A Review of Machine Learning Algorithms for Fraud Detection in Credit Card Transaction

Lim Kha Shing

Faculty of Science and Technology,

Quest International University Perak.

(E-mail: limkhashing@qiup.edu.my)

***Abstract* — The increasing number of credit card fraud cases has become a considerable problem since the past decades. This phenomenon is due to the expansion of new technologies, including the popularization of online banking transaction and e-commerce.** **In order to address the problem of credit card fraud detection, rule-based approach has been widely utilized to detect and guard against fraudulent activities. However, it requires huge computational power and high complexity in precisely identifying the fraud patterns. In addition, it does not come with intelligence and ability in predicting or analysing transaction data in looking for new fraud strategies. As such, Data Mining and Machine Learning algorithms are proposed in this paper. The aim of this paper is to highlight the important techniques and methodologies that are employed in fraud detection, while at the same time focusing on the existing literature. Methods such as Artificial Neural Networks (ANNs), Support Vector Machines (SVMs), naïve Bayesian, k-Nearest Neighbour (k-NN), Decision Tree and Frequent Pattern Mining algorithms are reviewed and evaluated for their performance in detecting fraudulent transaction.**

***Index Terms***—**Data Mining, Fraud Detection, Machine Learning, Supervised Classification**.

1. **INTRODUCTION**

Transactions between credit card and E-Businesses are the primary area that the fraudsters are abusing and exploiting. This is because that there exist various loopholes in the existing detection method, and the major factor – Business Negligence. There are multiple frauds that can be performed, such as deliberately underreporting or omitting income, overstating or claiming false amount of deductions, money laundering, swindle transactions and many more. To overcome the frauds that happening in the banking industry, a Rule-based system, also known as Production System or Expert System, are used as a way to store and manipulate fraud knowledges for helping interpret the information in a meaningful way.

A rule-based system is the simplest form of Artificial Intelligence that uses rules as the knowledge representation for knowledge coded into the system, and provide reasoning for the context of fraud detection. A typical Rule-based system consists of a list of rules, an inference engine or semantic reasoner, which infers information or acts based on the interaction of input and the rule base, a temporary working memory, and a User Interface which input and output signals are received and sent. In short, a Rule-based system is a system that infers the knowledges that stored in its rule base and mimic the reasoning of a human expert in solving a knowledge-intensive problem.

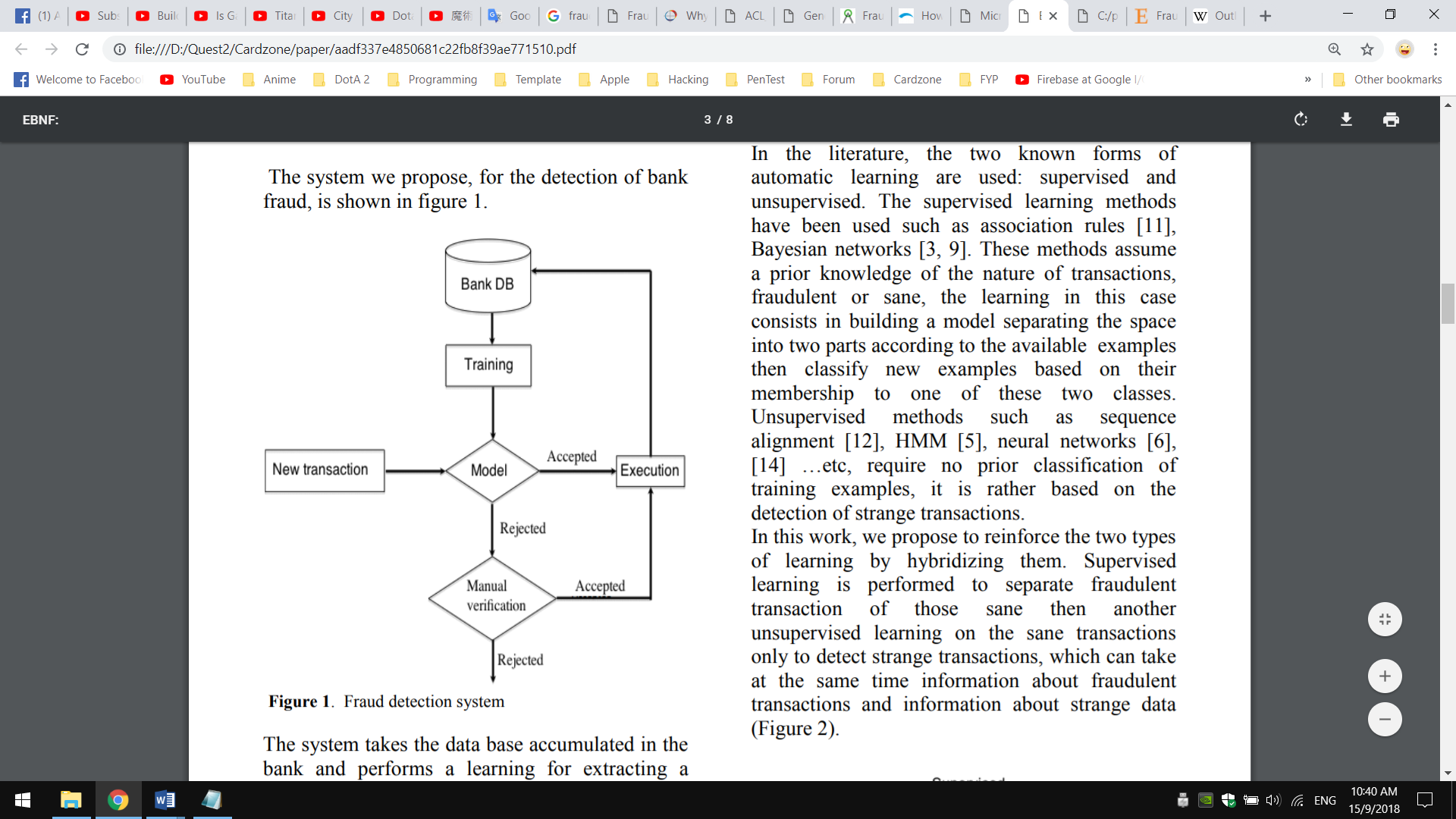
Although Rule-based system can be accurate in terms of its end result, however, it required a huge computational power for pattern matching and the rules had to be specially crafted to its working domain. The modification of rule base also is complicated, for example when introducing new knowledge to identify some specific fraud, it might introduce contradictions with the previous rules. Besides that, fraudsters are adaptive-able and given enough time, they will always be able to find ways to circumvent such prevention. For instance, they are capable to bypass simple pattern matching or rule-based detection and tricking the system into assuming that it was a genuine transaction, which often is not sufficient to detect those frauds accurately. The Rule-based system also does not come with any analysing and predicting capability that based on the data that it receives to help better identify new fraud in future. Hence, Data Mining and Machine Learning techniques were introduced to improve the detection in new fraud or offer a correlation in fraud data.

As Data Mining and Machine Learning are the branches of Artificial Intelligence, they are capable to analyse and discover the patterns in large datasets, and subsequently produce hidden insights, through learning from historical relationships and the trends in the data. There are multiple Data Mining and Machine Learning techniques that can be used in not only identify the fraud transaction, but also predicting the suspicious rate of transaction data that may transact over time. Methods such as Artificial Neural Networks (ANNs), Support Vector Machines (SVMs), Bayesian Classification, k-Nearest Neighbour (k-NN), Decision Tree and Frequent Pattern Mining algorithms are reviewed for discovering frauds in this paper. These methods involve distinguishing fraudulent financial data from authentic data, thereby disclosing fraudulent behaviour or activities, and enabling decision makers to develop appropriate strategies to decrease the negative impact of fraud.

In a nutshell, Data Mining techniques are proposed in this paper to overcome the difficulty in detecting fraud activities. This involves scrutinizing the behaviour of the user’s transaction history and in-dept analysing the transaction data to determine whether the transaction is genuine fraud. The ultimate goal of this paper is to provide a comprehensive review of Data Mining techniques that best in classifying fraudulent behaviours, identifying the major sources and characteristics of the data, based on the fraud detection that had been conducted.

1. **LITERATURE REVIEW**

In the context of AI, fraud detection is viewed as a classification problem [2], which the objective is to correctly classify the credit card transactions as legitimate or fraudulent. The detection methods were usually embedded in a Fraud Detection System, whereby the model contain the rules or knowledges that used to identify and preventing frauds. Fraud Detection System takes the accumulated data in the bank’s database, performs a training and learning, then output a model that best represent the characteristics of the transaction data.



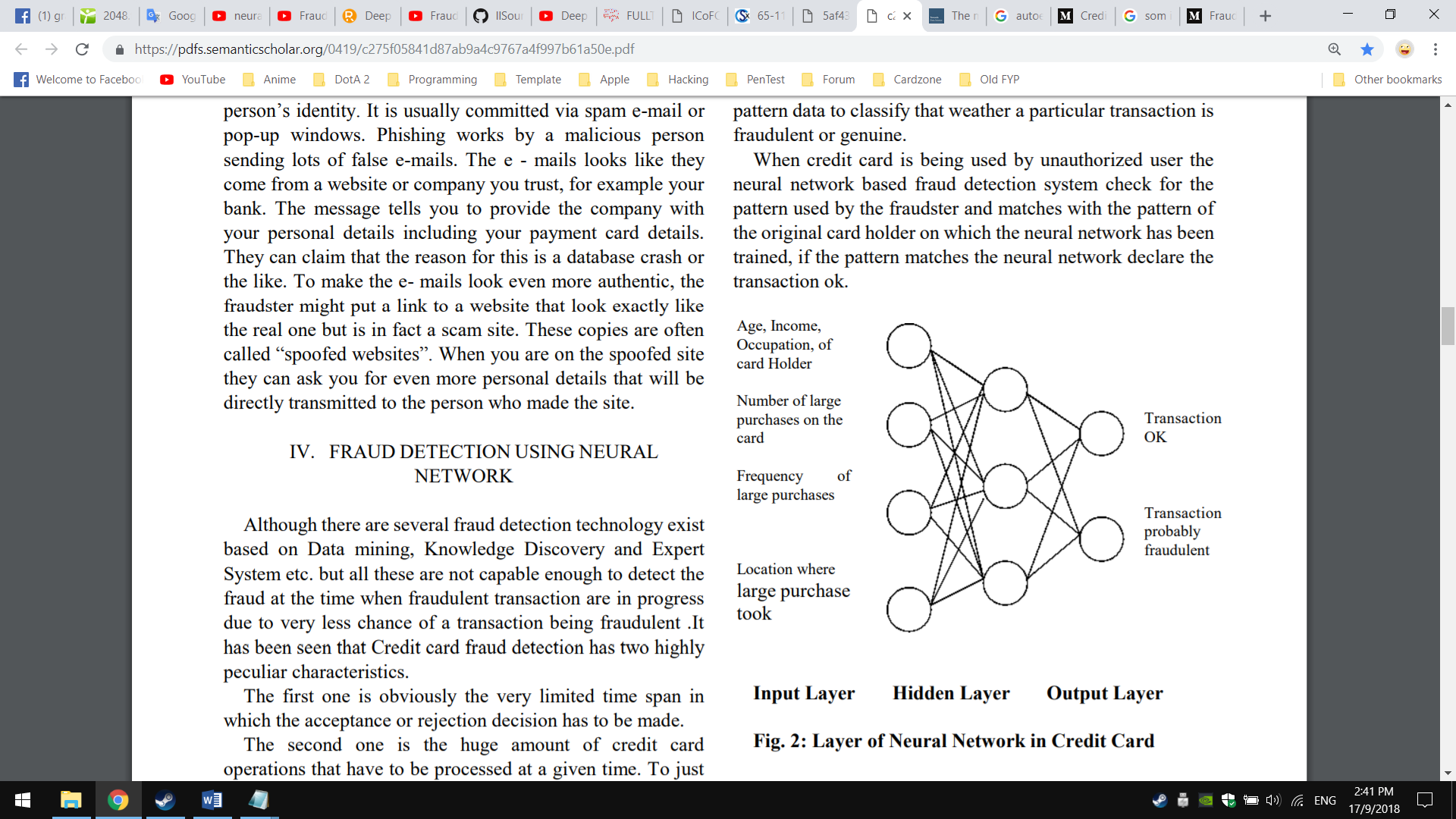
Figure

The outputted model is then used to decide new transactions, whether it can be accepted as genuine transaction, or reject it as fraudulent transaction. A transaction that is accepted by the model will be executed, follow by adding it to the database to improve the model, whereas rejected transaction will pass to manual check. If the rejected transaction is considered as normal after checking, the transactions are then executed and the information will be added into the bank’s database, otherwise the transaction is rejected

A big part of fraud detection and prevention process is to train and learn from the transaction data in order to identify new frauds. As such, it is important to design a model with best Data Mining and Machine Learning algorithm that suited in quickly detecting fraud and taking immediate action in preventing it. A good designed model would help in not only catching frauds, but also estimating the probability of fraudulent behaviour.

1. ***Artificial Neural Networks (ANNs)***

Artificial Neural Networks, or short for ANNs, are computing system that is inspired by the way of how biological neural networks works. ANNs composed of a large number of highly interconnected processing elements, called as artificial neuron, which receives signals, process it and transmit the processed signal to the next artificial neurons. A neural network when used for fraud detection, is typically a collection of neuron-like processing units with weighted connections between the units. With the ANNs’ remarkable ability to derive meaning from complicated or imprecise data, it is increasingly proposed as a state-of-the-art way to identify frauds [7].

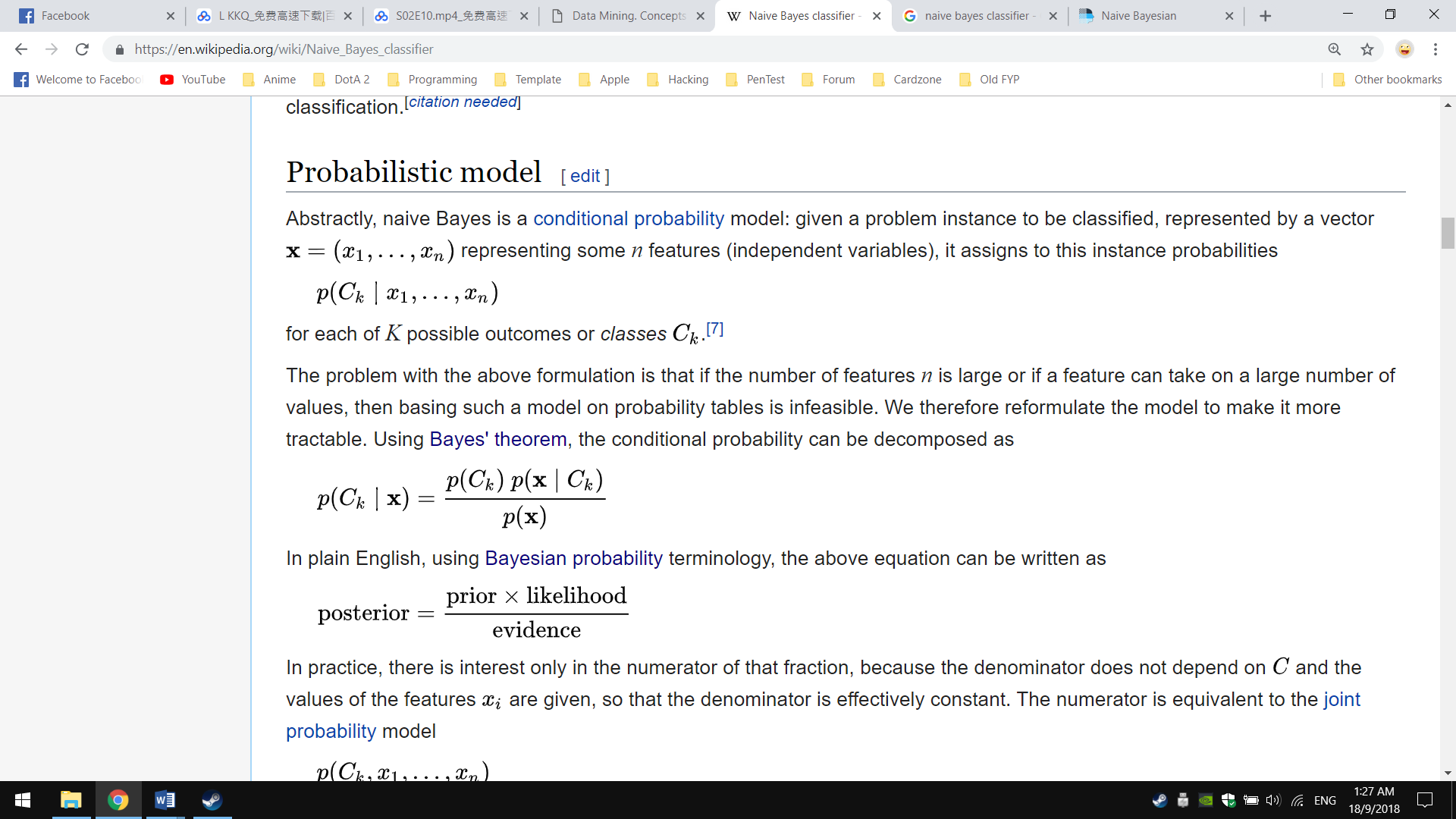


Figure

**Fraud Prevention Techniques, Fraud Scoring. Essentials of Online Payment Security and Fraud Prevention by David Montague (2012):** David Montague (2012) proposed Auto Encoder (AE) as one of the types of neural networks in fraud detection. An AE is divided into two parts: an encoder and a decoder. AE follows a feed-forward ANN architecture, except that the output layer has exactly the same number of neurons as the input layer. The transition between the first and second layer represents the encoder and the transition between the second layer and the third layer represents the decoder. AE will use the input data itself as its target value, and learn the common patterns that shared by the majority of the training data during the construction period. Since frauds will have a different distribution than a normal transaction, for the data points that the error is high and do not conform to those majority transaction patterns, these are the anomalies and might be the fraudulent transactions.

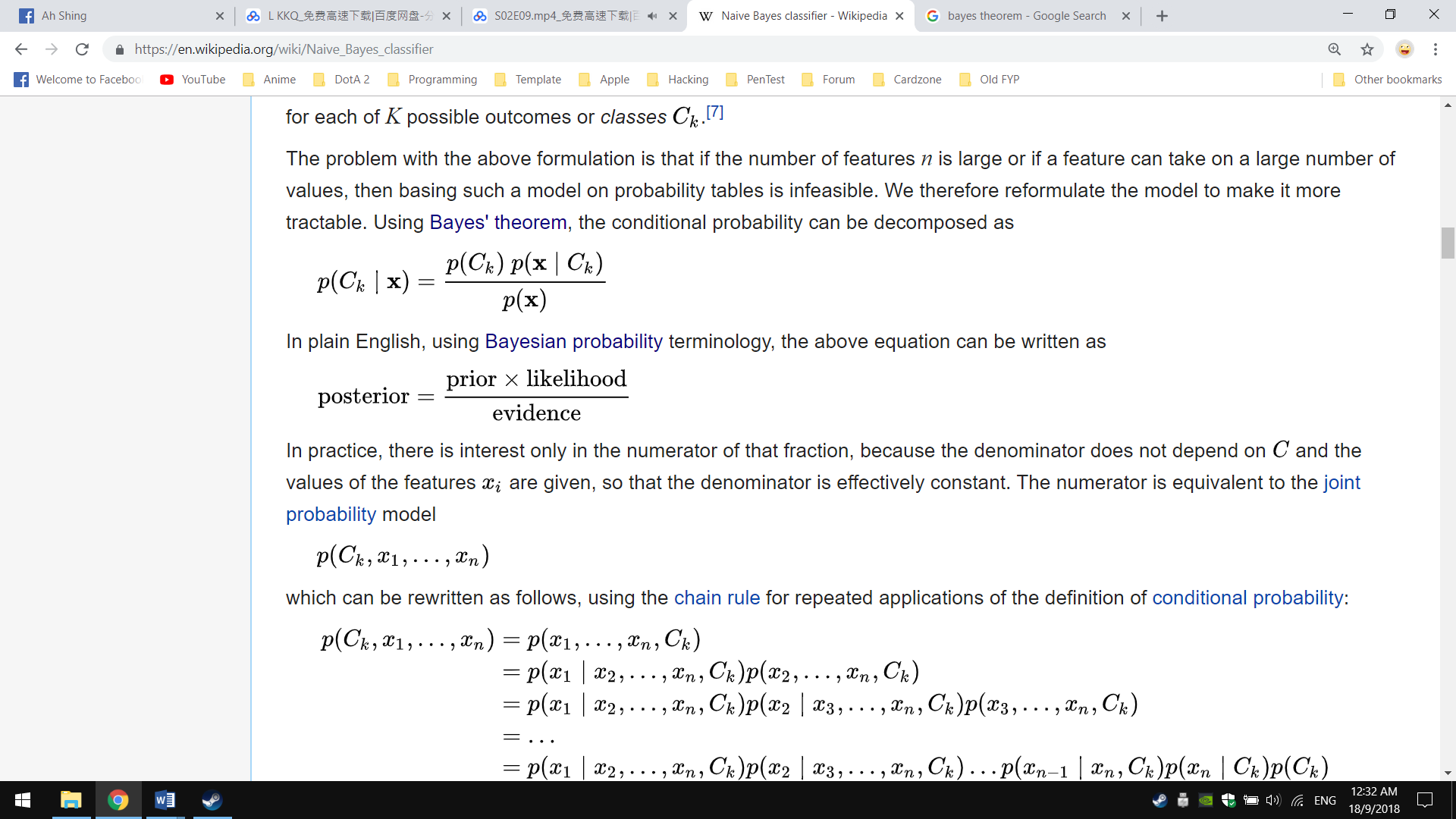
1. ***Naïve Bayesian Classifier***

Naïve Bayesian Classifier, or naïve Bayesian is a supervised Machine Learning method that uses a training dataset with known target classes to predict future or any incoming instance's class value. It can predict the class membership probabilities such as the probability that a given tuple belongs to a particular class. Naïve Bayesian is based on Bayes’ theorem of posterior probability. It assumes class-conditional independence, which means the effect of an attribute value on a given class is independent of the values of the other attributes. It is made to simplify the computations involved and, in this sense, is considered as “naïve”. Bayes theorem provides a way of calculating the posterior probability. A more description can be found in below:

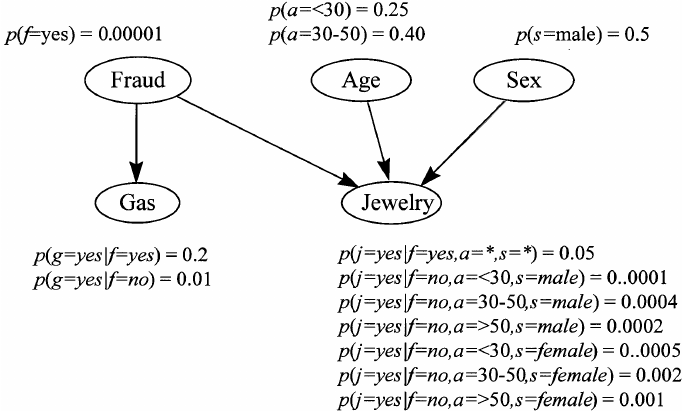


* **P(Ck|x)** is the posterior probability of class (target) given predictor (attribute).
* **P(Ck)** is the prior probability of class.
* **P(x|Ck)** is the likelihood which is the probability of predictor given class.
* **P(x)** is the prior probability of predictor

In plain English, using Bayesian theorem terminology, the above equation can be written as:



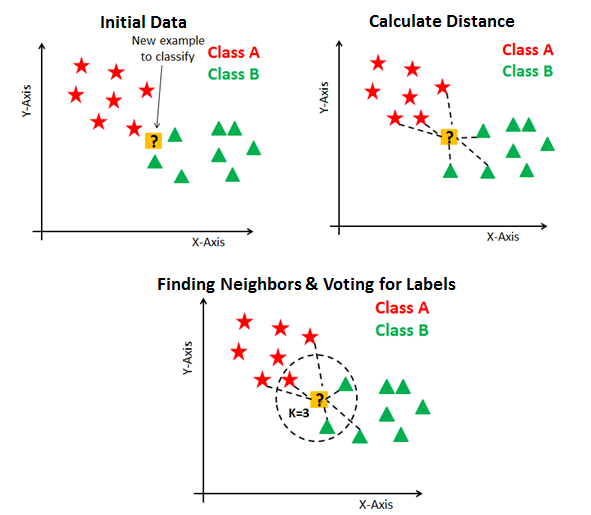
**Using Bayesian Belief Networks for credit card fraud detection by Lev Mukhanov (2008):** Lev Mukhanov (2008) proposed Bayesian Belief Networks (BNNs) to visualize naïve Bayesian classifiers. With BBNs, each node will represent a variable and is associated with the conditional fraud probability that given by its parents. It also represents the dependence between the variables and gives a compact specification of the joint probability distribution. To illustrate the process of building a BBNs in fraud detection, Lev Mukhanov consider the problem of detecting credit card fraud in whether there is a or jewellery gas purchase in the last twenty-four hour. BBNs begins by determining the variables to model and one of the possible choice of variables for the fraud detection would be: Fraud (F), Gas (G), Jewellery (J), Age (A), and Sex (S).



Figure

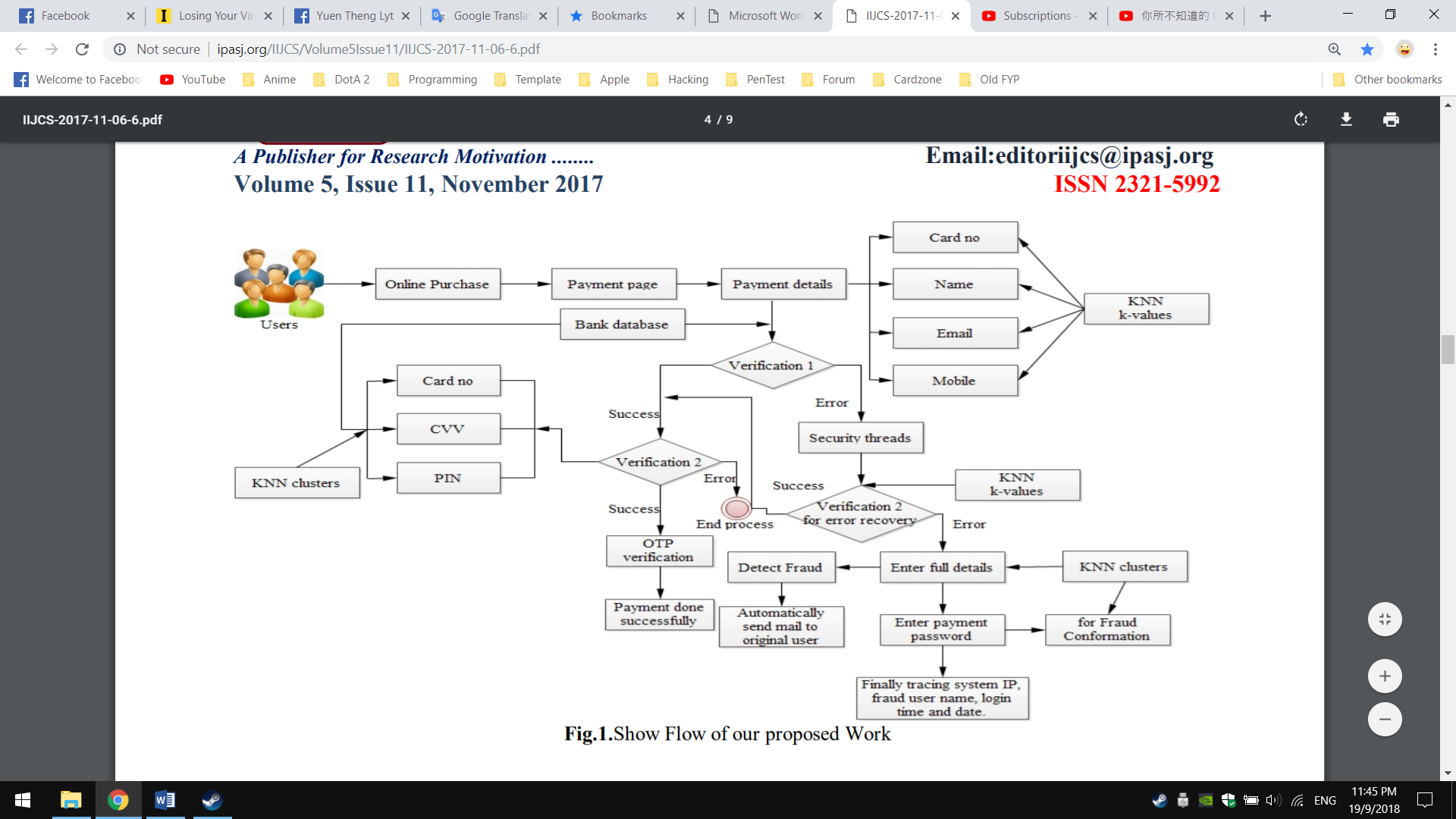
1. ***k-Nearest Neighbour (k-NN)***

k-Nearest Neighbour (k-NN) falls into the supervised Machine Learning family, and is an example of instance-based learning where it categorised objects based on closest feature space in the training set. k-NN is considered as a “lazy” learning algorithm as it does not use the training data points to do any generalization. There is no explicit training or learning phase in k-NN, and the training happens at the time when prediction is requested. New instance is compared with existing ones in the feature space by using Euclidean Distance, and the closest existing instance is used to assign the class to the new one. As k-NN algorithm is based voting scheme, which the most many neighbour is consider as winner and is used to label the query. An example of k-NN in classification are illustrates as in below figure 2.5:



Figure

**Credit Card Fraud Detection in Internet Using K-Nearest Neighbour Algorithm by C. Sudha & T. Nirmal Raj (2017):** C. Sudha & T. Nirmal Raj (2017) presented k-NN as Machine Learning model in helping detecting fraudulent transaction. In the process of k-NN during detecting fraudulent transaction, the model classifies any incoming transaction by calculating most many nearest neighbours to new transaction. If the most neighbours are fraudulent, then the transaction is classified as fraudulent, else, it is classified as legal transaction. They also described that various factors had to be considered before classifying the new transaction. For instance, the financial characteristics of a transaction such as card number, transaction amount, or time since last purchase needs to be collected first before perform classification. With those information, k-NN can then properly classify the new transaction as fraudulence or legitimate.

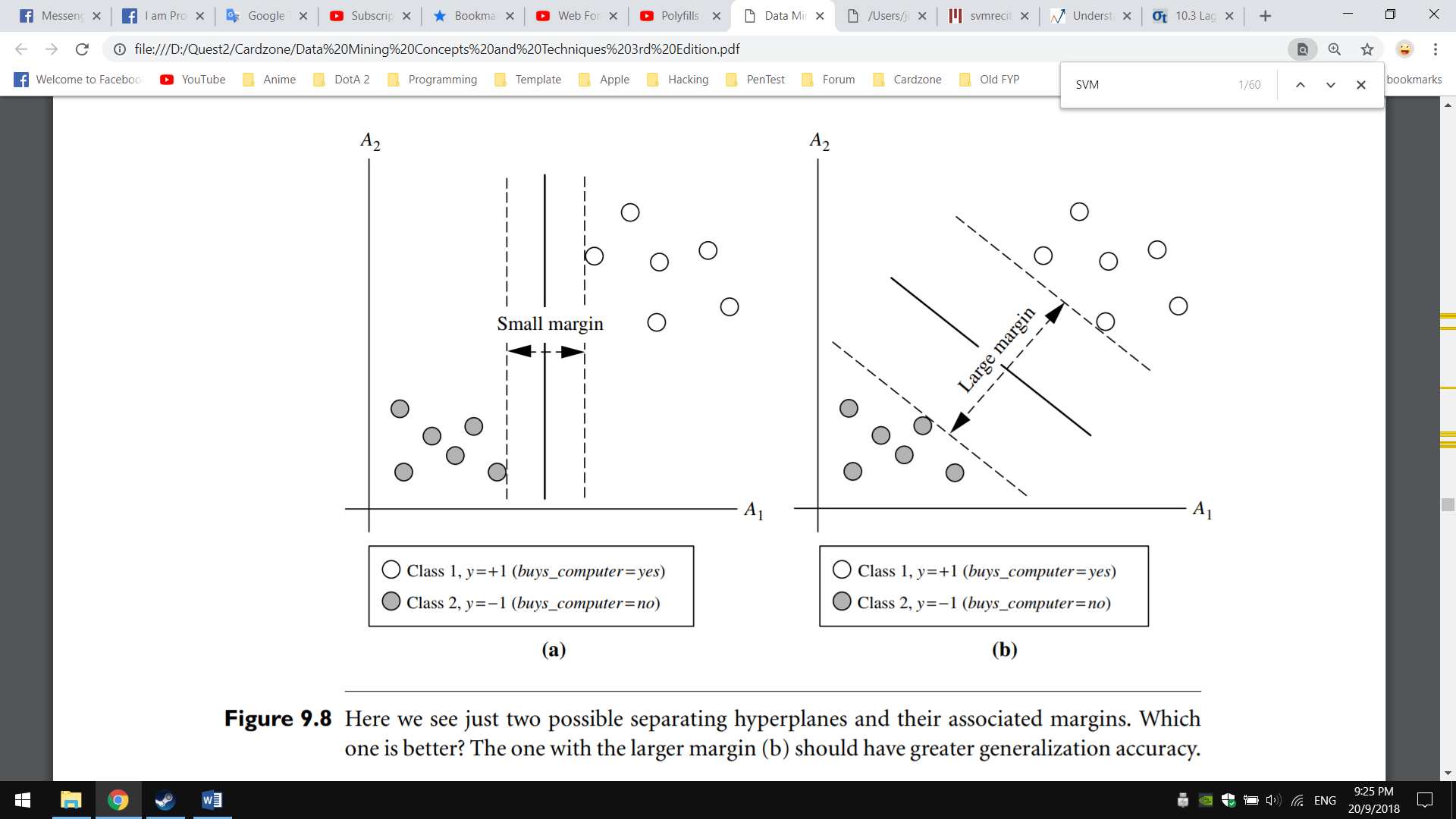


Figure

1. ***Support Vector Machines (SVMs)***

Support Vector Machines (SVMs) is a supervised Machine Learning algorithm that given a set of training samples, each marked as one or other categories, it will assign the best category to the new data. In classification, new data are mapped into the feature space and predicted to belong to which category based on which side of the gap they fall into. The algorithm is based on finding the Hyperplane in the feature space to divide the data points according to their categories or classes. Hyperplane, in human language, is a decision boundary that separates between a set of different classes.

The margin of the Hyperplane is the one that gives the greatest separation between the classes, and is known as maximum-margin Hyperplane. The instances or data points that are nearest to the maximum-margin of the Hyperplane are called Support Vectors, and there is always at least one Support Vector for each class, and often there are more. Figure 2.7 below shows the circle that touches dotted line are known as Support Vector, and a maximum-margin Hyperplane is drawn in between.



Figure

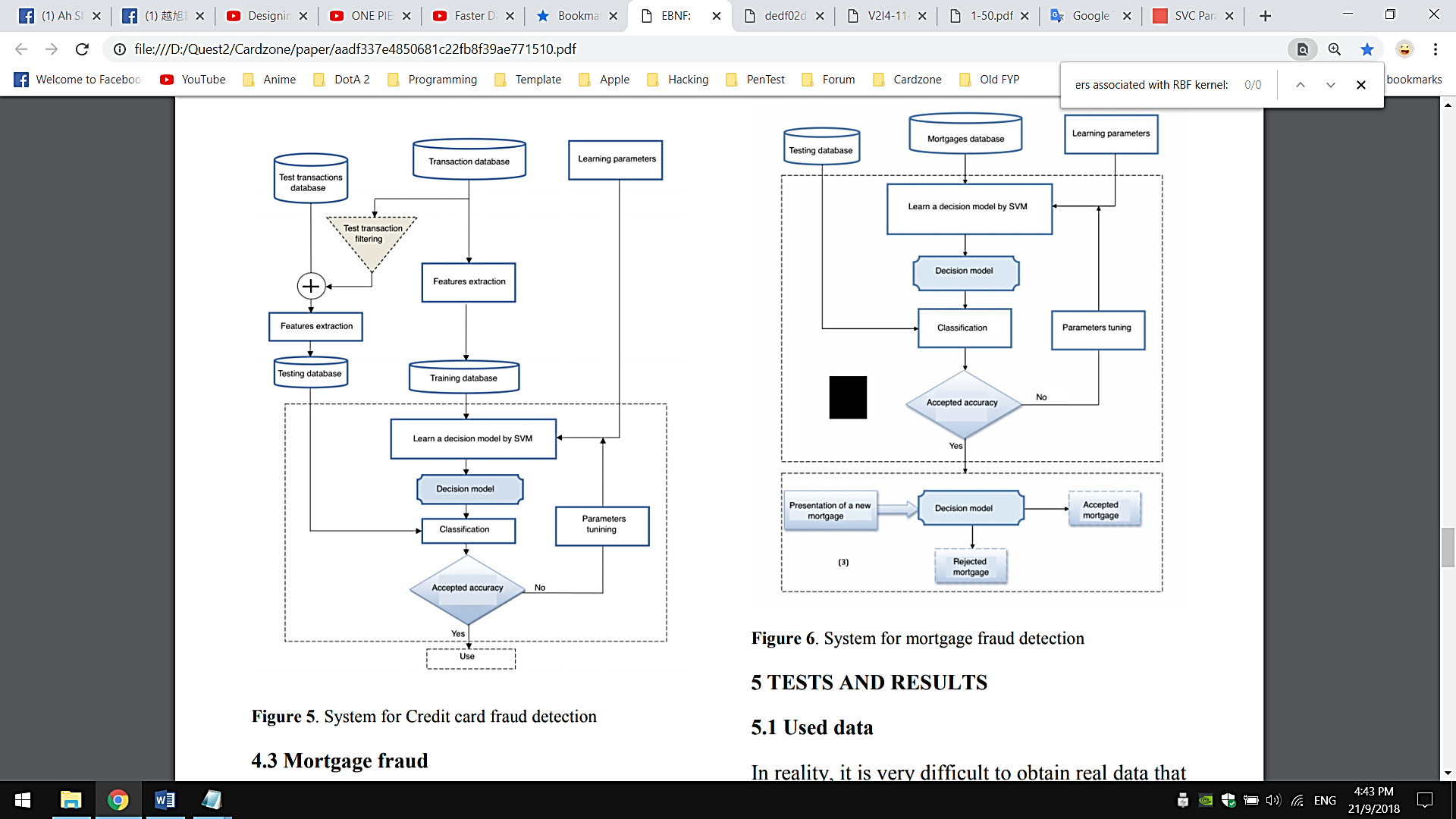
**Automatic Bank Fraud Detection Using Support Vector Machines by Djeffal Abdelhamid, Soltani Khaoula & Ouassaf Atika (2014):** Djeffal Abdelhamid, Soltani Khaoula & Ouassaf Atika (2014) proposed SVMs for credit card fraud detection, which in each test instance of the transaction data, SVMs classifier will used and determine the instance falls into which category. For example, if the instance falls into fraudulent class, it is declared as fraud, else it is declared as legitimate transaction. According to Djeffal Abdelhamid, Soltani Khaoula & Ouassaf Atika, they describe that there are multiple attributes that can used to improve detection rate. However, adding more or irreverent attributes can make the classifier inefficient. The important attributes that listed out in the paper and can be used for detecting fraud in SVM classifier are:

* Customer ID
* Transaction amount
* Frequency of card usage including Date and Time
* Average amount of transactions
* Place

They also utilized a trick which called as Radial Basis Function Kernel, or known as RBF Kernel in helping implicitly map the input vector into the feature space and find the non-linear decision boundary. RBF Kernel able map the and highly imbalance data points from non-linearity to linearity. The RBF Kernel is introduced as K(x,x^' )=exp⁡(-〖||x-x'||〗^2/〖2σ〗^2 ) , where:

* ||x−x′||2 is the squared Euclidean Distance between two data points x and x′
* σ is a standard deviation
* exp stands for exponent
* x and x’ are a vector containing data about a single observation

The next step – cross-validation is performed after running the SVMs classification algorithm with RBF Kernel onto transaction datasets. Cross-validation is executed for assessing how accurately a model will perform in practice. After validation is done, a learned model is then outputted



Figure

1. ***Decision Tree***

Decision Trees falls into supervised Machine Learning family whereby it learns the class-labelled training tuples and try to predicts a class for a given input vector. A Decision Tree is a flowchart-like tree structure, where it typically starts with a single topmost node, which branches into possible nodes. Each of those branched nodes leads to additional internal nodes, which test on an attribute and branch off into other nodes. By continuing branching into more nodes, this eventually gives it a treelike shape. Decision Tree will continue to expand until every branch reaches an endpoint. Meaning that there are no more conditions to consider. The endpoint, or called as lead node will represents the final labels or choices in the Decision Tree.

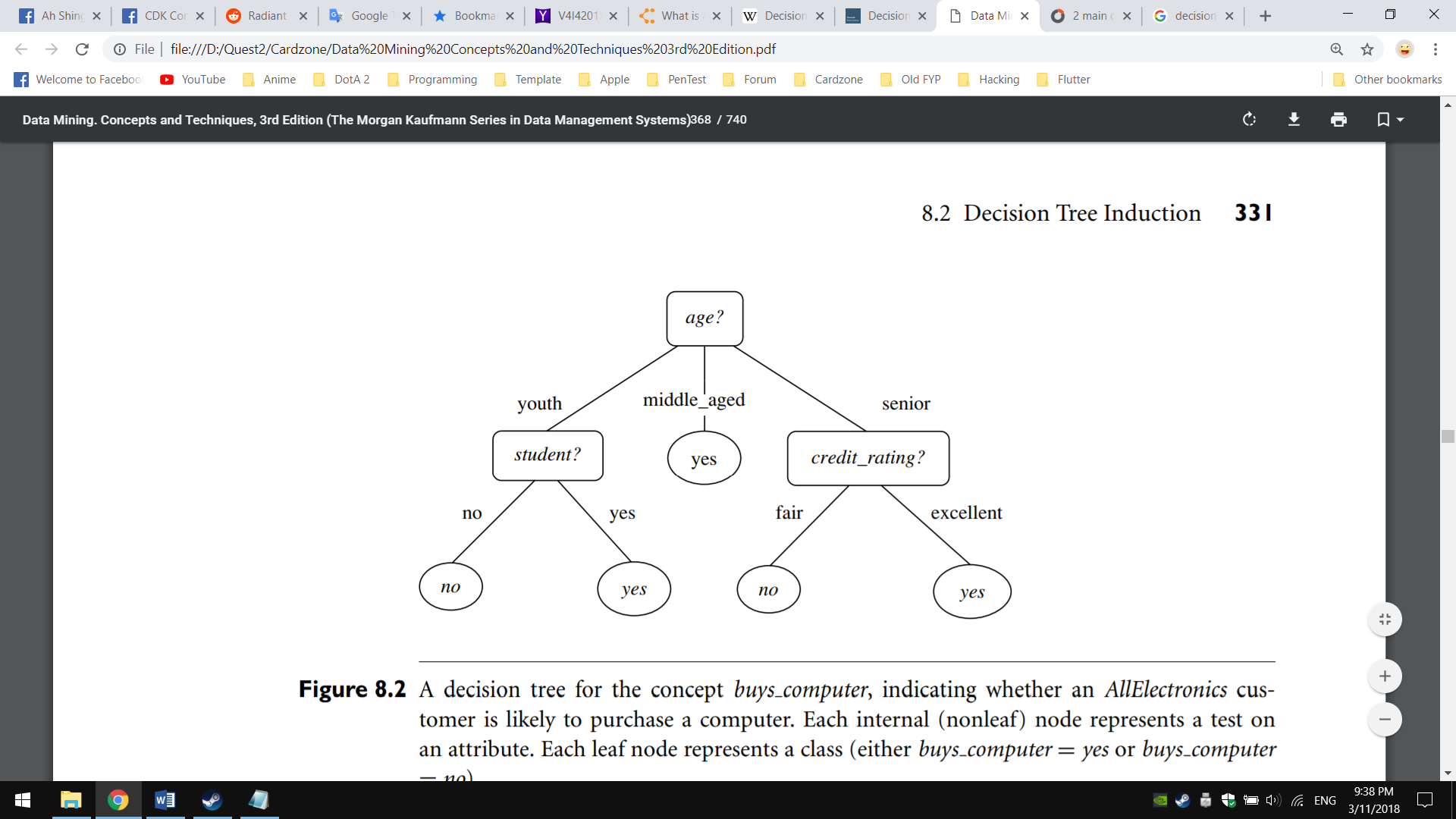


Figure :

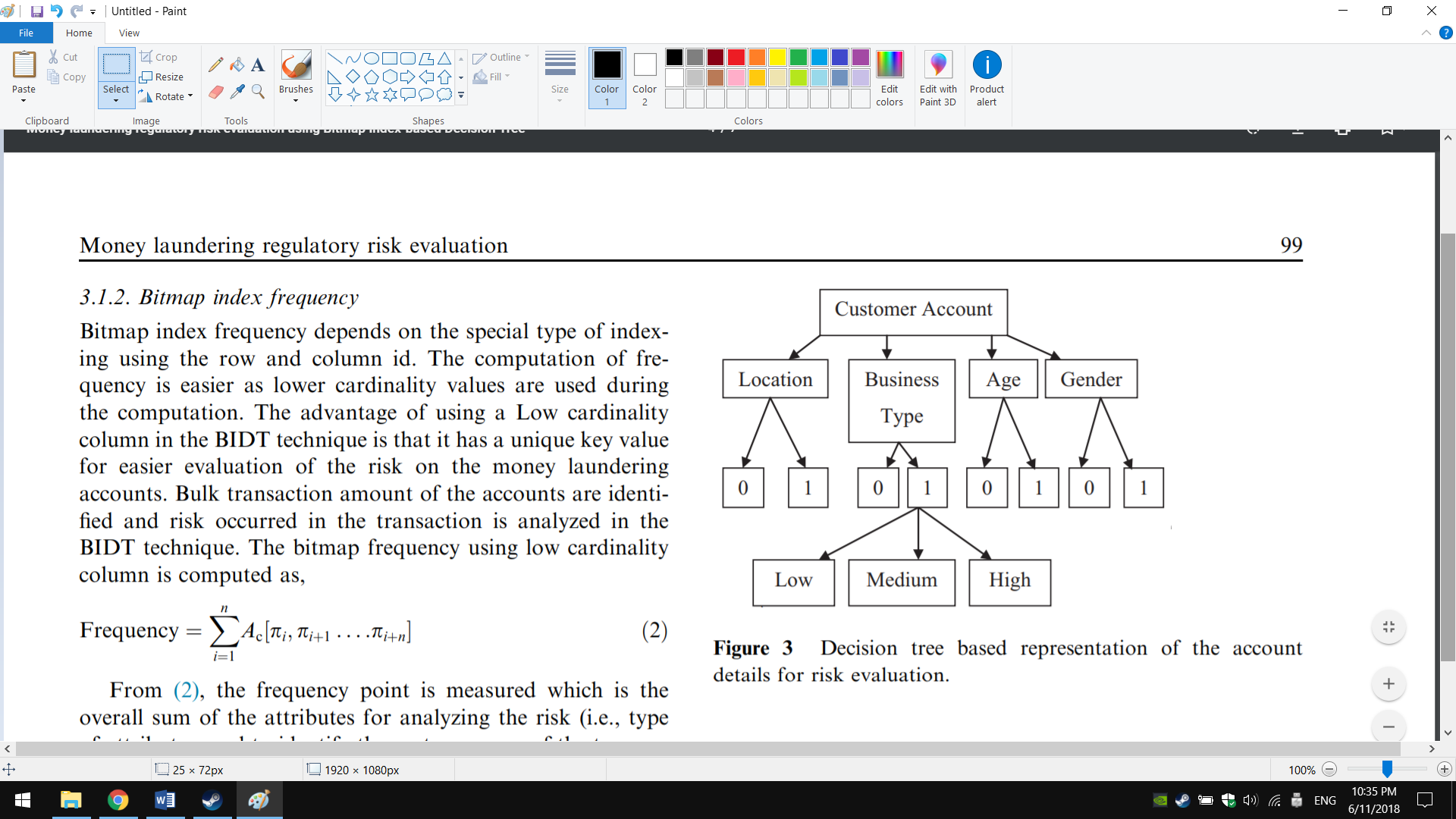
There are three main algorithms in Decision Tree has been devised to not only best classify a new data, but also best building the Decision Tree in an optimal way and generates rules out from it. The growing of Decision Tree algorithms is often related to Decision Tree Learnings which it deals with the construction of an optimal Decision Tree from class-labelled training dataset.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Characteristics (→)** | **Splitting Criteria** | **Attribute type** | **Missing values** | **Pruning Strategy** | **Outlier Detection** |
| **Algorithm (↓)** |
| **ID3** | Information Gain | Handles only categorical value | Did not handle missing values. | No pruning is done | Susceptible on outliers |
| **C4.5** | Gain Ratio | Handles both  categorical and numeric value | Handle missing values. | Error based pruning is used | Susceptible on outliers |
| **CART** | Gini Index | Handles both categorical and numeric value | Handle missing values. | Cost-complexity pruning is used | Can handle outliers well |

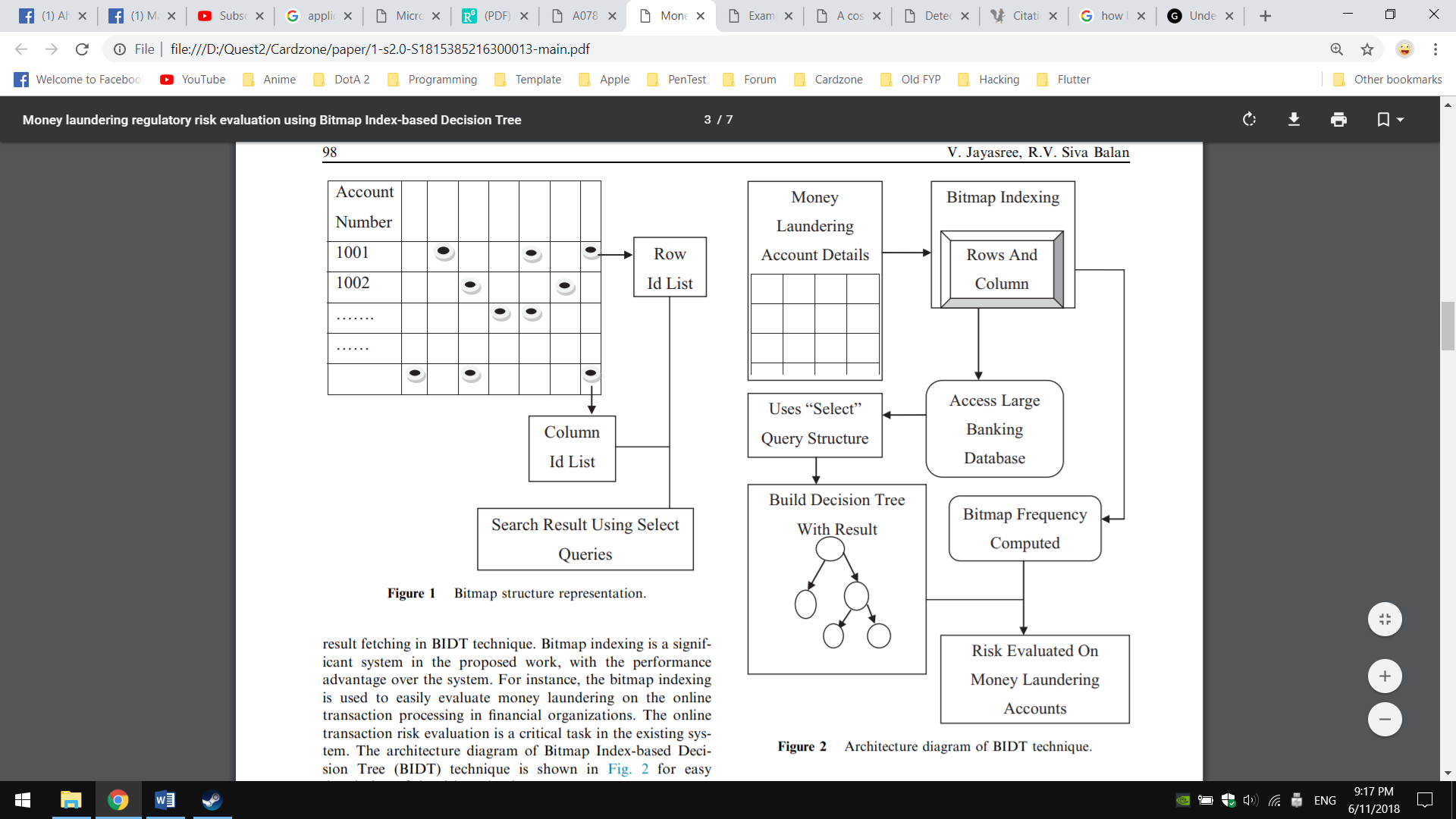
Table

**Money laundering regulatory risk evaluation using Bitmap Index-based Decision Tree by Jayasree and Balan (2017):** Jayasree and Balan (2017) presented fraud detection by using Decision Tree specifically for money laundering in credit card transaction. According to the research conducted by them, they take ID3 algorithm as a base and further improved it with their own algorithm, which named as Advanced Iterative Dichotomiser 3 (AID3). They integrated another method called Bitmap Index in helping getting the best attributes to split, instead of using Information Gain

Bitmap Index works by provide pointers to the rows and columns in the transaction tables, then stores the row ID, column ID and the key values. The attributes are subsequently analysed using the given distinct key value on each transaction for a particular account. The indexing procedure is applied on a dummy banking database that Jayasree and Balan obtained. Bitmap Index in AID3 is carried out by using the “SELECT” query and logical AND operator. The query results are then use to construct the Decision Tree and apply it onto fraud detection application.



Figure

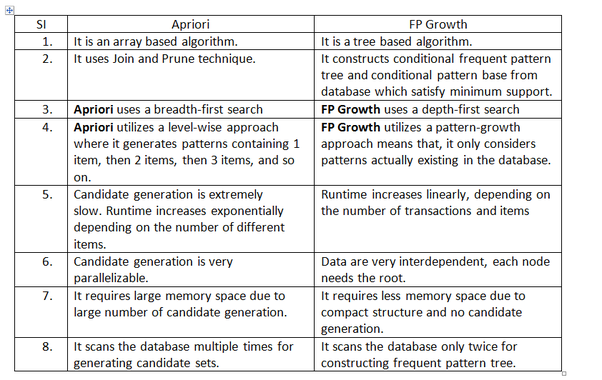


Figure

1. ***Frequent Pattern Mining***

Frequent Pattern Mining is one of the subfields in Data Mining in which it mines frequent sets of items and the interesting patterns in a given itemset. Here, the interesting patterns can be a set of items that appears frequently in a database, or want to discover rare and negative patterns that have not seen in other existing itemset. Application of Frequent Pattern Mining are largely inter-related with Association Rule Learning, whereby it finds all the itemset patterns and then post-process them into rules in helping solving problems. When using Frequent Pattern Mining, it can be stated as: Given a database D with itemset I1 ... In, determine all patterns P that are present in at least a fraction S of the itemset.

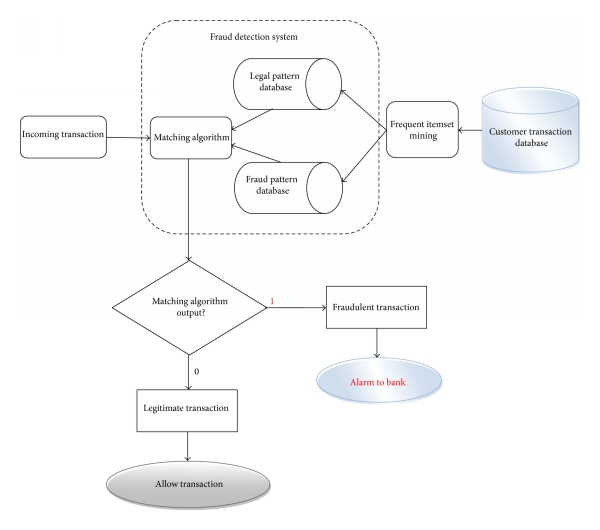
In this credit card fraud detection context, Apriori and FP-growth algorithm will be focused as these two algorithms are often used to detect new fraud patterns [3] [4] and generate those patterns as new rules in order to prevent same method of fraud happening in future. The differences between these two algorithms is described in below table 2.



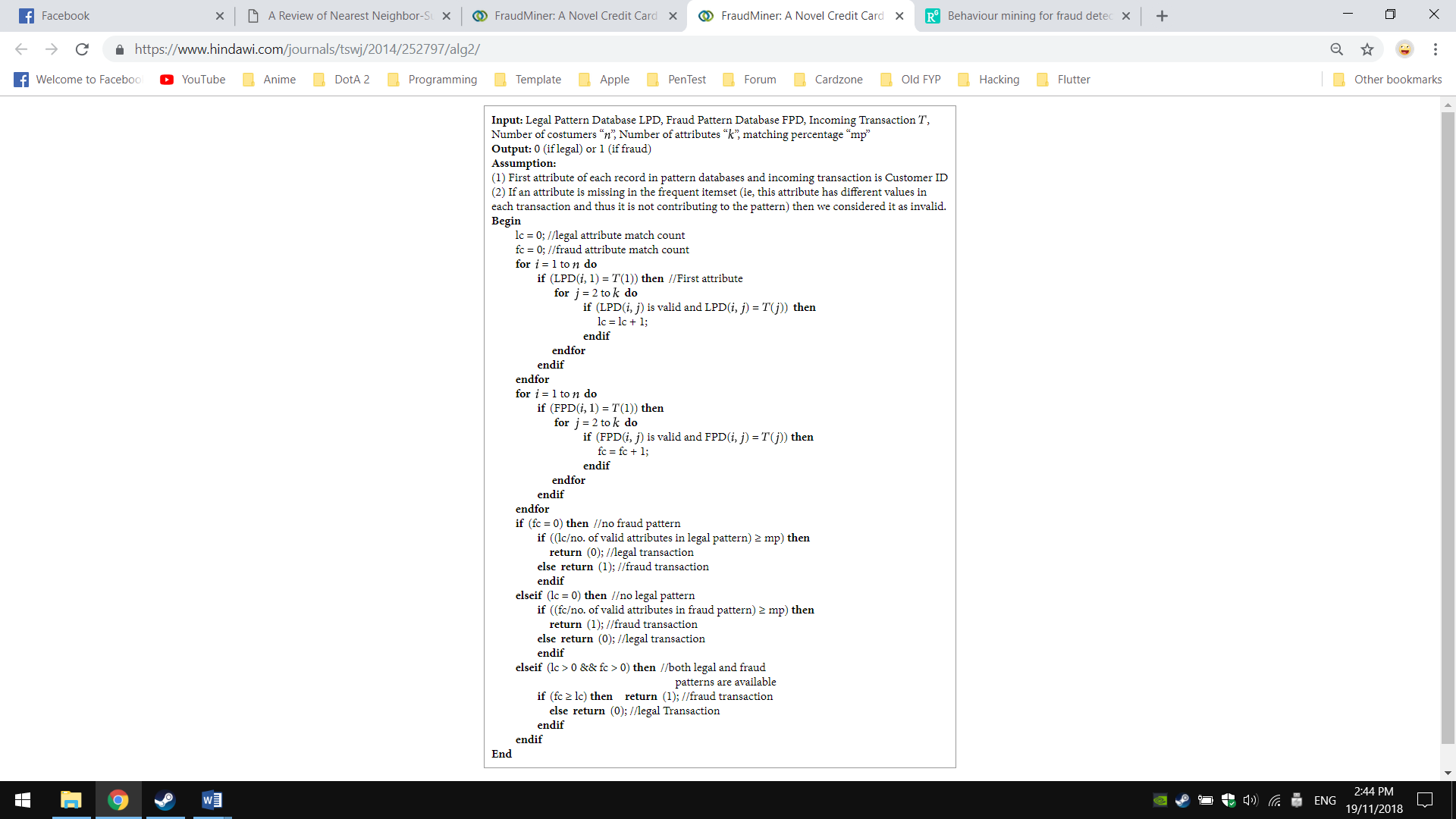
Table

**FraudMiner: A Novel Credit Card Fraud Detection Model Based on Frequent Itemset Mining by K. R. Seeja & Masoumeh Zareapoor (2014):** K. R. Seeja & Masoumeh Zareapoor (2014) utlized Apriori as the main Frequent Pattern Mining technique in mining the patterns from a database. According to them, they described that instead of finding patterns for fraudster behaviour, Apriori algorithm is use in identify buying patterns for fraud and legal transaction. They had set the MinSup as 0.9 and constructed two patterns for each customer – legal pattern and fraud pattern.

After finding the legal and fraud patterns and stored in the database, the Fraud Detection System then traverses these pattern databases in order to detect frauds. K. R. Seeja & Masoumeh Zareapoor developed a matching algorithm, which traverses the pattern databases for matching the incoming transaction in detecting fraud. If a closer match is found with legal pattern of the corresponding customer, then the matching algorithm returns “0” giving a green signal to the bank for allowing the transaction, else it returns as “1”, giving an alarm to the bank that this might be fraudulent transaction



Figure



Figure

1. **EXPERIMENTAL RESULT**
2. **DISCUSSION AND CONCLUSION**

Semantically representation of documents is the challeng-ing area for research in text mining. By proper implanta-tion of this will be improve the classification and the in-formation retrieval process.

1. MACHINE LEARNING TECHNIQUES

The documents can be classified by three ways, un-supervised, supervised and semi supervised methods. Many techniques and algorithms are proposed recently for the clustering and classification of electronic docu-ments. This section focused on the supervised classifica-tion techniques, new developments and highlighted some of the opportunities and challenges using the existing literature. The automatic classification of documents into predefined categories has observed as an active attention, as the internet usage rate has quickly enlarged. From last few years , the task of automatic text classification have been extensively studied and rapid progress seems in this area, including the machine learning approaches such as Bayesian classifier, Decision Tree, K-nearest neigh-bor(KNN), Support Vector Machines(SVMs), Neural Networks, Latent Semantic Analysis, Rocchio’s Algo-rithm, Fuzzy Correlation and Genetic Algorithms etc. Normally supervised learning techniques are used for automatic text classification, where pre-defined category labels are assigned to documents based on the likelihood suggested by a training set of labelled documents. Some of these techniques are described below.

*P*(*D*)

SVM with its outperformed classification effectiveness. Therefore, many active researches have been carried out to clarify the reasons that the naïve Bayes classifier fails in classification tasks and enhance the traditional ap-proaches by implementing some effective and efficient techniques [100] [102] [103] [104] [105].

architectures [106] [107]. These elements, namely artifi-cial neuron are interconnected into group using a mathe-matical model for information processing based on a con-nectionist approach to computation. The neural networks make their neuron sensitive to store item. It can be used for distortion tolerant storing of a large number of cases represented by high dimensional vectors.

*P*(*ci* | *D*)= *P*(*ci* )*P*(*D* | *ci* )

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  | *n* | |
|  | *P*(*D* | *ci* )=∏*P*(*d j* | *ci* ) | | | | |
|  |  |  |  | *j*=1 | |
| Where P(Ci)= | *P*(*C* = *ci* )= | *Ni* | |  | |
|  | *N* | | |
|  |  |  |
| and P(dj|ci) = | *P*(*d j* | *ci* )= |  | 1+ *N* *ji* | | |
|  |  | *M* |  |
|  |  |  | *M* + ∑ *Nki* | | |

*k* =1

1. Different types of neural network approaches have been implemented to document classification tasks. Some of the researches use the single-layer perceptron, which con-tains only an input layer and an output layer due to its

simplicity of implementing [108]. Inputs are fed directly

(3) to the outputs via a series of weights. In this way it can be considered the simplest kind of feed-forward network. The multi-layer perceptron which is more sophisticated, which consists of an input layer, one or more hidden lay-ers, and an output layer in its structure, also widely im-plemented for classification tasks [106].



Naïve Bayes has been one of the popular machine learn-ing methods for many years. Its simplicity makes the framework attractive in various tasks and reasonable per-formances are obtained in the tasks although this learning is based on an unrealistic independence assumption. For this reason, there also have been many interesting works of investigating naive Bayes. Recently the [83] shows very good results by selecting Naïve Bayes with SVM for text classification also the authors in [84] prove that Naive Bayes with SOM give very good results in cluster-ing the documents. The authors in [85] propose a Poisson Naive Bayes text classification model with weight-enhancing method, and shows that the new model as-sumes that a document is generated by a multivariate Poisson model. They suggest per-document term fre-quency normalization to estimate the Poisson parameter, while the traditional multinomial classifier estimates its parameters by considering all the training documents as a unique huge training document. The [86] presented that naive Bayes can perform surprisingly well in the classifi-cation tasks where the probability itself calculated by the naive Bayes is not important. The authors in a review

1. described that researcher shows great interest in naïve Bayes classifier for spam filtering. So this tech-nique is most widely used in email, web contents, and spam categorization.

Naive Bayes work well on numeric and textual data, easy to implement and computation comparing with other al-gorithms, however conditional independence assumption is violated by real-world data and perform very poorly when features are highly correlated and does not consider frequency of word occurrences.

1. *Artificial Neural Network*

Artificial neural networks are constructed from a large number of elements with an input fan order of magni-tudes larger than in computational elements of traditional

Fig. 5 Artificial Neural Network

The main advantage of the implementation of artificial neural network in classification tasks is the ability in han-dling documents with high-dimensional features, and documents with noisy and contradictory data. Further-more, linear speed up in the matching process with re-spect of the large number of computational elements is provided by a computing architecture which is inherently parallel, where each element can compare its input value against the value of stored cases independently from oth-ers [107].

The drawback of the artificial neural networks is their high computing cost which consumes high CPU and physical memory usage. Another disadvantage is that the artificial neural networks are extremely difficult to under-stand for average users. This may negatively influence the acceptance of these methods.

In recent years, neural network has been applied in doc-ument classification systems to improve efficiency. Text categorization models using back-propagation neural network (BPNN) and modified back-propagation neural network (MBPNN) are proposed in [54] for documents classification. An efficient feature selection method is used to reduce the dimensionality as well as improve the performance. New Neural network based document clas-sification method [68], was presented, which is helpful for companies to manage patent documents more effec-tively.

The ANN can get Inputs xi arrives through pre-synaptic connections, Synaptic efficacy is modelled using real

12 JOURNAL OF ADVANCES IN INFORMATION TECHNOLOGY, VOL. 1, NO. 1, FEBRUARY 2010

weights wi and the response of the neuron is a nonlinear function f of its weighted inputs.

The output from neuron j for pattern p is Opj where

netic algorithm. In the experimental analysis, they show that the improved method is feasible and effective for text classification.

1. *Support Vector Machine (SVM)*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *O pj* | ( *net* *j* ) = |  |  |  | 1 |  |
|  | + | *e* | − *λ* *net* | *j* |
|  | 1 | |  |  |

and

*net j* = *bias* \**Wbias* +∑ *O pkW jk*

*k*

Support vector machines (SVMs) are one of the discri-

1. minative classification methods which are commonly recognized to be more accurate. The SVM classification method is based on the Structural Risk Minimization principle from computational learning theory [109]. The idea of this principle is to find a hypothesis to guarantee
2. the lowest true error. Besides, the SVM are well-founded that very open to theoretical understanding and analysis [110].

Neural network for document classification produce good results in complex domains and suitable for both discrete and continuous data (especially better for the continuous domain). Testing is very fast however training is relative-ly slow and learned results are difficult for users to interp-ret than learned rules (comparing with Decision tree), Empirical Risk Minimization (ERM) makes ANN try to minimize training error, may lead to overfitting.

1. *Fuzzy correlation*

Fuzzy correlation can deal with fuzzy information or in-complete data, and also convert the property value into fuzzy sets for multiple document classification [69].

In [55] the authors explores the challenges of multi-class text categorization using one-against-one fuzzy support vector machine with Reuter’s news as the example data, and shows better results using one-against-one fuzzy sup-port vector machine as a new technique when compare with one-against-one support vector machine. [61] pre-sented the improvement of decision rule and design a new algorithm of f-k-NN (fuzzy k-NN) to improve categoriza-tion performance when the class distribution is uneven, and show that the new method is more effective. So the researchers shows great interest recently to use the fuzzy rules and sets to improve the classification accuracy, by incorporating the fuzzy correlation or fuzzy logic with the machine learning algorithm and the feature selection me-thods to improve the classification process.

1. *Genetic Algorithm*

Genetic algorithm [81] aims to find optimum characteris-tic parameters using the mechanisms of genetic evolution and survival of the fittest in natural selection. Genetic algorithms make it possible to remove misleading judg-ments in the algorithms and improve the accuracy of doc-ument classification. This is an adaptive probability glob-al optimization algorithm, which simulated in a natural environment of biological and genetic evolution, and is widely used for their simplicity and strength. Now several researchers used this method for the improvement of the text classification process. In authors in [82] introduced the genetic algorithm to text categorization and used to build and optimize the user template, and also introduced simulated annealing to improve the shortcomings of ge-

The SVM need both positive and negative training set which are uncommon for other classification methods. These positive and negative training set are needed for the SVM to seek for the decision surface that best sepa-rates the positive from the negative data in the n-dimensional space, so called the hyper plane. The docu-ment representatives which are closest to the decision surface are called the support vector. The performance of the SVM classification remains unchanged if documents that do not belong to the support vectors are removed from the set of training data [99].

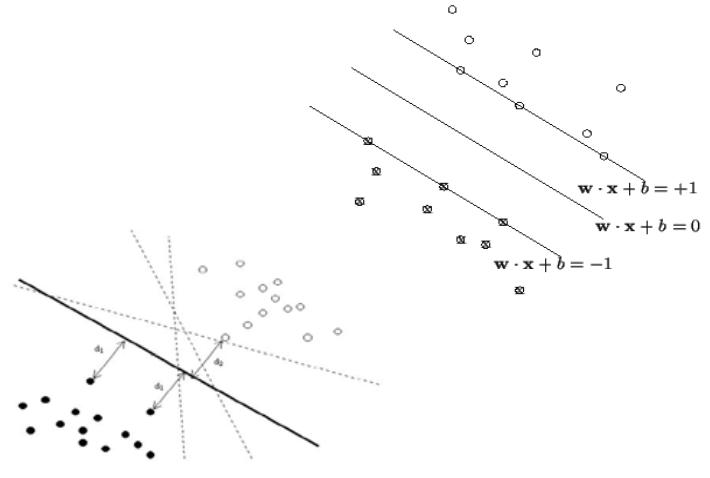
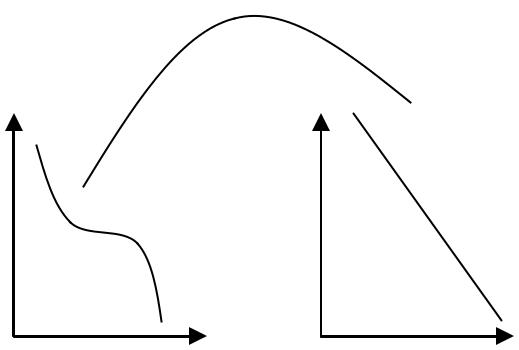


Fig. 6 Illustration of optimal separating hyper plane, hyper planes and support vectors

X



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 0 x |  |  |  |  | f | f |
|  |  | f |  |  |
| x |  |  |  | f |
|  | x | f |  |  |
| 0 |  |  |  |
|  |  |  |  |
|  | 0 |  | f |  | f |
|  |  |  |  |
|  |  | x |  |  | f |
|  |  |  |  |  |
|  |  | 0 |  |  |  |  |
|  |  | X |  |  |  | F |

Fig. 7 Mapping non linear input space onto high dimensional space

choices in pre-processing (stemming, etc.), indexing, di-mensionality reduction and classifier parameter values etc.

A performance compression in [115] presented a con-trolled study on a large number of filter feature selection methods for text classification. Over 100 variants of five major feature selection criteria were examined using four well-known classification algorithms: Naive Bayesian (NB) approach, Rocchio-style classifier, k-NN method and SVM system. Two benchmark collections were cho-sen as the testbeds: Reuters-21578 and small portion of Reuters Corpus Version 1 (RCV1), making the new re-sults comparable to published results. They presents that feature selection methods based on χ2 statistics consis-tently outperformed those based on other criteria (includ-ing information gain) for all four classifiers and both data collections, and that a further increase in performance was obtained by combining uncorrelated and high-performing feature selection methods. The results they obtained using only 3% of the available features are among the best reported, including results obtained with the full feature set. The empirical results of their study suggest that using filter methods which include the χ2 statistic, combining them with DF or IG, and eliminating the rare words. Such methods were consistently better.

In [116] the authors discussed, that some studies com-pared feature selection techniques or feature space trans-formation whereas some others compared the perfor-mance of different algorithms. Recently the rising interest towards the Support Vector Machine, various studies showed that SVM outperforms then other classification algorithms. So should we just not problem about other classification algorithms and opt always for SVM? They have decided to investigate this issue and compared SVM to k-NN and naive Bayes on binary classification tasks. An important issue is to compare optimized versions of these algorithms; from their results it shows all the clas-sifiers achieved comparable performance on most prob-lems. One surprising result is that SVM was not a clear winner, despite quite good overall performance. If a suit-able pre-processing is used with k-NN, this algorithm continues to achieve very good results and scales up well with the number of documents, which is not the case for SVM. As for Naive Bayes, it also achieved good perfor-mance.

The [117] deals with the performance of different classi-fication algorithms and the impact of feature selection algorithm on Logistic Regression Classifier, How it con-trols False Discovery Rate (FDR) and thus improves the efficiency of Logistic Regression classifier. As per the analysis support vector machine has more parameters than logistics regression and decision tree classifier, SVM has the highest classification precision most of the time, however SVM is very time consuming because of more parameters, demands more computation time. Compared to SVM, logistic regression is computationally efficient. Its results usually have static meaning. However it does not perform well when data set exhibits explicit data structures.

In [118] compression on four machine learning algo-rithms, which are Naive Bayesian (NB), neural network (NN), support vector machine (SVM) and relevance vec-tor machine ( RVM), are proposed for spam classification. An empirical evaluation for them on the benchmark spam filtering corpora is presented. The experiments are per-formed based on different training set size and extracted feature size. Experimental results show that NN classifier is unsuitable for using alone as a spam rejection tool. Generally, the performances of SVM and RVM classifi-ers are obviously superior to NB classifier. Compared with SVM, RVM is shown to provide the similar classifi-cation result with less relevance vectors and much faster testing time despite the slower learning procedure, they show that RVM is more suitable than SVM for spam classification in terms of the applications that require low complexity.

In [119] email data was classified using four different classifiers (Neural Network, SVM classifier, Naïve Baye-sian Classifier, and J48 classifier). The experiment was performed based on different data size and different fea-ture size. The final classification result should be ‘1’ if it is finally spam, otherwise, it should be ‘0’. This paper shows that simple J48 classifier which make a binary tree, could be efficient for the dataset which could be classi-fied as binary tree.

The [120] shows that two main research areas in statistic-al text categorization are: similarity-based learning algo-rithms and associated thresholding strategies. The combi-nation of these techniques significantly influences the overall performance of text categorization. After investi-gating two similarity-based classifiers (k-NN and Roc-chio) and three common thresholding techniques (RCut, PCut, and SCut), they described a new learning algorithm known as the keyword association network (KAN) and a new thresholding strategy (RinSCut) to improve perfor-mance over existing techniques. Extensive experiments have been conducted on the Reuters-21578 and 20-Newsgroups data sets, and shows that the new approaches give better results.

Comparing with ANN, SVM capture the inherent charac-teristics of the data better and embedding the Structural Risk Minimization (SRM) principle which minimizes the upper bound on the generalization error (better than the Empirical Risk Minimization principle) also ability to learn can be independent of the dimensionality of the feature space and global minima vs. local minima, How-ever there are some difficulties in parameter tuning and kernel selection.

VI DISCUSSION AND CONCLUSIONS

This paper provides a review of machine learning ap-proaches and documents representation techniques. An analysis of feature selection methods and classification algorithms were presented. It was verified from the study that information Gain and Chi square statistics are the most commonly used and well performed methods for feature selection, however many other FS methods are

proposed as single or hybrid technique recently, shown good results, and needs more exploration for efficient classification process. Several algorithms or combination of algorithms as hybrid approaches was proposed for the automatic classification of documents, among these algo-rithms, SVM, NB and kNN classifiers are shown most appropriate in the existing literature.

Most researchers in text classification assume the docu-ments representation as a Bag of Word (BOG), although according to [44] the statistical techniques are not suffi-cient for the text mining. Text representation is a crucial issue. Most of the literature gives the statistical of syntac-tic solution for the text representation. However the re-presentation model depend on the informational that we require. Concept base or semantically representation of documents requires more research. Better classification will be performed when consider the semantic under con-siderations, semantically and ontology base documents representation opportunities were discussed in this paper. With the addition of the ontology and semantic to represent the documents will be more improve accuracy and the classification process. So the identification of features that capture semantic content is one of the impor-tant areas for research. The general multiple learning is-sues in the presence of noise is a tremendously challeng-ing problem that is just now being formulated and will likely require more work in order to successfully develop strategies to find the underlying nature of the manifold.

Several algorithms or combination of algorithms as hybr-id approaches were proposed for the automatics classifi-cation of documents. Among these algorithms, SVM, NB

* kNN and their hybrid system with the combination of different other algorithms and feature selection tech-niques are shown most appropriate in the existing litera-ture. However the NB is perform well in spam filtering and email categorization, requires a small amount of training data to estimate the parameters necessary for classification. Naive Bayes works well on numeric and textual data, easy to implement comparing with other algorithms, however conditional independence assump-tion is violated by real-world data and perform very poor-ly when features are highly correlated and does not con-sider frequency of word occurrences.

SVM classifier has been recognized as one of the most effective text classification method in the comparisons of supervised machine learning algorithms [74]. SVM cap-ture the inherent characteristics of the data better and em-bedding the Structural Risk Minimization (SRM) prin-ciple which minimizes the upper bound on the generaliza-tion error (better than the Empirical Risk Minimization principle) also ability to learn can be independent of the dimensionality of the feature space and global minima vs. local minima, however, the SVM has been found some difficulties in parameter tuning and kernel selection.

If a suitable pre-processing is used with k-NN, then this algorithm continues to achieve very good results and scales up well with the number of documents, which is

not the case for SVM [122] [123]. As for naive Bayes, it also achieved good performance with suitable pre-processing. k-NN algorithm performed well as more local characteristic of documents are considered, however the classification time is long and difficult to find optimal value of k.

More works are required for the performance improve-ment and accuracy of the documents classification process. New methods and solutions are required for use-ful knowledge from the increasing volume of electronics documents. The following are the some of opportunities of the unstructured data classification and knowledge discovery.

* To improve and explore the feature selection methods for better classification process.
* To reduce the training and testing time of classifier and improve the classification accuracy, precision and re-call.
* For Spam filtering and e-mail categorization the user may have folders like electronic bills, e-mail from family, friends and so on, and may want a classifier to classify each incoming e-mail that’s automatically move it to the appropriate folder. It is easier to find messages in sorted folders in a very large inbox.
* Automatic allocation of folders to the downloaded ar-ticles, documents from text editors and from grid net-work.
* The use of semantics and ontology for the documents classification and informational retrieval.
* Mining trend, i.e. marketing, business, and financial trend (stock exchange trend) form e-documents (Online news, stories, views and events).
* Stream text quire some new techniques and methods for information management.
* Automatic classification and analysis of sentiment, views and extraction knowledge from it. The senti-ments and opinion mining is the new active area of text mining.
* Classification and clustering of semi-structured docu-ments have some challenges and new opportunities.
* An implementation of sense-based text classification procedure is needed for recovering the senses from the words used in a specific context.
* Informational extraction of useful knowledge from e-documents and Web pages, such as products and search results to get meaning full patterns.
* To identify or match semantically similar data from the web (that contain huge amount of data and each web-site represents similar information differently) is an im-portant problem with many practical applications. So web information, integration and schema matching needs more exploration.

REFERENCES

1. Marmo, R. (2013). Data Mining for Fraud Detection System. Encyclopedia of Data Warehousing and Mining, Second Edition,411-416. doi:10.4018/978-1-60566-010-3.ch065
2. Analysis of Credit Card Fraud Detection Techniques. (2016). International Journal of Science and Research (IJSR), 5(3), 1302-1307. doi:10.21275/v5i3.nov162099
3. J. Han, H. Cheng, D. Xin, X. Yan (2007). Frequent Pattern Mining: Current Status and Future Directions. 2007 Data Mining and Knowledge Discovery archive. Vol. 15 Issue 1, pp. 55 – 86
4. Ordonez, C., & Zhao, K. (2011). Evaluating association rules and decision trees to predict multiple target attributes. Intelligent Data Analysis, 15(2), 173–192.
5. Excell, D. (2012). Bayesian inference – the future of online fraud protection. Computer Fraud & Security, 2012(2), 8-11. doi:10.1016/s1361-3723(12)70018-0
6. Xu, Sung & Liu (2007). Behaviour Mining for Fraud Detection. Journal of Research and Practice in Information Technology, Vol. 39, No. 1, February 2007
7. Tom Sweers. (2018). Autoencoding Credit Card Fraud. Radboud University, Bachelor Thesis Computer Science. June 26, 2018