INTERNATIONAL JOURNAL OF RESEARCH AND DEVELOPMENT IN ENGINEERING SCIENCES
e-ISSN: 2582 - 4201 Volume: 2 | Issue: 6 | Year: 2020

Data Mining Classification Techniques on the Analysis of Text Mining

KANAPARTHI KANTHARAJU
Assistant Professor and HOD,
Department of Computer Science and Engineering,
GDMM College of Engineering and Technology, Nandigama
Jawaharlal Nehru Technological University, Kakinada, Andhra Pradesh, India
e-mail: kantharaju8@gmail.com

Abstract - *Data mining involves the searching of large information of the data or records to discover patterns and utilize these patterns in the prediction the future events. In most educational sectors such as high schools, polytechnics and universities; classification technique is a vital analytical mechanism in prediction of various levels of accuracy. Classification is one of the methods in data mining for categorizing a particular group of items to targeted groups. Main goal of classification is to predict the nature of an items or data based on the available classes of items. With the emergence of analytics as an overarching term for all data analyses, data mining is put back into its proper place a critical part of analytics continuum where the new discovery of knowledge happens. This mini-track has six papers, collectively illustrating the depth and breadth of data, text and Web mining, and their innovative applications to interesting and highly challenging business problems.*

Keywords: Data Mining, KDD, Text, Classifications

1. Introduction

Data mining includes the utilization of refined data analysis tools to find previously unknown, valid patterns and relationships in huge data sets. These tools can incorporate statistical models, machine learning techniques, and mathematical algorithms, such as neural networks or decision trees. Thus, data mining incorporates analysis and prediction. Depending on various methods and technologies from the intersection of machine learning, database management, and statistics, professionals in data mining have devoted their careers to better understanding how to process and make conclusions from the huge amount of data, but what are the methods they use to make it happen. In recent data mining projects, various major data mining techniques have been developed and used, including association, classification, clustering, prediction, sequential patterns, and regression [1].

Clustering is a division of information into groups of connected objects. Describing the data by a few clusters mainly loses certain confine details, but accomplishes improvement. It models data by its clusters. Data modeling puts clustering from a historical point of view rooted in statistics, mathematics, and numerical analysis. From a machine learning point of view, clusters relate to hidden patterns, the search for clusters is unsupervised learning, and the subsequent framework represents a data concept. From a practical point of view, clustering plays an extraordinary job in data mining applications.

For example, scientific data exploration, text mining, information retrieval, spatial database applications, CRM, Web analysis, computational biology, medical diagnostics, and much more. We can classify a data mining system according to the kind of databases mined. Database system can be classified according to different criteria such as data models, types of data, etc. And the data mining system can be classified accordingly.

Classification is a process of determining classes of given objects based on their characteristics, where semantic of classes are known beforehand. Typical applications of data mining classification are: Credit or Loan Approval- if a client is the safe or risky; Spam detection- If a message is valid or suspicious; what treatment applies to a patient- If Treatment A is suitable or Treatment B is more preferable or Treatment C is ideal?; Web-page categorization- which category a web page belongs business, entertainment or education.

Data mining is one important field that focuses on discovering the data set properties and also an analytical step of knowledge discovery in databases (KDD). Educational Data mining (EDM) is an important aspect of research that has assisted in the predicting of useful information obtained from the educational database which yields to an improvement in the performance and also enhances the possibilities of a better understanding of the students to have a better assessment of their learning process.

The modeling of user knowledge, user behavior and user experience is applied to EDM which is an aspect of data mining. Classification is a process of assigning new entities to existing defined class by examining the entities features. Classification makes decision from unseen cases by building of past decisions. Educational data mining uses many techniques such as k-nearest neighbor, linear regression analysis, naïve Bayes, neural networks, support vector machines, decision trees and many others. The KNIME analytics platform which is open source software for creating data science applications and services; intuitive, open and continuously integrating new developments, it makes understanding data, designing data science workflows and reusable components accessible to everyone. KNIME integrates various components from machine learning and data mining via its modular pipelining concepts of data [2].

2. Related Work

Prediction and Classification Prediction involves the search for the hidden patterns or the existing knowledge from the available historical data. For instance, the detection any fraudulent transaction from a person credit card, we simply analyze any extraordinary pattern from the person transaction historical data. Other typical applications of prediction include target marketing and medical diagnosis such that the predicting of suitable and best medicine for a patient based on patient medical history. Classification is a technique in data mining of generally known structure to apply to new data.

Decision Tree The concept used in this work for the classification technique of the student's mark is the decision tree. Decision tree is a tree like structure where internal node contains splits and splitting attributes. It represents test on the attribute.

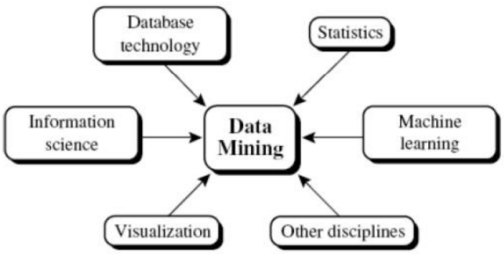

Figure.1: Data Mining

Decision trees are usually constructed from the training data set while the test data set is used to test or validate the accuracy of the decision tree. Decision tree is a flowchart-like tree structure which the following features:

- Each internal node also referred to as the non-leaf node denotes a test on an attribute;
- Each branch represents an outcome of the test ;
- Each leaf node or terminal node holds class label
- Topmost node of the tree is the root node.

The decision usually consists of the nodes that form a rooted tree, which means the directed tree with a node called root that has not incoming edges.

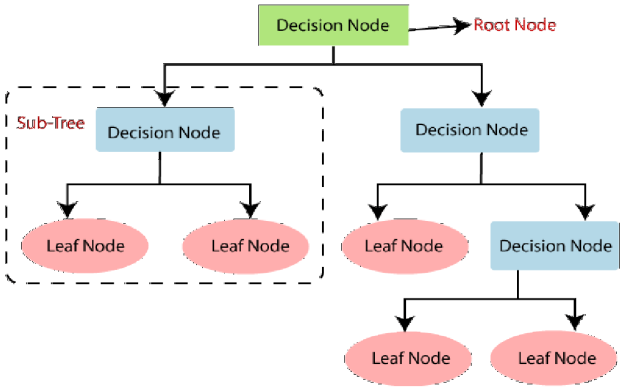

Figure.2: Decision Tree analysis

In 2012 Akhil jabbar et al. proposed "Heart Disease Prediction System using Associative Classification and Genetic Algorithm". They proposed efficient associative classification algorithm using genetic approach for heart disease prediction. The main advantage of genetic algorithm is the discovery of high level prediction rules is that the discovered rules are highly comprehensible, having high predictive accuracy and of high interestingness values. The proposed method helps in the best prediction of heart disease which even helps doctors in their diagnosis decisions [3].

In 2013 Akhil Jabbar et al. proposed "Classification of Heart Disease using Artificial Neural Network and Feature Subset Selection". They proposed a new feature selection method using ANN for heart disease classification. For rank the attributes which contribute more towards classification of heart disease they applied different feature selection methods, and indirectly reduce the no. of diagnosis tests to be taken by a patient. The proposed method eliminates useless and distortive data. In 2014 N. S. Nithya et al . proposed "Gain ratio based fuzzy weighted association rule mining classifier for medical diagnostic interface". They showed that earlier model based on information gain and fuzzy association rule mining algorithm for extracting both association rules and membership functions are not feasible. They used large number of distinct values. They modify gain ratio based fuzzy weighted association rule mining and improve the classifier accuracy.

In 2015 S. Olalekan Akinola, O. Jephthar Oyabugbe proposed "Accuracies and Training Times of Data Mining Classification Algorithms: An Empirical Comparative Study". They proposed study was designed to determine how data mining classification algorithm perform with increase in input data sizes. They used three data mining classification algorithms In 2016 Nikhil N. Salvithal & R.B. Kulkarni proposed "Appraisal Management System using Data mining Classification Technique". The proposed assorted classifier algorithms applied on Talent dataset to spot the talent set so as to judge the performance of the individual. Finally counting on accuracy one best suited classifier is chosen this method has been used to construct classification rules to predict the potential talent that for promotion or not. In 2016 Tanvi Sharma & Anand Sharma proposed "Performance Analysis of Data Mining Classification Techniques on Public Health Care". The proposed study focused on the application of various data mining classification techniques using different machine learning tools such as WEKA and Rapid miner over the public healthcare dataset for analyzing the health care system. The percentage of accuracy of every applied data mining classification technique is used as a standard for performance measure. The best technique for particular data set is chosen based on highest accuracy [4].

Erraguntla *et al.* propose an innovative applied data analytics and data mining research project, which was funded by U.S. Army Medical Research and Materiel. The goal of the project, creatively named as Data Integration and Predictive Analysis System (IPAS), is to enable prediction, analysis, and response management for incidents of infectious diseases. IPAS collects and integrates comprehensive datasets of previous disease incidents and potential influencing factors to facilitate multivariate, predictive analytics of disease patterns, intensity, and timing. IPAS supports comprehensive epidemiological analysis exploratory spatial and temporal correlation, hypothesis testing, prediction, and intervention analysis.

Innovative machine learning and predictive analytical techniques like support vector machines (SVM), decision tree-based random forests, and boosting are used to predict the disease epidemic curves. Predictions are then displayed to stakeholders in a disease situation awareness interface, alongside disease incidents, syndromic and zoonotic details extracted from news sources and medical publications. Data on Influenza Like Illness (ILI) provided by CDC was used to validate the capability of IPAS system, with plans to expand to other illnesses in the future. This paper presents the ILI prediction modeling results as well as IPAS system features.

This is written by Albashrawi et al., investigates mobile banking (MB) usage through the theoretical lens of UTAUT

model with its four pillars. The research model proposes to be tested using a hybrid neural networks-based structural equation modeling (SEM-NN) framework to identify the significant contributors/factors. Universal structural modeling (USM) can then be utilized to find the hidden paths and nonlinearity in our research model.

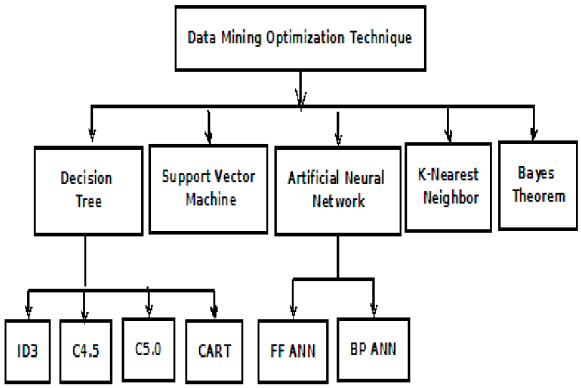

Figure.3: Data Mining Techniques

To the best of authors' knowledge, this is the first study to examine the role of subjective and objective experience on MB usage by utilizing a multi-analytical approach. Neural network (NN) and USM can identify the most significant determinants and hidden interaction effects, respectively. Thus, both techniques would help to complement SEM and increase our understanding of the influential factors on MB usage. The paper presents their preliminary results and discusses the implications. The third paper, written by Egger and Schoder, is about consumer-oriented tech mining, a relatively new an innovative concept in the world of analytics [5].

According to the authors, to avoid missing technological opportunities and to counteract risks, organizations ought to scan and monitor developments in the external environment through a structured process of technology intelligence. Previous approaches in tech mining—the applications of text mining for technology intelligence—have primarily focused on the elicitation of technical or legal information from Web, patent, or research databases. However, knowledge of consumers' needs, fears, and hopes is a prerequisite for the success of an emerging technology in the marketplace.

Thus, the authors claim that technology intelligence needs to also consider consumers' technology perceptions. Hence, they propose a novel and comprehensive approach to collect user-generated content from the Web and apply text mining to derive consumer perceptions. In doing so, they align with an established tech-mining process. This paper illustrates their approach on the emerging technology of autonomous driving and provides an initial indication of concurrent validity [6].

Support Vector Machine (SVM)

SVM is a very effective method for regression, classification and general pattern recognition. It is considered a good classifier because of its high generalization performance without the need to add a priori knowledge, even when the dimension of the input space is very high. It is considered a good classifier because of its high generalization performance without the need to add a priori knowledge even when the dimension of the input space is very high. For a linearly separable dataset, a linear classification function corresponds to a separating hyper plane f(x) that passes through the middle of the two classes, separating the two. SVMs were initially developed for binary classification but it could be efficiently extended for multiclass problems.

3. Conclusion

There are several classification techniques in data mining and each and every technique has its advantage and disadvantage. Decision tree classifiers, Bayesian classifiers, classification by back propagation, support vector machines, these techniques are eager learners they use training tuples to construct a generalization model. Some of than are lazy learner like nearest-neighbor classifiers and case-based reasoning. These store training tuples in pattern space and wait until presented with a test tuple before performing generalization.

References:

[1]. M.S. Mythili and A.R. Mohamed Shanavas, "An analysis of students" performance using classification algorithms", IOSR, Journal of Computing Engineering, volume 16, Issue 1, January 2014.
[2]. S.Lakshmi Prabha , A.R.Mohamed Shanavas, "Educational data mining applications", Operations Research and Applications: An international Journal (ORAJ), volume 1, No. 1, August, 2014.
[3]. DavinderKaur, Rajeev Bedi and S.K Gupta, "Review of decision tree data mining algorithms: ID3 and C4.5", pp. 5-8, july 2015
[4]. S.Ayesha, T.Musafa, A.Sattar and M.khan, "Data mining model for higher education system", European Journal of Scientific Research, vol. 43, no 1, 2010. pp. 24-29.
[5]. Open for Innovation KNIME Software, URL: https://www.knime.com/knime-software. 6. Electronic URL Source- https://slideplayer.com/slide/13868930/
[6]. Baradwaj, B.K. and Pal, S., 2011. "Mining Educational Data to Analyze Students' Performance". (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011