International Journal for Modern Trends in Science and Technology
Volume: 03, Special Issue No: 02, March 2017
ISSN: 2455-3778
http://www.ijmtst.com

Predictive Analytics Using Support Vector Machine

Ch.Sai Sindhu[1] | T.Hema Sai[2] | Ch.Swathi[3] | S.Kishore Babu[4]

[1,2,3,4]Department of IT, Andhra Loyola Institute of Engineering & Technology, Vijayawada, Andhra Pradesh, India.

To Cite this Article
Ch.Sai Sindhu, T.Hema Sai, Ch.Swathi, S.Kishore Babu, "Predictive Analytics Using Support Vector Machine ", *International Journal for Modern Trends in Science and Technology*, Vol. 03, Special Issue 02, 2017, pp. 19-23.

ABSTRACT

Predictive analytics is an area of data mining that deals with extracting information from data and using it to predict trends and behavior patterns. Analytics is a field of study that deals with analysing historical data, drawing inferences there from, and using the information so gathered in predicting the future. With the developments in the Information Technology and advancements in the e-commerce field, fraud is expanding causing huge financial losses. Though security prevention mechanisms such as CHIP and PIN are developed, these mechanisms do not prevent the most common fraud types. Fraud detection mechanism is the best way to stop such types of fraud. Predictive Analytics provides methods, techniques and tools helping us to learn automatically and to make accurate predictions based on past observations. This predictive analysis is done using R, a programming language and software environment for statistical Computing. In this project, Support Vector Machine is used to predict the fraudulent values based on past observations from Credit card data set.

I. Introduction

Data analytics (DA) is a science that combines data mining, machine learning, and statistics. DA examines raw data with the purpose of discovering useful information, suggesting conclusions, and supporting decison-making. DA has become popular as big data problems have emerged in biological science, engineering, business, and other fields. There are many techniques that have been developed in data analytics. DA is classified into Descriptive Analytics, Predictive Analytics.

1.1Predictive Analytics: Understanding the future

Predictive analytics has its root in the ability to "Predict" what might happen. These analytics are about understanding the future. Predictive analytics provides companies with actionable insights based on data. Predictive analytics provide estimates about the likelihood of a future outcome. It is important to remember that no statistical algorithm can "predict" the future with 100% certainty. Companies use these statistics to forecast what might happen in the future.

1.2What are Predictive Analytics-The Traditional View

Oracle Corporation et al [1] proposed that "Predictive analytics encompasses a variety of techniques from statistics, data mining and game theory that analyze current and historical facts to make predictions about future events". The variety of techniques is usually divided in three categories: predictive models, descriptive models and decision models. Predictive models look for certain relationships and patterns. Predictive models focus on a specific event or behavior, descriptive models identify as many different relationships as possible.

This classification is very practical; it provides an immediate understanding of the areas where predictive analytics add value. However, there are two problems with it:

- The classification is not exhaustive. If a new area of predictive analytics were to arise tomorrow the current classification would become invalid.
- The classification doesn't tell what the categories don't do. The limitations of each category are not clear.
- Categories are not clear.Traditionally, predictive analytics require a laborious process.

1.3Predictive Analytics

Predictive analytics is used to make the predictions about unknown future events and it uses many techniques from data mining, modeling, statics, artificial intelligence and machine learning to help analysts make future business forecasts. Predictive analytics can not tell us what will happen in the future but It forecasts what might happen in the future with an acceptable level of reliability.[6]Businesses collect vast amounts of real-time customer data and predictive analytics uses the historical data, combined with customer insight, to predict future events. Predictive analytics enable the organizations to use the big data (both stored and real-time) to move from a historical view to a forward-looking perspective of the customer.The regression techniques are used for prediction purposes.Regression analysis is a form of predictive modelling technique which investigates the relationship between **dependent**(target) and **independent variable (s)** (predictor).

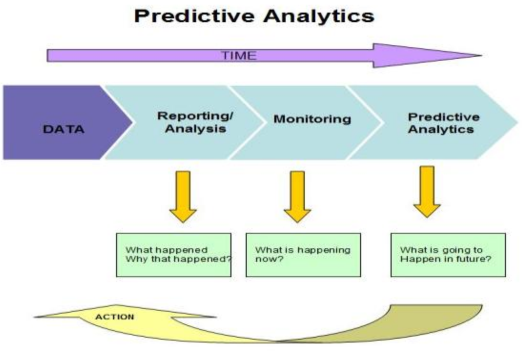

Fig 1:Predictive Analytics

1.4Predictive Analytics Algorithms

Various Predictive Analytics algorithms are:

1. Support Vector Machine
2. Gaussian Process
3. Artificial Neural Networks
4. Regression

1.4.1 Support Vector Machine

SVMs are a popular machine learning method for classification, regression, & other learning tasks. LIBSVM is a library for Support Vector Machines (SVMs). A typical use of LIBSVM involves two steps: first, training a data set to obtain a model & second, using the model to predict information of a testing data set.

SVM uses a technique called the kernel trick to transform your data and then based on these transformations it finds an optimal boundary between the possible outputs. Simply put, it does some extremely complex data transformations, then figures out how to seperate your data based on the labels or outputs you've defined.It is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples

Basic concept

SVMs were developed by Cortes &Vapnik (1995) for binary classification. Their approach may be roughly sketched as follows:

Class separation: Basically, we are looking for the optimal separating hyperplane between the two classes by maximizing the margin between the classes' closest points (see Figure 1)—the points lying on the boundaries are called support vectors, and the middle of the margin is our optimal separating hyperplane.

Overlapping classes: Data points on the"wrong"side of the discriminant margin are weighted down to reduce their influence ("soft margin").

Nonlinearity: When we cannot find a linear separator, data points are projected into an (usually) higher-dimensional space where the data points effectively become linearly separable (this projection is realised via kernel techniques).

Problem solution: The whole task can be formulated as a quadratic optimization problem which can be solved by known techniques.

A program able to perform all these tasks is called a Support Vector Machine.

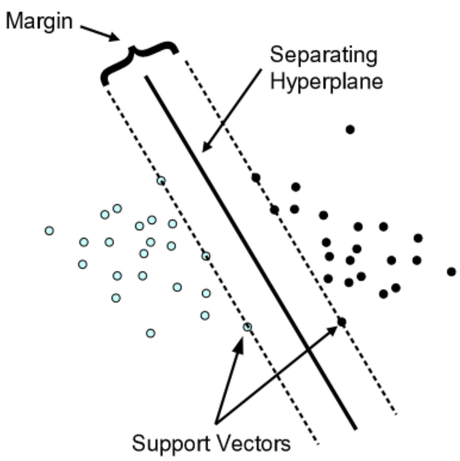

Fig 2:Classification (linear separable case)

Several extensions have been developed; the ones currently included in libsvm are:

ν-classification: This model allows for more control over the number of support vectors by specifying an additional parameter ν which approximates the fraction of support vectors;

One-class-classification: This model tries to find the support of a distribution and thus allows for outlier/novelty detection.

Multi-class classification: Basically, SVMs can only solve binary classification problems. To allow for multi-class classification, libsvm uses the one-against-one technique by fitting all binary subclassifiers and finding the correct class by a voting mechanism.

e-regression: Here, the data points lie in between the two borders of the margin which is maximized under suitable conditions to avoid outlier inclusion.

ν-regression: With analogue modifications of the regression model as in the classification case.

1.4.2 Gaussian Process

A Gaussian Process is a collection of random variables, any finite number of which has a joint Gaussian distribution. Gaussian process is a generalization of the Gaussian probability distribution. A typical Gaussian distribution concerns about a single random variable, whereas a Gaussian process is associated with a collection of random variables that produces a pool of functions relevant for prediction. In other words, Gaussian process is a distribution over functions.

1.4.3 Artificial neural networks

An artificial neural network (ANN) learning algorithm, usually called "neural network" (NN), is a learning algorithm that is inspired by the structure and functional aspects of biological neural networks. Computations are structured in terms of an interconnected group of artificial neurons, preprocessing information using a connectionist approach to computation. Modern neural networks are non-linear statistical data modeling tools.

1.4.4 Regression Analysis

Regression analysis is a form of predictive modeling technique which investigates the relationship between a dependent (target) and independent variable (8) (predictor). This technique is used for forecasting, time series modeling and finding the causal effect relationship between the variables. For example, relationship between rash driving and number of road accidents by a driver is best studied through regression. Regression analysis is an important tool for modeling and analyzing data.

II. PROBLEM DEFINITION

Fraud detection is necessary for any financial system. Fraud Identification in credit card system is most concerning issues in online transaction. With growing popularity on online shopping provided by various web services, various kind of fraudulent activities are being observed. With emergence of information technology & improvement in data communication, fraud is expanding causing huge financial losses. Complete elimination of banking fraud is not possible; however we can limit its occurring to certain level & prevent them from happening by artificial intelligence technique. Approach of artificial intelligence in credit card detection is newly used. Thus probability of effective results is likely to be more. So the fraud can be detected by extracting the insights from the data.

III. PROPOSED FRAMEWORK

Architecture flow

The following diagram is a generalized architecture involving data allocation to threads. Below architecture diagram represents mainly flow of data from the Preprocessing to the Predictive model. We preprocess the data to eradicate the noisy and inconsistent data. After that we apply some of the predictive models like Support Vector Machine (SVM) to predict the fraudulent data out of the legitimate ones. Its architecture is represented in the following way.

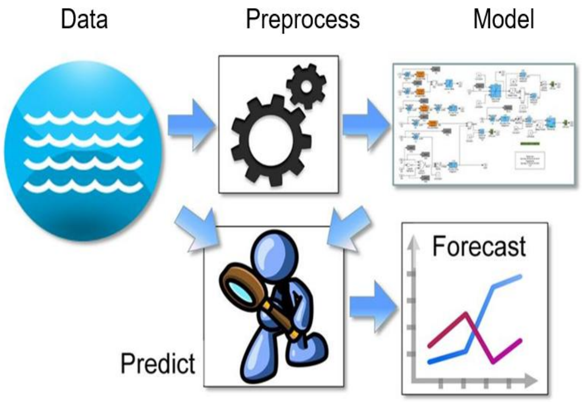

Fig 3: System Architecture

SYSTEM OVERVIEW

In this system, each account is monitored separately, and the transactions are attempt to be identified and flagged as legitimate or normal. The identification will be based on the suspicion score produced by the classifier models developed. When a new transaction is going, the SVM classifier can predict whether the transaction is normal or fraud.

Input: Dataset

- Read the input credit card data set for which we are going to build a predictive model.
- Pre-process the above data set using normalization and pre-processing transformation (centering, scaling etc.) methods.
- Create SVM Model for the Pre-processed data.
- Run Prediction and measure the execution time.
- Tune SVM to find the best cost and gamma.
- After finding the best cost and gamma, create svm model again and try to run.
- Evaluate the model.

IV. CONCLUSION

Fraud Identification in credit card system is most concerning issues in online transaction. It initially starts with the usage of clustering & outlier detection techniques. These techniques are considered to be the basis for finding data that does not belong to the current data pattern. These are further made accurate by the usage of SVM & behavior bases SVM. SVM, being a binary classifier helps in providing the user with a result of whether the current transaction is legitimate or fraudulent. Detecting the fraudulent process is the most important functionality that a bank could offer its customers. But this could prove to be a serious downside if the transaction detected as fraudulent by the system proves to be a legitimate one. This could lead to reduction in goodwill of the company. This project builds a predictive model that identifies fraudulent transactions. The concept of Support Vector Machine (SVM) is being used to solve the problem. Thus by implementation of this approach, financial losses can be reduced to greater extend.

V. FUTURE WORK

Fraud Identification in credit card system is most concerning issues in online transaction. This project builds a predictive model that identifies fraudulent transactions. The concept of Support Vector Machine (SVM) is being used to solve the problem. Thus by implementation of this approach, financial losses can be reduced to greater extend. The field of Support Vector Machine, Artificial Neural Networks, Gaussian process, Linear Regression model has diverse opportunities for future research in the predictive analytics. This project has implemented Support Vector Machine prediction technique to build the predictive model. In future, the research work can be extended by implementing various other techniques such as Artificial Neural Networks, Gaussian process and Linear Regression Model techniques. By implementing all these techniques, we can determine which technique is more reliable.

References

[1] Vijayshree B. Nipane, Poonam S. Kalinge, DipaliVidhate, Kunal War, Bhagyashree P. Deshpande, Fraudulent Detection in Credit Card System Using SVM & Decision Tree, ISSN: 2455-2631© May 2016 IJSDR

[2] Sitarampatel, Sunita Gond, Supervised Machine (SVM) Learning for Credit Card Fraud Detection, Volume 8 Number 3- Feb 2014.

[3] R. Dhanpal& P. Gayathiri, "Credit card fraud detection using decision tree for tracing email &ip," International Journal of Computer Science Issues, vol. 9, no. 2, 2012.

[4] R. D. Patel & D. K. Singh, "Credit card fraud detection & prevention of fraud using genetic algorithm," International Journal of Soft Computing & Engineering (IJSCE), vol. 2,no. 6, 2013.

[5] Y. Sahin& E. Duman, "Detecting credit card fraud by decision trees & support vector machines," Proceeding of theInternational MultiConfrence of Engineers & Computer Scientist, vol. I, 2011.

[6] J. Pun & Y. Lawryshyn, "Improving credit card fraud detection using a meta-classification strategy,"

International Journal of Computer Applications, vol. 56, no. 10, 2012.

[7] R. Patidar& L. Sharma, “Credit card fraud detection using neural network” International Journal of Soft Computing & Engineering (IJSCE), vol. 1, 2011.

[8] A. Srivastava & A. Kundu, “Credit card fraud detection using hidden markov model,” IEEE Transactions on Dependable & Secure Computing, vol. 5, no. 1, 2008.

[9] G. Singh, R. Gupta, A. Rastogi, M. Ch&el, & A. Riyaz, “A machine learning approach for detection of fraud based on svm,” International Journal of Scientific Engineering & Technology, vol. 1, no. 3, 2012.