Review on Decision Tree for Data Stream Classification

Priyanka Abhang
Department of Computer Engineering
DYPIET,Pimpri,Pune,Maharashtra,India

Dr. Pramod Patil
Department of Computer Engineering
DYPIET,Pimpri,Pune,Maharashtra, India

Bhagyashree Bhoyar
Department of Computer Engineering
DYPIET,Pimpri,Pune,Maharashtra, India

Abstract- Data stream classification is a great challenge with its core objective to classify stream data for better decision making purpose by utilizing limited amount of memory. Most challenging approach is based on decision tree which improves classification accuracy in less processing time. The classifier model built on training data is further used to classify unknown label data which utilizes limited memory by continuously analyzing stream data. Many split measures have been developed to choose attribute for splitting current node from finite data as it would be same in case of infinite data which also improves classification accuracy. The main aim of this paper is to review, analyze decision tree classification algorithms and methodologies.

Keywords – Classifier, Concept drift, Decision tree, Feature extraction, Split measures.

I. INTRODUCTION

Recently, data stream mining is being the most challenging area. These stream data have following characteristics:

1. It is continuous, massive data.
2. It changes fastly.
3. It is an infinite amount of data.

These huge amounts of data continue to grow and contain useful information to improve decision making in an organization. These stream data continuously arrive at high rate. Thus, traditional data mining techniques cannot be applied to process and analyze these streamed data due to limitation of memory and it also requires to scan the data multiple times. Many methods [13] such as clustering, classification, association are developed to process stream data. Among them, data stream classification is the most promising one. Many classification algorithms such as decision tree based, rule based, ensemble methods, nearest neighbor and SVM based [14] have been developed.

To effectively handle this stream data and to capture useful information from data, classification method has been getting increasing attention. To classify and analyze these data, methods based on decision tree construction play an important role. It should consider all requirements of processing stream data, i.e. use of limited memory to store all incoming data. Another challenge is to detect changes occurred in incoming data. Accordingly, data stream mining method should react to those changes.

Decision tree classification is a two step process[15]:

1. Training model or building classifier
2. Testing model or assigning class to unknown labeled data.

In a first step, classifier is built by analyzing training data set which consists of tuples, attributes and their corresponding class label. In second step, this built classifier is applied on the test data to check classification accuracy. If tuples are correctly classified by this classifier, this is used further to classify unknown labeled data. In

this process, attribute selected from sample data stream for splitting current node should be same as from infinite data stream. These classifier should be built in such way to improve classification accuracy and to reduce processing time.
The rest of the paper is organized as follows. Literature survey is explained in section II. Related work comparison is presented in section III. Research challenges for data stream classification are given in section IV. Conclusion is given in V section.

II. LITERATURE SURVEY

R. Bose, W. van der Aalst et al. [1] demonstrated that stream data are sequence of data examples that continuously arrive at time-varying and possibly unbound streams. Classification techniques fail to successfully process data streams because of two factors: their overwhelming volume and their distinctive feature known as concept drift. Concept drift is defined as changes in the learned structure that occur over time. The occurrence of concept drift leads to a drastic drop in classification accuracy. The recognition of concept drift in data streams has led to sliding-window approaches, instance selection methods, drift detection, ensemble classifiers. This work describes the various types of concept drifts that affect the data examples and discusses various approaches in order to handle concept drift scenarios. The aim of this work is to review and compare single classifier and ensemble approaches to data stream mining respectively and propose a methodology towards its contribution.

L. Kuncheva and W. Faithfull [2] proposed new method which is Principal Component Analysis (PCA) for change in data distribution. Because assumption about distribution of arriving data matches with the data which was used for training classifier is often incorrect. Proposed method is applied to training data set and used to mine stream of selected principal components. This method is for feature extraction and used to improve performance in change detection from multidimensional unlabeled incoming continuous data. It gives more insight into the potential of feature extraction for change detection in streaming data, data sets benefit significantly from using the PCA.

I. Zliobaite, A. Bifet et al.[3] proposed a strategies for active learning from stream data which is evolving over time. In process of classifying streaming data, it requires major effort to obtain the true labels and excessive cost. Active learning concentrates on learning accurate model with few labels. When changes are detected, old classifier model is replaced by new one and trained with new incoming data. Experimental evaluation is performed on real data streams. It shows that proposed method handle concept drift with low labeling cost, as it is not required to label each instance.

D. Brzezinski and J. Stefanowski [4]proposed new data stream classifier based on blocks, called Accuracy Updated Ensemble (AUE2). This reacts equally well to different types of drift. This is a combination of accuracy based weighting mechanism from block based with incremental nature of Hoeffding trees. In ensembles, component classifiers are generated from fixed size of data chunks. When new block data arrive, existing classifiers weights are updated. New learned classifier is added to ensemble and weaker classifiers are removed depending on evaluation result. This algorithm is compared with 11 stream methods. Out of them, AUE2 gives the best accuracy, less memory consuming. But disadvantage is whenever concept drift is detected, current model needs to be updated to maintain accuracy.

L. Rutkowski, M. Jaworski et al. [5] proposed an algorithm dsCART for streaming data which is based on Classification and regression tree(CART) algorithm. In constructing decision tree, main task is selecting best attribute to split. For this, Gaussian approximation is applied. This algorithm obtains high classification accuracy in less time. Main aim of this algorithm is to compute best attribute from sample data which is also same as if it is computed from entire stream data. It is proved that in two class problem number of examples in current node for each data concept, required for splitting, is smaller in proposed algorithm than in McDiarmid algorithm. This method is valid for any number of classes, numerical as well as categorical data. This also does not require prepruning. Theorem is also proved to solve problem of concept drift. Comparison between proposed method, McDiarmid decision tree and Gaussian tree is given by examining dependency between accuracy and size of tree. As number of leaves increases, accuracy also increases.

L. Rutkowski, L. Pietruczuk et al.[6] proposed a new method which performs better than McDiarmid tree algorithm. Main aim of constructing tree is to choose the best node to split into its further nodes. This selected node from finite sample data for splitting is same as in case of whole stream data. Focus of this method is to produce binary tree.

Many methods ware proposed to solve this problem. But they were either wrongly mathematically justified or time consuming. Proposed method does not address problem of concept drift. Experimental results shows that proposed method gives more accuracy than McDiarmid tree and ID3 takes more processing time than Gaussian tree.

L. Rutkowski, L. Pietruczuk [7] proposed two theorems. They represent the McDiarmid's bound for both Information Gain used in ID3 and Gini Index used in CART (Classification and Regression Trees algorithm) which select best attribute for splitting from available finite data sample which is also same as if it is from infinite data. Proposed theorems provide fast processing, low memory consumption and high accuracy. These make McDiarmid trees a proper tool for mining stream data.

M. Wozniak [8] proposed new algorithm which handles incremental learning of decision tree. This algorithm is called as NGE (Nested Generalized Exemplar) algorithm. This works in two phases. In first phase, tree is trained by using decision tree induction algorithm such as C4.5 and converted into hyper rectangles according to set of rules. In second phase, hyper rectangles are rebuilt as long as new data arrive.By using this algorithm, Classifiers obtained are very adaptable which are constructed to learn new incoming objects. This provides low computational complexity because of which it is easy to make decision. Incremental decision tree training NGE(iDTt-NGE) is proposed by doing certain modifications in NGE.

Hu, Hsiao-Wei et al. [9] proposed new algorithm for classifying structured data having continuous labels. These labels are distributed throughout hierarchical organization of decision tree during construction. This does not require discretization in preprocessing. In traditional methods, separate algorithm to build classifier for continuous label data was proposed. Also, another algorithm to construct decision tree with hierarchical label was proposed. Traditional methods are not applicable to handle hierarchical and continuous labels simultaneously. Thus, to handle labels which are continuous as well as hierarchical, author proposed hierarchical continuous label classifier algorithm(HCC) to construct decision tree. Discretization is performed dynamically during tree construction. This proposed approach has standard structure of previous methods such as C4.5 and ID3. Split measure developed to choose attribute is called goodness measure. It is used to select most appropriate attribute for splitting current data node having higher value of goodness measure. The comparison between C4.5 and proposed algorithm shows that proposed algorithm performs better than C4.5 with respect to its accuracy, specificity, complexity. HCC algorithm is superior to C4.5 considering overall tree performance.

Shukla, MsMadhu S. et al. [10] described working of classification algorithms such as Hoeffding tree, VFDT (Very Fast Decision Tree), Naive Bayesian and CVFDT (Concept Adapting Very Fast Decision Tree). Among these algorithms, only CFVDT algorithm handles concept drift and performs better than Hoeffding tree and VFDT in terms of accuracy.

Salperwyck, Christophe et al.[11] proposed new decision tree construction method which requires summaries in leaves. During tree construction, new examples arrive and go down into tree in summaries. These summaries in leaves are used to choose best splitting attribute and to construct classifier for each leaf node. In proposed method, Naïve Bayes classifier is built in each leaf node to improve prediction. In this method, tree memory is consumed because of these summaries. The Hoeffding bound is also used in this method for tree construction. To find groups and cut points for categorical and numerical values using summaries stored in leaves , MODL method is used. This criterion has an advantage that allows prepruning automatically in tree while in very fast decision tree, this pruning is implemented separately. Classifier built for each leaf node improves accuracy of whole tree mainly in the starting of training. This is fast to build and consumes memory. Experimental comparison is shown between proposed method and existing approaches. This shows that proposed tree performs better. Specifically local model improves prediction and accuracy in early stage.

HSSINA, Badr, et al [12] presented comparison between ID3 and C4.5 decision tree algorithm. This comparison is shown in the form of accuracy and execution time. C4.5 is an extended version of ID3. C4.5 overcomes limitation of ID3 such as use of missing values, use of continuous data, pruning tree. Also, theoretical comparison between C4.5 and C5.0, C5.0 and CART is given. This leads to confirm that C4.5 is the most powerful method in machine learning.

III. RELATED WORK

Table 1: Related work

Ref. No.	Algorithm/Method	Advantages	Disadvantages
[2]	Principal component analysis(PCA)	Detect changes in multidimensional unlabeled data and use for feature extraction.	Concept change is context dependent. Classification error increases with change in labels.
[3]	Active learning strategies	Low labeling cost, handles concept drift.	Does not redistribute dynamically labeling budget to the regions where changes are suspected.
[4]	Accuracy updated ensemble(AUE2)	Less memory consuming, handles different types of drifts, best accuracy.	Whenever concept change is detected, current model needs to be updated to maintain accuracy.
[6]	Based on Gaussian approximation	Better than Mcdiarmid tree algorithm in case of time consumption.	Does not handle concept drift.
[7]	Mcdiarmid tree algorithm	Fast processing, low memory consumption.	Time consuming.
[8]	Nested Generalized Exemplar (NGE)	Handles concept drift.Low computational complexity.	-

IV.RESEARCH CHALLENEGS IN DATA STREAM CLASSIFICATION

Several research issues and challenges are involved in data stream classification which are described below:

4.1.Data arriving at irregular rate and time:
Optimization and utilization of memory is one of research challenges in data stream classification because variant data arrive at irregular time and rate consuming memory space. It should be addressed by doing on the spot analysis of stream data.

4.2. Reacting to concept change:
This is the most important challenge which should be handled effectively because data which was relevant or important can become unneeded. Meaning or distribution of data may be changed. This should be handled in such way that accuracy of classification would not decrease.

4.3.Limitation of memory and huge data:
As a stream data continuously arrive to the system, it is not possible to store these large amounts of data due to limitation of memory. To optimize memory, novel techniques are required to handle and classify continuous flow of data.

V.CONCLUSION

This paper has described survey of previous researches conducted in the areas of Data stream classification based on decision tree construction approach. This process is an applicable for classification of numerical, categorical, structured continuous data.. Split measures such as Gini index, Information gain, and Gain ratio have been used to choose best attribute for splitting current node. Researches in this field shows that it yields better results in terms of accuracy and efficiency by modelling correct classifier which is used further to classify unknown data. Thus, a comprehensive survey of literature has been given help to provide more insight in this subject area.

REFERENCES

[1] R. Bose, W. van der Aalst, I. Zliobaite, and M. Pechenizkiy, “Dealing with concept drifts in process mining,” IEEE Trans. Neural Netw. Learn. Syst., vol. 25, no. 1, pp. 154–171, Jan. 2014.
[2] L. Kuncheva and W. Faithfull, “PCA feature extraction for change detection in multidimensional unlabeled data,” IEEE Trans. Neural Netw.Learn. Syst., vol. 25, no. 1, pp. 69–80, Jan. 2014.
[3] I. Zliobaite, A. Bifet, B. Pfahringer, and G. Holmes, “Active learning with drifting streaming data,” IEEE Trans. Neural Netw. Learn. Syst.,vol. 25, no. 1, pp. 27–39, Jan. 2014.

[4] D. Brzezinski and J. Stefanowski, "Reacting to different types of concept drift: The accuracy updated ensemble algorithm," IEEE Trans. Neural Netw. Learn. Syst., vol. 25, no. 1, pp. 81–94, Jan. 2014.
[5] L. Rutkowski, M. Jaworski, L. Pietruczuk, and P. Duda, "The CART decision tree for mining data streams," Int. J. Inform. Sci., vol. 266, pp. 1–15, May 2014.
[6] 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.
[7] 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
[8] M. Wozniak, "A hybrid decision tree training method using data streams," Knowl. Inform. Syst., vol. 29, no. 2, pp. 335–347, 2011.
[9] Hu, Hsiao-Wei, Yen-Liang Chen, and Kwei Tang. "A novel decision-tree method for structured continuous-label classification." Cybernetics, IEEE Transactions on 43.6 (2013): 1734-1746.
[10] Shukla, MsMadhu S., and MrKirit R. Rathod. "Stream Data Mining and Comparative Study of Classification Algorithms." Algorithms 3.1 (2013).
[11] Salperwyck, Christophe, and Vincent Lemaire. "Incremental Decision Tree based on order statistics." Neural Networks (IJCNN), The 2013 International Joint Conference on. IEEE, 2013.
[12] HSSINA, Badr, et al. "A comparative study of decision tree ID3 and C4. 5."International Journal of Advanced Computer Science and Applications 4.2 (2014).
[13] Gupta, Neha, and Indrjeet Rajput. "Stream data mining: a survey." Indrjeet Rajput, International Journal of Engineering Research and Applications (IJERA) ISSN (2013): 2248-9622.
[14] Aggarwal, Charu C. Data streams: models and algorithms. Vol. 31. Springer Science & Business Media, 2007.
[15] Han, Jiawei, MichelineKamber, and Jian Pei. Data mining: concepts and techniques: concepts and techniques. Elsevier, 2011.