IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Volume-4, Issue-6 E-ISSN: 2347-2693

Analysis of Classification Technique Algorithms in Data mining- A Review

S. Nagaparameshwara Chary[1*], Dr. B.Rama [2]

[1]*Department of Computer Science& Applications , Govt.Degree College, Satavahana University, TS India*
[2]*Department of Computer Science, Kakatiya University, TS India*

Available online at: www.ijcseonline.org

Received: May/15/2016 Revised: May/23/2016 Accepted: Jun/12/2016 Published: Jun/30/ 2016

Abstract: Data mining is the one of the most important research area in the field of Computer Science. By using Data mining techniques we can extract the hidden patterns from large amount of data. The Data mining is the process of categorizing valid, novel, potentially useful and understandable patterns in data. Data mining plays a major role in several application areas like business organizations, Educational institutions, Government sectors, Health care industry, scientific and Engineering. In Data mining techniques classification is one of the most important techniques. The classification technique has several algorithms like ID3,C4.5,Navie Bayes etc. In this paper we analyze the different data mining algorithms. By using these algorithms we can classify our data. Classification is used in every field of real life. The data mining classification technique is used to classify each item in a set of data into one of predefined set of classes or groups. The classification technique is used to predict group membership for data instances. In this Data mining the classification technique include several techniques such as Decision tree, Bayesian classification, Classification by Back propagation, Association Rue mining.

Keywords: Data mining, Classification, Decision tree, ID3,C4.5, Bayesian classification, Naive Bayes classification.

I. INTRODUCTION

Data mining is the process of extracting or mining knowledge from large amount of data. The various different meanings of Data mining are knowledge mining from Databases, knowledge extraction, data/pattern analysis. The most popular synonym for the data mining is Knowledge Discovery in Databases(KDD). Data mining is an essential step in the process of knowledge discovery in data bases[1].The main intension of Data mining is the extraction of hidden predictive information from large databases. In this process the data mining tools are predictive future trends and behaviors, that are allowing business to make proactive , knowledge-driven decisions. Data mining is a process of extracting previously unknown , valid and hidden patterns from large data sets. Data mining is a process of categorization of novel, potentially useful and ultimately understandable patterns in data.

II. SOME OF THE DATA MINING APPLICATION AREAS

The following are the application areas of Data mining:
1)Medicine
2)Finance
3)Marketing
4)Scientific Discovery
5)Engineering

The above mentioned areas are the most popular areas where Data mining concepts used for extracting hidden useful patterns in that fields.

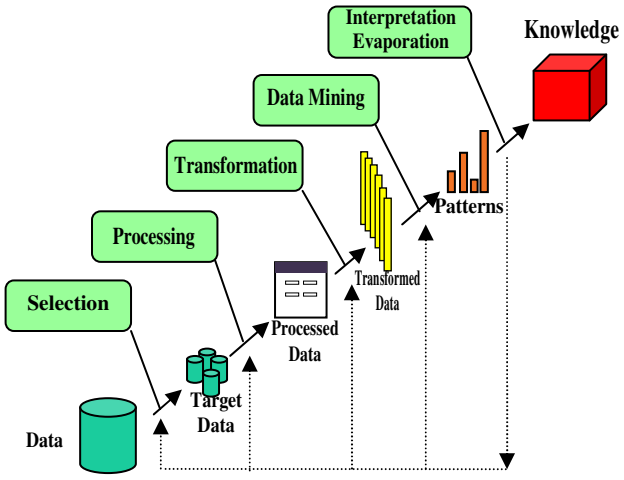
The following are basic steps for defining the Data mining process :
1) Requirement Analysis
2) Data selection and collection
3)Cleaning and preparing data
4)Data mining exploration and validation
5)Implementing , Evaluating and Monitoring
6)Result visualization

III. DATA MINING TECHNIQUES

The following are the major techniques used in the process of Data mining :
1)Classification
2)Association Rule
3)Clustering
4)Visualization technique

IV. DATA MINING PROCESS

The following Diagram illustrates the steps followed while transforming data into knowledge, it includesDataSelection,Preprosessing,Transformation,Datamining,Interpretation evaluation, Knowledge.



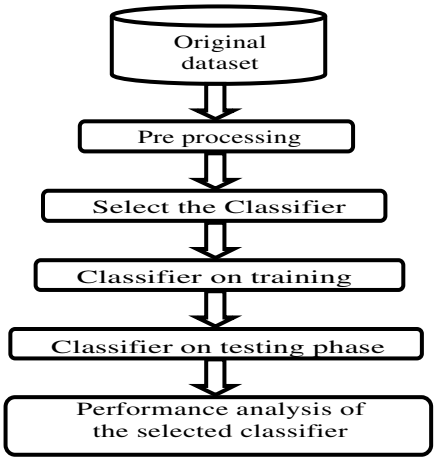
Figer.1 The Data mining process

V. THE CLASSIFICATION TECHNIQUE

In Data mining techniques the Classification is the one of the most important technique. The Classification is the process of finding a set of models, that describe and distinguish data classes or concepts, for the purpose of being able to use the model to predict the class objects whose class label is unknown. The derived model is based on the analysis of a set of training data, that is data objects whose class label is known[1].

In Data mining the classification technique is used to data analysis is that can be used to extract models describing important data classes or to predict future data trends preparing the data for classification have the following steps are applied to the data in order to help to improve the accuracy, efficiency and scalability of the classification.
1) Data cleaning
2) Relevance analysis
3) Data transformation

The Classification Technique it maps data into predefined groups or classes. In the Data mining classification Technique, the classes are indomitable before examining the data thus it is often mentioned as supervised learning. Classification is the process which classifies the collection of objects, data or ideas into groups, the members of which have one or more characteristics are common[8].

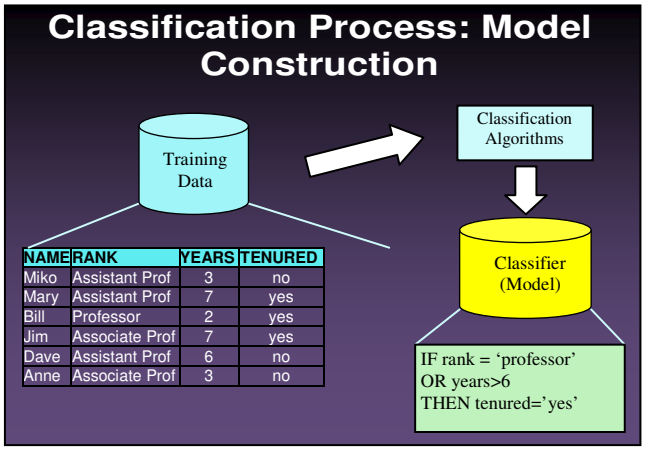


Figer.2 The Classification steps

Classification is one of the data mining methodologies used to predict and classify the predetermined data for the specific class. There are different classification methods proposed by researchers. The basic methods are given by

1)Decision tree induction
2)Bayesian classification
3)Rule based classification
4)Classification using Back propagation
5)Support Vector Machine
6)Classification using Association Rule[9].

Classification is one of the data analysis used to predict the categorical data. classification is a two phase process that is the training phase and the testing phase. In training phase the predetermine data and the associated class label are used for classification. The tuples used in training phase is called training tuples. this is also known as Supervised learning[9].



Figer.3 The Classification Process

VI. LITERATURE SURVEY

Classification is an important Data mining Technique. The input is a dataset of training records (also called training database), where in each record has several attributes. Attributes with numerical domains are numerical attributes and attributes whose domain is non-numerical are categorical attributes. There is also a distinguished attribute called the class label. This classification aims at building a concise model that can be used to predict the class label of future, unlabeled records. They are many classification models including Naive-Bayes, K-Nearest Neighbor, Decision Trees and Neural Networks have been proposed in the literature.

A classification task begin with build data for which the target values are known. Different classification algorithms use different techniques for finding relations between the predictor attributes values and the target attributes values in the build data. These relations are summarized in a model, which can then be applied to new cases with unknown target values to predicts target values. A classification model can also be used on build data with known target values to compare the predictions to the known answers, such data is also known as test data or evaluation data[3]. This technique is called testing a model, which measures the models predictive accuracy. The application of a classification model to new data is called applying the model, and the data is called apply data or scoring data. applying data is often called scoring data.

Pooja sharma, Divakar Singh, Sanju Singh in their Research about classification technique, they present the classification technique algorithms in their research papers.C4.5 is one of the most popular algorithms for rule base classification[2].There are many empirical features in this algorithm such as continuous number categorization, missing values handling etc. C4.5 is collection of algorithms for performing classifications in machine learning and data mining. It develops the classification model as a decision tree. C4.5 consists of three groups of algorithm: C4.5,C4.5 no-pruning and C4.5 rules. In this summary, they will focus on the basic C4.5 algorithm[3,4].

The following are the features of C4.5 algorithm discussed by the above authors are:
1) Speed: C4.5 is significantly faster than ID 3
2)Memory:C4.5 is more memory efficient than ID3
3)Size of Decision Trees:C4.5 gets smaller decision trees
4)Rule Set:C4.5 can give rule set as an output for complex decision tree
5)Over fitting Problem:C4.5 solve over fitting problem through reduce error pruning technique.

Sudeep D.Thepade, Madhura M.Kalbhor in their research they are said classification is a process which has a set of predefined classes and determines which class a new object belongs to ,there are large number of classifiers available which are used to classify the data such as bayes , function, rule ,lazy, meta, decision tree etc. Th ey discussed bayes, lazy, fuction, rule and tree family classifier[6].

Asli calis, Ahmet Boyaci said in their Research decision trees are data mining approaches that are frequently used in classification and estimation. Despite being capable of being used in classification of other methodologies like the nerve networks, the decision trees with their easy to make interpretations and ease of being understood provides advantage or decision makers[8].

Angelo Gargantini and Paolo Vavassori in their research about Decision tees they propose the use of decision trees for the suggestion of right tool for test generator. Decision trees are the traditional building blocks of data mining and the classic machine learning algorithm. since their development in the 1980's,decision trees have been the most widely deployed machine-learning based data mining model builder. their attraction lies in the simplicity of the resulting model, where a decision tree is quite easy to view, understand, and importantly, explain . classification tree structure is used in many different fields, such as medicine, logic ,problem solving and management science. It is also a traditional computer science structure for organizing data.[11].

Prashanth G.Shambharkar and MN Doja in their reseach they said in their research about Naive Bayes classifier, Bayes theorem is used when one want to do classification with the help of Naive Bayes Classifier, Bayes theorem uses previous knowledge with comparatively new evidences obtained from subsequent input data. The following equation gives the smaller form of the Bayes theorem that can use the conditional probabilities:

$$P(YVX) = \frac{P(\frac{X}{Y})P(Y)}{P(X)}$$

P(Y/X)denotes the posterior probability and P(Y) and P(X) denotes the prior probabilities of variable X & Y. Naive Bayesian Classifier uses a slight variant of formula that incorporate the conditional independence assumption.

$$P(YVX) = \frac{P(Y)\pi i = 1P(XtVY)}{P(X)}$$

If X_i is not relevant attribute then $P(X_i/Y)$ will be uniformly distributed. In such cases the probability of class condition for X_i has no effect on overall calculation of the posterior probability[5].

Dr.S.Vijayarani and Mr.S.Dhayanand in their research they discuss about Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes theorem with strong

independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model" this restricted individuality assumption infrequently clutches true in real world applications, hence the characterization as Naive yet the algorithm inclines to perform well and learn rapidly in various supervised classification problems[11].An advantage of the naive Bayes classifier is that it only requires a small amount of training data to estimate the parameters necessary for classification. Because independent variables are assumed, only the variance of the variables for each class need to be determined and not the entire covariance matrix[12].

VII. ANALYSIS OF CLASSIFICATION ALGORITHMS

This is the analysis of the following algorithms:
1) ID3 Algorithm
2) C4.5 Algorithm
3)Naive Bayes Algorithms

ID3 ALGORITHM

In the decision tree algorithms the Most preferred algorithm is ID3 i.e. Iterative Dichotomized ,it is developed by J.R Quinlan.ID3 uses information gain and entropy to classify data in tree structure. The ID3 algoithm follows Greedy approach.

ID3 Algorithm:
The algorithm produces decision trees using Shannon entropy.

Step:
a) Build Classification Attribute
b) Compute classification entropy

$$H(X) = -\sum_{i=1}^{n} P(X_i) \log_b P(X_i)$$

where ,X is current data set for which entropy can be calculated, n is set of classes in X and p(x) is proportion of the number of elements in class n to the number of elements in set X.

entropy is figured for each one remaining quality. The property with the littlest entropy is utilized to part the set on this iteration. The higher the entropy, the higher the possibility to enhance classification.
c) For each one attribute in table, compute
Information Gain utilizing classification attribute.

$$IG(A, X) = H(X) - \sum_{t=T} P(t) H(t)$$

where H(X)-Entropy of set X,T is subset created from splitting set X, p(t) is proportion of elements in to the number of elements in X.
The attribute which have highest information gain is used to split the set X on particular iteration.
d) Select attribute with the highest gain to following node in the tree(beginning from the Root hub).
e)Remove node attribute, making decreased table.
f)Repeat step3-5 until all attributes have been utilized, or the same classification values stays in rows of reduced table. At that point, smallest tree is preferred.
ID3 attempt in making short decision tree out of set of learning data, shortest is not generally the best classification. due to limitation, it is succeeded by Quinlan's C4.5 and C5.0 calculations[13].

C4.5 ALGORITHM

The another most popular algorithms for rule base classification is C4.5 . there are many empirical features in this algorithm such as continuous number categorization, missing value handling etc.C4.5 is collection of algorithms for performing classifications in machine learning and data mining. It develops the classification models a decision tree. c4.5 consists of three groups of algorithm:C4.5,C4.-no-pruning and C4.5-rule.In this summary, we will focus on the basic C4.5 algorithm. the resulting decision tree is generated after classification. the classifier is trained and tested first. then the resulting decision tree or rule set is used to classify unseen data.C4.5 is the newer version of ID3.The C4.5 algorithm follows Greedy approach.C4.5 algorithm has many features.
C4.5 algorithm Pseudo code:
1) Check for the base case
2)Construct a DT using random training data
3)Find the attribute with the highest info gain(A_Best)
4)A_Best is assigned with entropy minimization
5)Partition S into S1,S2,S3...
6)According to the value of A_Best
7)Repeat the steps for S1,S2,S3
8)For each ti€D, apply the DT
Base cases are the following:
1)All the examples from the training set belong to the same class(a tree leaf labeled with that class is returned)
2)The training set is empty(returns a tree leaf called failure).
3)the attribute list is empty (return a leaf labeled with the most frequent class or the disjunction of all the classes).

NAIVE BAYES CLASSIFIER

One of the most popular classifier is a Naive Bayes classifier, it is a simple probabilistic classifier that depends on Bayes' theorem with strong i.e. naive independence assumptions. It

is also be called as "independent feature model". In general terms, a naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. Naive Bayes classifiers are trained to work in supervised learning.

Naive Bayes classifier mainly pre assumes the effect of a variable value on predefined class that is not dependent on value of other variable. This is called as property of class conditional independence. It is particularly suited when the dimensionality of the inputs is high. NaïveBayesian is mainly used to form models with
predictive capabilities.

Bayes' Theorem:

$$\Pr obability(BgivenA) = \frac{\Pr obability(AandB)}{\Pr obability(A)}$$

Assume X as a data tuple. Let H be any hypothesis. P (H|X) be posterior probability of the H that is conditioned on X. In the same way, P (X|H) is the posterior probability of X condition on H.

$$P(H / X) = (\frac{P(X / H)P(H)}{P(X)})$$

P(H) is prior probability of H.

NAIVE BAYES ALGORITHM

1. Assume D to be training set of tuple. Every record can be represented by n-dimensional attribute vector i.e. X=(x1, x2…, xn),predicting n measurements on tuple from n attributes, i.e. A1 to An.
2. Let m number of class for prediction (C1, C2…....., Cm). As for record X, the classifier predict that X will belong to the class with maximum posterior probability that is
conditioned on X. Naïve Bayes predict that the tuple x will belong to class Ci only if P (Ci|X)>P (Cj|X). Therefore we have to maximize P(Ci|X).
By Bayes' theorem:

$$P(\frac{C_i}{X}) = \frac{P(\frac{X}{C_i}) * P(C_i)}{P(X)}$$

3. Because P(X) is constant in all classes, therefore P (X|Ci)* P (Ci) need be maximized.
4. As then assumption of class conditional independence is done. Therefore it is pre assumed that value of attributes are conditionally independent of each other.
Thus,

$$P(\frac{X}{Ci}) = \prod_{k=1}^{m} P(\frac{X_k}{C_i}) = P(\frac{X_1}{C_i}) * P(\frac{X_2}{C_i}).....P(\frac{Xm}{C_i})$$

5. To predict class of X, P(X|Ci)P(Ci) is calculated for each class Ci. Naive Bayes predict that class label of X is Ci class if

$$P(\frac{X}{C_i})P(C_i) > P(\frac{X}{Cj})P(Cj)\ for 1 \le j \le m, j \ne i$$

VIII.COMPARISON OF ID3,C4.5,NAIVE BAYES ALGORITHMS

Comparison	ID3	C4.5	Naive Bayes
Speed	Low	Faster than ID3	Fast
Advantages	Easy to understand	Memory efficient than ID3	Easy to handle of large amount of data
Dis Advantages	Can suffer from over fitting	High training samples are needed	Dependencies among variables

CONCLUSION

The Classification is the one of the most important technique in Data mining. The classification is the process of finding a set of models, that describe and distinguish data classes or concepts for the purpose of being able to use the model to predict the class objects whose class label is unknown. In this paper we, discussed about the various data mining technique algorithms like ID3,C4.5 and Naive bayes algorithms and also we compare these three algorithms based on speed, advantages, disadvantages. By this comparison, it is clear that the Naive Bayes classifier is the fast and it is easy to handle of large amount of data.

REFERENCES

[1] Jiawei Han, Macheline Kamber. "Datamining concepts an Techniques"- Elsevier Publisher-2001,ISBN:978-1-55860-489-6
[2] Pooja sharma,Divakar singh, anju singh "Classification Algorithms on a large continuous Random Dataset using Rapid Miner Tool" -IEEE Sponsored 2nd International Conference on Electronics and Communications System(ICECS 2015).
[3] Ali,M.M,Rajamani,L"Decision Tree induction:Priority classification"International conferance on Advances in engineering,Science and Management, ,pp.668-673, MArch 2012

[4] A.SGalathiya,A.P Ganatra,C.K Bhensdadia "classification with an improved Decision Tree Algorithm"International Journal of computer Applications,vol 46,No.23, ,pp1-6,May 2012
[5] Prashanth G.Shambharkar,M.N doja"Äutomatic classification of movie trailors using Datamining Techniques:a Review.-IEEE International Conference on Computing,Communication and Automation(ICCCA2015).
[6] Suddep D.thepade,Madhura M.Kalbhar"Extended performance Appraise of Bayes,function,Lazy,rule,Tree Data mining classifier in novel Transformed fracional content based image classification"-2015 IEEE International Conference on Pervasive Computing(ICPC).
[7] Asli calis, Ahmet Boyaci-"Datamining application in banking sector with clustering and classification methods".IEEE 2015 International Conference on Industrial engineering and Operations Management.
[8] Dr.S.Vijayarani,M.S Dhayanand.-"Datamining classification algorithms for kidney disease prediction" IJCI vol.4,No.4,August 2015.
[9] Vanajya.S,K.Rameshkumar-"Performance analysis of classification algorithms on Medical diagnoses-A Survey.
[10] Angelo Gargantini, Paolo Vavassori-"Using Decision Trees to aid Algorithm Selection in Combinatorial Interaction Tests Generation"IEEE 8th International conference on Software Testing,Verification and Validation workshops 2015.
[11] George Dimitoglou, "Comparison of the C4.5 and a Naive Bayes classifier for the prediction of Lung Cancer Survivability".
[12] Dr.S.Vijayarani,Mr.S.Dhayanand-"Data Mining Classification Algorithms for Kidney Disease Prediction"IJCI vol.4,No.4,August 2015.
[13] Monika Gandhi,Dr.shailendra Narayan Singh-"Pediction in heart Disease Using Techniques of Data mining"1st International Conference on futuristic trend in Computational analysis and Knowledge Management 2015.

