IJCAT International Journal of Computing and Technology, Volume 1, Issue 5, June 2014
ISSN : 2348 - 6090
www.IJCAT.org

Enhancing Novel Class Detection of Concept Evolving Data Stream

[1] Sakhi Gupta, [2] Brajesh Patel

[1] Dept. Computer Science, Shri Ram Institute, Of Technology- RGPV Bhopal, Jabalpur, Madhya Pradesh, India.

[2] Dept. Computer Science, Shri Ram Institute, Of Technology- RGPV Bhopal, Jabalpur, Madhya Pradesh, India.

Abstract - There is a lot of work is done for handling the challenges in data stream mining. Because of some practical aspects it is more challenging job. Main problem of data stream is its temporal behavior. Likewise, infinite length and concept drift are also two aspects. One important term is concept evolution. It is occur when new classes are invoking in data stream. As we know the exiting challenges of infinite length and concept drift, we address concept evolution detection in this paper. In this, enhance approach is used for detection of unseen classes in data stream using adaptive outlier detection, discrete Gini coefficient and multiple unseen classes detection.

Keywords **- Data stream, outlier, concept-evolution, unseen class, Novel Class.**

1. Introduction

Data stream is nothing but the high speed, evolving and uncertain sequence of continuous arriving data items. Data mining has recently growing field of multidisciplinary research. Databases, artificial intelligence, machine learning, automated scientific discovery, statistics, data visualization, high performance computing, decision science, etc. are combine research area of it.

In today's information society, computer users are used to gathering and sharing data anytime and anywhere. This concerns applications such as social networks, banking, telecommunication, health care, research, and entertainment, among others. As a result, a huge amount of data related to all human activity is gathered for storage and processing purposes. These data sets may contain interesting and useful knowledge represented by hidden patterns, but due to the volume of the gathered data it is very difficult to manually extract that knowledge[. Mining knowledge from the data which is generated by the continuous data generator or some real-time system is become a challenge in these day. We have large amount of data which is use in different areas. But due to their large structure it is very difficult to extract knowledge from it.

In the stream data, two characteristic is most important to extract the knowledge. First is infinite length, we all known that data stream is fast and continuous event that's why it is infinite in nature. In addition concept drift is the concept in which drift occur when the concept of stream is change over the time. Data stream also have concept evolution characteristic which means occurrences of unseen classes evolve in data. Example is intrusion detection in network traffic stream. If we consider each attack in network traffic as a class label then concept evolution occur at every new attack in network traffic. There are numerous techniques for handling concept drift but very limited focus on concept evolution and in limited way.

In [10] author, address novel and efficient technique for detection of novel class in presence of concept drift. Ensemble model method is used for classify an unlabeled data. It also use for detection of novel class. Author follows some basic steps for detection of novel class. First, during training decision boundary is built. Second, any test point occur outside the decision boundary declare as an outlier. Finally, analysis of enough cohesion among the outliers and separation processes of outlier from exiting class instance, but there be short of feature evolution problem. The problem of feature evolution and concept evolution both covered in [8]. Still there is scope for improve regarding to false alarm rate and multiple novel class detection in both [10] and [8]. This paper aim to address superior technique for outlier detection and detection of novel class, also reduce above drawback. In addition, paper is in fever of distinguish among two or more novel classes.

In this paper we addressing outlier detection technique and make it more accurate by adding slack space outside the decision boundary. Here, slack space is controlled by threshold. And Because of adapting nature of slack space it is helpful to reduce missed unseen classes and risk of

false alarm. Here unseen classes' detection approach is probabilistic using discrete Gini coefficient. Also graph based approach for detection occurrence of more than one unseen classes simultaneously.

2. Related Work

There is lot of work done on classification process to detect concept drift efficiently [1] [2] [3] [4] [5] [6] [7] [13] [11] [12] [13]. All of these worked on the infinite length and concept drift problem. From the above, every paper used some short of incremental approach to overcome this problem. [6] used a decision tree and [1] used micro clusters both are follow incremental update with incoming data. Paper [1] and [6] used single model incremental approach. They maintained data dynamically. There is one other incremental approach is hybrid batch incremental approach. In which batch learning technique is used to build a model.

In it there is a replacement policy for older model by newer when older one is become obsolete [3][4][11][7][12]. Hybrid approach required more simple operation to update the model than single model approach. Our approach is not only address concept drift and infinite length problem but also concept evolution. In [19] author used cluster based technique to detect unseen classes in data stream. They build normal model with the help of clustering, defined as hypersphere. If any cluster build outer the hypersphere, then it declared as a novel class. But there is drawback of it, we cannot directly applicable it to multiclass data stream classification. Hence it is one class classifier.

In [8] author address the feature evolution problem but in this approach rate of false positive is high (means false novel class detection rate) and missed novel class detection rates is also high. This is all about the decision boundary is rigid in nature. For eliminate this error rate we need flexible decision boundary. In addition, multiple novel class detection also not supported by it, means at same time if more than one novel class is appears then it cannot distinguish among them. In this, we addressing above issue by proposing flexible decision boundary and dynamic adaptation of that boundary. In addition solution for the multiple novel class detection.

3. Concept Evaluation and Detection

Here, we addressing enhancement of concept evolution. It can be done by three ways first, use of threshold for adaptive nature of outlier detection. Second, use of Gini coefficient for unseen class detection and finally simultaneous multiple novel class detection.

4. Overview

Here, E classification model are used for ensemble classifier in which our stream classify, M={M1……ME}. If any class B is appeared and if none of Mi is train by B then it is declared as exiting class, otherwise not. Divide the data stream into equal data chunks. kNN based classifier is use and train it with each label chunk. Semi-supervise K means are use for clustering. With the help of k means K cluster are built. After that summary of all clusters are saved as a Pseudopoint. Cluster summaries are contained centroid, weight (number of instances in cluster) and radius (distance from centroid to farthest instance of cluster). After creating summary row data point are discarded. Classification model is a constitution of these summaries (Pseudopoint). If new class is appeared then older one is replaced by newer one (usually highest error classifier). Further, Pseudopoint is corresponding to hypersphere having center is centroid of Pseudopoint and radius is also radius of Pseudopoint. Decision boundary for classifier is nothing but the constitute union of hypersphere. If any instance y is fall inside the decision boundary then it is exiting class otherwise it is F-outlier. Buffer is container of all F-outlier. When buffer contain sufficient number of outlier then unseen class detection procedure is invoke. This is for the checking whether outlier really not belong to exiting class. If it is not exiting class then F-outlier is tagged as a unseen class.

The unseen class detection is nothing but "A data point should be closer to the data points of its own class (*cohesion*) and farther apart from the data points of other classes (*separation*)" [4]. This procedure is based on measurement of cohesion among F-outlier in buffer and separation F-outlier from class which having exiting instance. And computing a unified measure of cohesion and separation also called q-Neighborhood Silhouette Coefficient or q-NSC.

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

Mean distance between two outlier is denoted by $\bar{D}_{c_{out},q}(x)$. And mean distance of x from exiting class is denoted by $\bar{D}_{c_{min},q}(x)$. Above expression taken from [14]

4.1 Outlier Detection

We know that if any instance is appeared outside the Pseudopoint in the ensemble of model then we call it as F-

outlier. If any instances close but outside the hypersphere of Pseudopoint still it is an outlier. This is due to noise; resulting false alarm rate is high. For this adaptive outlier detection is more efficient. In this case slack space is providing around each hypersphere. Slack space is in controlled by threshold, if any instance fall under this slack space then it is consider as an existing class instance. This threshold is adjustable.

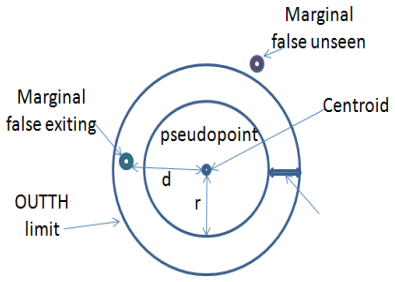

Figure1. Slack space representation around hypersphere[14].

Boundary of slack space is nothing but OUTTH. As figure 1 shows Pseudopoint having centroid in center with radius 'r'. Consider distance from centroid to instance if'd', than weight instance became weight(x) = $e^{r-d.}$ If r>=d then instance inside, otherwise it is outside Pseudopoint. If r>=d then weight(x) >=1, otherwise weight(x)<1. Exponential term is use specially for producing output in range (0,1).OUTTH value is in rang of (0,1) as shown in figure.1. if value of OUTTH is less than or equal to weight(x) then x consider as exiting instance. Otherwise it is outlier. OUTTH is adjustable. The initial value of OUTTH is 0.7, instance x is use for examination for adjusting the OUTTH. If instance x is false unseen instance then it must have been outlier, in this case weight(x) < OUTTH. Here there is small constant Ɛ, if OUTTH – weight(x) is less than Ɛ then instance x is consider as a false unseen instance. If instance is marginal false novel then we have to increase slack space. And future instance like these not to be fall outside the boundary. For increase slack space we have to decrease OUTTH value by small constant Ɛ.

4.2 Use of Gini Coefficient for New Class Detection

Appearance of outlier is due to noise or concept drift or due to concept evolution. For finding outlier occurrences due to concept evolution Gini coefficient is use. After finding outlier by OUTTH, find out q-NSC value for each outlier, if calculate q-NSC value is negative then don't consider it. It means that it is belong to exiting class. Calculate q-NSC for remaining in range [0,1]. Now we find the Nscore by computing compound measure as follows:

$$Nscore(x) = \frac{1-weight(x)}{1-minweight} q\text{-}NSC(x),$$

Above expression taken from [14]. Minweight is minimum weight in all outlier. It measured distance between outlier and nearest exiting class by higher value – greater distance, Also cohesion among outlier. In addition, separation outlier to exiting instance is calculating by it.

$$G(s) = \frac{1}{n}\left(n+1-2\left(\frac{\sum_{i=1}^{n}(n+1-i)y_i}{\sum_{i=1}^{n} y_i}\right)\right)$$

Above expression taken from [14]. For measuring statistic dispersion, Gini coefficient is usually used. It's value is in rang [0,1]. As dispersion high, value of Gini coefficient is also high. 'n' is equal intervals of Nscore value. And y_i be values of CDF (cumulative distribution function).

4.3 Multiple Unseen Class Detection

It may be happen to more than one unseen classes appeared simultaneously. It is really changing job to determining is their more than one class. Cohesion and separation are the properties by which we can determine this type of problem. For this, if there is more than one unseen classes, then separation between unseen classes is higher than cohesion among same class instance. For this work here use some unseen class instances and make Pseudopoint by using K means.

5. Conclusion

Here, several enhancements are show for existing classification and unseen class detection. problem of outlier detection, unseen class detection are stated in paper, also detection of multiple unseen classes is discuses. Problem of outlier detection is try to improved by slack space. Also address alternative solution for unseen classes detection.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] J. 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] Y. Yang, X. Wu, and X. Zhu, "Combining Proactive and Reactive Predictions for Data Streams," Proc. ACM SIGKDD 11th Int'l Conf. Knowledge Discovery in Data Mining, pp. 710-715, 2005.

[13] P. Zhang, X. Zhu, and L. Guo, "Mining Data Streams with Labeled and Unlabeled Training Examples," Proc. IEEE Ninth Int'l Conf. Data Mining (ICDM), pp. 627-636, 2009.

[14] 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.