International Journal of Engineering Science Invention (IJESI)
ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726
www.ijesi.org || PP. 07-11

Comparative study of Classification algorithms for Data Mining

A.LourduCaroline[1], Dr.S.Manikandan[2], D.Kanniamma[3]
[1]*Research Scholar, Department of Computer Science, Mother Teresa Women's University, Kodaikanal (caroethan@yahoo.co.in).*
[2]*Professor and Head, Department of Computer Science and Eng., Sriram Engineering College, Chennai (manidindigul@gmail.com).*
[3]*M.Phil Scholar, PG and Research Department of Computer Science, St.Joseph's College of Arts and Science(Autonomous), Cuddalore (kaniamma26@gmail.com).*

Abstract: *The classification algorithm is used to classifying data into dissimilar classes based on some condition. They are many algorithms like the support vector machine, ID3, C4.5, decision tree, neural network, Bayesian algorithm, k nearest neighbor's algorithm. Classification algorithms are used to compare the data set. The goal of this paper is to give the detail view of different classification methods in data mining.*

I. Introduction:

Nowadays there are hugequantity of data are used which are together and stored in the database. The ratio of database is high when year grows. To overcome this problem the data mining has using many methods like classification, association rule, clustering etc. This paper will focus on classification algorithm.

The term classification is clearly saying that data are divided by different part according to its nature characters. The classification is the task of data mining. In data mining the classification technique is processing in large data set. It is used to classify the data depending upon instruction set and class name or label, and it is also used to save the class label, and the newly available data also classified. The classification procedure is frequently using in newly decision making situation. The classification procedure is created from the set of data in which the class is already known is treaded as pattern restructuring. The three main research patterns are: *statistical, machine learning and neural network.* This all have some same objectivities.

Statistical procedure based approach:

The statistical procedure method is used in classes and some human process in variable collection and transformation and the structure of problem. There are two stages used in statistical procedure according to classification process. They are classical phase and modem phase.

Machine learning based approach:

The machine learning is used in computer operation like binary and logical process. This machine learning process is understood by human, because of classifying expressions. It is like statistical procedure, that knowledge is developed without the human interface.

Neural network:

The neural network is containing layers, which are interconnected by nodes, and that each node is producing the non-linear data, as input for the next node. The output is generating for network by some other nodes.

II. Discription Of Classification Algorithm:

The classification is the one of data mining technique; this classification is used to analysis the data sets and classify the data sets by the instance and separate the classes according to the instance of data sets. The classification also used to absorb the models which are important data in the data set. The classification has two types of process, firstly the models from data set or trained data set, developed by the applying of classification algorithm, and second process is the model is verified to the predefined model to get accurate and perfect data set or test set. The training set contains set of attributes. The process of classification is to find the model from the trained data set. The classification algorithm produces the relationship between two attribute. The main goal of classification algorithm is to maximize the accuracy ratio of the data set.

1. ID3 Algorithm:

The ID3 algorithm starts the process in the root node. If there is a set of values which is named as A, and the A is split to any condition as (age<25, age < 50, age >=50) and produce the subset for the information. The algorithm is used in the subset process. The subset will have the conditions like:

- Each element of subset will be positive or negative, and then it is named as class example.
- If there is no attribute is select3ed in the example, then it will be termed as most common class.

The working process of algorithm is:

a. Count the entropy of the attribute in the data set A.
b. Classify the data set A into two parts according to minimum and maximum values.
c. Do the decisions tree process in the attribute of data set.
d. Repeat the process on the3 each attribute of the subset.

The ID3 algorithm is processed in decision tree. The ID3 algorithm is based on supervised learning algorithm; it is trained by different classes. The ID3 attributes are different from one class to another class.
Advantages: The result is more accurate when compared to C4.5 algorithm. The memory space consuming is very low.
Disadvantages: Require more time for searching data. The ID3 is very sensitive when large number of data is processed.

2. C4.5 Algorithm:

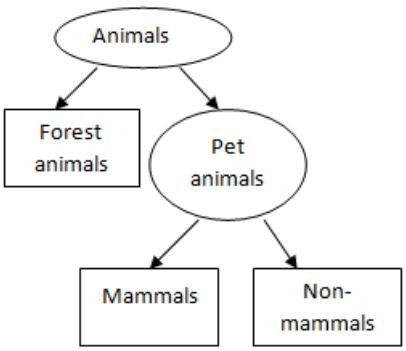
The C4.5 algorithm is used in decision tree which is advance of ID3 algorithm. The C4.5 algorithm is handling both continuous and discrete properties, missing values and pruning trees. The C4.5 algorithm creates the decision tree from the set of trained data. The algorithms access the trained data set and produce the trained and tested data.
The working process of C4.5 algorithm

a. If the attributes in the trained data set is belonged to same data set create the left node for the decision tree.
b. If no features are provided in data set the C4.5 creates the decision node and gives the expected value to the class.

Advantages: The implementation is very easy. It deals with noise data. The defined model is used early. The discrete and continuous values are used.
Disadvantages: Many decision treesare used to give small changes in data. The working process is slow in the data set and it is over fitting.

3. Decision Tree:

The decision tree in data mining technique is generated graphical design. The model is generated in decision tree is predictive or descriptive model. The decision tree handles large dimensional data.



The flow chart of decision tree is tree structure, in that the internal node is the test on the attribute, and branch is the output of the test, and leaf node is class name. The first node is root node. The decision tree uses the multidimensional data. The decision tree is simple and fast for learning and classification process. The decision tree algorithm in classification process is used in many areas like medicine, manufacturing and production molecular biology, financial analysis.

- A root node: Here there is no incoming edge, and has zero or more out coming edges.
- Internal node: It has exactly one incoming edge and two or more outgoing edges.
- Leaf node: The each node has one incoming edge but no outgoing edge.

4. K – Nearest Neighbors Algorithm:

The nearest neighbor is also called as the closest neighbor rule discuss the classification of unknown data which point to the nearest known data classes. The estimation of K indicates how many nearest data are

contains the same data point. The KNN utilize the nearest data in which they are belonging to same class. This data are stored in the memory of the system. The weight of the trained data is depend on distance of same data. The memory size is raised when the data set are stored in the system. The nearest neighbor training data set is utilizing the different system to overcome the memory limitation. The KNN implementation is done in ball tree, K-d tree, orthogonal search tree, principal axis search tree, nearest feature line (NFL).

The algorithm for KNN:

K-> number of nearest neighbors.

For each object X in the test data do the following process

Calculate distance D(X,Y) between X and every object Yin the trained set.

Neighborhood-> the K neighbor in trained set is X

Then X .class-> select class is neighbor

End for.

The KNN is first described in 1950, and get popularity in 1960. This KNN is widely used in pattern recognition. The nearest neighbor classifier is used for comparing the test tuple to training tuple. The training tuple is described by N attributes. There is an N-dimensional values in the attributes, the Euclidean distance between the two tuples X1 and Y1.

$$d(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Advantages: The class are not be in linear separable. The learning process is free. Multi model classes are well suitable.

Disadvantages: More time is taken to find the nearest neighbors data in large data set. The performance of algorithm is low when large data set is used.

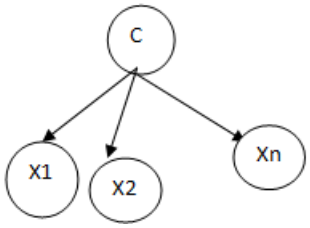
5. Bayes Classification Algorithm:

The Bayesian classifier is also called as statistical classifiers. They will predict the membership of the probability belong to particular class. The Bayesian classification is based on baye's theorem. The naïve Bayesian classifier is performed on decision tree, select the neural network. The Bayesian classifier is high accuracy and speed when applied in large data set also. The theorem is

$$P(H|X) = \frac{P(X|H)\,P(H)}{P(X)}$$

Let X is an data tuple. Let H be the hypothesis of the data tuple X. in classification problem we find the P(H|X) (i.e.) the probability of H hold the data tuple X. the P (H|X) is the posterior probability. The P (H) is the prior probability. The P (H|X) is posterior probability, which has many information then prior probability P (H). Similarly, P (X|H) is the posterior probability and P(X) is the prior probability.

Thus the baye's theorem is useful in providing the calculation of posterior probability P(H|X) from P(H), P(X|H) and P(X).



The principal of Bayesian algorithm is to count the each class in the trained data set.

Advantages: The implementation is very easy. It predicts the accurate result from classification problem.

Disadvantages: The accurate result of algorithm is decreased when the data is low. To get good result it require the large amount of data.

6. Support Vector Machine:

The SVM method for classification is done in both linear and non linear data. The SVM is used in numeric prediction and classification. The number of areas used in SVM are, hand written digit recognition, object recognition and speaker identification and in bench mark time series prediction tests.

The SVM are highly used in learning the classification and regression. The SVM is the statistical learning theory. The efficiency of SVM is not directly depending on classified entities. The accurate classification technique is applied in SVM but still it is robust in solving the problem. The convex quadratic

programming technique is used in SVM for analyzing data. But it takes more computationally expensive and time consuming is more.

The main goal of SVM is to find the perfect classification function, which is differing from the two classes in the training data. The key to find the best classification function is geometrically.

Advantages: it is high in accuracy level. The speed and size of process in test and trained data is high.

Disadvantages: the memory space for storing is high.

7. Artificial Neural Network:

The artificial neural networks are used large number of input to provide the approximate function which is generally unknown. The artificial neural network is the type of computer architecture. In artificial neural network there are many interconnected neurons which do the operation in the input. The neural network performs the operation by connecting different nodes in the biological brain. These nodes are constructed by digital computer system. The nodes are combined together and perform into different layers. The input is received by the input layer and the final output is produced.

The neural network is the multilayer approach. To process the specific data in neural network the mathematical function. If suppose the variables are week in the data set, the neural network will perform better when compare to other classification algorithm.

The neural network has the capacity of predicting the new data from the existing data. In neural network the nodes are connected to each other. The activity of the node is depending on the activity of other node and weight is connected at the edge. The node in neural network is arranged in layer form. The number of layer and number of nodes are same.

Dataset	No. of Variables	No. of Nominal Variables	No. of Continuous Variables	Classes	No. of Records
Adult	14	8	6	2	45222
House	8	0	8	3	25460
Credit	20	16	4	2	1000
Segment	19	0	19	7	2310
Vehicle	17	0	17	4	846
Cars	6	4	2	4	1728
NHANES	33	16	15	2	1346
White Wine	11	0	11	7	4898
Red Wine	11	0	11	6	1599

Advantages: It is easy to use and implement. It is capable of handling wide range of problems. It can handle both categorical and continuous data types. It provide the good result even in complicated domain.

Disadvantages: The input data is in the range of 1 and 0. It took long time to understand the data. The processing time is high because of large amount of data.

III. Exixting System:

In there are several data sets are used to compare the algorithms of classification in data mining. The data set is selected by number of record in class ratio, number of record in data set and class size. The data set is:

The c4.5 algorithm is better in five when compared to eight data set. The nb is best in cars an nn is best in vehicles. Then the result is:

Data set	Id3	C4.5	Dt	Knn	Ann	Svm	Nb
Adult	82.94	86.24	84.89	87.76	85.44	84.62	66.98
House	66.96	90.97	73.38	71.11	75.03	69.92	64.05
Credit	61.33	64.33	66.67	71.67	74.00	70.33	62.33
Segment	89.25	97.40	96.19	93.59	96.01	92.59	76.08
Cars	86.57	92.82	87.04	83.33	91.44	92.59	93.52
White wine	50.49	66.01	54.90	54.90	57.35	60.40	47.47
Red wine	57.39	67.67	55.14	59.90	57.39	60.40	56.14
vehicles	67.02	75.53	70.92	80.85	83.33	76.66	12.41

IV. Proposed System:

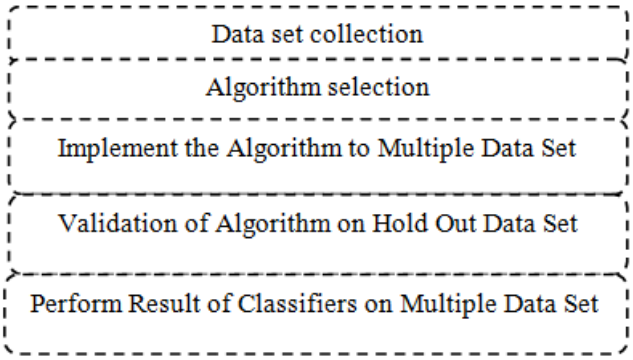
Hear I am going to use the student data set to compare the algorithm of classification. The data set of student is simple with classes as name, age, DOB, address, percentage.

By using the percentage the student are compared according to their family situation. The data set is:

Name	Age	DOB	Address	Persentage
Anu	21	1.1.95	nerust	80
Arun	20	2.3.96	Nerust	85
Anand	21	4.5.95	Kkst	75
Bala	21	6.7.95	Kkst	55
Basker	21	4.4.95	Nerust	90
Ramesh	22	9.9.94	Kkst	78
Suresh	21	7.8.95	Nerust	86
Sachin	21	6.8.95	Nerust	79
Vijay	22	5.6.94	Nerust	85
Vishnu	21	9.7.95	Kkst	67

By using this data set we can analysis the performance of the student by using the percentage. The village side student is low when compared to the city side student. The accurate ratio is find by comparing the perfect algorithm to the data set.

The methodology framework:



V. Conclusion:

Thus the classification algorithm is used in the student data set will give the perfect result according to my idea. There are many algorithm are presented in this paper for best use. Many advantages and disadvantages of the algorithm for the perfect analysis.

References:

[1]. **A Comparative Study ofClassificationTechniques in Data Mining AlgorithmsSagarS.Nikam** (Received: February 16, 2015; Accepted: April 10, 2015)
[2]. **Predictive modeling of trust to social media content Samuel Daniel Faculty of Science and Technology** University of Stavanger June 2014
[3]. **Data Mining Classification**:-FABRICIO VOZNIKA and LEONARDO VIANA
[4]. International Journal of P2P Network Trends and Technology (IJPTT) – Volume 4 Issue 8- Sep 2013ISSN: 2249-2615 http://www.ijpttjournal.org Page 369**Classification algorithm in Data mining: AnOverview**S.Neelamegam,Dr.E.Ramaraj
[5]. **An Exploration of ClassificationPrediction Techniques in Data Mining:The insurance domain**BySamuel OdeiDansoSeptember, 2006
[6]. **Comparing Classification Algorithms in Data Mining** -Sampson Adu-Poku= FEB 2012
[7]. www.googlescholer.com.
[8]. J. Han and M. Kamber, "**Data Mining Concepts and Techniques",**Elevier, 2011.