SVM based Feature Extraction for Novel Class Detection from Streaming Data

Arati Kale
P.G. Student, Computer Department, , JSPM's
JSCOE, Hadapasar, Pune, India

M.D. Ingle
Associate Professor, Computer Department,
JSPM's JSCOE, Hadapasar, Pune, India

ABSTRACT

World have huge amount of data. Data stream classification contain several problem such as Infinite Length , Concept Drift ,Concept Evolution and Feature Evolution. Infinite Length means data available in huge amount and it is difficult to store all historical data for training. Concept Evolution occurs as a result of new classes evolving in stream. Concept Drift occurs as a result of changes in underlying concepts. Feature Evolution occurs as new feature arises.

Traditional data stream classifier only addresses Infinite Length and Concept Drift. In this paper we propose ensemble classification framework where each classifier is equipped with novel class detector to address Concept Drift and Concept Evolution. Also increases accuracy of novel class detection techniques by using SVM based polynomial kernel.

Keywords:

Support vector machine, feature extraction, Novel class detection, and Polynomial kernel.

1. INTRODUCTION

Data Stream mining can be consider as sub-branch of machine learning , knowledge discovery and data mining approach. Data stream mining have many challenges such as Infinite length, Concept Evolution, Concept drift and Feature evolution.

Infinite Length:
Data stream is continuous and fast growing so it is infinite in length but practically it is impossible to store all historical data as a training data.

Concept Evolution:
In data stream new concept introduced, the concept evolution challenge is invoked.

Example-In network, different viruses are putting in different class label, except those unknown label class is detected in network is called as concept evolution.

Feature evolution:
When new feature introduced in data stream then old features migrated with new one.

In many techniques novel class detection problem has been addressed in presence of Infinite length and Concept drift. These techniques classify unlabelled data by ensemble model for novel class detection .Process of novel class detection consist three steps. In first step decision boundary built in training process, second step test the points which are falling outside of decision boundary are declared as outliers and in third step outliers are analyzed to find it is present in existing class or separate from existing class instance. But these techniques did not address challenge of feature evolution problem. This paper represents more efficient techniques for outlier detection and novel class detection to reduce false alarm rate and improve accuracy of classification. The main objective of paper is to detect outlier class by using Support Vector Machine based feature extraction. It is named as SVMiner which is mainly proposed for detecting outliers in class. It finds best outlier in class and then applied to detect novel class of data stream. Using Svm based polynomial kernel feature extraction we can improve outlier detection technique and improve efficiency of finding outlier by K-nn boundary and SVM boundary at a time.

The rest of paper is organized as follows:

The section 2.Related work gives related technologies, their advantages and disadvantages for novel class detection. The section3.Programmers Design focuses on proposed work of paper and explains how novel class detection is done through SVM based feature extraction using polynomial kernel. The section4.Conclusion describes SVM for feature extracting through polynomial kernel for streaming data for novel class detection. The section6.References contains references of paper that are referred for developing new technique for novel class detection.

2. RELATED WORK

Khan et al. [1] Data streams are continues data. Examples Twitter feedback, call center records, network traffic. In this data stream main challenge is classification of the data. Data stream classification has faced problems such as infinite length, concept drift, concept evolution and feature evolution.

Farid et al. [2] propose new approach for detecting novel class in data stream mining using decision tree. It identifies new instance belongs to novel class and create decision tree from training data which is updated so tree represent current concept in data stream.

Spinosa et al. [3] apply cluster based technique to detect novel class but they do not consider feature evolution challenge for classification of novel class.

M.M.Masuad et al. [4] addresses feature evolution and concept evolution problem to find novel class from stream data but it has drawback of finding false alarm rate. Penzhang et al. [5] propose framework for prediction model which contain labeled and unlabeled data. For prediction model divide data into different categories and propose relational K-means based transfer semi-supervised SVM learning framework.

Masud M. M. et al. [6] [7] proposes data stream classification model which combines novel class detection mechanism into traditional classifiers under time constrain factor. It build decision boundary during training phase. If any test instance falls outside decision boundary then it is consider as outlier.

Masud M.M. et al. [8] determines concept evolution problem in addition with infinite length and concept drift. Firstly it proposes adaptive threshold for outlier detection and then propose probabilistic approach for detecting novel class using Gini-coefficient.

Mohammad M. Masud et al. address all four challenges of data stream classification by proposing ensemble classification framework. To address feature evolution they propose feature set homogenization technique.

3. PROGRAMMERS DESIGN

3.1 Background:

3.1.1. K-NN Classifier:

K-Nearest Neighbors classifier is used to classify unknown data present in data stream. For classification it is used similarity function or distance function. It do classification using instance based classifier means it locates nearest neighbor in instance space and give name to test data with nearest neighbor class name. It consider data stream as vector having multiple dimensional feature space. It has two phases one training phase and other classification phase. In training phase it stores feature vector and class label of sample training data. In classification phase, unlabelled instance is classified by giving label which is most frequent in training sample nearest to test point. Using K-NN we are implemented decision boundary for data stream.

3.1.2. SUPPORT VECTOR MACHINE:

SVM is one of the machines learning technique used for feature extraction. In SVM, feature extraction is done through polynomial kernel. Polynomial kernel finds different- different relationship between feature set. It considers dataset as vectors and extracts features. After that it finds out relationship to identify outlier boundaries. It is heuristic in nature so to find relationship performs number of permutations and combination.SVM is known as NP-hard problem. It improves accuracy of finding outlier by using polynomial kernel.

3.1.3 q NSC (q-Neighborhood silhouette coefficient)

It is used for finding novel class detection in data stream. It assigns value to test instance as 0and 1 based on cohesion and separation function. If test data is nearer to ensemble boundary then it assigns 0 value and if test instance is far away from ensemble boundary then it assigns 1 value. When sufficient amount of outliers are collected then it detect novel class from collected outliers and put it into ensemble for future classification.

3.2 MATHEMATICAL MODEL:

Let, P= {d1, d2, d3 , ……………} be continuous data set stream for which we have to detect novel class.

The polynomial kernel is defined as:

$$k\,(x,y) = (x^T.y + c)\hat{}d \qquad (1)$$

Where x and y are vectors in input space i.e. vectors of features computed from training or test samples.

c>=0 is a constant trading off the influence of higher order versus lower order terms in polynomial.

C=0, then kernel is called homogeneous.

As a kernel k corresponds an inner product in a feature space based on some mapping Φ

$$K(x,y) = < \Phi(x), \Phi(y)>$$

For d=2,

$$K(x,y) = (\textstyle\sum_{i=1}^{n} xiyi + c)\hat{}2 \qquad (2)$$

$$K(x,y)=\textstyle\sum_{i=1}^{n} x_i^2\, y_i^2 + \sum_{i=2}^{n} \sum_{j=1}^{i-1} \sqrt{2}\, x_i y_i \sqrt{2} x_i y_i$$

$$+ \textstyle\sum_{i=1}^{n} \sqrt{2cx_i} \qquad \sqrt{2cy_i} + \; c^2 \qquad (3)$$

Algorithm SVM_Miner (P)

Input: p: continuous data stream

Output: Detection of novel class instance

Step1: on data set build k-NN classifier and perform clustering.

Step2: Build a feature set of centroid and cluster boundaries. Use these boundaries for detecting outliers.

Step3: Use the clusters generated in step 1 for feature extraction using SVM polynomial kernel in equation [1], [2] and [3].

Step4: Generate boundaries for classification using step 3.

Step5: Thus we have k-NN boundaries as well as SVM boundaries.

Step6: By using two boundary conditions we can improve outlier detection.

Step7: Build ensemble of classifier on continuous stream data.

Step8: Classify the continuous stream using both boundaries. If at least one of them terms data as outlier then test instance termed as outlier.

Step9: Use q-nsc for detecting novel class from outliers.

3.3 SYSTEM BLOCK DIAGRAM:

3.3.1. Using K-NN Classifier:

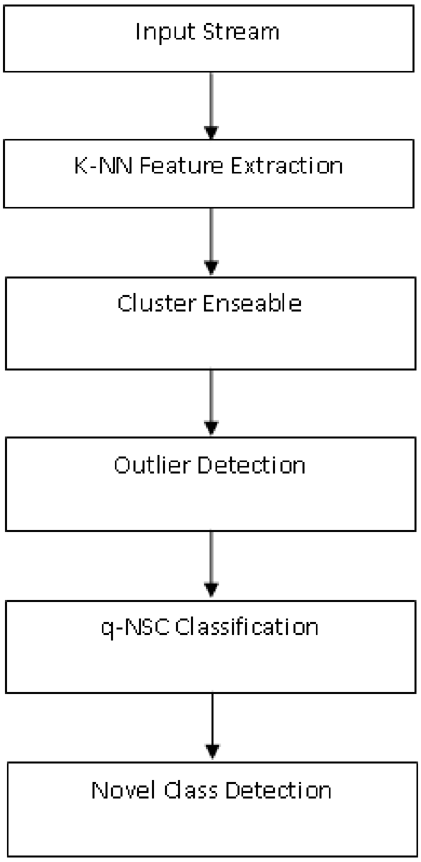

Figure 3.1 K-Nn Classifier

3.3.2. Using SVM Polynomial Kernal:

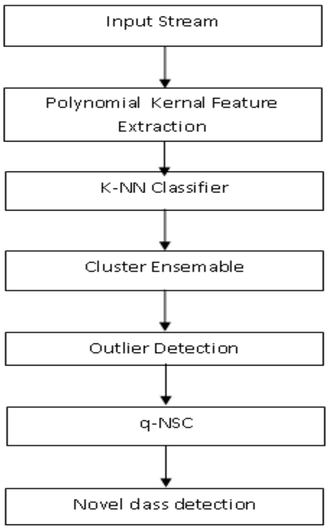

Figure 3.2 SVM Polynomial Kernal

4. CONCLUSIONS

We propose novel class detection technique for addressing challenges of data stream classification such as infinite length, concept drift, concept evolution, feature evolution. The present novel class detection technique does not address feature evolution problem and false alarm rate. We first find out cluster boundaries for outlier detection using k-NN classifier. In next step extract features using SVM polynomial kernel to build SVM boundaries. By using these two boundary conditions we can improve outlier detection. Finally used q-NSC for detecting novel class from outlier.

5. ACKNOWLEDGEMENT

The satisfaction achieves after successful completion of any task would be incomplete without mentioning those people who are responsible to complete that task. I'm grateful to many persons who contributed to the completion of this research. Particular I wish to thank my honorable guide and PG coordinator Prof. M.D.Ingle, H.O.D. Prof. S.M.Shinde and Principal Dr. M.G. Jadhav of Jayawantrao Sawant College Of Engg. Pune, for providing comments, information, and review of this report. And at last thanks to the most superior person of the universe, the almighty God, who always showered a blessing upon me.

6. REFERENCES

[1] Khan, L.Tools, “Data Stream Mining: Challenges and Techniques”, Artificial Intelligence (ICTAI), 2010 22nd IEEE International Conference on22010.

[2] Farid, D.M.; Rahman C.M. ,” Novel class detection in concept-drifting data stream mining employing decision tree”, Electrical & Computer Engineering (ICECE), 2012 7th International Conference on2012.

[3] 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.

[4] 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.

[5] Peng Zhang; Xingquan Zhu; Li Guo, “Mining Data Streams with Labeled and Unlabeled Training Examples”, Data Mining, 2009. ICDM '09. Ninth IEEE International Conference on2009.

[6] 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.

[7] 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.

[8] 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.

[9] Mohammad M. Masud, Qing Chen, Latifur Khan, Charu C. Aggarwal, “Classification and Adaptive Novel Class Detection of Feature-Evolving Data Streams” IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 7, JULY 2013.