ML UPDATE

Support Vector Machine is undoubtedly a very powerful algorithm because of its ability to separate non-linearly separable data using the kernel trick and its high classification accuracy. This non-parametric clustering algorithm is modeled using the Support Vector Machine (SVM). Even though in general SVM is a robust algorithm but its efficiency decreases when the size of the dataset used is large because while calculating the hyperplane it considers almost each and every point present in the dataset. Because of this, it becomes computationally very expensive to use SVM over large datasets. Support Vector machines cannot work with datasets having a more number of dimensions compared to a total number of samples. Clustering, an unsupervised learning algorithm, is very useful in real-life applications, and hence in most applications, large datasets are used; therefore it is essential to properly preprocess the data before using Support Vector Clustering. Principal Component Analysis is one of the ways to reduce the dimensionality of a dataset by transforming a large set of variables into a smaller one that still contains most of the information in the large set. Reducing the dimensionality reduces the accuracy of the model, but for simplicity and to reduce time complexity, this approach is very useful. Principle Component Analysis reduces the dimension of the dataset while preserving as much as information possible. Apart from Principal Component Analysis, autoencoders can also be used for extracting important features out of the dataset. Autoencoder is an unsupervised Artificial Neural Network that compresses the input dimensions to get reduced dimensions.

In this paper, the model described has two hyperparameters, namely q and C. The model defined is very sensitive to these parameters and does not perform well if values are not carefully selected. Even though the selection of q and C hyperparameters is very important, there is no proper method mentioned in this text. In another clustering algorithm which is K, means clustering. The algorithm identifies k number of centroids and then allocates every data point to the nearest cluster while keeping the centroids as small as possible. Initially, cluster points are selected randomly, and then iterative calculations are performed to optimize and update the centroids until they converge at some point. This algorithm takes the number of clusters(k) as one of the hyperparameters. The main challenge in this algorithm is to choose the optimal value of k. The Elbow graph is one of the ways to choose the value of optimal k.In the elbow, the idea is to calculate the Within-Cluster-Sum-of-Squares (WCSS) which is the sum of squares of the distance of each data point in all clusters to their respective centroids. Generally, a good clustering algorithm has a large intercluster distance between different clusters and a small intracluster distance. This elbow method can be used as one of the techniques to judge the values of q and C and how the model is performing at different values of q and C. Initially, q and C can be taken randomly but rationally, and using the elbow method, the hyperparameters can be tuned properly. Another method that is similar to the elbow method is the Silhouette method which takes into account the inter-cluster distances as well as intracluster distance and plots the values of the Silhouette coefficient at different values of k.

Support Vector clustering may sometimes overfit when the number of support vectors around a cluster is more than the bounded support vectors i.e. the algorithm is trying to fit all the points

(even outliers and noise) in some of the other clusters. This causes overfitting. Therefore the criteria for smooth boundaries without overfitting is that the number of support vectors should be less. In other words, we need to increase the value of q and C, which guarantees a minimum number of support vectors. This criterion can also be used to get a sense of direction while choosing the values of q and C in such a way that there is no overfitting. A function that calculates the number of SVs can be created, and a logical threshold for the number of Support Vectors can be calculated. A graph that shows how the number of support vectors varies with the value of q and C considered can give a better idea of whether the values of q and C should be increased or decreased.

This paper uses Gaussian kernel as the primary kernel for mapping the data points from input space to the higher dimensional feature space, instead of the Gaussian kernel if some other type of kernels like Radial Basis Kernel or polynomial kernel is used then the complexity of the model can be reduced and instead of q as a hyperparameter, we will have a degree of the polynomial as the hyperparameter which might prove fruitful where we have a large and complex dataset.

The Support Vector clustering algorithm can also be used in classification if hyperparameters are tuned carefully. Using Support Vector Clustering, we can obtain the optimal number of clusters. The obtained clusters can provide information about the region of the data points and their similarity with other points, this information can be readily utilized as features while classifying data points. Adding this feature to the dataset might increase the accuracy of the classification algorithm.

Support Vector Clustering algorithm is overall a robust clustering algorithm compared to other parametric algorithms which have specific shapes of clusters. There are drawbacks to this algorithm that can be overcome by using better optimization algorithms and other small techniques as proposed in the above ML update.