Classification Algorithm in Data Mining

B.V.Sudhakavya, Swathi.V
Students,
B.S in cloud computing and big data, School of Computer Science and Applications,
REVA University,Bangalore.
Dr. S. Senthil
Professor and Director,
School of Computer Science and Applications, REVA University, Bangalore.

ABSTRACT

Every day, Data generated from business, society, science and engineering, medicine and almost any other aspect of life. By using a large amount of data, we can identify the hidden knowledge by using the data mining process. Data mining consists of anomaly detection, association rule learning, clustering, classification, regression, and summarization. Classification is the main method in data mining and generally used in different field. Classification is a machine learning method used to predict group of data [1]. The main aim of this paper is to study the different classification algorithms in data mining. Classification algorithms are C4.5, ID3, k-nearest neighbor, Naïve Bayes, Support Vector Machine, and Artificial Neural Network. Normally a classification techniques, consists of three approaches Statistical procedure-based learning, Machine Learning and Neural Network.

Keyword: **Data mining, C4.5, ID3, ANN, SVM, k-nearest neighbor, Limitation and features of the classification algorithm.**

I. Introduction

Data mining is a tool that blends data analysis method with sophisticated algorithms for processing large volumes of data [2]. Classification procedures in data mining are equipped for preparing a large amount of data. It may be utilized to anticipate downright class labels and group of information dependent on the training set and class labels, it tends to be utilized for arranging new accessible data.

Age	Heart rate	Blood pressure	Heart problem
36	75	102/76	No
38	65	108/65	No

Training set

Age	Heart rate	Blood pressure	Heart problem
45	73	106/63	?
35	78	112/76	?

Test Set

Fig 1

IF (Age=65 AND Heart rate<70) OR (Age<60 AND Blood pressure<140/70) THEN Heart problem=no

Classification uses test set to find knowledge. Test set are expressed in the type of IF- THEN rules IF part consists of a conjunction of conditions and THEN part predicts a certain predictions attribute value for an item that satisfies the previous.

II. Classification Techniques

Classification techniques, consists of three approaches are Statistical procedure- based approach, Machine Learning based approach and Neural Network. All are attempted to develop the process that handles a large variety of problems and it is used in practical settings.

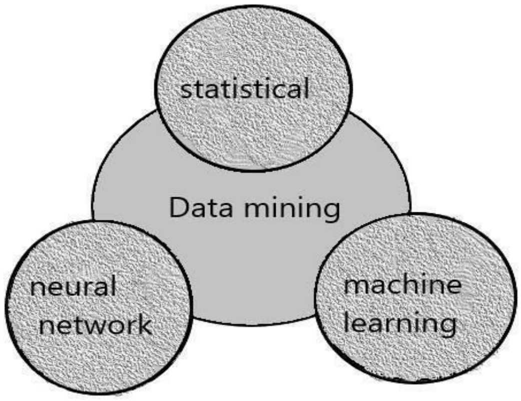

Figure 1.classification techniques in data mining

Statistical procedure-based approach:

In 2015 Sagar S. Nikam [2] had presented a statistical procedure-based approach, In that two primary phases of work on grouping can be perceived inside the factual network. In the first "classical" stage concentrated on direct separation.

In next step, "modern" stage is more concentrated on flexible classes of many models in which they try to present an approximation of the joint allocation within each class features to provide a classification rule [3]. Statistical procedures-based approach is normally characterized by having a defined basic probability model which provides a probability of being in every class. There was an expectation that the procedures will be utilized by analysts and consequently.

Machine learning based approach:

Machine learning based approach is usually worried about programmed processing methodology dependent on consistent or twofold activities that take in an assignment from a progression of precedents. Whereas in this, we are simply focusing on characterization thus consideration has concentrated on choice tree techniques in which grouping results from a request of intelligent advances.

Characterization results can imply the most compound issue given adequate information. Different methods, for example, hereditary calculations and inductive rational methodology (ILP) are as of now under dynamic advancement and its standard would allow us to manage progressively broad sorts of information covering situations where the amount of qualities may shift. Machine learning-based technique means to produce ordering terms sufficiently straightforward to be seen effectively by the human and must copy human thinking enough to give understanding into the outcome procedure. Comparable measurable methodologies foundation learning might be utilized being developed, however, activity is normal without human impedance.

Neural network:

Mr. Sudhir M [4] clearly explained about the neural network in data mining, neural network used a gradient descent method based on a biological nervous system having many organized processing elements. This is called as neurons [4]. Rules are extracted from the trained neural network to improve the interoperability of the learned network. neural network is used to solve the neurons problems which are organized processing elements [4].

neural network systems consist of interconnected hubs where each hub transport a non-straight efficiency of its data and giving to a different hub or straightforwardly from the data.

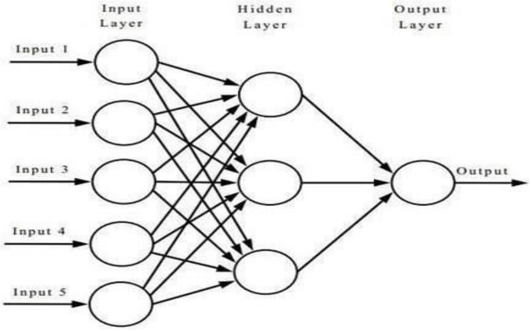

Fig 2: Neural network

Based on this model there are diverse applications for neural systems that include perceiving examples and settling on straightforward choices about neural system. During planes, we could utilize neural system as a fundamental autopilot where peruse signal from the different instruments and yield units, adjusting the plane's control properly and securely. Inside an industrial facility, we could utilize neural system for excellence power.

III. Classification algorithms

The main aim of this classification algorithm is to maximize the predictive accuracy [3]. classification is the supervised technique in which every instance belongs to a class. There are some of the classifications algorithms can be explained below,

Id3 Algorithm:

Sagar S. Nikam [2] had presented the Id3 algorithm, Id3 figuring starts with the main set of the root focus point. For each cycle of the estimation, Id3 algorithm underscores through each unused characteristic of the sets and figures of that entropy (data get IG (An)) of that properties. By using the quality which has deferential entropy regard, the set is S by then the part picked by quality to convey the subsets of the data. The computation proceeds to recurse on everything in the subset and taking into account, not to pick.

C4.5 Algorithm

SAGAR S. Nikam [2] gave a clear explanation about the C4.5 algorithm, C4.5 algorithm is used to convey an decision tree which is an augmentation for prior ID3 algorithm count. This improves the ID3 algorithm estimation by overseeing both steady and separating properties, losing the characteristics [2]. The decision trees made C4.5 algorithm to be used for the social affair along with the suggested accurate classifier. C4.5 algorithm settles the decision trees from a great deal of making data ready as same as Id3 count. when it is directed learning computation, it consist of plenty of planning points of reference which can be seen as a couple: input article and perfect yield regard (class).

The count separates the arrangement set and develops a classifier that must be able to decisively sort out both planning and investigations. A test point of reference is a data object and the figuring must envision yield regard. Consider the model getting ready instructive file S=S1, S2,...Sn which is starting now. Every precedent Si contains quality vector (x1, i, x2,i,..., xn, i) where xi addresses the attributes or features of the model and that class where Si falls. In middle, tree picks one normal for the data and most profitable parts, its course of action of tests into subsets of degree.

K Nearest Neighbors Algorithm

Srinivasan.B [4] presented the k- nearest neighbor algorithm. K-nearest neighbors (KNN), a classification scheme based on the expanse measure. The K- nearest neighbor technique assumes that not only the data in the training set but also the chosen classification for every data [4]. When classification is ready for a new data, then the distance to each data in the training set must be determined. Only the K nearest data in the training set are considered for further process[4]. The new data is placed in the class label that contains every data from this set of K-nearest data [4,5].

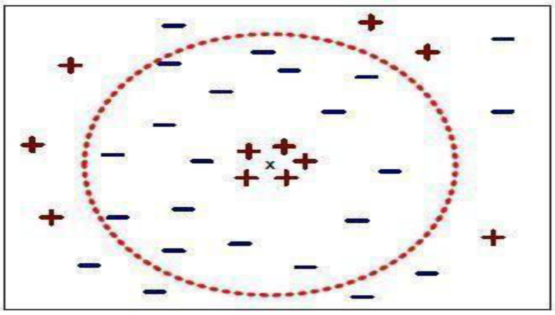

Fig 3: **K-Nearest Neighbors**

When we split our data into training and testing sets, we first normalize the data and then split it, Shown in fig 3[7].

SVM Algorithm

SAGAR S. Nikam [2] gave a clear explanation about the SVM algorithm; SVM has pulled into the big deal of thought in the latest decade and adequately associated with a variety of space application. SVM algorithm are routinely used for learning request, backslide or situating limit. SVM algorithm rely upon quantifiable learning speculation and fundamental danger minimization focal and have the purpose of choosing the territory of decision restricts generally called a hyper plane that provide a perfect parcel of class [2]. Expanding a edges and as such making the greatest possible division between the hyper plane and events on any side of it, it had exhibited to diminish an upper bound on the ordinary hypothesis bumble. The capability of SVM algorithm doesn't straightforwardly depend upon the part of portrayed substances. Regardless way that SVM algorithm is the mainly generous and accurate request policy, and also a couple of issues.

Information examination in SVM algorithm depend upon bent quadratic programming and quadratic programming strategies need huge framework assignments comparatively as monotonous numerical figuring’s [2].

Preparing for SVM algorithm scales quadratically in the measure of models, so gets some information about endeavor all the ideal open entryway for logically incredible arranging count, understanding a few assortments based estimation SVM can like way to be extended out to study non-direct choice, cutoff points by original anticipating information onto a high- dimensional segment space utilizing part works and describing a straight assembling problem in that include space.

The successive part space is a lot more noteworthy than the scope of the data set which is doubtful to store recognizable PCs. Original idea of disintegrating philosophy is to be the part of different portions, a set of free factors also called as working set, which can be resuscitated in each cycle. This technique is emphasized until the last condition is met. At first, the SVM algorithm was made for double assembling, and it is very hard to broaden for multi-class strategy issue.The multi-class issues consist of two or three class issues that would be in direct utilizing a couple of SVMs.

ANN Algorithm

Manikandan.S [8] expressed his taught by presenting the ANN algorithm, the artificial neural networks are used a large number of inputs to provide the estimated function which is generally unknown [8]. In the artificial neural network, there are many organized neurons which do the operation in the input.

The neural network performs the operation by connecting different nodes in the biological brain [8]. These nodes are constructed by a digital computer system.

The nodes are combined together and perform different layers and input is received by the input layer and output is produced. The neural network is the multilayer approach [8]. The process of specific data in a neural network is a mathematical function. If suppose the variables are a week in the data set, the neural network will perform improved when compared to another classification algorithm.

IV. CONCLUSION

This paper revolves around various request frameworks used in data mining. Data mining are used in a wide zone that consolidates techinques from a variety of fields includes machine learning, Network intrusion area, spam isolating, and man-made thinking, bits of knowledge and model affirmation for examination of far-reaching volumes of data.

Request systems are regularly strong in showing correspondences every one of the strategies can be used in different conditions as required, where one will be significant and the other will not or different way.

These portrayal estimations can be executed on various sorts of educational accumulations like money related information, and so on. In this manner, these request frameworks show how data can be settled and assembled when another course of action of information is open. Every technique has got its one of a kind segment and obstructions as given in the paper.

V. REFERENCES

[1]. Kesavaraj.G and Sukumaran.S," *A study on classification techniques in data mining"*, IEEE-31661.

[2] S. Nikam," A Comparative Study of Classification Techniques in Data Mining Algorithms", ORIENTAL JOURNAL OF COMPUTER SCIENCE & TECHNOLOGY, Techno Research Publishers, ISSN: 0974-6471, April 2015, Vol. 8, No. (1)

[3] D. Michie, D.J. Spiegelhalter, C.C. Taylor "Machine Learning, Neural and Statistical Classification", February 17, (1994).

[4] Sudhir M. Garade, Ankit Deo, and Preetesh Purohit, "A Study of Some Data Mining Classification Techniques", International Research Journal of Engineering and Technology (IRJET) e- ISSN: 2395 - 0056 p-ISSN: 2395-0072 Volume: 04 Issue: 04 | Apr - 2017.

[5] Srinivasan.B and Pavya.K," A Comparative Study on Classification Algorithms in Data Mining", IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 3 Issue 3, March 2016, ISSN 2348 – 7968.

[6] Suguna.N, and Thanushkodi.K," An Improved k-Nearest Neighbor Classification Using Genetic Algorithm", IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 4, No 2, July 2010.

[7] Kshitiz Sirohi," K-nearest Neighbors Algorithm, "K-nearest Neighbors Algorithm", https://towardsdatascience.com.

[8] LourduCaroline.A, Manikandan.S and Kanniamma.D, "Comparative study of Classification algorithms for Data Mining", International Journal of Engineering Science Invention (IJESI) ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726.