bolton, D. Hand, F. Provost, and L. Breiman, "Statistical Fraud Detection: A Review," *Statistical Science*, vol. 17, no. 3, pp. 235–255, 2002.

[3] S. Maes, K. Tuyls, B. Vanschoenwinkel, and B. Manderick, "Credit card fraud detection using Bayesian and neural networks," in *Proceedings of NF2002*, 2002.

[4] M. Gadi, X. Wang, and A. do Lago, "Credit card fraud detection with artificial immune system," *Artificial Immune Systems*, 2008.

[5] D. Sánchez, M. Vila, L. Cerda, and J. Serrano, "Association rules applied to credit card fraud detection," *Expert Systems with Applications*, vol. 36, no. 2, pp. 3630–3640, Mar. 2009.

[6] S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C. Westland, "Data mining for credit card fraud: A comparative study," *Decision Support Systems*, vol. 50, no. 3, pp. 602–613, Feb. 2011.

[7] D. J. Weston, D. Hand, N. M. Adams, C. Whitrow, and P. Juszczak, "Plastic card fraud detection using peer group analysis," *Advances in Data Analysis and Classification*, vol. 2, no. 1, pp. 45–62, Mar. 2008.

[8] T. Hastie and R. Tibshirani, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2009.

[9] J. V. Hulse and T. M. Khoshgoftaar, "Experimental Perspectives on Learning from Imbalanced Data," in *International Conference on Machine Learning*, 2007.

[10] D. Hand, C. Whitrow, N. M. Adams, P. Juszczak, and D. Weston, "Performance criteria for plastic card fraud detection tools," *Journal of the Operational Research Society*, vol. 59, no. 7, pp. 956–962, May 2007.

[11] C. Elkan, "The Foundations of Cost-Sensitive Learning," in *Seventeenth International Joint Conference on Artificial Intelligence*, 2001, pp. 973–978.

[12] G. Jayanta K., D. Mohan, and S. Tapas, "Bayesian Inference and Decision Theory," in *An Introduction to Bayesian Analysis*. Springer New York, Apr. 2006, vol. 13, no. 2, pp. 26–63.

[13] C. Whitrow, D. Hand, P. Juszczak, D. Weston, and N. M. Adams, "Transaction aggregation as a strategy for credit card fraud detection," *Data Mining and Knowledge Discovery*, vol. 18, no. 1, pp. 30–55, Jul. 2008.

[14] C. Phua, V. Lee, and K. Smith, "A comprehensive survey of data mining-based fraud detection research," *Artificial Intelligence Review*, 2005.

[15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[16] V. Sheng and C. Ling, "Thresholding for making classifiers cost-sensitive," in *Proceedings of the National Conference on Artificial Intelligence*, 2006.