Computing Science And Statistics

Volume 34

Geoscience and Remote Sensing

Edward J. Wegman
Amy Braverman
Editors

Proceedings of the
34th Symposium on the Interface
Montréal, Québec, Canada, April 17-20, 2002

INTERFACE
FOUNDATION
OF NORTH AMERICA

Novelty Detection in Mass Spectral Data using a Support Vector Machine Method

Christopher Tong[1,2] and Vladimir Svetnik[2]

[1]Dept. of Statistics, Purdue University
West Lafayette, IN 47907-1399
[2]Biometrics Research Department
Merck Research Laboratories
RY84-16, P. O. Box 2000
Rahway, NJ 07065
ctong@mailaps.org, vladimir_svetnik@merck.com

Abstract

Outlying samples are sought in a very high-dimensional data set, a library of mass spectra. Such samples are considered novel from the chemical structure point of view and are identified for further investigation of their potential biological activity. The support vector machine algorithm for domain description (Tax & Duin 1999; Schölkopf *et al.* 2000, 2001) is used to generate a list of potential outliers. The results are compared to those found by a sequential clustering procedure (Svetnik and Liaw, 2001). The results are quite reasonable.

1. INTRODUCTION

The multivariate novelty detection (outlier identification) problem is a standard one in classical statistics (e.g., Barnett and Lewis 1993) as well as in data mining (e.g., Bruenig *et al.* 2000; Guha *et al.* 2000; Ramaswamy *et al.* 2000). A recent contribution to the toolbox of nonparametric novelty detection methods is the *support vector domain description* (Tax & Duin 1999; Schölkopf *et al.* 2000, 2001). This approach is based on the Support Vector Machine algorithm that has stimulated recent interest in kernel-based learning methods (e.g., Vapnik 2000; Cristianini & Shawe-Taylor 2000). In this paper, the support vector machine approach is applied to a mass spectral data library. The results are compared to corresponding earlier results on the same data set, using the sequential clustering procedure for novelty detection (Svetnik & Liaw 2001).

The present work is motivated by a specific novelty detection problem originating from a drug discovery activity (An *et al.* 2001; Svetnik & Liaw 2001). Natural products, such as fungi and other microorganisms, are a rich source of potential drugs. Many fungi are difficult, if not impossible, to grow under lab conditions. An *et al.* (2001) described a technique in which the genome of a slow-growing fungus is randomly cut into about 1000 pieces, each of which is inserted into the genome of an easy-to-grow "host" fungus. This results in about 1000 transgenic fungi that are as easy to grow as the original host

fungi, but possibly possessing some novel characteristics of the donor fungus. In order to identify transgenic fungi with novel characteristics, their extracts are subjected to mass spectrometry. Mass spectrometry is an analytical chemistry technique that reveals important information about the molecular structure of a compound (see, e.g., De Hoffman and Stroobant 2001). Each extract has one mass spectrum consisting of intensity (abundance) measurements at 800 channels (mass-to-charge ratios). Many of these measurements have a maximum intensity across all samples that does not exceed a detection threshold, and thus the number of "informative" channels can be reduced by as much as 40%. In the data set examined here, measurements at only 468 channels each are used (although this still qualifies as a very high-dimensional data set). There are 764 samples in the library, and since all their mass spectra cannot be visually inspected by a scientist, an automated screening procedure for identifying the "novel" samples (those with spectra most unlike the bulk of the spectra) is required. Usually mass spectral data are not normally distributed, which suggests that nonparametric multivariate methods could be quite useful for data analysis. Several such procedures have been tried on this library, and here results for the support vector machine approach are reported.

In the next section, a brief review of the support vector machine algorithm for novelty detection is presented. In Section 3, the method is applied to the data set just described, and the results are discussed. In the concluding Section 4, these results are compared with those of Svetnik & Liaw (2001) using the sequential clustering procedure.

2. SUPPORT VECTOR MACHINE ALGORITHM FOR NOVELTY DETECTION

One basic approach to novelty (outlier) detection is based on estimating the support of the data, the domain in which most of the data lives. Samples that fall outside the estimated support are declared to be outliers. Therefore, the problem is considered to be one of "single class classification": either a data point belongs to the data domain, or it is an outlier. Since the data are unlabeled, this is an example of *unsupervised learning*.

Several *support vector machine* (SVM) approaches have been proposed to solve this problem (Tax & Duin 1999; Schölkopf *et al*. 2000, 2001; Campbell & Bennett 2001). For Gaussian kernels, the methods of Tax & Duin and Schölkopf *et al*. are essentially equivalent. The Tax & Duin approach is the simplest to describe: one could begin by drawing a sphere around the majority of the data, allowing a small fraction of the data to fall outside. In other words, find the minimum radius sphere (or hypersphere) enclosing the majority of points, while including a penalty in the objective function for outliers. Since a sphere is unlikely to be the most natural geometric object to describe the data domain, one can instead nonlinearly map the data into a higher dimensional feature space, fit the spherical domain estimate in the feature space, and then map the domain estimate back into the input space, where it becomes a non-spherical "blob" or set of disjoint "blobs". The mapping and inverse mapping process can be bypassed using the standard kernel trick: since the algorithm for fitting the spherical domain estimate can be formulated in a way such that the data points only appear via inner products with each other, the *kernel function* that specifies how to compute these inner products is the only aspect of the feature space that is required. A commonly used kernel function is a Gaussian Radial Basis Function (RBF) kernel,

$$K(\mathbf{x}, \mathbf{y}) = \exp(-\|\mathbf{x} - \mathbf{y}\|^2 / 2\sigma^2),$$

where σ^2 is a free parameter that may be varied to produce different estimates of the data domain. The Gaussian kernel satisfies Mercer's condition, which guarantees that for every value of σ^2, $K(\mathbf{x}, \mathbf{y})$ corresponds to the inner product of **x** and **y** after they are nonlinearly mapped into *some* Hilbert space that need not be identified explicitly. The Gaussian kernel is a particularly convenient one to use in the novelty detection problem (Tax and Duin 1999; Schölkopf *et al.* 2000, 2001).

The resulting domain estimate boundary is determined by a subset of the original data points, those that are on the boundary of the domain estimate, as well as the outliers themselves. Together these samples are called *support vectors*. Hence the representation of the domain estimate requires storing only a subset of the training data (unlike, say, nearest neighbor methods which require storing the entire data set). The computational solution for the estimate is fast and simple, since the optimization problem is a quadratic program with a convex feasible space, allowing only a single global minimum. Local minima, which plague tree-based and neural methods, are thus avoided.

Computationally, the formulation by Schölkopf *et al*. (2000, 2001) is simpler. Without going into details, one ends up with the following optimization problem:

$$\min_{\vec{\alpha}} \sum_{ij} \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j),$$

subject to the constraints,

$$\sum_i \alpha_i = 1, \qquad \forall i, 0 \le \alpha_i \le \frac{1}{\nu d},$$

where d is the dimensionality of the data and ν is a free parameter to be discussed shortly. The α_i's are Lagrange multipliers, one each associated with every data point. The task of the optimization procedure is to solve for these coefficients. Most of them are expected to be zero, and only those corresponding to support vectors will be nonzero. Once these Lagrange multipliers are determined, the decision function is written as

$$h(\mathbf{x}) = \sum_{i \in SV} \alpha_i k(\mathbf{x}_i, \mathbf{x}) - \rho,$$

where the sum is over the subset *SV* of data that are support vectors. (The threshold ρ is computed from the training data in a specified way; see Schölkopf *et al*. 2000.) The domain estimate is defined as the region where the above decision function takes a positive value. Classification of an individual data point **x** is done by evaluating the sign of the decision function h at that **x**. Outliers are those who have negative values of $h(\mathbf{x})$; they also have Lagrange multipliers attaining the upper bound $1/\nu d$. It has been suggested that the values of $h(\mathbf{x})$ for outliers can be used as a measure of degree of outlyingness (Schölkopf *et al.* 2000, 2001). Finally, it should be noted that the Schölkopf *et al.* (2000, 2001) formulation is based on a "maximal margin" idea that is inspired by the results of statistical learning theory (Vapnik, 2000).

In this work, we focus exclusively on the Gaussian kernel. There are then two free parameters in the SVM algorithm. The first, which we keep fixed, is a parameter ν

which serves simultaneously as an upper bound on the proportion of outliers and a lower bound on the proportion of support vectors (Schölkopf *et al.* 2000, 2001). It takes values in the range between 1/*n* and 1, where *n* is the number of samples. The second, σ^2, controls the sparseness of the data domain estimate. Generally, the larger its value, the more sparse is the domain estimate and the fewer support vectors are used. However, the behavior of, say, the number of support vectors as σ^2 varies is not guaranteed to be strictly monotonic.

The above has only been a brief overview of the SVM outlier detection methods. Please refer to the original papers (Tax & Duin 1999; Schölkopf *et al*. 2000, 2001; Campbell & Bennett 2001) for a more thorough discussion of the SVM outlier detection algorithms, as well as other applications thereof.

3. APPLICATION TO MASS SPECTRAL DATA LIBRARY

The mass spectral data library referred to in the introduction has been analyzed using the SVM Matlab package of Ma & Ahalt (2000), which in turn uses SVM MEX library files from LIBSVM written by Chang and Lin (2000). This package implements the outlier detection algorithm in addition to the usual SVM classification algorithm. Recall that there are 764 samples, each with 468 channels (dimensions). We fixed ν to be 0.10 and varied $\Gamma = 1/2\sigma^2$ over a range of values which produced consistent results (for extreme values of Γ, the proportion of outliers identified is not close to the pre-assigned value of ν). Also, in addition to applying the algorithm on the original data, we duplicated the procedure on normalized data. (In the latter, each sample is standardized with respect to its own mean and standard error.)

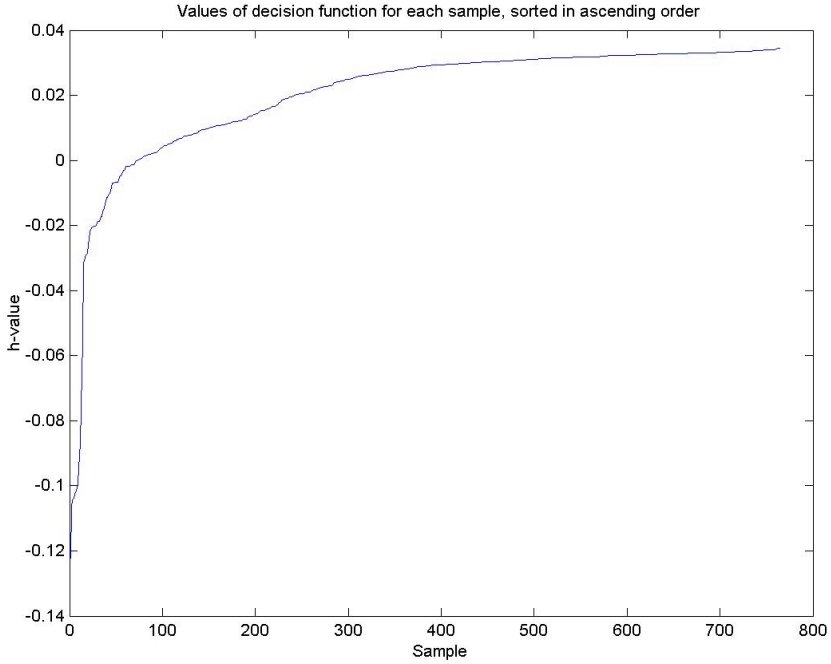
Once the SVM was trained, the original data can be classified as an outlier or not based on either the corresponding sign of the decision function *h* or based on the value of the corresponding Lagrange multiplier for that sample. Numerically, it was found that classifying based on the Lagrange multiplier was more reliable. This is because non-outlier support vectors have an *h*-value very close to zero and numerical error can result in small negative values for them.

On the non-normalized data, an outlier list was produced for five different values of Γ (each list had about 76 entries, consistent with the pre-assigned value of ν) and the results were pooled together to form a grand list of 88 outliers. Only one non-outlier support vector was identified in each run. The procedure was repeated for the normalized data, resulting in a grand list of 79 outliers, and no more than four non-outlier support vectors were identified in each run. The fact that so few non-outlier support vectors are used indicates that the outliers maintain a strong influence on the boundary of the domain estimate. In addition, the fact that the pooled outlier lists were not much larger than the size of an individual list of 76 indicates that the results were fairly stable within the range of Γ used.

The two lists of outliers can themselves be compared: together, they identify a total of 120 distinct samples as outliers. Of these, 39% were commonly identified by both lists, 34% were identified as outliers by only the non-normalized SVM list, and 27% were identified as outliers by only the normalized SVM list.

To illustrate the performance characteristics of the SVM algorithm, we focus now on a particular run of the SVM on non-normalized data, at the specific parameter value $\Gamma = 1 \times 10^{-13}$. After the SVM is trained, the decision function *h* is evaluated for each training sample. The resulting *h*-values can be sorted in ascending order, and a plot of these values is given in *Fig. 1*. As can be seen in the plot, the first 77 or so samples have a negative *h*-value and are the outliers.

Figure 1: Decision function values for mass spectra data using SVM with $\Gamma = 1 \times 10^{-13}$ *and ν = 0.1, sorted in ascending order.*



A median spectrum for the entire data set can be formed by taking the median of each channel across all samples; the resulting spectrum is therefore thought to represent a "typical" sample. The median spectrum for our data set is shown at the top of *Fig. 2*, along with the spectra of three examples of compounds declared to be outliers by the SVM. One can see that the outlier samples have spectra quite distinct from the median spectrum.

4. COMPARISON WITH RESULTS OF SEQUENTIAL CLUSTERING PROCEDURE

Because one does not know the "right" answer in unsupervised learning, it is difficult to estimate the error rate or confidence level (although these could be approximated using a resampling method). We propose instead to compare the results of the SVM algorithm to those of another outlier detection method. In any case, it is good practice to compare results from two competing methods when there is no reason to believe *a priori* that one method is superior to the other. In this case, we will compare the SVM results to those

of Svetnik and Liaw (2000) using a sequential clustering procedure (SCP). The latter method hierarchically clusters the data and identifies "small clusters" that are far away from "large clusters" as potential outlier data. Like SVMs, the SCP method can accommodate a wide range of similarity measures that could be chosen by the user, and it provides a measure of the degree of outlyingness for a given sample.

When the SCP was applied to the mass spectra library discussed above, three different similarity measures were utilized with average linkage clustering: Euclidean distance, Pearson correlation, and rank correlation. Combining the resulting outlier lists gives a grand list of 54 samples. This list overlaps with the SVM outlier list for non-normalized data by 61%. The SCP list overlaps with the SVM outlier list for normalized data by 83%. See *Table I* for a summary of these results.

Table I: Overlap of SVM outlier lists with SCP outlier list

SVM Distance measure:	**# in common w/ SCP:**	**% in common w/ SCP:**
Non-normalized data	33/54	61%
Normalized data	45/54	83%

Because two independent, distinct procedures (SVM and SCP) produce lists of outliers that significantly overlap, one could informally conclude that the results are quite reasonable. However, validation of these results remains for future work, including, e.g., performing a test for differences between the outlying spectra and a random sample of non-outlying spectra. The expert opinion of chemists should also be sought regarding whether there are real differences between the two sets of spectra.

One criticism of the present approach is that the ordering of the channels is ignored. Since the channels represent not masses, but rather mass-to-charge ratios, the physical interpretation of the ordering of the channels is not straightforward. Nonetheless, other outlier detection methods have been proposed to account for the ordering of the variables (P. L. Davies, personal communication).

Once validated, the above results will be of tremendous value, since the number of potentially novel chemical structures that should be further investigated has been reduced by nearly an order of magnitude. The use of these procedures can therefore be of great benefit for drug discovery activities.

ACKNOWLEDGMENTS

We are very grateful to Andy Liaw for his contributions to this work.

REFERENCES

Z. An, G. Harris, D. Zink, R. Giacobbe, R. Sangari, P. Lu, J. Greene, B. Gerald, C. Meyers, J. Armbruster, S. Smith, V. Svetnik, B. Gunter, A. Liaw, P. Masurarekar, J. Liesch, G. Steven, and W. Strohl, 2001: Expression of cosmid-size DNA of slow-growing fungi in *Aspergillus nidulans* for secondary metabolite screening. Submitted to *Nature Biotechnology*.

V. Barnett and T. Lewis, 1994: *Outliers in Statistical Data*. Third edition. Wiley.

M. M. Bruenig, H. P. Kriegel, R. T. Ng, and J. Sander, 2000: Identifying density-based local outliers. In *Proceedings of the ACM SIGMOD Conference on Management of Data.*

C. Campbell and K. P. Bennett, 2001: A linear programming approach to novelty detection. In *Advances in Neural Information Processing Systems 14.*

C.-C. Chang and C.-J. Lin, 2000: *LIBSVM – A Library for Support Vector Machines.* http://www.csie.ntu.edu.tw/~cjlin/libsvm/

N. Cristianini and J. Shawe-Taylor, 2000: *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods.* Cambridge.

E. De Hoffman and V. Stroobant, 2001: *Mass Spectrometry: Principles and Applications*. Second edition. Wiley.

S. Guha, R. Rastogi, and K. Shim, 2000: Cure: an efficient clustering algorithm for large databases. In *Proceedings of the ACM SIGMOD Conference on Management of Data*.

J. Ma and S. Ahahlt, 2000: *OSU SVM Classifier Toolbox*, version 2.00. http://eewww.eng.ohio-state.edu/~maj/osu_svm/

S. Ramaswamy, R. Rastogi, and K. Shim, 2000: Efficient algorithms for mining outliers from large data sets. In *Proceedings of the ACM SIGMOD Conference on Management of Data.*

B. Schölkopf, R. Williamson, A. Smola, J. Shawe-Taylor, and J. Platt, 2000: Support vector method for novelty detection. In *Advances in Neural Information Processing Systems 12.*

B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, 2001: Estimating the support of a high-dimensional distribution. *Neural Computation*, **13**, 1443-1471.

V. Svetnik and A. I. Liaw, 2001: Detection of novel samples in mass spectral data using cluster analysis. In *Interface 2001*.

D. M. J. Tax and R. P. W. Duin, 1999: Support vector domain description. *Pattern Recognition Letters*, **20**, 1191-1199.

V. Vapnik, 2000: *The Nature of Statistical Learning Theory*. Second edition. Springer.

Figure 2: Median mass spectrum for entire library (top), plus examples of three outlying samples' mass spectra, found using the SVM with $\Gamma = 1 \times 10^{-13}$ *and* $\nu = 0.1$.