Review on Gaussian Estimation Based Decision Trees for Data Streams Mining

Miss. Poonam M Jagdale[1], Asst. Prof. Devendra P Gadekar[2]

[1,2]Pune University, Pune

Abstract— *In recent year, mining data streams decision trees became one of the most popular tools for the mining data streams. The Hoeffding tree algorithm was projected in this paper. To decide the best attribute to split the considered node is a key point of the building decision tree. Many researchers work on this problem and present the methods which are used to solve this problem. But they are incorrect. Such as Hoeffding tree algorithm is justified mathematically incorrectly and McDiarmid tree algorithm which is very time consuming. To overcome this problem, this paper projected a new method which has solid mathematical basis and gives better performance than the McDiarmid tree algorithm. The selection of best attribute in the considered node with the help of finite data sample is similar as it would be in the case of the entire data stream with the high probability set by the user is make sure by the this method.*

In the data mining community mining data stream become an extremely demanding task researcher by C. Aggarwal, A. Bifet, R. Kirkby, W. Fan, Y. Huang, H. Wang, M.M Gaber, A. Zaslavsky, B. Pfahringer, Aggarwal et al. (2010). With the high rate data elements are incessantly entered in the data stream. That's why it has infinite size. Also here the concept drift is found which the concept of data involved in time is. Due to the concept drift in data streams, the data mining algorithm cannot apply directly. This paper describes the one of the data mining technique that is classification task (Bifet, 2009). Labeling the unclassified data and learning and learning from the training data set.

This literature projected a classification methods multitude for the static data neural networks by Rutkowski et al. (2004), decision trees by Breiman et al. (1993) and k-nearest neighbors Murty et al. (2011). This paper focus on the decision trees based classifier which is the one of the most effective. 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 nonbinary 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 binary case some sub set of number of elements of set is allocated to the each branch. If the attribute takes nominal values at that time only the non binary trees makes sense. To continue the process of tree growth the training set is divided into subsets on the basis of attribute values allocated to the branches and it is propel towards the equivalent children nodes. It is also used to allocate a class to the unclassified data elements. Selecting the best attribute to split the consider node is the key point of building the decision tree. 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 ID3 algorithm supports the attributes with big domain of probable values in the case of non-binary trees. It is the main disadvantage of the ID3 algorithm. This problem is solved by using C4.5 algorithm (J.R. Quinlan, 1993). To introduce the split-information function, which penalizes the attributes with large domains is the major thought. In the C4.5 algorithm, the split measure function is projected as a ratio of the information gain and the split information. The another impurity measure worth consideration is Gini index which is used in the CART algorithm by Breiman et al. (1993) which is used to develop a binary trees. Due to this it can be applicable to the nominal attribute values data as well as to the numerical data.

But this algorithm is appropriate for static data sets and cannot apply directly to the data stream and required important modification. Data streams have infinite size that's why the best attribute in each node establishment is the leading problem. By referring the two papers which *constitute the "state of the art" in this subject is the main and original result of this paper can be summarized here.*

Based on the McDiarmid's inequality (C. McDiarmid, 1989), Rutkowski et al. (2013) projected a method. In the considered node needs very large amount of data elements for the selection the best attribute.

Based on the multivariate delta method, the one more method is projected by Jin et al. (2003). Though the idea was promising, the result was wrong and not appropriate to the problem of building decision trees for data streams. To determine the best attribute in a node which ensure the highest value of the split-measure function with significantly high probability is done by statistical method is projected in this paper. The properties of the normal distribution and Taylor's theorem (L. Wasserman, 2005) are also used in this method (O. Kardaun, 2005).

I. LITERATURE REVIEW

Pattern Optimization for Novel Class in MCM for Stream Data Classification paper explains that the classification of stream data is someway hard. Existing data stream classification techniques presume that total number of classes in the stream is fixed. Therefore, instances belonging to a novel class are misclassified by the existing techniques. Because data streams have endless length, conventional multi pass learning algorithms are not suitable as they would require infinite storage and training time. Concept-drift occurs in the stream when the fundamental concept of the data changes over time. Thus, the classification model must be updated continuously so that it reflects the most recent concept.

An Adaptive Ensemble Classifier for Mining Concept-Drifting Data Streams paper discuss that traditional data mining techniques cannot be directly applied to the real-time data streaming environment. Existing mining classifiers therefore require to be updated frequently to adopt the changes in data streams. In this paper, we address this matter and propose an adaptive ensemble approach for classification and novel class detection in concept-drifting data streams. The planned approach uses traditional mining classifiers and updates the ensemble model automatically so that it represents the most current concepts in data streams. For novel class detection we consider the thought that data points belonging to the same class should be closer to each other and should be far apart from the data points belonging to other classes. If a data point is well separated from the existing data clusters, it is identified as a novel class instance.

Classification and Novel Class Detection in Concept-Drifting Data Streams under Time Constraints most existing data stream classification techniques ignore one significant feature of stream data: influx of a novel class. We address this subject and suggest a data stream classification technique that integrates a novel class detection mechanism into traditional classifiers, enabling automatic detection of novel classes before the true labels of the novel class instances arrive. Novel class detection problem becomes more challenging in the presence of concept-drift, when the fundamental data distributions evolve in streams. In order to decide whether an instance belongs to a novel class, the classification model sometimes wants to stay for more test instances to discover similarities among those instances.

A New Classification Algorithm for Data Stream projected a Associative classification (AC) which is based on association rules has exposed great promise over many other classification techniques on static dataset. Meanwhile, new challenges have been future in that the increasing fame of data streams arising in a wide range of advanced application. This paper describes and evaluates a new associative classification algorithm for data streams AC-DS, which is based on the opinion mechanism of the Lossy Counting (LC) and landmark window model. And AC-DS was applied to mining several datasets obtained from the UCI Machine Learning Repository which is effective and efficient.

Integrating Novel Class Detection with Classification for Concept-Drifting Data Streams proposed a novel and efficient technique that can automatically detect the appearance of a novel class in the presence of concept-drift by quantifying cohesion among unlabeled test instances, and separation of the test instances from training instances. Our approach is non-parametric, meaning; it does not suppose any underlying distributions of data. Comparisons with the state-of-the-art stream classification techniques prove the superiority of this approach.

A Nearest-Neighbor Approach to Estimating Divergence between Continuous Random Vectors is a method for divergence estimation between multidimensional distributions based on nearest neighbor distances is proposed. Given i.i.d. samples, both the bias and the variance of this estimator are established to disappear as sample sizes go to infinity. In experiments on high-dimensional data, the nearest neighbor approach usually show faster convergence compared to previous algorithms based on partitioning.

In A Framework for On-Demand Classification of Evolving Data Streams projected an on-demand classification process which can dynamically select the suitable window of past training data to build the classifier. The experiential results point to that the system maintains high classification correctness in a developing data stream, while providing an efficient solution to the classification job.

Accurate Decision Trees for Mining High Speed Data Streams paper study the problem of building accurate decision tree models from data streams. Data streams are incremental tasks that require incremental, online, and any-time learning algorithms. One of the most successful algorithms for mining data streams is VFDT.

In CVFDT ALGORITHM FOR MINING OF DATA STREAMS, system implementation is based on the decision tree of CVFDT Algorithm to address the inequalities projected in stream mining. The planned work also uses the network data streams to examine the attacks using splitting attribute with gain values. The builder tree can be used for classification of new observation. It gives better performance than the Hoeffding trees.

Forest Trees for On-line Data paper presents a hybrid adaptive system for induction of forest of trees from data streams. The Ultra Fast Forest Tree system (UFFT) is an incremental algorithm, with steady time for dispensation each example, works online, and uses the Hoeffding bound to decide when to install a splitting test in a leaf leading to a decision node.

Accurate Decision Trees for Mining High-speed Data Streams paper study the problem of building accurate decision tree models from data streams. Data streams are incremental tasks that require incremental, online, and any-time learning algorithms.

The rise and fall of redundancy in decoherence and quantum Darwinism inspects circumstances wanted for the formation of branching states and studies their termination through many-body interactions. They demonstrate that even forced dynamics can suppress redundancy to the values typical of random states on relaxation timescales, and show that these results hold precisely in the thermodynamic limit.

Data mining and knowledge detection in databases have been attracting a important quantity of investigate, manufacturing, and media attention of late. What is all the enthusiasm about? This article provides an impression of this emerging field, descriptive how data mining and knowledge discovery in databases are related both to each other and to related fields, such as machine learning, statistics, and databases. The article mentions particular real-world applications, specific data-mining techniques, challenges involved in real-world applications of knowledge discovery, and current and future research directions in the field.

Adaptive Mining Techniques for Data Streams using Algorithm Output Granularity is a resource-aware approach that is adaptable to available memory, time constraints, and data stream rate. The approach is generic and applicable to clustering, classification and counting frequent items mining techniques.

Knowledge Fusion for Probabilistic Generative Classifiers with Data Mining Applications says that if knowledge such as classification rules is extracted from sample data in a distributed way, it may be necessary to combine or fuse these rules. In a conventional approach this would typically be done either by combining the classifiers' outputs (e.g., in form of a classifier ensemble) or by combining the sets of classification rules (e.g., by weighting them individually). In this paper, introduce a new way of fusing classifiers at the level of parameters of classification rules. This technique is based on the use of probabilistic generative classifiers using multinomial distributions for categorical input dimensions and multivariate normal distributions for the continuous ones.

Compact Tree for Associative Classification of Data Stream Mining proposes a new scheme called Prefix Stream Tree (PST) for associative classification. This helps in compact storage of data streams. This PSTree is generated in a single scan. This tree efficiently discovers the exact set of patterns from data streams using sliding window.

The CART Decision Tree for Mining Data Streams proposes a new algorithm, which is based on the commonly known CART algorithm. The most important task in constructing decision trees for data streams is to determine the best attribute to make a split in the considered node. To solve this problem they apply the Gaussian approximation. The presented algorithm allows obtaining high accuracy of classification, with a short processing time. The main result of this paper is the theorem showing that the best attribute computed in considered node according to the available data sample is the same, with some high probability, as the attribute derived from the whole data stream.

Decision Tree Evolution Using Limited Number of Labeled Data Items from Drifting Data Streams proposes a new concept of demand-driven active data mining. In active mining, the loss of the model is either continuously guessed without using any true class labels or estimated, whenever necessary, from a small number of instances whose actual class labels are verified by paying an affordable cost. When the estimated loss is more than a tolerable threshold, the model evolves by using a small number of instances with verified true class labels.

An Intelligent Association Rule Mining Model for Multidimensional Data Representation and Modeling paper presents a new algorithm called Fuzzy-T ARM (FTA) to classify the breast cancer dataset. In this work, ARM is used for reducing the search space of the Multidimensional breast cancer dataset and Fuzzy logic is used for intelligent classification. The dimension of input feature space is reduced the instances from one third by using ARM. The FTA has applied to the Wisconsin breast cancer dataset to evaluate the overall system performance. This research demonstrated that the ARM can be used for reducing the dimension of feature space and the proposed model can be used to obtain fast automatic diagnostic systems for other cancer diseases.

A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise present the new clustering algorithm DBSCAN relying on a density-based notion of clusters which is designed to discover clusters of arbitrary shape. DBSCAN requires only one input parameter and support the user in determining and appropriate value for it.

On Reducing Classifier Granularity in Mining Concept-Drifting Data Streams show that reducing model granularity will reduce model update cost. Indeed, models of fine granularity enable us to efficiently pinpoint local components in the model that are exaggerated by the concept drift. It also enables us to derive new components that can easily integrate with the model to reflect the current data distribution, thus avoiding expensive updates on a global scale.

On Demand Classification of Data Streams projected a model for data stream classification sights the data stream classification problem from the point of view of a dynamic approach in which concurrent training and testing streams are used for dynamic classification of data sets. This model reflects real life circumstances effectively, since it is desirable to classify test streams in real time over an evolving training and test stream.

Data Clustering: A Review paper discuss that Clustering is the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters). The clustering problem has been addressed in many contexts and by researchers in many disciplines; this reflects its broad appeal and usefulness as one of the steps in investigative data analysis.

Classification and Novel Class Detection in Data Streams with Active Mining present ActMiner, which addresses four major challenges to data stream classification, namely, infinite length, concept-drift, concept- evolution, and limited labeled data.

Decision trees for mining data streams based on the McDiarmid’s bound exposed that the Hoeffding’s inequality is not suitable to solve the underlying problem. And also prove two theorems presenting the McDiarmid’s bound for both the information gain, used in ID3 algorithm, and for Gini index, used in CART algorithm. The results of the paper assurance that a decision tree learning system, practical to data streams and based on the McDiarmid’s bound, has the property that its output is nearly identical to that of a conventional learner.

Ubiquitous Data Mining (UDM) is the procedure of performing analysis of data on mobile, embedded and ubiquitous devices. It represents the next generation of data mining systems that will support the intelligent and time-critical information requirements of mobile users and will facilitate “anytime, anywhere” data mining. The underlying focus of UDM systems is to perform computationally intensive mining techniques in mobile environments that are forced by limited computational resources and varying network characteristics.

II. ID3 ALGORITHM

The background of our projected is provided by the ID3 algorithm. It is firstly developed to produce binary trees though it can be easily transformed to the binary mode. Here we focus on the binary case though this method can adapt to nonbinary trees also. This algorithm initiate with single node that is root node. Exacting subset of the training data set is procedure in each created node throughout the learning process. The node is tagged as a leaf and the split is not made if the all elements of the set are of the similar class or select the best attribute to split amongst the obtainable attributes in the considered node. The set of attribute values of some subset (A^i) is divided into two disjoint subsets A^i_L and A^i_R ($A^i = A^i_L \cup A_R{}^i$) for every obtainable attribute. The divider is symbolized additional only by $A_L{}^i$. The complementary subset $A_R{}^i$ is automatically determined by selection of $A_L{}^i$. In the ID3 algorithm split-measure function used as a maximizes information gain is a difference between the entropy and the weighted entropy. The maximizes value of the information gain is selected from the all likely partition of the set. For the subset of the training data set, the partition information gain is also called the optimal partition of number of element set. It is used to generate subset and this value called as an information gain of subset for attribute. One of the highest values of information gain is selected from the obtainable attributes in the node. The node split into two children nodes where the index of nodes created in the entire tree. The following two circumstances happened if the considered node is not split. They are all elements from the subset are from the same class. And only one element is present in the list of available attributes in the node. The problem of the concept drift is used as a part of the CVFDT algorithm by Hutten et al. (2001) in this paper. It also replaces the Hoeffding's bound which is used incorrectly in the CVFDT algorithm. The thought of CVFDT algorithm published initially by Domingo's and Hulten in 2001 is correct, though these authors incorrectly used the Hoeffding's bound in their paper.

III. C4.5 Algorithm

C4.5 is an algorithm used to produce a decision tree developed by Ross Quinlan. It is the extension of ID3 algorithm that accounts for unavailable values, incessant attribute value ranges, pruning of decision trees, rule derivation, and so on. It is also refer as a statistical classifier. In non binary case the ID3 algorithm favors the attributes with large domain of possible values. To manage up with this problem C4.5 algorithm is used. In the C4.5 algorithm the ratio of the information gain and the split information is projected as the split measure function. It has few base cases such as If all the samples in the list belong to the same class then it just creates a leaf node for the decision tree proverb to select that class. C4.5 creates a decision node higher up the tree using the predictable value of the class if not any of the features provide any information gain. C4.5 again creates a decision node higher up the tree using the predictable value if Instance of previously-unseen class encountered.

IV. RELATED WORK

It is extremely tricky to adapt the ID3 algorithm or any decision trees based on algorithm to data stream. The equivalent subsets of training data set incessantly cultivate due to this cause it is tricky to approximation the values of split-measure function in each node. On the basis of infinite training data set hypothetically the information gain values is intended in the data stream case. But it is impossible that's why these values predictable from the obtainable data sample in the considered node. Due to this only with some possibility, one can make a decision which attribute is the most excellent. Here this paper converse the few efforts to solve this problem.

"Hoeffdings trees" is the result of the P. Domingos and G. Hulten work result for the data mining streams. It was resulting from the Hoeffding's bound (W. Hoeffding, 1963), which states that with probability $1-\delta$ the accurate mean of a random variable of range does not vary from the predictable mean. To solve the difficulty of selecting the attribute according to which the split should be made Hoeffding's bound is not a sufficient tool. It is a correct tool only for numerical data, which does not of necessity have to be met become aware of by the Rutkowski et.al (2007). The split measures like information gain and Gini index form is the second problem. Both measures are uses the elements frequency and cannot be expressed as a sum of elements. One more method for finding the best attribute was projected in the work by Agrawal et al. (2003). One exacting node will be considered for the expediency of the following text situation. Hence, the node index q will be absent in all notations introduced before. To make a decision is there attribute gives higher value of information gain than other attribute, based on distribution, the suitable statistical test projected by the authors. Multivariate delta method is used to give good reason for the estimate. The correcting the mathematical foundations of Hoeffding's trees worked by the Rutkowski et al. (2013). Selecting the best attribute to make split in the node is extremely hard task and to solve this problem it is projected a McDiarmid's inequality.

V. GAUSSIAN DECISION TREES ALGORITHM

A Gaussian decision tree algorithm which is the modification of the Hoeffding tree algorithm projected in Domingos et al. (2000). The algorithm begins with a single root also called leaf and the input parameters are initialized. The statistics of elements collected in the root are initialized which enough to compute all the essential values. In the main loop of the algorithm, using the current tree it gets data or element from the stream and sorts it into a leaf. All statistics and majority class in leaf is updated. After that it ensures is there any class which is dominated to the other classes. It is also known as preprinting condition. The information gain values are calculated for each attribute if there is not establish any prepruning condition. After that the determination of the best attribute and second best attribute takes place. Then they calculate the value and verify that obtained values are enough or not to make a decision whether the split should be made or not. The leaf is replaced by a node with the attribute allocate to it if the answer is positive. At last the algorithm returns were a new data element from the stream is taken.

VI. CONCLUSION

With the decision trees application, the issues of data mining streams subjects measured in this paper. Selecting the best attribute to split the considered node is the key point in building the decision tree. This is solved by projecting the new method were if the best attribute determined for the current set of data elements in the node is also the best according to the entire stream. It is based on the properties of the normal distribution and Taylor's Theroms. Also C4.5 algorithm is use for building the decision tree which conquers the difficulty of ID3 algorithm. It is also essential mathematically. We also projected a GDT that is Gaussian Decision Tree algorithm. This 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.

REFERENCE

[1] Bifet, G. Holmes, G. Pfahringer, R. Kirkby, and R. Gavalda, "New Ensemble Methods for Evolving Data Streams," Proc. 15th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '09), June/July 2009.

[2] Pfahringer, G. Holmes, and R. Kirkby, "New Options for Hoeffding Trees," Proc. 20th Australian Joint Conf. Advances in Artificial Intelligence (AI '07), pp. 90-99, 2007.

[3] Aggarwal, Data Streams: Models and Algorithms. Springer, 2007.

[4] McDiarmid, "On the Method of Bounded Differences," Surveys in Combinatorics, J. Siemons, ed., pp. 148-188, Cambridge Univ. Press, 1989.

[5] G. Hulten, L. Spencer, and P. Domingos, "Mining Time-Changing Data Streams," Proc. Seventh ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 97-106, 2001.

[6] J.R. Quinlan, C4.5: Programs for Machine Learning. Morgan Kaufmann, 1993.

[7] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone, Classification and Regression Trees. Chapman and Hall, 1993.

[8] L. Rutkowski, L. Pietruczuk, P. Duda, and M. Jaworski, "Decision Trees for Mining Data Streams Based on the McDiarmid's Bound," IEEE Trans. Knowledge and Data Eng., vol. 25, no. 6, pp. 1272-1279, 2013.

[9] M.M Gaber, A. Zaslavsky, and S. Krishnaswamy, "Mining Data Streams: A Review," ACM SIGMOD Record, vol. 34, no. 2, pp. 18-26, June 2005.

[10] M. Narasimha Murty and V. Susheela Devi, Pattern Recognition: An Algorithmic Approach. Springer, 2011.

[11] P. Domingos and G. Hulten, “Mining High-Speed Data Streams,” Proc. Sixth ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining, pp. 71-80, 2000.

[12] R. Jin and G. Agrawal, “Efficient Decision Tree Construction on Streaming Data,” Proc. Ninth ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining, 2003.

[13] W. Fan, Y. Huang, H. Wang, and P.S. Yu, “Active Mining of Data Streams,” Proc. SIAM Int’l Conf. Data Mining (SDM ’04), 2004.