Faculty of Mathematics and Computer Science

Emanuel Bîscă

Machine Learning

Fraud Detection – a Survey from a Data Mining Perspective

2020

Department of Computer Science

Abstract

Nowadays, with the continuous progress of the financial technologies, the number of people wanting to earn undeserved profits has grown larger. The financial area is forced to improve their systems of detecting frauds, because these faults are costing billions of dollars, annually just in the United States. One answer when dealing with fraud detection is offered by data mining and its algorithms of knowledge discovery through pattern revealing.

After a detailed introduction, the paper contains an investigation on various methods with information about both the specific techniques and the limitations that break it. The results of two different experiments were presented at the end, in order to see the different accuracy of the selected algorithms being compared. The main goal was having an image of the existing techniques used in fraud detection and to improve the performances with ideas on further work.

Contents

Chapter 1

Introduction

The process that involves the patterns discovery is called **data mining**. It would be wise to mention since the beginning that *data mining* stands at the intersection of statistics, database systems and machine learning. As a consequence we are able to affirm that *data mining* is a field of both statistics and computer science with the main purpose of extracting information (using methods known for their intelligent approach) from given datasets. Further, this information needs to be transformed into a different structure, easy to understand by other tools, sometimes even humans.

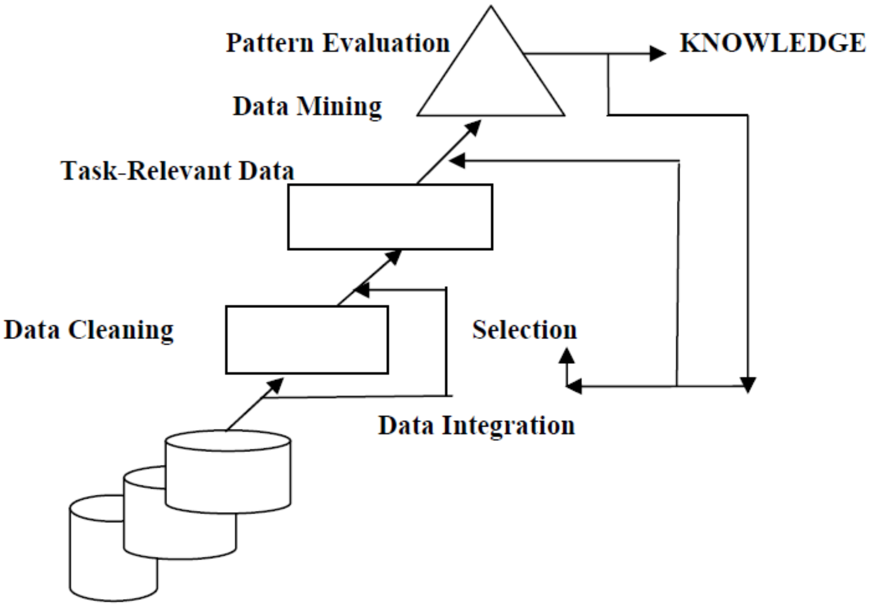


Figure 1.1: Data mining – the process of discovering knowledge

Being a misnomer, the *data mining* means using large datasets in order to extract knowledge and distinct patterns from them, not specifically the extraction (in other words *mining*) of data itself. These being said, we are able to affirm that the task of data mining is the automatic and sometimes semi-automatic analysis of big structures of data. This can be achieved by extracting the patterns previously mentioned, (which are previously unknown to the public) such as groups of abnormal records, – known as anomaly detection – data records, – simply cluster analysis – and dependencies – sequential pattern mining, association rule mining. Most of the time, the *spatial indices* are involved in this process, as a database technique.

Predictive analytics or machine learning are two major fields of knowledge that use these resulted patterns for further in depth analysis (J. Han et al. 2012). This is possible because the resulting patterns are interpreted as a summary of the given input data.

1.1 Motivation

It was proved that *data mining* is a good tool which is fit for use in a large area of expertise. Here are some subjects worth mentioning: the filtering of spam emails, the marketing of databases, the management needed by the risks of having a house loan, ultimately data mining has beginning to finds it purpose even in sentiment analysis or interpretation of opinions (Gayathri and Malathi 2013). Nowadays it has become essentially to detect, stop and even predict the fraudulent actions performed globally at a higher rate every year.

The economic experts are defining criminal or wrongful embezzlement intended to result in financial or personal gain as **fraud**. So in other words, let's consider an organization, it has to deal with fraud if there exist mishandling of its revenues with the intended will to avoid permissible consequences. In business, depending on its widespread, fraud can be an extremely serious problem to handle, especially if one cannot guarantee the existence of deterrence procedures. Today, overall speaking the fraud control and specifically the fraud detection have important tasks, forcing the area to be one of the most easily recognizable data mining application.

1.2 Framework

For a better understanding of the subject, I would emphasize the importance of knowing the sources of fraudulent activities and the most affected industries.

1.2.1 The sources of fraud

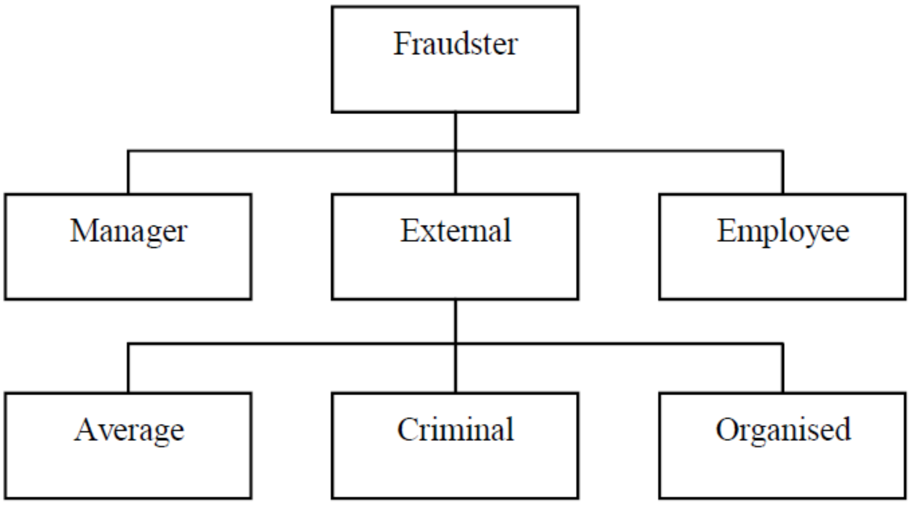


Figure 1.2: Types of fraudsters

Having a closer look at the Figure 1.2, it can be distinguished how a typically fraudster can be classified and interact with the selected business in order to commit fraud, being driven by the profit perspectives. Every business is vulnerable to fraud coming from its own management and also from management-free employees, these two point of view can be interpreted also as high-level, respective low-level corruption. Furthermore, the external environment can provide a wide range of fraudsters, interested in undeserved personal gain.

There are three basic profiles of an external fraudster: the average one, adopting a dishonest behaviour every time an opportunity exists, especially if experiencing a financial hardship; the criminal or the organised crime offenders. These last ones posses a greater risk, as they're experts in their area of expertise, perfecting themselves every single time, to fool the systems and to make their approximations of legal forms difficult to distinguish from authentic ones.

It is essential to be aware of these dangers, to analyse in detail how the career fraudsters interact with the algorithms used by detection systems. The average offenders are more likely to commit insurance or internal frauds, while it's highly probable that the credit and telecommunications fraud would be preferred by professionals. Think of a company that has millions of interactions with even more partners (seen as external parties) it's not practical to manually inspect the activities and identities for each and every external agent. Once more, the serious importance of data mining becomes undeniable, being used to determine visual anomalies, suspicion scores, rules so that the riskiest external agents can be pointed out for further investigations.

1.2.2 Affected areas

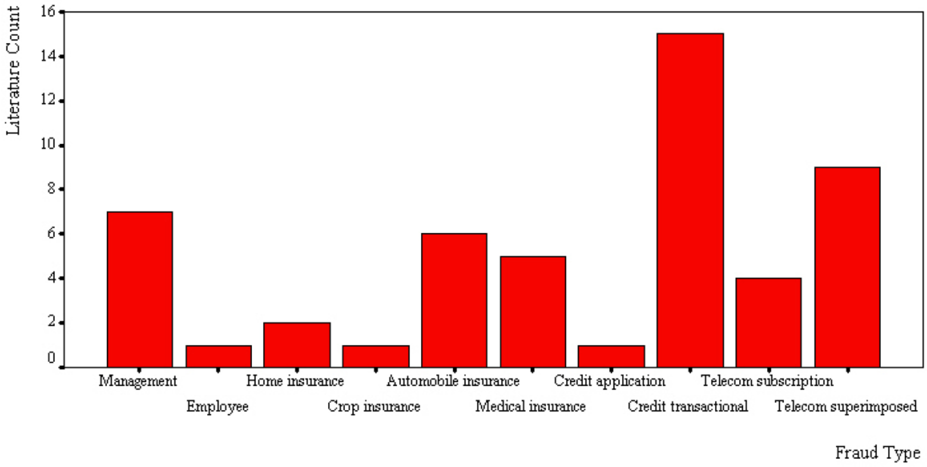


Figure 1.3: Fraud types as depicted in 51 different scientific papers on fraud detection

The information used for the chart presented in Figure 1.3 was retrieved from Phua et al. (2010). We can see in the figure, details related to different areas of interest for fraudsters and about the fraud detection research being conducted in these areas. If you're wondering why the researchers have shown such a great attention to detection of *Credit transactional fraud*, the next two passages can be revealing.

Modern times are requiring modern ways of doing things, while the comfort of humans must grow higher, so it's normal that the use of credit cards has increased beyond imaginations, in the last decade. The study conducted by (Robertson 2018) states that at the beginning of 2018 the number of people holding a credit card, a debit card or a prepaid card exceeded 20.48 billions.

Also, the currently growing number of fraudulent activities are the result of the imagination the fraudsters are putting into work – which makes it possible for them to design new methods for fraudulent purposes – and the high number of electronic payments performed daily at a world-wide level. According to Wang and D. Han (2018), there has been a tremendous growth of financial damages on account of electronic payment fraud, with the 7.6 billion dollars in 2010 reaching more than 21.8 billion dollars, five years later; so naturally we can affirm that the damages have increased with 300 percents in 2015, related to historical data from 2010.

That's why I considered it important that my survey must further focus on data mining techniques used for detection of credit card frauds.

Chapter 2

Investigation

The process that aims to identify fraudulent transactions involved when payments are done using credit cards, is known by specialists as credit card fraud detection. The main goal would be the partitioning of a large database into two different parts, one containing suspect transactions, with a higher possibility of being frauds, while the other one contains the transactions interpreted as genuine by the system. The **Survey** section debates several perspectives (which are naturally different) of dealing with fraud detection.

2.1 Fraud detection using neural networks

One important aspect to begin with, is that a fraud cannot be detected while a considered transaction is still processed, even if there are multiple ways in which data mining is used in the technology involved in fraud detection. The experts are encountering this because during a transaction the chance that a fraud would occur is minor.

The principle behind the fraud detection based on neural networks, is the way in which the human brain works. So it's natural to affirm that a system like this, uses this technology in order to think. We, as humans, use the previous experiences from our past to develop, in some way, a database of knowledge that help us in the decision making process. The same principle applies for the technology supporting the credit card fraud detection process. Every card holder who makes transactions leaves a trace behind, a pattern describing the behaviour of that specific card holder.

Fu et al. (2016) presented a new framework that uses a Convolutional Neural Network for mining hidden patterns of credit card fraudulent transactions. The authors present a high performance, relatively to the newest methods.

Constraints

Patidar and Sharma (2011) pointed out some problems that limit the neural networks use: before the training phase can start, a number of parameters must be set and moreover, there are certain rules that must be met when defining the parameters. The success of the training is decided by how well these parameters were defined.

Generally, the neural network is formed as more neurons join together, with every one coming with its own inputs, which are mapped to the related output. The networks are distinguished from one to another if we take a closer look at how they're interconnected. The performance of such a network is highly affected by his topology, but unfortunately, currently there are not many ways in which one can states the optimal topology to be used for a given problem, mainly as a consequence of the complexity of large networks. Ultimately the success of the training process is bounded by the parameters that are required, such as learning rate, network topology or the initial weights.

2.2 Fraud detection using decision tree

It's been long disputed that the decision tree logic can be used to develop something called a *similarity tree*. Being defined recursively, such a similarity tree has the following characteristics:

- Attribute names are considered labels for the nodes.
- The values of these attributes are considered labels for the edges, we must keep in mind that these values must assure some conditions.
- The definition of the *leaves* states that the strength of the connection is given as the division between the number of transactions satisfying the given condition and the total number of legitimate transactions

Despite being easy, this systems has the deficiencies that it's necessary to check every operation, one after another. This method is fit for other types of frauds, because there are proven results given by resemblance trees (Fan et al. 2001). It,s remarkable, the results being true even if with an inductive decision tree, used with the desire to initiate an intrusion detection system.

If we are considering the credit card transaction to be performed online, beside the verdict of a transaction being suspect or genuine, Dhanapal and Gayathiri (2012) tracked down the location of the client using the IP address. Sahin et al. (2013) was also interested in the study of decision trees and their ability to recognize frauds performed on credit card transactions; his dataset was worth 6 months of transactions from an important economic agent. The conclusion is that decision trees can be improved to offer higher performances.

Constraints

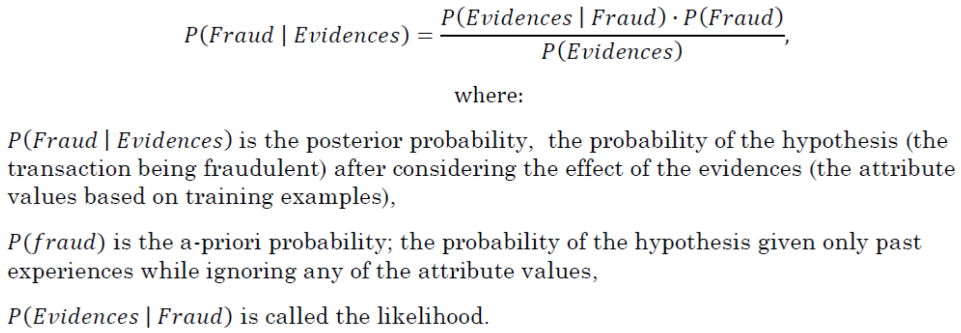
The first sign of attention when using decision tree learners, is the over-complexity of the generated tree, they are a great threat to the integrity of the system because, most of the time, they do not generalize well from the training data. Feeding the system with the proper information, both internal and external, affects the reliability of the information from the decision tree. Even minimal changes embedded in the input data, can fatally change a decision tree. Major loss of important data can force us to redesign our decision tree, if we change our variables without duplicating our information and including it or because of midway altered sequences.

The fundamentals of a decision tree consists of expectations, which makes the decision tree analysis weak and open to many errors. The contingencies arising from the decision making process, are not always part of the reality, so the consequence would be the tree to return a decision proved bad. In this case, the fact that the tree naturally uses a way by following related events it's of no help.

2.3 Fraud detection using a Naïve Bayes classifier

All the Naive Bayes classifiers form a group of easy *probabilistic classifiers*, in statistics. The principle behind them is the Bayes theorem, applied with a powerful – naïve – premise of independence between the features. It is common to use Kernel density estimation together with a Bayesian network model, mostly because the accuracy can be highly increased, even though the Naive Bayes classifiers are one of the easiest kind of a Bayesian network model.

John and Langley (1995) were the ones that first introduced this algorithm. The tests being done using real-world data, since then, have clearly proved that the Naive Bayes classifier is performing in a manner similar to other classical induction algorithms. These two have also conducted some experiments in which the Gaussian distribution was replaced by a kernel density estimation; their work concluded with the statement that in this case, the Naive Bayes classifier can obtain the same performance, even better ones in some particular cases, than the C4.5 algorithm, a decision tree based algorithm. Further, so that the classification can be done, the Bayes rule needs to be applied in order to find out the probability of the right class, taking into account the credit card transaction with its specific attributes:



$$P(Fraud \mid Evidences) = \frac{P(Evidences \mid Fraud) \cdot P(Fraud)}{P(Evidences)},$$

where:

$P(Fraud \mid Evidences)$ is the posterior probability, the probability of the hypothesis (the transaction being fraudulent) after considering the effect of the evidences (the attribute values based on training examples),

$P(fraud)$ is the a-priori probability; the probability of the hypothesis given only past experiences while ignoring any of the attribute values,

$P(Evidences \mid Fraud)$ is called the likelihood.

Figure 2.1: Bayes rule

Constraints

The hypothesis of class conditional independence does not remain true frequently. The Naive Bayes classifier cannot model the dependencies between the attributes. The increase of the size of the sample is often a reason why it is not handled with a high efficiency.

2.4 Fraud detection using k-NN

The most important feature used by the k-NN (meaning k-Nearest Neighbour) is the similarity measure, it uses it to classify the new possible cases. Together with the fact that it provides all the instances being available, the k-NN algorithm can be seen as a pretty straightforward instance-based learning method. Although all the other methods of learning are somehow considered *instance based*, for their initial training sequence which begins with a batch of instances, the instances themselves are representing what has been learned, in the case of the instance based learners.

The distance metric is applied to compare every existing instance with the instances in the process of arriving. The greater the distance between a new instance and an existing one provides a better probability of them being designated apart. If more than a single so called neighbour is used in this process, then all the nearest k neighbours would get designated to the new instance. The class being numeric would mean that the distance weighted average would get designated to this newly found instance.

Aha et al. (1991) introduced the framework of this algorithm. Even though a great period of time has passed since then, the Euclidean distance remains the standard way of calculating the distances between multiple numerical attributes. Of an essential importance is the problem of deciding which features possess the greater weight, most of the time the preferred solution is to normalize the data so that every attribute keeps the same influence on the results.

It must be well reflected in the metric of distance, by weighting the attributes, that some of them own a bigger importance. The instance based learning has a fundamental problem that needs further discussing, which is the obtaining of appropriate weights (related to the attributes, of course) derived from the training dataset.

Even though this method cannot offer us a pattern in the data, as the previous techniques, the distance metric is being combined with the possible cases in order to highlight the frontiers of the instance space, which makes it easier to tell the difference between two different classes. So that's been considered an explicit knowledge representation.

Constraints

Traditionally, the classification using k - nearest neighbour, encounters three main constraints:

- Great complexity of computations. The fact that we must compute every similarities existent among the samples of training in order to search for the k-nearest neighbour sample, explains this complexity. There cannot be a minimal number of samples of training, because the k nearest neighbour classifier is endangered to become no longer optimal. The same thing goes with too much samples, as more time would be necessary to compute these similarities.

- The subjection on the training set. Only the samples of training are being used when we generate the classifier, so because there's no supplementary data this forces the k-NN algorithm to rely excessively on the training data. A tiny change being performed on the training data would require immediate re-computation.

- The samples are not differentiate by weights. This is possible because the algorithm tends to treat all the training samples equally, even if among them ones have large numbers of data and others not.

2.5 Fraud detection using Support Vector Machines

The Support Vector Machines are known to the public as SVMs and the algorithm behind them was made acquainted when Cortes and Vapnik (1995) published their work. The SVM algorithm uses for classification maximum margin hyper plane, which is an interesting type of linear model. So it takes the instances from training and it's correctly classifying them into classes through this linear model, called simply the hyperplane. The support vectors are defined as those instances having the least distance to the maximum margin hyperplane – so in other words those being the closest to it. Every single class has, at least, one of such a vector, however, has been found out repeatedly that there are more than one.

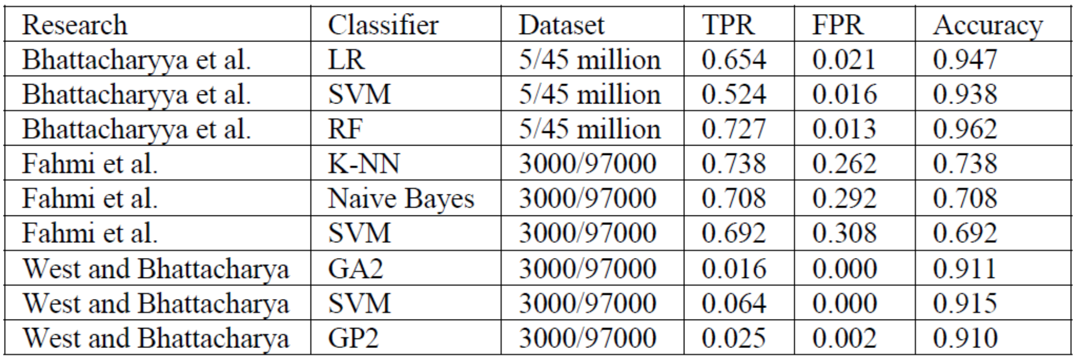
Moepya et al. (2015) has conducted a research on fraud detection that pointed out that the weighted SVMs are higher-level than the k-NN classifiers and the cost-sensitive Naive Bayes models.

Constraints

How to choose the appropriate kernel for a given problem is the greatest constraint meet by the SVM algorithm. There is still a hot topic for researcher world wide, as well as the solution for the optimal design of the multiclass Support Vector Machines classifiers.

2.6 Some experimental results

West and Bhattacharya (2016) and Fahmi et al. (2016) have conducted experiments and both had considered the same set of data, the one from "UCSD-FICO Data Mining Contest 2009". In this set, 97 percents of data were genuine credit card transactions, while the remaining 3 percents were fraudulent activities. The next figure shows their results:



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

Figure 2.2: View of distinct algorithms used in detection of credit card fraud

It's easy to see how the k - Nearest Neighbours model (Fahmi et al. 2016) had given better results than all the other models, by all means. On the other hand, the Support Vector Machines' model (West and Bhattacharya 2016) has won the race for the best performance, with accuracy at the highest level and a zero false positive rate (FPR).

2.7 Concluding remarks

The first part of my research was dedicated to the conceptual framework of data mining related to fraud detection. There can be found a classification of fraudsters, from different perspectives as well as the threats the society has to oppose.

In the second part, five different approaches on fraud detection based on data mining have been discussed. The accent is upon the constraints that each method encounters. These methods were Decision Tree, Neural Networks, K-NN (k - Nearest Neighbours), Naïve Bayes andd SVM (Support Vector Machines).

As a further idea of research, I believe it to be important that we would develop new methods which would deal in a better way with the incomplete sets of data, or which would deal with outliers in innovative ways. The scientists need to remain with a step ahead of the fraudsters.

References

Aha, D., Kibler, D, and Albert, M. (1991). “Instance-based learning algorithms”. In: *Machine Learning* 6, pp. 37–66.

Cortes, C. and Vapnik, V. (1995). “Support-vector networks”. In: *Machine Learning* 20, pp. 273–297.

Dhanapal, R. and Gayathiri, P. (2012). “Credit Card Fraud Detection Using Decision Tree For Tracing Email And IP”. In: *International Journal of Computer Science* 9.5-2, pp. 406–412.

Fahmi, M., Hamdy, A., and Nagati, K. (2016). “Data Mining Techniques for Credit Card Fraud Detection: Empirical Study”. In: *Sustainable Vital Technologies in Engineering and Informatics*, pp. 1–9.

Fan, W., Miller, M., Stolfo, S., Lee, W., and Chan, P. (2001). “Using Artificial Anomalies to Detect Unknown and Known Network Intrusions”. In: *Conference Proceedings of ICDM01.*

Fu, K., Cheng, D., Tu, Y., and Zhang, L. (2016). “Credit Card Fraud Detection Using Convolutional Neural Networks”. In: *Proceedings of International Conference on Neural Information*, pp. 483–490.

Gayathri, R. and Malathi, A. (2013). “Investigation of Data Mining Techniques in Fraud Detection: Credit Card”. In: *International Journal of Computer Applications* 89.9, pp. 12–15.

Han, J., Kamber, M., and Pei, J. (2012). *Data Mining: Concepts and Techniques.* 3rd ed. The Morgan Kaufmann Series in Data Management Systems. Elsevier.

John, H. and Langley, P. (1995). “Estimating Continuous Distributions in Bayesian Classifiers”. In: *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pp. 338–345.

Moepya, S., Akhouryand, S., and Nelwamondo, F. (2015). “Applying Cost-Sensitive Classification for Financial Fraud Detection under High Class-Imbalance”. In: *Proceedings of IEEE International Conference on Data Mining Workshop*, pp. 183–192.

Patidar, R. and Sharma, L. (2011). “Credit Card Fraud Detection Using Neural Network”. In: *International Journal of Soft Computing and Engineering* 1, pp. 32–38.

Phua, C., Lee, V., K., Smith, and Gayler, R. (2010). “A Comprehensive Survey of Data Mining-based Fraud Detection Research”. In:

Robertson, D. (2018). “Payment Cards Projected Worldwide”. In: *The Nilson Report* No. 1140.

Sahin, Y., Bulkan, S., and Duman, E. (2013). “A cost-sensitive decision tree approach for fraud detection”. In: *Expert Systems with Applications* 40, pp. 5916–5923.

Wang, C. and Han, D. (2018). “Credit card fraud forecasting model based on clustering analysis and integrated support vector machine”. In: *Cluster Computing,* pp. 1–6.

West, J. and Bhattacharya, M. (2016). “Some Experimental Issues in Financial Fraud Mining”. In: *Procedia Computer Science* 80, pp. 1734–1744.