Solving the One-Class Problem Using Neighbourhood Measures

Javier M. Moguerza[1] and Alberto Muñoz[2]

[1] University Rey Juan Carlos, c/ Tulipán s/n, 28933 Móstoles, Spain
j.moguerza@escet.urjc.es
[2] University Carlos III, c/ Madrid 126, 28903 Getafe, Spain
alberto.munoz@uc3m.es

Abstract. The problem of estimating high density regions from univariate or multivariate data samples is studied. To be more precise, we estimate minimum volume sets, whose probability is specified in advance. This problem arises in outlier detection and cluster analysis, and is strongly related to One-Class Support Vector Machines (SVM). In this paper we propose a new method to solve this problem, the Support Neighbour Machine (SNM). We show its properties and introduce a new class of kernels. Finally, numerical results illustrating the advantage of the new method are shown.

1 Introduction

The task of estimating high density regions from data samples arises explicitly in a number of works involving interesting problems such as outlier detection or cluster analysis (see for instance [7, 4] and references herein). One-Class Support Vector Machines (SVM) [7, 9] are designed to solve this problem with tractable computational complexity. We refer to [7] and references therein for a complete description of the problem and its ramifications.

In this work, a new algorithm to estimate high density regions from data samples is presented. The algorithm relaxes the density estimation problem in the following sense: instead of trying to estimate the density function at each data point, an easier to calculate data-based measure is introduced in order to establish a density ranking among the sample points.

The concrete problem to solve is the estimation of minimum volume sets of the form $S_\alpha(f) = \{x | f(x) \geq \alpha\}$, such that $P(S_\alpha(f)) = \nu$, where f is the density function and $0 < \nu < 1$. Throughout the paper, sufficient regularity conditions on f are assumed. For space reasons, proofs of propositions and theorems are omitted.

The rest of the paper is organized as follows. Section 2 introduces the Support Neighbour Machine and its properties. In Section 3, a kernel formulation of Support Neighbour Machines is shown. In Section 4, the performance of One-Class SVM and Support Neighbour Machines is compared on a variety of both artificial and real data sets. Section 5 concludes.

A. Fred et al. (Eds.): SSPR&SPR 2004, LNCS 3138, pp. 680–688, 2004.

2 The Support Neighbour Machine

There are data analysis problems where the knowledge of an accurate estimator of the density function $f(x)$ is sufficient to solve them, for instance, mode estimation [1], or the present task of estimating $S_\alpha(f)$. However, density estimation is far from trivial [8, 7]. The next definition is introduced to relax the density estimation problem: the task of estimating the density function at each data point is replaced by a simpler measure that asymptotically preserves the order induced by f.

Definition 1 (Neighbourhood Measures). *Consider a random variable X with density function $f(x)$ defined on $\mathbb{R}^d$. Let S_n denote the set of random independent identically distributed (iid) samples of size n (drawn from f). The elements of S_n take the form $s_n = (x_1, \cdots, x_n)$, where $x_i \in \mathbb{R}^d$. Let $M : \mathbb{R}^d \times S_n \longrightarrow \mathbb{R}$ be a real-valued function defined for all $n \in \mathbb{N}$. (a) If $f(x) < f(y)$ implies $\lim_{n\to\infty} P(M(x, s_n) > M(y, s_n)) = 1$, then M is a **sparsity measure**. (b) If $f(x) < f(y)$ implies $\lim_{n\to\infty} P(M(x, s_n) < M(y, s_n)) = 1$, then M is a **concentration measure**.*

Example 1. $M(x, s_n) \propto 1/\hat{f}(x, s_n)$, where $\hat{f}$ can be any consistent non-parametric density estimator, is a sparsity measure; while $M(x, s_n) \propto \hat{f}(x, s_n)$ is a concentration measure. A commonly used estimator is the kernel density one $\hat{f}(x, s_n) = \frac{1}{nh^d} \sum_{i=1}^{n} K(\frac{\|x-x_i\|}{h})$.

Example 2. Consider the distance from a point x to its k^{th}-nearest neighbour in s_n, $x^{(k)}$: $M(x, s_n) = d_k(x, s_n) = d(x, x^{(k)})$: it is a sparsity measure. Note that d_k is neither a density estimator nor is it one-to-one related to a density estimator. Thus, the definition of 'sparsity measure' is not trivial. Another valid choice is given by the average distance over all the k nearest neighbours: $M(x, s_n) = \bar{d}_k = \frac{1}{k}\sum_{j=1}^{k} d_j = \frac{1}{k}\sum_{j=1}^{k} d(x, x^{(j)})$. Extensions to other centrality measures, such as trimmed-means are straightforward.

Our goal is to obtain some decision function $h(x)$ which solves the problem stated in the introduction, that is, $h(x) = +1$ if $x \in S_\alpha(f)$ and $h(x) = -1$ otherwise. We will show how to use sparsity measures to build $h(x)$. To this aim a new algorithm, the Support Neighbour Machine, is introduced. Consider a sample $s_n = \{x_1, \ldots, x_n\}$. The SNM method works by solving the following optimization problem:

$$\begin{aligned} &\max_{\rho,\xi} \ \nu n\rho - \sum_{i=1}^{n} \xi_i \\ &\text{s.t.} \quad g(x_i) \geq \rho - \xi_i \,, \\ &\qquad \xi_i \geq 0, \qquad\qquad i = 1, \ldots, n \,, \end{aligned} \tag{1}$$

where $g(x_i) = M(x_i, s_n)$ is a sparsity measure, $x_i \in s_n$, ξ_i are slack variables, $\nu \in [0, 1]$ is a predefined constant and ρ is a variable whose role will become

clear below. Note that the calculus of $g(x)$ is not involved in the optimization process; it has to be determined in advance.

The SNM problem formulation is very similar to the linear One-Class SVM formulation (see [5]), but the solution to problem (1) is simpler and its continuity and differentiability is straightforward (while the use of the L_1 norm in the linear One-Class SVM problem implies non derivability, and using any other L_p norm would imply non linearity).

The motivation to adopt the name 'Support Neighbour Machines' is simple: 'sparsity' and 'concentration' are both neighbourhood measures.

The next proposition shows that the decision function $h(x) = sign(\rho^* - g(x))$ will be non-negative for at least a proportion equal to ν of the training s_n sample, where ρ^* is the value of ρ at the solution of problem (1). Following [7], this result is called ν-property.

Proposition 1 (ν-property). *At the solution of problem (1) the following two statements hold:*

1. $\frac{1}{n}\sum_{i=1}^{n} I(g(x_i) < \rho) \leq \nu \leq \frac{1}{n}\sum_{i=1}^{n} I(g(x_i) \leq \rho)$, *where I stands for the indicator function and $x_i \in s_n$.*
2. *With probability 1, asymptotically, the preceding inequalities become equalities.*

Remark 1. If $g(x)$ is chosen to be a concentration measure, then the decision function has to be defined as $h(x) = sign(g(x) - \rho^*)$.

Notice that ν in problem (1) represents the fraction of points inside the support of the distribution if $g(x)$ is a sparsity measure. If a concentration measure is used, ν represents the fraction of outlying points. The role of ρ becomes now clear: it represents the decision value which, induced by the sparsity measure, determines if a given point belongs to the support of the distribution.

As the next theorem states an asymptotical result, we will denote every quantity depending on the sample s_n with the subscript n. The theorem goes one step further from the ν-property, showing that, asymptotically, the SNM algorithm finds the desired α-level sets.

In order to formulate the theorem, we need a measure to estimate the difference between two sets. We will use the d_μ-distance. Given two sets A and B

$$d_\mu(A, B) = \mu(A \Delta B)\,, \tag{2}$$

where μ is a measure on $\mathbb{R}^d$, Δ is the symmetric difference $A\Delta B = (A \cap B^c) \cup (B \cap A^c)$, and A^c denotes the complementary set of A.

Theorem 1. *Consider a measure μ absolutely continuous with respect to the Lebesgue measure. The set $R_n = \{x : h_n(x) = sign(\rho_n^* - g(x)) \geq 0\}$ d_μ-converges to a region of the form $S_\alpha(f) = \{x | f(x) \geq \alpha\}$, such that $P(S_\alpha(f)) = \nu$. Therefore, the Support Neighbour Machine estimates a density contour cluster $S_\alpha(f)$ (which, in probability, includes the mode).*

We provide an estimate of a region $S_\alpha(f)$ with the property $P(S_\alpha(f)) = \nu$. Among regions S with the property $P(S) = \nu$, the region $S_\alpha(f)$ will have minimum volume as it has the form $S_\alpha(f) = \{x|f(x) \geq \alpha\}$. Therefore we provide an estimate that asymptotically, in probability, has minimum volume.

Finally, it is important to remark that the quality of the estimation procedure heavily depends on using a sparsity or a concentration measure (the particular choice is not – asymptotically – relevant). If the measure used is neither a concentration nor a sparsity measure, there is no reason why the method should work.

3 Kernel Formulation of SNM

In this section we will show the relation between SNM and One-Class SVM. In order to do so we have to define a class of neighbourhood measures.

Definition 2 (Positive and Negative Neighbourhood Measures). *$MP(x, s_n)$ is said to be a **positive sparsity (concentration) measure** if $MP(x, s_n)$ is a sparsity (concentration) measure and $MP(x, s_n) \geq 0$. $MN(x, s_n)$ is said to be a **negative sparsity (concentration) measure** if $-MN(x, s_n)$ is a positive concentration (sparsity) measure.*

Given that negative neighbourhood measures are in one-to-one correspondence to positive neighbourhood measures, only positive neighbourhood measures need to be considered. The following classes of kernels can be defined using positive neighbourhood measures.

Definition 3 (Neighbourhood Kernels). *Consider the mapping $\phi : \mathbb{R}^d \to \mathbb{R}^+$ defined by $\phi(x) = MP(x, s_n)$, where $MP(x, s_n)$ is a positive neighbourhood measure. The function $K(x, y) = \phi(x)\phi(y)$ is called a **neighbourhood kernel**. If $MP(x, s_n)$ is a positive sparsity (concentration) measure, $K(x, y)$ is a **sparsity (concentration) kernel**.*

Note that the set $\{\phi(x_i)\}$ is trivially separable in the sense of [7], since each $\phi(x_i) \in \mathbb{R}^+$. Separability is guaranteed by Definition 2.

The strategy of One-Class support vector methods is to map the data points into a feature space determined by a kernel function, and to separate them from the origin with maximum margin (see [7] for details). In order to build a separating hyperplane between the origin and the points $\{\phi(x_i)\}$, the quadratic One-Class SVM method solves the following problem:

$$\begin{aligned} \min_{w,\rho,\xi} \ & \frac{1}{2}\|w\|^2 - \nu n\rho + \sum_{i=1}^{n} \xi_i \\ \text{s.t.} \ & \langle w, \phi(x_i)\rangle \geq \rho - \xi_i\,, \\ & \xi_i \geq 0, \qquad\qquad i = 1, \ldots, n\,, \end{aligned} \tag{3}$$

where ϕ is the mapping defining the kernel function, ρ and ξ_i are variables whose meaning is the same as that in problem (1), and $\nu \in [0, 1]$ is an a priori fixed

constant. In the following we will refer to 'quadratic One-Class SVM' simply as 'One-Class SVM'.

The next theorem illustrates the relation between SNM and One-Class SVM when neighbourhood kernels are used.

Theorem 2. *Define the mapping* $\phi(x) = MP(x, s_n)$. *The decision function* $h(x) = sign(\rho_V^* - w^*\phi(x))$ *obtained from the solution* ρ_V^* *and* w^* *to the One-Class SVM problem (3) using the sparsity kernel* $K(x, y) = \phi(x)\phi(y)$ *coincides with the solution obtained by solving the SNM problem (1) using a positive sparsity measure.*

It remains open to show if the decision function obtained from One-Class SVM algorithms within the framework in [7, 5] can be stated in terms of positive sparsity or concentration measures. The next remark provides the answer.

Remark 2. The exponential kernel $K_c(x, y) = e^{-\|x-y\|^2/c}$ is neither a sparsity kernel nor a concentration kernel. For instance, consider a univariate bimodal density f with finite modes m_1 and m_2 such that $f(m_1) = f(m_2)$. Consider any positive sparsity measure $MP(x, s_n)$ and the induced mapping $\phi(x) = MP(x, s_n)$. As $n \to \infty$, the sparsity kernel $K(x, y) = \phi(x)\phi(y)$ would attain its minimum at (m_1, m_2) (or at two points in the sample s_n near to the modes). On the other hand, as the exponential kernel $K_c(x, y)$ depends exclusively on the distance between x and y, any pair of points (a, b) whose distance is larger than $\|m_1 - m_2\|$ will provide a value $K_c(a, b) < K_c(m_1, m_2)$, which asymptotically can not happen for kernels induced by positive sparsity measures. In this case, the neighbourhood kernel has four minima while the exponential kernel has the whole diagonal as minima. The reasoning for concentration kernels is analogous. A similar argument applies for polinomial kernels with even degrees (odd degrees induce mapped data sets that are non separable from the origin, which discards them).

Note that, while SNM work with every neighbourhood measure, the separability condition of the mapped data is necessary when One-Class SVM are being used, restricting the use of neighbourhood measures to positive or negative ones. This restriction and the fact that SNM provide a simpler linear approach make the use of SNM advisable when neighbourhood measures are being used.

4 Experiments

In this section we compare the performance of One-Class SVM and SNM for a variety of artificial and real data sets. Systematic comparisons of the two methods as data dimension increases are carried out. First of all we describe the implementation details concerning both algorithms.

With regards to One-Class SVM we adopt the proposal in [7], that is, the exponential kernel $K_c(x_i, x_j) = e^{-\|x_i - x_j\|^2/c}$ is used (the only kernel tested for experimentation in that work). To perform the experiments, a range of values

for c has been chosen, following the widely used rule $c = hd$ (see [6, 7]), where $h \in \{0.1, 0.2, 0.5, 0.8, 1.0\}$ and d is the data dimension.

Concerning SNM, two different sparsity measures have been considered:

- $M_1(x, s_n) = d_k = d(x, x^{(k)})$, the distance from a point x to its k^{th}-nearest neighbour $x^{(k)}$ in the sample s_n. The only parameter in M_1 is k, which takes a finite number of values (in the set $\{1, \cdots, n\}$). We have chosen k to cover a representative range of values, namely, k will equal the 10%, 20%, 30%, 40% and 50% sample proportions. Therefore we choose k as the closest integer to hn, where n is the sample size and $h \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$.
- $M_2(x, s_n) = \dfrac{1}{\sum_{i=1}^{n} \exp\left(-\frac{\|x-x_i\|^2}{2\sigma}\right)}$, where $\sigma \in \mathbb{R}^+$. The only parameter in M_2 is σ. We want σ to be related to the sample variability and, at the same time, to scale well with respect to the data sample distances. We choose $\sigma = hs$, where $s = \max d_{ij}^2/\varepsilon$, $h \in \{0.1, 0.2, 0.5, 0.8, 1.0\}$, $d_{ij}^2 = \|x_i - x_j\|^2$ and ε is a small value which preserves scalability in M_2. For all the experiments we have chosen $\varepsilon = 10^{-8}$.

Measure M_1 has been described in Example 2 in Section 2. Measure M_2 is of the type described in Example 1. M_2 uses as density estimator the Parzen window [8]. Note that Theorem 1 guarantees that asymptotically every sparsity measure (and in particular the two chosen here) will lead to sets containing the true mode.

4.1 Artificial Data Sets

In the first experiment we have generated 2000 points from a gamma $\Gamma(\alpha, \beta)$ distribution, with $\alpha = 1.5$ and $\beta = 3$. Figure 1 shows the histogram, the gamma density curve, the true mode $(\alpha - 1)/\beta$ as a bold vertical line, the SNM estimations with sparsity measure M_1 (five upper lines) and the One-Class SVM (five lower lines) estimations of the 50% highest density region. The parameters have been chosen as described at the beginning of Section 4, and lines are drawn for each method in increasing order in the h parameter, starting from the bottom. Being our goal to detect the shortest region of the form $S_\alpha(f) = \{x : f(x) > \alpha\}$ (that must contain the mode), it is apparent that the SNM regions improve upon the One-Class SVM regions. All the SNM regions contain the true mode and are connected. All the One-Class SVM regions are wider and show a strong bias towards less dense zones. Furthermore, only in two cases the true mode is included in the estimated SVM regions, but in these cases the intervals obtained are not simply connected. SNM using measure M_2 provide similar intervals to those obtained using measure M_1, and are not shown for space reasons.

In the second experiment a mixture of a normal $N(0, 1)$ and a uniform $U(6, 9)$ distribution is considered. Figure 2 shows the results. All the One-Class SVM (five lower lines) estimations spread part of the points in the uniform zone. However, all points in this zone have lower density than those found by the SNM procedure.

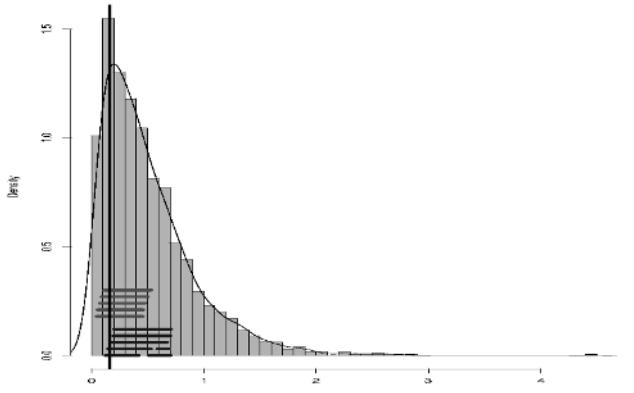

Fig. 1. Gamma sample with 2000 points. The figure shows the histogram, the density curve, a vertical line at the true mode, the SNM estimations with sparsity measure M_1 (five upper lines) and One-Class SVM (five lower lines) estimations of the 50% highest density region.

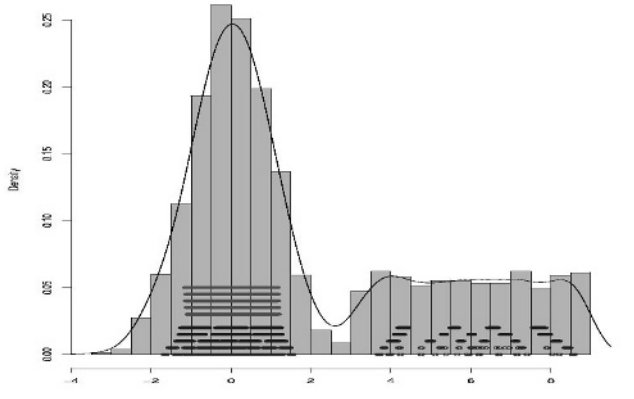

Fig. 2. Mixture sample with 3000 points. The figure shows the histogram, the estimated density curve, the SNM estimations with sparsity measure M_1 (five upper lines) and One-Class SVM (five lower lines) estimations of the 50% highest density region.

4.2 Increasing the Dimension of the Data Space

In this experiment we want to evaluate whether the performance of the SNM and SVM algorithms degrades as the data dimension increases. To this end, we have generated 20 data sets with increasing dimension from 2 to 200. Each data set contains 2000 points from a multivariate normal distribution $N(0, I_d)$, where I_d is the identity matrix in $\mathbb{R}^d$. Detailed results are not shown for space reasons. We will only show the conclusions. Since the data distribution is known, we can retrieve the true outliers, that is, the true points outside the support corresponding to any percentage specified in advance. For each dimension and

Table 1. Percentage of true outliers detected using the One-Class SVM, $d = 9$.

Results for One-Class SVM – Cancer Data					
$c = hd$	$0.1d$	$0.2d$	$0.5d$	$0.8d$	$1.0d$
Success %	53.7 %	61.3 %	72.4 %	78.6 %	79.7 %

Table 2. Percentage of true outliers detected using the SNM with M_1, $n = 683$.

Results for SNM with M_1 – Cancer Data					
$k \simeq hn$	$0.1n$	$0.2n$	$0.3n$	$0.4n$	$0.5n$
Success %	94.1 %	95.0 %	95.0 %	95.0 %	95.8 %

each method, we have determined, from the points retrieved as outliers, the proportion of true ones.

As the data dimension increases, the performance of One-Class SVM degrades: it tends to retrieve as outliers an increasing number of points. The best results for One-Class SVM are obtained for the largest magnitudes of the parameter c. Regarding the SNM procedure, robustness with regard to the parameter choice is observed. Dimension barely affects the performance of the SNM method, and results are consistently better than those obtained with One-Class SVM. For instance, for a percentage of outliers equal to 1%, the best result for One-Class SVM is 15%, against 100% using the SNM method (for all the sparsity measures considered). For a percentage of outliers equal to 5%, the best result for One-Class SVM is 68%, against 99% using the SNM method.

4.3 Cancer Data Set

This data set is given by a 699×9 matrix of clinical measures taken on breast cancer patients [2]. After removing cases with missing values, the matrix dimensions are 683×9 and the real class distribution becomes 65% benign and 35% malignant. The cancer data set is interesting for various reasons: relatively high dimension, overlap, unbalanced classes and different density distributions for each group. This data set has traditionally been approached from a supervised point of view, using classification schemes. Here we focus on a different point of view: we hypothesize that there is only one multivariate (approximate normal) distribution made up of bening cases (65% of the sample) and that the malignant cases are (asymmetrically distributed) outliers. This hypothesis is graphically supported by two-dimensional projections (see for instance [3]). Therefore, we run SNM and One-Class SVM on the cancer data set to detect the 65%-level set, and the percentage of outliers (malignant cases) detected by each of the algorithms is checked. Results are shown in Tables 1 and 2. The best One-Class SVM model detects 79.7% of the outliers, while the worst SNM model detects 94.1% of the outliers, with the best SNM result being equal to 95.8%. Once more the choice of the sparsitiy measure does not affect the results.

5 Conclusions

In this paper the Support Neighbour Machine, a new method to estimate minimum volume sets of the form $S_\alpha(f) = \{x|f(x) \geq \alpha\}$, has been proposed. The new algorithm introduces the use of neighbourhood measures. These measures asymptotically preserve the order induced by the density function f. In this way we avoid the complexity of solving a pure density estimation problem.

Regarding computational results, SNM performs consistently better than One-Class SVM in all the tested problems. The advantage that the SNM has over the One-Class SVM is due to Theorem 1 which guarantees that the SNM algorithm tends to (asymptotically) find the desired α-level sets. The suboptimal performance of One-Class SVM may arise from the fact that its decision function is not based on sparsity or concentration measures.

Acknowledgments

This work was partially supported by MCyT grant TIC2003-05982-C05-05 and by URJC grant PPR-2003-42 (Spain).

References

1. L. Devroye. *Recursive estimation of the mode of a multivariate density.* The Canadian Journal of Statistics, 7(2):159-167, 1979.
2. O.L. Mangasarian and W.H. Wolberg. *Cancer diagnosis via linear programming* SIAM News,Volume 23, Number 5, 1990, 1-18.
3. J.M. Moguerza, A. Muñoz and M. Martin-Merino. *Detecting the Number of Clusters Using a Support Vector Machine Approach.* Proc. ICANN 2002, LNCS 2415:763-768, Springer, 2002.
4. A. Muñoz and J. Muruzabal. *Self-Organizing Maps for Outlier Detection.* Neurocomputing, 18:33-60, 1998.
5. G. Rätsch, S. Mika, B. Schölkopf and K.R. Müller. *Constructing Boosting Algorithms from SVMs: an Application to One-Class Classification.* IEEE Trans. on Pattern Analysis and Machine Intelligence, 24(9):1184-1199, 2002.
6. B. Schölkopf, C. Burges and V. Vapnik. *Extracting Support Data for a given Task.* Proc. of the First International Conference on Knowledge Discovery and Data Mining, AAAI Press, 1995.
7. B. Schölkopf, J.C. Platt, J. Shawe-Taylor, A.J. Smola and R.C. Williamson. *Estimating the Support of a High Dimensional Distribution.* Neural Computation, 13(7):1443-1471 , 2001.
8. B.W. Silverman. *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, 1990.
9. D.M.J. Tax and R.P.W. Duin. *Support Vector Domain Description.* Pattern Recognition Letters, 20:1991-1999, 1999.