A Support Vector Method for Clustering

AsaBen-Hur[1]
https://www.cs.colostate.edu/~asa/
David Horn[2]
https://www.lifesci.dundee.ac.uk/people/david-horn
Vladimir Vapnik[3]
https://datascience.columbia.edu/people/vladimir-vapnik/
Hava T. Siegelmann [4]
https://www.cics.umass.edu/faculty/directory/siegelmann_hava

[1] Israel
[2] USA

Clustering is an important unsupervised machine learning technique. Clustering algorithm divides the data points into number of groups on the basis of their similarities.There two major types of clustering algorithms , parametric and non parametric.Parametric algorithms are usually limited in their expressive power.In this paper the authors have defined a non parametric algorithm based on support vector machine which is much more flexible than the other models.

Support Vector machines create a unique hyperplane having maximum margin to separate the data.For non linear datasets support vector machine uses the kernel trick.Kernel functions map the non-linearly separable data points from their input space to a higher dimensional feature space,where they might be linearly separable.This text uses the kernel trick of SVM algorithm to form clusters of the data points in the input space.The kernel used here is Gaussian Kernel.Gaussian kernel function maps the datapoints from input space to a high dimensionsal feature space,it computes a sphere of minimal radius enclosing the datapoints in the feature space.The procedure to find minimal radius sphere is called Support Vector Domain Description (SVDD) .This sphere is mapped back to the input space and it forms contours that encloses the data in input space.When datapoints are enclosed by the contours, the points belonging to the contours belong to the same cluster.

Mathematically, a dataset x_i of N points is considered in R^d input space.ϕ is defined as the non linear transformation of the dataset to a higher dimensional feature space.R is radius of the hyper sphere with centre at some point a in the feature space.For the data points to be enclosed in the sphere ,this condition is to be satisfied.

$$||\phi(x_i) - x_i|| < R^2 + \varepsilon \tag{1}$$

To solve this problem Lagrangian is introduced.

$$L = R^2 - \sum_j (R^2 + \varepsilon_j - ||\phi(x_j) - a||^2)\beta_j - \sum_j (\varepsilon_j \mu_j) + C\sum_j (\varepsilon_j) \tag{2}$$

In equation (1) ε is the slack variable which calculates how much an outlier lies outside the concentrated data distribution.C is penalty for Outliers. In equation (2)β_j and μ_j are lagrangian multipliers.On setting derivatives with respect R ,a and ε equal to zero and solving further ,following constraints are obtained. Any point with 0<β_j<C is referred to as support vectors and points with β_j=C are referred to as bounded Support Vectors. Any point with $\varepsilon > 0$ is outside the feature space sphere. It is very computationally expensive to map each point of the data to the feature space therefore the kernel functions are represented as the inner product of some function at the two points x_i andx_j Now the kernel function

$$K(x_i, x_j) = < \phi(x_i), \phi(x_j) > \tag{3}$$

Gaussian kernel is defined as

$$K(x_i, x_j) = e^{-q||x_i - x_j||^2} \tag{4}$$

In the Gaussian Kernel q is the width parameter and in the Lagrangian equation C is Regularization parameter which penalizes the outliers. This quadratic programming problem of cluster description algorithm is convex and has globally optimum solution. When C=1 (no outlier) , on increasing the value of q (width parameter of Gaussian Kernel) the shape and number of cluster changes.Since the Gaussian Kernel is such that K(x,x)=1,the hypersphere defined is a unit sphere. A point x_i is considered to be bounded support vector if $R(x_i) > R$. Number of Support vectors and bounded support vectors also play a crucial role in defining boundaries of the clusters.As C is decreased not only does the number of bounded SVs increase, but their influence on the shape of the cluster contour decreases . As the number of bounded support vector increase the probability of overfitting and overlapping of clusters decreases. The number of support vectors depends on both q and C. For fixed q, as C is decreased, the number of SVs decreases since some of them turn into bounded SVs and the resulting shapes of the contours become smoother. According the Proposition(2.1) mentioned in the paper

$$n_{bsv}(q, C) = max(0, 1/C - n_0) \tag{5}$$

$$n_{sv}(q, C) = a/C + b \tag{6}$$

where a, b, n_0 are functions of q and N

The clusters formed in input space are distinguished according to the gap in the support of the underlying probability distribution of data points.As q is increased the support is characterized by more detailed features, enabling the detection of smaller gaps.A continuous increase in q value may lead to overfitting of data. In many datasets clear separating boundaries of clusters are not visible,in such cases a slightly different approach is considered.The contour will enclose a small number of points which correspond to the maximum of the Parzen-estimated density. Hence in a high bounded Support vectors system we find a dense core of probability distribution. The second phase of SVC is cluster assignment in which an adjacency matrix $A(i, j)$ is computed with values 0 and 1.The values 0 and 1 determine whether the line segment joining two points x_i and x_j is outside or inside the hypersphere. SVC depends upon variability of parameters q and C.The value of q is initially taken small and slowly increased until a meaningful number of clusters are formed and there is no overfitting in the input space.

In this paper the algorithm is implemented on the very famous iris dataset in which authors have tried to cluster different classes.A table showing how the value of q and C and the number of support vectors , bounded support vectors affect the clustering algorithm is created.From this dataset it was concluded that as the dimensionality of the data increases a larger number of support vectors is required to describe the contours. Thus if the data is sparse, it is better to use SVC on a low dimensional representation. SVC is flexible with respect to parameters q and C it can deal with noise or outliers by a margin parameter.SVC has no explicit bias of either the shape or the number of clusters.This overall makes it a robust clustering algorithm

Paper link-*Please Click here*