Issues of Classification in Data Mining: Splitting Attribute Impacts the Performance Applying

Bineet Kumar Gupta[1], Yogesh Pal[2]
[1]*Associate Professor,* [2]*Assistant Professor,*
[12]*Department of Computer Application, Shri Ramswaroop Memorial University, Lucknow Deva Road, Barabanki, UP*

Abstract- Missing data not only affects performance of decision tree rather it is responsible for degrading system performance in addition it cause troubles during training phase and during classification process. Decision tree in classification plays crucial role because of easiness and effective use to sole real world problems. Rules are generated for the purpose of interpreting and understanding the scale for large data base. These large scale databases are chosen because the size of tree is independent of the size of database as well as every tuple in the must be filtered. The study in this paper observes the major part of data mining is to put attention towards decision making or prediction purpose rather focusing towards collection method of data or storage mechanism to system application.

Keywords- Missing Data; KNN; Decision Tree; Information Retrieval,

I. INTRODUCTION

Missing values are critical issues and it is advised that such issues must be handles carefully for the purpose of getting an accurate result otherwise it may cause to produce wrong result. Data are managed in their raw form that is no binning is required or recommended. In decision tree, there are three nodes i.e. root node and two child nodes [1]. Root node is the parent node which is divided into two children nodes and each child is again divided into two grandchildren. The maximal sized tree is removed back to the root split by split via the novel method of removing cost complexity. Accuracy of classification is crucial part in applied approaches which can be examined or evaluated with the help of calculating percentage of tuples placed in the acceptable class. It can be observed that cost may be associated with a wrong task to an incorrect class [3].

A. Outliers

There are frequently many terminologies of data mining which does not meet into the derived model responsible to become an additional issue for large database [4]. During the process of developing a model inclusion of these outliers may increase performance of system.

B. Interpretation of Result

Now a day's in the process of data mining output may have need of to export the result in correct form with regard to average database user.

C. Visualization of results

In the classification process the visualization plays crucial role to support formation of decision tree to the output of data mining algorithms.

D. Large datasets

Due to the rapid growth of network there is rapid growth of storing devices at the same time of datasets which are related with data mining algorithms designed for small datasets. Numbers of applications are available on the dataset sizes which are not competent for considering the larger datasets

II. RELATED WORK

A. Classification Tools

Two dissimilar classification results taking two different classification tools. These classification tools are able to determine best depends is applies by users to visualize the performance of classification algorithms. It can typically be judged by evaluating the correctness of the classification. However classification is frequently kind of fuzzy problem. Fuzzy problems such as evaluating the space and time exceed can be used to response the system accordingly [8].

An OC which stands for operating characteristics and ROC which stands for receiver operating characteristic are important terminologies in classification. Both OC and ROC are responsible for exposing the relationship between false positive and true positive to determine the true variable which can be applied for further processes. An OC curve was initially used in the communications area to be examined false alarm rates. A fallout stands for evaluating or observing percentage of retrieved that are not relevant. It has also been used in information retrieval to observe fallout verse recall (percentage of retrieved that are relevant). There is none of either category, at the beginning of evaluating an example while at the end there is 100 percent of each, when evaluating the results for a specific sample [10].

The name of the field of data in the decision tree is the object of analysis which is generally arranged instances by sorting them based on feature values.Instances are categorized at theinitial root node and sorted while analysingtheir feature values in a decision tree to shows a value of node [17,18].

- KNN Algorithm
- CART
- Concept of IDE3
- C4.5
- Decision Tree

1 KNN Algorithm
2 CART
3 Concept of IDE3
4 C4.5
5 Decision Tree

Decision Tree Algorithms

B. Statement of Problem

Distance based algorithm is the part of classification which is responsible in the process of mapping to the similar class. Such mapping is considered similar to the further objects in that class than the items found in other classes. Hence similarity or distance is applies to identify the dissimilarity of various objects or items in the data base.

Sometimes there is a situation where there is no idea about presence of the classes. If this is known in advance it will be easier to use a similarity measure for classification. Before performing this step the classes should be are predefined. Again, think of the IR example. Each IR query provides the class definition in the form of IR query itself. So the classification problem then becomes one of determining similarity not among all tuples in the database but between each tuple and the query [19].

C. Simple Approach

Using the IR techniques, if there are representative of each class, user can perform classification by assigning each tuple to the class to which it is most similar. Assume here that each tuple, tp_i, in the database is defined as a vector $< tp_{i1}, tp_{i2}, ..., tp_{ik} >$ of numeric values likewise, in all possible aspects assume that each class Cl_j is defined by a tuple $< Cl_{j1}, Cl_{j2}, ..., Cl_{jk} >$of numeric values. The classification problem is then restated.

Definition: Given a database $DB = \{tp_1, tp_2, ..., tp_n\}$ of tuples where each tuples $t_i =< tp_{i1}, tp_{i2}, ..., tp_{ik} >$ contains numeric values as well as set of classes represented by $Cl =< Cl_1, Cl_2, ..., Cl_m >$ where each class $Cl_j =< Cl_{j1}, Cl_{j2}, ..., Cl_{jk} >$ has numeric values. The classification problem is to allocate each with the terminology t_i to the class C_j such that $(tp_i, Cl_j) \geq sim(tp_i, Cl_1) \forall Cl_i \in Cl$ where $Cl_i \neq C_j$ There are represented vectors in these approaches. These are represented in the form of C or for very class for the purpose to calculate similarity measures. Referring to the three classes as shown in figure, determine a representative for each class by calculating the centre of each region. Thus class A is represented by < 4, 7.5 >, class B by < 2, 2.5 >, class C by <6, 2.5 >. Each item is identified in the class o find out the similarity of classes among defined class. The similarity of class is represented by closest set of items. However the approached of pattern recognition are applied to find out out similarity and dissimilarity of classes to characterize the each class in pattern recognition methods. The item which has been classified is compared based on the predetermine pattern. The items will be place in the class after evaluating the similarity of related objects and to focus on largest value.

The following example states a meaningful approach of distance based algorithm for each class, C_i where it is represented by its centre or sometimes by centroid. In algorithm the use of Cl_i is for its class for the reason that every tuple must be compared to the centre due to the small number of classes.

D. DBSAI (Distance Based Simple Approach Input)

Centers for each class= $Cl_1, Cl_2, ..., Cl_m$

Tuples are input to classify tp

Output: Class where tuple tp is assigned Cl

```
Distance= $\infty$
fori=0; i<=m do
{
if $dist(cl_i, tp) <$ distance
$Cl = I;$
 Distance = $dist(cl_i, tp)$
 }
```

E. Decision Tree Based Algorithms

The decision tree based algorithm methodology is very much applicable to solve classification problems. There are two types of tuples which are applied. The first one is each tuple in the database and the tuple whose values are used to build a tree. There are two kind of phase in the technique, the first one known as building the tree and the second one is known by applying the tree to the database. The basic motive behind constructing decision tree is to generate such result in the form tree so that users have the choice to utilize their input in the form of output. Labeling concept is also used in decision tree based algorithm.

The decision tree method of classification is responsible to partition the search space in the form of various rectangular regions in addition a tuple is classified into region assuming the similarity of region. An example of binary tree can be considered in the form of DT and leading nodes could be labeled with the predicates themselves in binary decision methods i.e. with yes or no.

III. DEFINITION

Given a database $Db = \{tp_1,....., tp_n\}$ where and the database schema contains the following attributes {A_1 , A_2 ,…………,A_h}. Also given is a set of classes $Cl = < Cl_1, Cl_2,...., Cl_m >$.

A decision tree is represented by symbol DT is a tree which is related with data D. The internal node, arc and leaf node are used to create a decision tree. Internal node is represented in the form of an attribute and leaf is represented in the form of class. It has the following characteristics.

- Each internal node-> A_i (an Attribute)
- Each arc is -> a predicate
- Each leaf node->Cl_j (Class)

It has the two steps.

1. Decision tree induction: It means the process of creating DT

2.For each tp $\in$ db

According to definition, the constructed DT represents the logic needed to perform the mapping based on our definition of the classification problem. Thus, it implicitly defines the mapping. A different DT could yield a different classification. User does not consider the second part of the problem. A decision tree can further be relatively straight forwarded to carry out the process. Instead, focus on algorithms to construct decision trees. Several algorithms are surveyed in the following subsections.

This filtration is done through the tree. Many attributes are used to construct a tree. Training data and decision tress are represented at Dt and Tr respectively.

A. Decision Tree Build Algorithm

```
Input: ->Dt
Output :->Tr
        Tr = 0;
Determine best splitting criterion;
Tr = Create root node ‘node’ and label with splitting attribute;
Tr = Add arc to root node for each split predicate and label,
For each arc do
Dt = Data base created by applying splitting predicate to Db;
If stopping point reached for this path, then
Tr’ = Create leaf node and label with appropriate class,
else
Tr’ = Decision_Tree_Build (Db);
Tr = Add Tr’ -> arc;
```

IV. RESULT

Some attributes are better than others and which attribute to use for splitting attributes impacts the performance applying the built DT in the name attribute definitely should not be used and the gender may or may not be used.

The choice of attributes involves not only examination of data in the training set but also the informed input of domain experts. The creation of the tree definitely stops when the training dataare perfectly classified. There may be situations when stopping earlier would be desirable to prevent the creation of larger trees. There may be situations when stopping earlier would be desirable to prevent the creation of larger trees.

V. REFERENCES

[1]. D. Donko and A. Dzelihodzic, “Data mining techniques for credit risk assessment task,” Recent Advances in Computer Science and Applications, pp. 105–110, 2013

[2]. Ibrahim M. El-Hasnony, Hazem M. El-Bakry, Ahmed A. Saleh,," Classification of Breast Cancer Using Soft computing Techniques", International Journal of Electronics and Information Engineering, Vol.4, No.1, PP.45- 54, Mar. 2016.

[3]. J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, Burlington, MA, USA, 2012.

[4]. J. Xia, F. Xie, Y. Zhang, and C. Caulfield, “Artificial intelligence and data mining: algorithms and applications,” Abstract and Applied Analysis, vol. 2013, Article ID 524720, 2 pages, 2013.

[5]. Journal of Computer Applications, vol. 54, no. 13, pp. 21–25,

[6]. Journal of Computer Applications, vol. 54, no. 13, pp. 21–25

[7]. M. Andrecut, “Parallel GPU Implementation of Iterative PCA Algorithms”, Institute of Biocomplexity and Informatics, University of Calgary, Canada, 2008.

[8]. M. Bowles, “Machine Learning in Python: Essential Techniques for Predictive Analytics”, John Wiley & Sons Inc., ISBN: 978-1-118- 96174-2

[9]. P. Harrington, “Machine Learning in Action”, Manning Publications Co., Shelter Island, New York, ISBN 9781617290183, 2012.

[10]. R. Arora and S. Suman, “Comparative analysis of classification algorithms on different datasets using WEKA,” International Journal of Computer Applications, vol. 54, no. 13, pp. 21–25, 2012.

[11]. S. B. Hiregoudar, K. Manjunath, K. S. Patil, “A Survey: Research Summary on Neural Networks”, International Journal of Research in Engineering and Technology, ISSN: 2319 1163, Volume 03, Special Issue 03, pages 385-389, May, 2014

[12]. S. S. Shwartz, Y. Singer, N. Srebro, “Pegasos: Primal Estimated sub -Gradient Solver for SVM”, Proceedings of the 24th International Conference on Machine Learning, Corvallis, OR, 2007.

[13]. Intelligence and data mining: algorithms and applications

[14]. Intelligence and data mining: algorithms and applications

[15]. V. Sharma, S. Rai, A. Dev, “A Comprehensive Study of Artificial Neural Networks”, International Journal of Advanced Research in Computer Science and Software Engineering, ISSN 2277128X, Volume 2, Issue 10, October 2012

[16]. W. H. Inmon, Building Data Warehouse, QED/Wiley, Hoboken, NJ, USA, 2005.

[17]. Witten, I. & Frank, E. (2005), "Data Mining: Practical machine learning tools and techniques",2nd Edition, Morgan Kaufmann, San Francisco, 2005.

[18]. X. Zhu, A. B. Goldberg, “Introduction to Semi – Supervised Learning”, Synthesis Lectures on Artificial Intelligence and Machine Learning, 2009, Vol. 3, No. 1, Pages 1-130.

[19]. Yun Wan, Dr. QigangGao," An Ensemble Sentiment Classification System of Twitter Data for Airline Services Analysis",2015 IEEE Volume 3, Issue 1, January-February-2018.

[20]. Zhou, Z. (2004), Rule Extraction: Using NeuralNetworks or For Neural Networks?, Journal of Computer Science and Technology, Volume 19, Issue 2, Pages: 249 – 253.