International Journal of Soft Computing 2 (2): 307-312, 2007

Automatic Model Selection for One-Class SVM

[1]Qun Chang, [1]Xiaolong Wang, [2]Yimeng Lin and [3]Daniel S. Yeung
[1]Department of Compute Science and Technology, Harbin Institute of Technology, Harbin, 150001, China
[2]MiLeS Lab, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, 518055, China
[3]Department of Computing, Hong Kong Polytechnic University, Kowloon, Hong Kong, China

Abstract: One-class SVM splits the feature space into two parts by a sphere. Inside or on the sphere are the normal data objects while outside the sphere are outliers. Since no labels are available for normal data and outliers, training one-class SVM belongs to unsupervised learning. In previous research the model parameters for one-class SVM are chosen *a priori* and manually and in any case, such dealings are not persuasive as an underlying mechanism. We approach this problem by iteratively optimizing the primary objective function through genetic algorithms and thus automatically implement model selection for one-class SVM. The model selection procedure embodies the principle of structural risk minimization for one-class SVM. The algorithms and their performance are validated by experiments.

Key words: One-class SVM, model selection, genetic algorithms, outliers

INTRODUCTION

One-class SVM splits the feature space into two parts by a sphere Scholkopf *et al.* (1999; 2001) D.M.J. (1999), The part inside or on the sphere contains most of the data objects which are called normal data, while the part outside the sphere covers the unrepresentative data objects which are called outliers. The mechanism is illustrated with Fig. 1.

Given m data patterns x_i, i = 1,..., m, the sphere is obtained by an optimization problem which is

$$\begin{aligned} \min \quad & R^2 + \frac{1}{vm}\sum_i \xi_i \\ \text{s.t.} \quad & 0 < v \le 1,\ \xi_i \ge 0, \\ & \left\|\Phi(x_i) - c\right\|^2 \le R^2 + \xi_i \end{aligned} \tag{1}$$

where R is the sphere's radius, c is the sphere's center and ξ_i are the slack variables which allow for the outliers outside the sphere. The parameter v controls the tradeoff between the minimum volume of the sphere and the number of the outliers (Scholokpf *et al.*, 2001; Scholkpf and Smola, 2002). The above problem leads to the dual,

$$\begin{aligned} W(\mathbf{a}): \quad & \min \quad \sum_{i,j}\alpha_i\alpha_j k(\mathbf{x}_i,\mathbf{x}_j) \\ & \text{s.t.} \ \ 0 < \alpha_i \le \frac{1}{vm} \ \text{and} \ \sum_i \alpha_i = 1 \end{aligned} \tag{2}$$

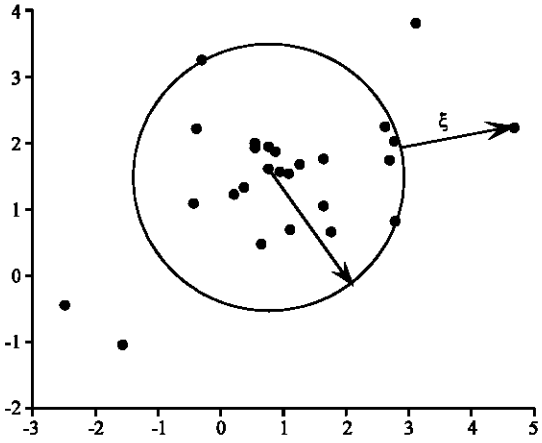

Fig. 1: One-class SVM. Outside the sphere are outliers and the rest are normal data

if the Gaussian kernel is adopted which is in the form of

$$\begin{aligned} & k(\mathbf{x}_i,\mathbf{x}_j) = \Phi(\mathbf{x}_i)^T\Phi(\mathbf{x}_j) = \exp \\ & \left(-\lambda\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2\right),\ \lambda > 0 \end{aligned} \tag{3}$$

One-class SVM has its applications in data domain description (D.M.J., 2004; 1999; Clustering Ben-Hur *et al.*, 2001; Chiang and Hao, 2003) novelty detection (Scholkopf *et al.*, 2003; Cao *et al.*, 2003), distribution estimation (Scholkopf *et al.*, 2001), classification, (D.M.J., 2001; Manevitz and Yousa, 2001), so on. Here the model

Corresponding Author: Qun Chang, Department of Compute Science and Technology, Harbin Institute of Technology, Harbin, 150001, China

parameters for one-class SVM are v and λ. The previous studies have explored the parameters' role and their effects on the sphere and nevertheless, the model parameters were empirically and manually assigned. This syudy proposes an automatic model selection method for one-class SVM, which is based on the genetic algorithms by optimizing the objective function (1). The rest of the syudy is organized as follows.

Model selection for one-class SVM: While v and λ (Gaussian kernel) are assigned to some values, it is easy to obtain a solution for one-class SVM (Scholkopf and Somla, 2002). Hereof the solution for the center c is

$$c = \sum_i \alpha_i \Phi(\mathbf{x}_i) \tag{4}$$

corresponding to the decision function of the form

$$\begin{aligned} f(\mathbf{x}) &= \operatorname{sgn}\left(R^2 - \left\|\Phi(\mathbf{x}) - c\right\|^2\right) \\ &= \operatorname{sgn}\left(R^2 - \sum_{i,j} \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) + 2\sum_i \alpha_i k(\mathbf{x}_i, \mathbf{x}) - 1\right) \end{aligned} \tag{5}$$

where R^2 is computed such that for any x_i with $0 < \alpha_i < 1/(vm)$ the argument of the sgn is zero. The pattern x_i with $\alpha_i \neq 0$ are called support vectors. As stated in (Scholkopf and Smola, 2002; Scholkopf *et al*., 2001) the property of v is that

- v is an upper bound on the fraction of outliers.
- v is a lower bound on the fraction of support vectors.
- Suppose the data were generated independently from a distribution which does not contain discrete components. Suppose, moreover, that the kernel is analytic and non-constant. With probability 1, asymptotically, v equals both the fraction of support vectors and the fraction of outliers.

Always based on the above proposition, *v* was assigned a *priori* in model selection (D.M.J., 1999; Ben-hur *et al*., 2001; Scholkopf *et al*., 2001; Unnporsson *et al*., 2003) Generally, in fact, v was always discussed in the condition of $v \geq 1/m$. If $v < 1/m$, it seems not allowing for the outliers' existence. However, our automatic model selection algorithms show that the optimal value for v tends to drop in $v < 1/m$. Note: another version for *v* is the penalty factor C which just replaces v by $C = 1/vm$, so both have the same use and we only discuss v in this syudy. Here we ask: what is the criterion to choose the v's value? To this end, we consider to choose a function, from a hypothetical functional space, which can catch most of the data domain. The basic intuition here is that if the functional space is too large, in the sense of containing too many wild functions, it is impossible to construct any useful approximating function using empirical data (Poggio *et al*., 2004). So, minimizing the radius of the sphere containing the data objects is to restrict the wild functions and obtain a tight data domain description, that is, the complexity of the approximating function is minimized. On the other side, the outliers don't belong to the normal data, so they shouldn't affect the selection of the approximating function, that is, the empirical risk is minimized. At this point, just as in two-class SVM (Vladimir, 1995; 1998), the optimization problem in one-class SVM is also based on a Structural Risk Minimization (SRM) principle which defines a trade-off between the complexity of the approximating function and the empirical risk. This tradeoff is controlled by v, which should be chosen in consistence with the principle of SRM. Of course, the kernel parameter λ also affects the structure of the approximating function and the scale of empirical errors of the outliers, so is a model parameter. Thus, it is a natural idea to take (1) as the objective function for one-class SVM model selection, that is, the objective function is

$$B = R^2 + \frac{1}{vm}\sum_i \xi_i \tag{6}$$

We propose automatic model selection for one-class SVM. The iterative procedure is described as follows.

- Initialize v and λ to some values.
- Using standard algorithms for one-class SVM by optimizing (2), find

$$\mathbf{a}^0 = \arg\min_{\mathbf{a}} W(\mathbf{a}, \lambda, v)$$

- Using standard genetic algorithmsholdberg, 1989, to optimize the objective function 6, cf. 1, find

$$(\lambda^0, v^0) = \arg\min_{\lambda, v} B(\mathbf{a}, \lambda, v)$$

4 Go to 2 or stop the iterative procedure according to some stop condition on *B*.

Now we discuss how to compute the B in each iterative step. For convenience, several notations are introduced as follows. In some iterative step, let N_{sv}, N_{usv} and N_{bsv} denote the number of Support Vectors (SVs), the number of Unbounded Support Vectors (USVs) on the sphere with $0 < \alpha_i < 1/(vm)$, the number of Bounded Support Vectors (BSVs) outside the sphere with $\alpha_i = 1/vm$, respectively. Obviously, $N_{sv} = N_{usv} + N_{bsv}$. Let i (or j), u and

b index SV, USV and BSV, respectively. In the step 3), we use all the USVs to obtain the mean approximating value for R^2. That is,

$$R^2 = \frac{1}{N_{usv}} \sum_u \left\| \Phi(\mathbf{x}_u) - c \right\|^2 = \frac{1}{N_{usv}} \sum_u \left(1 - 2\sum_i \alpha_i k(\mathbf{x}_i, \mathbf{x}_u) + \sum_{i,j} \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) \right) \quad (7)$$

Considering (1) and (7), we have

$$\xi_{bsv} = \left\| \Phi(\mathbf{x}_{bsv}) - c \right\|^2 - R^2 = 1 - 2\sum_i \alpha_i k(\mathbf{x}_i, \mathbf{x}_{bsv}) + \sum_{i,j} \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) - R^2 \quad (8)$$

In view of (7) and (8), (6) is easily computed. In fact, (6) can be regarded as the generalization error bound for one-class SVM.

Outliers problem: Though one-class SVM has different applications, e.g., clustering, distribution estimation, novelty detection and classification, all the applications are summarized to be a task which aims to separate the normal data from the outliers. What are outliers? Ripley (1996), defines outliers as examples which did not (or thought not to have) come from the assumed population of examples. Bernett and Lewis (1994), have almost the same definition for outliers: An observation (or subset of observations) which appear to be inconsistent with the remainder of the set of data. Most definitions of outliers regard outliers as unrepresentative data which seems from a different distribution than the normal data. The outliers' problem also holds for one-class SVM. That is, we don't know what proportion of outliers in one dataset is. This specifies training one-class SVM to be unsupervised learning. At this point, the automatic model selection for one-class SVM is more reasonable than *a priori*-oriented.

Experiments: The model selection for one-class SVM (unsupervised learning) is more difficult than for two-class SVM (supervised learning). The reason also comes from how to test the validity and the performance of the model selection algorithms for one-class SVM? Accuracy is useless for (unsupervised learning) one-class SVM. However, in order to validate the model selection algorithms, we have to resort to the supervised learning elements, i.e. accuracy, training dataset and testing dataset, to test the validity and the performance of our idea. The trick is as follows. We artificially construct a dataset including the normal data and outliers and also, the number of normal data objects is significantly bigger than that of the outliers. Then a sphere is obtained by one-class SVM training algorithms. Obviously, most of normal data objects will be inside or on the sphere while the outlier region is outside the sphere. The accuracy is computed in this way: The same dataset is predicted by the just obtained sphere. That is, the sum of the number of the normal data objects dropping inside or on the sphere and the number of the outliers dropping outside the sphere is divided by the total number of the dataset objects and then the result is accuracy. Of course, the error rate is the difference between 1 and accuracy. The model selection algorithms adopt as subroutines two sets of software. One is the genetic algorithms from Matlab 7.0 tool box. The setting is that each generation has 20 individuals and the maximal iterations (also the stop condition for model selection) are 50. The other is the standard one-SVM algorithms from LIBSVM Chih-chung-Chang and Chih-Jin-Lin, 2001.

Artificial dataset: The artificial data is generated from Gaussian distribution. The normal data with mean vector (1, 1) has 500 patterns and the outlier data has 10 patterns with mean vector (-2, 0) and 10 patterns with mean vector (4, 2). Both normal data and outlier data have the variance with 1. The artificial data is illustrated in Fig. 2.

The curves of error rate and object function (cf. (6)) following the iterative step may refer to Fig. 3, 4 and 5, respectively.

Figure 3-5 and Table 1 show that though different initial parameter (v, λ) value pairs correspond to different optimal parameter value pairs, the final accuracies with different optimal parameter value pairs

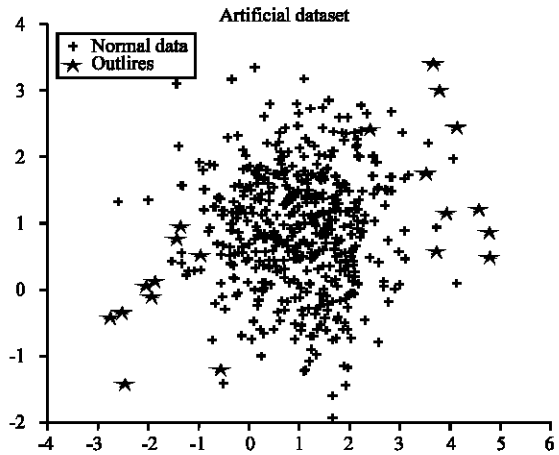

Fig. 2: Artificial data with 500 normal patterns and 20 outliers

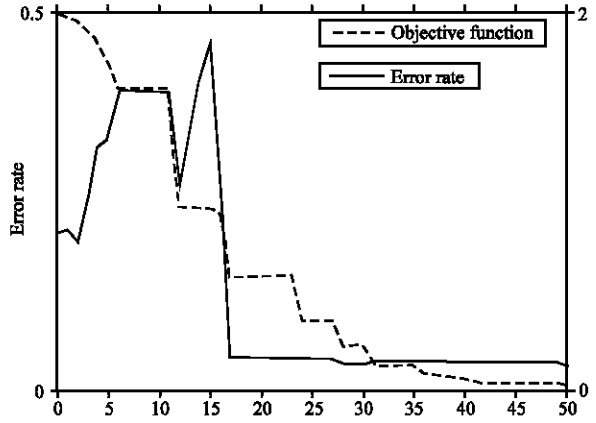

Fig. 3: The initial values for (v, λ) are (0.01, 1), respectively

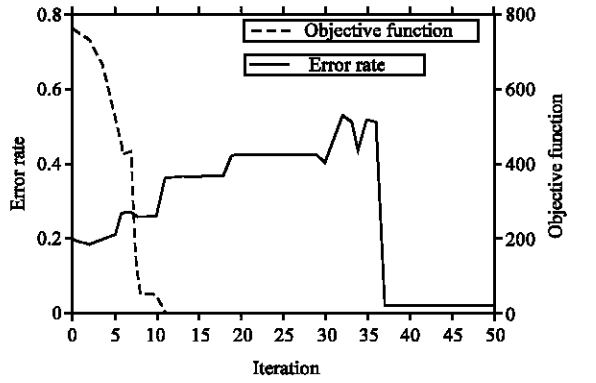

Fig. 4: The initial values for (v, λ) are (0.2,1), respectively

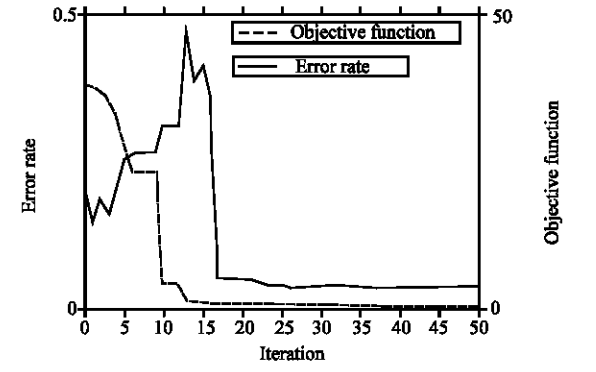

Fig. 5: The initial values for (v, λ) are (0.05,1), respectively

Table 1: Model selection results with different initial values on artificial dataset

Initial values (v, λ)	Op_v (v, λ)	O_v B	Accuracy
(0.01,1)	(0.0019,0.0005)	0.0177	97.8846%
(0.05,1)	(0.0019,0.0010)	0.0341	97.8846%
(0.1,1)	(0.0019,0.0566)	0.6377	97.5%
(0.2,1)	(0.0008,0.0020)	0.3952	97.8846%
(0.4,1)	(0.0018,0.0010)	0.0376	97.8846%

Op_v: Optimal parameter values; O_v: Objective value

give almost the same accuracy when model selection is finished. That is, the model selection algorithms converge to a best accuracy, though different optimal parameter values exist. More important is that after 40 iterations in model selection, the curves of objective function and error rate tend to be parallel. This is enough for model selection algorithms which shows that the optimal parameters belong to a region and not a point. Also, 50 as the maximum of iterations is a reasonable stop condition for model selection.

Heart disease: Heart disease data is a benchmark dataset (Brazdil and Statlog Databset, 2005) which has 13 attributes, 120 positive samples and 150 negative samples. Here with 150 negative samples taken as normal data, the outlier class are randomly selected from the positive class which has 120×0.08%≈10, 120×0.06%≈7, 120×0.04%≈5 outliers in test, respectively. In the tests the initial values of (v, λ) are set to be (0.01, 1). The test results are illustrated in Fig. 6, 7, 8 and Table 2.

Figure 3 to 8 show that in the initial iterative process the error rate curve don't descend monotonically with the descent of the objective function. This

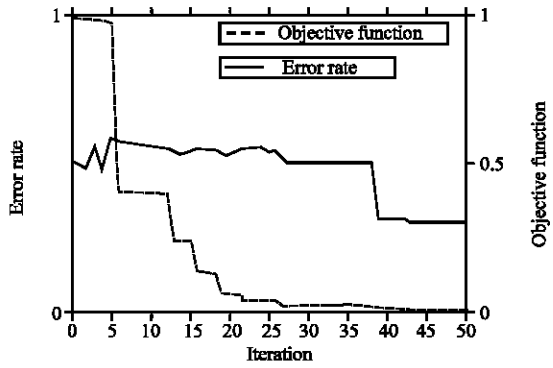

Fig. 6: Model selection with 10 outliers

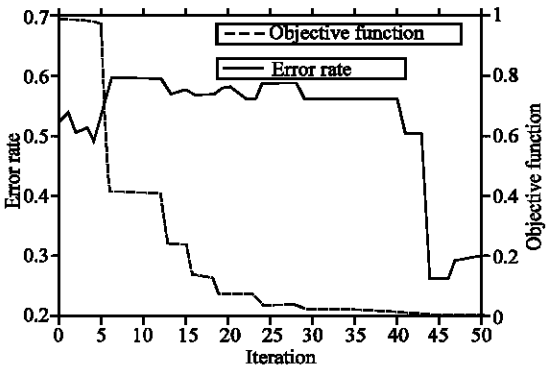

Fig. 7: Model selection with 7 outliers

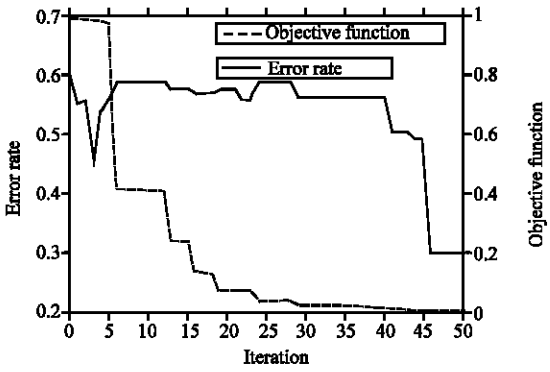

Fig. 8: Model selection with 5 outliers

Table 2: Model selection results with different number of outliers on Heart disease data

Initial values (v, λ)	Op_v (v, λ)	Outliers B	O_v	Accuracy
(0.01,1)	(0.0062,0.0001)	10	0.0019	70%
(0.01,1)	(0.0064,0.0001)	7	0.0018	70%
(0.01,1)	(0.0065,0.0001)	5	0.0013	70%

Op_v: Optimal parameter values; O_v: Objective value

phenomenon is from the nature of the iterative algorithms. The model selection algorithms alternatively optimize the model parameters (v, λ) and the coefficients (α_i) for support vectors. When the sphere is found in one iteration, we assume that the support vectors (i.e. coefficients) are unchangeable and only based on this assumption, the model parameters (v, λ) are optimized. However, such assumption is far from what the fact is in the initial iterative process. Nevertheless, the assumption asymptotically holds with the iteration procedure runs. When the iterative steps reach nearly 50, the error rate is asymptotically convergent to a stable value and this is what we want. Besides, the optimal model parameters are far less than the previous work assumed, i.e., $v << 1/m$.

Outlier detection: After the model selection algorithms are validated by the above constructive experiments, the outlier detection on Heart Disease is implemented automatically through model selection algorithms. As seem from the above experiments, the initial values for model parameters affect the objective function but not the outlier detection accuracy. So, we set the initial values (0.01, 1) for (v, λ) and 50 as maximum iterations as stop condition. After model selection, we predict that there are 12 outliers in 270 samples. These 12 outliers are more reasonable than did the previous work by manually stipulating the percentage of the outliers.

CONCLUSION

Training one-class SVM is a process of unsupervised learning. In previous research the model parameters for one-class SVM are chosen *a priori* and manually and in any case, such dealings are not persuasive as an underlying mechanism. The automatic model selection algorithms we proposed embody, in nature, the principle of SRM. On the other side, the SRM principle rationalizes the model selection algorithms as an underlying mechanism for one-class SVM. The optimal value of v is always less than 1/m. This phenomenon is counter-intuitive since discussing the problem with $v < 1/m$ seemed very popular in the previous research. In experiments, that the error rate curve converges to a smallest value demonstrates the validity and excellent performance of the model selection algorithms. It seems not very necessary for v and λ to reach the precise values and just a good value pair is sufficient to construct one-class SVM. With reasonable model selection algorithms, one-class SVM will have better applications in multi-fields.

ACKNOWLEDGMENT

This work is supported by the key program project (No. 60435020) of National Natural Science Foundation of China (NSFC).

REFERENCES

Barnett, V. and T. Lewis, 1994.Outliers in statistical data. John Wiley and Sons, New York.

Ben-hur, A., D. horn, H.T. Siegelmann and V. Vapnik. 2001. Support vector clustering. J. Mach. Learining Res., 2: 125-137.

Brazdil, P. and Statlog Dataset, 2005. http: //www. niaad. liacc. up. pt/old/statlog.

Cao, L.J., H.P. Lee and W.K. Chong, 2003. Modified support vector novelty detector using training data with outliers. Pattern Recog. Lett., 24: 2479-2487.

Chih-Chung Chang and Chih-Jen Lin, 2001. LIBSVM : A library for support vector machines. Software available at http://www.csie.ntu.edu.tw/ ~cjlin/libsvm

Chiang, J.H. and P.Y. Hao, 2003. A new kernel-based fuzzy clustering approach: support vector clustering with cell growing. IEEE. Trans. Fuzzy Systems, Vol., 11.

Goldberg, E., 1989. Algorithms in Search, Optimization and Machine Learning. Addison-Wesley Pub.

Manevitz, L.M. and M. Yousef, 2001. One-class SVMs for document classification. The J. Mach. Learn. Res., 2: 139-152.

Poggio, T., R. Rifkin, S. Mukherjee and P. Niyogi, 2004. General conditions for predictivity in learning theory. Nature, 428: 419-422.

Ripley, B.D., 1996. Pattern recognition and neural networks. Cambridge University Press.

Schölkopf, B., R. Williamson, A.J. Smola and J. Shawe-Taylor.,1999. Single-class support vector machines. In J. Buhmann, W. Maass, H. Ritter and N. Tishby, (Eds.), Unsupervised Learning, Dagstuhl-Seminar-Report, 235: 19-20.

Schölkopf, B., R. Williamson, A. Smola, J. Shawe-Taylor and J. Platt. 2000. Support vector method for novelty detection. Advances in Neural Inform. Process. Sys., 12: 582-588,

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

Schölkopf, B. and A.J. Smola, 2002. Learning with kernels. MIT Press, Cambridge, MA.

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

Tax, D.M.J. and R.P.W. Duin, 2004. Support vector data description. Mach. Learn., 54: 45-66.

Tax, D.M.J., 2001. One-class classification. Ph. D Thesis, Delft University of Technology. Available at: http://www.ph.tn.tudelft.nl/~davidt/thesis.pdf.

Unnþórsson, R., T.P. Runarsson and M.T. Jonsson, 2003. Model selection in one-class v-SVMS using RBF kernels. Proceedings of COMADEM, Vaxjo University Press, Sweden, pp: 135-145.

Vladimir, N., 1995. Vapnik. The Nature of Statistical Learning Theory. New York: Springer-Verlag.

Vladimir, N., 1998. Vapnik. Statistical Learning Theory. New York: Wiley.