A Detailed Study on Ensemble Learning Techniques on Predicting Novel Class For Outlier Data

Sukanya.K[1] Dr.N.Ranjith[2]
[1]Research Scholar [2]Head and Assistant Professor
[1,2]Department of Computer Applications
[1,2]KSG College of Arts and Science, Coimbatore, India

Abstract— In traditional data mining based learning approaches, learning of exploring and evolving data streams produces the various challenges in terms of continuous stream of data, unbounded data and high Speed data characteristics from the various data repositories. In order to handle those difficulties, many scalable and effective learning models have been employed for data classification after feature extraction and reduction process. However default classifier requires effective learning model to classify the constantly evolves data over time on data streams. In this paper, a detailed study has been carried out on ensemble learning techniques for classifying the data streams. Those analysing model cope will consider the variation of the data in term of concept drift and feature drift. Further those drift data has been analysed on basis of concept and semantic of data. Moreover features of data which evolve will be efficiently handled on ensemble classification model. Finally performance of the models has been evaluated on data evolved as error driven representativeness learning along various constrained on the classification through computation of association of the feature and its weight on adaptive sliding windows of the data streams.

Keywords: Evolving Data Streams, Outlier Data Classification, Learning Model, Drift Data

I. INTRODUCTION

Data classification is becoming current research in the data mining domain as nowadays streaming data is evolving in nature due to its inherent characteristics such as volume of the data on length and velocity of data propagation of acceleration. Many state of art approaches implements learning model for training data under streaming with infinite length on several phases. In familiar, hybrid classification model has been defined for classification of the large streaming data with several benefits such as utilization of very limited training data for each classifier and time consumption for classification is less to produce high accuracy. Hybrid model can be scaled to any size of the data streams especially to data from social networks along new concepts and sentiments. Time varying concepts is difficult on predicting the outlier data.

Classification of outlier data on concept and Sentiment drift in data streams can be employed to multi level learning model which adapts to time varying data streams to produce effective class prediction to handle the concept drift and feature drift simultaneously on increase the prediction accuracy of the model on quick adaptation. Moreover time changing concept drift and feature drift can change the learning representative, it has been tackled using ensemble model on automatically generating the dynamic changing representative of the drift data through learning model.

The rest of the paper is organized into following sections. In this Section discusses problem statement of the work and whereas section 3 discusses in detail about the review analysis of literature on existing learning model for evolving data streams has been carried out. Proposed outline of the current research model is described in the section 4 and finally conclusion of the paper is presented in the section 5.

II. PROBLEM STATEMENT

In this section, existing problem on different classification model on the time varying large data streams with drift has been identified and it is represented on terms of various evaluation metrics which is as follows

A. Major Challenges of learning models

- It consumes more time in determining a suitable decision boundary for the class.
- Existing classifier is non-trivial in terms of feature grouping based on the decision making function of the classifier
- The Classifier performance of the existing model degrades on time varying data
- Data structure of large data streams is complicated and computationally expensive is processing it with distance measures
- The Target objective function of the classifier has not been defined locally in existing model which leads to class overlapping issues.
- Ensemble learning is weighted base classifiers which is capable for static data streams and leads to multiclass problem.

III. REVIEW OF LITERATURES

In this section, existing literatures on classifying the static data streams model has been analysed against various performance measure with different data chucks. The model analysed as follows

1) Junming Shao et.al proposed a model named as Prototype-based Learning against static data Streams in order to determine the classes on dynamically changing concepts along the time varying data distributions. In addition, this technique identifies concept drifts in data streams using incorporating a familiar unsupervised learning model named as Principle Component Analysis (PCA) technique. Incorporated model determines the drift and association of the data with class structure.
2) Yu Zhang et.al proposed a model named as fast online learning algorithm for classification of distributed data streams. This model uses the multiple learning models with fixed sliding windows to mine chunks of the incoming data. On exploration of the correlations between the features on the data extracted, learning

models determines the effective classes to the feature instance. Further, it uses the optimization function to increase the prediction accuracy.

3) P.Srimani et.al proposed a model named as Instance based Regression Model to identify the dynamic patterns and data regularities in the evolving data streams. In addition, it uses the various weight function on the feature instance extracted from the static data streams. The class function uses the weighted least square estimators to minimizing the variances on the drifted data. The algorithm automatically adjusts its parameters based on the data.
4) Hongfu Liu Et.al proposed model named as Spectral Ensemble Clustering via Weighted K-means for classifying dynamic data distributions. Learning model employs the co-association matrix to determine the variance and correlation of the data to to eliminate the graph partition problem and which decreases the time and space complexities of of various dynamic learning models. It is considered as high efficient solution for dynamic data classification.
5) Xuebing Yang et.al proposed a model named as Over-Sampling Technique for Multi-Class Imbalanced Problems. This model uses on underlying skewed distribution of multiple classes to determine the lack of representative data and mixed-type data. This model handles the outlier problem effectively. Distance-based over-sampling concept included in this process to capture the covariance structure of the minority class and identify the majority classes on over-sampling strategy. Finally Resampling of the classifier has employed to balance the class distribution.

IV. Outline of the Proposed Model

On observation of various existing technique, an efficient dynamic data distribution ensemble classification model has been outlined as proposed model on inclusion on multilevel learning representatives for large volume, large velocity and time-changing data streams with high accelerations. Adaptive decision boundary has to fix class boundaries with high flexibility. Proposed model can also include error driven representativeness learning and constrained classification on association and weight determination on decision boundaries of classes.

V. Conclusion

A detailed study on ensemble learning model for dynamic data distribution to predict the novel classes to outlier data has been carried out. The efficient learning models and its approaches towards classifying dynamic data streams has been analysed in this work. Classification model with multi-class problem has been employed with adaptive sliding window on decision boundary of the classes. Further these models eliminate the concept drift and feature drift in order to identify the outlier data has been efficiently mentioned. Finally experimental analysis on the dynamic data streams using different strategies have been demonstrated the effectiveness and robustness of the approaches.

References

[1] A. Bifet, G. Holmes, B. Pfahringer, R. Gavalda and R. Kirkby,"New Ensemble Methods for Evolving Data Streams" Proceeding of 15th ACM SIGKDD International Conference of . Knowledge Discovery and Data Mining, pp. 139-148, 2009.
[2] W. Fan, "Systematic Data Selection to Mine Concept-Drifting Data Streams," Proceedings in 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 128-137, 2004.
[3] G. Hulten, L. Spencer, and P. Domingos, "Mining Time-Changing Data Streams," Proceeding in 7th ACM SIGKDD International Conference in Knowledge Discovery and Data Mining, pp. 97-106, 2001
[4] I. Katakis, G. Tsoumakas, and I. Vlahavas, "Dynamic Feature Space and Incremental Feature Selection for the Classification of Textual Data Streams," Proceeding at International Workshop on Knowledge Discovery from Data Streams, pp. 102-116, 2006.
[5] B. Wenerstrom and C. Giraud-Carrier, "Temporal Data Mining in Dynamic Feature paces," Proc. Sixth Int'l Conf. Data Mining (ICDM), pp. 1142-1145, 2006.
[6] M.M. Masud, Q. Chen, J. Gao, L. Khan, J. Han, and B.M. Thuraisingham, "Classification and Novel Class Detection of Data Streams in a Dynamic Feature Space," In Proceeding in . European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 337-352, 2010.
[7] Y. Zhang, D. Sow and D. Turaga "A fast online learning algorithm for distributed mining of BigData" ACM SIGMETRICS Performance Evaluation Review, pp: 90-93, 2014.
[8] M. J Hosseini, A. Gholipour, and H. Beigy. An ensemble of cluster-based classifiers for semi-supervised classification of non-stationary data streams. Knowledge and Information Systems, 46(3):567–597, 2016.
[9] H. Wang, W. Fan, P.S. Yu, and J. Han, "Mining Concept-Drifting Data Streams Using Ensemble Classifiers," Proceeding in 9th ACM SIGKDD International Conference in Knowledge Discovery and Data Mining, pp. 226-235, 2003.
[10] Y. Yang, X. Wu, and X. Zhu, "Combining Proactive and Reactive Predictions for Data Streams," Proceeding in 11th ACM SIGKDD International Conference in Knowledge Discovery in Data Mining, pp. 710-715, 2005.

