IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 6, Ver. VI (Nov.-Dec. 2016), PP 30-33
www.iosrjournals.org

Classification Techniques in Data Mining-Case Study

Swathi Agarwal[1], G.L.Anand Babu[2], Dr.K.S.Reddy[3]

[1,2,3] *Department of Information Technology, Anurag Group of Institutions, Hyderabad ,Telangana, India.*

Abstract: *Data Mining was a basic technique for examine the data in a choice of perception and categorizes it and finally to compress it. Classification was used to expect cluster relationship for data instances in data mining. Many key types of classification techniques including KNN, Bayesian classification, and Decision tree (DT) induction, C4.5, ID3, SVM, and ANN are used for classification. The major purpose of this assessment was to grant a complete examination of various classification mechanisms in data mining.*

Keywords: *C4.5, ID3, SVM, Decision Tree, KNN.*

I. Introduction

Generally Data mining is an iterative and influencing innovation process. Nowadays, data mining is given an immense agreement of fear and awareness in areas of information industry and in society as a whole.

Data mining responsibilities are generally classified as clustering, association, classification and prediction [4]. Within data mining, classification and prediction are two varieties of data analysis used to extort models to describe essential data module or to expect future data trends. The classification method has two parts: the first part is learning practice, in which training data will be analyzed. The learned type or classifier shall be characterizing in the shape of classification regulations. The other level of classification practice, in which test information to calculate approximately the exactness of classification style or classifier. If the exactness is acceptable, the regulations can be useful to classification of new data.

In fact, classification method is supervised learning, which is class level or prophecy target is already known. As a result, the classification form which is represented through rules structures will be constructed in the classification method. In this case, the created model will be representing the valuable information and is utilize for upcoming planning.

Classification is one kind of analytical modeling. More particularly, classification is a concept for conveying latest objects to predefined type or classes: from a collection of labeled records, construct the prototype such as a decision trees and estimates labels for upcoming not labeled records.

Various classification techniques are ID3, decision tree (DT) induction, C4.5, Bayesian classification, SVM, ANN, KNN classifier, are studied and provided a comprehensive assessment for different classification approaches in data mining [1].

II. Classification Methods

Classification analyses given dataset and selects each instance and allocates the instance to particular class, so that classification errors are minimized. It is a two level process. During first phase the model was created by related classification algorithm on data set then in second phase the extort form was experienced against a predefined test dataset to decide the model trained performance and accuracy. So classification was the method to be assigned a class label from dataset where class label is not known. Classification algorithms had involved significant attention in research areas of data mining [5]. Some of the famous classification models are:

1) ID3
2) C4.5
3) K-Nearest-Neighbor
4) SVM
5) Bayesian Classification

Among which few of the algorithms features, limitations are studied and offered in the paper.

1. ID3

Ross Quinlan has proposed ID3 algorithm and its purpose is to create decision tree from dataset. ID3 was forerunner to C4.5 algorithm. ID3 explores all attributes of training dataset, this extort the element from the dataset. ID3 stops, if the attributes rightly categorize the training datasets; otherwise it operates recursively on n detachment subsets to get the "finest" attribute. The algorithm applies greedy search, it selects the finest attribute and earlier choices are not considered.

ID3 Algorithm

```
Function ID3 ( AI, OT, TD) {
/* Where AI = Input Attributes
OT= Output attributes
TD= training data
The out put of the function is decision tree
*/
If ( TD is blank ) {
 Gives distinct node is “Failure”;
}
If (every record in TD has similar value for OT) {
Gives only node by that value;
}
If (AI is blank) {
Gives distinct node by the majority value of OT in TD;
}
Calculate Gain for each attribute in AI comparative to TD;
Let X be attribute with largest Gain(X,TD) of attributes in AI;
Let { x_ j | j = 1,2,3,……, m) is the value of X;
Assume subsets of TD are { TD_ j | j = 1,2,3,…….,m) and is accordingly partitioned significance of X;
Returns a tree with root node labeled with X and arcs labeled x_1, x_2, x_3, …….., x_m, then arcs go to trees ID3( I-X, OT, TD_1), ID3( I-X, OT, TD_2), …….., ID3( I-X, OT, TD_m);
}
```

Advantages of ID3

- ID3 was included in a number of profit-making packages. Few precise applications comprise classification of soybean diseases, and web search classification.
- ID3 is simple to exercise.
- ID3 is efficient.

Limitations of ID3

- ID3 does not assurance best possible result;
- ID3 can avoid over fitting by preferring smaller decision trees over larger decision trees.
- ID3 was difficult to use on uninterrupted data.

2. K-Nearest-Neighbor Classification

K-Nearest-Neighbor algorithm is simple that supplies all existing cases and categories on a resemblance measure. It was useful in statistical evaluation and pattern detection.

KNN is a non *parametric lazy learning* algorithm. When a technique is non parametric, it implies that it will not build any assumptions on the underlying data distribution [2].

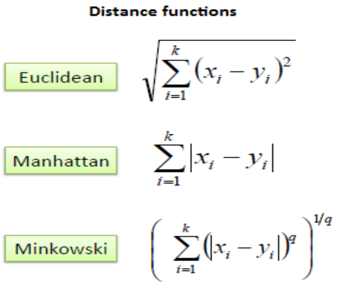
KNN is a sluggish algorithm. This says that it will not utilize the training datasets to achieve any generalization.

K-Nearest-Neighbor Algorithm

A collection was categorized by a popular choose of its neighbors, it’s K-Nearest-Neighbors calculated by a distance function. If k=1, in such case basically allocate to the group of its nearest neighbor.

Summary of KNN algorithm is as follows:

1. An integer S is precise, along with a new sample.
2. Select S entries in the database which are nearby to the new sample.
3. Find recognizable classification of entries.
4. This was the classification that gives to the sample.



Distance functions

Euclidean $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$

Manhattan $\sum_{i=1}^{k}|x_i - y_i|$

Minkowski $\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$

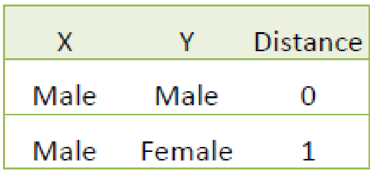
It should be prominent, that distance actions valid for uninterrupted variables. In the occurrence of definite elements Hamming distance is used.

Hamming Distance

$$D_H = \sum_{i=1}^{k} \left| x_i - y_i \right|$$

$$x = y \Rightarrow D = 0$$

$$x \neq y \Rightarrow D = 1$$



X	Y	Distance
Male	Male	0
Male	Female	1

Selecting the best value for M is done by probing of data. Historically, most datasets are between 3-10 for the optimal K. This produces best solutions than 1NN.

Properties of K-Nearest-Neighbor Algorithm

1. KNN algorithm is simplest classification algorithm.
2. KNN algorithm gives highly aggressive results.
3. KNN algorithm can be uses for regression problems.
4. The algorithm was versatile algorithm and is used in many fields.

Limitations of K-Nearest-Neighbor Algorithm

1. Which group of distance to use and which kind of attribute to use to produce the best result is not understandable by using Distance based learning?
2. Because of calculation cost was high; researchers need to calculate distance of every query. Some indexing may reduce this computational cost.

3. 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. The splitting principle gives normalized information gain (variation in entropy). In order to formulate the decision, the attribute by maximum standardize information increase is chosen. It is then returning on the minor sub lists [3]. C4.5 is set of algorithms for accomplish classifications in data mining. It extends the classification model as a decision tree.

C4.5 Algorithm

C4.5 is executing recursively with the following order:

1. ensure if algorithm satisfies extinction
Condition
2. for all attributes use Computer information theoretic condition
3. select best attribute depending to the information theoretic condition
4. construct a decision node by the outstanding attribute in step 3
5. create data group found in recently produced decision node in step 4
6. The entire sub dataset in step 5, call C4.5 algorithm to get a sub tree
7. connect the tree acquired in step 6 to the decision node in step 4
8. Return tree

Properties of C4.5

- C4.5 has supplementary characteristics considering tree pruning, enhanced utilize of continuous attributes, missing values handling and inducing rule set.

III. Conclusion

This paper focuses on diverse classification methods recycled in data mining and a study on each of them. Numerous techniques are recommended for the construction of group of classifiers. Even though or possibly as many methods of group constructions has been proposed, there is as yet no clear picture of which method is best. Classification methods are typically strong in modeling communications. Each of these methods is applied in various situations as needed where one tends to be useful while the other may not and vice-versa. These classification algorithms can be implemented on unlike kinds of datasets like share market data, data of patients, financial data, etc. Hence these classification techniques show how a data is determined and grouped when a new group of data is available. Each technique has got its own feature and limitations as given in the paper.

References

[1]. KS Reddy, GPS Varma, MK Reddy, An Effective Methodology for Pattern Discovery in Web Usage Mining - International Journal of Computer Science and Information Technologies, 2012

[2]. Boser, B. E., I. Guyon, and V. Vapnik (1992). A training algorithm for optimal margin classifiers. In Proceedings of the Fifth Annual Workshop on Computational Learning Theory, pages. 144 -152. ACM Press 1992.

[3]. Dr.Sudarson Jena, Santhosh Pasuladi, Karthik Kovuri, G.L.Anand Babu, An Investigation And Evaluation on Précised Decision For Scientific Data Using New Approaches in DMDW, International Journal Of Computer Applications In Engineering, Technology And Sciences (Ij-Ca-Ets), Issn: 0974-3596 | April 2012- September 2012 | Volume 4 : Issue 2)

[4]. KS Reddy, GPS Varma, SSS Reddy, Understanding the scope of web usage mining & applications of web data usage patterns, International Conference on Computing, Communication and Applications, 2012

[5]. L. Breiman, L. Friedman, R. Olshen, and C. Stone. Classification and Regression Trees. Wadsworth and Brooks, 1984.