International Refereed Journal of Engineering and Science (IRJES)
ISSN (Online) 2319-183X, (Print) 2319-1821
Volume 4, Issue 2 (February 2015), PP.01-07

Novel Class Detection Using RBF SVM Kernel from Feature Evolving Data Streams

Arati Kale[1], Prof. M.D.Ingle[2]
[1]*P.G. Student, Computer Department, JSPM's JSCOE, Hadapasar, Pune, India*
[2]*Associate Professor, Computer Department, JSPM's JSCOE, Hadapasar, Pune, India*

Abstract:- In the data mining field the classification of data stream creates many problems. The challenges faces in the data stream are infinite length, concept drift, concept evaluation and feature evolution. Most of the existing system focuses on the only first two challenges. We propose a framework in which each classifier is prepared with the novel class detector for addressing the two challenges concept drift and concept evaluation and for addressing the feature evolution feature set homogeneous technique is proposed. We improved the novel class detection module by building it more adaptive to evolving the stream. SVM based feature extraction for RBF kernel method is also proposed for detecting the novel class from the steaming data. By using the concept of permutation and combination RBF kernel extracts the features and find out the relation between them. This improves the novel class detect technique and provide more accuracy for classifying the data

Keywords:- Support vector machine, feature extraction, feature evolution, Novel class detection, and RBF Kernel.

I. INTRODUCTION

Now a day's tremendous amount of data is available and from this data knowledge extraction is very difficult task. To extract knowledge from this data, one technique is classification. But data stream classification possesses different challenges that are infinite length, concept drift, concept evolution and feature evolution. Infinite length problem is due to tremendous growth of data in length and practically it is very difficult to store all data for training purpose. Concept drift issue arises due to concepts are changes rapidly. Concept evolution is nothing but occurrence of new concept. Feature evolution is occur due to new features are introduces in day by day life.

Many techniques are available but some identified only infinite length and concept drift challenge. Few tackle infinite length and concept drift challenge along with concept evolution but very few handle feature evolution problem of data stream classification. Novel class detection contains two main modules one is outlier detection and novel class detection. To find novel classes' first decision boundary is form and test that which instances are within boundary or which instances are out off the boundary. Those instances fall outside boundary is declared as outlier but this kind of technique has more false alarm rate .This paper represent more effective technique to tackle all classification challenges and to reduce false alarm rate. Our approach uses DBSCAN to detect outlier in outlier detection module using density based function and use RBF kernel of SVM classifier for detecting multi novel classes in novel class detection module. SVM maximizes margin between boundaries due to this it is easy to identify which instances are within a boundary or which instances are outside the boundary .SVM provides more accuracy and improve novel class detection technique. It also reduces false alarm rate of outliers.

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 section 3.Background gives detail of background concept used for novel class detection. The section 4.Programmers Design focuses on proposed work of paper and explains how novel class detection is done through SVM based feature extraction using RBF kernel. The section 5.Results describes experimental result on data set. The section 6.Conclusion describes SVM for feature extracting through RBF kernel for streaming data for novel class detection and reducing high alarm rate of outlier. References contain references of paper that are referred for developing new technique for novel class detection.

II. RELATED WORK

M.M.Masuad et al. [1] Classification and Adaptive Novel Class Detection of Feature Evolving Data Streams. They propose an ensemble classification framework, where each classifier is equipped with a novel class detector, to address concept-drift and concept-evolution. To address feature evolution, they propose a feature set homogenization technique.

M.M. Masud, J. Gao, L. Khan, J. Han, and B.M. Thuraisingham [2] Integrating Novel Class Detection with Classification for Concept-Drifting Data Streams. propose a novel and efficient technique that can automatically detect the emergence of a novel class in the presence of concept-drift by quantifying cohesion among unlabelled test instances, and separation of the test instances from training instances.

Charu C. Aggarwal et al. [3] A Framework for Classification and Segmentation of Massive Audio Data Streams. They discuss the details of such an online voice recognition system. For this purpose, we use our micro-clustering algorithms to design concise signatures of the target speakers. One of the surprising and insightful observations from our experiences with such a system is that while it was originally designed only for efficiency, we later discovered that it was also more accurate than the widely used GMM. This was because of the conciseness of the micro-cluster model, which made it less prone to over training. This is evidence of the fact that it is often possible to get the best of both worlds and do better than complex models both from an efficiency and accuracy perspective.

J. Kolter and M. Maloof et al. [4] Using Additive Expert Ensembles to Cope with Concept Drift. Present the additive expert ensemble algorithm AddExp, a new, general method for using any online learner for drifting concepts. We adapt techniques for analyzing expert prediction algorithms to prove mistake and loss bounds for a discrete and a continuous version of AddExp.

M.M. Masud, Q. Chen, L. Khan, C. Aggarwal, J. Gao, J. Han, and B.M. Thuraisingham [5]. Addressing Concept Evolution in Concept Drifting Data Streams, Proposed the concept evolution phenomenon is studied, and the insights are used to construct superior novel class detection techniques. First, we propose an adaptive threshold for outlier detection, which is a vital part of novel class detection. Second, we propose a probabilistic approach for novel class detection using discrete Gini Coefficient, and prove its effectiveness both theoretically and empirically. Finally, we address the issue of simultaneous multiple novel class occurrence, and provide an elegant solution to detect more than one novel class at the same time. We also consider feature evolution in text data streams, which occurs because new features (i.e., words) evolve in the stream.

A. Bifet, G. Holmes, B. Pfahringer, R. Kirkby, and R. Gavalda [6], new ensemble methods for evolving data streams, Proposes a new experimental data stream framework for studying concept drift, and two new variants of Bagging: ADWIN Bagging and Adaptive Size Hoeffding Tree (ASHT) Bagging. Using the new experimental framework, an evaluation study on synthetic and real-world datasets comprising up to ten million examples shows that the new ensemble methods perform very well compared to several known methods.

H. Wang, W. Fan, P. S. Yu, and J. Han. Mining concept drifting data streams using ensemble classifiers [7], in this paper, we propose a general framework for mining concept drifting data streams using weighted ensemble classifiers. We train an ensemble of classification models, such as C4.5, RIPPER, naive Bayesian, etc., from sequential chunks of the data stream. The classifiers in the ensemble are judiciously weighted based on their expected classification accuracy on the test data under the time-evolving environment.
S. Hashemi, Y. Yang, Z. Mirzamomen, and M. Kangavari [8], Adapted One-versus-All Decision Trees for Data Stream Classification. This paper advocates some outstanding advantages of OVA for data stream classification. First, there is low error correlation and hence high diversity among OVA’s component classifiers, which leads to high classification accuracy. Second, OVA is adept at accommodating new class labels that often appear in data streams. However, there also remain many challenges to deploy traditional OVA for classifying data streams.

W. N. Street and Y. Kim. A streaming ensemble algorithm (sea) for large scale classification [9] , The method presented in this paper takes advantage of plentiful data, building separate classifier on sequential chunks of training points. These classifiers are combined into fixed size ensemble using heuristic ensemble strategies.

Y. Yang, X. Wu, and X. Zhu [10], Combining Proactive and Reactive Predictions for Data Streams. In a proactive mode, it anticipates what the new concept will be if a future concept change takes place, and prepares prediction strategies in advance. If the anticipation turns out to be correct, a proper prediction model can be launched instantly upon the concept change. If not, it promptly resorts to a reactive mode: adapting a prediction model to the new data. A system Reactive-Proactive is proposed to implement these new ideas.

III. BACKGROUND

3.1 DBSCAN CLASSIFIER:

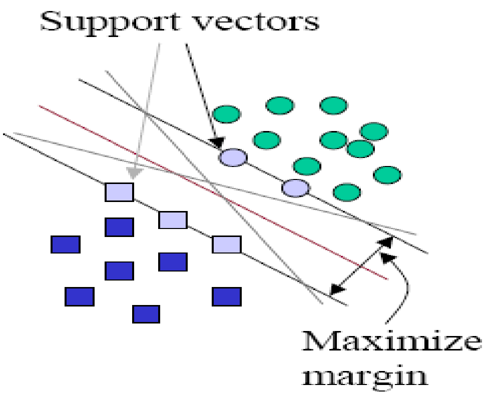
It is clustering algorithm. it takes input as € (radius) and minpts (number of instances present in cluster). Using this input it calculates core points, neighborhood points and border points referring Euclidean distance. After calculating these point density reachable and density connect is identified and core points are given as input to next chunk in data stream. DBSCAN find cluster of arbitrary size and shape so outlier detection accuracy rate is high.

3.2 K-Means CLASSIFIER:

It is partitional clustering algorithm. It takes K as input. It contains four main steps such as first select centroid randomly from data stream. Second assign each instance to cluster with neighbor centroid. Third calculate each centroid as average of instances assigned to it. Fourth step repeat previous two steps still no change occur in cluster formation. It uses Euclidean distance formula for calculating distance between two instances.It is more efficient clustering algotithm.

3.3 SUPPORT VECTOR MACHINE:

Kernel-Transforms input data to high dimensional space. SVM find decision surface that maximizes margin between data points of two classes. Support vectors are data points that lie closest to decision surface. Thus are most difficult to classify. SVM provide optimal surface for linearly separable pattern.



SVM algorithm work in four steps such as first step is choosing kernel function. Second choose value for C. Third solve quadratic programming problem. Fourth construct discriminate function from support vector.

IV. PROGRAMMERS DESIGN

4.1 MATHEMATICAL MODEL:

$$\left.\begin{aligned} \text{Epsilon} &= \frac{\sum_{i=1}^{k} \text{Avg Intra } (Ci)}{k} \\ \text{where} \quad \text{Avg Intra } (Ci) &= \frac{\sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \text{Dist}\,(Oi,Oj)}{2\times n} \end{aligned}\right\} \quad (1)$$

$$\left.\begin{aligned} \text{Min Pts} &= \text{Avg No. of Objects with a distance of Epsilon} \\ &\quad \text{from a object in cluster of smallest density} \\ \text{where Density } (Ci) &= \frac{\text{No. of Object } (Ci)}{\text{Radius } (Ci)} \end{aligned}\right\} \quad (2)$$

$$\left.\begin{aligned} W_j &= \begin{cases} 0 & if\ D_j=0 \\ \dfrac{1}{\sum_{t=1}^{h}\left[\frac{D_j}{D_t}\right]^{\frac{1}{\beta-1}}} & if\ D_j\neq 0 \end{cases} \\ where D_j &= \sum_{l=1}^{k}\sum_{i=1}^{n} u_{i,l}\, d(x_{i,j}, z_{l,j}) \end{aligned}\right\} \quad (3)$$

$$\sum_{j=1}^{m} w_j = 1 \text{ and } 0 \leq w_j \leq 1 \quad (4)$$

In the above equation k is the number of clusters, n is the total number of objects comprises all clusters. Ci represent ith cluster and Oi represent ith object. Dist () is the distance between two objects. We have taken Euclidian distance in our experiment for implementation. Based on these equation parameters of DBSCAN are updated in MinPts and epsilon updation selection.

We calculate the RBF SVM functionality with the following equation:

$$K(x_i, x_j) = \exp\left(-\frac{||xi - xj||^2}{2\sigma^2}\right)$$

Where, xi, xj are support vector and testing data point, σ be the area of influence this support vector has over the data space.

4.2 PROPOSED ALGORITHM:

Algorithm SVMRBF_miner (P)

Input: p: continuous data stream

Output: Detection of novel class instance

Step 1: On data set perform feature selection algorithm to select feature from data

Step 2: Divide data in chunks

Step 3: Perform clustering of data with DBSCAN algorithm

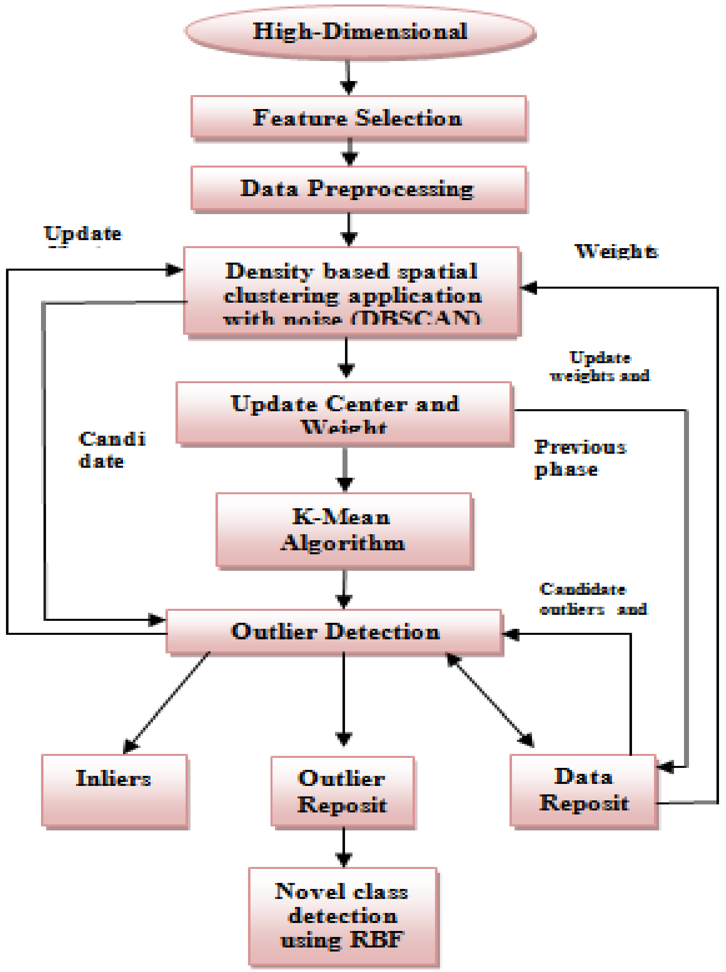
Step 4: Perform K-means on point list

Step 5: Detect outlier from data set

Step 6: When buffer size get full using NN-search technique classes has detected.

Step 7: Using RBF kernel SVM classification multiple novel classes has been detected.

4.3 SYSTEM ARCHITECTURE:



V. RESULT AND DISCUSSION

5.1 DATASET:

Forest cover data set from UCI repository (forest). Data set contains geospatial descriptions of different types of forests. It contains 7 classes, 54 attributes and around 581,000 instances.

5.2 PARAMETERS SETTINGS:

Numeric features are used.
K(number of pseudopoints per chunk) = 50,
S(chunk size) = 1,000,
L(ensemble size) = 6,
q(minimum number of f-outliers required to declare a novel class) = 50.

5.3 EVALUATION

Let Enovexist = total novel class instances misclassified as existing class, Eexistnov = total existing class instances misclassified as novel class and Eother = total existing class instances misclassified as another existing class, N = total number of instances, and V = total number of novel class instances.

The performance metrics for evaluation:

$M_{new} = (100)(Enovexist)/V$ i.e. % of novel class instances misclassified as existing class.

$F_{new} = (100)(Eexistnov) / N - V$ i.e. % of existing class instances misclassified as novel class.

ERR = (100) (Enovexist + Eexistnov + Eother) / N i.e. Overall error.

TABLE 1

Result Summary

Dataset	Method	ERR	M_{new}	F_{new}	AUC
Forest	Mine Class	3.6	8.4	1.3	0.97
	MCM	3.1	4.0	0.68	0.99
	O-F	5.9	20.6	1.1	0.74
	SVM	2.48	2.39	0.54	0.99

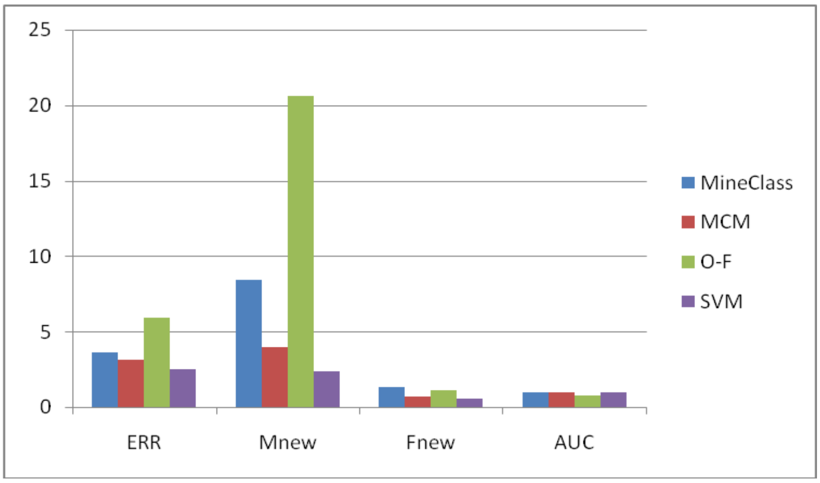
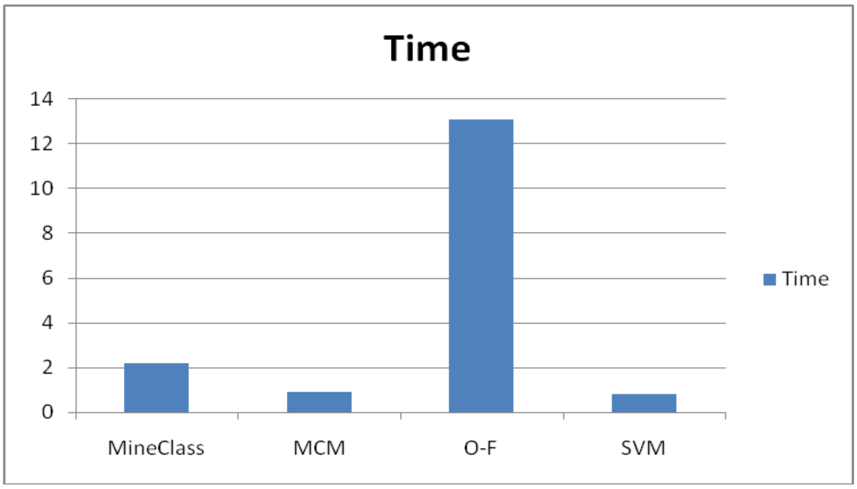


TABLE 2 Multimode class detection Result

Dataset		Occurrence		Total
Forest 581K Instance 7 Classes		**1**	**2**	
	TP	**474**	**529**	**1003**
	FP	**127**	**294**	**421**
	TN	**245**	**497**	**742**
	FN	**94**	**272**	**387**
	Prec.	**0.78**	**0.64**	**0.71**
	Recall	**0.83**	**0.66**	**0.75**

TABLE 3 Running Times (in seconds)

Dataset	**Mine Class**	**MCM**	**O-F**	**SVM**
Forest	2.2	0.9	13.1	0.8



VI. CONCLUSIONS

We propose multi novel class detection technique which addresses all data stream classification issues such as infinite length, concept drift, concept evolution and feature evolution. The some present multi novel class detection techniques are addresses all issues but they have false alarm rate i.e. wrong identification of novel classes .we first uses DBSCAN and K-means for outlier detection module and RBF of SVM is used for detecting multi novel classes present in data stream. so using these two approaches we can reduce false alarm rate and increase accuracy of outlier detection.

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 am 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. H.A. Hingoliwala 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.

REFERENCES

[1]. 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.

[2]. 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.

[3]. Charu C. Aggarwal "A Framework for Classification and Segmentation of Massive Audio Data Streams". KDD'07, August 12–15, 2007, San Jose, California, USA Copyright 2007 ACM 978-1-59593-609-7/07/0008.

[4]. J. Kolter and M. Maloof, "Using Additive Expert Ensembles to Cope with Concept Drift," Proc. 22nd Intl Conf. Machine Learning (ICML), pp. 449-456, 2005.

[5]. 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 Intl Conf. Data Mining (ICDM), pp. 929-934, 2010.

[6]. A. Bifet, G. Holmes, B. Pfahringer, R. Kirkby, and R. Gavalda', "New Ensemble Methods for Evolving Data Streams," Proc. ACM SIGKDD 15th Intl Conf. Knowledge Discovery and Data Mining, pp. 139-148, 2009.

)

[7]. H. Wang, W. Fan, P. S. Yu, and J. Han. Mining concept drifting data streams using ensemble classifiers. In SIGKDD, pages 226□235, 2003.

[8]. 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.

[9]. W. N. Street and Y. Kim. A streaming ensemble algorithm (sea) for large scale classification. In Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD), pages 377-382, 2001.

[10]. Y. Yang, X. Wu, and X. Zhu, “Combining Proactive and Reactive Predictions for Data Streams,”Proc. ACM SIGKDD 11th Intl Conf. Knowledge Discovery in Data Mining, pp. 710-715, 2005.