ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 228, 2009

*Michał Trzęsiok**

ON SOME PROPERTIES OF SUPPORT VECTOR CLUSTERING

Abstract. The aim of this paper is to analyse the relatively new clustering method – Support Vector Clustering (SVC) in terms of fulfilling admissibility conditions. The results are compared within a group of four other clustering methods.

Since it is not possible to assess which clustering method is the "best" in general, given a specific problem the user can decide which method to apply considering some properties of clustering methods, known as *admissibility conditions*. This paper expands the knowledge about the properties of clustering methods with the properties of SVC.

Keywords: support vector machines, clustering, admissibility conditions.

I. INTRODUCTION

The Support Vector Machines were introduced as a powerful tool for classification. They are also suitable for regression and novelty detection. There is a natural way of turning SVMs for novelty detection (i.e. in the case of one–class classification) into a clustering method (as proposed by Ben-Hur, Horn, Siegelmann and Vapnik in Ben-Hur *et al.* (2001)). The problem of novelty detection can be translated into the issue of finding the multi–dimensional quantile function. Using the kernel trick (a standard technique for the support vector approach), we can search for the smallest hypersphere enclosing the image of the data in the high–dimensional feature space. By setting the radial kernel parameter large enough, we can force the hypersphere to split into several components, when we map it back to data space. These components can be interpreted as clusters.

Although using the quantile estimation method we treat all the observations as representing only one class, it is possible to make the algorithm able to predict whether the given pair of observations belongs to the same cluster or not. It can be performed by checking whether there is a point from the line segment connecting that two observations, which lies outside the multi–dimensional quantile.

* M.Sc., Department of Mathematics, Karol Adamiecki University of Economics, Katowice.

It turns out that SVC is a very flexible method. It can handle clusters with very irregular shapes without the need to make any arbitrary assumptions about the number and the shape of the clusters.

There are many different clustering methods applicable in different situations. It does not seem possible to point to the one which outperforms the others. Having no information about the number and shape of clusters (which is usually the case), the choice of a clustering method can be based on the knowledge about the properties of a particular method.

In Section II the algorithm of the Support Vector Clustering is briefly presented. In Section III the definitions of selected properties, known as *admissibility conditions* of clustering methods are given. Then in Section IV the results of the analysis of properties of SVC are presented. Additionally the properties of other clustering methods are cited to enable the comparison and further conclusions.

II. THE OVERVIEW OF THE SVC ALGORITHM

The Smallest Enclosing Hypersphere

Following Ben-Hur *et al.* (2001) we present briefly the main ideas of the SVC methodology. Let $D = \{\mathbf{x}^1,\ldots,\mathbf{x}^N\}$, $\mathbf{x}^i \in \mathbf{R}^d$, $i \in \{1,\ldots,N\}$ be the data set of N points. First we transform data points to higher dimensional feature space using the nonlinear mapping $\varphi : \mathbf{R}^d \to \mathbf{Z}$. Then we find the smallest hypersphere enclosing the image of the data in this feature space. We denote the center of the hypersphere by $\mathbf{a}$ and its radius by R.

The problem of finding the smallest enclosing hypersphere of radius R can be written as an optimization task with ν–parametrization ($0 < \nu \leq 1$) and soft constraints as in Schölkopf and Smola (2002):

$$\begin{aligned} &\underset{R\in\mathbf{R},\mathbf{a}\in\mathbf{Z},\xi_i\geq 0}{\text{minimize}}\ R^2 + \frac{1}{\nu N}\sum_{i=1}^{N}\xi_i, \\ &\text{subject to}\ \left\|\varphi(\mathbf{x}^i)-\mathbf{a}\right\|^2 \leq R^2+\xi_i,\quad \xi_i \geq 0,\quad i\in\{1,\ldots,N\}. \end{aligned} \tag{1}$$

The solution of the problem can be found using Lagrange multipliers method. The dual form of the Lagrangian is:

$$\begin{aligned} &\underset{\alpha}{\text{minimize}} \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j \mathrm{K}(\mathbf{x}^i,\mathbf{x}^j) - \sum_{i=1}^{N}\alpha_i \mathrm{K}(\mathbf{x}^i,\mathbf{x}^i), \\ &\text{subject to } 0 \le \alpha_i \le \frac{1}{vN}, \quad \text{and} \quad \sum_{i=1}^{N}\alpha_i = 1, \quad i \in \{1,\ldots,N\}, \end{aligned} \tag{2}$$

Where $\mathrm{K}(\mathbf{u},\mathbf{v}) = \varphi(\mathbf{u}) \cdot \varphi(\mathbf{v})$ already denotes the *kernel function* representing the dot product in high dimensional feature space *Z*. The solution has the following form:

$$\begin{aligned} a &= \sum_{i=1}^{N}\alpha_i \varphi(\mathbf{x}^i), \\ R^2 &= \mathrm{K}(\mathbf{x}^s,\mathbf{x}^s) + \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j \mathrm{K}(\mathbf{x}^i,\mathbf{x}^j) - 2\sum_{i=1}^{N}\alpha_i \mathrm{K}(\mathbf{x}^i,\mathbf{x}^s), \end{aligned} \tag{3}$$

where $\mathbf{x}^s$ denotes any of the identified *support vectors*, i.e. observations corresponding to nonzero Lagrange multipliers $(\alpha_s > 0)$.

Now we use the derived hypersphere to define a decision function *f*:

$$f(\mathbf{x}) = \operatorname{sgn}\left(R^2 - \left(\mathrm{K}(\mathbf{x},\mathbf{x}) + \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j \mathrm{K}(\mathbf{x}^i,\mathbf{x}^j) - 2\sum_{i=1}^{N}\alpha_i \mathrm{K}(\mathbf{x}^i,\mathbf{x}) \right) \right). \tag{4}$$

This function will be used for cluster assignment in the next subsection.

Cluster Assignment

The hypersphere when mapped back to data space forms a set of contours. Points enclosed by each contour are associated with the same cluster and the contours are interpreted as cluster boundaries. Formally, the contours consist of all points $\mathbf{x}^i \in \mathbf{R}^d$ for which the decision function *f* equals zero ($f(\mathbf{x}) = 0$).

We still need to know how to distinguish two different clusters (because the data points are now enclosed by the contours but still not labeled). To do the labeling we first note that all points from the ball in the feature space correspond only to the points in data space enclosed by the contours. So if we connect two points $\mathbf{x}^i$ and $\mathbf{x}^k$ from two different clusters with a line segment $\overline{\mathbf{x}^i\mathbf{x}^k}$ we find $\mathbf{y} \in \overline{\mathbf{x}^i\mathbf{x}^k}$, that is not enclosed by any contour, which means that its image lies outside the ball in the feature space. With the use of the decision function

f defined in (4) we can easily check if the image of the given point $\mathbf{y}$ lies outside the ball because it is equivalent to checking whether $f(\mathbf{y}) = -1$. For the cluster assignment we check the line segments connecting every pair of points from the data set D by sampling the number of points. The results are stored in an *adjacency matrix* $A = [a_{ik}]_{i,k=1,\ldots,N}$:

$$a_{ik} = \begin{cases} 1, & \text{iff} \quad f(\mathbf{y}) = 1 \quad \underset{\mathbf{y} \in \overline{\mathbf{x}^i \mathbf{x}^k}}{\forall}, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Note that since the clusters are not necessarily *convex*, the value "0" in the matrix A does not mean that two corresponding points belong to two different clusters. Therefore clusters are not defined directly by the matrix A, but as the connected components of the graph induced by A.

Implementational Details

We performed SVC based on the svm(...) function implemented in **R** package e1071. However, this function is designed only for supervised classification, regression and one-class classification (novelty detection). Nevertheless, Remark 1 allows to apply this function to clustering:

Remark 1. It can be shown (as in Schölkopf and Smola (2002)) that the use of the RBF kernel $K(\mathbf{u}, \mathbf{v}) = \exp(-\gamma \|\mathbf{u} - \mathbf{v}\|^2)$ (or any other translation invariant kernel) makes the problem of finding the smallest enclosing *hypersphere* equivalent to the task of finding the optimal *hyperplane* separating the image of the data points from the origin in the feature space.

Remark 1 indicates that when we use the RBF kernel, the first part of the SVC algorithm – the identification of the optimal hypersphere enclosing the image of the data – can be performed using the function for one-class classification. It is possible since the function svm(..., type="one-classification") derives the optimal hyperplane separating the image of the data points from the origin and this is equivalent to finding the hypersphere. Based on the results of the svm(...) function we developed the **R** code responsible for the second part of the SVC algorithm i.e. for the cluster assignment part.

III. DEFINITIONS OF SELECTED PROPERTIES OF CLUSTERING METHODS

The choice of the clustering method can be based on the knowledge of the properties of a particular method. These properties are known as *admissibility conditions* (see Fisher and Van Ness (1971)). Let us briefly present the definitions.

Image. The algorithm is said to fulfil the *image* admissibility condition when the result of the clustering does not change if the observations in a data set are permutated (the algorithm is independent from data points order).

Convex. The algorithm is said to be *convex admissible* if the convex hulls of the identified clusters are disjoint.

Well-structured. The algorithm is said to be *well-structured* if all within cluster interpoint distances are smaller than all between cluster distances.

Repeatable. This condition applies only to these algorithms which have corresponding discriminant analysis algorithms. The algorithm is said to be *repeatable admissible* if for all $\mathbf{x}$ from the data set D the cluster assignment produced by the clustering method is the same as the prediction obtained for $\mathbf{x}$ after performing the corresponding discrimination method on the data set $D \setminus \{\mathbf{x}\}$, where $\mathbf{x}$ was removed and class labels were taken as results of the clustering. So if the point $\mathbf{x}$ is always put back in its original cluster by the corresponding discriminant method, the algorithm is said to be repeatable admissible.

Cluster omission. The algorithm is said to fulfil *cluster omission* condition if the cluster boundaries resulting from performing the clustering on a whole data set are the same as the ones obtained from clustering on a reduced data set, where one of the previously identified clusters was omitted (of course the cluster boundaries should be the same but for the omitted cluster).

IV. RESULTS

Experiments were conducted on artificial benchmark data sets smiley, circle, twonorm, spirals from **R** package mlbench.

The simulations showed that SVC is repeatable admissible. First, clustering was performed using SVC. Having the classes labeled, we performed discrimination analysis using SVM on training set $D \setminus \{\mathbf{x}\}$. Then we asked the SVM model to predict the class for $\mathbf{x}$ and we observed that SVM put every point $\mathbf{x}$ back to the cluster which this point was taken from.

The models resulting from applying the Support Vector technique are defined by the kernel function (RBF kernel used in all the experiments) and the identified *support vectors*. Therefore, comparing SVC with the image admissi-

bility condition it was enough to check if the set of support vectors had changed. The experiments confirm that SVC fulfils this condition.

Since it turned out that the SVC does not meet the three other admissibility conditions we provide simple counterexamples to prove it. We used the data set smiley since it is simple and suitable as a counterexample for all three admissibility conditions.

Fig. (1) presents a counterexample to convex admissibility of SVC. Fig. (2) shows that SVC is not well-structured. Fig. (3) illustrates that SVC does not fulfil the cluster omission admissibility condition.

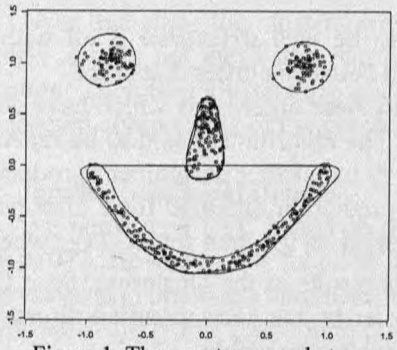

Figure 1. The counterexample to convex admissibility of SVC – presented convex hulls of two clusters are not disjoint.

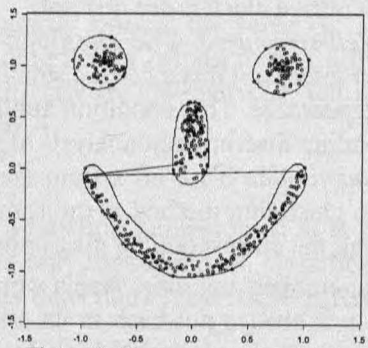

Figure 2. The counterexample showing that SVC is not well-structured. The shorter line segment represents one of the between cluster Euclidean distances, the longer one – the within cluster distance.

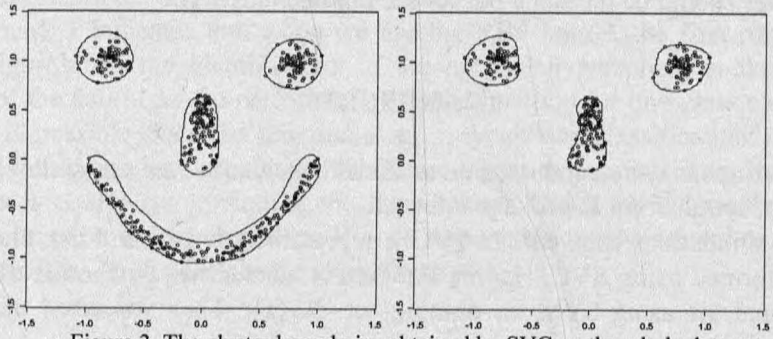

Figure 3. The cluster boundaries obtained by SVC on the whole data set and on the data set with one cluster excluded – the shapes of the contour boundaries are different and therefore SVC does not meet the cluster omission admissibility condition.

The results of the analysis of selected properties of SVC are summarized in Table 1.

Table 1. Admissibility table

Clustering method	PROPERTIES				
	Image	Convex	Well-structured	Repeatable	Cluster omission
SVC	YES	NO	NO	YES	NO
Nearest neighbour	YES	NO	YES	YES	YES
Furthest neighbour	YES	NO	YES	NO	YES
Average linkage	YES	NO	YES	NO	YES
Ward linkage	YES	YES	YES	NO	YES

Source: the properties of Nearest neighbour, Furthest neighbour, Average linkage and Ward linkage methods were taken from Fisher and Van Ness (1973). The properties of SVC are own results.

V. CONCLUSION

There are many clustering methods applicable in different situations. Since it is very hard to indicate a clustering method that would give the best results in every situation, the properties of clustering algorithms need to be considered. Given the knowledge about admissibility conditions fulfilled by different methods, the user can choose the proper method to tackle the particular problem.

The new clustering method known as Support Vector Clustering seems to be a flexible tool. It can handle very irregular shapes without making any assumptions about the number of clusters and their shapes. However, these features strongly depend on the kernel width parameter selection. The disadvantage of SVC is still the lack of an effective algorithm for choosing the value of the kernel width parameter.

The Support Vector Clustering satisfied only two of the analyzed admissibility conditions. However, meeting a certain condition (e.g. convex) is not always the property required by a user. Therefore, it may indicate the high flexibility of this method, but the control over this flexibility (kernel parameter selection) remains a crucial problem with. Moreover, solving the optimization problem and the process of the cluster assignment are computationally very expensive. This makes SVC unsuitable for large data sets. Taking into consideration these limitations, SVC should be applied with caution.

REFERENCES

Ben-Hur A., Horn D., Siegelman H.T., Vapnik V. (2001), Support Vector Clustering, *Journal of Machine Learning Research*, 2, 125–137.

Fisher L., Van Ness J.W. (1971), Admissible Clustering Procedures, *Biometrika*, 58, 91–104.

Fisher L., Van Ness J.W. (1973), Admissible Clustering Procedures, *Biometrika*, 60, 422–424.

Schölkopf B., Smola A. (2002), *Learning with Kernels. Support Vector Machines, Regularization, Optimization, and Beyond*, MIT Press, Cambridge.

Vapnik V. (1998), *Statistical Learning Theory*, John Wiley & Sons, N.Y.

Walesiak M. (2004), Clustering methods (in polish). In: Gatnar E., Walesiak M. (eds.), *The Methods of Multivariate Statistical Analysis in Marketing Research*, Wrocław University of Economics Publishing House, Wrocław, 316–350.

Michał Trzęsiok

ANALIZA WYBRANYCH WŁASNOŚCI TAKSONOMICZNEJ METODY WEKTORÓW NOŚNYCH

Celem referatu jest przedstawienie analizy wybranych formalnych własności taksonomicznej metody wektorów nośnych (SVC). Wyniki dotyczące nowej metody SVC zestawiono i porównano z własnościami innych znanych metod taksonomicznych.

Ponieważ na ogół nie jest możliwe wskazanie, która z metod taksonomicznych daje najlepsze rezultaty, stojąc wobec konkretnego problemu, badacz musi dokonywać wyboru metody w oparciu o wiedzę dotyczącą ich własności. Zadaniem badacza jest wtedy ustalenie preferencji w zbiorze własności metod by następnie użyć ich przy doborze odpowiedniego narzędzia. Wiedza dotycząca formalnych własności metod taksonomicznych jest w referacie rozszerzona o nową – taksonomiczną metodę wektorów nośnych.