International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08, Special Issue | Oct 2021 www.irjet.net p-ISSN: 2395-0072

International Conference on Recent Trends in Science & Technology-2021 (ICRTST - 2021)

Organised by: ATME College of Engineering, Mysuru, INDIA

Survey on Algorithms of Data Mining

Lavanya N[1], Akshatha A[2]

[1,2]*Assistant Professor, Dept. of CSE, ATMECE, Mysuru*

Abstract— Data mining is a technical process where it analyses a bulk of scattered information to extract relevant information or data. This survey paper emphasis the 5 extremely recycled algorithms of data mining in the environment of research are :K-means, Apriori, NaïveBayes, Expectation Maximization(EM), KNN algorithm. A summary of each algorithm is disposed with a real time example and prospective of individual methods are discussed. These innovative methods stands at the leading edge in the field of data mining exploration and advancement such as classification, clustering, statistical learning, association analysis, and link mining.

Keywords— Data mining techniques, Text Analytics, Information Retrieval, Association rules.

Introduction

In other words, mining of data involves discovering new patterns and rules of association to determine the concealed knowledge from data bases which are derived from unique and various resources. The concepts of data mining can be applied in various fields, like science and research .The word data mining describes the process of information retrieval and discovering new patterns from the databases. It has some fundamental characteristics and they are:

- Automatic pattern predictions based on behavior analysis.
- Prediction based on likely outcomes.
- Creation of decision-oriented information.
- Focus on large data sets and databases for analysis.
- Clustering based on finding and visually documented groups of facts not previously known.

Fig: Techniques of Data Mining

SURVEY

A. K-Means Algorithm

It is one of the eminent clustering analysis approach in data mining which handles and extract relevant dataset. The process of arranging different objects into classes of similar objects is termed as clustering. K –means works on this clusters i.e,

Initially it creates k group of objects which forms the cluster. In other words, clusters are nothing but the group of synonyms.For example, if we apply k-means for the dataset consists of education related data means, student name, department, academic year, USN etc. K-means algorithm works as follows:

i. Clusters the data into *k* groups where *k* is predefined.

ii. Select *k* points at random as cluster centers.

iii. Assign objects to their closest cluster center

iv. Calculate the centroid or mean of all objects in each cluster.

v. Repeat steps 2, 3 and 4 until the same points are assigned to each cluster in consecutive rounds.

In the first step clusters are formed i.e, group of similar objects are formed .Based on the department of the student clusters are formed and the number of groups will be denoted as k. Then will select the k points as cluster centers. Then calculate the centroid of all clusters. The figure depicts the clusters formed before and after applying K-means algorithm. It is a vector representation.

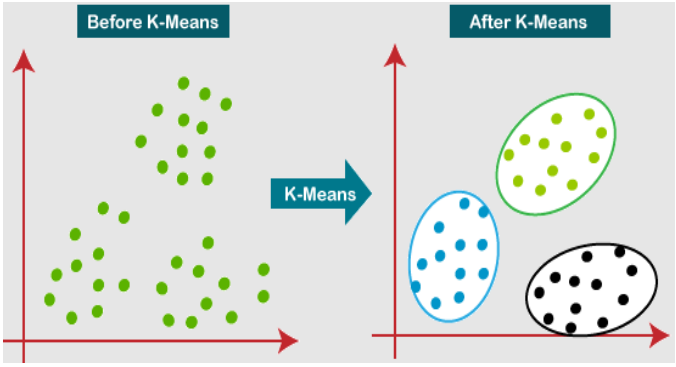


Fig: Cluster Analysis

In K-means algorithm implementation is easy. It calibrates large set of data. It assures convergence. It can be easily accommodated to new examples. These are some of the advantages of this algorithm. On the other hand one of main disadvantage of this algorithm is dependent on values i.e, k and it is hard to form clusters of varying size . The order of the data has an impact on the final results.

B. Apriori Algorithm

It is applied to a dataset containing a large number of transactions. The algorithm is based on frequent item set, association rules.

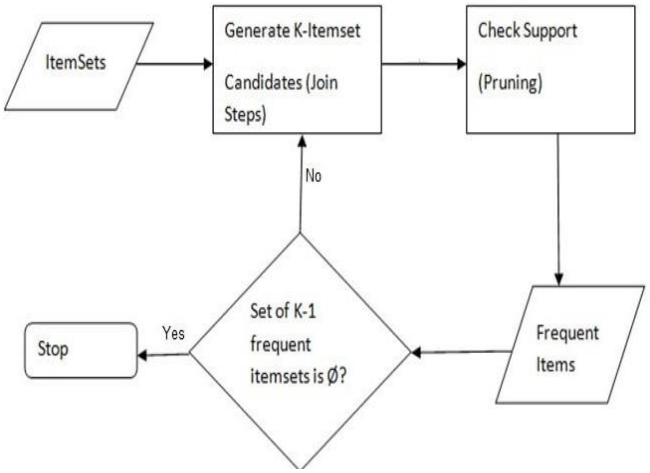


Fig :Apriori algorithm

Itemset is nothing but set of items. For example, if any item set consists of n-items then it is called n-items set. The set of items generated are known as frequent iteme set. The frequent item set must assure the least possible i.e, minimum threshold value for support and confidence. In this algorithm the support defines the actions with items that are acquired simultaneously in a single action. Confidence is defined by actions where the items are acquired or collected consecutively. The rules of association are applied to discover the relationship between the attributes. An association rule, A=> B, will be of the form for a set of transactions, some value of itemset A determines the values of itemset B under the condition in which minimum support and confidence are met.

$$\text{Support (A)} = \frac{\text{Number of transaction in which A appears}}{\text{Total number of transactions}}$$

$$\text{Confidence (A}\longrightarrow\text{B)} = \frac{\text{Support(AUB)}}{\text{Support(A)}}$$

This algorithm can applied data analysis, market basket analysis, etc. The figure depicts the example for apriori algorithm. It has transactions, frequent item set, Support using association rules. Apriori algorithm has pros and cons. Some of the advantages are: Apriori algorithm is the uncluttered and convenient in association with the other algorithms in data mining. The outcomes of this algorithm are spontaneous and are straightforward to get connection with the end user. An abundant development are expected for distinct use cases based on this implementation—consider this scenario, there are association learning algorithms that take into account the ordering of items, their number, and associated timestamps. The algorithm is exhaustive, so it finds all the rules with the specified support and confidence.

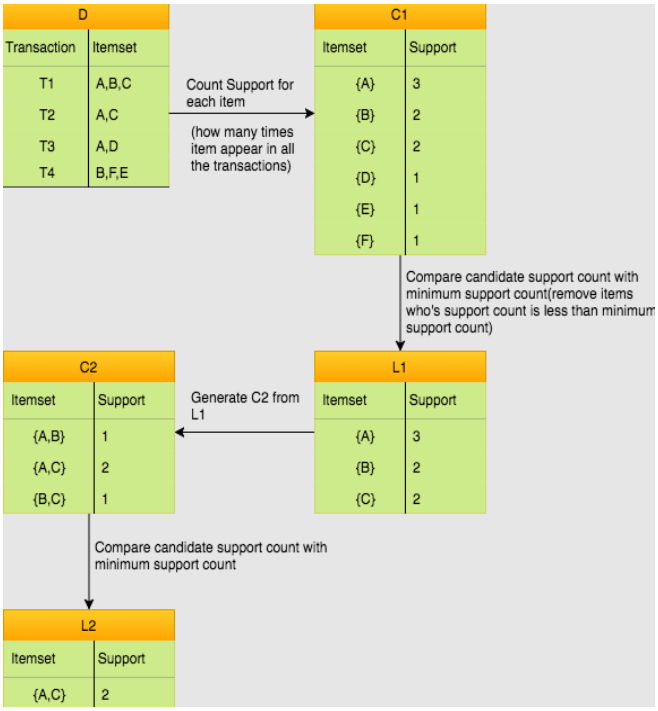
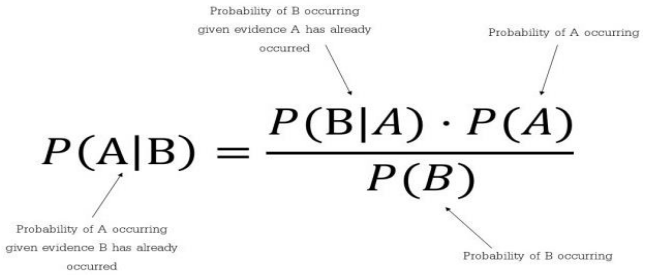


Fig: Example for apriori algorithm

C. Naïve Bayes Algorithm

In this algorithm we consider a classifier as unique characteristics of a class which is not abided by the existence of alternative attributes. For an instance, consider a car as a class which has the attributes color, model and brand. Here, we can observe that each of these attribute rely on each other or else the presence of the alternative attribute. All of these features individually provides support to the probability that this car belongs to some brand and that is known as "Naïve". Naive Bayes model is easy to build and particularly useful for very large data sets.



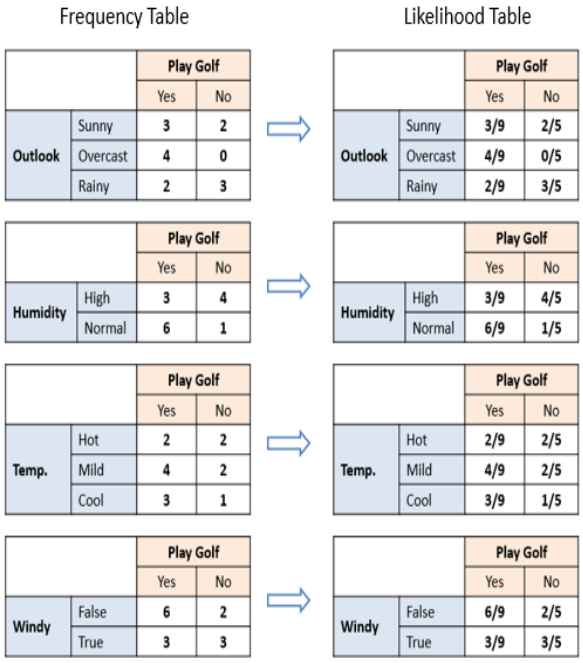


Fig: Example for algorithm

While we compare other algorithms with this naïve bayes algorithm, it is less complicated and easy to understand when we assume that the characteristics are independent. In few situation ,speed is favored over high accuracy. Naive bayes algorithm works adequately with large databases like classification of text, text analytics, email spam detection.

D. Expectation Maximization (EM) Algorithm

The Expectation Maximization (EM) algorithm is applied for the cases where the mathematical statement cannot be resolved at the first attempt .It is used to calculate the maximum similar parameters of a statistical model. These models requires variables from unknown resources, latent variables as well as the data known form the observations. It means the algorithm, one of the two i.e, the missing data which remain in the middle of the data or simply by overbearing the unrecognized points of data. For example, a mixture model can be described more simply by assuming that each observed data point has a corresponding unobserved data point, or latent variable, specifying the mixture component to which each data point belongs.

Finding a maximum likelihood solution typically requires taking the derivatives of the likelihood function with respect to all the unknown values, the parameters and the latent variables, and simultaneously solving the resulting equations. In statistical models with latent variables, this is usually impossible. Instead, the result is typically a set of interlocking equations in which the solution to the parameters requires the values of the latent variables and vice versa, but substituting one set of equations into the other produces an unsolvable equation. The EM algorithm proceeds from the observation that there is a way to solve these two sets of equations numerically. It is always guaranteed that likelihood will increase with each iteration. The E-step and M-step are often pretty easy for many problems in terms of implementation.

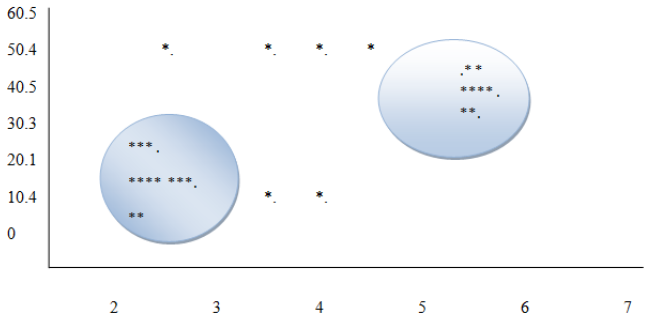


Fig: Example for EM algorithm

D. K Nearest Neighbor Algorithm

It is a classification algorithm which vary from the other class as previously explained it's as low-going algorithm. A slow-going will keep only the trained data during training process. It will categories when only a new unlabeled data is given as input. However, a fast learner builds a categorization as same as during training. Now the new un labeled data will be loaded, so now the data will feed into categorization model.

KNN works on two steps: First, this check which is the nearest labeled training data points. Second, it will check with the neighbors' classes, KNN now improve the way how the new data must be categorized. To finding the nearest data, KNN will take the help of distance metric like Euclidean distance. The way of deciding metric for the distance largely depends on how much data is been fed. Still few will suggest learning a distance metric based on the training data. Already detail study on KNN distance metrics has been done and so many papers are also present on the same. The data which is distinct, the way is to convert the data from distinct data into continuous data. Two examples which is present is: Using Hamming distance as a metric for the "closeness" of two text strings and now converting distinct data into binary format.

The Two known method for choosing the class is when the adjacent class is not having same class means take a simple majority vote from the adjacent class. Now check which class is having the greatest number of votes, will be the class for the new data point. In the same way, except this time hand big weight to adjacent that are nearer. The easiest way to find by using a reciprocal distance. Example, if the neighbor is 5 units away, then the weight of its vote is 1/5. As the neighbor gets further away, the reciprocal distance gets smaller and smaller. KNN is over see learning algorithm which provides the labeled training dataset.

Even this algorithm has some pros and cons: KNN algorithm is accurate when we consider distance as one of the attribute. It can be implemented easily and very easy to understand. The drawbacks are Noisy data can throw off kNN classifications. If we go with a large database KNN algorithm may be expensive in computation as we are attempting to find out the nearest neighbors. kNN generally requires greater storage requirements than faster classifiers. Selecting a good distance metric is crucial to kNN's accuracy.

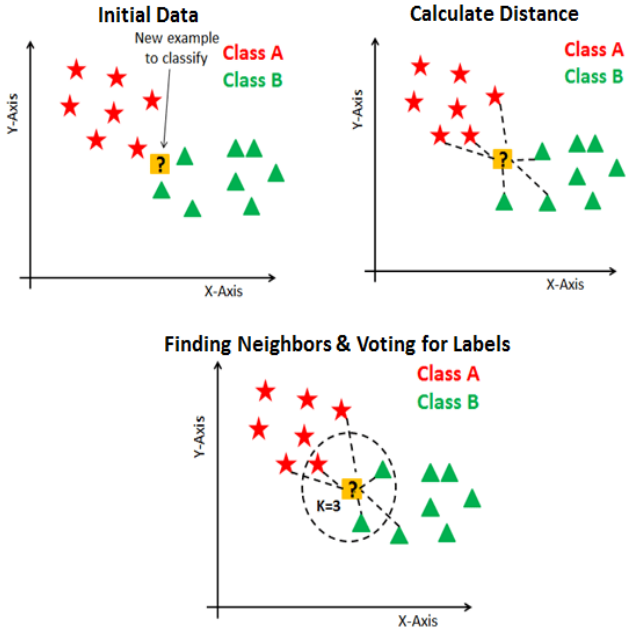


Fig: Example for KNN Algorithm

CONCLUSION AND FUTURE SCOPE

Data mining techniques and data mining involves the process of analyzing the data, identifying new patterns from a large set of data. To make efficient decision we have to analyze the result. The algorithms of data mining supports decision making. In future, the combination of data mining, artificial intelligence, machine learning performs the analysis over a large databases and which would impact to the field of research and development.

REFERENCES

[1] http://www.kdnuggets.com/2015/05/top-10-data miningalgorithms-explained.html

[2] Shafiq Aslam,Imran Ashraf International Journal of Advance Research in Computer Science and Management Studies Volume 2, Issue 7, July 2014 pg. 50-56

[3] www.mdpi.com/journal/applsci

[4] B.K. Bharadwaj and S. Pal. "Data Mining: A prediction for performance improvement using classification",

[5] International Journal of Computer Science and Information Security (IJCSIS), Vol. 9, No. 4, pp. 136-140, 2011.

[6] Botia, J.A., Garijo, M.y Velasco ,J.R., Skarmeta, A.F., "A Generic Data mining System basic design and implementation Guidelines" A technical Project Report of CYCYT project of Spanish Government .1998.

[7] website:http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.53.19353.

[8] Chapman , P.,Clinton, J., Kerber, R., Khabaza, T., "CRISP-DM 1.0 : Step by Step data mining guide,NCR Systems Engineering Copenhagen (USA and Denmark), Daimler Chrysler AG (Germany) , SPSS Inc. (USA) and UHRA Verzekeringenen Bank Group B.V (The Netherlands),2000".

[9] Lasore,D.T., "Discovering Knowledge in Data: An introduction to Data Mining", ISBN 0-471-66657-2, John Wiley & Sons, Inc,2005.

[10] Fayyad, U., Piatetsky-Shapiro, G., and Smyth P., "From Data Mining to Knowledge Discovery in Databases," AI Magazine ,American Association for Artificial Intelligence,1996.

[11] Bernstein , A and Provost, F., "An intelligent Assistant for Knowledge Discovery Process". Working paper of the Center for Digital Economy Research ,New York University and also Presented at the IJCAI 2001 Workshop on Wrappers for Performance Enhancement in Knowledge Discovery in Databases.

[12] Xindong Wu, Senior Member, IEEE "Data Mining: An AI Perspective" vol.4 no 2 (2004)

[13] Satvika Khanna et al. "Expert Systems Advances in Education" NCCI 2010 -National Conference on Computational Instrumentation CSIO Chandigarh, INDIA, 19-20 March 2010

[14] Kaijun Xu." Dynamic neuro-fuzzy control design for civil aviation aircraft in intelligent landing system. Dept. of Air Navig. Civil Aviation Flight Univ. of China 2011.

[15] Eike.F Anderson.,"Playing smart artificial intelligence in computer games" The National Centre for Computer Animation (NCCA) Bournemouth University UK.

[16] K.R. Chaudhary "Goals, Roots and Sub-fields of Artificial Intelligence. MBM Engineering College, Jodhpur, India 2012

[17] Text Analytics: the convergence of Big Data and Artificial Intelligence(2016)-Universal Autonoma de Madrid and Instituto de Ingeniería del Conocimiento, Madrid, Spain 2ZED Worldwide, Madrid, Spain.

[18] Information Retrieval Using Artificial Intelligence & Fuzzy Logic For Documents Through OCR(2015) – PandeyM.K And Nandan Singh.

[19] Google – Word2vec (2016): http://arxiv.org/pdf.