

# Improving Support Vector Clustering with Ensembles

Wilfredo J. Puma-Villanueva, George B. Bezerra, Clodoaldo A. M. Lima, Fernando J. Von Zuben

LBiC/DCA/FEEC/Unicamp

C.P. 6101

Campinas/SP, Brazil, 13083-970

+55 19 3788-3885

{wilfredo, bezerra, moraes, vonzuben}@dca.fee.unicamp.br

**Abstract:** Support Vector Clustering (SVC) is a recently proposed clustering methodology with promising performance for high-dimensional and noisy data sets, and for clusters with arbitrary shape. However, setting the parameters of the SVC algorithm is a challenging task. Instead of searching for a single optimal configuration, the proposal involves generation, selection, and combination of distinct clustering solutions that guides to a consensus clustering. The purpose is to deal with a wide variety of clustering problems without the necessity of searching for a single and dedicated high-performance solution.

## I. PROBLEM STATEMENT

Support Vector Machines (SVM) are high-performance supervised learning machines based on the Vapnik's Statistical Learning Theory, and successively extended by a number of researchers to deal with clustering problems. The SVM variants are generally competitive among each other, even when they differ on formulation, solution strategy, and/or the choice of the kernel function. Under the availability of multiple learning machines, there are many theoretical and empirical reasons to implement an ensemble.

Ensembles involve the generation, selection, and linear/nonlinear combination of a set of individual components designed to simultaneously cope with the same task. This is typically done through the variation of some configuration parameters and/or employment of different training procedures, such as bagging and boosting. Such ensembles should properly integrate the knowledge embedded in the components, and have frequently produced more accurate and robust models. The effectiveness of the ensemble will strongly depend on the diverse behaviour and accuracy of the learning machines taken as components.

For a sample of size  $N$  composed of  $p$ -dimensional real-valued vectors, clustering is a procedure that divides the  $p$ -dimensional vectors in  $m$  disjoint groups. Data points within each group are more similar to each other than to any data point in other groups.

Each clustering procedure may produce diverse solutions depending on its parameters setup. In cases where no a priori knowledge is available, it becomes quite difficult to attest the consistency of a single solution. Cluster

boundaries tend to be fuzzy, and clustering results will significantly vary at transitory regions.

The resulting diversity among clustering proposals can be explored to synthesize an ensemble of clustering solutions. The main aspects to be explored are:

- Reuse of the knowledge implicit in each clustering solution.
- Clustering over distributed datasets in case where the data cannot be directly shared or grouped together because of restrictions due to ownership, privacy, and storage.
- Attribution of a confidence level to each cluster.

The ensemble proposed here will combine partitions produced by SVC (Support Vector Clustering) [1] [2]. SVC is used to map the data set into a higher dimensional feature space using a Gaussian kernel, and then searches for the minimal enclosing sphere. When the sphere is mapped back to the original data space, it will automatically separate data into clusters. The SVC methodology can generate clusters with arbitrary shape and size. Besides, it also has a unique mechanism to deal with outliers, making it especially adequate for noisy datasets.

Yang et al. [20] proposed a mechanism to improve the performance of the original SVC [1] by adopting proximity graph tools at the cluster assignment stage, thus increasing the accuracy and providing scalability to deal with large data sets by a considerable reduction of the required processing time.

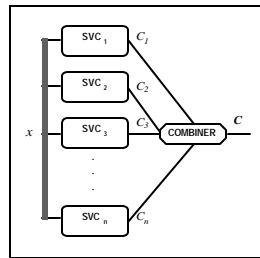
## II. CURRENT RESEARCH

Researches involving combination of multiple clustering are relatively few in the machine learning community. Park et al. [16] adopted several values for the width parameter of the Gaussian kernel in SVC, aiming at obtaining various adjacency matrixes, and then combined them via Spectral Graph Partitioning to obtain one consensus adjacency matrix. The following works, though relevant to the intended application, did not apply kernel-based approaches to perform clustering. Fisher [5] analyzed methods for iteratively improving an initial set of hierarchical clustering solutions. Fayyad et al. [4] obtained multiple approximate  $k$ -means solutions in main memory after making a single pass through a database

and combining these means finding a final set of cluster centers. Fred et al. [6] presented an evidence accumulation framework, wherein multiple  $k$ -means with a much higher value of  $k$  than the original anticipated answer were run on a common data-set. Kargupta et al. [8] combined multiple clustering solutions based only on a partial set of features. Johnson and Kargupta [7] presented a feasible approach for combining distributed agglomerative clustering solutions. Kargupta et al. [9] introduced a distributed method of PCA for clustering.

### III. KEY AVENUES

The research group has been involved with clustering, ensembles, and SVM approaches for a while [3] [10] [11] [12] [13] [14] [15]. Based on the classical ensemble for classification approach, an extension for clustering can be conceived as outlined in Fig. 1. Three phases are involved in the implementation: Generation, Selection and Combination. In what follows, we describe the essence of the proposal.



$x$  : Input data  
 SVC $_i$  : Particular clustering approach (SVCs with distinct kernels)  
 $C_i$  : Clustering solution produced by SVC $_i$   
 $C$  : Consensus solution

Figure 1. Ensemble of SVCs

#### Generation strategy

A high rate of diversity among the components of the ensemble will generally allow improving the final result. Thus, we can generate diversity by choosing among a variety of kernel functions (see Table 1) available in the literature.

Table 1. Types of kernel functions.

Gaussian Radial Basis Function	$K(x, y) = \exp\left(-\frac{(x-y)^2}{2\sigma^2}\right)$
Exponential Radial Basis Function (RBF)	$K(x, y) = \exp\left(-\frac{ x-y }{2\sigma^2}\right)$
Hyperbolic Tangent	$K(x, y) = \tanh(b(x \cdot y) + c)$
Fourier Series	$K(x, y) = \frac{\sin\left(d + \frac{1}{2}\right)(x-y)}{\sin\left(\frac{1}{2}(x-y)\right)}$
Linear Splines	$K(x, y) = 1 + xy + xy \min(x, y) - \frac{(\alpha + \beta)}{2} (\min(x, y))^2 + \frac{1}{3} (\max(x, y))^3$
Bn-splines	$K(x, y) = B_{2n+1}(x-y)$

A kernel function is a function that represents the inner product in a higher dimensional space, named feature space.

No other approach in the literature has tried distinct kernel functions in an ensemble of SVCs.

#### Selection of the clustering solution

This part is still to be properly defined. The selection of appropriate candidates among the clustering solutions is a challenging task, because distinct solutions may present a diverse number of clusters, and defining the optimal number of clusters is still an open question. In SVC, after computing the minimum radius of the sphere that contains the whole dataset in feature space, the corresponding number of clusters in the original space is very sensitive to the parameters of the kernel function. Some strategies to estimate the proper number of clusters have been proposed in the literature, such as Akaike's and Bayesian Information Criterion, and Minimum Message Length. The most used is Minimum Description Length (MDL), originally proposed by Rissanen [18], which has been widely applied in the field of neural networks. Robust Growing Neural Gas, proposed by Qin and Suganthan [17], adopted MDL to their constructive clusterization model. However, the extension of such approach to deal with SVC seems to be very computationally demanding.

#### Combination strategy

Strehl and Ghosh [19] presented a rich framework for combining clustering solutions, in terms of an optimization problem, and proposed up to four combination methods: direct and greedy optimization, cluster-based similarity partitioning algorithm (CSPA), hyper-graph partitioning algorithm (HGPA), and meta-clustering algorithm (MCLA).

CSPA: Finds a relationship between data points in the same cluster used to establish a measure of pairwise similarity. An induced similarity measure is then proposed to recluster the data points, reaching a consensus clustering solution.

HGPA: The objective is to approximate the maximum mutual information criterion with the minimum number of edges to be cut. Basically, the cluster ensemble problem is posed as a partitioning problem of a suitably defined hypergraph having hyperedges that represent clusters.

MCLA: Groups of clusters, denoted metaclusters, have to be identified and collapsed.

#### IV. REFERENCES

- [1] A. Ben-Hur, D. Horn, H. T. Siegelmann, and V. Vapnik. Support vector clustering. *Journal of Machine Learning Research*, no. 2, pp. 125-137, 2001.
- [2] A. Ben-Hur, H. T. Siegelmann, and V. N. Vapnik. A support vector clustering method. *Proceedings of International Conference on Pattern Recognition*, vol. 2, 2000, pp. 728-732.
- [3] A.L.V. Coelho, C.A.M. Lima, F.J. Von Zuben. GA-based Selection of Components for Heterogeneous Ensembles of Support Vector Machines. 2003 Congress on Evolutionary Computation (CEC'2003), Canberra, Australia, 8<sup>th</sup>-12<sup>th</sup> December, vol. 3, pp. 2238-2245, 2003.
- [4] U. M. Fayyad, C. Reina, and P. S. Bradley. Initialization of iterative re-norm clustering algorithms. In *Proc. 14th Intl. Conf. on Machine Learning (ICML)*, pages 194-198, 1998.
- [5] Fisher Doug. Iterative optimization and simplification of hierarchical clusterings. *Journal of Artificial Intelligence Research*, 4:147-180, 1996.
- [6] A. L. N. Fred and A. K. Jain. Data clustering using evidence accumulation. *Proceedings of ICPR*, 2002.
- [7] E. Johnson and H. Kargupta. Collective, hierarchical clustering from distributed, heterogeneous data. In M. Zaki and C. Ho, editors, *Large-Scale Parallel KDD Systems*, volume 1759 of Lecture Notes in Computer Science, pages 221-244. Springer-Verlag, 1999.
- [8] H. Kargupta, B. Park, D. Hershberger, and E. Johnson. Collective data mining: A new perspective toward distributed data mining. In Hillol Kargupta and Philip Chan, editors, *Advances in Distributed and Parallel Knowledge Discovery*. MIT/AAAI Press, 1999.
- [9] H. Kargupta, W. Huang, Krishnamoorthy, and E. Johnson. Distributed clustering using collective principal component analysis. *Knowledge and Information Systems Journal Special Issue on Distributed and Parallel Knowledge Discovery*, 3:422-448, 2001.
- [10] C.A.M. Lima, A.L.V. Coelho, F.J. Von Zuben. Fuzzy Systems Design via Ensembles of ANFIS. *Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE'2002)*, vol. 1, pp. 506-511, in the 2002 IEEE World Congress on Computational Intelligence (WCCI'2002), Honolulu, Hawaii, May 12-17, 2002.
- [11] C.A.M. Lima, A.L.V. Coelho, F.J. Von Zuben. Ensembles of Support Vector Machines for Regression Problems. *Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN'2002)*, vol. 3, pp. 2381-2386, in the 2002 IEEE World Congress on Computational Intelligence (WCCI'2002), Honolulu, Hawaii, May 12-17, 2002.
- [12] C.A.M. Lima, A.L.V. Coelho, F.J. Von Zuben. Model Selection based on VC-dimension for Heterogeneous Ensembles of Support Vector Machines. *Proceedings of the 4th International Conference on Recent Advances in Soft Computing (RASC2002)*, Nottingham, United Kingdom, pp. 459-464, December 2002.
- [13] C.A.M. Lima, A.L.V. Coelho, W.J. Puma-Villanueva, F.J. Ensembles of Support Vector Machines for Classification Tasks with Reduced Training Sets, *WSEAS Transactions on Systems*, Issue 2, Volume 2, pp. 370-375, April 2003.
- [14] C.A.M. Lima, A.L.V. Coelho, W.J. Puma-Villanueva, F.J. Von Zuben. Gated Mixtures of Least Squares Support Vector Machine Experts Applied to Classification Problems. *Proceedings of the 5th International Conference on Recent Advances in Soft Computing (RASC2004)*, Nottingham, United Kingdom, pp. 494-499, December 2004.
- [15] C.A.M. Lima, W.J. Puma-Villanueva, E.P. dos Santos, F.J. Von Zuben. A Multistage Ensemble of Support Vector Machine Variants. *Proceedings of the 5th International Conference on Recent Advances in Soft Computing (RASC2004)*, Nottingham, United Kingdom, pp. 670-675, December 2004.
- [16] J. H. Park, X. Ji, H. Zha and R. Kasturi, Support Vector Clustering Combined with Spectral Graph Partitioning. *Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04)*, 2004.
- [17] A. K. Qin and P. N. Suganthan. Robust growing neural gas algorithm with application in cluster analysis. *Neural Networks special issue on Recent Developments in Self-Organizing Systems*, vol. 17, no. 8-9, pp. 1135-1148, Oct.-Nov. 2004.
- [18] J. Rissanen, *Stochastic complexity in statistical inquiry* World Scientific: series in computer science, 15, 1989.
- [19] A. Strehl and Ghosh J. Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *Journal on Machine Learning Research (JMLR)*, 3:583-617, December 2002b.
- [20] J. Yang, V. Estivill-Castro and S.K. Chalup. Support Vector Clustering Through Proximity Graph Modelling. *Special Session on Support Vector machines: 9th International Conference on Neural Information Processing ICONIP 2002*. November 18-22, 2002, Singapore.