Concept-adapting Very Fast Decision Tree with Misclassification Error

Shubhangi.A. Jadhav, Prof.S.P. Kosbatwar

***Abstract*—In data stream mining, to learn time-changing concepts many algorithms are evolved. CVFDT algorithm based on ultra-fast VFDT decision tree learner efficiently addresses the problem of mining the concept-changing data streams. CVFDT grows an alternative tree for questionable oldest data and replace the old with the new one when new becomes more accurate. In this paper we used new splitting criterion based on misclassification error which removes the incorrect use of hoeffding bound in CVFDT.This assures that the best attribute computed on available data sample is same as the attribute derives from the whole infinite data stream.**

***Keywords*—Decision Trees, Misclassification Error, Splitting Criterion, Stream Data.**

I. Introduction

A. Data Mining

In recent years data mining is favorite for research purpose. Huge amount of data generated due to daily operations or transactions in the various organizations, web, and social networking sites. This huge data contains important information which is useful to organization for their respective purpose. It was question how to deal with this huge data to extract information from it and Data mining is the answer for it. According to Oracle, Data mining [1] is the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple analysis. Data mining uses sophisticated mathematical algorithms to segment the data and evaluate the probability of future events. Data mining is also known as Knowledge Discovery in Data (KDD).

Data mining is broad term which has many sub domains for specific purposes. Data mining consist of Classification, Clustering, Association Rules, Unsupervised Learning, Supervised learning, Feature selection, and Ranking etc. sub domains or sub topics. Each sub topic has different purpose. Clustering is used to cluster the data i.e. all similar records are gathered into one cluster. Association rules find out the association between the items in the data. In the learning procedure, if label of the instance is available the learning is call as supervised learning otherwise unsupervised learning. If data available for learning is high dimensional data or data contains the data of irrelevant feature then there is need to remove those irrelevant features, feature selection is used to detect the irrelevant features.

Classification is important technique in data mining, its details is as below.

B. Data Classification:

Data classification is a method to assign class to instance of the data. Instance in the data mining can be defined as, set of values of attributes/features. For example, I have something which is round in shape, color is red and edible, and so it's an apple. Here shape, color and is edible are the attributes or features of that thing. Round, red and edible are the values of those attributes. Apple is the class of the instance in this example. We have knowledge about apple i.e. it is round, red and edible therefore we are able to classify that instance. How machine will have this knowledge? To get knowledge about these machines needs learning. Classification method has two parts training and testing. Classification uses supervised approach of learning, in which training data has class labels. In training part, knowledge is extracted from training data and extracted knowledge is used to classify test data.

There are many algorithms for data classification. Each algorithm has different strategy for training /learning and Representation of knowledge. Most popular methods are neural network [2], k nearest neighbor [3], naïve Bayes, decision trees [4].

Details of decision trees are as follows:

C. Decision Tree

In decision trees, extracted knowledge in training step is represented using tree structure. There are many algorithms to form decision tree. All decision trees have same testing technique but each algorithm has its distinct method to form a tree i.e. each algorithm has different training method. Notable algorithms are ID3, C4.5, and CART.

ID3: ID3 stands for Iterative Dichotomiser 3 and it is invented by Ross quinlan [5] [6]. Which attribute from the attribute will be the root element and which attribute will be the child of the parent attribute is decided using Information gain of the attribute in iterative manner. Example of the how to form a decision tree using ID3 is explained at [6]. Properties of the ID3: (i) ID3 is non-incremental method i.e. new classes are not derived; only predefined classes will appear in the tree. (ii)Classes created by ID3 are inductive i.e. tree formed using small dataset is also applicable for future

Manuscript received June 2016

***Shubhangi A. Jadhav**, Smt. Kashibai Navle College of Engineering, Vadgaon Bk,Pune, City Pune, Country India,*
***Prof.S.P. Kosbatwar,** Smt. Kashibai Navle College of Engineering, Vadgaon Bk,Pune, City Pune, Country India,*

instances. (iii) Only applicable to nominal attributes i.e. attributes whose values are nominal (If Gender is attribute then nominal values are male, female).

C4.5: C4.5 is improved version of the ID3. It removes limitations of the Id3 and also adds other important features. It also uses the same criteria i.e. Information gain to decide parent and child in the decision tree.

Newly added features to the C4.5

- C4.5 can deal with numeric attribute .i.e. attributes which has continuous numeric value (Temperature is numeric attributes whose values are continuous)
- Can deal with incomplete data at training time and testing time also.
- Pruning of the decision tree is done by replacing a whole sub tree by a leaf node. The replacement takes place if a decision rule establishes that the expected error rate in the sub tree is greater than in the single leaf.

D. Existing Splitting Criteria

In the construction of the decision tree it is very important which attribute will be root or parent and which attribute will be child, this decision is taken based on some measure and this measure is called as splitting criteria measure. The basic fundamental of choosing attribute to find out attribute which will increase the accuracy and decrease errors. There are different measures in literature which are used for splitting decision in ID3 and C4.5 information entropy or information gain is used. In CART [7] algorithm gini index is used. Misclassification error is another measure which is used rarely in the literature for decision tree construction and this criteria has strong mathematical foundations and allows determining the optimal attribute to split the node of the tree. The misclassification error can be expressed a follows:

$G(S) = 1 - \frac{\boldsymbol{max}_{\mathbf{k}\in\{\mathbf{1},\ldots,\mathbf{k}\}}\{\boldsymbol{n^k(S)}\}}{\boldsymbol{n(S)}}$, Where, G(S) is a misclassification error of S, k is number of classes, n(S) is the cardinality of S, nk is number of elements of set S from kth class.

E. Decision Trees for Data streams

Data stream can be defined as flow of the data which is generated continuously. Examples of data streams are ATM transaction, Transaction in malls, social network such as twitter or Facebook, sensor reading, cctv cameras etc. To analyze this data stream standard data mining approaches cannot be applied because standard data mining algorithms are designed with assumption of static nature of the data. New field is arise to analyze data streams and called as data stream mining. There are many challenges to analyze the stream of the data.

- Requirement of On-the-spot data analysis because all data cannot be stored.
- Speed of analysis must be fast because data is coming continuously.
- Concept drift problem.

To fulfill the above mentioned requirements CVFDT (Concept adapting Very Fast Decision Tree Leaner) algorithm is helpful and advantageous. We used best available splitting criteria, based on Misclassification error attribute selection method which avoids the hoeffding bound problem in CVFDT and gives the better and efficient results.

II. Literature Review

Paper [8] uses Hoeffding tree and Hoeffding bound to construct the decision tree on high speed data streams and method is named as Very Fast decision tree (VFDT). Hoeffding trees can be learned in constant time per example (more precisely, in time that is worst-case proportional to the number of attributes), while being nearly identical to the trees a conventional batch learner would produce, given enough examples. Hoeffding bound evaluates how many examples are enough to decide the node in the decision tree. Information gain or gini index is used for attribute selection. From the experiments it is clear that, VFDT gives practically good solutions to form decision tree on data stream.

Paper[9] proposed an efficient algorithm called CVFDT (Concept-adapting Very Fast Decision Tree learner), for mining decision trees for continuously changing data streams based on ultra-fast VFDT decision tree learner. VFDT makes assumption that training data is random sample drawn from stationary distribution, but data streams available for mining violate this assumption. Instead of assuming data was generated by single concept, it is more accurate to assume that the data was generated by multiple concepts with time varying parameters. Concept drift problem (Relevancy of data over time) is addressed with high speed and accuracy.

In paper [10], the important challenge in data stream mining i.e. Concept Drift (unforeseen changes of the data stream's underlying data distribution) is addressed. This paper propose a new data stream classifier, called the Accuracy Updated Ensemble (AUE2), which aims at reacting equally well to different types of drift. AUE2 combines accuracy-based weighting mechanisms known from block-based ensembles with the incremental nature of Hoeffding Trees. It Provides bet average classification accuracy with less time consuming approach.

In Paper [11], L Rutkowski and his team proved that, Hoeffding bound used in literature is not enough to form the decision tree on the data streams and also stated that Hoeffding bound is mathematically wrongly proved in the literature. The Hoeffding bound is applicable only for the sum or average of random variables, whereas the information gain or Gini gain are nonlinear functions, which cannot be expressed in such a simple form. In this paper they suggested replacement for hoeffding bound as McDiarmid's bound. After evaluation they got equal result for information gain with Hoeffding bound and worse result for gini index with hoeffding bound.

Again in Jan 2014, L. Rutkowski [12] and his team worked on same issues but this time they come up with another approach. In this work they linearized the splitting functions using the Taylor's theorem and proposed new Gaussian Decision Tree Algorithm which is the modification of Hoeffding tree and which is designed on the basis of the ID3. From the examples they proved that GDT outperforms than McDiarmid's bound in the field of time consumption.

Numerical simulations proved that the GDT algorithm is able to give satisfactory accuracies in data streams classification problems. Unfortunately GDT is applicable for two class problem only.

In paper [13], new decision tree algorithm dsCART is designed for data stream which is based on CART algorithm. Paper proves that required number instances to split the node is lesser than the GDT also time requirement of dsCART is less than GDT. dsCART is applicable to any number of classes. The only limitation of this work is Gini gain the proportionality coefficient was dependent on the number of classes K and reached large values.

In MAY 2015 Leszek Rutkowski and his team concluded [14] that the problems in decision trees for data streams are arise due non linearity of attribute selection methods (Information Gain, Gini Index). In this paper they used Misclassification error as attribute selection method and named as mDT. To combine the advantages of misclassification error and gini index new attribute selection method is proposed as hybrid criterion.

III. System Architecture

In data stream mining data occurs continuously, so new examples coming through stream of data are taken for the further process. As continuous data is keep coming from the stream, we have considered a window of examples i.e. bunch of examples at a time for the processing. When a new example arrives, it is checked that whether the number of sufficient examples is exceeded or not. If window is overflow then the oldest example is forgotten. The decision of forgetting particular example is made on the basis of statistics maintained at every node. CVFDT monitors the validity of its old example by maintaining sufficient statistics at every node in a tree.

As the concept is changing, selected split attribute may go outdated because now an alternative attribute has higher gain. In this case CVFDT grows an alternative sub tree with a new best attribute at its root (CVFDT Grow). The best attribute is considered based on highest misclassification error and then it takes the decision whether to replace the old sub tree with the new alternate sub tree or not.

In normal CVFDT which is used hoeffding bound incorrectly, this decision of replacing the old sub tree with the new promising one is done by using hoeffding bound, where difference between the attribute showing highest gain and the attribute with second highest gain is compared with splitting criterion based on hoeffding bound. This leads to less effective tree generation. In proposed system instead of this measure,splitting criterion based on misclassification error as split evaluation function is used and it can be expressed a follows:

$G(S) = 1- \frac{\max_{k\in\{1,....,k\}}\{n^k(S)\}}{n(S)}$,

Where, G(S) is a misclassification error of S, k is number of classes, n(S) is the cardinality of S, nk is number of elements of set S from kth class. This is the split evaluation function and $z(1-\delta)\sqrt{\frac{1}{2n(S)}}$is a splitting criteria. Where δ is 1 – desired probability to select correct attribute at any node,$z(1-\delta)$ is the (1-δ) quintile of the standard normal distribution N(0,1) then with probability (1-δ) the ith attribute (Δgi(S)) would give higher value of the accuracy gain than the jth attribute (Δgj(S)) for the infinite dataset.

.

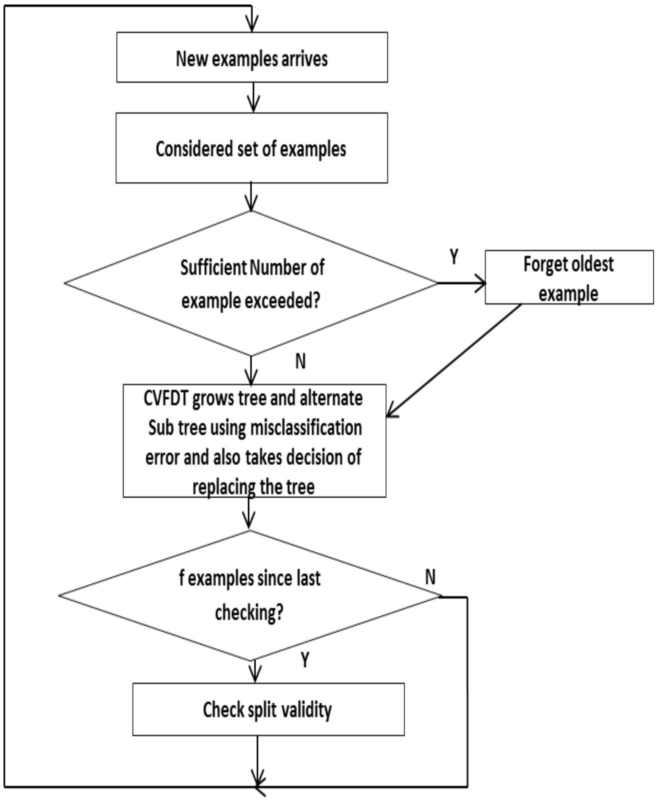

New examples arrives
Considered set of examples
Sufficient Number of example exceeded?
Y
Forget oldest example
N
CVFDT grows tree and alternate Sub tree using misclassification error and also takes decision of replacing the tree
f examples since last checking?
N
Y
Check split validity

Fig. 1. System Arcitecture

The decision of replacing the attribute is made with the following condition.

$\Delta gi(S) - \Delta gj(S) > z(1-\delta)\sqrt{\frac{1}{2n(S)}}$

It is possible that the tree may have changed since example was initially incorporated. A unique increasing ID is assigned to each node when they are created. An example's effects are forgotten by decrementing the counts in the sufficient statistics at every node the example reaches in tree whose ID is <= the stored ID. Then Checksplit validity checks whether to start the new alternate sub-tree or not based on best split attribute. The process executes iteratively.

IV. Mathematical Model

1. In decision tree data comes in records of the form

   (x,Y)=(x1,x2,………..,xn,Y)
   Where Y is target variable we are trying to classify
   X =(x,,x2,…….,xn) is attribute set used for classification

2. Synthetic data is presented as,
   SD={x,h,r,c,w,v} …………...….(1)
   Where
   SD= synthetic data

X={x1,x2,……..xn} is set of n dimension/attributes
H={h1,h2,……hn} is set of n hyperplane
R number of records to be generated
C concept drift parameter
W weight assigned
V value assigned for each attribute

3. Synthetic data is generated by using rotating hyperplane which is nothing but n dimensional space that is the set of points X that satisfy
   $\sum_{i=1}^{n}$ WiXi= Wo …..……………..........(2)
   Where w is weight
   Xi is ith co-ordinate of the X
4. Concept drift data generated by using following formula
   H=wo+c ……………..…..(3)
   Where c concept drift parameter
5. CVFDT using hoeffeding bound use information gain as split evaluation function which is given by
   $g(S)= -\sum_{k=1}^{K} pk(S) \log pk(S)$ ……………..…..(4)
   where,
   g(s) impurity measure or gain of s
   $Pk(S)= n^k(S) / n(S)$
   n(S) is the cardinality of S
   $n^k(S)$ be the number of elements of set S from the kth class.
6. Hoeffeding bound (ϵ) is given by
   $\epsilon(n(s), \delta) = R\sqrt{\frac{\ln(\frac{1}{\delta})}{2n(S)}}$ ……………….....(5)
7. hoeffeding bound criteria to check which attribute gives highest accuracy is given by
   $\Delta gi(S) - \Delta gj(S) > \epsilon(n(S), \delta)$ ……………….....(6)

8. CVFDT using misclassification error use misclassification error as impurity measure given by
   $g(S) = 1 - \frac{\max_{k\in\{1,\dots,k\}}\{n^k(S)\}}{n(S)}$, ……………….....(7)
   where,
   g(s) - misclassification error of set s
   k - class
   n(s) - cardinality of s
   $n^k(S)$ - number of element from set s belongs to kth class

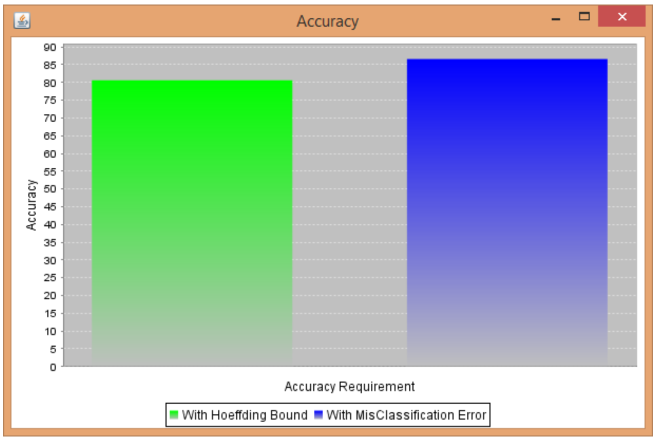
9. Misclassification error based criteria to check which attribute gives highest accuracy given by
   $\Delta gi(S) - \Delta gj(S) > z(1-\delta)\sqrt{\frac{1}{2n(S)}}$ ……………….....(8)
   Accuracy of classification given by
   Accuracy = (Nc/ N) * 100; ...…………....(9)
   Where,
   Nc= Number of correctly classified instances
   N= Total number of instances classified
10. Time Requirement for building classifier calculated by
    T = Tn – Ts ……………..(10)
    Where,
    Tn=End time
    Ts=start time
11. Memory Requirement calculated by
    Memory = Me – Ms ………..….(11)
    Where,
    Me= Free memory at the end of classifier building process
    Ms=Free memory at the start of the classifier building process.

V. Results

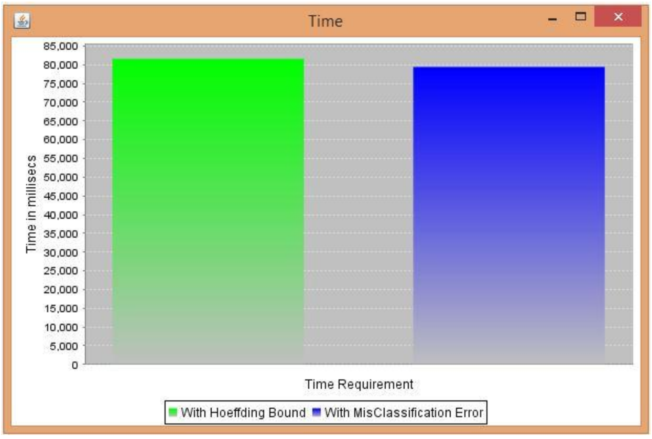
The system is developed for data stream mining. The system built to learn the time changing concepts and it addresses the concept-drift problem of data stream mining and also addresses the pitfalls from existing approach and improves the performance of stream mining. For the purpose of stream data mining, synthetic dataset of total 500000 instances is used for learning classifier. After performing the experiment the results showed that CVFDT with Misclassification error perform better than CVFDT with hoeffding bound in Accuracy and efficiency aspects.CVFDT with misclassification error takes 79 msectime and 1488 MB memory and CVFDT with hoeffding bound takes 81 msec time and1164 MB memory. Proposed approach significantly improves the accuracy of classifier learning on time changing data streams.

Table 1 Results for proposed and existing system

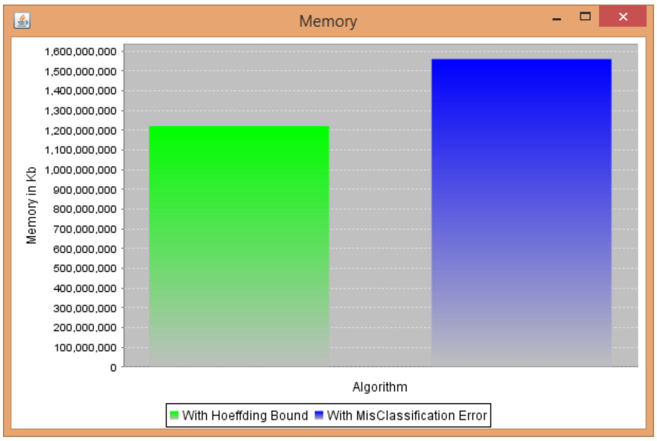
Measure	CVFDT with MSE	CVFDT with Hoeffding bound
Time (msec)	79	81
Memory (KB)	1220606992	1560717000
Accuracy (%)	86.59	80.60



Graph1 Comparison of Accuracy gained for proposed and existing system



Graph 2 Comparison of Time requirement for proposed and existing system



Graph3 Comparison of Memoryrequirement for proposed and existing system

VI. Conclusion

This paper discussed the problems in the data stream mining and also discussed various approaches used to solve them. The Paper particularly focuses on how misclassification error can be good attribute split measure due to its linearity. This paper used a decision tree algorithm and introduced the use of Misclassification error as attribute selection measure.

REFERENCES

[1] http://docs.oracle.com/cd/B28359_01/datamine.111/b28129/process.htm

[2] A. Martin and P. Bartlett, Neural network learning: Theoretical Foundations. Cambridge, U.K.: Cambridge Univ. Press, 2009. (Partly available on Google books)

[3] T. Cover and P. Hart, “Nearest neighbor pattern classification,” IEEE Trans. Inf. Theory, vol. 13, no. 1, pp. 21–27, Jan. 1967.

[4] 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. (Available on IEEE)

[5] Faculty of Information Technology, MONASH UNIVERSITY, “CSE5230 Tutorial: The ID3 Decision Tree Algorithm”

[6] http://www.cise.ufl.edu/~ddd/cap6635/Fall-97/Short-papers/2.htm

[7] L. Breiman, J. Friedman, R. Olshen, and C. Stone, Classification and Regression Trees. London, U.K.: Chapman and Hall, 1993.

[8] P. Domingos and G. Hulten, “Mining high-speed data streams,” in Proc. 6th ACM SIGKDD Int . Conf. Knowl. Discovery Data Mining, 2000, pp. 71–80.

[9] G. Hulten, L. Spencer, and P. Domingos, “Mining time-changing data streams,” in Proc. 7th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2001, pp. 97–106.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] Leszek Rutkowski, Fellow, IEEE, Maciej Jaworski, Lena Pietruczuk, and Piotr Duda “A New Method for Data Stream Mining Based on the Misclassification Error” May 2015.

ShubhangiAJadhav Pursuing ME (computer engineering) from Smt.Kashibai Navale College of Engineering, Pune, India. Savitribai Phule Pune University, Pune, Maharashtra, India -411007. I have completed BE(Information technology) from PVPIT budhgaon Sangli. My area of interest is data mining and information retrieval.

Prof.S.P. Kosbatwarreceived his ME. (Computer) Degree from M.G.M, College of Engineering,Nanded, Maharashtra, India. He received his B.E (Computer)Degree from Govt college of Engineering, Aurangabad,Maharashtra, India. He is currently working as Asst Prof (Computer) at Department of Computer Engineering, Smt. Kashibai Navale College of Engineering, and Pune, India. Savitribai Phule Pune University, Pune, Maharashtra, India -411007. His area of interest is data mining and information retrieval.