International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970)
Volume-2 Number-4 Issue-7 December-2012

An efficient Support Vector Clustering with combined core outlier and boundary value for pre-processing of data

Deepak Kumar Vishwakarma[1], Anurag Jain[2]
Department of Computer Science & Engineering, RITS, Bhopal, India[1,2]

Abstract

The performance of support vector clustering suffered Due to noisy data. The pre-processing of data play important role in support vector cluster. In support vector clustering the mapping of data from one sphere to another sphere found some unwanted behaviour of data, these behaviour are boundary point, core and outlier. These data point degrade the performance and efficiency of support vector clustering. For the reduction of core, outlier and boundary value, we combined all dissimilar data and form COB model and data passes through genetic algorithm for collective collection of COB and reduce the COB value in data pre-processing phase. After reduction of COB support vector clustering are applied. Our empirical evaluation shows that our method is better than incremental support vector clustering and SSN-SVC.

Keywords

SVC, COB, GA, SVM

I. Introduction

Support vector clustering is great achievement over unsupervised clustering technique. The mapping and grouping of cluster are better in compression of partition clustering, density clustering and hierarchal clustering technique. Division of patterns, data items, and feature vectors into clusters with different shapes, they still cannot produce arbitrary cluster boundaries to adequately capture or represent the characteristics of clusters in the dataset [1, 9]. A group (clusters) is a complicated task since clustering does not presume any prior knowledge, which is the cluster to be searched for. There exist no class label attributes that would tell which classes exist. Some of the traditional clustering techniques are Hierarchical clustering algorithms, Partitioned clustering algorithms, nearest neighbour clustering, and Fuzzy clustering [5]. Clustering algorithms are capable of finding clusters with all shapes, dimensions, densities, and even in the presence of noise and outliers in datasets. Although these algorithms can handle. Support Vector Clustering (SVC), which is inspired by the support vector machines, can overcome the limitation of these clustering algorithms. SVC algorithm has two main steps a) SVM Training and b) Cluster Labelling [9]. SVM training step involves construction of cluster boundaries and cluster labelling step involves assigning the cluster labels to each data point. Solving the optimization problem and cluster labelling is time consuming in the SVC training procedure [4]. The performance of clustering suffered from data dissimilarity point value, the data dissimilarity point generates a irregular pattern and unshaped cluster. For the improvement of data processing used combined core, outlier and boundary value for reduction of noise in data. The selection process of reduces data passes through genetic algorithm. Genetic algorithm is heuristic function, the nature of heuristic function id give optimal result in single objective function. The rest of paper is organized as follows. In Section II discuss related work of support vector clustering. The Section III discusses the COB model. The section IV discuss proposed model. In section V discuss experimental result analysis Followed by a conclusion in Section VI.

II. Related work Support Vector Clustering

In this section we discuss some related work for support vector clustering for data pre-processing for efficient mechanism for cluster generation. There are many techniques exist in literature to reduce time complexity of cluster labelling step such as complete graph (CG) strategy [1], modified complete graph (SVG) strategy [9], proximity graph modelling [25], 2-phase cluster labelling strategy [07]. From literature, we found that many research efforts have been conducted to improve the effectiveness of cluster labelling. Because the computation of cluster labelling is considerably expensive, many researchers have engaged in reducing time complexity of this step. Yang *et al.* [25] 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 labelling 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 Sheaved [2] 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 proposed an efficient pre-processing procedure for SVC [4]. 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.

III. COB Model

We consider there is an data point collection Ec with N (N > 1) individual cluster {Ci, i = 1, 2, …, N}. For the convenience of using the simple majority voting rule, we set N as an odd number: N = 2K + 1, where K is a natural number. We further assume there is a testing dataset **X** with n items {(**x**j, yj), j = 1, 2, …, n}. Each input item **x**j is a vector with m features (variables) {xjk, k = 1, 2, …, m} and each output yj is a class label in {–1, 1}[27]. For each input item Xj, each individual classifier Ci predicts an ouput Cij. We set Zij=$\begin{cases} 1\ cij = yj \\ 0\ cij \neq y \neq j \end{cases}$ so the data point collection method E predicts the item Xj correctly if and only if $(\sum_{i=1}^{n} Zij)$ >K by the majority voting rule. We denote Pi=$\sum_{j=1}^{n} Zij/n$ as the predication accuracy of each classifier Ci and Pi=count$\left((\sum_{i=1}^{n} Zij) > K\right)/n$ as the predication accuracy of the data point collection E. A general assumption in data point collection learning is that individual cluster is independent of each other since the items are sampled from a dataset uniformly. If the accuracies of all individual cluster are the same, say, pi = p, i = 1, 2, …, N, then these cluster follow the binomial distribution and the accuracy of the data point collection method can be calculated as

$$p_B = \sum_{i=K+1}^{n} \binom{N}{i} p^i (1-p)^{N-i} \quad \ldots(1)$$

If we consider pB as a function of p, it can be shown that for any given n > 1, pB strictly increases when p increases. In the COB model, we assume that a dataset consists of three subsets: core, outlier, and boundary. For a dataset with items from multiple classes, some items may be buried by items from the same class, some may be buried by items from other classes, and some may be surrounded by mixed items from the same and other classes (the unfilled but encircled point i. We classify those items in the first case as a core subset, those in the second case as an outlier subset, and those in the last case as a boundary subset. From another point of view, a core subset contains items that are clearly different from items in other classes, an outlier subset contains items that are classified by mistake, and a boundary subset contains items that are similar to some items in other classes and can be correctly predicted or misclassified. We note that it may not be possible to classify a highly noisy dataset into these three subsets, as the errors will be overwhelming The COB model models these three subsets separately. We define the numbers of items in the core, boundary, and outlier subsets as n1, n2, and n3, respectively, so n1 + n2 + n3 = n. For an data point collection method E with N individual cluster {Ci, i = 1, 2… N}, all clusters always predict the items in the core subset correctly and predict the items in the outlier subset incorrectly. We define the accuracy on the core subset as q.

IV. Proposed Model for SVC

In this paper we proposed an optimized data point cluster for the reduction of noise, core point and outlier in individual cluster for performing an SVC cluster. For the process of optimization genetic algorithm are used. Genetic algorithm is heuristic function and the nature of heuristic function is gets optimal result. In the process of SVC of individual cluster combined with selection of feature vector of data. The multiple support vector machines combined with feature vector and spread of data in form of noise, core and outliers are calculated with binomial distribution function. The combined data of noise core and outlier passes through simple genetic algorithm and form a new class of COB and improved the voting ratio of SVC cluster. For the input of genetic algorithm create a sub set of COB features set. We randomly assigned population of genetic algorithm according to selection of COB feature set. We define COB on

the variable id which matrix contain the COB upper and lower value set. For the selection of COB population used fitness function given by

$$F(xi) = \frac{f(xi)}{\sum_{i=0}^{n} F(xi)} \quad \text{.....(2)}$$

Where f (xi) is the fitness of individual xi and F (xi) is the probability of that individual Cob selected. Here in the process of genetic algorithm crossover phase are not required. For the process of mutation we fixed the value of variable probability p=0.07. And finally gets the optimized set of cluster. Proposed model of our process shown in figure.

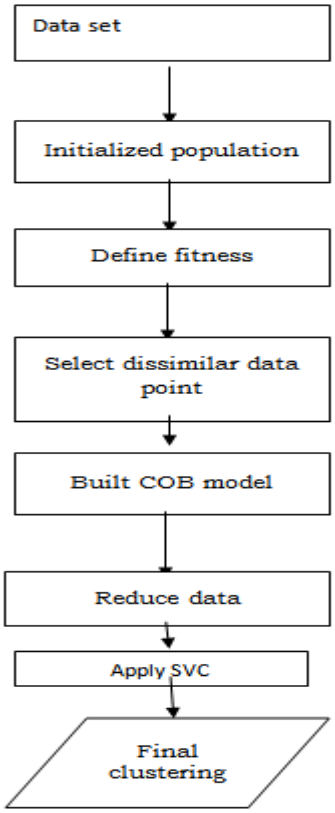

Figure 1: Flowchart

Steps for algorithm
Input data set s number of cluster M

1. For d=1 to n
2. Rd=random sample from feature set
3. Md=M(Rd)
4. Calculate COB with binomial distribution
5. Find upper and lower COB value and along with difference set weight parameter
6. Generate random population of COB matrix
7. Check fitness constraints
8. Apply mutation probability p=0.07
9. SVC output
10. Exit

V. Experimental Result Analysis

In the experimental process, I have measured classification accuracy, execution time of SVC procedure. To evaluate these performance parameters I have used six datasets from UCI machine learning repository [10] namely Wisconsin breast cancer (original), Iris, Glass identification, Page blocks classification, White wine quality, and Yeast dataset. Out of these six dataset, two are small dataset namely Iris and Glass identification dataset; and remaining four are large datasets namely Wisconsin breast cancer, Page blocks classification, White wine quality, and Yeast dataset.

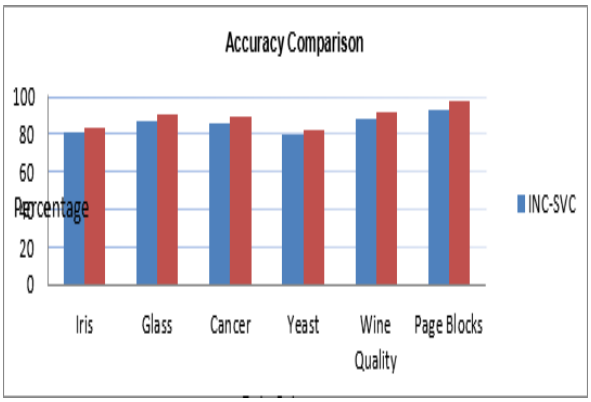

Figure 2: shows that comparative result analysis of incremental support vector clustering and cob-svc accuracy the accuracy of COB-SVC are increased due to reduce core, outlier and boundary point value

The execution time of our proposed method is decrease.

VI. Conclusion and Future Work

This proposed technique work focuses the drawbacks of SVC for dealing with large datasets. INC based data pre-processing procedure for SVC results in loss of data. To overcome these drawbacks COB_GA based data pre-processing procedure is used to eliminate noise and irrelevant data from the dataset. To measure performance of the proposed algorithm I have used six datasets namely iris, glass identification, Wisconsin breast cancer, yeast, wine quality, and page blocks classification dataset. From experimental results, it is observed that the proposed algorithm improves accuracy and efficiency of SVC for iris, glass identification, Wisconsin breast cancer, yeast, wine quality, and page blocks classification dataset without altering the final cluster configurations. By our proposed method, the classification accuracy of

all the six dataset is better than INC-SVC method. From experimental results on the six datasets shows that the proposed algorithm can be served as an effective tool in dealing with classification problems. Our proposed pre-processing procedure passes quality data to SVC training procedure and it results in increase in accuracy of SVC. Hence, it improves ability of SVC in dealing with classification problems. In the future, I would verify my proposed procedure on diverse real-world applications.

References

[1] 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.

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

[3] J. S. Wang, J. C. Chiang, "A Cluster Validity Measure with a Hybrid Parameter Search Method for Support Vector Clustering Algorithm", Pattern Recognition, 2008, pp. 506-520.

[4] J. S. Wang, J. C. Chiang, "An Efficient Data Pre-processing Procedure for Support Vector Clustering", Journal of Universal Computer Science, 2009, pp. 705-721.

[5] A. Jain, M. Murty, P. Flynn, "Data Clustering: A Review", ACM Computing Surveys, 1999, pp. 264-323.

[6] J. C. Platt, "Fast training of support vector machines using sequential minimum optimization", Advances in Kernel Methods Support Vector Learning, 1998, pp. 185-208.

[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] K. Jong, E. Marchiori, and van der Vaart, "Finding Clusters using Support Vector Classifiers", ESANN Proceedings. Europian Symposium on ANN, 2003, pp. 223-228.

[9] A. Ben-Hur, D. Horn, H.T. Siegelmann, V. Vapnik, "Support Vector Clustering", Journal of Machine Learning Research 2, 2001, pp. 125-137.

[10] C. Blake, E. Keogh, C. Merz, "UCI Repository of Machine Learning databases", Department of Information and Computer Sciences, University of California, Irvine, 1998.

[11] Tom Mitchell, "Machine Learning", McGraw Hill, Computer Science Series. 2005, Page no. 2-4, 81-95, 238-245.

[12] Nils J. Nilsson, "Introduction to machine learning", 1997, page no 1-15.

[13] Arun Pujari, "Data Mining Concepts", Universities press, page no 2-25, 2001.

[14] Steve Gunn, "Support Vector Machine for Classification and Regression", Technical Report, Faculty of Engineering, Science and Mathematics, School of Electronics and Computer Science, 10 May 1998, Page No. 2-8, 19-23.

[15] Ethan Alpayadin, "Introduction to machine learning", MIT press Cam-bridge, 2005.

[16] Christopher M. Bishop, "Pattern Recognition and Machine Learning", Springer Publication, Singapore, 2006, Page no. 1-3, 308-320.

[17] Ian H. Witten, Eibe Frank, "Data Mining-Practical Machine Learning Tools and Techniques", Morgan Kaufmann Publishers, Second Edition, 2005, pp. 7-9.

[18] V. Vapnik, "The Nature of Statistical Learning Theory" Springer, N.Y., 1995.

[19] J. P. Lewis, Tutorial on SVM, CGIT Lab, USC, 2004.

[20] E. Osuna, R. Freund, and F. Girosi, "Support Vector Machines: Training and Applications", A.I. Memo No. 1602, Artificial Intelligence Laboratory, MIT, 1997.

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

[22] 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.

[23] 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.

[24] 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.

[25] 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.

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

[27] Xueyi Wang" A New Model for Measuring the Accuracies of Majority Voting Ensembles" in IEEE World Congress on Computational Intelligence in 2012.