WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

DATA MINING CLASSIFICATION TECHNIQUES AND ALGORITHM EXPLORATION USING WEKA TOOL

S.Thennammai [1] , Dr. A. Kanagaraj[2]

[1]*M.Phil. Scholar, Department of Computer Science, Nallamuthu Gounder Mahalingam College, Pollachi-642 001, India*
[2]*Asst. Prof., Department of Computer Science, Nallamuthu Gounder Mahalingam College, Pollachi-642 001, India*

ABSTRACT

In this paper we are going to see about data mining, what is data mining techniques and various data mining techniques and in that we have taken text mining. Text mining is nothing but we are analysing the data set in text format and finding the results and accuracy. In that we have taken classification technique of data mining and analysed the dataset with classification types likes Bayes, lazy, rules and trees methods in that we have taken one option from each method for instance in Bayes we have taken Naïve Bayes, lazy method we have taken KNN and in weka it is represented as IBK, IBK is Linear N neighbour search using the K method, rules we identified PART algorithm and in trees we have taken J48 method. We have taken a dataset of weather numeric and using weka tool we have analysed the dataset correct accuracy percentage of each method. From these we have identified KNN method – IBK which gives a good accuracy of correct instance compared to all other methods.

Keywords *– Data Mining, Text Mining, Techniques, Classification, Naïve Bayes, PART, KNN, J48*

1. INTRODUCTION

Data Mining is the process of identifying patterns in huge datasets involving methods at intersection of machine learning, statistics and database systems. Data mining is an inter disciplinary subfield of statistics and computer science. Its overall goal is to extract the information in specified pattern. The term "data mining" is to extract the goal and achieve the pattern and knowledge information. It is mostly applied to any form of large data to find the information.

Text mining is used in organizations for knowledge driven. Text mining is the process of examining large collections of documents to discover new information or to help to find solution for specific research questions. Text mining identifies facts, relationships and assertions which would remain hidden textual big data which is in large. Once information is extracted the information is converted to a structured form which can be further analysed or presented directly using the clustered Tables, mind maps, charts etc. Text Mining services different variety of techniques one of the most is Natural language processing. The structured data created by text mining can be included into databases, data warehouses or business intelligence consoles and its used for descriptive, prescriptive or predictive analytics.[2]

2. LITERATURE REVIEW

Chandrasekhar Rangu, Shuvojit Chatterjee and Srinivasa Rao valluru of HCL Technologies Ltd Hyderabad , India in 2017 has suggested a model which enhances the product quality using the Text Mining Approach. To identify issues of a product related to text mining process the paper gives a good idea about different text mining techniques which they have used different natural language processing techniques which used to reduce the vocabulary size and to build forceful classifier. They have tried classification techniques like SVM Classifier, Bayes, Random Forest and based on the data knowledge which they gained they have chose SVM as best classifier techniques for data. The disadvantage of this approach is this classification model expects removing of characters of ten to fifteen words while doing pre-processing. So that smaller size review comments are not taken into consideration. [3]

Nikhita Mangaonkar and Sudarshan Sirsat in 2017, suggested a neuro linguistic programming method to calculate customer product experience. They introduced a set of technique required to review and filter product analysis, in a specific behaviour its to understand the customer experience for overall based on the text mining and neuro linguistics programming. The main aim is to provide a language a medium to communicate through which a customer can be profiled and retained for a life time, benefitting the organization business. While reading the views of customer/ product feedback is finding words which falls in the categories of visual description, auditory description and kinaesthetic description. [4]

Arun Manicka Raja, Godfrey Winster, Swamynathan S in 2016 suggested a analyser review system based on the sentimental words performance analysis for sentiment classification. From the experimental results on the datasets of digital camera review they have confirmed that the accuracy of existing classifiers at sentence level out performs Bag of Words(BOW) method. With these results of this method the patterns of adjective or adverb or

noun phrases combination can be extended by checking it with the individual comments weight which can be computed. Then the weight computed can be used to calculate the product models with the help of weight based ranking. [5]

Made Kevin Bratawisnu, Refi Rifaldi windya Giri, Rudi Rinaldi in 2017 suggested a analysis on text network, Based on the perception of consumers well the research methodology focuses on finding and compare and analyse the network relationship between words, sentences and system to model interactions which generate new knowledge or information modelling. The pros of network text analysis rather than other method like Multi – dimensional Scaling(MDS) to see lower cost of the consumer perception it reaches larger area than more real time. So that business enterprise can use to enhance CRM customer relationship management since it allows public opinion to be followed in business about its brand and respond quickly intervening to customer issues. [6]

3. DATA MINING OVERVIEW

Data Mining is a discovery process which is generally iterative and interactive. The main aim of the process is to mine patterns, associations, changes, anomalies and statistically find major structures from large amount of data. Moreover, the mined results should be valid, unique, beneficial and reasonable. These qualities which are placed on the process and outcomes of data mining are significant for a number of reasons, and can be described as following[1]

Valid – it is crucial that patterns, rules and models that are discovered are valid not only in data samples but it represent and remain valid in future new data samples that already examined. So that the model and rules obtained is considered meaningful. [1]

Unique – it is desirable with patterns, rules and models that are discovered are not known to already known experts. Other hand it would yield very little new understanding of the data samples and problem at hand. [1]

Beneficial – it is desirable that the patterns, rules and models are discovered and allowed us to take some useful actions. For an instance it allows to make reliable predictions on future events. [1]

Reasonable – it is desirable that the patterns, rules and models that are discovered makes an new overview on the data samples and the problems are being analysed. [1]

Figure 1: Data Mining Process

The general data mining process is represented in the above Figure 1.

1) Original Dataset – Understanding the data mining process and collect the information to make the data mining process to be successful and the target data set is achieved. [1]

2) Target Data – Data mining relies on the availability of suitable data that reflects the basic diversity, order and structure of the problem is being analysed. The target data is in proper structure [1]

3) Pre-processed Data – The pre-processed data removes the noisy data, outliers and missing value and it cleans the data [1]

4) Transformed Data – The data is transformed to another form and made it ready to find the patterns. [1]

5) Patterns – By applying data mining concepts can find the patterns from the transformed data. Its used to identify the best patterns. [1]

6) Knowledge – using patterns its deployed to find the proper outcome and identify the information or knowledge. [1]

4. DATA MINING TECHNIQUES

Data mining technique is process of uncovering the data mining patterns and finding anomalies and relationships in large datasets which can be used to make predictions about future trends. The main purpose of data mining is to make a outcome which have the valuable information from available data. [7]

There are different data mining technique in that seven major data mining technique is tracking patterns, classification, clustering, Association rule, outlier detection, regression, prediction. [7]

i) **Tracking Patterns** – tracking patterns is one of the basic data mining techniques in learning to recognize the patterns in data set. It is usually a recognition of some abbreviation in data happening at frequent intervals or flow of certain variable over a time. [7]

ii) **Classification** – classification data mining technique is the most difficult which forces to collect the different attributes together into visible categories which can be used to find a further conclusion. [7]

iii) **Association** – Association is related to pattern tracking, but it is more specific to variable linked which is dependent. [7]

iv) **Outlier detection** – In many cases this simply recognize the prime pattern can't give a clear understanding of data set. In this also need to be able to identify anomalies or data in outliers. [7]

v) **Clustering** – similar to classification is clustering but involves grouping chunks of data together based on the connections of data which is similar. [7]

vi) **Regression** – regression is basically used as form to plan and model. It is used to identify the likelihood of certain variable which is given based on the presence of other variables. [7]

vii) **Prediction** – prediction is one of the most valuable data mining techniques, since its used to project the types of data which will be viewn in the future. In many times its just to identify and understand the historical trends which is enough to chart a prediction of what will happen in advance in a accurate way. [7]

5. CLASSIFICATION TECHNIQUE

In this paper we will see about classification techniques of Naïve Bayes, J48, PART, KNN

5.1 Classification – Naïve Bayes

It is a classification techniques based on the bayes theorem with an thought of independence among predictors. [10] In simple terms a naïve bayes classifier predicts the presence of particular feature in a class is unrelated to the presence of any other feature. [11] Naïve bayes is known for the out performance of highly classy classification method. [12]

5.2 Classification – Lazy – KNN

K-Nearest Neighbour (KNN) is one of the simplest algorithms used in machine learning for regression and classification problems. KNN algorithm is used in data and classifies a new data points based on similarity measures. Classification is done by its neighbours vote in major way. The data is assigned to class which has the nearest neighbours. When the neighbours which is nearer is increased the value of K, accuracy might increase. [13] IBK is the linear neighbour search which we have used in weka tool. It implies the KNN Method.

5.3 Classification – Rules – PART

Rule based classification is used to set if then rules for classification. IF condition then conclusion. The IF part is the rule antecedent or precondition. [14] PART is the rule consequent. This rule consequent contains a class prediction. To extract the rule from the decision tree the root of a leaf node is created for each path.[15] The leaf node holds the class prediction forming the consequent rules. Rule induction using a sequential covering algorithm is sequentially learned from training data without having a decision tree first. [16]

5.4 Classification – Trees – J48

Venkatesan in 2015 proposed Quinlan's c4.5 algorithm which actualizes J48 to create a trimmed C4.5 decision tree. The every aspect of the information is to fragmented the information by choosing an attribute. [17] To review the attribute extreme standardized data gained is utilized. The algorithm is returned by minor subsets. [18] J48 develops a decision node utilizing the expected estimations of the class. J48 decision tree can be dealed with specific characteristics, lost or missing attributes estimations of the data and varying attribute costs. Here accuracy can be expanded by pruning.[9]

6. CLASSIFICATION ALGORITHMS

6.1 Algorithm of Naïve Bayes

$$\boldsymbol{P(c|x) = \frac{P(x|c)P(c)}{P(x)}}$$

$P(c|x)$ - Posterior Probability

$P(x|c)$ - Likelihood

$P(c)$ - Class Prior Probability

$P(x)$ - Predictor Prior Probability

The working principles of algorithm naïve Bayes is we have to change the data set into a frequency table then in next step we have to create a likelihood table by finding the probabilities like playing probability and windy probability then next step is we use the naïve Bayesian equation to calculate the posterior probability for each class. [20] The outcome of prediction which will be highest is the class with highest posterior probability. [21]

6.2 Algorithm of KNN – IBK

$$C_n^{1nn}(x) = Y_{(1)}$$

x – One nearest neighbour classifier is assigned to point x

Then we have to load the input data and initialize the K to chosen number of neighbours then for each instance we have to calculate the query example[22] and current example from the data and add the distance and index of the instances to an ordered group then we have to sort the ordered group collection of distance and indices from smallest to largest by the distances then we have to take the first k entries from the sorted collection then we have to get the labels of selected k entries and for classification we have to return the mode of the k labels. [23]

6.3 Algorithm of PART

C4. 5 algorithm is mentioned below

$$H(X)=\sum_{i=1}^{n} P(x_i)log_b\ P(x_i)= -\sum_{i=1}^{2}\left(\frac{1}{2}\right)log\left(\frac{1}{2}\right) = -\sum_{i=1}^{2}\left(\frac{1}{2}\right)x(-1) = 1$$

First separate and conquer the data then build a partial C4.5 decision tree. [24] Then from the partial decision tree best iteration is taken into rule at each time. Then identified the accuracy. [25]

6.4 Algorithm of J48

$$H(X)=\sum_{j=1}^{k} P_j log_2\ \frac{1}{P_j} = -\sum_{j=1}^{k} P_j log_2 P_j$$

X is the attribute and p is each element and j is position of each element of x.

H(x) indicates that attribute X is more random.

First step if the data instances belong to similar class the leaf is labelled with same class. [26] In next Step the potential data will be explored for each attribute and the gain in the data will be taken from the attribute test. The next process is best attribute is choosed last based on the current selection parameter. [27]

7. RESULT AND DISCUSSIONS

Using weka tool percentage split of a weather data set is analysed here and prepared a visualization charts and Percentage Split of Classification Correct accuracy value.

Classification Method	Correctly Classified Instances	Percentage Accuracy
Naïve Bayes	**9**	64.29%
IBK	**11**	78.57%
PART	**5**	35.71%
J48	**9**	64.29%

Table 1: Percentage Split Accuracy of correct instances

In this table 1 the classification methods is specified with the correctly classified instances and percentage of accuracy. By using weka tool we have identified the correct instances of classification methods and percentage accuracy of classification methods. For naïve bayes and J48 the instance and percentage is 9 and 64.2857 is the percentage accuracy. For KNN correct instance is 11 and percentage of accuracy is 78.57

7.1 Visualization Chart

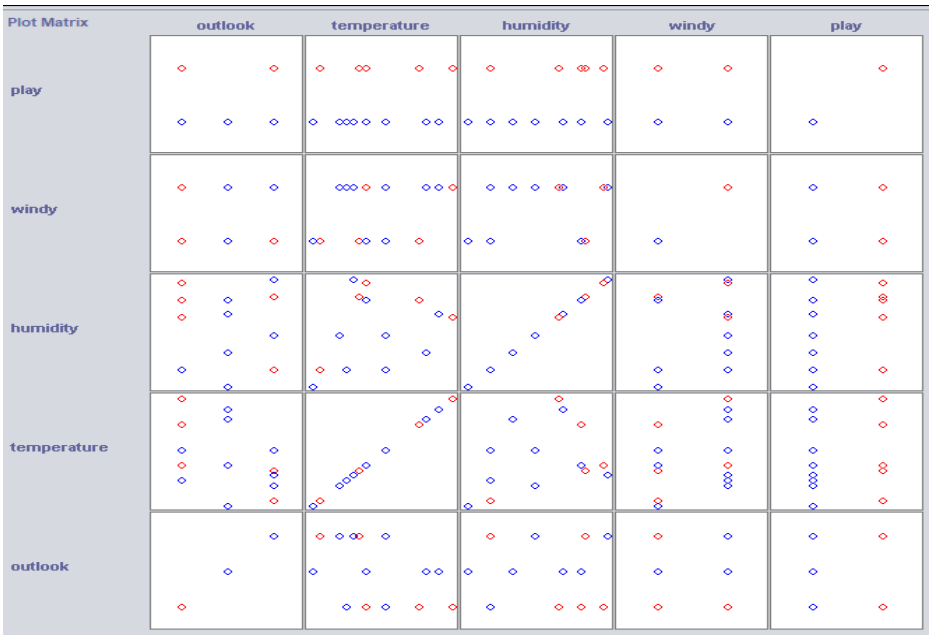

Figure 2: Visualization chart

In this visualization chart it represents the comparison plot using weka tool. For example, with plot chart in instance the outlook value is compared with temperature, humidity, windy and play which are yes is viewed in blue colour and when it is no it is represented by red colour.

7.2 Accuracy Chart

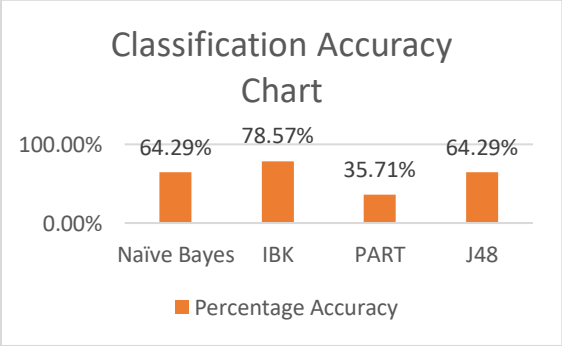

Figure 3 : Accuracy Chart

From the above Figure 3 we got percentage accuracy of each classification in Naïve Bayes its percentage of accuracy is 64.29%, 78.57% for IBK, 35.71% for PART and J48 64.29% from results we can identify PART gives the low accuracy while IBK gives the best accuracy results.

8. CONCLUSION

In paper we have compared and identified the results of weather dataset using weka and identified the correct instance and accuracy of each method. We have compared a dataset using classify method and analysed Naïve Bayes method, IBK, PART and J48. From that we have analysed and got a result in which IBK gives a best accuracy compared to all other classification method for our dataset. In weka tool the accuracy of the Naïve Bayes and the J48 are similar and PART gives the lowest accuracy value while IBK gives the best accuracy value. The best accuracy value which we have got from the classification algorithm is 78.57%.

REFERENCES

[1] Mohammed J. Zaki, Limsoon Wong, DATA MINING TECHNIQUES, August 9, 2003 12:10 WSPC/Lecture Notes Series: 9in x 6in.

[2] https://www.linguamatics.com/what-text-mining-text-analytics-and-natural-language-processing

[3] Chandrasekhar Rangu, Shuvojit Chatterjee and Srinivasa Rao Valluru, HCL Technologies Ltd Hyderabad, India, 2017

[4] Nikhita Mangaonkar and Sudarshan Sirsat, 2017. Neuro Lingusitic Approach to evaluate Customer Product Experience Based On Neuro Linguistic Programming.

[5] Arun Manicka Raja, Godfrey Winster , Swamynathan S, 2016.Review Analyzer System Based On Sentiment Analysis

[6] Bruce, R., and Wiebe, J. 2017. Recognizing Subjectivity: A Case Study of Manual Tagging. Natural Language Engineering

[7] https://www.datasciencecentral.com/profiles/blogs/the-7-most-important-data-mining-techniques

[8] Bharati M. Ramageri - Indian Journal of Computer Science and Engineering Vol. 1 No. 4 301-305

[9] N.saravanan,V.Gayatri International Journal of Computational Intelligence and Informatics, Vol. 7: No. 4, March 2018

[10] Deepali Kharche, K. R. (2014). Comparison Of Different Datasets Sing Various Classification Techniques With

[11] Modified WEKA. International Journal of Computer Science and Mobile Computing, IJCSMC, 389 – 393.

[12] Dr. Anjali B Raut, A. A. (2017). Students Performance Prediction Using Decision Tree Technique. International

[13] Journal of Computational Intelligence Research, 1735-1741.

[14] Gaganjot Kaur, A. C. (2014). Improved J48 Classification Algorithm forthe Prediction of Diabetes. International

[15] Journal of Computer Applications, 22.

[16] Gang Kou, Y. L. (2012). Evaluation Of Classification Algorithms Using Mcdm And Rank Correlation. International

[17] Journal of Information Technology & Decision Making .Jayasimman, L. (2015). Performance Accuracy of Classification Algorithms for Web Learning System. International Journal of Computer Applications.

[18] Kalpesh Adhatrao, A. G. (2013). Predicting Students performance using id3 and c4.5 classification algorithms. International Journal of Data Mining & Knowledge Management Process (IJDKP) .

[19] Kameswara Rao N K, D. V. (2014). Classification Rules Using Decision Tree for Dengue Disease. International Journal of Research in Computer and Communication Technology, 340-343.

[20] Minyechil Alehegn, R. J. (2017). Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble approach. International Research Journal of Engineering and Technology (IRJET) .

[21] Mohammed M Mazid, A. B. (2013). Improved C4.5 Algorithm for Rule-Based Classification. Recent Advances In Artificial Intelligence, Knowledge Engineering And Data Bases .

[22] Nagaparameshwara chary, S. D. (2017). A Survey on Comparative Analysis of Decision Tree Algorithms in Data Mining. International Conference on Innovative Applications in Engineering and Information Technology(ICIAEIT-2017) .International Journal of Computational Intelligence and Informatics, Vol. 7: No. 4, March 2018

[23] Prerna Kapoor, R. R. (2015). Efficient Decision Tree Algorithm Using J48 and Reduced Error Pruning. International Journal of Engineering Research and General Science .

[24] Sagar, N. S. (2015). A Comparative Study of Classification Techniques in Data Mining Algorithms. Oriental Journal Of Computer Science & Technology , 13-19.

[25] Srishti Taneja. (2014). Implementation of Novel Algorithm (SPruning Algorithm). IOSR Journal of Computer Engineering (IOSR-JCE), 57-65.

[26] Tina, R. P. (2013). Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification. International Journal Of Computer Science And Applications.

[27] Venkatesan, E. V. (2015). Performance Analysis of Decision Tree Algorithms for Breast Cancer Classification. Indian Journal of Science and Technology.