Joint Split Criterion Based Data Stream Classification Technique

Anushree R Bhat Kanak, Prof. H R Shashidhara

***Abstract*— The quantity of data that needs to be analyzed has lead to a new field data stream mining. The goal of most of data stream mining application is to predict the class or value of new instances in data stream which gives some knowledge about the class members. Data classification is one of the important techniques used in data mining.**

This paper involves developing a new method for constructing decision tree stream data. Along with the existing methods, misclassification error is used as another measure to build hybrid technique which provides an accurate measure of impurity. A hybrid technique is developed using Chi-square and misclassification error to calculate impurity measure. Impurity measure is then used in split criteria to build the decision tree classifier to classify the stream data. Accuracy is calculated during the data streaming process. It is shown that the proposed system i.e. the hybrid technique provides satisfactory accuracies at any time of data stream processing.

***Index Terms*— Classification, data stream, decision trees, splitting criterion.**

I. Introduction

Recently the quantity of data that needs to be analyzed is growing fast. New research field called data stream mining is created mainly because of the unlimited data that need to be processed. In recent years, classification learning for data stream has become an important and active research topic. The new research field, Data Stream mining which is attracting many researchers is growing very fast[1].The goal of most of data stream mining application is to predict the class or value of new instances in data stream given some knowledge about the class members. Standard approach to the problem of data mining cannot be applied to the data stream mining methods as one has to face new difficulties while analyzing the stream of data elements.

Data classification is one of the important techniques used in data mining. Whenever a new observation has been done classification has to identify to which of a set of categories it belongs depending upon the training set of data containing observations whose category membership is known. Wide varieties of methods are used for data classification. Most significant methods are k- nearest neighbours, neural networks and decision trees [2]. Data stream classification using decision tree is the main subject of this paper. A decision tree is a tree –like graph where in a test on an attribute is denoted by internal node, the outcome of the test is represented by the branch and the class label is held by each leaf node. High dimensional data can be handled using decision trees. A good accuracy is achieved using decision tree classifiers.

The best split measure for construction of decision tree use different metrics for different algorithms. CART algorithm makes use of Gini index as impurity measure [3]. For each attribute Gini index considers binary split. The attribute which has a minimum Gini Index is selected as the splitting attribute. Gini index is biased towards multivalued attributes and when the number of classes is large it has a difficulty.

In this paper a new approach is proposed where in a new criterion for splitting the nodes based on the impurity measure called misclassification error is developed. A decision tree is constructed using Chi-Square and misclassification error. Combining the advantages of both methods a new hybrid algorithm is developed.

Rest of the paper is organized as follows. In section II Related work is noted down. Section III describes the joint splitting criterion for data stream classification. Results obtained in the simulation are presented in section IV. In Section V, the conclusion of this paper is drawn.

II. Related work

The main task in decision tree construction process is to determine which attribute is best to split the considered node. The quality of split is evaluated according to the split measure function. In majority of decision tree algorithms it is defined as reduction of some impurity measure or gain of information measure. Several forms of impurity measures exist such as information entropy in ID3 algorithm, Gini index in CART algorithm.

Decision trees for mining data streams based on Hoeffding bound tree algorithm is wrong tool as it is wrongly mathematically justified [4]. Decision tree learning system applied to stream of data based on McDiamid’s bound [4] has the property that its output is nearly identical to that of a conventional learner. McDiarmid’s tree algorithm is time consuming. Gausian Approximation [5] outperforms the McDiarmid tree algorithm and ensures with a high probability set by the user. Rutkowaski, Jaworski proposed the new splitting criterion [6] by combining the advantages of both misclassification error and Gini index. The algorithm based on this hybrid criterion provided the highest accuracy among all the considered algorithm.

Summarizing, the previous work carried out is either has wrong mathematical background or require impractically high number of instances to reach relatively low value. Hence proposing a new splitting criterion which overcomes this is the key of this paper.

Anushree R Bhat Kanak, Department of CSE, RNSIT, Bengaluru, India. 7795195344.,
Prof. H R Shashidhara, Department of CSE, RNSIT, Bengaluru, India, 9880822991.,

III. Methodology

In this paper a new approach is presented for data stream classification technique. A new criterion for splitting the tree nodes based on the impurity measure called misclassification error is proposed. A hybrid approach is developed to construct a decision tree using Chi-square and misclassification error. It makes use of the advantages of both the methods while developing a joint split criterion.

The proposed hybrid algorithm can be explained using the Figure 1[6]. From the flowchart it can be seen that the data elements from the stream are read one by one into the decision tree with only one leaf i.e. the root node. Then it is sorted to an appropriate leaf and the statistics is updated. Split measure for all the attribute is calculated which provides the measure of split and split criterion is applied for the same. If the split criterion is true then chosen partition is initialized with the new leaves and the process continues. If the split criterion is false then the next data element is read and the iteration continues. The split criterion which is shown in the figure 1 is the new split criterion i.e. Hybrid split criterion. The Joint split criterion makes use of Chi-square and misclassification error for splitting the attributes.

Chi-square provides the statistical significance between the differences between sub-nodes and parent node. Two or more splits can be performed using it. Higher the value of Chi-Square higher the statistical significance of differences between sub-node and Parent node. Chi-Square is calculated as follows:

$$\textit{Chi-square} = \sqrt{\frac{(Actual-Expected)^2}{Expected}} \quad \text{---------------} \; (1)$$

Using Equation 1 Chi-Square for individual node is calculated using the deviation for success and failure both. Chi-Square of split is calculated using sum of success and failure of each node of the split.

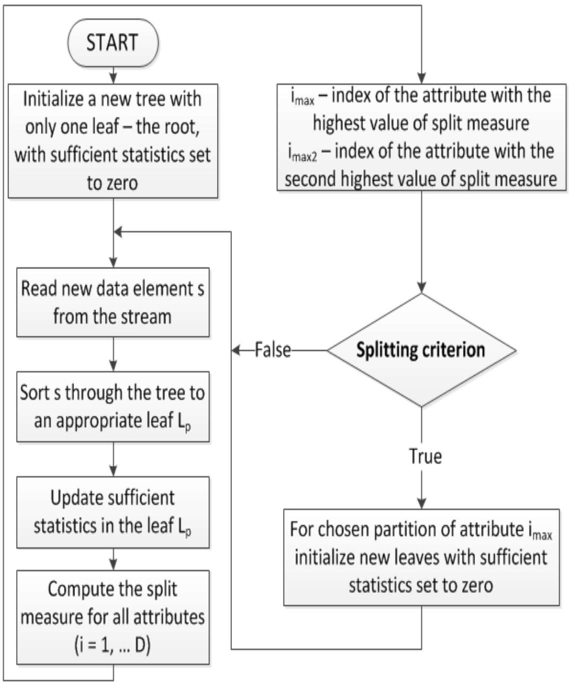
START
Initialize a new tree with only one leaf – the root, with sufficient statistics set to zero
Read new data element s from the stream
Sort s through the tree to an appropriate leaf L_p
Update sufficient statistics in the leaf L_p
Compute the split measure for all attributes (i = 1, ... D)
i_{max} – index of the attribute with the highest value of split measure
i_{max2} – index of the attribute with the second highest value of split measure
Splitting criterion
False
True
For chosen partition of attribute i_{max} initialize new leaves with sufficient statistics set to zero

Fig 1: System design for Hybrid algorithm

There exists another impurity measure, which is rarely mentioned in literature survey called misclassification error and is given in a very simple form,

$$g(S) = 1 - \max_{k\in\{1,\ldots,K\}}\{pk(S)\} \text{ -------------------}(2)$$

Where S is a finite set of data elements and g(S) denote the misclassification error of S.

The only difference between the misclassification error based decision tree, Chi-square based decision tree and Hybrid algorithm based decision tree lies in the splitting criterion. The criterion for the misclassification based decision tree algorithm is stated as follows:

$$\Delta g_i(S) - \Delta g_j(S) > z(1-\delta)\sqrt{\frac{1}{2n(S)}} \text{ ----------------------}(3)$$

Where g(s) denotes the misclassification error for set S, *z(1-δ)* is the *(1-δ)* quantile of the standard normal distribution N(0,1).

Misclassification based algorithm gives highest accuracy at the beginning stages and the Chi-Square algorithm performs better when the tree becomes more complex. Hence these two are considered for joint split criterion proposed in this paper.

IV. Experimental Results

In this section, the performance of the proposed method is discussed. Different datasets are considered such as Airlines datasets, Electricity, CoverType datasets [7] and KDD CUP 99[8]. One of the considered datasets is Airline dataset which consists of 8 attributes and 539383 data elements. One more dataset KDD CUP 99 consists of 4898431 data elements. Data are described by 41 attributes, 7 of which are nominal and 34 are numerical. To imitate the stream of data properly, appropriate data sets should contain as many as data elements possible.

The performance of the proposed method was tested on different real data sets as described earlier in this section. The major benefit of using the hybrid decision tree algorithm is higher accuracy than for the misclassification error based decision tree. Figure 2 gives the graph for hybrid algorithm where in the accuracy is plotted along with data elements.

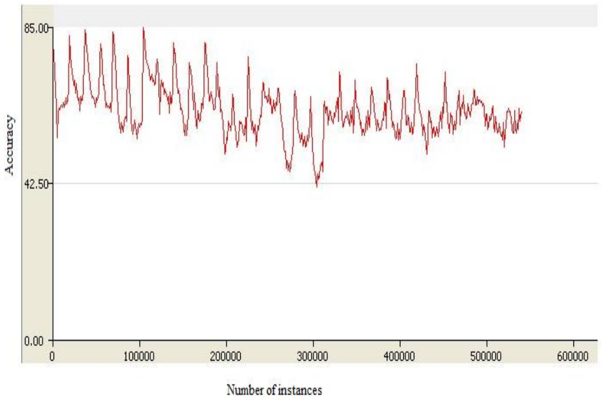

Fig 2: Accuracy of the hybrid algorithm

Table 1 shows the accuracy obtained for different datasets for hybrid algorithm (HDT) and misclassification error based decision tree (mDT).

Table 1
Mean accuracy for hybrid algorithm.

	Mean Accuracy	
Dataset	HDT	mDT
Airlines	62.29	61.44
KDD CUP 99	98.00	97.23
Covertype	71.05	65.72
Electricity	70.59	74.08

Datasets were divided into training and testing subsets while carrying out the classification.

The conducted experiments demonstrate the high usability of misclassification error impurity measure. As can be seen in the majority of cases, mean accuracy for hybrid algorithm is notable compared to the misclassification error.

ACKNOWLEDGMENT

Authors would like to thank Visvesvaraya Technological University, Belagavi and RNS Institute of Technology, Bengaluru for the opportunity. Authors are thankful and grateful for all the people who have supported and guided in successful accomplishment of the work.

REFERENCES

[1] J. Gama, “A survey on learning from data streams: Current and future trends,” Prog. Artif. Intell., vol. 1, no. 1, pp. 45–55, Apr. 2012.

[2] J. Gama, R. Fernandes, and R. Rocha, “Decision trees for mining data streams,” Intell. Data Anal., vol. 10, no. 1, pp. 23–45, Mar. 2006

[3] Jiawei Han and Micheline Kamber: Data Mining - Concepts and Techniques, 2nd Edition, Morgan Kaufmann Publisher, 2006.

[4] L. Rutkowski, L. Pietruczuk, P. Duda, and M. Jaworski, “Decision trees for mining data streams based on the McDiarmid’s bound,” IEEE Trans. Knowl. Data Eng., vol. 25, no.6, pp. 1272–1279, Jun. 2013.

[5] L. Rutkowski, L. Pietruczuk, P. Duda, and M. Jaworski, “Decision trees for mining data streams based on the Gaussian approximation,” IEEE Trans. Knowl. Data Eng., vol. 26, no. 1, pp. 108–119, Jan. 2014.

[6] L. Rutkowski, M. Jaworski, L. Pietruczuk, P. Duda, “A New Method for Data Stream Mining Based on the Misclassification Error,” Ieee Transactions On Neural Networks And Learning Systems, Vol. 26, No. 5, May 2015.

[7] (2014). *Massive Online Analysis* [Online]. Available: http://moa.cms.waikato.ac.nz/datasets/

Anushree R Bhat Kanak is a Post Graduation student. She is pursuing her Master of Technology in Computer Science and Engineering from RNS Institute of Technology Bengaluru, India. She has completed her Bachelors of Engineering degree in Computer Science and Engineering from Don Bosco Institute of Technology, Bengaluru, India.

Prof. H R Shashidhara is an Assistant Professor at RNS Institute of Technology, Bengaluru, India. He has completed his BE and MTech in Computer Science and Engineering. He has a work experience of 24 years in teaching and has published 14 papers. He is a gold medal awarder from Vasvi Union Bengaluru for MTech course.