Proceedings of the
International Conference on Mechanical Engineering 2005
(ICME2005) 28- 30 December 2005, Dhaka, Bangladesh

ICME05-AM-35

OUTLIERS DETECTION IN STATISTICAL QUALITY CONTROL USING SUPPORT VECTOR DATA DESCRIPTION

Birajashis Pattnaik[1] and S. S. N. Murty[2]

Department of Mechanical Engineering, Indian Institute of Technology, Kharagpur-721302, India
[1]biraja@mech.iitkgp.ernet.in, [2]ssnm7@mech.iitkgp.ernet.in

ABSTRACT

Outliers have been informally defined as observations in a data set which appear to be inconsistent with the remainder of that set of data, or which deviate so much from other observations so as to arouse suspicions that they were generated by a different mechanism [1]. The identification of outliers can lead to the discovery of useful knowledge and has a number of practical applications. The boundary of a dataset can be used to detect novel data or outliers. Outlier detection belongs to the most important tasks in data analysis. The outliers describe the abnormal data behavior, i.e. data that are deviating from the natural data variability Outlier detection has many applications, such as data cleaning and fraud detection. Frequently, outliers are removed to improve accuracy of the estimators. But sometimes the presence of an outlier has a certain meaning which explanation can be lost if the outlier is deleted. Often outliers are of primary interest, for example in geochemical exploration they are indications for mineral deposits. The cut-off value or threshold, which divides anomalous and non-anomalous data numerically, is often the basis for important decisions. It is tried here to find the best representation of a dataset such that the target class may best be distinguish from the real data. In this paper outlier in a dataset of percentage of different indigrends for copper production is studied using support vector data analysis for its quality study.

Keywords: Multivariate data analysis, Outlier detection, Support vector machine, Support vector data description.

1. INTRODUCTION

The outliers describe the abnormal data behavior, i.e. data that are deviating from the natural data variability. Multivariate outliers can be identified as points with large Mahalanobis distances based on robust estimates of population scatter and location.

The issue of robust estimation and/or outlier detection has been researched by many authors (Campbell, 1980, 1982; Davies, 1987; Develin et al., 1981; Hadi 1992, 1994; Hampel et al., 1986; Huber, 1981; Lopuhaä, 1989; Maronna, 1976; Rocke and Woodruff, 1993; Rousseeuw and Leroy, 1987; Rousseeuw and van Zomeren, 1990; Tyler, 1983, 1991).Rousseeuw (1985) introduces minimum volume estimator (MVE) and minimum covariance determinant (MCD) and use them for outlier detection. Other authours have used the concept of MVE or MCD in their outlier detection methods. Atkinson (1994), in his outlier detection method, considered forward search from random elemental sets and choose a partition of the data that had the smallest “half” sample ellipsoid volume. Rocke and Woodruff (1996) obtained a hybrid algorithm utilizing steepest descent procedure of Hawkins (1993) for obtaining the MCD which was used as a starting point for forward search algorithm of Atkinson (1993) and Hadi (1992). Rocke and Woodruff performed extensive simulations and observed that it is very difficult to detect outliers in data with a contamination fraction of 35%, and almost impossible in data with a contamination fraction of 40% or 45%.

In this paper we explain detection of outlier technique named as support vector data description (SVDD, Tax and Duin, 1999), which is based on support vector method developed by Vapnik. A spherically shaped decision boundary around a set of objects is constructed by a set of support vectors describing the sphere boundary. It has the possibility of transforming the data to new feature spaces without much extra computational cost. By using the transformed data, this SVDD can obtain more flexible and more accurate data descriptions. The error of the first kind, the fraction of the training objects, which will be rejected, can be estimated immediately from the description without the use of an independent test set, which makes this method data efficient. The support vector data description is compared with other outlier detection methods on real data. In this paper a dataset of an electrolysis process of copper production was taken into consideration. A large dataset of eight variables of different metal impurities (ppm) in the process and three hundred seventy observations are taken into account to define the data domain and outliers. It is tried here to find the best representation of a dataset such that the target class may best be distinguish from the outlier class.

2. SUPPORT VECTOR MACHINE (SVM)

SVMs are a new learning method introduced by V.Vapnik et al. They are well-founded in terms of computational learning theory and very open to theoretical understanding and analysis. The foundations of Support Vector Machines (SVM) have been developed by Vapnik [2] and are gaining popularity due to many attractive features, and promising empirical performance [3]. They are used in many real world pattern recognition problems. Initially they are proposed for binary classification.

The classification problem can be restricted to consideration of the two-class problem without loss of generality. Consider the problem of separating the set of training vectors belonging to two separate classes. Let $x_i, i = 1,2,...,l, x_i = \Re^n$ with corresponding class levels $y_i \in \{-1,1\}$ be the training vectors. y_i are also called the desired values in classical supervised learning. The class levels are discrete (e.g. Boolean) values for 1classification problem. Here l is the number of training observations and n is the dimension of each observation.

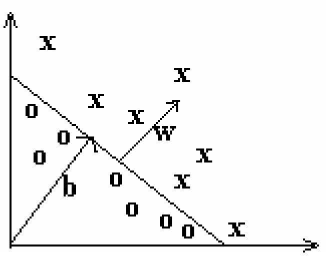

Fig 1. A separating hyperplane $(\boldsymbol{w}, \boldsymbol{b})$ for a two dimensional training set.

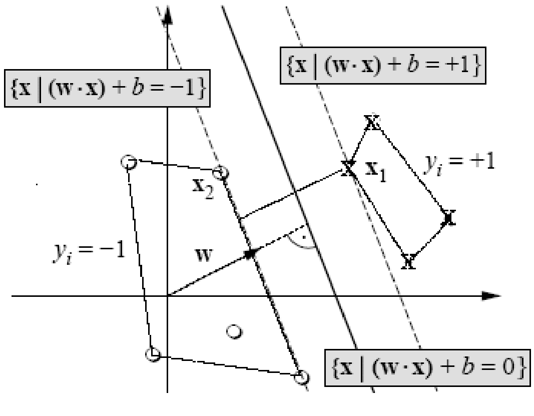
$\{\mathbf{x} \mid (\mathbf{w}\cdot\mathbf{x}) + b = -1\}$
$\{\mathbf{x} \mid (\mathbf{w}\cdot\mathbf{x}) + b = +1\}$
$\{\mathbf{x} \mid (\mathbf{w}\cdot\mathbf{x}) + b = 0\}$
$\mathbf{x}_1$
$\mathbf{x}_2$
$y_i = +1$
$y_i = -1$
w

Fig 2. A binary classification problem with maximal margin by SVM.

3. SUPPORT VECTOR DATA DESCRIPTION (SVDD)

Support vector machine is primarily for binary classification. But support vector data description (SVDD) is used for ‘one-class classification’ [5, 6]. For data domain description not the optimal separating hyperplane has to be found, but the sphere with minimal volume (or minimal radius) containing all objects. A sphere with minimum volume, containing all (or most of) the data objects is to be found out. This is very sensitive to the most outlying object in the target data set. When one or a few very remote objects are in the training set, a very large sphere is obtained which will not represent the data very well. Therefore, we allow for some data points outside the sphere and introduce slack variables ξ_i . The sphere is described by center a and radius R; the radius can be minimized by

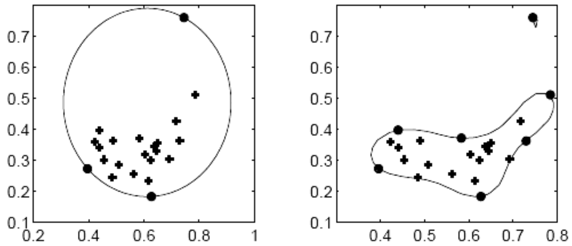

Fig 3. Data Description of a small data set, (left) normal spherical description, (right) description using a Gaussian kernel.

$$F(R,a,\xi_i) = R^2 + C\sum_i \xi_i \tag{1}$$

where the variable C gives the trade-of between simplicity (or volume of the sphere) and the number of errors (number of target objects rejected). ξ_i is a slack variable. This has to be minimized under the constraints

$$(x_i - a)^T (x_i - a) \le R^2 + \xi_I \quad \forall_{i,} \xi_i \ge 0 \tag{2}$$

where a is center of the sphere. Incorporating these constraints in $\sum_{i=1}^{m} \alpha_i y_i = 0$, we construct the Lagrangian,

$$L(R,a,\alpha_i,\xi_i) = R^2 + C\sum_i \xi_i$$

$$-\sum_i \alpha_i \left\{R^2 + \xi_i - \left(x_i^2 - 2ax_i + a^2\right)\right\} - \sum_i \gamma_i \xi_i \tag{3}$$

with Lagrange multipliers $\alpha_i \ge 0$ and $\gamma_i \ge 0$. Setting the partial derivatives to 0, new constraints are obtained:

$$\sum_{i=1}^{l} \alpha_i = 1, \; a = \frac{\sum_{i=1}^{l} \alpha_i x_i}{\sum_{i=1}^{l} \alpha_i} = \sum_i^1 \alpha_i x_i$$

$$C - \alpha_i - \gamma_i = 0 \quad \forall_i \tag{4}$$

Since $\alpha_i \ge 0$ and $\gamma_i \ge 0$ we can remove the variables γ_i from the third equation in (4) and use the constraints $0 \le \alpha_i \le C \quad \forall_i$.

Rewriting Eq. (3) and resubstituting Eq. (4) give to maximize with respect to α_i:

$$L=\sum_{i=1}^{l}\alpha_i\left(x_ix_i\right)-\sum_{ij}^{1}\alpha_i\alpha_j\left(x_ix_j\right) \tag{5}$$

with constraints $0\le\alpha_i\le C$, $\sum_{i=1}^{l}\alpha_i=1$

The second equation in (4) states that the center of the sphere is a linear combination of data objects, with weight factors α_i , which are obtained by optimizing Eq. (5). Only for a small set of objects the equality in Eq. (2) is satisfied: these are the objects, which are on the boundary of the sphere itself. For those objects the coefficients α_i will be non-zero and are called the support objects. Only these objects are needed in the description of the sphere. The radius R of the sphere can be obtained by calculating the distance from the center of the sphere to a support vector with a weight smaller than C . Objects for which α_i . C have hit the upper bound in Eq. (4) and are outside the sphere. These support vectors are considered to be outliers.

To determine whether a test point z is within the sphere, the distance to the center of the sphere has to be calculated. A test object z is accepted when this distance is smaller than the radius, i.e., when $\left(z-a\right)^T-\left(z-a\right)\le R^2$. Expressing the center of the sphere in terms of the support vectors, we accept objects when

$$\left(z.z\right)-2\sum_{i=1}^{l}\alpha_i\left(z.x_i\right)+\sum_{i=1}^{l}\alpha_i\alpha_j\left(x_i.x_j\right)\le R^2 \tag{6}$$

The method just presented only computes a sphere around the data in the input space. Normally, data are not spherically distributed, even when the most outlying objects are ignored. So, in general, we cannot expect to obtain a very tight description. Since the problem is stated completely in terms of inner products between vectors (Eqs. (5) and (6)), the method can be made more flexible, analogous to (Vapnik, 1995). Inner products of objects $(x_i\cdot x_j)$ can be replaced by a kernel function $K(x_i,x_j)$ when this kernel $K(x_i,x_j)$ satisfies Mercer's theorem. This implicitly maps the objects x_i into some feature space and when a suitable feature space is chosen, a better, more tight description can be obtained. No explicit mapping is required; the problem is expressed completely in terms of $K(x_i,x_j)$ .This in general does not give a tight description of the dataset, hence a kernel is used and most generally a Gaussian kernel is used to get a tight or a superior description as in comparison to polynomial kernel.

Therefore, we replace all inner products $(x_i\cdot x_j)$ by a proper $K(x_i,x_j)$ and the problem of finding a data domain description is now given by (see (5))

$$L=\sum_{i=1}^{l}\alpha_iK\left(x_ix_i\right)-\sum_{ij}^{1}\alpha_i\alpha_jK\left(x_ix_j\right) \tag{7}$$

Using a Gaussian kernel

$$K_G(x_i-x_j)=\exp(-(x_i-x_j)^2/s^2) \tag{8}$$

the above equation will be

$$L=1-\sum_{i=1}^{j}\alpha_i^2-\sum_{i\neq j}\alpha_i\alpha_jK_G(x_i,x_j)$$

and Eqn.(6), becomes

$$-2\sum_i\alpha_iK_G(z,x_j)\le R^2-C_X-1 \tag{9}$$

where C_X (trade-off parameter) depends on support vectors and α_i .

The Gaussian kernel contains one parameter s called as width parameter in Eq. (8). It is noticed that with smaller value of s the boundary description will be tighter and with larger it will be more like a sphere.

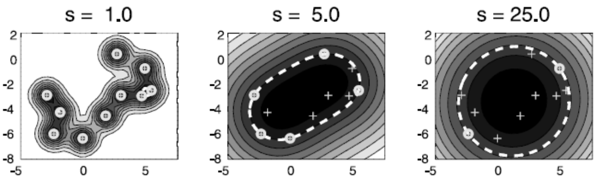

Fig 4. With width parameter s = 1.0 (smaller value) the boundary is very tight and the support vectors are more in comparison to of s = 25.0 (larger value).

In general a good representation of the target class and the outlier class can be identified.

4. QUALITY CONTROL ANALYSIS

When quality of a final product in an industry depends on more than two variables and when the depending variables are large in numbers and they do not follow a particular distribution the study of quality characteristic is be very difficult in simple statistical methods. In multivariate normal distribution the dataset give an elliptical shape and in multidimensional they give an ellipsoid look.

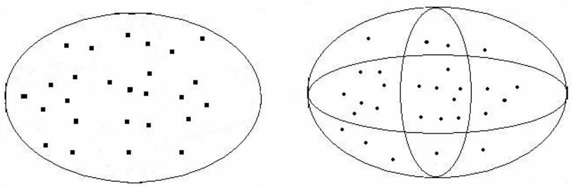

Fig 5. Dataset in a bivariate, (left) and a multivariate, (right) normal distribution.

To get the shape of a distribution of dataset with nonparametric distribution is difficult. The target class and the outlier in the distribution in multidimensional case can be well judged by using support vector data description. Even the percentage of data outside the boundary can be calculated by SVDD.

5. CASE STUDY

A study of raw data [15] of an electrolysis process is taken for the analysis. The samples are taken during one year of copper production, and two samples are taken each day. Each cathode was left in the electrolyte for 10 days and during this period the amount of copper was allowed to grow continually. For each sample the levels of eight metal impurities Ag, Ni,Pb, Bi, Sb, As, Te, and Se are recorded. The impurities should not more than 5%. Here the data are of 370 observations and with 8 variables (370×8).

6. FEATURES EXTRACTION

The SVDD is applied to the dataset and outliers are detected. In figure (6) distance of different points from kernel center is shown. Outliers are clearly visible in the figure (9). Number of support vectors for different value of beta (Lagrangian multiplier) is also shown in the figure (6).

In this analysis the trade off parameter as in the figure (7) below, the number of data points inside and outside the support vector boundary with their radius shows the percentage of data to be considered as outliers.

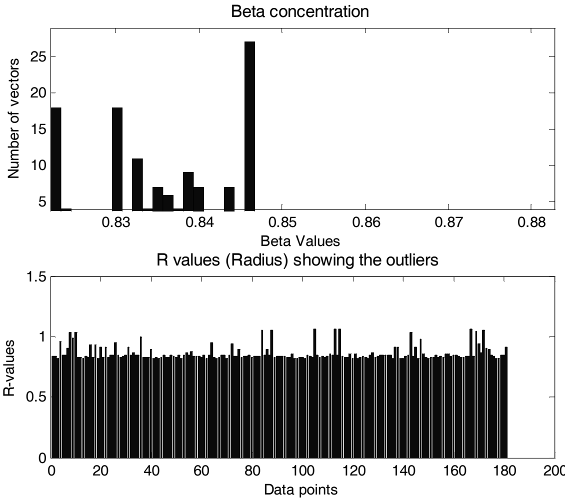

Fig 6. Number of vectors for different values of beta, (upper) and distance of data from kernel centre (R-values).

The SVDD analysis can well be presented and understood in a two dimensional case. In figure (8) all 8 variables are considered and the support vectors are shown with $\otimes$ and data within the support vectors are with O sign. The outliers are shown with $+$ sign. Here it is difficult to understand the support vectors as all 8 variables are represented in a two dimensional figure.

In figure (10) first and second variables and in figure (11) second and third variables are represented. Here the support vector boundary can well be drawn classifying the target and outliers class.

In our analysis the trade-off parameter C_X is taken as 0.25 to get a tight descript boundary. For a good data description, two requirements have to be fulfilled: (1) a low target rejection rate and (2) a low outlier acceptance rate. When we are given only examples of the target set, the first term can be estimated by the number of support vectors that we obtain in the minimization of Lagrangian (17).

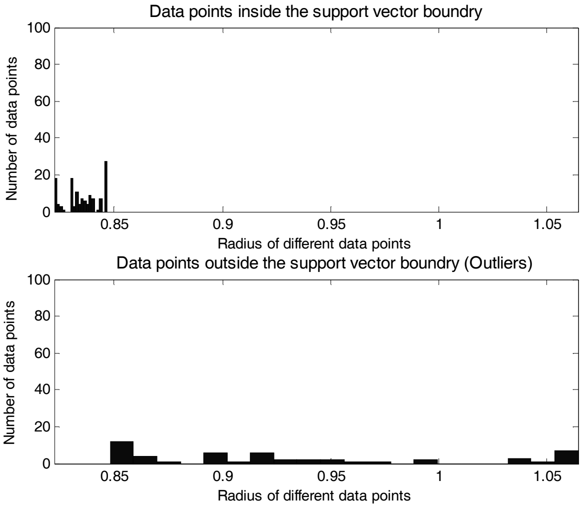

Fig 7. Data inside and outside the support vector boundary with their radius (kernel distance).

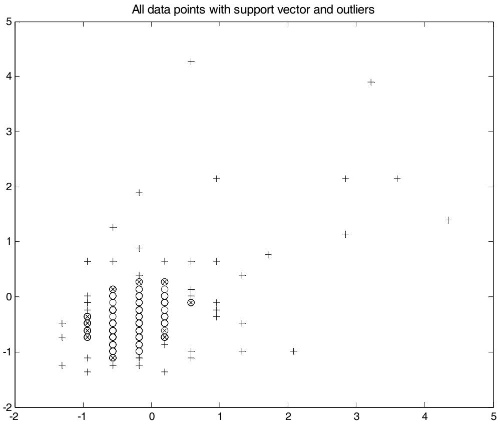

Fig 8. All eight variables are considered and support vectors classify the target and outliers class.

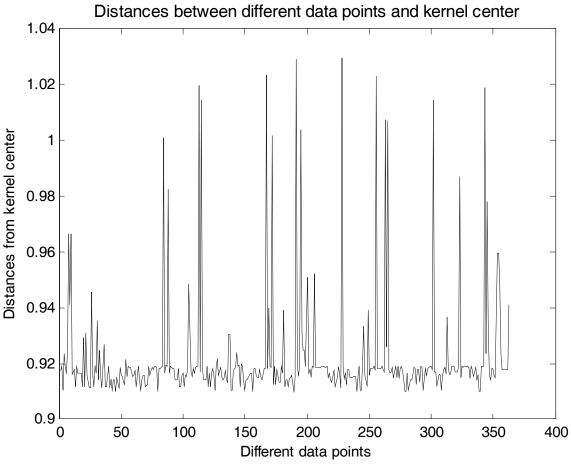

Fig 9. All variables with 370 samples are projected with their kernel distances. The visible peaks are the outliers.

7. CONCLUSIONS

Best representation of the electrolysis process data is made by support vector data analysis and the target class and outliers are found out. Statistics does not help for higher degree of data with nonparametric distribution. Hence in industrial process where the process variables are many and do not follow a particular distribution SVDD can play as a good decision maker to identify the outliers so that the process can be modified to get good quality of product.

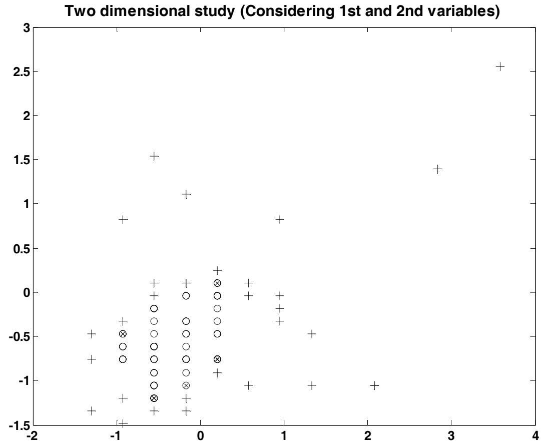

Fig 10. First and second variables are considered and support vectors classify the target and outliers class.

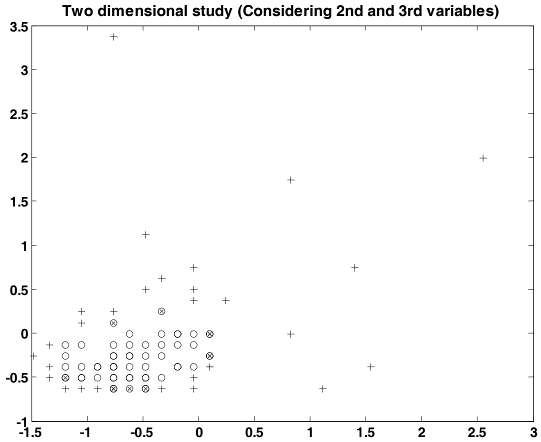

Fig 11. Second and third variables are considered and support vectors classify the target and outliers class.

In this paper the work of identification of target class and outliers is explained but the work can be extended to find out the percentage of data rejection with respect to the trade-off parameter. This Gaussian kernel contains one extra free parameter, the width parameter s in the kernel (from definition (8)). For small values of s the SVDD resembles Parzen density estimation, while for large s the original hypersphere solution is obtained [6]. As shown in [6] this parameter can be set by setting a priori the maximal allowed rejection rate of the target set, i.e. the error on the target set. Secondly, we also have the trade-off parameter C. We can define a new variable:

$$\upsilon = \frac{1}{NC} \tag{10}$$

Scholkopf [17] showed that this is an upper bound for the fraction of objects outside the description. The exact influence of s and (or C) on the SVDD is to be investigated.

8. REFERENCES

1. Hawkins, D.M., 1980, *Identification of Outliers*, Chapman and Hall, London.
2. Vladimir N. Vapnik., 1995, *The Nature of Statistical Learning Theory*, Springer, New York.
3. C. Cortes and V. Vapnik, "Support-vector networks", Machine Learning, 20:273-297, November 1995.
4. N.Crisianini, J.Shawe-Taylor, 2000, *An introduction to Support Vector Machines and other kernel-based learning methods*, Cambridge University Press.
5. Scholkopf, B. Mika, S. Burges et.al, 1999, "Input space versus feature space in kernel-based methods", IEE transactions on neutral networks, 10, 1000-1017.
6. David M.J. Tax, Robert P.W. Duin, 2004, "Support Vector Data Description", Machine Learning, 54, 45–66.
7. David M.J. Tax, Robert P.W. Duin, 1999, "Support vector domain description", Pattern Recognition Letters 20 1191-1199.
8. D. M. J.Tax, A.Ypma, and R.P.W. Duin, 1999a, "Pump failure detection using support vector data description", *Proceedings of the Third International Symposium on Advances in Intelligent Data Analysis*, Amsterdam, Netherlands, pp. 415–425.
9. David M.J. Tax, Alexander Ypma and Robert P.W. Duin, 1999b, "Support Vector Data Description applied to machine vibration analysis", *Proceedings of the Fifth Annual Conference of the Advanced School for Computing and Imaging*, Heijen, Netherlands, pp. 398–405.
10. David M.J. Tax, Robert P.W. Duin, 1999, "Data Domain Description using Support Vectors", *ESANN'1999 proceedings-European Symposium on Artificial Neural Networks*, Bruges (Belgium), D-Facto public, ISBN 2-600049-9-X, pp. 251-256.
11. Ritter, G., Gallegos, M.T., 1997, "Outliers in statistical pattern recognition and an application to automatic chromosome classification" Pattern Recognition Letters 18, 525-539.
12. Tax, D. Duin, R., 1998. Outlier detection using classifier instability. In: Amin, A., Dori, D., Pudil, P., Freeman, H. (Eds.), *Advances in Pattern Recognition Proc. Joint IAPR Internat. Workshops SSPR'98 and SPR'98*, Sydney, Australia. Lecture Notes in Computer Science, Vol. 1451. Springer, Berlin, pp. 593-601.
13. Tax, D., Duin, R., 1999, "Data domain description using support vectors", *In: Verleysen, M. (Ed.), Proc. European Symposium Artificial Neural Network*, D. Facto, Brussel, pp. 251-256.
14. Vapnik V., 1995, *The Nature of Statistical Learning Theory*, Springer, New York.
15. Ypma, A., Pajunen, P., 1999, "Rotating machine

vibration analysis with second-order independent component analysis", *In: Proc. 1st Internat. Workshop Independent Component Analysis and Signal Separation*, ICA'99, pp. 37-42.

16. Repository of test dataset for multivariate quality control, http://www.umetrics.com/methodtech_resources.asp.
17. David M.J. Tax, Robert P.W. Duin, *Outliers and data description*.
18. B. Scholkopf P. Bartelett, A.J. Smola and R. Williumson. Shrinking the tube: A new support vector regression algorithm. M.S. Kearns, S.A. Solla, and D.A. Cohn, editors, Advances in Neural, 1999.