International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Efficient Concept Evolution Detection in Data Stream

Mahesh R. Hirde[1], Piyush K. Ingole[2]

Student, Department of CSE, G. H. Raisoni College of Engineering, G. H. Raisoni College of Engineering, Nagpur, India Nagpur, India

Professor, Department of CSE, G. H. Raisoni College of Engineering, G. H. Raisoni College of Engineering Nagpur, India Nagpur, India

Abstract: ***In current years we have seen extraordinary growth in the application of data mining. Lots of work is done on infinite length, concept drift, and concept evolution. In that case we try to solve problem of continuous and fast data stream plus drift in. Previous work had done on the detection of unseen class. In this project we enhance the unseen class detection process with the additional method. Here we used flexible decision boundary and gini coefficient for batter result. In addition we are working on the simultaneous multiple unseen class detection process.***

Keywords: Data stream, new class detection, flexible decision boundary, outlier detection, concept evolution.

1. Introduction

In the resent year there is lot of work is done on the classification. For the classification point of view it is very important to classify each instance into their accurate class. But in the data there is no certain surety that there is not any novel class into it. Means by means occurrence of the new class it is very difficult to accurately classify the instance into it appropriate class. These are the main problems in the classification task. Before that many work is done on the classification and new class detection. Many author cover the error which degraded the performance of the system. Many of the author implement ensemble approach for the improving the efficiency of the classification system, also some use the outlier detection system to make classification efficient. Still there is a scope to detect outlier by which classification can do in efficient manner [1]

.

Classification on the dynamic data is itself a difficult task. Over that data stream is having some problems. Data stream is nothing but the continuous data sequences. It is very difficult to store data and used entire in practically. Classification of data stream is suffered from problem of infinite length, concept drift, and concept evolution. Several incremental learners have been proposed to address this problem [8], [5].

Many authors solve first two problems but problem of concept evolution is still in process for efficient way. Concept evolution is nothing but the occurrences of new class in the data stream. For example consider the intrusion detection in network. If we already label some kind of attack and at some time completely new kind of attack is occur then this kind of situation is concept evolution [3]. This problem is address by very few author, we address some solution to detect and improve it.

2. New Class Detection

First to which class we called existing and to which we called new class. Here we used ensemble E classifier $\{M_1 \ldots M_E\}$. If class c is occur and if it is belong to the any classifier model in ensemble then it consider as a existing class, otherwise it is consider as a non existing class.

For the system convenient purpose we divide the data into chunks and label it. After that we use it as a training data for the classifier. A basic step in the system is as follows [17]. In the first step we divide the data set in to chunks them provide that chunk to the outlier detection module after that the outlier detection module find out the outliers in the data chunks and if any instance is not belong to the any class then it will be store in the buffer. And if it is not

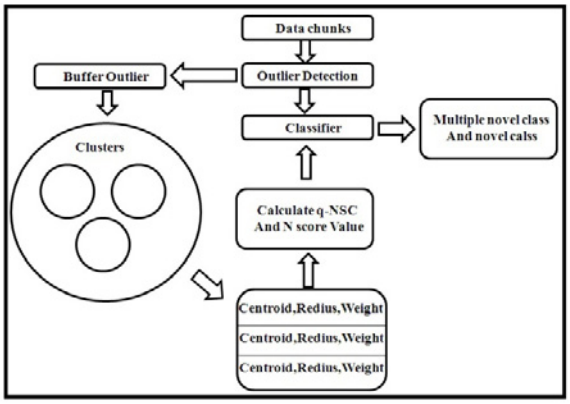

Figure 1: System architecture

Outlier then it will be forward to the classification module and classify as per the class. After sufficient outlier store in a buffer we perform clustering on it. In the result of the clustering outlier are cluster as well. Then we are making the pseudopoint of the cluster. In the case each pseudopoint contains centroid, radius and weight of the cluster. And after that we find out the q-NSC value for that pseudopoint. And after that we find out the N-score for the instance. And then we classify the it as a existing class or new class by calculating value of the N-score. Also find out the simultaneous new class detection. Formula for calculating q-NSC value.

$$\text{q-NSC(h)} = \frac{(\bar{D}\text{min}_{\text{-}}q(h) - \bar{D}\text{out}_{\text{-}}q(h))}{(\max(\bar{D}\text{min}(h), \bar{D}\text{out}_{\text{-}}q(h))}$$

q-NSC value find out the cohesion among the outlier instance in the same cluster and separation among the another cluster instance. The value calculated by the q-NSC is yield in -1 to +1 range. If value is positive then it indicate that x is closer to the outlier instance. Otherwise it is close to the existing class[22].

.

$$\text{N-score} = \frac{1 - \text{instance\_weight}(x)}{1 - \text{minweight}} \; q - NSC$$

minweight is the outlier weight having q-NSC value. N-score value is measure the cohesion among the two outlier and separation among the existing class. And how outlier is far from the existing class. Value of N-score is in the range of [0, 1]. As value increases major the chances of class is of new. Otherwise it is existing class [22].

3. Flexible Decision Boundary

Here, we address flexible decision boundary. Because of the static nature of the decision boundary an error rate of the system get increase. Due to this reason we are address flexible decision boundary.

We are use slack space concept for it. It is nothing but the flexible boundary space around the hypersphere. When any instance occurs in that space then we consider as a existing class. It also have threshold we call it as OUTTH. Initially OUTTH value is set to 0.7. First we check false new instance by taking any label instance x. Then it should be an outlier. In that case inst_weight(x) < OUTTH. If OUTTH – inst_weight is less then constant value ∂ then x is a marginal false new instance. In that case we need to increases the OUTTH limits for in future instance like these not to be fall outside boundary [22].

Algorithm: Adjust-threshold(x, OUTTH)
Input: x: most recent labeled instance
OUTTH: current outlier threshold
Output: OUTTH: new outlier threshold

step1: if false-novel(x) && OUTTH - inst_weight(x) < _ **then**
step 2: OUTTH _ ¼ _ //increase slack space
step 3: else if false-existing(x) && inst_weight(x) - OUTTH < _**then**
step 4: OUTTH þ ¼ _ //decrease slack space
step 5: **end if**

4. Data Set Used

The Forest Cover type dataset is the classification problem. There are Seven Classes in this data set. The actual Forest Cover type for a given observation is 30 x 30meter cell was determined from US Forest Service (Region2 Resource Information System (RIS) data. Independent variables were derived from data originally obtained from US Geological Survey (USGS) and USFS data. Data is in raw form and contains binary columns of data for qualitative independent variables. It consists of 500000 of instances 12 attributes.

5. Result

Here, we show result of our system with other systems.

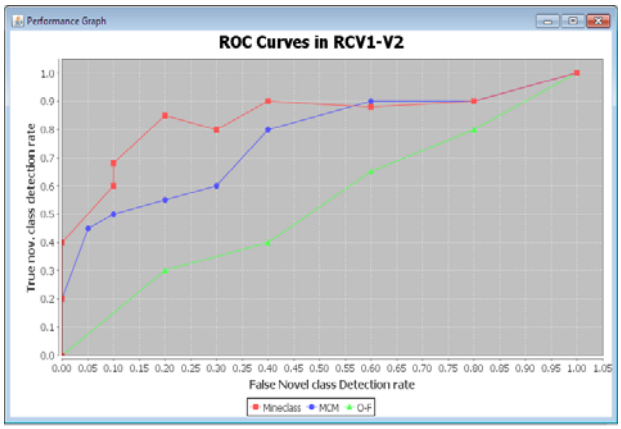

Figure 2: Result of our system

y-axis define true new class detection and x-axis shows false new class detection. In blue line is result of our approach. In fig.3 we show simultaneous new class detection. It shows class type 2, class type1, and class type 7 simultaneously.

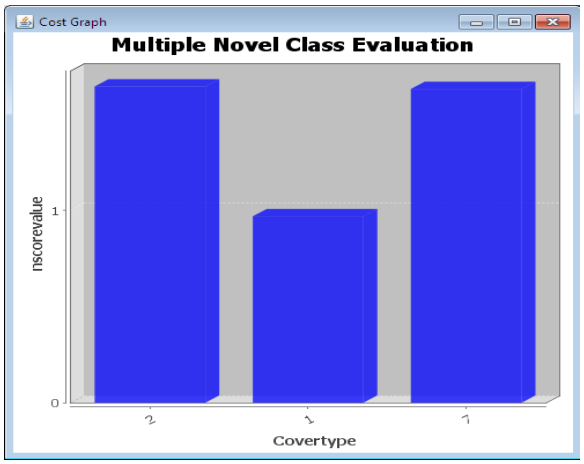

Figure 3: Multiple class detection graph

6. Conclusion

We propose a enhance concept evolution technique for data streams that addresses three major challenges, namely, infinite length, concept-drift, concept-evolution.. The existing novel class detection techniques for data streams suffer from high false alarm rate and false detection rates in many scenarios. We identify 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, we propose an improved technique for outlier detection by defining a slack space outside the decision boundary of each classification model, and adaptively changing this slack space based on the characteristic of the evolving data. We also propose a better alternative approach for identifying novel class instances using discrete Gini Coefficient, and theoretically establish its usefulness. Finally, we propose detection of simultaneous multiple novel classes. We apply our technique on data streams that experience concept-evolution and achieve much better performance than existing techniques.

7. Future Scope

An interesting and relevant question here is what will happen if one class split into several classes. If after split, they occupy the same feature space, meaning, the feature space they were covering before split is the same as the union of the feature spaces covered after split, none of the new classes will be detected as novel, because our novel class detection technique detects a class as novel only if it is found in the previously unused (unoccupied) feature spaces. However, if part of one or both of the new classes occupies a new feature space, then those parts will be detected as novel. An interesting future work would be to identify this special case more precisely to distinguish from the actual arrival of a novel class.

References

[1] C.C. Aggarwal, "On Classification and Segmentation of Massive Audio Data Streams," Knowledge and Information System, vol. 20, pp. 137-156, July 2009.
[2] C.C. Aggarwal, J. Han, J. Wang, and P.S. Yu, "A Framework for On-Demand Classification of Evolving Data Streams," IEEE Trans. Knowledge and Data Eng., vol. 18, no. 5, pp. 577-589, May 2006.
[3] Bifet, G. Holmes, B. Pfahringer, R. Kirkby, and R. Gavalda`, "New Ensemble Methods for Evolving Data Streams," Proc. ACM SIGKDD 15th Int'l Conf. Knowledge Discovery and Data Mining, pp. 139-148, 2009.
[4] S. Chen, H. Wang, S. Zhou, and P. Yu, "Stop Chasing Trends: Discovering High Order Models in Evolving Data," Proc. IEEE 24th Int'l Conf. Data Eng. (ICDE), pp. 923-932, 2008.
[5] W. Fan, "Systematic Data Selection to Mine Concept-Drifting Data Streams," Proc. ACM SIGKDD 10th Int'l Conf. Knowledge Discovery and Data Mining, pp. 128-137, 2004.
[6] J. Gao, W. Fan, and J. Han, "On Appropriate Assumptions to Mine Data Streams," Proc. IEEE Seventh Int'l Conf. Data Mining (ICDM), pp. 143-152, 2007.
[7] S. Hashemi, Y. Yang, Z. Mirzamomen, and M. Kangavari, "Adapted One-versus-All Decision Trees for Data Stream Classification," IEEE Trans. Knowledge and Data Eng., vol. 21, no. 5, pp. 624-637, May 2009.
[8] G. Hulten, L. Spencer, and P. Domingos, "Mining Time-Changing Data Streams," Proc. ACM SIGKDD Seventh Int'l Conf. Knowledge Discovery and Data Mining, pp. 97-106, 2001.
[9] Katakis, G. Tsoumakas, and I. Vlahavas, "Dynamic Feature Space and Incremental Feature Selection for the Classification of Textual Data Streams," Proc. Int'l Workshop Knowledge Discovery from Data Streams (ECML/PKDD), pp. 102-116, 2006.
[10] Katakis, G. Tsoumakas, and I. Vlahavas, "Tracking Recurring Contexts Using Ensemble Classifiers: An Application to Email Filtering," Knowledge and Information Systems, vol. 22, pp. 371-391, 2010.
[11] Kolter and M. Maloof, "Using Additive Expert Ensembles to Cope with Concept Drift," Proc. 22nd Int'l Conf. Machine Learning (ICML), pp. 449-456, 2005.
[12] D.D. Lewis, Y. Yang, T. Rose, and F. Li, "Rcv1: A New Benchmark Collection for Text Categorization Research," J. Machine Learning Research, vol. 5, pp. 361-397, 2004.
[13] X. Li, P.S. Yu, B. Liu, and S.-K. Ng, "Positive Unlabeled Learning for Data Stream Classification," Proc. Ninth SIAM Int'l Conf. Data Mining (SDM), pp. 257-268, 2009.
[14] 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," Proc. European Conf. Machine Learning and Knowledge Discovery in Databases (ECML PKDD), pp. 337-352, 2010.
[15] M.M. Masud, Q. Chen, L. Khan, C. Aggarwal, J. Gao, J. Han, and B.M. Thuraisingham, "Addressing Concept-Evolution in Concept- Drifting Data Streams," Proc. IEEE Int'l Conf. Data Mining (ICDM), pp. 929-934, 2010.
[16] M.M. Masud, J. Gao, L. Khan, J. Han, and B.M. Thuraisingham, "A Practical Approach to Classify Evolving Data Streams: Training with Limited Amount of Labeled Data," Proc. IEEE Eighth Int'l Conf. Data Mining (ICDM), pp. 929-934, 2008.
[17] M.M. Masud, J. Gao, L. Khan, J. Han, and B.M. Thuraisingham, "Integrating Novel Class Detection with Classification for Concept-Drifting Data Streams," Proc. European Conf. Machine Learning and Knowledge Discovery in Databases (ECML PKDD), pp. 79-94, 2009.
[18] M.M. Masud, J. Gao, L. Khan, J. Han, and B.M. Thuraisingham, "Classification and Novel Class Detection in Concept-Drifting Data Streams under Time Constraints," IEEE Trans. Knowledge and Data Eng., vol. 23, no. 6, pp. 859-874, June 2011.
[19] E.J. Spinosa, A.P. de Leon F. de Carvalho, and J. Gama, "Cluster- Based Novel Concept Detection in Data Streams Applied to Intrusion Detection in Computer Networks," Proc. ACM Symp. Applied Computing (SAC), pp. 976-980, 2008.
[20] H. Wang, W. Fan, P.S. Yu, and J. Han, "Mining Concept-Drifting Data Streams Using Ensemble Classifiers," Proc. ACM SIGKDD Ninth Int'l Conf. Knowledge Discovery and Data Mining, pp. 226-235, 2003.
[21] P. Wang, H. Wang, X. Wu, W. Wang, and B. Shi, "A Low-Granularity Classifier for Data Streams with Concept Drifts and Biased Class Distribution," IEEE Trans. Knowledge and Data Eng., vol. 19, no. 9, pp. 1202-1213, Sept. 2007.
[22] Mohammad M. Masud,, "Classification and Adaptive Novel Class Detection of Feature-Evolving Data Streams" IEEE transactions on knowledge and data engineering, vol. 25, no. 7, july 2013.

Author Profile

Mahesh R. Hirde has received Bachelor of Engineering degree in Computer science and engineering from SGB Amravati university, Maharashtra, India. in year 2011. Currently pursuing M.Tech. (Final Year) in Computer Science and Engineering from G. H. Raisoni College of Engineering Nagpur, Maharashtra, India. His research area of interests in Data mining. At present he is engaged in "enhancement of concept evolution in data stream" under the guidance of Assit. Prof. Piyush K. Ingole.