Faculty of Mathematics and Computer Science

Emanuel Bîscă

Advanced Methods of Data Analysis

# Credit Card Fraud Detection Using Clustering

2020

Department of Computer Science

# Abstract

Financial frauds are a real problem which makes scientists to unite in defending the security of all types of systems. The unprecedented number of fraudulent activities are the result of the growing number of online transactions, often desired by people because it's much easier and they have a sense of comfort. Understanding the fact that the credit cards payments are the most popular to be attacked by fraudsters, is essential in protecting all the citizens who make transactions daily. In this paper, I show the use of clustering in the detection of these credit card frauds. The first approach was based on the expenditure of the owners, to predict suspicious actions performed. The second approach presented a hybrid method which used the cluster analysis applied on qualitative data.

# Contents

# Chapter 1

# Introduction

In our modern times, the use of credit cards has increased exponentially. Nowadays, it has become the most popular payment method. A study (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.

## 1.1 Purpose

The credit card transactions have escalate extensively because of the online shopping, the result of globalization which amplifies the use of internet. The most notable consequence of the growing numbers of transactions performed by credit card holders, is the massive rise of fraudulent activities. The logical conclusion, would be that it becomes clearly necessary to develop mechanisms being prepared to offer assistance in detection-fraud systems.

Also, the more and more fraudulent activities are the result of the imagination possessed by the fraudsters – which makes it possible for them to design new methods for fraudulent purposes – and the increased number of electronic payments performed world wide. According to Wang and Han (2018), there has been a tremendous increase 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 basically the damages has increased with 300 percents in 2015, related to 2010.

A very difficult task for engineers all over the world is detecting these frauds and even predicting them. The experts in cybersecurity often use the clustering of data, when dealing with these issues. The scope of this paper would be to offer to the reader a range of clustering models used in experiments regarding the detection or predicting of frauds performed on credit card payments.

## 1.2 Clustering Types

Clustering, as a data mining task, is used when it's desired to divide the available information into different groups so that the data from the same group resembles highly similar characteristics while the distinct groups are considered as unlike as possible. Another important concept involved is the *cluster analysis* which consist of breaking down the studied data with the wish of making the data pattern easy visible; the related components contained in each cluster are of help when performing this. (Madhulatha 2012)

In the next figure, it's presented a classification of various methods used when clustering data provided. We will see that different methods are fit for different purposes, some approaches even use hybrid methods or techniques in predicting the credit card frauds.

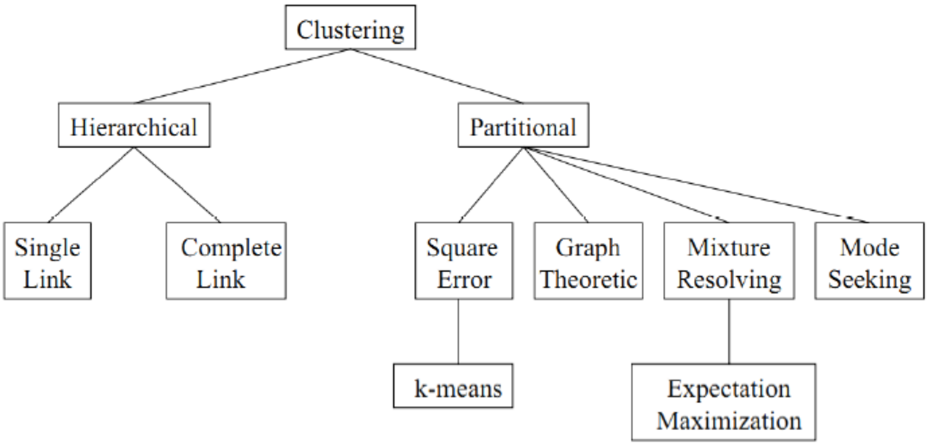


Figure 1.1: Types of Clustering

# Chapter 2

# Survey

In the next sections, different experiments regarding the use of clustering in detection or prediction of credit card financial frauds, are presented.

## 2.1 Clustering and HMM

Sathyapriya and Thiagarasu (2019) proposed a model in which the duration of time is kept to a minimum, the main idea being detecting fraudulent actions among the transactions while a high efficiency in the process of analyzing huge amount of data is provided.

The system is based on analyzing customers credit card payments in the purpose of discovering patterns which later will lead to financial frauds. The *k-means clustering* algorithm is the one that facilitates the validation process of these patterns. This algorithm is part of the unsupervised learning area and uses the past activities to predict the normal usage pattern of card holders (Dragan et al. 2016). So these transactions performed in the past must be stored in a database and trained aiming to possess in the end a clear image of predicted future transactions used to detect fraud patterns. This experiment is focused on offering a way of developing new prototypes for detection-fraud systems, by creating clusters based on a specific dataset and studying them in order to find anomalies.

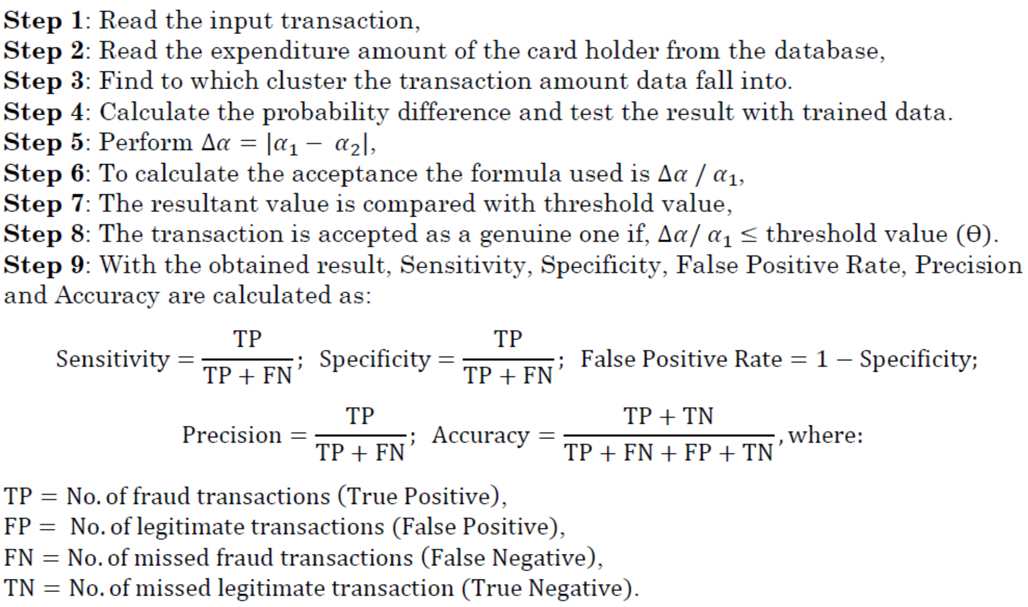
### 2.1.1 System Design

The main purpose of designing such a system, is to detect fraudulent activities involving cred card payments, with the aim to reduce false alarm rates, which can be divided into two different steps. More information regarding the concepts and use of the each step can be found below the next paragraph. Training happens in the first step, while the second step is saved for the actual detection.

For the purpose to become fulfilled, the expenditure likelihood pattern is used to predict incoming transactions behaviour of every subject, based on the history of the most recent credit card activities performed by each one of them. Essentially it's all about the way in which we extract the knowledge regarding the expenditure behaviour of the analyzed card holders and use it in implementing a behavior likelihood model, able to detect or predict credit card frauds. Let's see more information on each step:

1. The first step consist of a training phase, where the data is presumed to be composed of the behavior of expenditure related to every considered credit card holder. After all the data is being stored, we want to apply the *k-means clustering* process to obtain the clusters of activities using the individual expenditure pattern. Once the clustering is done, a *Hidden Markov Model* – HMM – is used to generate these payments as observation symbols, for example see the sequence of states low, medium, high; where we consider the payment amount used by any given card holder to be recorded on previously mentioned the sequence of states.

2. The second step is of a greater importance, being the so called *detection phase.* Now the idea is to compare the already collected data on the expenditure level of each card holder, with the likelihood of a new payment. Also, now the observation symbols are being generated for the recently incoming payments. Keeping a specific threshold value in the game, if the obtained result values is contradicting the threshold it's considered a fraudulent actions which must attract attention upon. Otherwise nothing happens, as the system assumes that the transaction was authentic. With the desire of giving the best accuracy over credit card actions, this model is developed with the help of the algorithm presented in the next figure:



**Step 1**: Read the input transaction,
**Step 2**: Read the expenditure amount of the card holder from the database,
**Step 3**: Find to which cluster the transaction amount data fall into.
**Step 4**: Calculate the probability difference and test the result with trained data.
**Step 5**: Perform $\Delta\alpha = |\alpha_1 - \alpha_2|$,
**Step 6**: To calculate the acceptance the formula used is $\Delta\alpha \,/\, \alpha_1$,
**Step 7**: The resultant value is compared with threshold value,
**Step 8**: The transaction is accepted as a genuine one if, $\Delta\alpha/\,\alpha_1 \leq$ threshold value ($\Theta$).
**Step 9**: With the obtained result, Sensitivity, Specificity, False Positive Rate, Precision and Accuracy are calculated as:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}};\ \ \text{Specificity} = \frac{\text{TP}}{\text{TP} + \text{FN}};\ \ \text{False Positive Rate} = 1 - \text{Specificity};$$

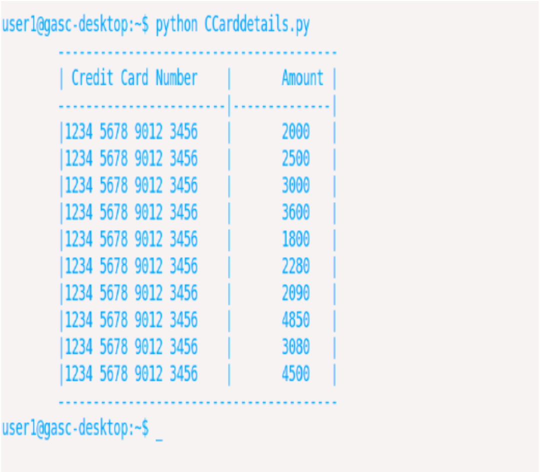
$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FN}};\ \ \text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}}, \text{where:}$$

TP = No. of fraud transactions (True Positive),
FP = No. of legitimate transactions (False Positive),
FN = No. of missed fraud transactions (False Negative),
TN = No. of missed legitimate transaction (True Negative).

Figure 2.1: Detection Fraud Algorithm

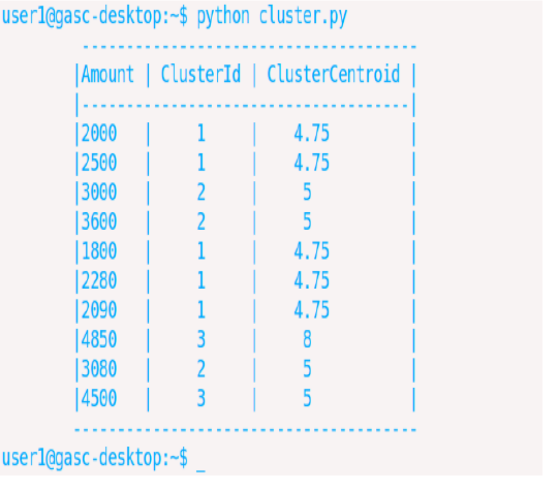
### 2.1.2 Experimental Results

The detection of credit card frauds is the purpose of the suggested system. The development of the system takes into account several improvements over the performance metrics. The software framework used to create the dataset was Spark. Firstly, before any action, the system features the data parameters as the previously discussed observation symbols, then the history of payments are stored in the database. The next move would be using *k-means clustering* to group the data in three distinct clusters defined as low, medium and high related to the expenditure of every considered card owner.



```
user1@gasc-desktop:~$ python CCarddetails.py
       ---------------------------------------
       | Credit Card Number    |      Amount |
       ------------------------|-------------|
       |1234 5678 9012 3456    |       2000  |
       |1234 5678 9012 3456    |       2500  |
       |1234 5678 9012 3456    |       3000  |
       |1234 5678 9012 3456    |       3600  |
       |1234 5678 9012 3456    |       1800  |
       |1234 5678 9012 3456    |       2280  |
       |1234 5678 9012 3456    |       2090  |
       |1234 5678 9012 3456    |       4850  |
       |1234 5678 9012 3456    |       3080  |
       |1234 5678 9012 3456    |       4500  |
       ---------------------------------------
user1@gasc-desktop:~$ _
```

Figure 2.2: Representation of the dataset



```
user1@gasc-desktop:~$ python cluster.py
          ----------------------------------------
          |Amount | ClusterId | ClusterCentroid |
          |---------------------------------------|
          |2000   |    1      |   4.75          |
          |2500   |    1      |   4.75          |
          |3000   |    2      |   5             |
          |3600   |    2      |   5             |
          |1800   |    1      |   4.75          |
          |2280   |    1      |   4.75          |
          |2090   |    1      |   4.75          |
          |4850   |    3      |   8             |
          |3080   |    2      |   5             |
          |4500   |    3      |   5             |
          ----------------------------------------
user1@gasc-desktop:~$ _
```

Figure 2.3: The Clusters

In Figure 2.2 and Figure 2.3, it is shown the idea of how the data looks like, together with the payment amounts being interpreted into three distinct clusters. Also, for those interested, an analysis of how well the proposed system works, is presented in the next figures.

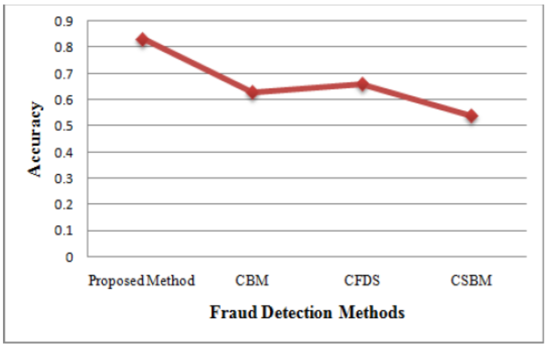


Figure 2.4: Fraud Detection Methods

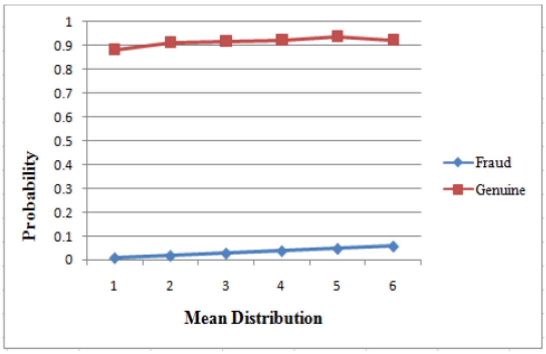


Figure 2.5: Mean Distribution

### 2.1.3 Remarks

The system job is to offer a high performance and accuracy based on the previous behavioural expenditure of the credit card owners. Also, it is desirable that the recently developed systems would give a better handling of big volumes of data, this system being no exception. With today's world, full of billions of financial transactions every week, it becomes really difficult to acknowledge when a false alarm is encountered. The experimental results are showing, that the proposed system can reduce the rate of alarms proved false, by studying the relation existent between the activities flagged as real frauds and the activities with a high probability of being fraudulent ones. The logical conclusion, would be that accuracy is served once the system owns the power of reducing the false alarms rate.

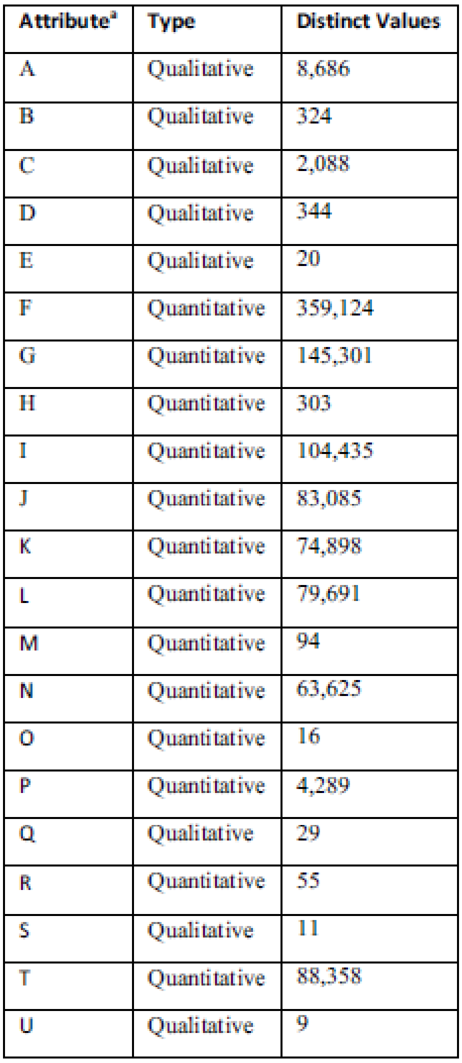
## 2.2 Clustering and ANN

The whole idea of just detecting the financial fraud activities would be in vain without the attempt to predict or even prevent these fraudulent activities. Billions of dollars are gone each year as a result of frauds that involve credit card transactions. So it became necessity for engineers to develop software capable of managing these actions and to prevent them from happening. The Artificial Neural Networks were proved to work harmoniously together with the data mining technique of clustering.

This section is reserved for the model proposed by Carneiro et al. (2015) which debates the fraud detection through the results obtained using the Cluster Analysis – CA – and Artificial Neural Networks – ANN. As well, we will discuss some aspects regarding the technology needed when developing the system handling the detection and preventing of the credit card fraudulent actions.

### 2.2.1 Problem Statement

The first challenge was finding a proper classification technique capable of using historical data in order to recognize transactions with a high probability of being frauds. For this specific task, an MLP (in other words a multilayer perceptron) classifier was considered. The data of transactions performed using a credit card possess characteristics concerning the quantity and the quality.



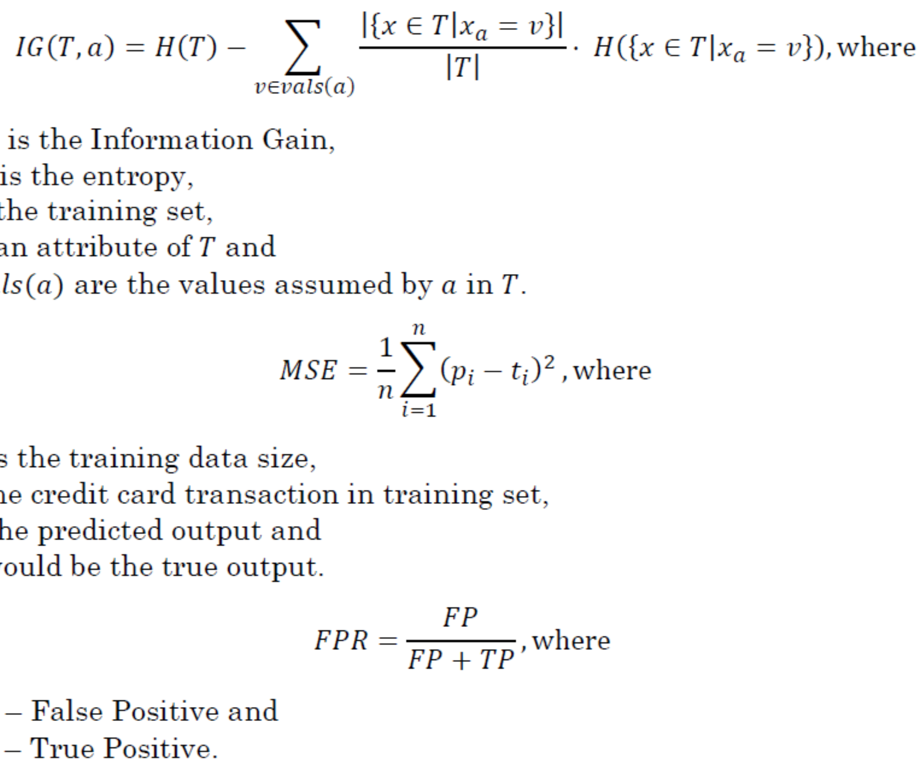
| Attribute[a] | Type | Distinct Values |
|---|---|---|
| A | Qualitative | 8,686 |
| B | Qualitative | 324 |
| C | Qualitative | 2,088 |
| D | Qualitative | 344 |
| E | Qualitative | 20 |
| F | Quantitative | 359,124 |
| G | Quantitative | 145,301 |
| H | Quantitative | 303 |
| I | Quantitative | 104,435 |
| J | Quantitative | 83,085 |
| K | Quantitative | 74,898 |
| L | Quantitative | 79,691 |
| M | Quantitative | 94 |
| N | Quantitative | 63,625 |
| O | Quantitative | 16 |
| P | Quantitative | 4,289 |
| Q | Qualitative | 29 |
| R | Quantitative | 55 |
| S | Qualitative | 11 |
| T | Quantitative | 88,358 |
| U | Qualitative | 9 |

Figure 2.6: Characteristics of Data

In the Figure 2.6 an overview about existing data characteristics is provided. A non-disclosure agreement keeps the name of the attributes unknown. The input values that are used by the MLP algorithm must be between 0 and 1. So a normalization of the data needs to be considered. The normalization the attributes related to quantity was done by rescaling, the used formula was:

$$x' = \frac{x - min(x)}{max(x) - min(x)};$$

Regarding the normalization of the qualitative data, most scientists prefers the method of creating for every different value a different input, which will let us with 27461 different input. In our experiment another manner was desired due to the improved performance, it consists of grouping the similar values into groups and creating input for each group instead for every single value. The normalization of the qualitative data was done using CA (cluster analysis).



$$IG(T,a) = H(T) - \sum_{v \in vals(a)} \frac{|\{x \in T | x_a = v\}|}{|T|} \cdot H(\{x \in T | x_a = v\}), \text{where}$$

- $IG$ is the Information Gain,
- $H$ is the entropy,
- $T$ the training set,
- $a$ an attribute of $T$ and
- $vals(a)$ are the values assumed by $a$ in $T$.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(p_i - t_i)^2 \text{, where}$$

- $n$ is the training data size,
- $i$ the credit card transaction in training set,
- $p$ the predicted output and
- $t$ would be the true output.

$$FPR = \frac{FP}{FP + TP}\text{, where}$$

- $FP$ – False Positive and
- $TP$ – True Positive.

Figure 2.7: Theoretical Notions

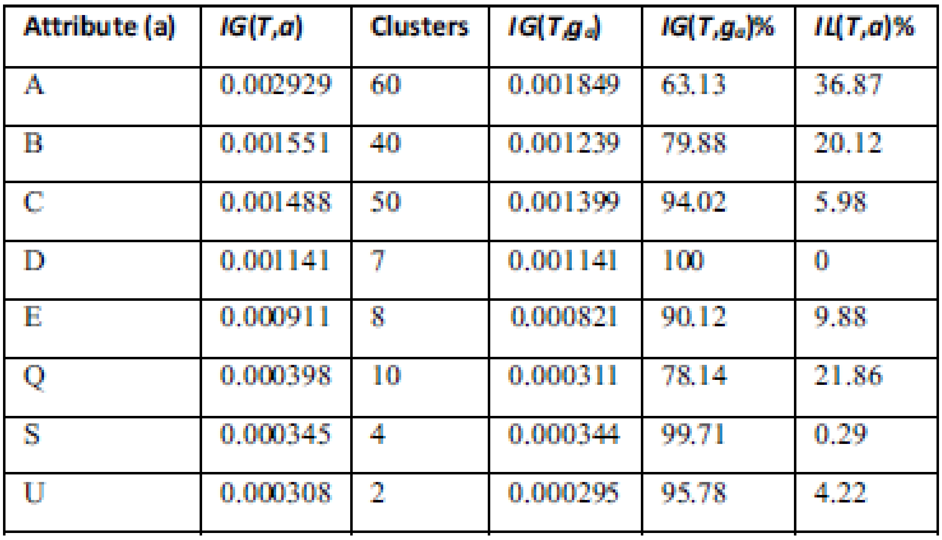
The use of cluster analysis can have results in the loss of information, so to prevent this from altering the experiment, the Kullback-Leibler divergence was applied. The formula and some explanations are present in Figure 2.7, we will call the Kullback-Leibler divergence, **Information Gain** from now on. Furthermore, the Mean Squared Error togeher with the False Discovery Rate were defined in Figure 2.7. These two have applications in the evaluations of the performance of the trained classifier.

### 2.2.2 Experimental Results

A total number of 645,538 transactions provided by the internet users of credit cards were used in the experiment. From all these, 37,359 were reported as frauds. Taking a closer look to statistics, we will see that from the total number of activities studied, 80 percents of them are being used in the training phase, while the remaining 20 percents are equally divided between validation and testing purposes.

As of the clustering of the qualitative data, an interesting algorithm was used, named INBIAC, which means Iterative Naïve Bayesian Inference Agglomerative Clustering. In few words, the Naïve Bayes classifiers would appear to give a great importance to the idea of independence. Their use is to study the relation existent betwixt dependent variable and the independent ones, in order to deliver a probability for each relation.

This algorithm is part of a bigger project, called the Clustering Engine. The main reason of the existence of this system, is the normalization of the inputs with the desire to obtained a trained MLP.



| Attribute (a) | $IG(T,a)$ | Clusters | $IG(T,g_a)$ | $IG(T,g_a)$% | $IL(T,a)$% |
|---|---|---|---|---|---|
| A | 0.002929 | 60 | 0.001849 | 63.13 | 36.87 |
| B | 0.001551 | 40 | 0.001239 | 79.88 | 20.12 |
| C | 0.001488 | 50 | 0.001399 | 94.02 | 5.98 |
| D | 0.001141 | 7 | 0.001141 | 100 | 0 |
| E | 0.000911 | 8 | 0.000821 | 90.12 | 9.88 |
| Q | 0.000398 | 10 | 0.000311 | 78.14 | 21.86 |
| S | 0.000345 | 4 | 0.000344 | 99.71 | 0.29 |
| U | 0.000308 | 2 | 0.000295 | 95.78 | 4.22 |

Figure 2.8: The results of CA

In the Figure 2.8, we can see how the results of the cluster analysis phase look like. For the C attributes, it was observed even a increasing data loss, once with the increasing number of clusters. These were possible because the time of processing needed for the MLP classifier was minimized. In the next figure, a graph showing the variance of IG related to distinct number of cluster is shown.

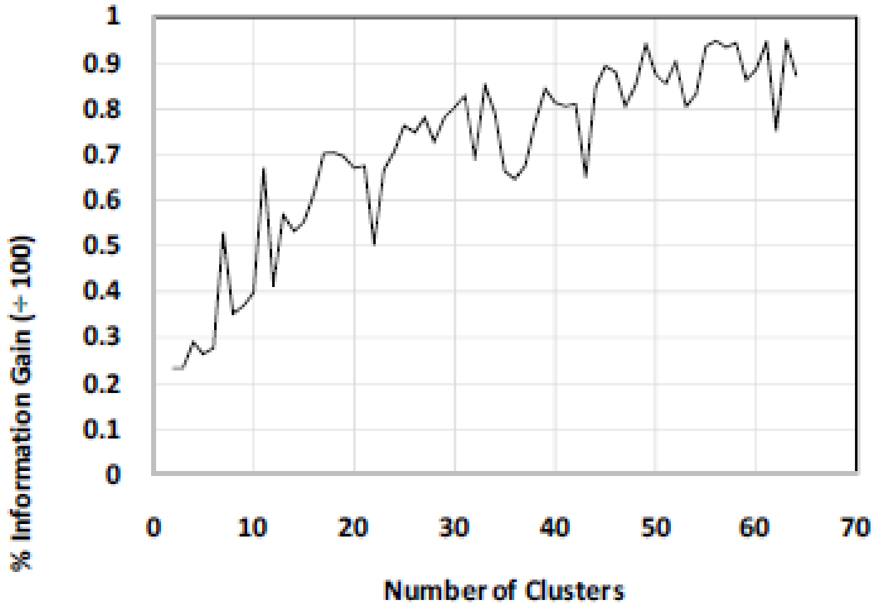


Figure 2.9: Variance

In the end all these transaction which are normalized now, are given to the MLP in order to be trained. The next figure shows the variation of the MSE during the training phase. For performance reasons, the FPR is 24.73 percents.

The threshold values is proved to have a powerful influence on the FPR as it can have values of 50 percents to 87.5 percents, as the preliminary data suggests. Further work should concentrate on the information loss. One important conclusion is that the more research needs to be done in order to prove the higher performance of the ANNs which use trained data, than the ones using raw data.

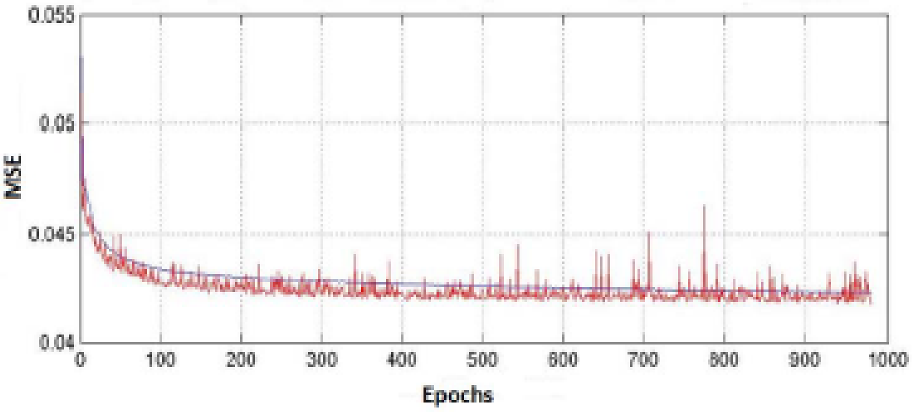


Figure 2.10: MSE Training and Validation Phases

## 2.3 Concluding Remarks

Today, engineers from all over the world are trying to put their effort into developing new methods in which the fraudulent actions are detected and prevented. In this paper I aimed to show two experiments in which the clustering is used. The first one was mainly about clustering, while the second experiment involved a Multilayer Perceptron Artificial Neural Networks and Cluster Analysis, which makes it a hybrid approach.

# References

Carneiro, E.M., Vieira Dias, L.A., da Cunha, A.M., and Mialaret, L.F. (2015). "Cluster Analysis and Artificial Neural Networks – A Case Study in Credit Card Fraud Detection". In: *The Conference Proceedings of the 12th International Conference on Information Technology - New Generations,* pp. 123–126.

Dragan, F., Borlea, I., and Precup, R. (2016). "On the Architecture of a Clustering Platform for the Analysis of Big Volumes of Data". In: *The Conference Proceedings of the 11th IEEE International Symposium on Applied Computational Intelligence and Informatics,* pp. 145–150.

Madhulatha, T. (2012). "An Overview on Clustering Methods". In: *IOSR Journal of Engineering,* 2.4, pp. 719–725.

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

Sathyapriya, M. and Thiagarasu, V. (2019). "A Cluster Based Approach for Credit Card Fraud Detection System using Hmm with the Implementation of Big Data Technology". In: *International Journal of Applied Engineering Research,* 14.2, pp. 393–396.

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.