RESEARCH ARTICLE OPEN ACCESS

Novel Drift Detection on Streaming Data Using Enhanced SVM Technique

Mrs. M.Kavitha [1], Mrs.R.Baby [2]
Assistant Professor [1], M.phil Research Scholar [2]
PG and Research Department of Computer Science
Tiruppur Kumaran College for Women, Tirupur
Tamil Nadu - India

ABSTRACT

Data stream classification has been an extensively study research problem in recent years. The active and developing nature of data streams need competent and effectual method that are considerably different from still data categorization techniques. Two problem of the most demanding and well studied distinctiveness of data streams are its unlimited length and concept-drift. Data stream categorization poses much challenge for the data mining society. In this research process, the focal point is on four key challenges, such as, concept-evolution, infinite duration, concept-drift, and the final one is feature-evolution. Concept-drift is a general inclination in data streams, which occurs as an consequence of changes in the underlying concepts. When a new class is arrived then here concept evolution occurs. Characteristic development is an over and over again happening process in most of the streams, such as text streams, in which novel feature appear as the stream development. Most existing data stream categorization method address only the first two challenges that is the drift and length, and ignore the latter two. If we consider among each type of data those are need for classification for best results, so we have proposed our SVM in our paper for feature extraction and KNN Classifier for data stream from beginning to end data set. A class label, then concept-evolution occurs when a completely new kind of stream occurs in the dataset.

Keywords :- *Data Streams, Concept Drift, Novel Class, Feature Evolution, SVM Classification.*

I. INTRODUCTION

A data stream is a well thought-out series of instances that reach your destination at a rate that does not allow to everlastingly storing them in memory. Data streams are potentially abundant in size construction them not possible to process by the majority data mining methods. The major distinctiveness of the information stream model imply the subsequent constraints. It is not achievable to store all the in sequence from the data stream. Only small summaries of data streams can be premeditated and stored, and the relax of the data is thrown away. The entrance speed of information stream tuples services each exacting element to be processed fundamentally in real time, and then discarded. The sharing generating the items can change over moment. Thus, information from the precedent may happen to irrelevant or even destructive for the recent summary. Constraint limits the extent of memory that algorithms in service on information streams can use, while constraint confines the time in which an item can be processed.

The initial two constraints led to the augmentation of information stream summarization method. Constraint is more major in some applications than in others. Many of the preliminary data stream mining techniques ignored this characteristic and created the group of static data stream learning algorithms.

Other researches well thought-out constraint as a key characteristic and dedicated their work to evolving information stream learning. Most of data stream investigation, querying, classification, and clustering applications need some variety of summarization methods to keep happy the earlier mentioned constraints.

Summarization mechanisms are second-hand for producing approximate answers from huge information sets in the main through information reduction and synopsis manufacture. This can be inclusive by selecting only a subset of incoming information or by using sketching, consignment shedding, and aggregation methods. In the subsequently paragraphs, we present the crucial methods used to minimize data stream amount and speed for analysis purposes.

In this journal we think about the problem of concept drift in supervised learning where the accurate classification for each instance (label) is delayed. In particular, we are fascinated in the training set formation strategy which is able to reform the training sets after in view of each concept drift , this will lead to achieving adaptively to concept drift.

Change detection is not just a task to decide whether or not the distribution change, but also it must analyze and give exact reasons about the change. This is important to choose the suitable adaptivity strategy. Most of the work to date on both drift detection and drift handling assumes that the true class of all instances in the data stream will be known shortly after classification. Under such assumption the incoming new data can be regularly used to update the model. Some works like use active learning technique which is used to build classifiers from large collections of unlabeled examples with the assistance of a human expert. The human expert is asked to label only those examples that are deemed to be most informative to the training process.

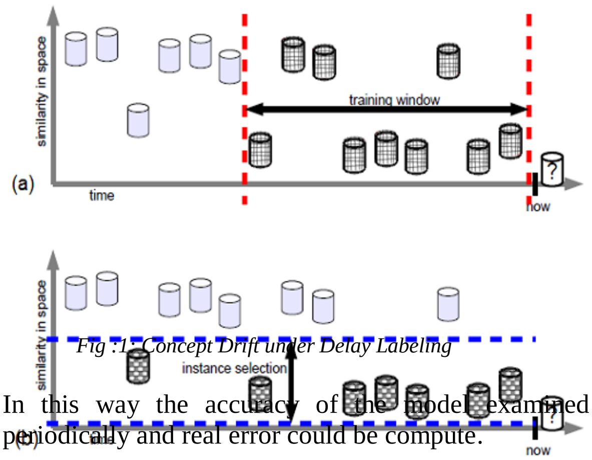

Fig :1 Concept Drift under Delay Labeling

In this way the accuracy of the model examined periodically and real error could be compute.

II.RELATED WORK

In this investigate "A Framework for On-Demand Classification of Evolving Data Streams" [2] presented by the authors Charu C. Aggarwal, elder Member, IEEE, Jiawei Han, elder Member, IEEE, Jianyong Wang, a Member, IEEE, and Philip S. Yu, Fellow, IEEE , Current models of the categorization problem do not successfully handle bursts of particular classes coming in at diverse times. In fact, the current model of the classification problem purely concentrates on methods for one-pass classification modeling of very huge data sets.

Their representation for data stream classification views the data stream classification problem from the point of view of a forceful approach in which simultaneous training and test streams are used for forceful classification of data sets. This model reflects real life situations successfully, since it is desirable to classify test streams in real point in time over an evolving guidance and test stream.

The most important intend here is to cause a categorization system in which the training model can fiddle with rapidly to the changes of the primary information stream. In order to accomplish this aim, they anticipated an on-demand classification procedure which can enthusiastically choose the suitable window of precedent training information to build the classifier. The empirical consequences indicate that the system maintains high classification exactness in an evolving data stream, while provided that an efficient solution to the classification assignment

In the paper with the "New Ensemble Methods For Evolving Data Streams" [3] the authors of the paper.Albert Bifet, Geoff Holmes, Bernhard Pfahringer, Richard Kirkby Ricard Gavaldà stated that sophisticated analysis of data streams is rapidly becoming a key area of data mining look into as the number of applications demanding such processing raises.

In the paper titled as "Using Additive Expert Ensembles to Cope with Concept Drift" [4] the authors Jeremy Z. Kolter and Marcus A. Maloof well thought-out online learning where the objective concept can revolutionize over time.

Learning algorithms often necessitate working in dynamic environments, which are varying unexpectedly. One popular belongings of these algorithms is their capability of incorporating innovative information If the information creating procedure is not strictly stationary (as applies to most of the authentic world applications), the primary concept, which we are predicting, may be unreliable over time. The capability to adapt to such concept drift can be seen as a expected extension for the incremental learning systems [Giraud-Carrier 2000] that gain knowledge of predictive models example by example. Adaptive learning algorithms can be seen as difficult incremental learning algorithms that are intelligent to adjust to progression of the data generating development over occasion.

When there are more new classes formed, then the classes with less comfortable are discarded so that the item of classes is maintained within a given limit and this avoids the infinite complication.

II. PROPOSED APPROACH OUTLIER DETECTION

When the data arrived is more and the classes formed out of them increases the problem is termed as infinite length problem. This is to be avoided. Each incoming instance in the data stream is first examined by an outlier detection module to check whether it is an outlier. If it is not an aberration, then it is confidential as an existing class using bulk polling among the confidential in the altogether. If it is an outlier, it is briefly stored in an error.

CONCEPT DRIFT IDENTIFICATION

The words and the category to which it belongs are added in the 'category' table. A client application is developed in which the text content is sent to the server application which updates the incoming message.

The words are extracted and the words fell in the given category are identified and counted. If there are more words in the category and the word count reduced in the successive incoming messages, then the concept is found to be reduced and when the number of words reduced to zero, the concept is said to be drifted. The number of observation time count is set so that when the number of word count is zero for that given number of time, then the concept is said to be drifted.

NOVEL CLASS DETECTION

During the concept evolution phase, the novel class detection module is invoked. If a novel class is found, the instances of the novel class are tagged accordingly. Differently, the occurrences in the error are considered as an older process class and classified normally using the altogether of models. The words occurred frequently but not matched with any of the division available, and then the word is studied to be fallen in new class.

Algorithm Used

Adjust-Threshold(x, OUTTH)

Input: x, OutTh

Which are most recent labeled instance and OutTh is current outlier threshold

Process:

i. Populate the class labels
ii. Check the incoming X data is matched with any of the class
iii. If not fallen in any of the class, then new class is said to be occurred. It is increased with a slack variable.

Output: OUTTH

New outlier OutTh threshold.

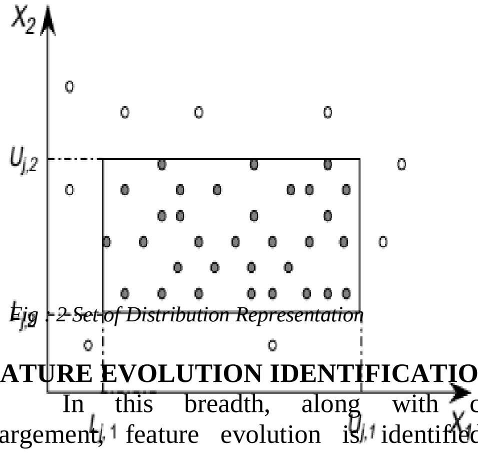

Fig 2 Set of Distribution Representation

FEATURE EVOLUTION IDENTIFICATION

In this breadth, along with concept enlargement, feature evolution is identified. The imitated process is identified in the accepted information and if it is found that more number of accepted information contains the patterns, then it is said that feature evolution occurs.

The following output forms are available.

1.Concept Evolution Identification
2.Feature Evolution Identification

CONCEPT EVOLUTION IDENTIFICATION

In this form, the words occurred frequently but not matched with any of the category available, and then the word is considered to be fallen in new class. A notify icon is displayed when new concept is evolutes.

FEATURE EVOLUTION IDENTIFICATION

The repeated patterns are identified in the received messages and if it is found that more number of received messages contains the patterns, then it is said that feature evolution occurs. A notify icon is displayed when new concept is evolves..

K-NN CLASSIFIER ALGORITHM

In accumulation to the class label outputted by k-NN classifier, we personalized , so it can output two additional class labels, y′′ and y′′′ for the matching example. The basic proposal of the algorithm does not revolutionize, but we add two more computations, one for y′′ and the other y′′′. The rationale of doing this computing is to make a decision later what class label should be fixed to the given drift example. The particulars of this development and how the standards of y′ ′and y′′′ are used will be explained in the subsequently section.

CLOSEST CLASS

We develop this computation as a heuristic to facilitate us to get the nearest class to in progress available classes. Many other methods estimate the distance between centers openly to get how much one class is outlying from the others. These methods may not work fit when the distribution of the instance points fit in to one class label is scattered and non-intensive. This heuristic guides the algorithm to make a decision how to change the class label when there is a drift especially when the drift is gradual.

SUPPORT VECTOR MACHINE CORRECTION WITH DRIFTED DATA

This section introduces the most important contribution of this paper. A method to correct SVM models opposite sudden drift is described. To correct SVM models, the method uses a modification based on the divergence between the preliminary dataset and the innovative drifted dataset. Let D1 correspond to the un-drifted dataset through labeled instances and D2 the drifted dataset devoid of labels. The SVM model M1 learns simply from D1. The meta-parameters of the SVM are preferred by cross validation.

IV.PERFORMANCE ANALYSIS

The document shows how to solve the structural trouble with a *cutting plane* technique, where as an alternative of solving a complex convex issue, we solve an estimation of it consisting of a number of hyper planes (these are the cutting planes); this is similar to the idea behind bundle methods, where it turns out that SVMperf can be viewed as a special case of a bundle method. We can solve the structural SVM trouble iteratively by keeping a functioning set *W* of constraint indices, and solving the trouble limited to the constraints in *W*. Each element of *W* is a vector *w* 2 *f*0;1*gn*, and so can be idea of as some combination of training point indices; when $W = f0;1gn$, then we solve the accurate structural trouble.

This working set is updated at each iteration to contain the indices for the points that are presently misclassified: we add a new binary vector *w* to *W*, where *wi* denotes if the *i*th training example is misclassified or not. The algorithm terminates when it is within ρ of the optimal primal solution, and this is shown to happen in at most $O(1{=}\lambda\rho2)$ iterations, and so SVMperf , like Pegasos, depends on the difficulty of the task (λ ! 0 the training time increases). Unlike Pegasos, the dependence on $1{=}\rho$ is quadratic, meaning that it takes longer to reach an approximate solution.

VI. CONCLUSION

The project identifies two key mechanisms of the novel class detection technique, namely, outlier detection, and identifying novel class instances, as the prime cause of high error rates for previous approaches. To solve this problem, the project proposes an improved technique for outlier detection by defining a slack space outside the decision boundary of each K-NN classification model, and adaptively changing this slack space based on the characteristic of the evolving data using SVM Technique. It also proposes a better alternative approach for identifying novel class.

In this journal, we addressed the problem of mining time evolving data streams. We defined the main individuality of information streams and converse unusual types of changes that occur in streaming information. During our conversation we paying attention on non-random class definition changes called concept drift. We reviewed existing single classifier and ensemble approaches to mining data streams with concept drift. Our analysis led to the development of a new algorithm called Accuracy Diversified Ensemble, which is based on our critique of the earlier developed Several real data streams are collected and concept-drift is found out.

In this journal, we addressed a problem of supervised learning over time when the data is changing (concept drift) and label of new instances is delayed. have different drift types (sudden, gradual, incremental reoccurrences) with different speed of change. Experimental evaluation confirms improvement in classification accuracy as compared to ordinary classifier for all drift types.

V.FUTURE ENHANCEMENT

For input setting parameters like number of neighborhood k and number of most recent instances w recent, these parameters have been determined by application designer. It is better to automatically determine these parameters to preserve self-adaption. Extending our algorithm so it can add or remove classes. This is important where in some domains, there are classes that disappear by time and must be removed or vice versa. We expect such strategies could increase flexibility in adaptation to drifts in cases of small sample size. Exploring some ideas to enhance the proposed strategy to improve the results accuracy. A very high classification accuracy can be provided if we build a customized version to deal with each drift individually. Finally, concept drift problems are heterogeneous from the application perspective. We believe that the future research on adaptively to concept drift has prospects and demand to come closer to specializing in application groups.

REFERENCES

[1] Aha, D.; Kibler, D. and Albert, M.;"Instance-Based Learning Algorithms," Machine Learning, vol.6 no.1, pp.37-66, Jan 1991.

[2] Brzezinski, D.; "Mining data streams with concept drift ", Master thesis, Poznan University of Technology, 2010.

[3] Chao, S. and Wong, F.; "An Incremental Decision Tree Learning Methodology regarding Attributes in Medical Data Mining," Proceedings of International Conference on Machine Learning and Cybernetics, IEEE Computer Society, pp. 1694-1699, 2009.

[4] Cohen, L.; Avrahami, G.; Last, M. and Kandel, K.; "Info-fuzzy algorithms for mining dynamic data streams," Applied Soft Computing, vol.8 no.4, pp.1283-1294, September 2008.

[5] Crespo, F. and Weber, R.;" A methodology for dynamic data mining based on fuzzy clustering, Fuzzy Sets and Systems," pp.267-284. 2005.

[6] Delany, S. J.; Cunningham, P.; Tsymbal, A. and Coyle, L.; "A case-based technique for tracking concept drift in spam filtering," Knowledge-Based Systems, vol 18. 2005.

[7] Dries, A. and Ruckert, U.; "Adaptive concept drift detection," In SDM, pp. 233-244. 2009.

[8] Fan, W.; Huang, Y.; Wang, H.; and Yu, P. S. "Active mining of data streams," Proceedings of 4th SIAM ICDM. 2004.

[9] Harries, M.; "Splice-2 comparative evaluation: Electricity pricing," Technical report, The University of South Wales, 1999. URL http:/www.liaad.up.pt/~jgama/ales/ales_5.html (last accessed Oct 8, 2011).

[10] Katakis I.; Tsoumakas G. and Vlahavas I.; "Tracking recurring contexts using ensemble classifiers: an application to email filtering," Knowlegde and information systems, vol.22 no.3, pp.371–391. 2010.

[11] Campbell, N. A., Mahon, R. J. (1974). A multivariate study of variation on two species of rock crab of genus Leptograspus. Australian

Journal of Zoology , 1974, 22 , pp. 417-425.

[12] Campigotto , P., Passerini, A., Battiti, R. (2010). Handling concept drift in preference learning for interactive decision making. Proc. of the First International Workshop (HaCDAIS, December 2010, Vancouver, Canada), Handling Concept Drift in Adaptive Information Systems: Importance, Challenges and Solutions, pp. 29-40.

[13] Dries, A., Rückert, U. (2009). Adaptive Concept Drift Detection. Journal Statistical Analysis and Data Mining Best of SDM'09 Volume 2 Issue 5-6, December 2009, pp. 311-327.

[14] Jacques, J., Biernacki, C. (2010). Extension of model-based classification for binary data when training and test population differ. Journal of Applied Statistics, March 2010, Volume 37, Issue 5, pp. 749-766.

[15] Rüping, S. (1999). Incremental Learning with Support Vector Machines. Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD, August 1999, San Diego, USA), pp. 317-321.