A Systematic Literature Review of Data Classification Techniques

Neha Nigam
Shri Vaishnav Institute of Information Technology
Indore, India

Anand Rajavat, PhD
Shri Vaishnav Institute of Information Technology
Indore, India

ABSTRACT

The data mining and their different applications are becomes more popular now in these days a number of large and small scale applications are developed with the help of data mining techniques i.e. predictors, regulators, weather forecasting systems and business intelligence. Many of classification algorithms are available to analyze data. Classification is used to classify each item in a data set into one of a predefined set of classes or groups. Classification is the chore of identifying a model or function. There are two kinds of model are available for namely supervised and unsupervised. The performance and accuracy of the supervised data mining techniques are higher as compared to unsupervised techniques therefore in sensitive applications the supervised techniques are used for prediction and classification. In this presented work the supervised learning based data mining techniques for classification and prediction are analyzed.

Keywords

Data Mining, Classification, Decision Tree, KNN Classification, supervised learning

1. INTRODUCTION

Data mining is the process of extracting hidden patterns from data set. As large amount of data is gathered, with the amount of data doubling every three or four years, data mining is becoming an increasingly important tool to transform this data or data set into knowledge. It is mostly used in a wide range of applications, such as marketing, scientific discovery and fraud detection. Data mining can be used to data sets of any size, and while it can be used to discover hidden patterns, it cannot discover patterns which are not already present in the data set.

Knowledge Discovery in Databases (KDD) [3] is an automated extraction of novel, understandable and potentially useful patterns implicitly stored in huge databases, data warehouse and other massive information storehouse. KDD (Figure 1) is a multi-disciplinary field drawing work from areas including database technology, high performance computing, artificial intelligence, machine learning, neural networks, statistics, pattern recognition, information retrieval and data visualization.

Data mining [1] has become an essential technology for businesses and researchers in many fields, the number and variety of applications has been growing gradually for several years and it is predicted that it will carry on to grow. A number of the business areas with an early embracing of DM into their processes are banking, insurance, retail and telecom. More lately it has been implemented in pharmaceutics, health, government and all sorts of e-businesses.

[13] Reports a work dealing with understanding student data using data mining. Here decision tree algorithms are used for predicting graduation, and for finding factors that lead to graduation.

2. LITERATURE REVIEW

Venkata et al. [4] As the cost of the data processing and Internet accessibility increases, more and more organizations are becoming vulnerable to a wide range of cyber threats. Most current offline intrusion detection systems are focused on unsupervised and supervised machine learning approaches. In this system, Information Gain (IG) and Triangle Area based KNN are used for selecting more discriminative features by combining Greedy k- means clustering algorithm and SVM classifier to detect Network attacks. This system achieves high accuracy detection rate and less error rate of KDD CUP 1999 training data set.

Esh et al. [5] In present time many intrusions in network and the activities of intrusion is the goal of the security policy system. The unsupervised learning techniques using the machine learning for intrusion detection datasets, we know that Clustering is the best techniques on the efficient data mining for intrusion detection. The k-mean clustering algorithm is widely used for intrusion detection, because it gives efficient results.

Deepika et al. [6] Intrusion detection is an awfully exigent area of research in a current scenario. Now- a-days find a novel pattern of intrusion and detection of this pattern are exceedingly demanding job. In this project the object is to affect a method for intrusion detection using KNN classification and Dempster theory of evidence.

Prabhu et al. [7] Network intrusion detection is a way to separate normal behaviors from the attacked ones. The proposed system is based on the adaboost algorithm with Naive Bayes classifier to detect network intrusions with high detection rates and low false-alarm rates. This results in low computational complexity and error rates.

Nagarajan et al. [8] IDS which are increasingly a key part of system defense are used to identify abnormal activities in a computer system. In general, the traditional intrusion detection relies on the extensive knowledge of security experts, in particular, on their familiarity with the computer system to be protected. To reduce this dependence, various data-mining and machine learning techniques have been used in the literature

Nasser et al. [9] The rapid growth of Internet malicious activities has become a major concern to network forensics and security community. With the increasing use of IT technologies for managing information there is a need for stronger intrusion detection mechanisms. Critical mission systems and applications require mechanisms able to detect any unauthorized activities.

Debdutta et al.[10] In multi-hop wireless systems, the need for cooperation among nodes to relay each other's packets exposes them to a wide range of security attacks. A particularly devastating attack is the wormhole attack, where a malicious node records control traffic at one location and tunnels it to another compromised node, possibly far away, which replays it locally.

Dianbo et al. [11] Neural Networks approach is an advanced methodology used for intrusion detection. As a type of Neural Network, Self-organizing Maps (SOM) is getting more attention in the field of intrusion detection.

Hazem et al. [12] E-government is an important issue which integrates existing local area networks into a global network that provide many services to the nation citizens. This network requires a strong security infrastructure to guarantee the confidentiality of national data and the availability of government services.

Todd Heberlein [13] proposed an intrusion detection system called network system monitor. This system is based on the concept of analyzing network instead of the system log entry.

Teng, Chen, And Lu [14], proposed time based inductive machine to capture or store user behavior. Inductive generalization is also a part of the process.

Anderson D, Lunt TF, Javitz H, Tamaru A, Valdes [15], proposed a network intrusion detection expert system. This system learns from the training data and predicts the test data.

Lane and Brodley [16] applied the concept of the instant based learning. Lee W. and Stolfo S. and Mok [17] proposes a novel data mining based framework for intrusion detection. This model is based on the concept of the utilizing the contents of the audited programs.

Debar, H., Dacier, M., And Wespi [18] proposes taxonomy of the intrusion detection systems. This classification is done according to the property of the intrusion detection system.

In this paper *R. Thanigaivel et al [19]* survey different papers in which one or more algorithms of data mining used for the prediction of heart disease. Result from using neural networks is nearly 90%. So that the prediction by using data mining algorithm given efficient results. Applying data mining techniques to heart disease treatment data can provide as reliable performance as that achieved in diagnosing heart disease.

Authors in [20][21][22] used the concept of data classification to predict the employee performance in an organization. K-Nearest neighbors algorithm is used. It also includes Artificial neural network,Decision Tree, Logistic regression. Efficiency and Effectiveness of employees is determined using Data Mining Classification Technique with basic paramenters. It does not include confidential information about employee.

Shantakumar, et al. [23] have done a research work in which the intelligent and effective heart attack prediction system is developed using Multi-Layer Perceptron with Back-Propagation. Accordingly, the frequency patterns of the heart disease are mined with the MAFIA algorithm based on the data extracted.

Yanwei, et.al [24] have built a classification method based on the origin of multi parametric features by assessing HRV (Heart Rate Variability) from ECG and the data is pre-processed and heart disease prediction model is built that classifies the heart disease of a patient.

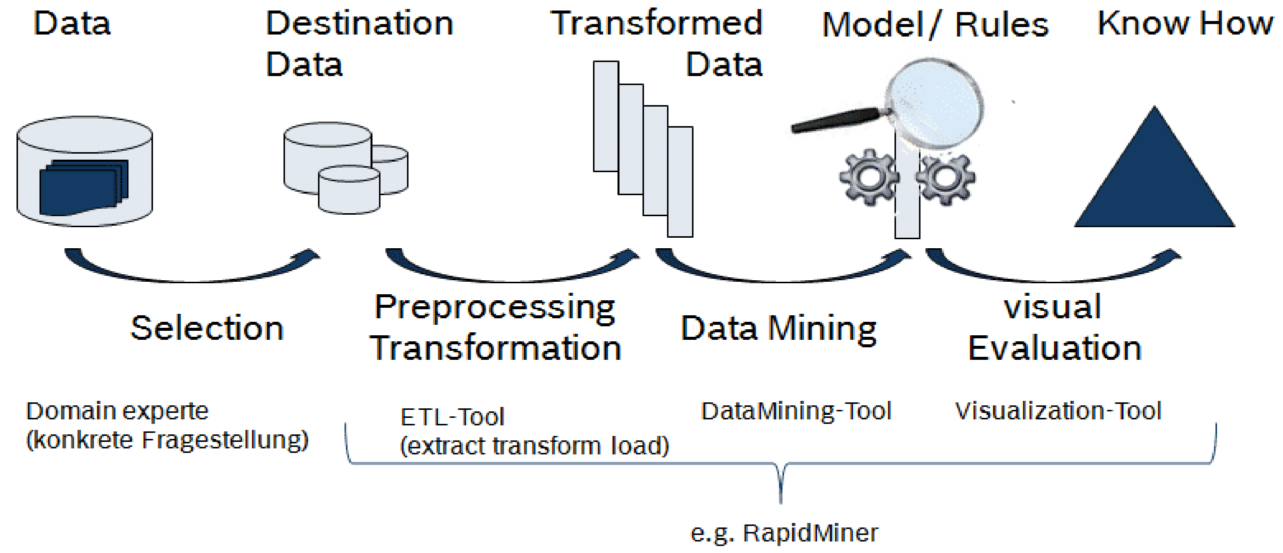

Fig. 1. The process of knowledge discovery in databases [1]

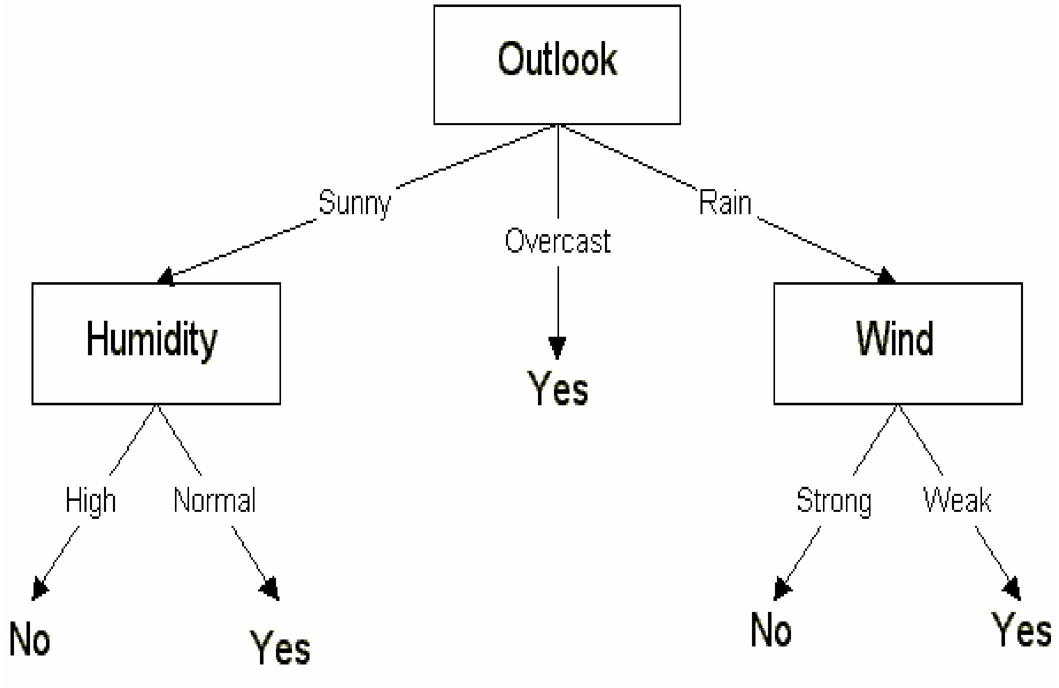

Figure 2 : A Sample decision tree-Partial view [13]

3. CONCLUSION

The data mining is helpful for analysis the data, when the manually analysis of the data is not feasible then the data mining techniques are applied for analysis. The data mining techniques are the computer based algorithms which identify the relationship among the data and extraction of the similar pattern data on which they are trained. This paper presented a critical review of various data mining based techniques for the classification and prediction of data.

4. REFERENCES

[1] Tan P.-N., Steinbach M., and Kumar V. ―Introduction to data mining, Addison Wesley Publishers‖. 2006 .

[2] Fayyad U. M., Piatetsky-Shapiro G. and Smyth, P. ―Data mining to knowledge discovery in databases, AI Magazine‖. Vol. 17, No. 3, pp. 37-54, 1996.

[3] https://www.sas.com/en_us/insights/analytics/data-mining.html

[4] Venkata Suneetha Takkellapati G.V.S.N.R.V Prasad, "Network Intrusion Detection system based on Feature Selection and Triangle area Support Vector Machine", International Journal of Engineering Trends and Technology- Volume3, Issue 4, 2012.

[5] Esh Narayan, Pankaj Singh and Gaurav Kumar Tak, "Intrusion Detection System Using Fuzzy C-Means Clustering with Unsupervised Learning via EM Algorithms" VSRD-IJCSIT, Vol. 2 (6), 502-510, 2012.

[6] Deepika Dave, Prof. Vineet Richhariya, "Intrusion detection with KNN classification and DS- theory", IRACST Vol. 2, No.2, April 2012.

[7] P.S. Prabhu, "Network Intrusion Detection Using Enhanced Adaboost Algorithm", International Journal of Communications and Engineering Volume 3, No.3, Issue:02 March 2012.

[8] R. Shanmugavadivu, Dr.N.Nagarajan, "Network Intrusion Detection System Using Fuzzy Logic" IJCSE Vol. 2 No. 1, 2011.

[9] Nasser S. Abouzakhar And Abu Bakar, "A Chi-Square Testing-Based Intrusion Detection Model",CFET, 2010.

[10] Debdutta Barman Roy, Rituparna Chaki, Nabendu Chaki, "A New Cluster-Based Wormhole Intrusion Detection Algorithm for Mobile Ad-Hoc Networks", IJNSA, Vol 1, No 1, April 2009.

[11] Dianbo Jiang, Yahui Yang, Min Xia, "Research on Intrusion Detection Based on an Improved SOM Neural Network", IEEE 2009.

[12] Hazem M. El-Bakry, Nikos Mastorakis, "A Real-Time Intrusion Detection Algorithm for Network Security", Wseas Transactions on Communications Issue 12, Volume 7, December 2008.

[13] Todd, H. L., Gihan V.D., Karl N.L., Biswanath, M., Jeff, W. and David, W. "A network security monitor," in Proceedings of Symposium on Research in Security and Privacy, Oakland, CA, pp. 296–304, 1990.

[14] Teng, H., Chen, K. and Lu, S. "Adaptive real time anomaly detection using inductively generated sequential patterns', IEEE Computer Society Symposium on Research in Security and Privacy, California, IEEE Computer Society, pp. 278-84, 1990.

[15] Anderson, J.B. and Mohan, S. "Sequential coding algorithms: A survey and cost analysis", IEEE Transactions on Communication, Vol.32, pp. 169-176, 1984.

[16] Lane, T. and Brodley, C.E. "Temporal sequence learning and data reduction for anomaly detection", ACM Transactions on Information and System Security, Vol. 2, No. 3, 1999.

[17] Lee, W., Stolfo, S. and Mok, K. "Adaptive intrusion detection: A data mining approach", Artificial

Intelligence Review, Kluwer Academic Publishers, Vol. 14, No.6, pp. 533-567, 2000.

[18] Debar, H., Becker, M. and Siboni, D. "A neural network component for an intrusion detection system," in IEEE Symposium on Research in Computer Security and Privacy, pp. 240-250, 1992.

[19]] R.Thanigaivel, Dr. K.Ramesh Kumar, "Review on Heart Disease Prediction System using Data MiningTechniques", Asian Journal of Computer Science and Technology (AJCST)Vol.3.No.1 2015 pp 68-74.

[20] Rahul edida, RakshitVahe, Rahul reddy, Rahul j, Abhilash, DeeptiKulkarni,"Employeeattrition prediction", Management journal, 2018.

[21] KedirEyasu Abdul Kadir, FuleaAmenaTolfsa, "Predict and analysis of employee performance in bank using classification algorithms", International journal of interdisciplinary current advanced research(IJICAR), Vol. 1, No .1, Feb 2019.

[22] Ananya Sarkar, S.M.Shamim, Dr. Md. Shahiduz Zama, Md. MustafizurRahman,"Employees performance analysis and prediction using K means Clustering and decision tree algorithm", Global Journals, Vol.18, No. 1,2018.

[23] Ersen Yilmaz and Caglar Kilikcier, "Determination of Patient State from Cardiotocogram using LS-SVM with Particle Swarm Optimization and Binary Decision Tree", Master Thesis, Department of Electrical Electronic Engineering, Uludag University, 2013.

[24] Nidhi Singh and Divakar Singh, "Performance Evaluation of K-Means and Hierarchal Clustering in Terms of Accuracy and Running Time", Ph.D Dissertation, Department of Computer Science and Engineering, Barkatullah University Institute of Technology,2012.