BIG DATA ANALYTICS AS AN OPPORTUNITY FOR FRAUD MANAGEMENT IN THE INSURANCE SECTOR

A Hybrid Machine Learning Approach in medical insurance

By

Margaret Oyugi and Dr.John Olukuru (Both from Kenya)

This paper is compiled for presentation at the International Congress of Actuaries (ICA 2018) in berlin, Germany

Table of Contents

List of Figures

List of Abbreviations

ANN - Artificial Neural Network

MLP - Multilayer Perceptron

NN - Neural Networks

TCP - Protocol

IP - Internet Protocol

MIP - Medical Insurance Provider

MSPs - Medical Service Providers

IRA - Insurance Regulatory Authority

SOFM - Self-Organizing Feature Map

SVM - Support Vector Machine

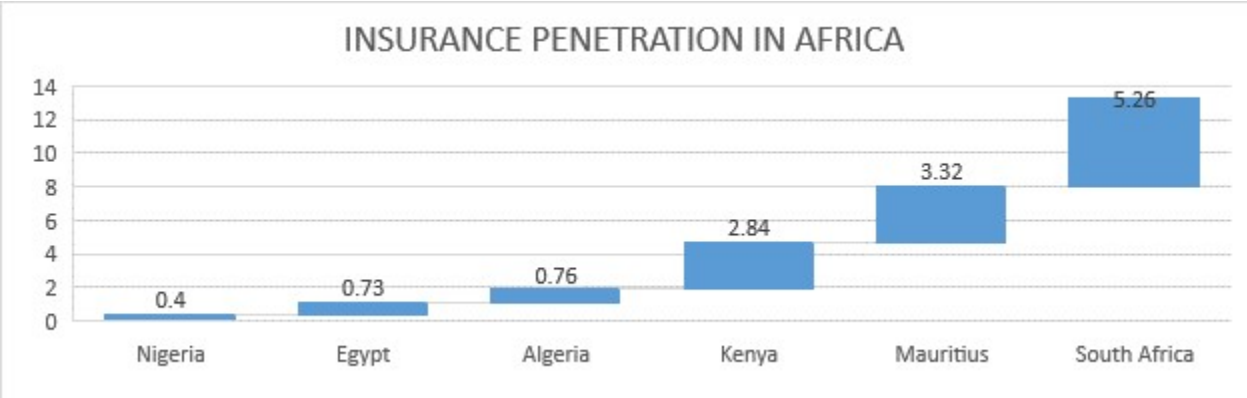
CHAPTER 1: INTRODUCTION

1.1. Background

1.1.1. Insurance Industry in Africa

The growing share of the insurance sector in the aggregate financial sector in almost every developing and developed country has shifted attention to insurance-growth nexus. The insurance markets have witnessed a tremendous growth in the last decade with the world premium in US dollars increased by 175% between the year 2000 and 2008, significantly outpacing worldwide economic growth (Outreville, 2011Outreville, J.F., (2011). World insurance premiums rose by 6 percent from $4.3 trillion in 2010 to $4.57 trillion in 2011, according to Swiss Reinsurance Company (2015 Swiss Reinsurance Company, (2015). Zurich: Sigma Publication, No. 5/2006.) latest study of world insurance. It rose by 2.4 percent from $4.57 trillion in 2011 to $4.61 trillion in 2012 and it rose by 1.4 percent to the tune of $4.64 trillion by 2013.

The sample countries chosen were based on availability of common long data span, and it was observed that the insurance penetrations (volume of premium divided by the gross domestic product of each county) show variations across the selected countries. For instance, insurance penetration, as a measure of insurance demand, shows large variations among the selected African countries.



It should be noted that colonial heritage and the colonial settlers in South Africa contribute greatly to the huge insurance development in the country. South Africa consists of both white and black race.

In Kenya, medical insurance management firms with a regional presence have, for the fourth time in a row, posted losses. They blame the trend on escalating hospital bills and the high incidences of fraud, even as the uptake of medical cover remains low.

Figures from Kenya's Insurance Regulatory Authority (IRA) show that the sector posted a loss ratio of 76 per cent in the third quarter of last year across the 18 firms offering medical insurance covers.

Despite the premiums growing to $250.79 million in that quarter as compared with $222.4 million over the same period in 2015, losses grew to $5.37 million.

Data from IRA shows that 13 of the 18 firms offering medical covers posted losses, meaning that they paid more claims than the premiums they received.

The insurance firms say that the medical claims have consistently exceeded the segment's premiums, while management fees attributed to agent and broker commissions, as well as price undercutting, have also been blamed for the dwindling net revenues. Association of Kenya Insurers chairman Patrick Tumbo blames the industry losses mainly on fraud.

1.1.2. Medical Insurance Fraud

Fraud and abuse are widespread and very costly to Kenya's health-care system and in Africa at large. The premium in medical business has been increasing but the business continues to make losses, a situation attributed to medical insurance fraud. According to Statistics from the Insurance Regulatory Authority a total of 143 cases of medical insurance fraud were reported in 2012 and out of the Sh253.6 million lost, only Sh5.2 million recovered. In the year 2010, the loss ratio for this class of business stood at 81% with an average loss ratio of 78% over a five-year period from 2006 to 2010.

It is known that fraud exists but the level of fraud is not known. This is due to inadequate information on the prevalence of medical insurance fraud particularly on the forms and financial loses in Kenya and the region. It is against this background that the Association of Kenya Insurers (AKI) contracted MaxWorth Associates to undertake a health insurance fraud survey to establish the extent of fraud in the Kenyan health industry, identify the perpetrators, and recommend strategies for minimizing the same, if not eliminating it. Structured questionnaires, focused group discussions and in-depth interviews were carried out to obtain information from beneficiaries, health service providers, regulators, health insurers, associations, health insurance facilitators and third party administrators, investigators, the general public and industry captains.

The survey targeted beneficiaries, Health Service Providers, regulators, Health Insurers, Associations, health insurance facilitators and third party administrators, investigators, and industry captains. A total of 1000 out of 1500 respondents from 326 companies participated in the survey. The survey reports on fraud awareness levels among respondents; various forms and sources of fraud in the health industry; existing health insurance detection, prevention and mitigation practices and challenges; and the estimated impact of fraud on the business operations; lessons learnt and recommendations to minimize or eliminate fraud within the health insurance industry in Kenya.

In the fight against fraud, Baesens, Van Vlasselaer, & Verbeke, (2015), lists two components that are essential parts of almost any effective strategy to fight fraud: fraud detection and fraud prevention. Fraud detection involves distinguishing fraudulent financial data from authentic data, thereby disclosing fraudulent behaviour or activities and enabling decision makers to develop appropriate strategies to decrease the impact of fraud (Ngai, Hu, Wong, Chen, & Sun, 2011). Fraud prevention, on the other hand, refers to measures put in place to prevent or reduce fraudulent activity.

Fraud Management is inevitable to insurance companies as this informs the level of risk covered which in turn affects the premiums being charged. Over the years, with evolution of fraud detection methods, perpetrators of fraud have also been evolving their practices to avoid detection. This means as insurance premiums are increasing so is the level of fraud.

With fraudsters advancing their gimmick continuously, insurers need to embrace big data techniques that can quickly and accurately access, link and analyze massively increasing volume, velocity and granular data. For insurance companies, big data is a collection of all the data from both traditional sources (proposal forms, sales records, systems used for underwriting, claims, finance etc.) and digital sources both inside and outside the organization. Big data helps insurers have a superior understanding of their entire business environment and their client behaviors in all dimensions which informs fraud management techniques. This is made possible by the use of big data analytics tools; Machine learning is such a tool that addresses the question of how to build computers that improve automatically through experience. It is one of today's most rapidly growing technical field lying at the intersection of computer science and statistics and at the core of artificial intelligence and data science.

1.1.3. Resolution Insurance

Established in 2002 as Resolution Health, a Medical Insurance provider, Resolution Health has undergone a 360-degree revolution to become Resolution Insurance in 2013.

Known for her innovation and creative industry benchmarks (Recently awarded the most innovative insurance company in Kenya), Resolution Insurance has several firsts to her credit including being the first insurance company in Kenya to offer medical insurance to individuals suffering HIV/AIDS in 2005 when all other insurance companies were rejecting this members.

Their products are distributed by a passionate team of over 900 external agents & brokers and an in-house sales team of over 200 business consultants who believe in the brand. Over 70,000 members comprising of corporate, social groups, families and individuals have put their trust in the brand since 2002 and the Company strives to constantly develop products that are effective and relevant to the consumers. The product offering includes medical insurance, travel insurance, property fire and theft insurance, car insurance and workmen's compensation under WIBA among others.

Resolution insurance is an award winning brand that has achieved great heights since 2012 when it moved from being a medical insurance provider (MIP) to an insurance company; with key awards listed below;

2012: Winner of the Medical Insurance Provider of the Year Award in the 3rd Annual Think Business Insurance Awards.

2012: 2nd Runner up of the Training Award in the 3rd Annual Think Business Insurance Awards

2013: 2nd Runner up of the Marketing Initiative of the year award in the 4th Annual Think Business Awards

2014: Awarded in the 2014 Chartered Institute of Marketing (CIM) Excellence Awards for: Attaining the 30% Gender diversity on our board and continued support to marketing activities.

2015: 2nd Runners up of the Customer Service Award in the 6th Annual Think Business Insurance Awards

2016: Winner of Customer service award in Medical insurance at the Annual Think Business Awards

2016: 2nd Runner up of the Marketing Initiative of the year award in the Annual Think Business Awards

2017: Winner of the most innovative insurance company award in the Annual Think Business Awards

2017: 1st Runner up of the Customer service award in Medical insurance award in the 4th Annual Think Business Awards

2017: 2nd Runner up of the Training award in the Annual Think Business Awards

1.1.4. Fraud Detection & Machine learning

The classical approach used to detect fraud is an *expert-based approach*[1]. It typically involves a manual investigation of a suspicious case that may be signalled, for instance by a customer complaining of exhaustion of their medical cover limit when they have not had any claim. Such an occurrence may indicate a new fraud mechanism. Comprehension of this mechanism or pattern allows extending the fraud detection mechanism that is often implemented as a *rule base or engine*[2] that describes the newly detected fraud mechanism. This approach however, suffers from a number of disadvantages. One of such is that fraudster can also learn the business rules that block or expose them through trial and error and find ways to by-pass them. Moreover, since the rules are based on past experience, new emerging fraud patterns are not automatically flagged or signalled (Baesens et al., 2015).

Fraud detection and prevention tools have been evolving over the past decade in order to address the problem and to provide reliable solutions. Machine learning is such a tool that addresses the question of how to build computers that improve automatically through experience (Waltz, 1988). Recent progress in machine learning has been driven both by the development of new learning algorithms and theory and by the ongoing explosion in the availability of online data and low-cost computation (Jordan & Mitchell, 2015). It has been a major contributor for detecting different types of financial fraud through its diverse methods, such as, logistic regression, decision tree, Support Vector Machine (SVM) and naïve Bayes (Ngai et al., 2011).

According to Ngai et al. (2011), an effective fraud-detection and prevention system combines different tools such as - supervised learning, unsupervised learning and semi-supervised learning – which have different possibilities and limitations and therefore reinforce each other when applied in a combined setup.

In supervised methods, models are trained to discriminate between fraudulent and non-fraudulent behaviour, so that new observations can be assigned to classes in order to optimize some measure of classification performance. This requires one to be confident about the true classes of the original data used to build the models; uncertainty is introduced when legitimate transactions are

[1] It builds on the experience, intuition and business or domain knowledge of the fraud analyst
[2] A set of If-Then rules that trigger an alert or signal when a rule is broken

mistakenly reported as fraud or when fraudulent observations are not identified as such. Supervised methods require that one has examples of both classes, and they can only be used to detect frauds of a type that have previously occurred. In addition, these methods also suffer from the problem of unbalanced class sizes: in fraud detection problems, the legitimate transactions generally far outnumber the fraudulent ones and this imbalance can cause misspecification of models (Bolton, Hand, & others, 2001).

In contrast, unsupervised methods simply seek those accounts, customers, etc. whose behaviour is 'unusual' (Bolton et al., 2001). One would model a baseline distribution that represents normal behaviour and then attempt to detect observations that show greatest departure from this norm. These can then be examined more closely. Outliers are a basic form of nonstandard observation that can be used for fraud detection. It is particularly useful to notice that unsupervised learning techniques can detect both old and new fraud types since they are not bounded to the fraud patterns that are encapsulated in the labelled training samples like supervised learning techniques do.

The third major emerging machine-learning paradigm is reinforcement learning. In this method, the information available in the training data is intermediate between supervised and unsupervised learning (Horvitz & Mulligan, 2015). Instead of training examples that indicate the correct output for a given input, the training data in reinforcement learning are assumed to provide only an indication as to whether an action is correct or not; if an action is incorrect, there remains the problem of finding the correct action. Although simplified versions of reinforcement learning known as bandit problems are studied, where it is assumed that rewards are provided after each action, reinforcement learning problems typically involve a general control-theoretic setting in which the learning task is to learn a control strategy (a "policy") for an agent acting in an unknown dynamical environment, where that learned strategy is trained to choose actions for any given state, with the objective of maximizing its expected reward over time (Commission et. al., 2013). The ties to research in control theory and operations research have increased over the years, with formulations such as Markov decision processes and partially observed Markov decision processes providing points of contact. (Jordan & Mitchell, 2015).

1.1. Problem Statement

The modern approach to fraud detection in Medical Insurance is to use statistical methods such as machine learning. For example Neural Networks (NN) are used in industrial products by Visa and MasterCard (Patidar, Sharma, & others, 2011). Zhdanova, Repp, Rieke, Gaber, & Hemery (2014), however, point out the deficiencies in applying only supervised learning since they can only be used to detect frauds of a type that have previously occurred. To add to this, the chosen fraud scenario poses additional difficulties since fraudsters tend to camouflage ML activities, so that they would be statistically indistinguishable from the behaviour of a regular user. As such the method developed needs to also consider the lack of prior information about fraud patterns or samples of ML transactions. Nevertheless, it should be able to show comparable (or better) recognition performance in terms of the precision and recall metrics.

This study proposes a hybrid approach that considers both supervised and unsupervised learning and that has been noted in recent years to show superior performance across different applications. By combining both techniques, the intention is to raise detection rates of known fraudulent activity and decrease the false positive rate for unknown fraud cases. In contrast to using only neural networks (supervised learning) which are prone to local minima, overfitting and noise (Meyer, Leisch, & Hornik, 2003), the study sets out to utilize SVMs which can obtain global solutions with good generalization error and incorporates an additional technique, Self-Organizing Feature Maps (SOFM) to improve the performance of the proposed approach.

1.2. Research Objectives

i. To construct a Medical Insurance fraud detection model that utilizes NNs
ii. To construct a hybrid Medical Insurance fraud detection model that utilizes SVMs
iii. To provide a comparative analysis on the use of NNs and SVMs

1.3. Research Questions

i. How does one apply NNs in Medical Insurance fraud detection?
ii. How does one apply a hybrid approach using SVMs in Medical Insurance fraud detection?
iii. What is the difference in performance between NNs and SVMs methods?

1.4. Significance of study

Due to Insurance Regulatory Authorities (RA) regulations in most countries, it is compulsory for Medical Insurance providers to report any fraudulent activities. Therefore, Medical Insurance Fraud detection is vitally important for Medical Insurance Providers to be able to run Insurance services and prevent reputational risks. Because of this stigma associated with fraud, a successful fraud detection system or strategy is seen as an important advantage in the industry and has led to growing investment and research into the matter globally, both from academia, industry and government (Baesens et al., 2015). Despite this however, the results are seldom published in the public domain which in turn only serves to hamper overall progress in Medical Insurance Fraud research.

This paper addresses some of the problems associated with detecting Medical Insurance Fraud, and publishes comparative results on the performance of NN based machine learning methods and SVM. It is interesting to note that, in most cases, researchers claim that SVMs match or outperform neural networks in classification problems (Patidar et al., 2011). This claims shall be put to the test.

CHAPTER 2: LITERATURE REVIEW

In this chapter a brief description of the medical Insurance ecosystem and definition of the fraud scenario is first given. This is then followed by a discussion on the fraud detection techniques to be used while highlighting the practical application of each method.

2.0. Medical Insurance Ecosystem

Medical Insurance service is a complex ecosystem that involves a MIP, an employer, Members, Brokers the country's IRA, a Re-Insurance Company and MSPs. An MIP provides medical coverage for medical Risks to all parties interested including independent members in partnership with a re-insurance company of choice. Such is the case in the partnership between Resolution Insurance Group and Swiss-Reinsurance Company in Providing Medical Insurance coverage. Medical Insurance provides coverage for Medical bills incurred by an insured party in exchange of a charge called Insurance Premium. Upon occurrence of a medical insured risk, clients will access the selected panel of MSP for medical services, the claim will be documented and sent to the MIP for payment.(Peter Wehrwein et al., 2016).

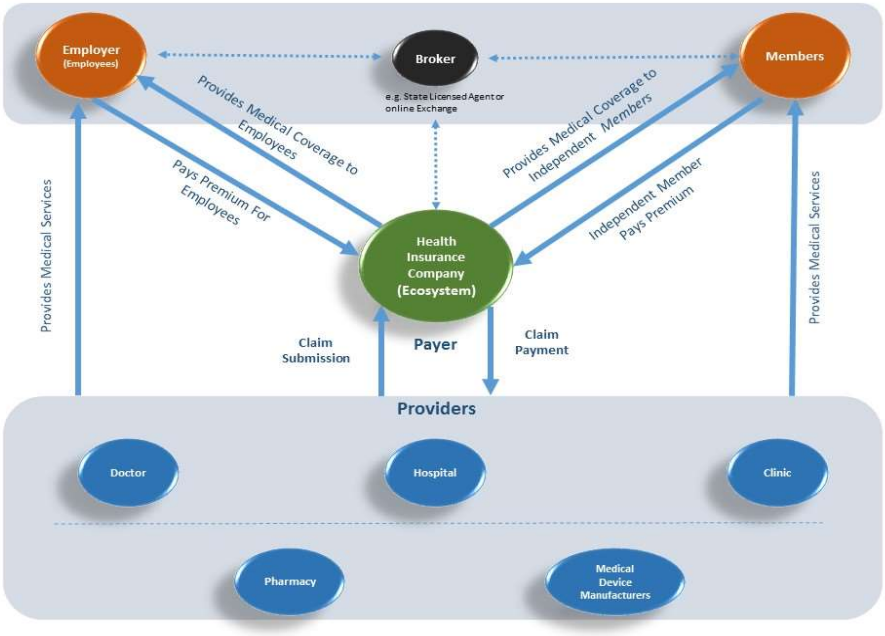


Figure 1: Medical Insurance Ecosystem

2.1. Medical Insurance Fraud

According to the National Medical Insurance Anti-Fraud Association (2011) report, it is estimated that the financial losses due to Medical insurance fraud are in the tens of billions of dollars each year. Although they constitute only a small fraction, those fraudulent claims carry a very high price tag.

Whether you have employer-sponsored health insurance or you purchase your own insurance policy, Medical insurance fraud inevitably translates into higher premiums and out-of-pocket expenses for consumers, as well as reduced benefits or coverage. For employers-private and government alike-Medical insurance fraud increases the cost of providing insurance benefits to employees and, in turn, increases the overall cost of doing business. For many Americans, the increased expense resulting from fraud could mean the difference between making health insurance a reality or not.

However, financial losses caused by Medical insurance fraud are only part of the story. Medical insurance fraud has a human face too. Individual victims of Medical insurance fraud are sadly easy to find. These are people who are exploited and subjected to unnecessary or unsafe medical procedures. Or whose medical records are compromised or whose legitimate insurance information is used to submit falsified claims.

Health care fraud, like any fraud, demands that false information be represented as truth. An all too common health care fraud scheme involves perpetrators who exploit patients by entering into their medical records false diagnoses of medical conditions they do not have, or of more severe conditions than they actually do have. This is done so that bogus insurance claims can be submitted for payment. Unless and until this discovery is made (and inevitably this occurs when circumstances are particularly challenging for a patient) these phony or inflated diagnoses become part of the patient's documented medical history, at least in the health insurer's records.

Examples of Medical Insurance Frauds include;

- Billing for services that were never rendered-either by using genuine patient information, sometimes obtained through identity theft, to fabricate entire claims or by padding claims with charges for procedures or services that did not take place.

- Billing for more expensive services or procedures than were actually provided or performed, commonly known as "upcoding"-i.e., falsely billing for a higher-priced treatment than was actually provided (which often requires the accompanying "inflation" of the patient's diagnosis code to a more serious condition consistent with the false procedure code).
- Performing medically unnecessary services solely for the purpose of generating insurance payments-seen very often in nerve-conduction and other diagnostic-testing schemes.
- Falsifying a patient's diagnosis to justify tests, surgeries or other procedures that aren't medically necessary.
- Unbundling - billing each step of a procedure as if it were a separate procedure.
- Billing a patient more than the co-pay amount for services that were prepaid or paid in full by the benefit plan under the terms of a managed care contract.
- Accepting kickbacks for patient referrals.
- Waiving patient co-pays or deductibles for medical or dental care and over-billing the insurance carrier or benefit plan (insurers often set the policy with regard to the waiver of co-pays through its provider contracting process; while, under Medicare, routinely waiving co-pays is prohibited and may only be waived due to "financial hardship") among others

2.2. Fraud Chain

Zhdanova et al. (2014), present summarised definitions of key terms in examining Fraud in which we shall base to Medical Insurance in this context. They first define a *fraud chain* as a group of identified insured's that misuse the medical insurance service to hide or disguise the origin or cause of a claim in order to receive payments for a claim they are not entitled to. They specifically evade the standard claim reporting process.

The Fraud chain may consist of an internal fraudster (employee), an insurance agent or broker/Insurance agents and a medical practitioner. The length of a fraud chain is determined by the number of participants involved in performing fraudulent transactions. Additionally, this can be defined as a Fraud *activity* that consists of one or more fraudulent operations occurring at arbitrary time intervals during the observation period.

Finally, they define a Fraud *operation* as a complete structured transfer of the fraud proceeds between two fraudsters and consisting of several individual transactions between the fraudsters and intermediaries belonging to the fraud chain. If fraudsters want to act smart and use different intermediaries for every fraudulent operation, then, the longer the observation period and the higher the length of the fraud chain.

2.3. Fraud Detection Techniques

Many techniques have been investigated for fraud detection, mainly from the statistical and Machine Learning (Sudjianto et al., 2010). The easiest way is to use thresholds on transaction amounts or any other statistical value, such as transaction or expenditure rate (Bolton & Hand, 2002). Among the data-mining techniques, mostly used are NNs, SVMs, Bayesian network models, and naïve Bayes scoring. In this section we discuss some of the Machine Learning methods in more detail while examining the results in their application.

2.3.1. Neural Networks

Neural networks were widely used in fraud detection, particularly in credit cards. They are the classifiers underlying commercial systems such as VISA which implements neural networks in the fraud detection tool RST performing real-time scoring of transactions (Patidar et al., 2011). Some financial models such as Mobile Money service have deployed Minotaur™ Fraud Management Solution based on the use of business rules and neural networks[3]. Typically, feed-forward networks with only three layers (input, hidden, and output layers) are used in fraud detection. The input to the neural network is the vector of features. The signal emitted by the output unit is the probability of the activity being criminal, which is used as a suspicion score. Given enough hidden units and proper nonlinearities and weights, three-layer neural nets are able to implement a universal function approximate (Haykin, 2009). Haykin (2009), states that back propagation is commonly used for training. The weights are initialized with random values, which are then changed in the direction that minimizes training error. More complex setups with two hidden layers, or strategies other than backpropagation are possible, but uncommon. Looking also at its application across various fields, Husain & Vohra (2017) give an overview of existing applications of machine

[3] "Minotaur™ Fraud Detection Software,"

learning in the financial sector and include a discussion of how Big Data technologies can be applied in the finance sector and traces some leading application paths.

With developments in this field such as recurrent, feedforward, and convolutional NNs, Artificial Neural Networks (ANNs) are gaining in popularity again, and at the same time winning many prizes in recent pattern recognition contests. Because the advanced versions of ANNs require even more processing power, they are implemented commonly on graphics processing units (Buczak & Guven, 2016).

Neural networks, however, also suffer from drawbacks. One major issue is the need to select and adjust the structure of the network. The choice of the number of hidden states must be made to optimize learning and generalization. Further, the performance of the classifier is very sensitive to the vector of features chosen, so significant attribute selection and pre-processing (e.g., normalization) are necessary. Bivens et al. (2002), describe a complete Intrusion Detection System that employs a pre-processing stage, clustering the normal traffic, normalization, an ANN training stage, and an ANN decision stage. The first stage used a SOFM, which is a type of unsupervised ANN, to learn the normal traffic patterns over time, such as commonly used Transmission Control Protocol (TCP)/Internet Protocol (IP) port numbers. In this manner, the first stage quantized the input features into bins, which were then fed to the second stage, a Multilayer Perceptron (MLP) ANN. The MLP network parameters, such as the number of nodes and layers, were determined by the first stage SOM. Once the MLP training was completed, it started predicting the intrusions. The system can be restarted for a new SOM to learn a new traffic pattern and a new MLP attack classifier to be trained. The study used TCP/IP data from the DARPA 1999 challenge such as Lippmann et al. (2000), where the data set consisted of network packet-level data. Unlike the previous study by Cannady (1998), which classified each packet-level data separately, this system used time windows to perform the detection and classified a group of packets. Thus, the system was able to detect attack types of longer duration. Because the input was low-level network packet data (as opposed to NetFlow data), the granularity is high and the produced predictions still correspond to short durations. Bivens et al. (2002), reported successfully predicting 100% of the normal behaviour. Their overall approach is promising, even though some attacks were not fully predicted and the FAR for some attacks reached 76%.

Sudjianto et al. (2010), point out that training neural networks is time consuming for large training datasets, especially if the model is intended to be retrained very often. In addition, backpropagation-trained multilayer perceptrons are prone to overfitting; a number of algorithms exist that address this problem, at the expense of adding complexity to the training process. Finally, neural networks are often treated as "black boxes," and their results can be difficult to interpret.

2.3.2. Self-Organizing Feature Maps

Sharma, Kalita, & Borah (2016), describe SOFM as a special class of unsupervised learning ANNs that describes a mapping from a higher-dimensional input space to a lower-dimensional map space. The procedure for placing a vector from data space onto the map is to find the node with the closest (smallest distance metric) weight vector to the data space vector.

A SOFM consists of components called nodes or neurons. Associated with each node are a weight vector of the same dimension as the input data vectors, and a position in the map space. The weights of the closest unit and its neighbours are updated after each iteration during the training process. The usual arrangement of nodes is a two-dimensional regular spacing in a hexagonal or rectangular grid. Once the training process is over, each input vector has a corresponding output vector and the Euclidean distance between the input and each unit. The unit with the smallest distance is called the Best Matching Unit (Kohonen, 1990).

2.3.3. Support Vector Machines

An SVM maps the input vector into a higher dimensional space. It is a binary classification technique that classifies input instances into two classes. Only the Support Vectors determine the optimal separating hyper-plane to classify input instance into one of the two classes. Support Vectors are the points closest to the separating hyper-plane. During classification, mapped input vectors placed on one side of the separating hyper-plane in the feature space falls into one class and placed on the other side of the plane falls into the other class. In case the data points are not linearly separable, SVM uses suitable kernel function to map them into higher dimensional space, so that, in that higher dimensional space they become separable (Sharma et al., 2016).

Drawing from literature reviewed from Wiese & Omlin (2009), it is possible to get a good overview of SVMs. Boser, Guyon, & Vapnik (1992), first discussed the theory and practicalities

behind the optimal margin classifier for separable training data - a discussion which was extended to the non-separable case by Cortes & Vapnik (1995). Burges (1998), later gave a detailed and complete tutorial on support vector machines when used for pattern recognition, and a more in depth account of the theory involved. The paper by Lee, Lin, & Wahba (2001), then tackled the task of extending SVMs to multi-category problems, i.e. problems where the data set can be divided into more than two classes. Not long after the work by Keerthi & Lin (2003) emerged useful in investigating what happens when an SVM displays tendencies towards asymptotic behaviour, for example when the separating hyper planes take on very large or small values. This helps in understanding how to prevent an SVM from overfitting or under fitting the training data. Perhaps more practical was Hsu, Chang & Lin (2003), who gave important information on successfully using SVMs in practice. They went further to discuss the cross-validation technique and proposed a grid-search method for obtaining the most appropriate kernel parameters for a new classification problem. In 2004, Kroon & Omlin went ahead to provide an overview of SVM theory and application, including discussions on the kernel trick used to generalise the SVM to high-dimensional feature space in an attempt to transform non-separable data into separable data. Their paper on SVM application is also useful in getting insight into using the popular libSVM toolkit (Chang & Lin, 2012).

The survey on data mining and machine learning techniques by Buczak & Guven (2016), proved to be very informative on matters to do with the application of SVMs. They highlight the Basic Security Module in Lippmann et al. (2000), which showed portions from the DARPA 1998 data set were used to pre-process training and testing data. The study yielded good classification performance in the presence of noise (such as some mislabelling of the training data set) and reported 75% accuracy with no false alarms and 100% accuracy with a 3% FAR. Hu, Liao, & Vemuri (2003), used the Robust SVM, a variation of the SVM where the discriminating hyperplane is averaged to be smoother and the regularization parameter is automatically determined, as the anomaly classifier in their study.

Most intriguing however, is the approach described by Shon & Moon (2007), as a framework for the detection of novel attacks in network traffic. Their approach is a combination of the SOFM, Genetic Algorithm, and SVM. Data are classified using the Enhanced SVM, which is derived from a one-class SVM, and the supervised soft-margin SVM. The first provides the unlabelled

classification capability of the one-class SVM, and the second provides the high detection performance of the supervised SVM. The study used the 1999 DARPA intrusion detection data set. To make the data set more realistic, the subset chosen consisted of 1% to 1.5% attacks and 98.5% to 99% normal traffic. The results from the Enhanced SVMs had 87.74% accuracy, a 10.20% FP rate, and a 27.27% FN rate. Those results were substantially better than those from the one-class SVMs, but not as good as soft-margin SVMs. However, the Enhanced SVM can detect novel attack patterns, whereas a soft-margin SVM cannot.

It can easily be argued that using a similar approach would be presenting an over complication of the problem. That is reason why in this paper a similar approach is used however only employing the pre-processing stage that utilised SOFM and the soft-margin SVM.

2.4. Theoretical Framework

ANNs are inspired by the brain and composed of interconnected artificial neurons capable of certain computations on their inputs (Hornik, Stinchcombe, & White, 1989). The input data activate the neurons in the first layer of the network whose output is the input to the second layer of neurons in the network. Similarly, each layer passes its output to the next layer and the last layer outputs the result. Layers in between the input and output layers are referred to as hidden layers. When an ANN is used as a classifier, the output layer generates the final classification category (Buczak & Guven, 2016).

ANN classifiers are based on the perceptron (Rosenblatt, 1958) and were very popular until the 1990s when SVMs were invented by Boser et al. (1992). Compared to the convex quadratic optimization applied in an SVM, ANNs often suffer from local minima and thus long runtimes during learning. Unlike an SVM, as the number of features in an ANN increase, it's learning runtime increases. With one or more hidden layers, the ANN is able to generate nonlinear models (Buczak & Guven, 2016).

SOFMs, on the other hand, differ from other artificial neural networks as they apply competitive learning as opposed to error-correction learning (such as backpropagation with gradient descent), and in the sense that they use a neighbourhood function to preserve the topological properties of the input space. This makes SOFMs useful for visualizing low-dimensional views of high-dimensional data, akin to multidimensional scaling. The artificial neural network introduced by

the Finnish professor Teuvo Kohonen in the 1980s is sometimes called a Kohonen map or network (Kohonen, 1982; Kohonen & Honkela, 2007) . The Kohonen net is a computationally convenient abstraction building on work on biologically neural models from the 1970s (Malsburg, 1973) and morphogenesis models dating back to Alan Turing (Turing, 1990) in the 1950s.

Crossing over, several papers have shown that SVMs statistical learning techniques, with strong theoretical foundation have displayed successful application in a range of problems (Cristianini & Shawe-Taylor, 2000). They are closely related to neural networks, and through use of kernel functions, can be considered an alternate way to obtain neural network classifiers.

Rather than minimizing empirical error on training data, SVMs seek to minimize an upper bound on the generalization error. As compared with techniques like neural networks which are prone to local minima, overfitting and noise, SVMs can obtain global solutions with good generalization error. They are more convenient in application, with model selection built into the optimization procedure, and have also been found to outperform neural networks in classification problems (Meyer et al., 2003). Appropriate parameter selection is, however, important to obtain good results with SVM.

Bhattacharyya, Jha, Tharakunnel, & Westland (2011), highlight that the strength of SVMs come from two important properties they possess — kernel representation and margin optimization. In SVMs, mapping to a high-dimensional feature space and learning the classification task in that space without any additional computational complexity are achieved by the use of a kernel function. Schölkopf & Smola (2002), explains that a kernel function can represent the dot product of projections of two data points in a high-dimensional feature space. The high-dimensional space used depends on the selection of a specific kernel function. The classification function used in SVMs can be written in terms of the dot products of the input data points. Thus, using a kernel function, the classification function can be expressed in terms of dot products of projections of input data points in a high-dimensional feature space. With kernel functions, no explicit mapping of data points to the higher-dimensional space happens while they give the SVMs the advantage of learning the classification task in that higher-dimensional space. The second property of SVMs is the way the best classification function is arrived at. SVMs minimize the risk of overfitting the training data by determining the classification function (a hyper-plane) with maximal margin of

separation between the two classes. This property provides SVMs very powerful generalization capability in classification.

Zhdanova et al. (2014), further suggests that semi-supervised or unsupervised detection tools are set to provide the solution to the shortcomings of supervised methods, as they do not need a training database that might be difficult to obtain. As such classical supervised machine learning tools, even though not demonstrating the best possible detection rates, in some application scenarios might not be as efficient as other techniques. It is because of this that SOFMs are employed. This method does not depend on any prior information about fraud patterns or samples of ML transactions (Shon & Moon, 2007). At the same time, it acts as a simple and efficient method for defining normal classes within the data pool. This method is used as the first step in clustering the data then feeds the output into the supervised method.

2.5. Conceptual Framework

Over the years, along with the evolution of fraud detection methods, perpetrators of fraud have also been evolving their fraud practices to avoid detection (Bolton et al., 2001). Therefore, Medical Insurance fraud detection methods need constant innovation. In this study, we evaluate two advanced Machine Learning approaches, support vector machines and neural networks, together with the SOFMs, as part of an attempt to improve Medical Insurance fraud detection.

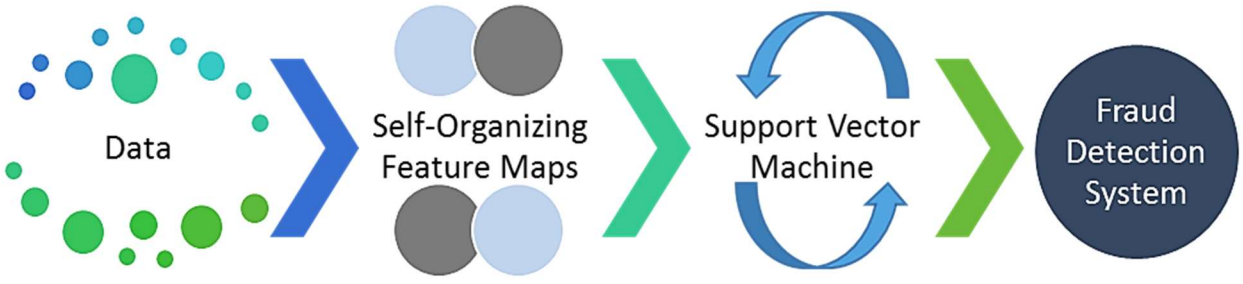


Figure 2: Conceptual Model

This study aims at providing a comparative analysis on the use of neural networks and support vector machines. The choice to focus on SVMs in conjunction with SOFM as opposed to other Machine Learning methods is influenced by existing literature that claims that SVMs are already known as one of the best learning algorithm for binary classification.

A series of experiments will be run on the data sets using NNs and SVMs, during which the resulting performances of each algorithm will be computed. These results will then be analytically compared to see how the two algorithms compare to each other when applied to a data set.

CHAPTER 3: METHODOLOGY

3.0. Introduction

This chapter presents the research methodology that entails: the research design, data population and sampling as well as the methodology used in data analysis and processing.

3.1. Research Design

The study takes an applied exploratory approach to investigate the effectiveness of using SVMs in Medical Insurance Fraud detection in comparison to NNs. Furthermore, it incorporates SOFMs as a tool for unsupervised learning to reinforce the shortcomings of the supervised method (SVM).

3.2. Population and Sampling

The study seeks to use a real life data set of Medical Insurance transactions. For the purpose of this study, this dataset of all transactions shall be called dataset U and much smaller subsets of this large dataset: dataset A, B and C. To compare Medical Insurance Fraud prediction using the proposed techniques, sets of transactions of both known fraudulent and undetected or observed legitimate transactions will be needed.

The sample was chosen from claims paid and registered under resolution Insurance Limited between the periods 1st Jan 2015 to 31st Dec 2016.

There were challenges with regards to past known fraudulent transactions. However, this presents a good challenge for SOFMs to reinforce SVM as a measure of identifying fraudulent claims.

3.3. Data Processing

The Machine Learning methods will be subjected to three phases:

i. Training (A)
ii. Validation (B)
iii. Testing (C)

Buczak & Guven (2016), point out that Machine Learning and Data Mining methods often have parameters such as the number of layers and nodes for a NN. After the training is complete, there are usually several models (e.g., NNs) available. To decide which one to use and have a good estimation of the error it will achieve on a test set, there should be a third separate data set, the validation data set. The model that performs the best on the validation data should be the model used, and should not be fine-tuned depending on its accuracy on the test data set. Otherwise, the accuracy reported is optimistic and might not reflect the accuracy that would be obtained on another test set similar to but slightly different from the existing test set. For a binary classification problem such as this, the classification metric used will be the confusion matrix.

3.4. Data Analysis

3.4.1. Profiling using SOFM

Shon & Moon, (2007) explain that collecting normal data is one of the most important aspects in the field of machine learning because this collected data is required for training the supervised learning system. Moreover, the normal data can provide an intelligent criterion to an unsupervised machine learning system. In the proposed machine learning framework the machine learning algorithm SOFM which is capable of making clusters as in Stateczny & Wlodarczyk-Sielicka (2014), is employed. The aim is to generate several clusters and distinguish these with normal features.

A SOFM converts non-linear statistical relationships between data points in high-dimensional space into geometrical relationships between points in a two-dimensional map. In addition, it differs from competitive learning in that neighbouring neurons in SOFM learn to recognize neighbouring sections of the input space while it belongs to a family of competitive learning algorithms. Thus, SOFM learns from both the distribution and topology of the input vectors used for training. The learning procedure of SOFM is similar to that of competitive learning networks.

In other words, a similarity measure is selected and the winning unit is considered to be the one with the largest activation. However, for Kohonen feature maps, we update all of the weights, including the winning unit's weights, in the neighbourhood surrounding the winning units. Generally, the size of neighbourhood decreases slowly with each iteration. The sequential description of how to train a Kohonen self-organizing network is as follows:

Step 1: Select the winning output unit as the one with the largest similarity measure (or smallest dissimilarity measure) between all weight vectors w_i and input vector x. If the Euclidean distance is chosen as the dissimilarity measure, then the winning unit c satisfies the following equation:

$$\left||x - w_c|\right| = min_i \, ||x - w_c||,$$

Where the index c refers to the winning unit and the index i refers to the output unit.

Step 2: Let NB_c denote a set of indices corresponding to a neighbourhood around winner c. The weights of the winner and its neighbouring units are then updated by:

$$\Delta w_i = \eta(x - w_i), i \in \mathrm{NB}_c$$

Where η is the small positive learning rate. Instead of defining the neighborhood of a winning unit, we use a neighbourhood function $\Omega_c(i)$ around the winning unit c. For instance, the Gaussian function can be used as the neighbourhood function:

$$\Omega_c(i) = \exp\left(\frac{-\left||p_i - p_c|\right|^2}{2\sigma^2}\right)$$

Where p_i and p_c are the positions of the output units i and c, respectively, and σ reflects the scope of the neighbourhood. By using the neighbourhood function, the update formula can be rewritten as:

$$\Delta w_i = \eta\Omega_c(i)(x - w_i)$$

Where i is the index for all output units. To achieve better convergence, the learning rate η and the size of neighbourhood (or σ) should be decreased gradually each iteration.

When the high-dimensional input is provided to the above, SOFM can use topology and distance to transform it into two-dimensional space and generate the various clusters. In the problem

domain, a variety of SOFM clusters will be made in accordance with the attributes of packets. At this point in the process, if we label the generated clusters using a statistical method or a verification procedure (distinguishing normal from abnormal by off-line analysis), then some of the clusters would be used as a normal packet profile, or a profile of packets exhibiting normal packet features. Not only that, an unsupervised clustering using SOFM does not require any other measures to make clusters because it only relies on the original characteristics of the given inputs. Therefore, SOFM can act as a simple and efficient method for defining normal packet classes.

3.4.2. Soft-Margin SVM

Borrowing much from Shon & Moon (2007), they describe a soft margin SVM learning algorithm written by Cortes & Vapnik, (1995), which is sometimes called c-SVM. This SVM classifier has a slack variable and penalty function for solving non-separable problems. First, given a set of points $x_i \in R^d, i = 1, \ldots, l$ and each point x_i belongs to either of two classes with the label $y_i \in \{-1,1\}$. These two classes can be applied to anomaly attack detection with the positive class representing normal and negative class representing abnormal. Suppose there exists a hyper-plane $w^T x_i + b = 0$ that separates the positive examples from the negative examples. That is, all the training examples satisfy:

$$w^T x_i + b \geq +1 \; for \; all \; x_i \in P$$

$$w^T x_i + b \leq -1 \; for \; all \; x_i \in N$$

w is an adjustable weight vector, x_i is the input vector and b is the bias term.

Equivalently:

$$y_i(w^T x_i + b) \geq 1 \; for \; all \; i = 1, \ldots, N$$

In this case, we say the set is linearly separable.

In Fig. 2, the distance between the hyper-plane and $f(x)$ is $\frac{1}{||w||}$. The margin of the separating hyper-plane is defined to be $\frac{2}{||w||}$. Hence, the learning problem is reformulated as minimizing $||w||^2 =$

$w^T w$ is subject to the constraints of linear separation. This is equivalent to maximizing the hyper-plane distance between the two classes, of which the maximum distance is called the support vector. The optimization is now a convex quadratic programming problem.

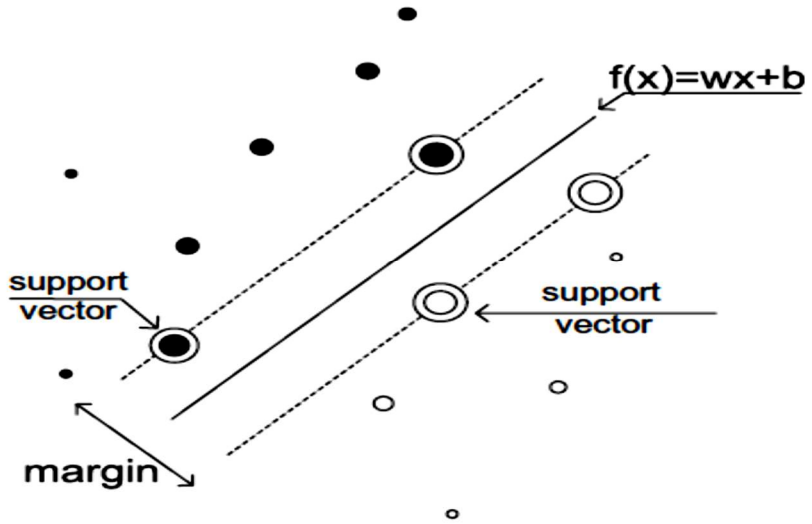


Figure 3: SVM hyper-plane

$$\text{Minimize}_{w,b} \quad \phi(w) = \frac{1}{2}||w||^2$$

$$\text{Subject to} \qquad y_i(w^T x_i + b) \geq 1 \; for \; all \; i = 1, \ldots, l$$

Since $\phi(w) = \frac{1}{2}||w||^2$ is convex in w and the constraints are linear in w and b, we can be sure that this problem has a global optimum solution. This has the advantage that parameters in quadratic programming affect only the training time, and not the quality of the solution. This problem is tractable, but anomalies in fraud detection show characteristics of non-linearity, and as a result they are more difficult to classify. In order to proceed to such non-separable and non-linear cases, it is useful to consider the dual problem as outlined below.

The Lagrange for this problem is

$$L\,(w, b, \Lambda) = \frac{1}{2}||w||^2 - \sum_{i=1}^{l} \lambda_i [y_i(w^T x_i + b) - 1]$$

Where $\Lambda = (\lambda_1, \ldots, \lambda_l)^T$ are the Lagrange multipliers, one for each data point. The solution to this quadratic programming problem is given by maximizing L with respect to $\Lambda \geq 0$ and minimizing with $\frac{1}{2}||w||^2$ respect to w and b. Note that the Lagrange multipliers are only non-zero when

$y_i(w^T x_i + b) = 1$, and the vectors for this case are called support vectors, since they lie closest to the separating hyper-plane. However, in the non-separable case, forcing zero training error leads to poor generalization. We introduce the soft margin SVM using a vector of slack variables $\mathcal{Z} = (\psi_1, \ldots, \psi_l)^T$ that measure the amount of violation of the constraints, taking into account the fact that some data points may be misclassified.

The equation is now:

$$\text{Minimize}_{w,b,\mathcal{Z}} \qquad \phi(w, b, \mathcal{Z}) = \frac{1}{2}||w||^2 + C \sum_{i=1}^{l} \psi_i^k$$

$$\text{Subject to} \qquad y_i(w^T \phi(x_i) + b) \geq 1 - \psi_i, \psi_i \geq 0, \quad i = 1, \ldots, l$$

where C is a regularization parameter that controls the trade-off between maximizing the margin and minimizing the training error. The effects of C are crucial. If C is too small, insufficient stress is placed on fitting the training data. If C is too large, the algorithm will over fit the dataset.

In practice, a typical SVM approach, such as the soft margin SVM, showed excellent performance more often than other machine learning methods (Akande, Owolabi, Twaha, & Olatunji, 2014). For intrusion detection applications, supervised machine learning approaches based on SVM methods proved superior to intrusion detection approaches using artificial neural networks. Therefore, the high classification capability and processing performance of the soft margin SVM approach is useful for anomaly detection. However, because the soft margin SVM is a supervised learning approach, it requires datasets to be labelled.

3.4.3. SOFM and SVM procedure

In this study, incurred claims are clustered using the SOFM algorithm followed by applying SVM algorithm to each of the clusters to classify those claims into “legitimate” and “fraudulent” classes.

Clustering is being done with respect to Disease, Medical Service Provider, Service Type (Lab, Consultation and Prescriptions), gender and age.

SOFM Steps:

Training occurs in several steps and over many iterations:-

1. Select the size and type of the map. The shape can be hexagonal or square, depending on the shape of the nodes required. Typically, hexagonal grids are preferred since each node then has 6 immediate neighbours.
2. Initialise all weight vectors randomly.
3. Choose a random data point from training data and present it to the SOFM.
4. Find the "Best Matching Unit" (BMU) in the map – the most similar node (winning unit).
5. Determine the nodes within the "neighbourhood" of the BMU. The size of the neighbourhood decreases with each iteration.
6. Adjust weights of nodes in the BMU neighbourhood towards the chosen data point. The learning rate decreases with each iteration. The magnitude of the adjustment is proportional to the proximity of the node to the BMU.
7. Repeat Steps 3-6 for N iterations / convergence.

After the training phase is complete, SVM algorithm will be applied to predict into which class the new incoming claims data will fall whether “legitimate’ or “fraudulent” based on duplicate claims, up coding of services, up coding of items, unnecessary services and billing for service not rendered.

SVM Steps:

1. Define two class labels namely “legitimate” or “fraudulent”
2. Classify claims into two classes using the training data set.
3. Choose support vectors and find the maximum marginal hyper plane that separates the claims into two classes.
4. Identify the new incoming claims into either “legitimate’’ or “fraudulent” class.

3.4.4. Feedforward Neural Network

The neural network that will be used in this fraud detection framework is a three-layer, feed-forward network that uses two training passes through the data set and borrows heavily from the work by Patidar et al., (2011). The first training pass involves a process of prototype cell commitment in which exemplars from the training set are stored in the weights between the first and second (middle) layer cells of the network. A final training pass determines local a posteriori probabilities associated with each of these prototype cells. The objective of the neural network training process is to arrive at a trained network that produces a fraud score that gives the best ranking of the claims transactions. If the ranking were perfect, all of the high scoring transactions down to some threshold would be fraud; below this threshold, only valid transactions would be ranked. However, perfect separation of frauds from goods is not possible due to the inherently non-separable nature of the fraud and good distributions in the selected pattern recognition space.

The back propagation learning rule is a standard learning technique. It performs a gradient descent in the error/ weights space. To improve the efficiency, a momentum term is introduced, which moves the correction of the weights in the direction compliant with the last weight correction.

It is a multi-layer feed forward network that is trained by supervised learning. A standard back propagation network consists of 3 layers, an input, an output and a hidden layer. The processing elements of both input and output layer are fully connected with the processing elements of the hidden layer. The fact that it is feed forward means that there are no recurrent loops in the network. The output of a node never returns at the same node, because cycles are not allowed in the network. In standard back propagation this can never happen because the input for each processing element always comes from the previous layer (except the input layer, of course). This, again, is a large simplification compared with the real brain because the brain itself appears to contain many recurrent loops.

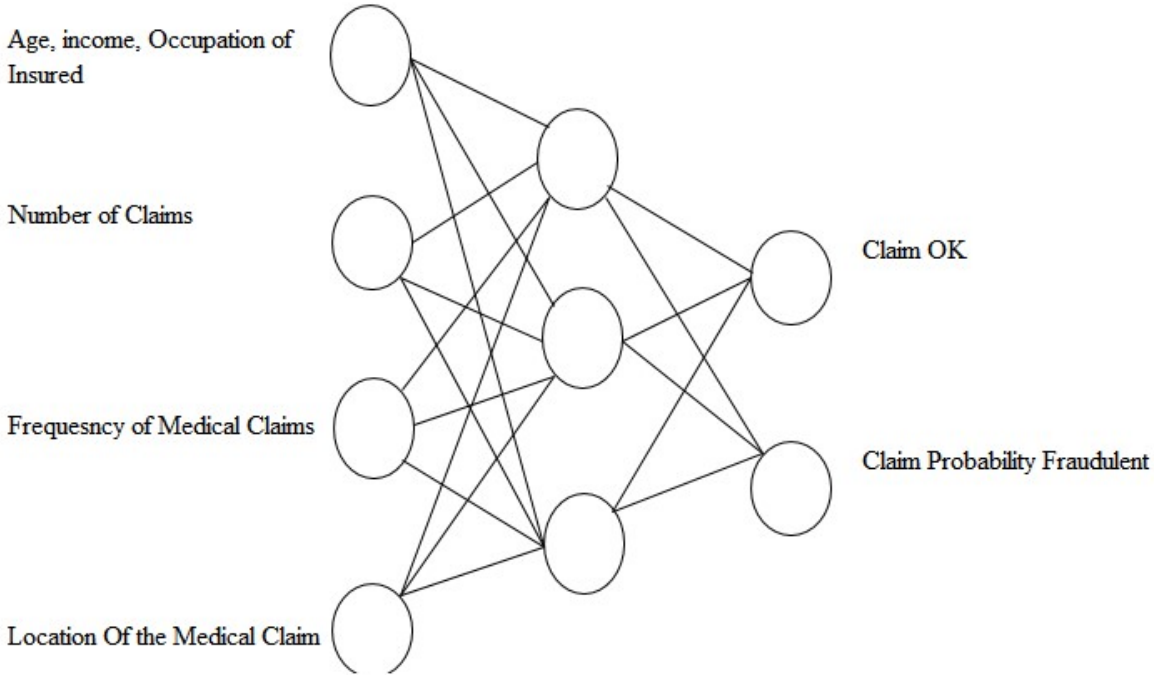


Figure 4: NN Framework

Supervised learning means that the network is repeatedly presented with input/output pairs (I, O) provided by a supervisor, where O is the output the network should produce when is presented with input I. These input/output pairs specify the activation patterns of the input and output layer. The network has to find an internal representation that results in the wanted input/output behaviour. To achieve this, back propagation uses a two-phase propagate-adapt cycle.

i. *First Phase*: In the first phase the input is presented to the network and the activation of each of the nodes (processing elements) of the input layer is propagated to the hidden layer, where each node sums its input and propagates its calculated output to the next layer. The nodes in the output layer calculate their activations in the same way as the nodes in the hidden layer.

ii. *Second Phase*: In the second phase, the output of the network is compared with the desired output given by the supervisor and for each output node the error is calculated. Then the error signals are transmitted to the hidden layer where for each node its contribution to the total error is calculated. Based on the error signals received, connection weights are then adapted by each node to cause the network to converge toward a state that allows all the training patterns (input/output pairs) to be encoded.

REFERENCES

Akande, K. O., Owolabi, T. O., Twaha, S., & Olatunji, S. O. (2014). Performance comparison of SVM and ANN in predicting compressive strength of concrete. *IOSR Journal of Computer Engineering*, *16*(5), 88–94.

Baesens, B., Van Vlasselaer, V., & Verbeke, W. (2015). *Fraud analytics using descriptive, predictive, and social network techniques: a guide to data science for fraud detection*. John Wiley & Sons.

Birk, D., Gajek, S., Grobert, F., & Sadeghi, A.-R. (2007). Phishing phishers-observing and tracing organized cybercrime. In *Internet Monitoring and Protection, 2007. ICIMP 2007. Second International Conference on* (pp. 3–3). IEEE. Retrieved from http://ieeexplore.ieee.org/abstract/document/4271749/

Bivens, A., Palagiri, C., Smith, R., Szymanski, B., Embrechts, M., & others. (2002). Network-based intrusion detection using neural networks. *Intelligent Engineering Systems through Artificial Neural Networks*, *12*(1), 579–584.

Bolton, R. J., & Hand, D. J. (2002). Statistical fraud detection: A review. *Statistical Science*, 235–249.

Bolton, R. J., Hand, D. J., & others. (2001). Unsupervised profiling methods for fraud detection. *Credit Scoring and Credit Control VII*, 235–255.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory* (pp. 144–152). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=130401

Buczak, A. L., & Guven, E. (2016). A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications Surveys & Tutorials*, *18*(2), 1153–1176.

Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, *2*(2), 121–167.

Cannady, J. (1998). Artificial neural networks for misuse detection. In *National information systems security conference* (pp. 368–81). Retrieved from http://pld.cs.luc.edu/courses/intrusion/fall05/cannady.artificial_neural_networks_for_misuse_detection.pdf

Chang, C., & Lin, C. (2012). LIBSVM: a library for support vector machines," 2001. Software available at http://www. csie. ntu. edu. tw/∼ cjlin/libsvm. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.552.5513

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines*. Cambridge University Press Cambridge. Retrieved from http://library02.embl.de/InmagicGenie/DocumentFolder/TableOfContents_H898.pdf

Haykin, S. (2009). *Neural networks and learning machines* (3. ed). New York: Pearson.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359–366.

Horvitz, E., & Mulligan, D. (2015). Data, privacy, and the greater good. *Science*, *349*(6245), 253–255.

Hsu, C.-W., Chang, C.-C., Lin, C.-J., & others. (2003). A practical guide to support vector classification. Retrieved from http://www.datascienceassn.org/sites/default/files/Practical%20Guide%20to%20Support%20Vector%20Classification.pdf

Hu, W., Liao, Y., & Vemuri, V. R. (2003). Robust Support Vector Machines for Anomaly Detection in Computer Security. In *ICMLA* (pp. 168–174). Retrieved from http://web.cs.ucdavis.edu/~vemuri/papers/rvsm.pdf

Husain, R., & Vohra, R. (2017). APPLYING MACHINE LEARNING IN THE FINANCIAL SECTOR. *International Education and Research Journal*, *3*(1). Retrieved from http://ierj.in/journal/index.php/ierj/article/view/637

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260.

Keerthi, S. S., & Lin, C.-J. (2003). Asymptotic behaviors of support vector machines with Gaussian kernel. *Neural Computation*, *15*(7), 1667–1689.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, *43*(1), 59–69. https://doi.org/10.1007/BF00337288

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, *78*(9), 1464–1480.

Kohonen, T., & Honkela, T. (2007). Kohonen network. *Scholarpedia*, *2*(1), 1568.

Kroon, S., & Omlin, C. W. (2004). Getting to grips with support vector machines: Theory. *South African Statistical Journal*, *282*, 159–172.

Lee, Y., Lin, Y., & Wahba, G. (2001). Multicategory support vector machines. In *Proceedings of the 33rd Symposium on the Interface*. Retrieved from https://pdfs.semanticscholar.org/3a6d/9e4c0ef0ac8f3b1f4cc2da4463ef6fc473cd.pdf

Lippmann, R. P., Fried, D. J., Graf, I., Haines, J. W., Kendall, K. R., McClung, D., … others. (2000). Evaluating intrusion detection systems: The 1998 DARPA off-line intrusion detection evaluation. In *DARPA Information Survivability Conference and Exposition, 2000. DISCEX'00. Proceedings* (Vol. 2, pp. 12–26). IEEE. Retrieved from http://ieeexplore.ieee.org/abstract/document/821506/

Malsburg, C. (1973). Self-organization of orientation sensitive cells in the striate cortex. *Kybernetik*, *14*(2), 85–100. https://doi.org/10.1007/BF00288907

Meyer, D., Leisch, F., & Hornik, K. (2003). The support vector machine under test. *Neurocomputing*, *55*(1), 169–186.

Minotaur™ Fraud Detection Software. (n.d.). Retrieved May 30, 2017, from https://swipx.com/apps/minotaur-fraud-detection-software/

Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*, *50*(3), 559–569.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386.

Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press. Retrieved from https://books.google.com/books?hl=en&lr=&id=y8ORL3DWt4sC&oi=fnd&pg=PR13&dq=B.+Sch%C2%A8olkopf,+A.+Smola,+Learning+with+Kernels,+MIT+Press,+2002.&ots=bKBW8AW2Dy&sig=GFqQRAgK9fHI0ydwCwWY2BnLTR4

Sharma, R. K., Kalita, H. K., & Borah, P. (2016). Analysis of machine learning techniques based intrusion detection systems. In *Proceedings of 3rd International Conference on Advanced Computing, Networking and Informatics* (pp. 485–493). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-81-322-2529-4_51

Shon, T., & Moon, J. (2007). A hybrid machine learning approach to network anomaly detection. *Information Sciences*, *177*(18), 3799–3821.

Stateczny, A., & Wlodarczyk-Sielicka, M. (2014). Self-organizing artificial neural networks into hydrographic big data reduction process. In *International Conference on Rough Sets and Intelligent Systems Paradigms* (pp. 335–342). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-319-08729-0_34

Sudjianto, A., Nair, S., Yuan, M., Zhang, A., Kern, D., & Cela-Díaz, F. (2010). Statistical methods for fighting financial crimes. *Technometrics*, *52*(1), 5–19.

Turing, A. M. (1990). The chemical basis of morphogenesis. *Bulletin of Mathematical Biology*, *52*(1–2), 153–197.

US Dept of Justice, N. D. I. C., & America, U. S. of. (2008). ML in Digital Currencies. Retrieved from https://www.ncjrs.gov/App/Publications/abstract.aspx?ID=245786

Waltz, D. L. (1988). The prospects for building truly intelligent machines. *Daedalus*, 191–212.