JCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Volume-4, Issue-3 E-ISSN: 2347-2693

Decision Trees for Mining Data Streams Based on the Gaussian Approximation

S.Babu[1*] and G.Fathima[2]

[1*,2]*Department of CSA, SCSVMV University ,Enathur, Kanchipuram, India*

www.ijcseonline.org

Received: Feb/26/2016 Revised: Mar/06/2016 Accepted: Mar/17/2016 Published: Mar/31/ 2016

Abstract—Decision Trees are considered to be one of the most popular approaches for representing classifiers. Researchers from various disciplines such as statistics, machine learning, pattern recognition, and data mining have dealt with the issue of growing a decision tree from available data. The key point of constructing the decision tree is to determine the best attribute to split the considered node. Several methods to solve this problem were presented so far. However, they are either wrongly mathematically justified or time-consuming. The primary comparison parameters are time and accuracy. Also shown efforts made for handling the change in the concept and they are compared in terms of memory, technique and accuracy. Our method ensures, with a high probability set by the user, that the best attribute chosen in the considered node using a finite data sample is the same as it would be in the case of the whole data stream.

Keywords—Data steam, decision trees, information gain, Gaussian approximation

I. INTRODUCTION

Decision tree contained nodes, branches and leaves which are used to take the decision. Trees may be either binary were nodes are split into the two children nodes or non-binary were nodes have many children as the number of elements of set. The nodes which are not terminal nodes (end nodes) are accompanied by some attribute. The parent nodes and children nodes are connected to each other through the branches.

In the majority of the projected algorithm, the selection is based on some contamination gauge of the data set. The impurity of the data set before the split and weighted impurity of the resulting subsets are calculated for all the probable dividers of the node. Split measure function is the difference of these values. Considered node is assigned by the best attribute which is nothing but an attribute which gives the highest value of this function. Impurity measure is taken as information entropy in the ID3 algorithm. The matching split-measure function is also called as entropy reduction or the information gain.

The ratio of the information gain and the split information is proposed as the split measure function in the CART algorithm. The Gini index is another impurity measure worth the consideration. It is used in the CART algorithm, which is intended to develop binary trees. Therefore, it can be applied to the numerical data as well as to the data with nominal attribute values.

Objectives of study:

- To develop a theoretical tool to deal with data streams, in particular replace the Hoeffding bound (wrong technique) by another approach.
- To design a decision tree learning system for stream data such that its output is nearly identical to that of a conventional learner.
- To find (analytically) a number of samples of the infinite stream data such that a split in decision tree learning system can be made.

II. REVIEW OF LITERATURE

The decision tree is a popular classification method. It is a tree like structure where each internal node denotes a decision on an attribute value. Each branch represents an outcome of the decision and the tree leaves represent the classes. Decision tree is a model that is both predictive and descriptive. A decision tree displays relationships found in the training data. In data mining and machine learning, a decision tree is a predictive model; that is, a mapping from observations about an item to conclusions about its target value.

Classification: The given data instance has to be classified into one of the target classes which are already known or defined. Estimation- Like classification, the purpose of an estimation model is to determine a value for an unknown output attribute. However, unlike classification, the output attribute for an estimation problem are numeric rather than categorical. Prediction- It is not easy to differentiate prediction from classification or estimation. The only difference is that rather than determining the Literature review on data mining research current behaviour, the predictive model predicts a future outcome. The output attribute can be categorical or numeric.

An interesting hidden rules called association rules in a large transactional data base is mined out. For e.g. the rule {milk, butter->biscuit} provides the information that whenever milk and butter are purchased together biscuit is also purchased, such that these items can be placed together for sales to increase the overall sales of each of the items. The main application areas of data mining are in Business analytics, Bio-informatics, Web data analysis, text analysis, social science problems, biometric data analysis and many other domains where there is scope for hidden information retrieval.

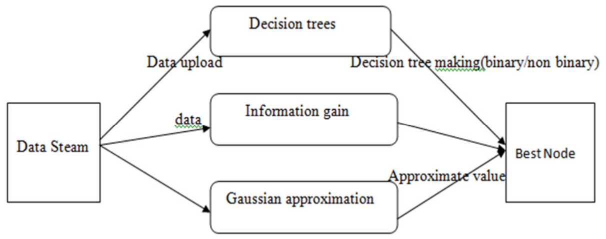
III. METHODOLOGY

Proposed System

Advantage

- The attribute which gives the highest value of this function is called the best attribute, and is assigned to the considered node.
- The decision tree is choosing the best attribute to split the considered node.
- Time-consuming

Proposed system Architecture



Input Design

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document.

The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy.

Objectives

1. Input Design is the process of converting a user-oriented description of the input into a computer-based system.

2.It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors.

3. When the data is entered it will check for its validity.

Output Design

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

1. Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements.
2. Select methods for presenting information.
3. Create document, report, or other formats that contain information produced by the system.

Implementation:

- Login and Registration
- Data Stream
- Decision trees
- Information gain
- Gaussian approximation

Login & Registration

Login: The registered user will Login by providing the user name and the password by entering it correctly and will proceed with the other work of sharing the data with data stream.

Registration: If the user didn't register in the registration module then the user can enter the details of username, password, conform password, mobile number in the Register module.

Data Stream: The data sets to be directly applied to the data streams, and significant modifications are needed. The dominant problem is to establish the best attribute in each node, since the stream is of infinite size. Given the data set of n elements in the considered node, we want to know if the best attribute computed from this data set is also the best attribute for the whole data stream.

Decision Trees: The decision tree is composed of nodes, branches, and leaves. Every node, which is not terminal, is accompanied by some attribute ai. The tree can be either binary or non-binary. If the tree is binary, the node is split into two children nodes (or leaves). In the second case, the node has as many children as the number of elements of set. Children are connected with their parent nodes by branches. To each branch a value of attribute ai (in the non-binary case) or some subset of Ai (in the binary case) is assigned. It is obvious, that the non-binary trees make sense only if the attributes take nominal values.

Information Gain: The information entropy is taken as the impurity measure. The corresponding split-measure function is called the information gain or the entropy reduction .The ratio of the information gain and the split information is proposed as the split measure function. The Gini index is another impurity measure worth the consideration. It is used in the CART algorithm, which is intended to develop binary trees. Therefore, it can be applied to the numerical data as well as to the data with nominal attribute values.

Gaussian Approximation: Gaussian approximation is to reduce the number of data required to make a split. The main reason of the dramatic decrease of required elements to make a split, when using the Gaussian approximation versus the MacDiarmid's method, is the fact that the latter gives too coarse bound for information gain. Decision tree is the choice of the best attribute to split the considered node. We proposed a new method for deciding if the best attribute determined for the current set of data elements in the node is also the best according to the whole stream.

IV. RESULT AND DISCUSSION

Fig: 1. Gaussian Decision Tree and McDiarmid Tree algorithms.

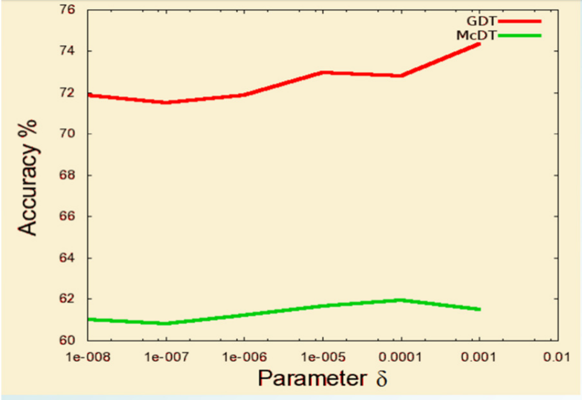


Fig: 2. Gaussian Decision Tree and the ID3 algorithm.

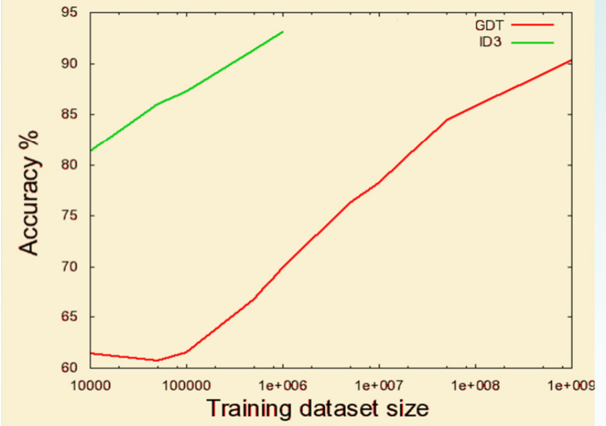
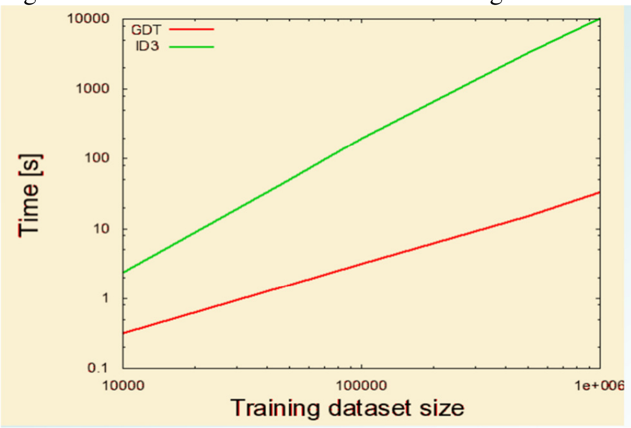


Fig: 3. Gaussian Decision Tree and the ID3 algorithms.



- Developed a theoretical tool to deal with data streams; in particular we replaced the Hoeffding bound (wrong technique) by another approaches.
- Designed a decision tree learning system for stream data such that its output is nearly identical to that of a conventional learner.
- Found analytically a number of samples of the infinite stream data such that a split in decision tree learning system can be made

V. CONCLUSION

In this paper, we considered the issue of mining data streams with the application of decision trees. The key point in construction of decision tree is the choice of the best attribute to split the considered node. We proposed a new method for deciding if the best attribute determined for the current set of data elements in the node is also the best according to the whole stream. The method is based on the Taylor's Theorem and on the properties of the normal distribution. It is mathematically justified by the theorem presented in the paper. Following the idea presented in, we proposed the GDT algorithm. GDT algorithm radically outperforms the McDiarmid tree algorithm in the field of time consumption. The GDT algorithm is able to give acceptable accuracies in data streams classification problems is shows by the numerical simulations. Numerical simulations proved that the GDT algorithm is able to give satisfactory accuracies in data streams classification problems.

VI. REFERENCES

[1] Caiyan Dai and Ling Chen, "An Algorithm for Mining Frequent Closed Itemsets with Density from Data Streams", International Journal of Computer Sciences and Engineering, Volume-04, Issue-02, Page No (40-48), Feb -**2016**, E-ISSN: 2347-2693

[2] P. Argentiero, R. Chin and P. Beaudet, "An automated approach to the design of decision tree classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.* PAMI-4, 51-57 (**1982**).

[3] P. Fletcher and M.j.D. Powell,"A rapid decent method for minimization," *Computer Journal*, Vol.6, ISS.2, 163-168 (**1963**).

[4] Rudolf Ahlsmede and Ingo Wegeru, *Search problems,* Wiley-Interscience, **1987**.

[5] K.S. Fu, *Sequential methods in pattern recognition and machine learning,* Academic press, **1998**.

[6] D. E. Gustafson, S. B. Gelfand, and S. K. Mitter, " A nonparametric multiclass partitioning methods for classification," in proc. 5th int. conf. pattern Recognition, 654-659 (**1980**).

[7] E. G. Henrichon,Jr. and K. S. Fu, "A nonparametric partitioning procedure for pattern classification," *IEEE Trans. Computer.*, Vol. C-18, 604-624,(**1969**).

[8] G. R. Dattatreya and V. V. S. Sarma,"Bayesian and decision tree approaches for pattern recognition including feature measurement costs," *IEEE Trans. Pattern Anal. Mach. Intell. V*ol. PAMI-3, 293-298, (**1981**).

[9] R. L. P. Chang and T. Pavlidis, "Fuzzy decision tree algorithms," *IEEE Trans Syst. Man Cybernet.*, vol. SMC-7, 28-35 (**1977**)

[10] J. Aczel and J. Daroczy, *On measures of information and their characterizations,* New York: Academic, **1975**.

AUTHORS PROFILE

S.Babu MCA,. Mphil working as an Assistant Professor in the Department of Computer Science and Applications in Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya, Enathur, Kanchipuram, Tamil Nadu. He has published more than 12 Papers in International journals and conferences. His research interest lies in the area of Data Mining, Distributed Computing and Database Management System. mailto: babulingaa@gmail.com

Ms.G.Fathima pursed Bachelor of Science from VIT University, Vellore in 2007 and Master of Science from Alagappa University, Karaikudi in year 2012. She is currently pursuing M.Phil in SCSVMV university, Enathur, Kanchipuram.