Ant Colony Optimization Technique: Robust Data Pre-processing Procedure Support Vector Clustering

Pradeep Jha[1], Krishan Kant Lavania[2],Deepak Dembla[3]
[1] M.Tech. Student Department of CSE ,AIET ,Jaipur, INDIA
[2] Assoc. Prof , Department of CSE ,AIET ,Jaipur, INDIA
[3] Assoc. Prof & HOD, Department of CSE, AIET, Jaipur, INDIA
pradeep.jha198@gmail.com, k@lavaniaa.in,deepak_dembla@yahoo.com

Abstract: This paper presents robust data preprocessing and minimization of Core, Outlier and Noise Procedure for Support Vector Clustering(SVC) .Cluster labeling is time consuming in the SVC training procedure. Our main objective of research is to reduce the execution time of SVC procedure as well as to improve the ability of proposed SVC scheme in dealing with classification problems. The procedure contains a Ant colony optimization (ACO) technique .We have used Ant Colony Optimization (ACO) based data preprocessing step to remove noise points, outliers, and insignificant points from the dataset Experiments showed reduction in the execution time of SVC procedure without altering the final cluster configurations. Using our proposed method, the classification accuracy of all dataset is better than the SNN-SVC method..

Keywords: Support Vector Clustering, Ant Colony Optimization.

1 INTRODUCTION

Support Vector Clustering was begun to come into reality in research during 2000s. Many researches had been done till date. Clustering has always been a challenging task in pattern recognition. Many clustering algorithms have been proposed in the past years. Division of patterns, data items, and feature vectors into groups (clusters) is a complex task since clustering does not assume any previous knowledge, which are the clusters to be searched for. There exist no class label attributes that would tell which classes exist. Some of the conventional clustering techniques are Hierarchical clustering algorithms, Partitional clustering algorithms, Nearest neighbor clustering, and Fuzzy clustering . Clustering algorithms are capable of finding clusters with different shapes, sizes, densities, and even in the presence of noise and outliers in datasets. Although these algorithms can handle clusters with different shapes, they still cannot produce arbitrary cluster boundaries to sufficiently capture or represent the characteristics of clusters in the dataset [1, 2]. Support Vector Clustering (SVC), which is inspired by the support vector machines (SVM), can overcome the limitation of clustering algorithms. SVC algorithm has two main steps[3]. (a) SVM Training and (b) Cluster Labeling .SVM training 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 [4]. Many of the research efforts have been taken to improve the efficiency of cluster labeling step. Only little work is done to improve the accuracy and efficiency of SVC training procedure. In recent time, specialists have made use of different cluster labeling techniques and different preprocessing procedures for improving the efficiency of SVC procedure. Preprocessing 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.

1.1 Support Vector Clustering

Given a set of data points $x_j \epsilon \ \mathfrak{R}^d$, j = 1,…, N and a nonlinear mapping function $\Phi : \mathfrak{R}^d \rightarrow F$, the objective is to find a hypershpere with the minimal radius R, such as

$$\|\Phi(x_j) - \alpha\|^2 \leq R^2 + \xi_j , \qquad \text{----------(1)}$$

where α is the center of the hypersphere and $\xi_j \geq 0$ are the slack variables allowing soft constraints. The primal problem is solved in its dual form by introducing the Lagrangian

$$L = R^{2\ -} \Sigma_j \sum_{i=1}^{N} \alpha_i (R^{2\ +} \xi_j - \|\Phi\ (x_j) - \alpha\|^2)\beta_j - \Sigma_j\ \xi_j\mu_j + C\ \Sigma_j\ \xi_j, \qquad \text{----------(2)}$$

where $\beta_j \geq 0$ and $\mu_j \geq 0$ are Lagrange multipliers and $C\ \Sigma_j\ \xi_j$ is a penalty term with C as a regularization constant [1, 2]. The dual form of the constrained optimization is constructed as

$$\max W = \Sigma_j\ \Phi(x\ ) - \Sigma_{i,j}\ \beta_i\beta_j\ \Phi(x_i)\cdot\Phi(x_j), \qquad \text{----------(3)}$$

subject to the constraints:

$$(1)\ \ 0 \leq \beta_j \leq C, \qquad \text{----------(4)}$$

$$(2)\ \ \Sigma_j\beta_j = 1 \text{ for j=1, 2, N} \qquad \text{----------(5)}$$

Using the kernel representation $k(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)$, Eq. (1.3) is rewritten as

$\max W = \Sigma_j\, k(x_j, x_j)\beta_j - \Sigma_{i,j}\, \beta_i\beta_j\, k(x_i, x_j),$ ---------- (6)

The Gaussian kernel $k(x_i, x_j) = e^{-q||x_i -x_j||2}$ is usually used for SVC algorithms, while polynomial kernels do not generate tight contour representations of clusters [1, 2].

Furthermore, for each data point x, the distance of Φ(x) to the center is calculated as

$R^2(x) = \| \Phi(x) -\alpha \|^2 = k(x, x) - 2\, \Sigma_j\beta_j k(x_j, x) - \Sigma_{i,j}\, \beta_i\beta_j k(x_i, x_j)$ ----------(7)

The points that lie on the cluster boundaries are defined as support vectors (SVs), which satisfy the conditions ξ $j = 0$ and $0 < \beta_j < C$. The points with $\xi_j > 0$ and $\beta_j = C$ lie outside the boundaries and are called bounded support vectors (BSVs). The rest of the data points lie inside the clusters. Note that the increase of the Gaussian kernel width parameter σ can increase the number of SVs, therefore causing the contours to change shape. By iteratively decreasing (increasing) σ from a certain large (small) value, SVC can form agglomerative (divisive) hierarchical clusters [1].

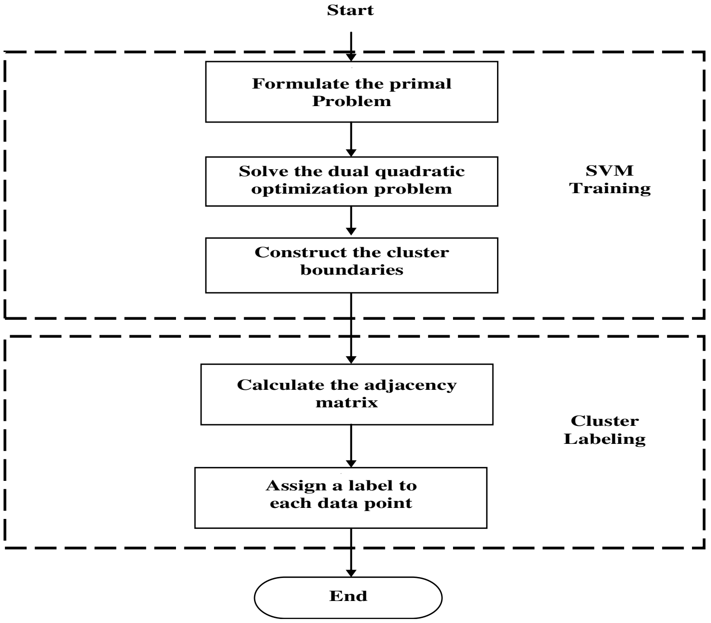
Start
Formulate the primal Problem
Solve the dual quadratic optimization problem
Construct the cluster boundaries
SVM Training
Calculate the adjacency matrix
Assign a label to each data point
Cluster Labeling
End

Figure 1.6: Flowchart of SVC algorithm [3]

The data points are clustered together according to the adjacency matrix A, which is based on the observation that any corresponding path in the feature space, which connects a pair of data points belonging to different clusters, must exit from the hypersphere. Given each pair of x_i and x_j, their adjacency value is defined as

$A_{ij} = 1, \text{ if } R\,(x_i + \gamma\,(x_j - x_i)) \leq R,\ \gamma \in [0, 1]$ ---------(8)

0, otherwise.

1.2Ant Colony Optimization

Dorigo et. [5] Al adopted this concept and proposed an artificial colony of ants algorithm, which was called the Ant Colony Optimization (ACO) meta-heuristic [6], to solve hard combinatorial optimization problems.. The ACO has been successfully applied to the optimization problems including data mining, telecommunications networks, vehicle routing, etc.

1.3 Nearest Neighbor Clustering

Since proximity plays a key role in our intuitive notion of a cluster, nearest neighbor distances can serve as the basis of clustering procedures. An iterative procedure was proposed in Lu and Fu; it assigns each unlabeled pattern to the cluster of its nearest labeled neighbor pattern, provided the distance to that labeled neighbor is below a threshold. The process continues until all patterns are labeled or no additional labeling occur. The mutual neighborhood value (described earlier in the context of distance computation) can also be used to grow clusters from near neighbors [5].

2 Existing Method: SNN-SVC

Wang and Chiang's scheme [4] 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 training datasets, the computational effort for solving the optimization problem can be greatly decreased. To fulfill the idea, they first explore the shared nearest neighbor (SNN) algorithm [7, 8] to eliminate noise points. Subsequently, the concept of unit vectors [9] is employed to reduce the core points of clusters and to retain the data points near the cluster boundaries. Based on these two methods, they developed an efficient data preprocessing procedure for SVC to reduce the size of the training datasets without altering the cluster configuration of the datasets

2.1SNN-SVC: SNN Based Preprocessing Procedure for SVC

Solving the optimization problem and labeling the data points with cluster labels are time-consuming in the SVC training procedure. This makes using the SVC algorithm to process large datasets inefficient [4]. 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. Researchers 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. An idea of Wang and Chiang [4] 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 training datasets, the computational effort for solving the optimization problem can be greatly decreased. To fulfill the idea, Wang and Chiang first explore the shared nearest neighbor (SNN) algorithm [7, 10] to eliminate noise points. Subsequently, the concept of unit vectors [6] is employed to reduce the core points of clusters and to retain the data points near the cluster boundaries. Based on these two methods, Wang and Chiang developed an efficient data preprocessing procedure for SVC to reduce the size of the training datasets without altering the cluster configuration of the datasets [4].

2.2 Elimination of Noise Points by SNN Algorithm

A shared nearest neighbor (SNN) algorithm proposed by Jarvis and Patrick [7] first finds the nearest neighbors of each data point, and then computes the similarity between pairs of points in terms of how many nearest neighbors each pair of the data points shares. The SNN algorithm can help to eliminate noise and outliers, and to identify core points that are the representative points from the regions with relatively high densities [4].

2.3Elimination of Core Points by the Concept of Unit Vectors

After the SNN algorithm is performed, most of noise points or outliers are removed from the datasets. Wang and Chiang's data preprocessing procedure [4] does not significantly alter the final cluster configurations but can save the computational time of SVC. Therefore, they need to eliminate non-support vector data points, such as core points. To achieve the objective, they further propose a method based on the concept of unit vectors [9] to eliminate the core points and retain the representative data points that are near the cluster boundaries.

3 PROPOSED METHOD: ACO-SVC

We proposed a new data point subset selection method for finding similarity matrix for clustering without alteration of support vector clustering. The proposed data point subset selection method is based on ant colony optimization; ant colony optimization is very famous meta-heuristic function for searching/finding similarity of data. In this method, we have introduced continuity of ants for similar data points and dissimilar data points collect into next node. In this process, ACO finds optimal selection of data point subset. Suppose ants find data points of similarity in continuous root. Every ant of data points compares their property value according to initial data point set.

When deciding data is noise and outlier, we should consider two factors: importance degree and easiness degree of noise and outliers. While walking ants deposit pheromone on the ground according to importance of the outlier and follow, in probability pheromone previously laid by other ants and the easiness degree of the noise.

Let D be the dataset and m be the number of ants, importance degree a_1, a_2,…., a_n is c_1, c_2 ,c_3 ……………..c_n, the appetency of solutions searched by two ants is defined as

$$\text{App (i, j)} = \frac{1}{ci - cj} \; \frac{1}{Ci - Cj} \qquad \text{……..(1)}$$

where c_i and c_j is the importance of noise and outlier path. The concentration of the solution (1) is defined as

$$\text{Con (i+j)} = \frac{\delta i + \delta j}{m} \frac{\delta i+\delta j}{m} \qquad \text{……..(2)}$$

where δ_i and δ_j is the number of ants whose appetency with other ants is bigger than α; α can be defined as m/10, then the incremented pheromone deposited by ants is

$$\Delta\tau_i = Q.\beta_i / \text{Con (i+j)} \qquad \text{……..(3)}$$

where Q is constant.

Each level of pheromone modeled by means of a matrix τ where $\tau_{ij}(t)$ contains the level of pheromone deposited in the node i and j at time t, ant k in node i will select the next node j to visit with probability,

$$P_{ij}^{k}(t) = \begin{cases} \dfrac{[\tau_{ij}(t)]^{\alpha}[\eta_{ij}]^{\beta}}{\sum_{l\varepsilon J_i^k}[\tau_{ij}(t)]^{\alpha}[\eta_{ij}]^{\beta}} & \text{if } j\varepsilon J_i^k \\ 0 & \mathit{otherwise} \end{cases} \qquad \text{…….. (4)}$$

where η_{ij} represents heuristic information about the problem which can be defined as the easiness of the path. The heuristic desirability of traversal and edge pheromone levels are combined to form the so-called probabilistic transition rule is given in equation (4), denoting the probability of an ant at data point i choosing to travel to data point j at time t.

Direct search in the best solution need global update rule applied as:

$$\tau\,(t+1) = (1-\rho).\ \tau_{ij}\,(t) + \rho.\Delta\ \tau_{ij} \qquad \text{……..(5)}$$

where *$\rho(0<\rho\leq 1)$* is a parameter that control the pheromone evaporation.

The steps of the proposed ACO based data preprocessing procedure for SVC (ACO-SVC) are as follows:

Step 1: Initialization of ants and degree of importance for the acceptance of data point selection: The appetency of solutions searched by two ants is defined as

$$\text{App (i, j)} = \frac{1}{Ci - C_j} \qquad \text{……..(1)}$$

where c_i and c_j is the importance of noise and outlier path.

Step 2: Find the acceptance solution on given parameter of degree of acceptance: The concentration of the solution (3.1) is defined as

$$\text{Con (i+j)} = \frac{\delta i + \delta j}{m} \qquad \text{……..(2)}$$

where δ_i and δ_j is the number of ants whose appetency with other ants is bigger than α; α can be defined as m/10, where m is the number of ants.

Step 3: Check the acceptancy of the data point and update the value of pheromone with amount of $\Delta\tau_i$: the incremented pheromone deposited by ants is

$\Delta\tau_i = Q.\beta_i / \text{Con(i+j)}$ ……..(3) where Q is constant.

Step 4: Generate the data point selection matrix after the increment of pheromone value and selected data points: Each level of pheromone modeled by means of a matrix τ where $\tau_{ij}(t)$ contains the level of pheromone deposited in the node i and j at time t, ant k in node i will select the next node j to visit with probability,

$$P_{ij}^{k}(t)=\begin{cases}\dfrac{\left[\tau_{ij}(t)\right]^{\alpha}\left[\eta_{ij}\right]^{\beta}}{\sum_{l\varepsilon J_i^k}\left[\tau_{ij}(t)\right]^{\alpha}\left[\eta_{ij}\right]^{\beta}} & \text{if } j\varepsilon J_i^k \\ 0 & otherwise\end{cases} \qquad \text{……..(4)}$$

where η_{ij} represents heuristic information about the problem which can be defined as the easiness of the path (η_{ij} is the heuristic desirability of choosing data point j when at data point i), J_i^k is the set of neighbor nodes of node i which have not yet been visited by the ant k. $\alpha > 0$, $\beta > 0$ are two parameters that determine the relative importance of the pheromone value and heuristic information, and $\tau_{ij}(t)$ is the amount of virtual pheromone on edge (i,j).

Step 5: Iterate and check data point matrix for processing of SVC mapping: Direct search in the best solution need global update rule applied as:

$$\tau\,(t+1) = (1\text{-}\rho).\ \tau_{ij}\,(t)\,_{+}\ \rho.\Delta\ \tau_{ij} \qquad \text{……..(5)}$$

where $\rho(0<\rho\underline{<}1)$ is a parameter that control the pheromone evaporation.

Step 6: Finally, data point matrix is passed to SVC algorithm for obtaining final clustering results.

4. RESULT AND DISCUSSION

We have used six datasets namely Wisconsin breast cancer (original), Iris, Glass identification, Yeast, Wine quality, and Page blocks classification dataset which are taken from UCI Machine Learning Repository [11]. Experimental result demonstrates that the proposed ACO-based data preprocessing procedure reduces the execution time of SVC procedure. The experimental results also demonstrate that the proposed ACO-SVC method is more robust than SNN-SVC. As classification accuracy of ACO-SVC is better than SNN-SVC, it shows that ACO based data preprocessing procedure is able to deal with noise and outliers precisely than SNN-SVC.

4.1 Performance Parameters

4.1.2 Execution Time

The proposed preprocessing procedure eliminates noise, outliers, and redundant points from the dataset. Hence, the size of the dataset is reduced and then it is passed to SVC algorithm. Thus, increase in effort due to preprocessing is negligible when compared to decrease in effort at later stages. Proposed algorithm decreases execution time significantly for small as well as large dataset.

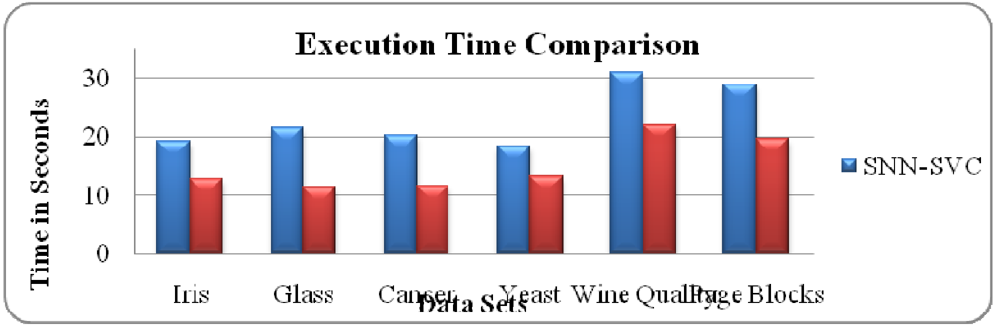

Figure 4.2: Execution Time Comparison Graph of SNN-SVC and Proposed Algorithm

Dataset	SNN-SVC	Proposed Algorithm	Difference
Iris	19.67 sec	12.89 sec	6.78 sec
Glass	21.38 sec	11.84 sec	9.54 sec
Cancer	20.28 sec	11.65 sec	8.63 sec
Yeast	18.31 sec	13.28 sec	5.03 sec
Wine Quality	30.87 sec	22.13 sec	8.74 sec
Page Blocks	28.81 sec	19.62 sec	9.19 sec

Table 4.8: Execution Time Comparison of SNN-SVC and Proposed Algorithm

4.2.3 Robustness

Robustness is one of the performance parameters criteria for clustering methods. Real world datasets are generally noisy containing errors or outliers. Data preprocessing is essential to remove noise points, or outliers from the dataset. Robustness of clustering method can be defined as its ability to deal with noise and outliers. From experimental results it is obvious that our proposed ACO based data preprocessing procedure is more robust than SNN based data preprocessing procedure as data passed to SVC by proposed data preprocessing procedure is more noise and outlier free. Hence accuracy of SVC is increased and execution time required for SVC procedure is decreased as compared to SNN-SVC.

4. CONCLUSION

This proposed research work focuses the drawbacks of SVC for dealing with large datasets. SNN based data preprocessing procedure for SVC results in loss of data. To overcome these drawbacks ACO based data preprocessing procedure is used to eliminate noise and irrelevant data from the dataset. The proposed data preprocessing procedure reduced the size of the dataset significantly which results in decrease in execution time of 13.30% required by SVC. That is the proposed preprocessing procedure reduces noise and irrelevant data from the dataset.

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

[3] R. X. Donald, C. Wunsch, “Clustering”, IEEE Press Series on Computational Intelligence, 2009, pp. 172-187.

[4] J. S. Wang, J. C. Chiang, “An Efficient Data Preprocessing 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. 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.

[7] 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

[8] L. Ertoz, M. Steinbach, and V. Kumar, " A New Shared Nearest Neighbor Clustering Algorithm and its Applications", Workshop on Clustering High Dimensional Data and its Applications, Second SIAM International Conference on Data Mining, Arlington, VA, 2002.

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

[10] 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.

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