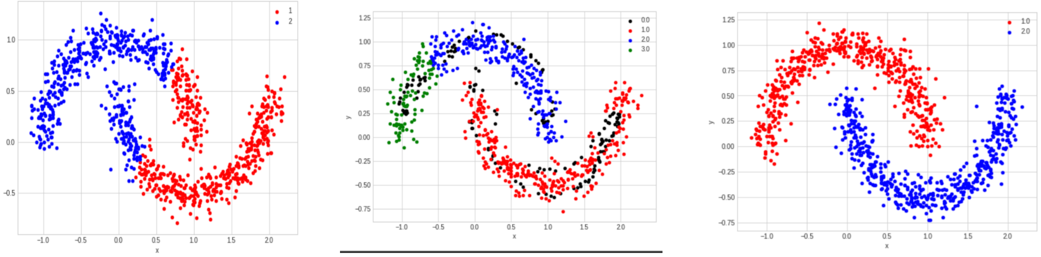
Clustering and SVM belong to different categories of Machine learning . Clustering techniques can be used in various areas or fields of real-life examples such as data mining, web cluster engines, academics, bioinformatics, image processing & transformation, and many more and emerged as an effective solution to above-mentioned areas.On the other hand SVM is one most robust supervised learning algorithm because of its convex optimization and high accuracy in most cases.

The paper ‘A Support Vector Method for Clustering’ written by AsaBen-Hur,David Horn, Hava T. Siegelmann and Vladimir Vapnik discuss how Support Vector Machine can be used for clustering .SVM is best known for its ability to separate linear and non -linear data points using the kernel trick.It is based on the ideology of finding a hyperplane with good margin that best separates the features into two domains.

In this paper the author proposes a model for clustering data points in the input space using Gaussian kernel.The basic idea used in this model is that ,the data points in the input space are mapped in the feature space(a higher dimensional space) where a sphere with minimal radius is calculated which encloses the data points in feature space.The sphere with radius R is then mapped back to the input space where it forms contours enclosing the data points.SVC uses Support Vector Domain Detection (SVDD) to find the hypersphere in the feature space.The Gaussian kernel has two key parameters which play a crucial role in forming required number of Clusters without overfitting and noise.q is the width parameter which basically assign a score proportional to the nearness of the query point to the support vector points.C ,regularization parameters, penalizes the outliers of data while forming clusters.As the width parameter in model increases ,keeping C=1 to neglect outliers for simplicity, it is observed that the number of clusters increases gradually.Whereas if C is increasing the algorithm starts to consider outliers while forming cluster which leads to overlapping clusters.Therefore it can be said that while forming apt clusters, hyper parameter should be tuned appropriately.After choosing suitable parameters an Adjacency matrix is calculated to label the cluster.

In the research paper the authors have taken the iris data,which is a standard benchmark in pattern recognition literature, to check the performance of the algorithm.Through the example is was proved that increase in the number of support vectors and bounded support vectors required to obtain contour splitting. 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 a non-parametric model which gives it more flexibility with respect to shape of cluster boundaries.

SVC gives better results than many other clustering algorithms like K-Means ,when hyperparameters are carefully tuned.The following graphs show the results of K Means Clustering and Support Vector Clustering on Two Moons dataset.The ideal result should be the two half moons separated as two different clusters.



(a) (b) ©

Figure (a) shows clustering using K-means algorithm ,Figure(b) shows clustering using SVC without carefully chosen hyper parameters; the black points in figure b represent the boundary of the clusters.The Figure(c) shows results obtained from SVC after proper tuning of hyper parameters.

Most of the other clustering methods like fuzzy based clustering,hierarchical clustering,density based clustering etc. However for a linearly separable data space, most of these methods require data with a regular distribution. These methods often have an unstable performance when extracting appropriate cluster boundaries and identifying the exact number of clusters, whereas SVC can generate arbitrarily shaped cluster boundaries and avoids artificially predefining the number of clusters.SVC outperforms conventional methods because it can describe clusters with irregular shape and imbalance distribution.The model proposed in this paper has ability to deal with outliers as well as noise .This model is less prone to overfitting if hyperparameters are properly tuned.Since it is a convex optimization quadratic problem ,convergence is guaranteed.

Even though this model performs better than many other clustering algorithms there are a lot of which are faced while implementing this model.When dealing with complicated data we need to separate the data points using very high dimensional space, but picking a kernel with more parameters/dimensions will be computationally more expensive.In the case of most real-world problems and strongly overlapping clusters, the SVM-Internal Clustering algorithm above can only delineate the relatively small cluster cores.Working in a higher-dimensional feature space increases the generalization error of support-vector machines.Support vector clustering algorithm works well in general, but its performance degrades when applied on big data. Even though hyperparameter tuning is very important ,the paper does not explain any proper method for finding the best q and C.

There is ongoing research on how to solve these bottleneck problems.Many researchers have proposed some global optimization strategies to systematically and optimally select parameter value.Since this model cannot deal with very large datasets properly ,algorithms for efficient cluster labeling and for efficient boundary detection of clusters are being proposed which can boost the performance of the original algorithm.Huina Li and Yuan Ping in their paper ‘Recent Advances in Support Vector Clustering:Theory and Applications’ introduce various complex and flexible algorithms to improve the performance of the basic SVC model.

Hence it can be concluded that non-parametric SVM-based clustering methods may allow for much improved performance over parametric approaches, particularly if they can be designed to inherit the strengths of their supervised SVM counterparts.