DOI: http://dx.doi.org/10.26483/ijarcs.v8i7.4153

ISSN No. 0976-5697

Volume 8, No. 7, July – August 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

ANALYTICAL FINDINGS ON DATA MINING ALGORITHMS

Vinita Malik
Information Scientist,
Central University of Haryana

Mamta Ghalan,
Assistant Professor
M.S.I.T,Delhi

Sukhdip Singh
Assistant Professor ,D.C.R.U.S.T,Murthal

Anu Saini
Assistant Professor ,G.B.Pant College of Engineering ,Delhi

Neeti Sangwan
Assistant Professor ,M.S.I.T,Delhi

Abstract**—**The heart of issue pumps out several classification algorithms based on decision tree which is again one of the most prominent area in data mining .Data Mining constitutes discovery or extraction of various patterns or relationships by large data clusters. For this we can employ several strategies which can be based on statistics or artificial intelligence. Here we use classification as well as prediction methodologies which follows decision tree concept. Various classical approaches i.e. ID3 or C4.5 have been discussed and we discuss algorithms SLIQ ,SPRINT and FUDT which will help in avoidance of costly sorting by further discussing their characteristics, challenges they pose, advantages as well as disadvantages.

Keywords**—** Decision Tree, SLIQ, SPRINT, Big Data, FUDT

I. INTRODUCTION

As the technology has grown rapidly, the database has also grown up proportionately. So we really have an urgent requirement for usage of new technology and tools for extraction of such a large volumes of data and for further cleaning of it and transform it into the required one by processing it automatically and intelligently .So to process the data into required information the research has been going prominent in the field of data mining. Data classification is as much as mandatory as data prediction. These both can be utilized together for extraction of data from which future predictions can be done . By the help of classification we can predict categorical (which can be further divided into discrete and unordered) labels and same way we can use prediction models which help in outputting continuous valued functions. In past we have proposed several machine learning techniques for doing recognition and drawing of statistics.

Most of the algorithms are memory resident in nature, so requires a small data size. Data mining in social media also has been gaining importance over times [1].

In the proposed paper, we have done analytical study of various algorithms supporting data mining concepts i.e.ID3, C4.5, C5.0 for structured data and SLIQ ,SPRINT ,FUDT to understand their characteristics, challenges , advantages and disadvantages.

A. Types of Data

The data within the data bases and throughout the world is not consistent. Data can be Structured (dimensional data) or Unstructured (non-dimensional data).The Unstructured data is also called as Dark data. Meaning of *structured data* basically pertains to kind of data which have a well defined fixed length and fixed format. Various examples may be numbers, dates, and groups of words and string [2] .For example: customer name or address. Various text mining algorithms have been used in past[3,4] .Experts are agreed on the terms that such type of data accounts for only 20 % of the data available [2]. Structured data can be stored in a database and can get the output by using a language like structured query language (SQL).

Some sources of structured data can have following types of data -

Generated by Computer system : Data can be generated by machine without utilizing manpower [2].

Generated by Human: It is the data produced by the human beings while interaction with the computer .

Unstructured data does not have any specified format .In any organization ratio of unstructured to structured data is generally found as 80:20.So the data requires further analysis and processing. Just like structured data we have only machine generated or system generated computer data .

B. Privacy Issue

The main concern is how to extract the useful information from the data provided with the complete preservation of data privacy .Here, we require to publish the data with right amount of distortion or decomposition so that privacy

© 2015-19, IJARCS All Rights Reserved

of individuals should not be compromised. The most important assumption is that private data is collected for proper consolidation and analysis of data [5]. In publishing patterns for privacy preservation we find that the central question that is addressed is mainly how can we publish frequent patterns without telling all sensitive and secure information about the given data.

In case of **pattern hiding** the main motive is basically to change the data in a certain way that some of patterns can not be found out by mining. We can also use **secure multiparty** mining over other distributed type of datasets which also shows that data on which type of mining is to be performed which can be further divided into horizontally or vertically The data that is divided into several parts is not shared but the output of mining on the union of data is shared among various parties and hence in this way we can further extract useful data while doing preservation of the original data [5] .

II. ALGORITHMS FOR MINING DATA : SURVEY

Below are given several algorithms that are really helpful for data mining for conversion of structured data into unstructured one .

- ID3
- C4.5
- C5.0
- SPRINT
- SLIQ
- FUDT

ID3 Algorithm

ID3 means Iterative Dichotomiser3. This algorithm was invented by Ross Quinlan .By this algorithm we can generate a decision tree[6] .This ID3 algorithm can be typically used for machine learning and also for natural language processing.

The decision trees can be built in following steps:

Step 1: If instances in C are taken as positive, then we need to create YES node and then we require to halt. In another condition if all instances in C are taken as negative we need to create a NO node and then we require to halt.

Otherwise we have to select a feature, F with various values like v1, ..., vn and then we need to create a decision node.

Step 2: After this we require partitioning of the training instances in C into subsets C1, C2, Cn as per the values of V.

Step 3: In step 3 we can use the methodology recursively to each of the sets Ci. In such cases the expert takes the decision which features can be selected.

ID3 is basically an improvement over CLS by adding a feature selection heuristic. ID3 searches through the attributes of the training instances and extracts the attribute that best separates the given examples. If the attribute perfectly classifies the training sets then ID3 gets stopped; otherwise it can recursively operate on the n (where n is the number of possible values of an attribute) which can be partitioned subsets to get their "best" attributes [6]. The algorithm can use a greedy search means it can pick one of the best attribute and will not look back for reconsideration of previous choices.

ID3 can be described by following examples :

EXAMPLE 1

If we consider S as a bundle of 14 examples with 9 Yes and 5 NO examples then in such case we can find out the Entropy function as given below :

$$H(S) = -\sum_{x \in X} p(x) \log_2 p(x)$$

H(s) is the entropy.

Entropy(S) by considering above formula can be found out as 0.940.We can assume entropy value as 0 if all of the members of S pertains to the same class or in such case if the data is perfectly classified. Entropy change can be varied from 0 ("perfectly classified") to 1 ("totally random").

Gain(S, A) which is also called as information gain of set S on attribute A can be defined as:

Gain(S, A) = Entropy(S) - S ((|S_v| / |S|) * Entropy (S_v))

Where:

S is defined as each value v of all possible values of attribute A

S_x = It is the subset of S for which attribute A has the value x

|S_v| = element count in S_v

|S| = element count in S

EXAMPLE 2

Consider S be a set of 14 examples out of which one of the attributes can be taken as wind speed. The type of values of wind can be either *weak* or it can be *strong*. If we do the classification of these 14 examples then we can have 9 YES and 5 NO. For an attribute like wind, consider there can be 8 occurrences of Wind as Weak and the 6 occurrences of Wind as Strong. For Wind as Weak, 6 of the examples can be YES and 2 a can be as NO. For Wind taken as Strong, 3 are found to be YES and 3 are found to be NO. So , in this case we can calculate gain as Gain(S,Wind)=Entropy(S)-(8/14)*Entropy(S_{weak})-(6/14)*Entropy(S_{strong}) Which is calculated as 0.048 .Entropy (S $_{weak}$) can be calculated as 0.811.Entropy (S $_{strong}$) is found out to be 1.00.

For each of the attribute we can calculate the gain and then the highest gain is used in the decision node.

Advantages we can have by ID3 algorithm:

- By the training data we can formulate the clear prediction rules
- Helps in building the fastest tree
- Can form the short tree
- We only require to test various attributes so that whole data can be classified.
- We can also find leaf nodes which will enable test data for pruning by complete reduction of number of test cases.

Disadvantages we can have by ID3 algorithm:

- Data can be over classified if tested with a small sample .
- For decision we can take only one attribute for testing.
- Continuous data classification can be computationally expensive so that as many trees can be generated .

C4.5 ALGORITHM

C4.5 is an extended version of the Quinlan's or ID3 algorithm. The decision trees can be generated by C4.5 which can further be used for classification purpose. C4.5 can be referred as a statistical classifier also.C4.5 algorithm can use information gain/ entropy as one of splitting criteria. It can also accept the data with various continuous or discrete values [1]. For handling of continuous values it can generate the threshold value and then it can divide all attributes with values above the threshold as well as values equal to or below the threshold [7].

C4.5 algorithm can also handle various missing values of the attributes because missing attribute values can not be utilized in gain calculations by C4.5.

C4.5 had several improvements on ID3. Some of these may be:

- C4.5 is more advantageous while taking both kinds of the continuous and discrete attributes. For handling continuous attributes C4.5 can create a threshold and then used to split the list into a list where the attribute value is also above the threshold[6]and values which are less than or equal to it.
- Missing values of attributes can be handled well by C4.5.It allows to mark missing values as "?"
- Attributes which have the different costs can be handled well by C4.5
- Once the tree is created then c4.5 can prune tree by replacement with leaf nodes.

C5.0 ALGORITHM

C5.0 algorithm is an extended version of C4.5 algorithm. It can be used to efficiently handle multi-valued and missing attributes [8]. It is the classification algorithm which can be applied on the very big data sets. It gives better performance than C4.5 in terms of various parameters like speed, memory and also in terms of efficiency. This model can work by splitting the sample data based on the field that can provide the maximum information gain. This model can also split all the samples on basis of the biggest information gain field. The sample subset that can be achieved can be divided afterwards [8]. The process will be continued till the samples are not divided. Lastly , we can examine the lowest level of those sample subsets which do not have any contribution to the model .Given below is table 1 which gives the comparison among three algorithms mentioned above :

Table 1: Data mining algorithms comparison

Data Mining Algorithm	**Id3**	**C4.5**	**C5.0**
Types of Data	Categorical	Continuous and Categorical	Continuous and Categorical ,dates, times
Speed	Low	Faster than ID3	Highest
Missing values	Can't deal with	Can't deal with	Can deal With
Pruning	No	Pre -pruning	Pre-pruning
Formula	Use information gain ,entropy	Can use the split info as well as gain ratio	Same as C4.5

Now we can consider some more advanced level algorithms that can mine data efficiently .So here we will discuss two more algorithms SLIQ and SPRINT .

III. ALGORITHMS IN DEMAND NOW : DISCUSSION AND OVERVIEW

A.SLIQ ALGORITHM

SLIQ which means Supervised Learning in Quest, this algorithm was developed by IBM's Quest project team. This algorithm is designed to train and classify large volumes of data [9]. It can use a pre-sorting technique which further can be employed in the tree growth phase. It will help in avoiding high cost search at each of the node. This algorithm utilizes a different list for each of attribute and a different class list. For each entry in the class list that corresponds to a data item has a different class label and also node name that belongs to a decision tree .All entries in the sorted attribute list will have an attribute values and data item indexing in the class list .

We use attribute list in sorted manner and further data item indexing in the class list .By help of this algorithm ,decision tree can grow in **breadth-first** manner [7]. For every attribute, it can go through by the corresponding list in sorted order and can also evaluate all entropy values of nodes of the decision tree simultaneously. Once the entropy values gets calculated

© 2015-19, IJARCS All Rights Reserved

for each attribute then one attribute can be chosen for every split for each node and they can be expanded to have a new frontier. After this one more scan of the sorted attribute list is also performed so that class list for the new nodes can be updated[9]. Although SLIQ can handle disk-resident data which is too large to fit into the memory but it still require some more information to stay memory-resident which again can also grow in the direct proportion to the number of input records by putting a hard-limit on the training data size[9]. The Quest team has been recently designed a new decision-tree-based classification algorithm which is also called as SPRINT (Scalable Parallelizable induction of decision **t**rees) which can remove all of the memory restrictions.

B. SPRINT ALGORITHM

Algorithm: In SPRINT algorithm all attributes are maintained in a list of attributes which is sorted in nature. In this attribute list and its corresponding records, once we identify attribute for splitting a node in a classification tree then each attribute list has to be divided according to split decision . For every attribute we create an attribute list and form a table where table entries are called as attribute records .Each record is consisted of an attribute value and also class label index of the record. The initial attribute lists are also associated with the root[10]. As the tree grows the attribute lists are divided belonging to each node and also associated with the children of the nodes. Histograms can be created for each attribute and by the use of Gini index we can find out the splint point. SLIQ does not support separate sets of attribute lists for every node . Text mining algorisms are a major concern for mining data [11].

Advantages of SPRINT algorithm :

1) Inexpensive to build
2) Easy interpretation
3) Easy integration with the commercial database
4) Better accuracy

Disadvantages of SPRINT algorithm :

1) It does not work efficiently for handling very large data sets.
2) Memory limitations also exists.
3) Low processing speed is also one of drawback

C. FUDT ALGORITHM

This algorithm is based on fuzzy entropy which helps in selection of best split point .It can be used as an efficient measure for decision tree related classification issues. First we divide the uncertain decision tree into two parts i.e. training phase and testing phase . In the first phase we build the tree from the uncertain data by help of fuzzy entropy value . We use the best feature vector value from the fuzzy entropy value for selection of best split point .Once the construction is completed then the tree is evaluated by the testing of the uncertain data[12] .

IV. CONCLUSIONS

Algorithms based on classification of data for data mining can be included as decision tree based algorithms , k-Nearest Neighbor algorithm , Bayesian and Neural-Net based classification. We have discussed several algorithms based on decision tree structure in past and 2 recent algorithms . ID3 algorithm comprises disadvantage of tending to select attributes with several values. It also can have the issue pertaining to over classification. We have discussed various improvements on ID3 algorithm which gives re-optimized form of ID3, C4.5 and C5.0 algorithm .SLIQ , SPRINT,FUDT are algorithms that attempt to reduce costly sorting at each node by pre-sorting continuous attributes in the beginning.

V. REFERENCES

[1] J. Sang, Y. Gao, B. Bao, C. Snoek, Q. Dai, "Recent advances in social multimedia big data mining and applications",Springer , Multimedia Systems , 2016,vol 22

[2] W. brand , " Big_data_for_dummies"

[3] J. Rani , A. R. Shah , S. Ramachandran , "pubmed.mineR: An R package with text-mining algorithm to analyse PubMed abstracts", Indian Academy of Sciences, 2015

[4] W.Zhao, J.J. Chen, R.Perkins, "Erratum to: A novel procedure on next generation sequencing data analysis using text mining algorithm, Division of Bioinformatics and Biostatistics, National Center for Toxicological Research,2016

[5] NESSI White Paper, "Big Data :A New World of Opportunities", 2012

[6] http://en.wikipedia.org/wiki/ID3_algorithm

[7] S. Ghosh, S. Roy, S. K. Bandyopadhyay , "A tutorial review on Text Mining Algorithms" ,International Journal of Advanced Research in Computer and Communication Engineering , Vol. 1, Issue 4, June 2012

[8] B. R Patel, Kaushik, " A Survey on Decision Tree Algorithm For Classification" , International Journal of Engineering Development and Research, 2014 ,Volume 2, Issue 1

[9] "SLINQ: An optimally efficient algorithm for the single-link cluster method",2014

[10] J.Shafer, R. Agrawal, "SPRINT: A Scalable Parallel Classifier for Data Mining" ,IBM Almaden Research Center,1996

[11] http://en.wikipedia.org/wiki/Text_mining

[12] S. Meenakshi, V. Venkatachalam, "FUDT: A Fuzzy Uncertain Decision Tree Algorithm for Classification of uncertain Data", Arab J Sci Eng , 2015

© 2015-19, IJARCS All Rights Reserved