International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied Research)

ISSN (Print): 2279-0020
ISSN (Online): 2279-0039

International Journal of Engineering, Business and Enterprise Applications (IJEBEA)

www.iasir.net

A Survey of Support Vector Clustering based Labeling

Rinki Baghele, Prof. Aparajit Shrivastava
Department of Computer Science
Shri Ram College of Engineering & Management
Banmore, INDIA

Abstract: *Division of patterns, data items, and feature vectors into groups (clusters) is a complicated task since clustering does not assume any prior knowledge that is the clusters to be searched for. Clustering algorithms are capable of finding clusters with diverse sizes, shapes, densities, and constantly in the existence of noise and outliers in datasets. Although these algorithms can handle clusters with variety of shapes, they yet cannot generate arbitrary cluster boundaries to adequately capture or represent the characteristics of clusters in the dataset. Support Vector Clustering (SVC), which is inspired by the support vector machines, can overcome the limitation of these clustering algorithms. In this paper we are presenting various support vector clustering schemes that are offered by several researchers.*

Keywords: *SVM, Clustering, Gaussian Kernel, SVC, SNN.*

I. Introduction

Because of the rapid development of computer and information technologies, databases become larger and larger that increases the necessity of more efficient and effective diagnostic tools to analyze and recover useful information or knowledge from databases. Clustering algorithms are functional for discovering groups and distributions in large databases and have been extensively adopted in diverse scientific fields and commercial sectors.

Clustering is a division of data into groups of comparable objects. Each one of them called cluster that consists of objects that are similar amongst them and dissimilar compared to object of other groups. Representing data by smaller amount clusters unavoidably loses convinced fine details, but attains generalization. It characterizes many data objects by a small number of clusters, and therefore it models data by its clusters [1]. SVC algorithm has two main steps one is SVM Training and another is Cluster Labelling. 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.

A group (clusters) is a complicated task since clustering does not presume any earlier information, which are the clusters to be investigated for. Some of the conventional clustering methods are Partitional clustering algorithms, nearest neighbor clustering, Hierarchical clustering and Fuzzy clustering.

Supervised Clustering Task: Clustering is sometimes applied to multiple sets of items, with each set being clustered separately. For example, in the noun-phrase co reference task, a single document's noun-phrases are clustered by which noun phrases refer to the same entity, and in news article clustering, a single day's worth of news articles are clustered by topic. In this method, users provide complete clustering of a few of these sets to express their preferences, e.g., provide a few complete clustering of several documents' noun-phrases, or several days' news articles [2].

SVM based clustering: The structural SVM algorithm provides a general framework for learning with complex structured output spaces [3]. This work shares many similarities with the semi supervised clustering, which attempts to form desirable clustering's by taking user information into account, typically of the form 'these items do/do not belong collectively.' Clustering is valuable to numerous sets of items, with each set being clustered separately. Several supervised clustering techniques alter a clustering algorithm so it satisfies constraints [4]. In this, users provide a few complete clustering's of several documents' noun-phrases, or several days' news articles.

In Support Vector Clustering (SVC) scheme data points are plotted from data space to a high dimensional characteristic or feature space via Gaussian kernel. In feature space looks for the least sphere that includes the data image. This sphere is mapped back to data space. Where this data sphere is forms a set of contours that surrounds the data points. These set of contours are inferred as cluster boundaries. Points together with this by each split contour are associated with the equivalent cluster. As the width parameter of the Gaussian kernel is reduced, the number of disjointed contours in data space raises, leading to a growing number of clusters. In view

of the fact that the contours can be deduced as delineating the support of the core probability distribution, this algorithm can be viewed as one identifying valleys in this prospect distribution [5].
In the SVC algorithm, data points are mapped from the data space to a high dimensional feature space using Gaussian kernels. The goal 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. Generally SVC algorithm comprises of three main levels. First one is finding the hyper-sphere for the problem solution; second one is identifying the clusters by labeling the data points according to cluster labels and third one is searching a satisfactory clustering result by tuning kernel parameters [6].

II. Ant Colony Optimization

Ant Colony Optimization (ACO) is a paradigm for designing met heuristic algorithms for combinatorial optimization problems. The first algorithm which can be classified within this framework was presented 1991and, since then, many diverse variants of the basic principle have been reported in the literature. The essential trait of ACO algorithms is the combination of a priori information about the structure of a promising solution with posterior information about the structure of previously obtained good solutions."

1. The first ant finds the food source (F), via any way (a), then returns to the nest (N), leaving behind a trail pheromone (b)
2. Ants indiscriminately follow four possible ways, but the strengthening of the runway makes it more attractive as the shortest route.
3. Ants take the shortest route; long portions of other ways lose their trail pheromones.

III. Background

Clustering has always been a challenging task in pattern recognition. Many clustering algorithms have been proposed in the earlier periods. Division of patterns, feature vectors, and data items into groups (clusters) is a complex task since clustering does not assume any previous knowledge that are the clusters to be investigated for. In it there is no class label attributes that would tell which classes exist. Some of the conventional clustering techniques are Hierarchical clustering algorithms, nearest neighbor clustering, Partitioned clustering algorithms and Fuzzy clustering. Kernel-based learning algorithms have become increasingly important in pattern recognition and machine learning, particularly in supervised classification and regression analysis with the introduction of support vector machines. Clustering algorithms are capable of finding clusters with diverse densities, sizes, shapes, and even in the existence of noise and outliers in datasets. Although these algorithms can handle clusters with dissimilar shapes, they still cannot construct capricious cluster boundaries to adequately capture or represent the characteristics of clusters in the dataset.

IV. Related Work

In year 2001, Ben-Hur et al presented a concept of support vector machine. They offered a non-parametric clustering algorithm based on the support vector approach. A support vector algorithm was used to characterize the support of a high dimensional distribution. A unique advantage of this algorithm is that it can generate cluster boundaries of arbitrary shape, whereas other algorithms that use a geometric representation are most often limited to hyper-ellipsoids. SVC can deal with outliers by employing a soft margin constant that allows the sphere in feature space not to encircle all points. For great values of this parameter, they can also deal with overlapping clusters [5].

Support vector clustering (SVC) [5] can overcome the limitation of these clustering algorithms. The SVC algorithm, first 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 [5].
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.
They have proposed a novel clustering method, SVC, based on the SVM formalism. This method has no explicit bias of either the number or the outline of clusters. It is having two parameters, allowing it to obtain various clustering solutions. The parameter q of the Gaussian kernel determines the scale at which the data is investigated, and as it is enlarged clusters begin to divide. Another parameter, p, is the soft margin constant that controls the number of outliers. This parameter 'p' enables analysing noisy data points and separating between overlapping clusters [5].

In year 2006, J.S. Nath and S.K. Shevade offered a pre-processing step that eliminates data points from the training data that are not crucial for clustering. This is a new scheme under title of An Efficient Clustering Scheme Using Support Vector Methods [6]. They present a novel method that efficiently eliminates data points from the training data that are not crucial for determining the cluster labels. The computational effort in clustering (with such pre-processing) will be less. In Support Vector Clustering (SVC) [5], data points are mapped from the data space to a high dimensional space called feature space. In feature space, the smallest hyper sphere that encloses the images of most of the data points is identified. This hyper sphere when mapped back to data space forms a set of disjoint contours which enclose most of the data points. Such types of contours are inferred as cluster boundaries and the points enclosed by each contour are associated with the same cluster [6].

The NSVs (non-support vector points)do not influence the inference function. Also, it can be shown that the optimum dual solution (for a given (C, σ)) will be the same even if one uses the SVs instead of all the data points in solving the optimization. Thus, if the set of NSVs is known a priori, then one can eliminate them from training set and achieve a reduction in the computational effort. But, it is not possible to determine the exact set of SVs without solving the optimization problem. However, one can eliminate some data points that have a high chance of being NSVs. If in NSV processes some SVs are eradicated, then the optimum dual solution will not be the same as that with all data points. However, if the final clustering obtained is not significantly different from that obtained with the entire training set, then such elimination can be employed [6].

S. Wang and J.-C. Chiang proposed A Cluster Validity Measure with Outlier Detection for Support Vector Clustering [7] and A Cluster Validity Measure with a Hybrid Parameter Search Method for Support Vector Clustering Algorithm [8]. They have developed an effective parameter search algorithm to automatically search suitable parameters for the SVC algorithm. However, there is a common agreement in SVC research community—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 our review of literature, we found that many research efforts have been conducted to improve the effectiveness of cluster labeling. Because the computation of cluster labeling is considerably expensive, many researchers have engaged in reducing time complexity of this aspect [7] [8].

In year 2009, Wang and Chiang suggested An Efficient Data Preprocessing Procedure for Support Vector Clustering. This procedure ameliorates the drawbacks of the SVC algorithm for dealing with large datasets. The preprocessing procedure utilizes a shared nearest neighbor (SNN) algorithm for eliminating the noise points, and the concept of unit vectors for removing the core points from the dataset. Since the size of the dataset is concentrated, the computational load for solving the optimization problems as well as cluster labeling can be greatly decreased [9].

Solving the optimization problem and labeling the data points with cluster labels are time-consuming in the SVC training procedure. This composes by means of the SVC algorithm to process large datasets inefficient. Thus, how to exclude redundant data points from a dataset is an important issue for minimizing the time spent in solving the optimization problem of the SVC algorithm. This research challenge in this topic is how to identify insignificant data points so that the removal of these data points does not significantly alter the final cluster configuration. This idea is to eliminate insignificant data points, such as noise and core points, from the training datasets, and use the remaining data points to do the SVC analysis. Due to the size reduction of the exercise datasets, the computational attempt for answering the optimization problem can be greatly decreased [9].

Deepak Kumar Vishwakarma and Anurag Jain presented a review of Support Vector Clustering with different Kernel function for Reduction of noise and outlier for Large Database. 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. From literature, they found that many research efforts have been conducted to improve the efficiency of cluster labeling. Because the computation of cluster labeling is considerably expensive, many researchers have engaged in reducing time complexity [10].

Vishwakarma and Jain describe the procedure for constructing cluster based SVM, i.e. CK-SVM. In this regard we have introduced a cluster based simple and fast training algorithm to solve outliers and computational cost problem. In addition, CK-SVM has provided efficiency for fast classification and continuous outputs via weighted distances for multiclass classification. Outlier detection encompasses aspects of a broad spectrum of procedures. Several techniques occupied for detecting outliers are fundamentally identical but with different names chosen by the authors [10].

Yang et al. [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 cut-off 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. Currently there are two data pre-processing procedures are available in literature for support vector clustering (SVC). These pre-processing techniques remove noise points, outliers and insignificant points which are not important for clustering. They reduce the size of the training dataset. After pre-processing, Sequential Minimal Optimization (SMO) algorithm is applied on the reduced dataset for solving the optimization problem. Next, labelling of each data point with appropriate cluster labels is done using cluster labelling method [6, 9].

In year 2011, C. D. Wang et al offered an incremental support vector machine [12]. They suggested an incremental support vector clustering (ISVC), which constructs a sphere in an incremental manner. The incremental (unsupervised) clustering is more challenging than the supervised incremental learning [13], [14], which hinders the development of incremental clustering. In incremental clustering, without any cluster label or other prior knowledge about the class distribution, it is difficult to decide what summary information of the historical data should be used for accurately learning an updated model.

Inspired by the previous work on incremental SVM, the first time proposes an incremental support vector clustering (ISVC) algorithm. The basic idea is to use SVs as the representative information obtained so far to learn a new sphere. At each step, only the SVs are persevered, which are added to the data points of the new chunk to form the current data so as to learn a modernized sphere. Hypothetical analysis has exposed that the proposed ISVC algorithm can generate cluster structure (i.e. sphere) the same as SVC, assuming that no outlier is presented in the data; meanwhile it dramatically reduces time consumption. Experimental results have validated the theoretical analysis [12].

By regarding the data as arriving over time in chunks, only the support vectors of the historical data and the data points of the new data chunk are used to learn an updated sphere. They have theoretically shown that the proposed ISVC approach can generate the same cluster structure as SVC. Computational complexity analysis has also revealed that our approach is several magnitudes faster and requires much lower memory consumption in sphere construction than the conventional SVC method [12]. After the SNN algorithm is executed, a large amount of noise points or outliers are detached from the datasets. They [9] hope that the proposed data preprocessing procedure does not significantly alter the final cluster configurations but can save the computational time of SVC. As a result, they want to reduce non-support vector data points, such as core points. To achieve the intention, they additionally suggest a method based on the concept of unit vectors [6] to eliminate the core points and retain the representative data points that are near the cluster boundaries [9].

Cluster Validity Measure with OD for SVC: Several cluster validity indexes have been presented. However, none of them considers the special properties of the SVC algorithm. Many of the validity techniques that compare the inter-cluster versus intra-cluster variability tend to favour configurations with ball-shaped well-separated clusters. Using the existing cluster validity measures for irregularly shaped clusters is problematic because the existing validity measures are not able to measure the distance between two clusters with nonlinearly separable. Arbitrary shapes. In addition, the performance of these measures usually degrades when the data sets contain noise or outliers, which means that they lack an effective mechanism to deal with noise or outliers [10].

In general, following their rule to make reasonable adjustments for these two parameters may result in desirable clustering outcomes. However, the time-consuming procedure of iterative executions of the SVC algorithm with different parameter selections is necessary for obtaining a desirable outcome. Moreover, varying the value of *C* to allow for the existences of outliers may increase the chance of preferable contour separations, but it may create subsidiary clusters that hinder the chance for discovering the physical cluster configuration. Hence, the clustering result is sensitive to the value of *C*, and some trial-and-error efforts are usually inevitable for reaching a desirable outcome when Ben-Hur's [5] heuristic rule is applied. To prevent these two drawbacks while maintaining a minimal number of clusters and assuring smooth cluster boundaries, Wang and Chiang proposed a systematic approach that integrated a new cluster validity measure, an outlier detection method, and a cluster merging mechanism [7].

V. Conclusion

Clustering has always been a tricky task in pattern classification. Many clustering algorithms have been proposed in the few years back. Data items, division of patterns and characteristic vectors into clusters with unlike shapes, they still cannot generate arbitrary cluster boundaries to adequately capture or represent the characteristics of clusters in the dataset. Support Vector Clustering (SVC), which is inspired by the support vector machines, can overcome the limitation of these clustering algorithms. In recent time, specialists have made use of different cluster labeling techniques and different pre-processing procedures for improving the efficiency of SVC procedure. Pre-processing procedures used for SVC to reduce SVC training set are Heuristics for Redundant-point Elimination (HRE) and Shared Nearest Neighbor (SNN) technique result in loss of data. This paper represents various SVC techniques.

VI. References

[1] Han J. and Kamber M. "Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers, 2001.

[2] Thomas Finley and Thorsten Joachims "Supervised Clustering with Support Vector Machines", Proceedings of the 22nd International Conference on Machine Learning, pp. 217 – 224, 2005.

[3] Tsochantaridis, I., Hofmann, T., Joachims, T., & Altun, Y. "Support vector machine learning for interdependent and structured output spaces", Proceedings of the twenty-first international conference on Machine learning (ICML - 04), pp. 104, 2004.

[4] Wagstaff, K., Cardie, C., Rogers, S., & Schroedl, S., "Constrained k-means clustering with background knowledge", Proceedings of the Eighteenth International Conference on Machine Learning (ICML -01), pp. 577 – 584, 2001.

[5] Asa Ben-Hur, David Horn, Hava T. Siegelmann and Vladimir Vapnik "Support Vector Clustering", Journal of Machine Learning Research vol. 2, pp. 125 - 137, 2001.

[6] J. Saketha Nath, S.K. Shevade, "An Efficient Clustering Scheme Using Support Vector Methods", journal of Pattern Recognition, vol. 39, issue 8, pp. 1473-1480, 2006.

[7] J. S. Wang and J.-C. Chiang, "A Cluster Validity Measure with Outlier Detection for Support Vector Clustering"; IEEE Transaction Systems, Man, and Cybernetics-Part B, vol. 38, issue 1, pp. 78-89, 2008.

[8] J. S. Wang and J. C. Chiang, "A Cluster Validity Measure with a Hybrid Parameter Search Method for Support Vector Clustering Algorithm"; journal of Pattern Recognition, vol. 41, issue 2, pp. 506-520, 2008.

[9] Jeen-Shing Wang and Jen-Chieh Chiang "An Efficient Data Preprocessing Procedure for Support Vector Clustering", Journal of Universal Computer Science, vol. 15, no. 4, 705-721, 2009.

[10] Deepak Kumar Vishwakarma and Anurag Jain "A review of Support Vector Clustering with different Kernel function for Reduction of noise and outlier for Large Database", International Journal of Advanced Computer Research, ISSN: 2277-7970, Vol. 2, No. 4, Issue-7, pp. 144 – 150, December-2012.

[11] J. Yang, V. E. Castro, S. K. Chalup, "Support Vector Clustering Through Proximity Graph Modeling", In Proceedings of 9th Internal Conference on Neural Information Processing, pp. 898-903, 2002.

[12] Chang-Dong Wang, Jian-Huang Lai, Dong Huang "Incremental Support Vector Clustering", 11th IEEE International Conference on Data Mining Workshops, pp. 839 – 846, 2011.

[13] J. Lin, M. Vlachos, E. J. Keogh, and D. Gunopulos, "Iterativeincremental clustering of time series," in Proc. of EDBT, pp. 106–122, 2004.

[14] B. Liu, Y. Shi, Z. Wang, W. Wang, and B. Shi, "Dynamic incremental data summarization for hierarchical clustering," in Proc. of WAIM, pp. 410–421, 2006.

