iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 6 **Issue:** VII **Month of publication:** July 2018

DOI: http://doi.org/10.22214/ijraset.2018.7066

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue VII, July 2018- Available at www.ijraset.com

Analysis the Data Mining Classification Algorithm

Disha A. Katariya[1] Prof. U.R.Gandhi[2]
[1, 2]*PG Department of Computer Science & Technology, Hanuman Vyayam Prasarak Mandal Amravati,Maharashta*

Abstract: Data mining deals with large voluminous data. Classification of data is an important task in data Mining process. Data mining is a process of searching data from a pool of data like database, web-servers, cloud based servers etc. Classification is method of generalizing the data consistent according to different instances. It is a technique which is used primarily for discovering unknown patterns and that converts raw data into user understandable information. There are different classification algorithms including k-nearest neighbor, naïve bays, and Decision tree algorithms. In this paper we study different algorithms of classification and there advantage & Disadvantage.

I. INTRODUCTION

Data mining is the technique of retrieving the data according to expectation from the collection of data using different Data mining techniques and algorithms. The gained data can be used for applications ranging from market analysis, fraud detection, and customer retention, to production control and science exploration.

Data mining could be a method of extracting or mining the useful pattern or information and relationships within massive amounts or huge volumes of data. The term data processing is additionally referred to as "Knowledge mining from data" [12].

Fig 1.1 shows the Knowledge discovery process in databases. Steps in KDD process id shortly explain as follows:

A. Data cleaning (To clear the data from noise and unwanted data
B. Data selection (where data relevant to the analysis task are retrieved from the database)
C. Data transformation (where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance)
D. Data mining (a process where methods are applied in order to extract data patterns)
E. Pattern evaluation (to identify the truly interesting patterns representing knowledge based on some interestingness measures)
F. Knowledge presentation (where visualization and knowledge representation techniques are used to present the mined knowledge to the user)

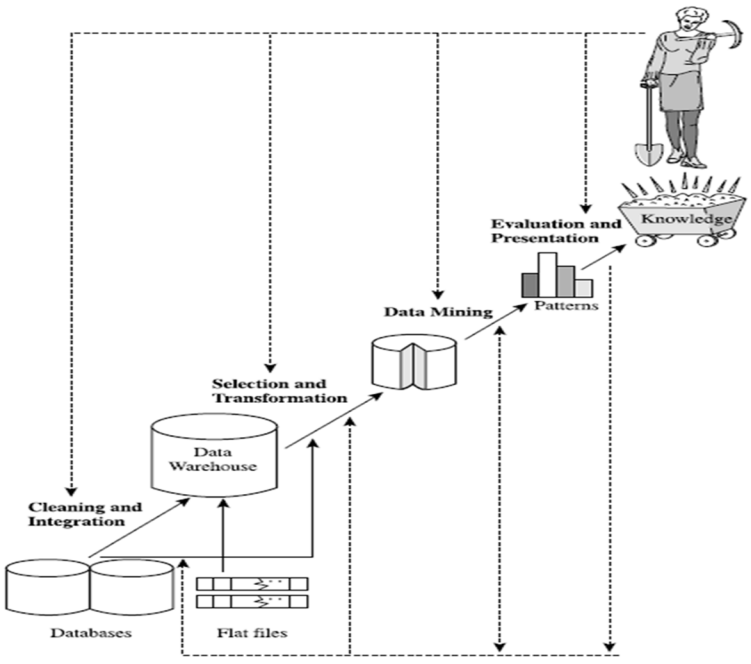


Figure 1.1 Data Mining as a step in the process of knowledge discovery

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue VII, July 2018- Available at www.ijraset.com

There are several data mining techniques like classification, association rule mining, regression etc. Classification of data is a vital task in data mining process. In machine learning, classification denotes to a step by step process for categorizing a given input data into one among the various categories.

II. CLASSIFICATION ALGORITHMS USED IN DATA MINING

Data classification is the process of formation of model or classifier to predict the categorical labels that is the labels of the distinguished data classes. These models so created are used to delineate the labels of the new input data or the data whose class labels are not known. Data mining techniques broadly classified into two categories. They are predictive and descriptive. Both of these methods are used to extract the hidden patterns from huge amount of data. Decision tree, classification rules or if then rules, mathematical formulae, neural networks etc. are customarily used to emblematize classification model.
Classification algorithms had involved significant attention in research areas of data mining [1]. Some of the famous classification models are:

A. ID3 (Iterative Dichotomiser 3)

Decision tree is powerful classification algorithm in data mining. There are several popular decision algorithms such as Quinlan's ID3, C4.5, C5.
ID3 algorithm begins with the first set because it is the root node. On each iteration of the algorithm, it going through every unused attribute of the set and calculates the entropy IG(A) of that attribute.
Ross Quinlan has proposed ID3 algorithm which is used to design the Decision tree to retrieve the data from the dataset. ID3 was forerunner to C4.5 algorithm. ID3 analyze all attributes of training dataset, this analyze the element from the training dataset.
The algorithm applies greedy search, it selects the finest attribute and earlier choices are not considered. [3]
ID3 algorithm selects the best attribute based on the concept of entropy and information gain for developing the tree.

1) There are some limitations of ID3 algorithm.

a) It is possibility of data over-fitted or over-classified, if a small sample is tested.
b) Only one attribute at a time is tested for making a decision.
c) Does not handle numeric attributes and missing values.

B. C4.5 Algorithm

Ross Quinlan developed C4.5 algorithm and produce decision tree. It is an expansion of Quinlan's earlier ID3 algorithm. C4.5 is frequently called as statistical classifier. From a group of training data C4.5 fabricate decision trees, in the similar manner like ID3. C4.5 is set of algorithms for accomplish classifications in data mining. It extends the classification model as a decision tree. [3]
The C4.5 algorithm made a number of changes to improve ID3 algorithm. Some of these are:

1) Handling training data with missing values of attributes
2) Handling differing cost attributes
3) Pruning the decision tree after its creation
4) Handling attributes with discrete and continuous values

C. K-Nearest Neighbor classifier

K-Nearest Neighbor classifier is also known as a distance based classifier. Nearest neighbor classifiers are based on analogy learning. So that means, it is comparing given test tuples with training tuples that are similar to it. The unknown tuple is assigned to most common class among its K-nearest neighbors. When K = 1, the unknown tuple is assigned the class of the training tuple that is closest in the pattern space. It is also used for numeric prediction, that is, it returns real-value prediction for unknown tuple. This method is also called the lazy learner method. [4]

1) Properties of K-Nearest-Neighbor Algorithm

a) KNN algorithm is simplest classification algorithm.
b) KNN algorithm gives highly aggressive results.
c) KNN algorithm can be uses for regression problems.
d) The algorithm was versatile algorithm and is used in many fields

D. Naive Bays Algorithm:

The approach used in Naïve Bayes classifier is very simple. With the help of small amount of training data it is possible to classify the given instances [13]. The naive bays classifier is statistical algorithm providing astonishingly higher results. Bayesian filter has been used widely in building spam filters. The Naïve Bays classifier is predicted on the Bays' rule of conditional probability.[6]
The approach used in Naïve Bayes classifier is very simple. With the help of small amount of training data it is possible to classify the given instances. For example to predict the fruit as "apple", based on the color red, and its shape round it is classified as apple which shows it as an independent model. This method is also suitable for complex situations.

III.ADVANTAGES AND DISADVANTTAGES OF CLASSIFICATION ALGORITHMS

ALGORITHMS	ADVANTAGES	DISADVANTTAGES
ID3	1. It produces the high accuracy result than the C4.5 algorithm. 2. ID3 algorithm typically uses nominal attributes for classification with no missing values.	1 It has long searching time. 2. It takes the more memory than the C4.5 to large program execution.
C4.5 Algorithm	1. It produces the correct result. 2. It takes the less memory to massive program execution. 3. It takes less model build time.	1. Empty branches. 2. Insignificant branches. 3. Over fitting.
K-Nearest Neighbor classifier	1. It is an easy to understand 2. Training is very fast. 3. Robust to noisy training data.	1. Memory limitation. 2. Being a supervised learning lazy algorithm
Naive Bays Algorithm	1.To improves the classification performance by removing the unrelated options. 2.Good Performance & short computational time	1. The naive bays classifier requires a very large number of records to obtain good results. 2. Threshold value must be tuned

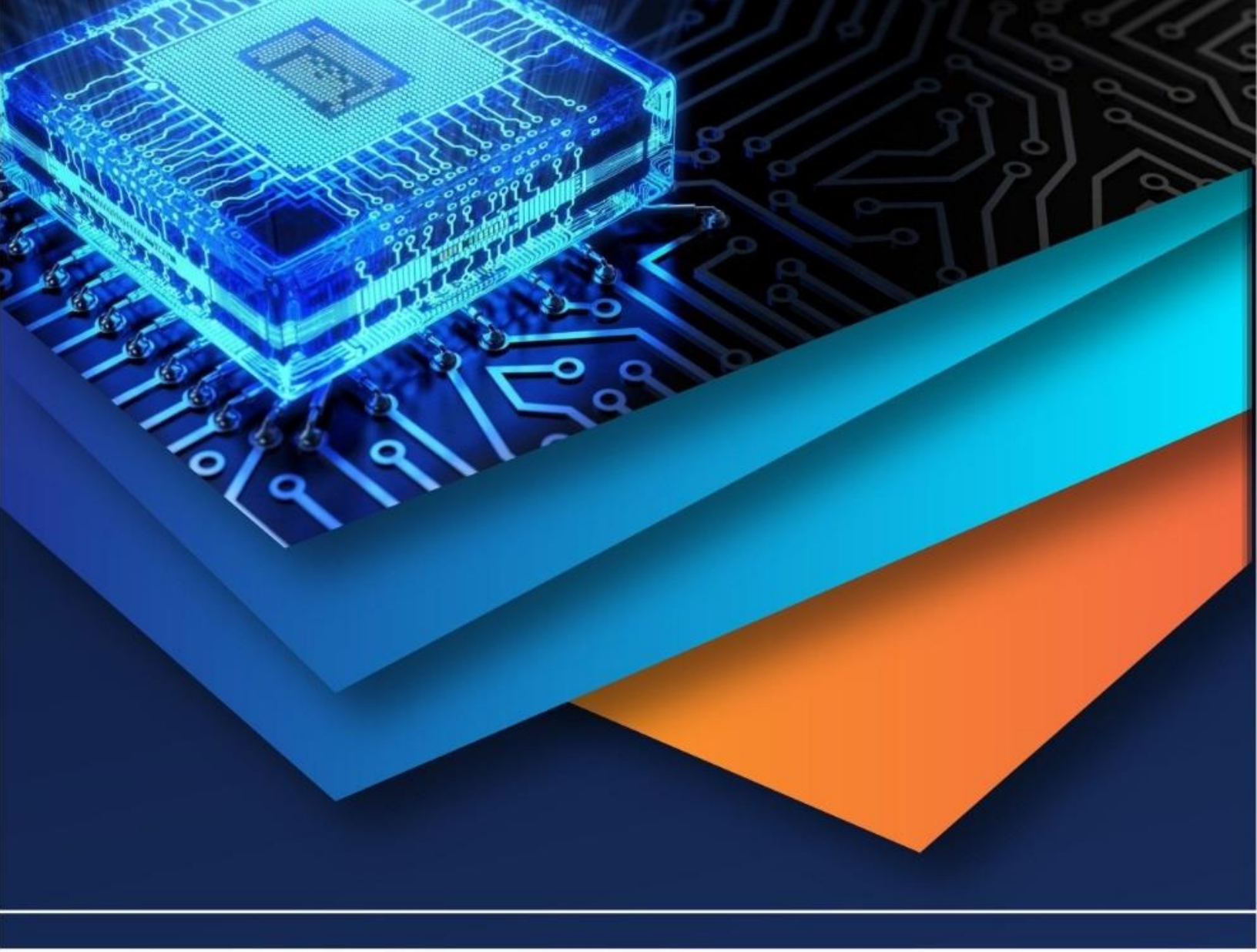
IV.CONCLUSION

Classification techniques have helped to mine the data in time constraint manner and have provided faster extraction of data. It has reduced the time required to extract the knowledge from the data warehouses. It has improved the efficiency of data extraction. Its techniques can be implemented in Medical fields, Business, Risk management, Financial corporations. It has made the data mining simple and easy by classifying the data first and then predicts the data for the analysis. This paper deals with numerous classification techniques employed in data mining and a study on every one of them. We study the Decision tree algorithms ID3 & C 4.5, K-Nearest Neighbor Classification Algorithm, Naïve bays Algorithm. We also discuss the advantage and disadvantage of technique. Classification methods are popular for extraction of data. Each of these methods can be used in various situations.

REFERENCES

[1] S.Umadevi, Dr.K.S.Jeen Marseline," A Survey on Data Mining Classification algorithms",International Conference on Signal Processing and Communication (ICSPC'17) – 28th & 29th July 2017.
[2] Y.Jeya Sheela S.H.Krishnaveni," A Comparative Analysis of various Classification Trees",International Conference on circuits Power and Computing Technologies [ICCPCT],2017
[3] Swathi Agarwal, G.L.Anand Babu, Dr.K.S.Reddy," Classification Techniques in Data Mining-Case Study," IOSR Journal of Computer Engineering (IOSR-JCE) Volume 18, Issue 6, (Nov.-Dec. 2016
[4] Mr. Chintan Shah , Dr. Anjali G. Jivani," Comparison of Data Mining Classification Algorithms for Breast Cancer Prediction," Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), July 2013.
[5] Viswanath, P. ; Sarma, T.H.,"An improvement to k-nearest neighbor classifier", Recent Advances in Intelligent Computational Systems (RAICS), 2011 IEEE , Sept. 2011.

[6] N. Chandra Sekhar Reddy, K. Sai Prasad and A. Mounika," Classification Algorithms on Datamining: A Study", International Journal of Computational Intelligence Research ,Volume 13, Number 8 (2017).
[7] R. Savundhariyalachmi, N. Pandimeena, P. Ramya," Study of Classification algorithm in Data mining", International Journal of Science and Research.
[8] S.K. Yadav, B. Bharadwaj, and S. Pal, "Data mining applications: A comparative study for predicting student's performance," International Journal of Innovative Technology and Creative Engineering, vol. 1(12), 2011.
[9] Shivam Agarwal," Data Mining Concepts And Techniques," International Conference On Machine Intelligence Research And Advancement,2013.
[10] Samiddha Mukherjee, Ravi Shaw, Nilanjan Haldar, Satyasaran Changdar," A Survey of Data Mining Applications and Techniques", International Journal of Computer Science and Information Technologies, Vol. 6 ,2015.
[11] Mustafa Agaoglu, "Predicting Instructor Performance Using Data Mining Techniques in Higher Education," IEEE Access , Volume: 4 ,2016.
[12] Jiawei Han, Micheline Kambar, Jian Pei, "Data Mining Concepts and Techniques" Elsevier Second Edition.
[13] R. Savundhariyalachmi, N. Pandimeena, P. Ramya," Study of Classification algorithm in Data mining", International Journal of Science and Research



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY