Improvement of Pre-processing Capacity of Support Vector Clustering using Neural Network Kernel Functionfor Stream Data Classification

Ritika Chatterjee
Department of computer science and engineering
Lakshmi Narain College of Technology
Bhopal (M.P) India

Shweta Shrivastav
Department of computer science and engineering
Lakshmi Narain College of Technology
Bhopal (M.P) India

Vineet Richhariya, Ph.D
Department of computer science and engineering
Lakshmi Narain College of Technology
Bhopal (M.P) India

ABSTRACT

Pre-processing of data before generation of pattern or classification is major steps. In the phase of pre-processing reduces the noise level of data using different technique of data mining. In current research trend support vector clustering is used for efficient data processing for noise reduction and pattern generation. Support vector clustering is new paradigm of data mining tools. It combined with supervised learning and unsupervised learning. for the success story behind support vector clustering technique is kernel function. The better selection of kernel function produces better result in terms of noise reduction and classification. In this paper proposed an improved support vector clustering method using neural network kernel function for stream data classification. The neural network function work as data optimizer and data selector in support vector clustering.

General Terms

Data mining, clustering, classification

Keywords

Stream data, Support vector clustering (SVC), neural network.

1. INTRODUCTION

Data stream classification is more challenging than classifying static data because of several unique properties of data streams. First, data streams (web clicks, credit card transactions etc) are assumed to have infinite length, which makes it unfeasible to accumulate and use all the past data for class labeling. Hence, common learning methods are not directly applicable to stream data. In the process of stream data classification infinite length is not single problem it comes along data drift, concept evaluation and feature selection over time. With the advent of advanced data streaming technologies [1], we are able to continuously collect large amounts of data in various application domains, e.g., daily punctuations of stock market, traces of dynamic processes, credit card transactions, web click stream, network traffic monitoring, position updates of moving objects in location-based services and text streams from news etc [2].Due to its potential in industry applications, data stream mining has been studied intensively in the past few years. The general approach is to first learn one or multiple classification models from the past records of the evolving data, and then use a selected model that best matches the cur-rent data to predict the new data records. All the existing data stream classification techniques assume that at each time stamp there are both large amounts of positive and negative training data available for learning. The goal of data stream classification is to learn a model from past labeled data, and classify future instances using the model. There are many challenges in data stream classification. First, data streams have infinite length, and so, it is impossible to store all the historical data for training. Therefore, old and traditional learning method for stream data classification faced numerous problems during data categorization. A classification model must adapt itself to the most recent concept in order to cope with concept-drift. Third, novel classes may appear in the stream, which we call concept-evolution. Data stream classifiers may either be single model incremental approaches, or ensemble techniques, in which the classification output is a function of the predictions of different classifiers. Ensemble techniques have been more popular than their single model counterparts because of their simpler implementation and higher efficiency. Most of these ensemble techniques use a chunk-based approach for learning which they divide the data stream into chunks, and train a model from one chunk. We refer to these approaches as “chunk-based” approaches. An ensemble of chunk-based step involves construction of cluster boundaries and cluster labeling step involves assigning the cluster labels to each data point. Solving the optimization problem and cluster labeling is time consuming in the SVC training procedure. This paper is divided into five sections. Section-I gives the introduction of stream data classification and support vector models is used to classify unlabeled data. These approaches usually keep a fixed-sized ensemble, which is continuously updated by replacing an older model with a newly trained model. Some chunk-based techniques, such as, cannot detect novel classes, whereas others can do so. Support Vector Clustering is Kernel-Based Clustering. Division of patterns, data items, and feature vectors into groups (clusters) is a complicated task since clustering does not assume any prior knowledge, which are the clusters to be searched for labeling. if you not found in any class passes through new generation of class. Thus clustering serves in particular for exploratory data analysis with little or no prior knowledge. SVC algorithm has two main steps a) Support vector machine (SVM) Training and b) Cluster mapping. Support vector machine involves estimating the outer value of and cluster labeling step involves assigning the cluster labels to each data point. Solving the optimization problem and cluster labeling is time consuming in the SVC training procedure. This paper is divided into five sections. Section-1. gives the introduction of stream data classification and support vector clustering. Section- 2. Gives the information about related work. Problem formulation in section-3. In section-4.discuss the proposed method and finally conclusion and future work in section 5.

2. RELATED WORK

In this section discuss some related work of stream data classification and support vector clustering. For the improvement of support vector clustering along with stream data classification used various algorithm in different stage such as boundary collection, noise reduction and cluster labeling. Some another method are used as kernel optimization for support vector clustering. The proposed algorithm dynamically maintains multiple spheres in a multi sphere set, which are used to represent the summary information of the historical data elements. According to the SVDD theory, the multi sphere representation provides a very compact and accurate data description of the historical data chunks, so it has limited memory consumption with precision guaranteed. The support vectors are used to construct cluster boundaries of random layer. to gain the dynamically change of stream data, when a new data chunk arrives, the multi sphere set is updated [1]. By allowing for bounded support vectors (BSVs), the proposed algorithm is capable of partitioning overlapping clusters.
[2]In this paper proposed an adaptive incremental feature combination selection method IFCS to address the "siren pitfall". Our method is stable that the results is re-producible under the same parameter setting, robust that it can seamlessly integrate with non-parametric methods and does not depend on any prior knowledge, and also low in computational cost.

The SVC algorithm, first proposed by Ben-Hur [3], identifies the cluster contours with arbitrary geometric representations, and automatically determines the number of clusters for a given dataset by a unified framework. The SVC algorithm has been widely researched in both theoretical developments and practical applications due to its outstanding features. In the SVC algorithm, data points are mapped from the data space to a high dimensional feature space using Gaussian kernels. The objective of the SVC algorithm is to look for the smallest sphere that encloses the images of data points in the feature space. This sphere is then mapped back to the data space, where a number of contours which enclose the data points are formed. These contours are interpreted as cluster boundaries. In general, the SVC algorithm involves three main steps: a) finding the hyper-sphere by solving the Wolfe dual optimization problem, b) identifying the clusters by labeling the data points with cluster labels, and c) searching a satisfactory clustering outcome by tuning kernel parameters. Wang and Chiang [5] have developed an effective parameter search algorithm to automatically search suitable parameters for the SVC algorithm. The cluster structure obtained by support vector clustering is controlled by two parameters –the parameter of kernel function (q) and the soft margin function constant of lagrangian function denoted as c. However, there is a common agreement in SVC research community that Because the computation of cluster labeling is considerably expensive, many researchers have engaged in reducing time complexity of this step. solving the optimization problem and labeling the data points with cluster labels are time-consuming in the SVC training procedure. The above limitations make the SVC algorithm inapplicable for large datasets. From literature, we found that many research efforts have been conducted to improve the effectiveness of cluster labeling. Yang [11] used proximity graphs to model the proximity structure of datasets. Their approach constructed appropriate proximity graphs to model the proximity and adjacency. After the SVC training process, they employed cutoff criteria to estimate the edges of a proximity graph. This method avoids redundant checks in a complete graph, and also avoids the loss of neighbourhood information as it can occur when only estimating the adjacencies of support vectors.Lee and Lee [7] created a new cluster labeling method based on some invariant topological properties of a trained kernel radius function. The method they proposed consisted of two phases. The first phase was to decompose a given data set into a small number of disjoint groups where each group was represented by its candidate point and all of its member points belong to the same cluster. The second phase was then to label the candidate points. Nath and Shevade [4] presented a novel approach that increases the efficiency of the SVC scheme. The geometry presented in the clustering problem was exploited to reduce the training data size. Their experiments showed that the pre-processing procedure drastically decreased the run-time of the cluster algorithm. However, different pre-specified parameters could produce totally different clustering results Wang and Chiang [6] proposed an efficient pre-processing procedure for SVC. This procedure reduces the size of the dataset by eliminating noise, outliers, and insignificant points from the dataset. Then SMO algorithm is applied on the reduced training set.

3. PROBLEM FORMULATION

In this section discuss some problem related to support vector clustering for stream data classification. some problem related to algorithm approach and some problem related to data attribute .

3.1 Correctness of data- Basically the stream data is multi-attribute data generated by stream source, some time some attribute value loss and induced noise, now various method are used for data correctness .

3.2 Every time value of parameter q width of kernel has to change-In support vector clustering the classification ratio and correctness of data depend of kernel function, if the size of kernel increase the classification rate also increase, but physically it cannot possible change the value of kernel function.

3.3 There is no standard method to estimate boundary value feature & svc feature coefficient mapping relation-Outlier and boundary value of data decreases the classification ratio because the boundary value and outlier data not participate in classification because the range of data from kernel is very high.

3.4 Stream data are continuous source of data but svc predefined process. It cannot create dynamically new feature-The infinite length of stream data some time generates new feature evaluation, but the process of classification is predefined process. The new evolved features are not are part of predefined class so new feature participation of confusion matrix.

3.5 Mapping of sphere data into cluster takes more time during new feature

evolution -The conversion of data one space to another space takes more time for clustering.

4. PROPOSED METHODOLOGY

In this section described a proposed method for improved SVC algorithm for feature reduction cum classification technique. The huge amount of feature process through our sample selection process, the sample selection process used correlation factor for estimated feature value for reduction process. The nature of mixture data correlation of attribute used in SVC algorithm. The combination of RBF and SVC algorithm perform well feature reduction cum classification process over stream data. The RBF function increases the size of sample selection. The increased size of sample selection increases the range of feature attribute of intruder data. RBF function is creating for sample selection for reduces and unreduced categories data sample for dealing out of SVC classification. The input processing of training phase is data sampling technique for classifier. Single-layer RBF networks can potentially learn virtually any input output relationship; RBF networks with single layers might learn complex relationships more quickly. The function RBFSVC creates forward networks. The network-layer network also has connections from the input to all cascaded layers. The additional connections might improve the speed at which the network learns the desired relationship.RBF artificial intelligence model is similar to feed-forward back-propagation neural network in using the back-propagation algorithm for weights updating, but the main indication of this network is that each layer of neurons related to all previous layer of neurons .The process of feature reduction and classification steps given below.

1. estimate the feature correlation attribute as $\mathrm{Rel}(a,b) = \frac{\mathrm{cov}(a,b)}{\sqrt{\mathrm{var}(a)\times \mathrm{var}(b)}}$ Here a and b the feature attribute of input data
2. the estimated correlation coefficient data passes through RBF function as

$$x(t) = w0 + \sum_{j=1}^{\text{total data}} wj \exp\left(\frac{-(\text{total} - xj)}{\sigma^2}\right)$$

3. create the relative feature difference value

$$Rc = \sum_{k=1}^{r} \sum_{i=1}^{m} (hi - h)(eik - et)$$

4. After sampling of feature data get reduces set of feature attribute of feature matrix.
5. generate feature attribute of each matrix
6. compute the entropy of feature attribute for the root node
7. $\mathrm{Entropy}(D) = -\sum_{i=1}^{N} pi \log pi \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots..(1)$
8. compute the maximum margin of feature $$FB(v) = \sum_{j=1}^{N} Pj[pi \log pi] \ldots\ldots\ldots\ldots..(2)$$
9. compute the gain of each feature attribute
   Gain (v) =Entropy (D)-FB (v)
10. determine maximum gain of feature value and split encode feature in Gaussian form
11. estimate support vector
12. data are classified
13. estimate the classification ratio
14. exit.

The process of stream data classification is done on the biases of feature classification and selection of near data in concern of kernel function. the estimated kernel function find the relational attribute parameter for selection.

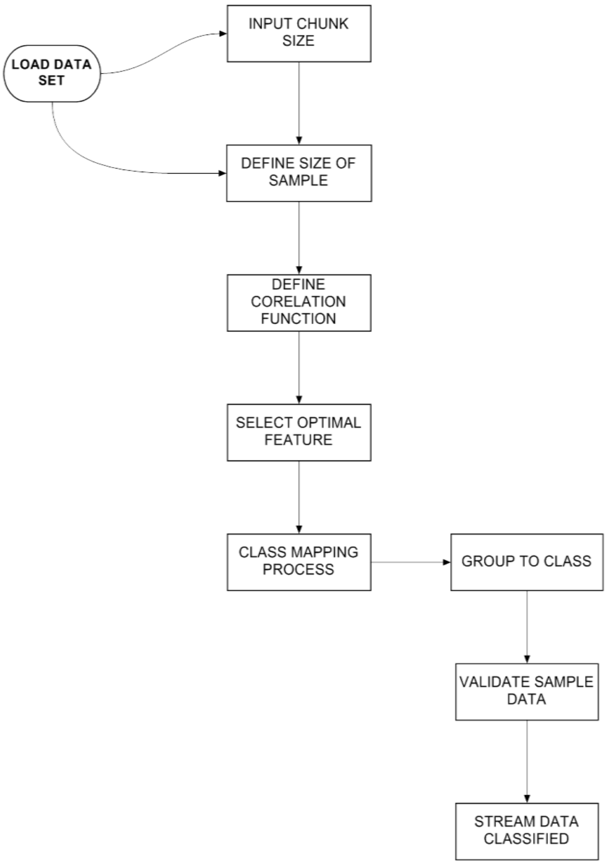

Figure 1 proposed model for stream data classification using support vector clustering.

5. CONCLUSION AND FUTURE WORK

In this paper proposed a novel method for support vector clustering for stream data classification. The proposed algorithm gives the better performance in diverse nature of data. For the collection of data used neural network based kernel function. The neural network based finds the correlation factor of data point and processing for boundary value estimation. The major problem of stream data classification is feature evaluation and concept evaluation which is controlled by neural network kernel function. Evaluation of new feature creates a problem in feature selection during the classification process of support vector clustering. This paper reduces these problems using neural networks, neural network is used to control new feature evolution problem. Neural network creates a feature prototype for cluster used in classification. The controlled feature evaluation process proposed a modified support vector cluster called SVC-NN. For future implementation neural network will improve the ratio of classification and for parameter analysis chunk size M, tradeoff parameter C and multi-sphere parameters will be used.

6. REFERENCES

[1]Chang-Dong Wang, Jian Huang La, Dong Huang, Dong Huang “SVStream: A Support Vector-Based Algorithm for Clustering Data Streams” IEEE TRANSACTIONS

ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, 2013. Pp 1410-1425.

[2]Xin Xu, Wei Wang, Guilin Zhang, Yongsheng Yu “An Adaptive Feature Selection Method for Multi-class Classification” 2010. Pp 225-230.

[3]A. Ben-Hur, D. Horn, H.T. Siegelmann, V. Vapnik, “A Support Vector Clustering Method”, In Proc. of Int. Conf. on Pattern Recognition, 2000, pp. 724-727.

[4]J. Saketha Nath, S.K. Shevade, “An Efficient Clustering Scheme Using Support Vector Methods”, Pattern Recognition, 2006, 1473-1480.

[5]J. S. Wang, J. C. Chiang, “A Cluster Validity Measure with a Hybrid Parameter Search Method for Support Vector Clustering Algorithm”, PatternRecognition, 2008, pp. 506-520.

[6]J. S. Wang, J. C. Chiang, “An Efficient Data Preprocessing Procedure for Support Vector Clustering”, Journal of Universal Computer Science, 2009, pp. 705-721.

[7]J. Lee, D. Lee, “An Improved Cluster Labeling Method for Support Vector Clustering”, IEEE Trans. Pattern Analysis and Machine Intelligence, 2005, pp. 461-464

[8]Jiawei Han and Micheline Kamber, “Data Mining: Concepts and Techniques”, Morgan Kaufmann Publishers, Second Edition, 2006, pp. 355.

[9]L. Ertoz, M. Steinbach, V. Kumar, “Finding Clusters of Different Sizes, Shapes, and Densities in Noisy, High Dimensional Data”, In Proc. of SIAM Int. Conf. on Data Mining, 2003, pp. 1-12.

[10]R. A. Jarvis, E. A. Patrick, “Clustering Using a Similarity Measure Based on Shared Nearest Neighbors”, IEEE Trans. Computers, C-22, 11, 1973, pp. 1025-1034.

[11]J. S. Wang, J. C. Chiang, “A Cluster Validity Measure with Outlier Detection for Support Vector Clustering”, IEEE Trans. Systems, Man, and Cybernetics-Part B, 38, 1, 2008, pp. 78-89.

[12]J. Yang, V. E. Castro, S. K. Chalup, “Support Vector Clustering Through Proximity Graph Modeling”, In Proc. of 9th Int. Conf. on Neural Information Processing, 2002, pp. 898-903.

[13]D. Tax and R. Duin, “Support vector domain description”, Pattern Recognition Letters, vol. 20, 1999, pp. 1191-1199.