ISSN No. 0976-5697

Volume 7, No. 6(Special Issue), November 2016

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

Support Vector Clustering with RBF Gaussian Kernel Parameter Optimization

Madhurbain Singh, Husanbir Singh Pannu

Thapar Institute of Engineering and Technology, Patiala India

Email: mbs22madhur@gmail.com

Abstract- Grouping the data according to their mutual similarities is called clustering which is an unsupervised technique. We have introduced an improved version of Support Vector Clustering (SVC) technique with automatic parameter tuning in this paper. Using Gaussian RBF kernel, the data points are mapped into a higher dimensional feature space. The objective function becomes to find the minimal enclosing sphere for all data points in the kernel space. This sphere, when mapped back to the data space, separates into several components with irregular boundaries, enclosing separate clusters of points. There are two tuning parameters in SVC, soft margin penalty constant and width of the Gaussian kernel which are varied and used to attain smooth cluster boundaries. It is difficult to accurately select both of these parameters manually using k-fold Cross Validation. So, we have employed an optimization technique to find out RBF sigma parameter tuning which was originally proposed by Cheng Hsuan Li et.al for Support Vector Machines (SVM). The optimal kernel RBF sigma value for SVC is calculated which is more accurate and computationally efficient and accurate compared to k-fold cross validation method. We have performed experiments on artificial and public datasets from UCI public repository. Promising outcome results show that this technique is effective in kernel parameter optimization for cluster analysis.

Keywords - Support Vector Machine, RBF kernel, cross validation, unsupervised learning

1. INTRODUCTION

Collection of unlabelled datasets with similar and dissimilar attributes is often analyzed using clustering technique. The data and objects present in the same cluster are all but similar to each other when compared to other clusters. So, the goal is to create homogeneous groups of the data objects. Major research areas of clustering are exploratory data mining and analyzing data statistically in the field including Machine Learning, recognition of different patterns, analyzing images, retrieval of information, bioinformatics, data compression, and computer graphics. Clustering has an application in anthropology which was derived by Driver and Kroeber in 1932, also A Psychologist Zubin found itsapplication in psychology [1][2]. It was also applied by Cattell in the 1943[3] for trait theory classification in personality psychology.

There are different types of clustering algorithms such as: connection based clustering, centroidal based, density based, distribution based clustering. In connectivity based clustering, the cluster indices are decided based upon relative distances among the objects. It is also known as hierarchical clustering. In centroidal based clustering, clusters are represented by centroidal vectors. For given k clusters, k centers are found and then individual data points are assigned respective cluster indices based upon their relative distances from those centroids. Distribution based clustering involves grouping the data objects depending upon their underlying data distribution characteristics. For example an Expectation Maximization (EM) algorithm works well for the Gaussian distribution because it uses Gaussian model for clusters. Density based clustering involves grouping of the data which has more density than rest of the dataset. Density based spatial clustering of applications with noise (DBSCAN) ismostly used algorithm of this category.

External Evaluation

It is defined as when the result of clustering is evaluated by the data that was not used in clustering. The Rand Index can be computed by the following formula:

$$RI = \frac{TP + TN}{TP + FP + FN + TN}$$

where TP = True positive, TN = True Negative, FP=False Positive, FN = False Negative

In recent years, a modern clustering technique has been developed, known as Support Vector Clustering (SVC) [4]. It is quite efficient in detecting the clusters with irregular boundaries using non linear mapping of data points into the higher dimensional feature space using Gaussian kernel [2]. In the higher dimensional feature space, the objective is to include maximum data points into the minimum enclosing sphere. This sphere when mapped back into the original space gets split into different irregular boundary contours which enclose the data points known as clusters. The performance of SVC is based on parameters used for Gaussian kernel and soft margin penalty (C) [6][9]. Selecting these parameters manually by method of k-cross validation method (CV) is very

978-93-85670-72-5 © 2016 (RTCSIT)

CONFERENCE PAPER

International Conference on

Recent Trends in Computer Science & Information Technology (RTCSIT-2016)

21st August 2016

Guru Nanak College Budhlada, Punjab India

tiring and for large data sets it is almost impossible to get an ideal result. In this paper an automatic technique for tuning of the RBF kernel function parameters is proposed.

The paper is organized in 6 sections. Section 2 explains about related works, section 3 discusses SVC technique, section 4 explains about proposed automatic tuning for SVC, section 5 explain experimental results and section 6 discusses conclusion and future works.

Internal evaluation

When the result of clustering is decided by the data used inside the cluster itself is called internal evaluation. This type of clustering is helpful when different clusters are very different from each other. The Davies–Bouldin index can be evaluated by the following formula:

where n, c_x,$\boldsymbol{\sigma}(x)$, $d(c_i,c_j)$ represent the count of clusters, centroid of x, average distance between all elements in x to centroid(c_x) and distance between c_i and c_j respectively.

2. RELATED RESEARCH

There have been studies on the automatic selection of kernel function parameters recently [5]. It was used in SVM for automatic parameter tuning successfully. In [8] a clustering approach based upon Information Maximization clustering based upon tuning parameter selection is discussed. It is based upon squared-loss variant of mutual information and eigenvalue decomposition. Another advanced method for automatic parameter tuning is used in [13] in context of SVM in which the RBF kernel function sigma value is calculated through optimization. In this research, we propose this automation for clustering technique for automated tuning of parameters in SVC.

3. SUPPORT VECTOR CLUSTERING

The objective is to define a sphere of the smallest radius that can enclose the given data points in the higher dimensional RBF kernel feature space mapped by . Thus the goal is to minimize $R^{2,}$ where R is radius of sphere with respect to constraints

Setting up the Lagrange function using the multiplier are defined as follows:

To find out the dual objective function, first we need to minimize *L.* So calculating the partial derivatives of *L* with respect to *R, €* and *a* we get:

(3)

(4)

(5)

Applying the Karush–Kuhn–Tucker (KKT) [14] conditions we get complementary equation

(6)

$= 0$

(7)

By using Eq. (7) it can be seen that the point x_i lies outside the feature space only if $€_i>0$ and $\beta_i>0$. From Eq.(6) we can say that $\mu_i=0$ as $€_i>0$, hence from Eq. (5) we can conclude that $\beta i = C$. This is called BSV (Bounded support vector). When $€_j=0$ any point x would be marked on the surface or inside the feature space. From Eq.(7) it can be said that its image $Ø(x_i)$ is present on the boundary of the feature space under the constraint $0 < \beta i < C$. This point is known a support vector (SV). So, now we can say that Support Vectors(SV) lie on the boundary whereas the boundary support vectors(BSV) lie outside the boundary and the rest lie inside the boundary line. There exists no BSV's due to constraint 3, provided that $C \geq 1$. Using the above relations we eliminate *R, a* and μ_j, and thus convert the Lagrange's form to the Wolfe's Duals

$$W = \sum Ø(x_j)2\beta_j - \sum \beta_i\beta_j Ø(x_i)Ø(x_j) \quad (8)$$

Using Constraint,

$$0 \leq \beta_j \leq C,\ j=1, 2, 3\ldots\ldots..N \quad (9)$$

From non linear SVM extension we can convert the dot product of $Ø(xi)\cdot Ø(xj)$ by a certain Kernel function represented by $K(xi,xj)$. We will use the concept of kernel function in this research paper.

(10)

where is RBF width parameter. The Lagrangian function *W* is defined as:

The image distance of each point x from the center of the sphere created in the feature space is represented by:

(12)

According to Eq.(4) And using kernel function :

(13)

Sphere's radius is given by R

$$R = \{R(x_i) | \ x_i \ is \ SV\} \quad (14)$$

The boundary that encloses the points in the data space is defined by the set

$$\{x | \ R(x) = R\} \quad (15)$$

They are interpreted by us as forming cluster boundaries. In view of equation (14), SVs lie on cluster boundaries, BSVs are outside, and all other points lie inside the clusters.

Cluster Assignment Phase

The theory which we have discussed beforehand does not differentiate between points of different clusters. To do this we use a geometrical theory in which we will include *R(x),* now by using the following observation: we have been given a pair of data points which are from distinct clusters; so any of the paths which connect them has to exit from the sphere in feature space. Thus, a pathway containing a segment of the points **y** such that $R(\mathbf{y}) > R$ should lie within the sphere. We end up with the definition of the adjacency matrix B_{ij} among the points x_i and x_j whose images lie inside or on the boundary of the sphere in feature space:

Because of this square matrix representation during this SVC's cluster assignment phase, it becomes a bottleneck during the calculations for the space and time complexity. So SVC is efficient to detect quite irregular cluster boundaries for smaller datasets only.

Comparison with K-means

Now we briefly discuss an important existing clustering method known as K-means clustering. This would help to contrast the features of SVC over other conventional clustering methods. In recent times there have been detailed studies on them [5]. K-means Method is very simple unsupervised learning algorithms which help us solve a well-known problem of clustering. In this we first differentiate between a given data set through a definite number of clusters (k clusters) and then name k centers, one for each cluster. These centers need to decide in such a way that each location or place causes un-similar results. It is be useful to place them at a maximum possible distance from each other. Firstly, we need select each point from a defined data set connect it to the closest k-center. When all the points have been connected the early part of our job is done. Now, we need to find out new k-centroids as barycenter of the clusters found earlier. Now that we have found the new k-centroids we need to establish a new binding between the data set and the closest centers. As a result a loop was produced. After this we notice that the k-center have changed their location, now as no more changes are done we observe that centers do not move anymore. Finally, using this algorithm we aim to minimize an objective function (squared error function):

where,

'$||x_i - v_j||$' is the Euclidean distance between x_i and v_j.

'c_i' is the number of data points in i^{th} cluster.

'c' is the number of cluster centers.

Limitations of K-means

1. This method of clustering is not robust to non-linear data distribution. It gives different results with different representations. This is represented in the form of coordinates.
2. To learn this algorithm we require apriori specification of the number of k's i.e. the cluster centers.
3. Overlapping data cannot be resolved by k-means method.
4. It only is defined for a definite mean which means it will fail for categorized data.
5. A Cluster center cannot be chosen randomly.
6. This algorithm is not applicable for non-linear data set.
7. This algorithm is not able to handle noisy data as well as outliers.

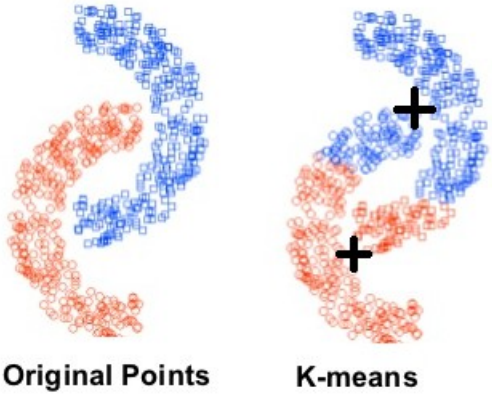

Fig 2: k-means method fails for non-linear data set.

978-93-85670-72-5 © 2016 (RTCSIT)

CONFERENCE PAPER
International Conference on
Recent Trends in Computer Science & Information Technology (RTCSIT-2016)
21st August 2016
Guru Nanak College Budhlada, Punjab India

Due to the above disadvantages, we choose SVC and employ automatic parameter tuning to make it self-dependable clustering algorithm. It would take care of above limitations and there is accurate selection of Kernel Parameters.

4. PROPOSED ALGORITHM FOR RBF KERNEL OPTIMIZATION

Using this algorithm we propose that one should be able to find out the most ideal parameters for our Kernel Function used in SVC. This theorem was previously used in Support Vector Machines to for automatic tuning of parameters of Kernel Function [5], but in this research paper we try to apply it on SVC.

Let us assume that A_i is the set of training samples in the set i, where $i=0,1,2\ldots\ldots L$.[11].There are two important properties of the RBF kernel function: $K(x_i,x_i,\sigma) = 1$ for all $i=1,2\ldots\ldots\ldots n$, i.e., the norm of every sample in the feature space is 1. $0<K(x_i, x_j, \sigma)\leq 1$ for all i,j=1,2…….,n., i.e., the cosine value of two training samples and in the feature space can be computed by and it determines the similarity between these two samples. RBF Kernel function:

Where $||w_i||$ is the number of training samples in class i. The parameter should be determined such that closes to 1.It is easy to find that $0 < w(\sigma) \leq 1$ and . Hence, the optimal σ can be obtained by solving the Following optimization problem:

5. Dataset Results

Quantitative testing of the proposed technique is performed on an artificial and publically available datasets. The artificial dataset contains 200 points in two dimensions as shown below in Fig 3. Using the parameter automation approach [13] we got three clusters detected as shown in Fig. 3. Values of parameters are listed in the Table 1. Fig.4 shows the optimal RBF sigma calculated through [13] for the artificial dataset. gives the minimum value of .

The experimental results have also been illustrated for other available at UCI repository [11]. In all of these datasets, k-means clustering fails to detect the appropriate clusters but the proposed technique achieves good results as shown in the Table 1. We have performed the experiment on ten other standard datasets and obtained results with high accuracy through automatic parameter tuning but only listing five of them due to space constraints.

The automation of RBF sigma [11] saves a great deal of time complexity as compared to traditional Cross Validation (CV) technique. The overall average clustering accuracy from comes out to be 90% which makes it well suitable to detect clusters of irregular boundaries. K-means algorithm fails to detect such irregular clusters successfully.

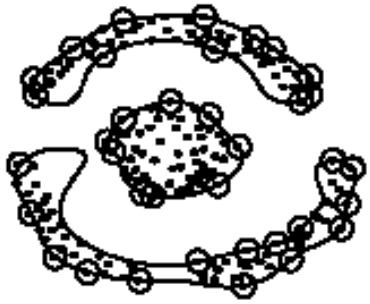

Fig 3: Artificial dataset with 200 points using proposed technique detected 3 clusters, with $\sigma = 0.15$

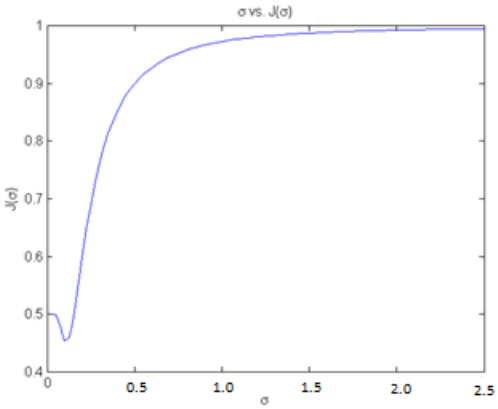

Fig 4: RBF sigma value calculated through [13] for the artificial dataset

Table1: Clustering results on three datasets where K-means clustering fails

Dataset	Instances	Sigma	C	Accuracy
IRIS	150	0.28	0.1	0.86
AAAI'13	150	0.11	40	0.92
Artificial	183	0.15	100	0.95
SEEDS	210	0.34	10	0.9
AAAI'14	399	0.45	12	0.87

6. CONCLUSION

RBF sigma parameter optimizations have been proposed in this paper for Support Vector Clustering. For small but irregular shaped datasets, SVC is quite efficient to detect clusters compared to other conventional algorithms such as k-means. Kernel

parameter in SVC is done using a Chen-Hsuan et al.'s technique which was proposed for SVM. Results show that clusters have been detected with high accuracy of average 89%. SVC although is a good algorithm for clustering, but it uses a square matrix during the cluster assignment phase. So it is not efficient for time and space complexity. In future we are planning to extend SVC for massive datasets by using advanced cluster assignment strategies.

7. REFERENCES

1. Bailey, Ken (1994). "Numerical Taxonomy and Cluster Analysis". *Typologies and Taxonomies*. p. 34. ISBN 9780803952591.
2. *Tryon, Robert C. (1939). Cluster Analysis: Correlation Profile and Orthometric (factor) Analysis for the Isolation of Unities in Mind and Personality. Edwards Brothers.*I.S. Jacobs and C.P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G.T. Rado and H. Suhl, Eds. New York: Academic, 1963
3. Cattell, R. B. (1943). "The description of personality: Basic traits resolved into clusters".*Journal of Abnormal and Social Psychology* **38**: 476–506. doi:10.1037/h0054116.
4. *Asa Ben-Hur, David Horn, Hava T. Siegelmann, Vladimir Vapnik*; 2(Dec):125-137, 2001.M.
5. Mingyuan Zhao,; Ke Tang,; Mingtian Zhou,; Fengli Zhang,; Ling Zeng, "Model parameter selection of support vector machines", *Cybernetics and Intelligent Systems, 2008 IEEE Conference on,* On page(s): 1095 – 1099
6. An Efficient k-means Clustering Algorithm: Analysis and Implementation by Tapas Kanungo, David M. Mount,Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman and Angela Y. Wu.
7. Ortega, Joaquín Pérez, Ma Del Rocío Boone Rojas, and María J. Somodevilla. "Research issues on K-means Algorithm: An Experimental Trial Using Matlab."
8. Wu, Xindong, et al. "Top 10 algorithms in data mining." Knowledge and information systems 14.1 (2008):1-37.
9. Teknomo, Kardi. "K-means clustering tutorial." Medicine 100.4 (2006): 3.k-means clustering by kechen.
10. Chapelle, Olivier, et al. "Choosing multiple parameters for support vector machines." *Machine learning* 46.1-3 (2002): 131-159.
11. Bache, K., and M. Lichman. "UCI Machine Learning Repository [http://archive. ics. uci. edu/ml]. University of California, School of Information and Computer Science." *Irvine, CA* (2013).
12. Sugiyama, Masashi, et al. "On information-maximization clustering: Tuning parameter selection and analytic solution." *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*. 2011.
13. Li, Cheng-Hsuan, et al. "An automatic method for selecting the parameter of the RBF kernel function to support vector machines." Geoscience and Remote Sensing Symposium (IGARSS), 2010 IEEE International. IEEE, 2010.
14. Gordon, Geoff, and Ryan Tibshirani. "Karush-kuhn-tucker conditions."*Optimization* 10.725/36 (2012): 725.

978-93-85670-72-5 © 2016 (RTCSIT)

CONFERENCE PAPER
International Conference on
Recent Trends in Computer Science & Information Technology (RTCSIT-2016)
21st August 2016
Guru Nanak College Budhlada, Punjab India