Ulrich Flegel, Michael Meier (Eds.)

Detection of Intrusions and Malware & Vulnerability Assessment

GI Special Interest Group SIDAR Workshop, DIMVA 2004
Dortmund, Germany, July 6-7, 2004
Proceedings

Gesellschaft für Informatik 2004

Lecture Notes in Informatics (LNI) - Proceedings
Series of the Gesellschaft für Informatik (GI)

Volume P-46

ISBN 3-88579-375-X
ISSN 1617-5468

Volume Editors
Ulrich Flegel
University of Dortmund,
Computer Science Department, Chair VI, ISSI
D-44221 Dortmund, Germany
ulrich.flegel@udo.edu
Michael Meier
Brandenburg University of Technology Cottbus,
Computer Science Department, Chair Computer Networks
P.O. Box 10 13 44, D-03013 Cottbus, Germany
mm@informatik.tu-cottbus.de

printed by Köllen Druck+Verlag GmbH, Bonn

Intrusion detection in unlabeled data with quarter-sphere Support Vector Machines

Pavel Laskov and Christin Schäfer
Fraunhofer-FIRST
Kekuléstr. 7
12489 Berlin, Germany
{laskov,christin}@first.fhg.de

Igor Kotenko
SPIIRAS
14th Liniya 39
199178 St. Petersburg, Russia
ivkote@spiiras.nw.ru

Abstract: Practical application of data mining and machine learning techniques to intrusion detection is often hindered by the difficulty to produce clean data for the training. To address this problem a geometric framework for unsupervised anomaly detection has been recently proposed. In this framework, the data is mapped into a feature space, and anomalies are detected as the entries in sparsely populated regions. In this contribution we propose a novel formulation of a one-class Support Vector Machine (SVM) specially designed for typical IDS data features. The key idea of our "quarter-sphere" algorithm is to encompass the data with a hypersphere anchored at the center of mass of the data in feature space. The proposed method and its behavior on varying percentages of attacks in the data is evaluated on the KDDCup 1999 dataset.

1 Introduction

The majority of current intrusion detection methods can be classified as either misuse detection or anomaly detection [NWY02]. The former identify patterns of known illegitimate activity; the latter focus on unusual activity patterns. Both groups of methods have their advantages and disadvantages. Misuse detection methods are generally more accurate but are fundamentally limited to known attacks. Anomaly detection methods are usually less accurate than misuse detecion methods — in particular, their false alarm rates are hardly acceptable in practice — however, they are at least in principle capable of detecting novel attacks. This feature makes anomaly detection methods the topic of active research.

In some early approaches, e.g. [DR90, LV92], it was attempted to describe the normal behavior by means of some high-level rules. This turned out to be quite a difficult task. More successful was the idea of collecting data from normal operation of a system and computing, based on this data, features describing normality; deviation of such features would be considered an anomaly. This approach is known as "supervised anomaly detection". Different techniques have been proposed for characterizing the concept of normality, most notably statistical techniques, e.g. [De87, JLA+93, PN97, WFP99], and data mining techniques, e.g. [BCJ+01, VS00]. In practice, however, it is difficult to obtain clean data to implement these approaches. Verifying that no attacks are present in the training data may

be an extremely tedious task, and for large samples this is infeasible. On the other hand, if the "contaminated" data is treated as clean, intrusions similar to the ones present in the training data will be accepted as normal patterns.

To overcome the difficulty in obtaining clean data, the idea of *unsupervised* anomaly detection has been recently proposed and investigated on several intrusion detection problems [PES01, EAP+02, LEK+03]. These methods compute some relevant features and use techniques of unsupervised learning to identify sparsely populated areas in feature space. The points — whether in the training or in the test data — that fall into such areas are treated as anomalies.

More precisely, two kinds of unsupervised learning methods have been investigated: clustering methods and one-class SVM. In this contribution we focus on one-class SVM methods and investigate the application of the underlying geometric ideas in the context of intrusion detection.

We present three formulations of one-class SVM that can be derived following different geometric intuitions. The formulation used in previous work was that of the hyperplane separating the normal data from the origin [SPST+01]. Another formulation, motivated by fitting a sphere over the normal data, is also well-known in the literature on kernel methods [TD99]. The novel formulation we propose in this paper is based on fitting a sphere centered at the origin to normal data. This formulation, to be refered to as a *quarter-sphere*, is particularly suitable to the features common in intrusion detection, whose distributions are usually one-sided and concentrated at the origin.

Finally, we present an experimental evaluation of the one-class SVM methods under a number of different scenarios.

2 One-class SVM formulations

Support Vector Machines have received great interest in the machine learning community since their introduction in the mid-1990s. We refer the reader interested in the underlying statistical learning theory and the practice of designing efficient SVM learning algorithms to the well-known literature on kernel methods, e.g. [Va95, Va98, SS02]. The one-class SVM constitutes the extension of the main SVM ideas from supervised to unsupervised learning paradigms.

We begin our investigation into the application of the one-class SVM for intrusion detection with a brief re-capitulation and critical analysis of the two known approaches to one-class SVM. It will follow from this analysis that the quarter-sphere formulation, described in section 2.4, could be better suited for the data common in intrusion detection problems.

2.1 The plane formulation

The original idea of the one-class SVM [SPST+01] was formulated as an "estimation of the support of a high-dimensional distribution". The essense of this approach is to map the data points x_i into the feature space by some non-linear mapping $\Phi(x_i)$, and to separate the resulting image points from the origin with the largest possible margin by means of a hyperplane. The geometry of this idea is illustrated in Fig. 1. Due to nonlinearity of

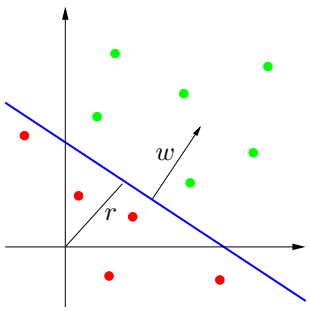

Figure 1: The geometry of the plane formulation of one-class SVM.

feature space, maximization of the separation margin limits the volume occupied by the normal points to a relatively compact area in feature space. Mathematically, the problem of separating the data from the origin with the largest possible margin is formulated as follows:

$$\begin{aligned} \min_{w\in\mathcal{F},\xi\in\mathbb{R}^l,r\in\mathbb{R}} \quad & \tfrac{1}{2}||w||^2 + \tfrac{1}{\nu l}\textstyle\sum_{i=1}^{l}\xi_i - r, \\ \text{subject to:} \quad & (w\cdot\Phi(x_i)) \geq r - \xi_i, \\ & \xi_i \geq 0. \end{aligned} \tag{1}$$

The weight vector w, characterizing the hyperplane, "lives" in the feature space $\mathcal{F}$, and therefore is not directly accessible (as the feature space may be extremely high-dimensional). The non-negative slack variables ξ_i allow for some points, the anomalies, to lie on the "wrong" side of the hyperplane. Instead of the primal problem (1), the following dual problem, in which all the variables have low dimensions, is solved in practice:

$$\begin{aligned} \min_{\alpha\in\mathbb{R}^l} \quad & \textstyle\sum_{ij=1}^{l}\alpha_i\alpha_j k(x_i,x_j), \\ \text{subject to:} \quad & \textstyle\sum_{i=1}^{l}\alpha_i = 1, \\ & 0 \leq \alpha_i \leq \tfrac{1}{\nu l}. \end{aligned} \tag{2}$$

Once the solution α is found, one can compute the threshold parameter $r = \sum_j \alpha_j k(x_i,x_j)$ for some example i such that α_i lies strictly between the bounds (such points are called *support vectors*). The decision, whether or not point x is normal, is computed as:

$$f(x) = \text{sgn}\left(\textstyle\sum_i \alpha_i k(x_i,x) - r\right). \tag{3}$$

The points with $f(x) = -1$ are considered to be anomalies.

2.2 The sphere formulation

Another, somewhat more intuitive geometric idea for the one-class SVM is realized in the sphere formulation [TD99]. The normal data can be concisely described by a sphere (in a feature space) encompassing the data, as shown in Fig. 2. The presence of anomalies in the

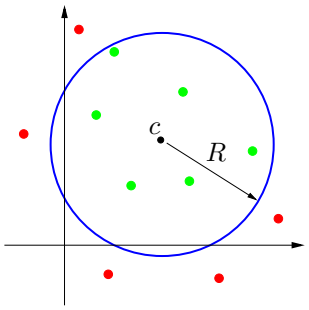

Figure 2: The geometry of the sphere formulation of one-class SVM.

training data can be treated by introducing slack variables ξ_i, similarly to the plane formulation. Mathematically the problem of "soft-fitting" the sphere over the data is described as:

$$\begin{aligned} \min_{R\in\mathbb{R},\xi\in\mathbb{R}^l,c\in\mathcal{F}} \quad & R^2 + \tfrac{1}{\nu l}\textstyle\sum_{i=1}^{l}\xi_i, \\ \text{subject to:} \quad & ||\Phi(x_i) - c|| \leq R^2 + \xi_i, \\ & \xi_i \geq 0. \end{aligned} \tag{4}$$

Similarly to the primal formulation (1) of the plane one-class SVM, one cannot directly solve the primal problem (4) of the sphere formulation, since the center c belongs to the possibly high-dimensional feature space. The same trick can be employed — the solution is sought to the dual problem:

$$\begin{aligned} \min_{\alpha\in\mathbb{R}^l} \quad & \textstyle\sum_{ij=1}^{l}\alpha_i\alpha_j k(x_i,x_j) - \sum_{i=1}^{l}\alpha_i k(x_i,x_i), \\ \text{subject to:} \quad & \textstyle\sum_{i=1}^{l}\alpha_i = 1, \\ & 0 \leq \alpha_i \leq \tfrac{1}{\nu l}. \end{aligned} \tag{5}$$

The decision function can be computed as:

$$f(x) = \text{sgn}\left(R^2 - \textstyle\sum_{ij=1}^{l}\alpha_i\alpha_j k(x_i,x_j) + 2\sum_{i=1}^{l}\alpha_i k(x_i,x) - k(x,x)\right). \tag{6}$$

The radius R^2 plays the role of a threshold, and, similarly to the plane formulation, it can be computed by equating the expression under the "sgn" to zero for any support vector.

The similarity between the plane and the sphere formulations goes beyond merely an analogy. As it was noted in [SPST+01], for kernels $k(x,y)$ which depend only on the difference $x-y$, the linear term in the objective function of the dual problem (5) is constant, and the solutions are equivalent.

2.3 Analysis

When applying one-class SVM techniques to intrusion detection problems, the following observation turns out to be of crucial importance: *A typical distribution of the features used in IDS is one-sided on* $\mathbb{R}_0^+$. Several reasons contribute to this property. First, many IDS features are of temporal nature, and their distribution can be modeled using distributions common in survival data analysis, for example by an exponential or a Weibull distribution. Second, a popular approach to attain coherent normalization of numerical attributes is the so-called "data-dependent normalization" [EAP+02]. Under this approach, the features are defined as the deviations from the mean, measured in the fraction of the standard deviation. This quantity can be seen as F-distributed. Summing up, the overwhelming mass of data lies in the vicinity of the origin.

The consequences of the one-sidedness of the data distribution for the one-class SVM can be seen in Fig. 3. The one-sided distribution in the example is generated by taking the

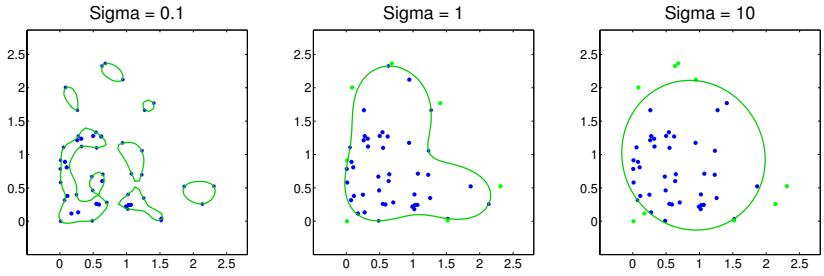

Figure 3: Behavior of the one-class SVM on the data with a one-sided distribution.

absolute values of the normally distributed points. The anomaly detection is shown for a fixed value of the parameter ν and varying smoothness σ of the RBF kernel. The contours show the separation between the normal points and anomalies. One can see that even for the heavily regularized separation boundaries, as in the right picture, some points close to the origin are detected as anomalies. As the regularization is diminished, the one-class SVM produces a very ragged boundary and does not detect any anomalies.

The message that can be carried from this example is that, in order to account for the one-sidedness of the data distribution, one needs to use a geometric construction that is in some sense asymmetric. The new construction we propose here is the quarter-sphere one-class SVM described in the next section.

2.4 The quarter-sphere formulation

A natural way to extend the ideas of one-class SVM to one-sided non-negative data is to require the center of the fitted sphere be fixed at the origin. The geometry of this approach is shown in Fig. 4. Repeating the derivation of the sphere formulation for $c = 0$, the

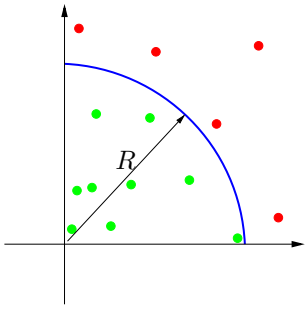

Figure 4: The geometry of the quarter-sphere formulation of one-class SVM.

following dual problem is obtained:

$$\begin{aligned} \min_{\alpha \in \mathbb{R}^l} \quad & -\sum_{i=1}^{l} \alpha_i k(x_i, x_i), \\ \text{subject to:} \quad & \sum_{i=1}^{l} \alpha_i = 1, \\ & 0 \leq \alpha_i \leq \tfrac{1}{\nu l}. \end{aligned} \tag{7}$$

Note that, unlike the other two formulations, the dual problem of the quarter-sphere SVM amounts to a linear rather than a quadratic program. Herein lies the key to the significantly lower computational cost of our formulation.

It may seem somewhat strange that the non-linear mapping affects the solution only through the norms $k(x_i, x_i)$ of the examples, i.e. that the geometric relations *between* the objects are ignored. This feature indeed poses a problem for the application of the quarter-sphere SVM with the distance-based kernels. In such case, the norms of the points are equal, and no meaningful solution to the dual problem can be found. This predicament, however, can be easily fixed. A well-known technique, originating from kernel PCA [SSM98], is to center the images of the training points $\Phi(x_i)$ in feature space. In other words, the values of image points are re-computed in the local coordinate system anchored at the center of mass of the image points. This can be done by subtracting the mean from all image values:

$$\tilde{\Phi}(x_i) = \Phi(x_i) - \tfrac{1}{l} \sum_{i=1}^{l} \Phi(x_i).$$

Although this operation may not be directly computable in feature space, the impact of centering on the kernel values can be easily computed (e.g. [SSM98, SMB$^+$99]):

$$\tilde{K} = K - 1_l K - K 1_l + 1_l K 1_l, \tag{8}$$

where K is the $l \times l$ kernel matrix with the values $K_{ij} = k(x_i, x_j)$, and 1_l is an $l \times l$ matrix with all values equal to $\frac{1}{l}$. After centering in feature space, the norms of points in the local coordinate system are no longer all equal, and the dual problem of the quarter-sphere formulation can be easily solved.

3 Experiments

To compare the quarter-sphere formulation with the other one-class SVM approaches, and to investigate some properties of our algorithm, experiments are carried out on the KDDCup 1999 dataset. This dataset comprises connection record data collected in 1998 DARPA IDS evaluation. The features characterizing these connection records are pre-computed in the KDDCup dataset.

One of the problems with the connection record data from the KDDCup/DARPA data is that a large proportion (about 75%) of the connections represent the anomalies. In previous work [PES01, EAP+02] it was assumed that anomalies constitute only a small fraction of the data, and the results are reported on subsampled datasets, in which the ratio of anomalies is artificially reduced to 1-1.5%. To render our results comparable with previous work we also subsample the data. The results reported below are averaged over 10 runs of the algorithms in any particular setup.

3.1 Comparison of one-class SVM formulations

We first compare the quarter-sphere one-class SVM with the other two algorithms. Since the sphere and the plane formulations are equivalent for the RBF kernels, identical results are produced for these two formulations.

The experiments are carried out for two different values of the parameter σ of the RBF kernel: 1 and 12 (the latter value used in [EAP+02]). These values correspond to low and moderate regularization. As the evaluation criterion, we use the portion of the ROC curve between the false alarm rates of 0 and 0.1, since higher false alarm rates are unacceptable for intrusion detection. The comparison of ROCs of the three formulations for the two values of σ are shown in Fig. 5. It can be easily seen that the quarter-sphere formulation

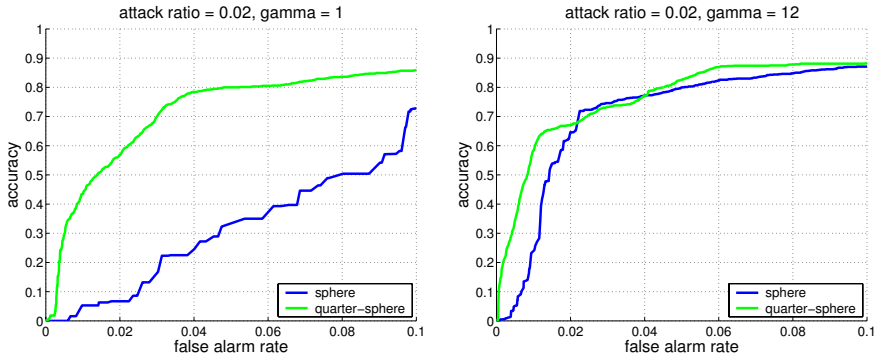

Figure 5: Comparison of the three one-class SVM formulations.

consistently outperforms the other two formulations; especially at the low value of regularization parameter. The best overall results are achieved with the medium regularization

with $\sigma = 12$, which has most likely been selected in [EAP+02] after careful experimentation. The advantage of the quarter-sphere in this case is not so dramatic as with low regularization, but is nevertheless very significant for low false alarm rates.

3.2 Dependency on the ratio of anomalies

The assumption that intrusions constitute a small fraction of the data may not be satisfied in a realistic situation. Some attacks, most notably the denial-of-service attacks, manifest themselves precisely in a large number of connections. Therefore, the problem of a large ratio of anomalies needs to be addressed.

In the experiments in this section we investigate the performance of the sphere and the quarter-sphere one-class SVM as a function of the attack ratio. It is known from the literature [TD99, SPST+01] that the parameter ν of the one-class SVM can be interpreted as an upper bound on the ratio of the anomalies in the data. The effect of this parameter on the quarter-sphere formulation is different: it specifies that *exactly ν fraction of points is expected to be the anomalies*. This is agreeably a more stringent assumption, and methods for the automatic determination of the anomaly ratio must be further investigated. Herein we perform a simple comparison of the algorithms under the following three scenarios:

- the parameter ν matches exactly the anomaly ratio,
- the parameter ν is fixed whereas the anomaly ratio varies,
- the ratio of anomalies is fixed and the parameter ν varies.

Under the scenario that ν matches the anomaly ratio it is assumed that perfect information about the anomaly ratio is available. One would expect that the parameter ν can tune both kinds of one-class SVM to the specific anomaly ratio. This, however, does not happen, as can be seen from Fig. 6. One can observe that the performance of both formulations noticeably degrades with the increasing anomaly ratio. We believe that the reason for this lies in the data-dependent normalization of the features: since the features are normalized with respect to the mean, having a larger anomaly ratio shifts the mean towards the anomalies, which leads to worse separability of the normal data and the anomalies.

Under the scenario with fixed ν it is assumed that no information about the anomaly ratio is available, and that this parameter is simply set by the user to some arbitrary value. As one can see from Fig. 7, the performance of both formulations of one-class SVM degrades with increasing anomaly ratio similarly to the scenario with ν matching the true anomaly ratio. Notice that the spread in the accuracy, as the anomaly ratio increases, is similar for both scenarios. This implies that, at least for the data-dependent normalization as used in the current experiments, setting the parameter ν to a fixed value is a reasonable strategy.

Under the scenario with fixed anomaly ratio and the varying ν we investigate what impact the adjustment of the parameter has on the same dataset. As it can be seen from Fig. 8, varying the parameter only has an impact on the sphere one-class SVM, the best accuracy

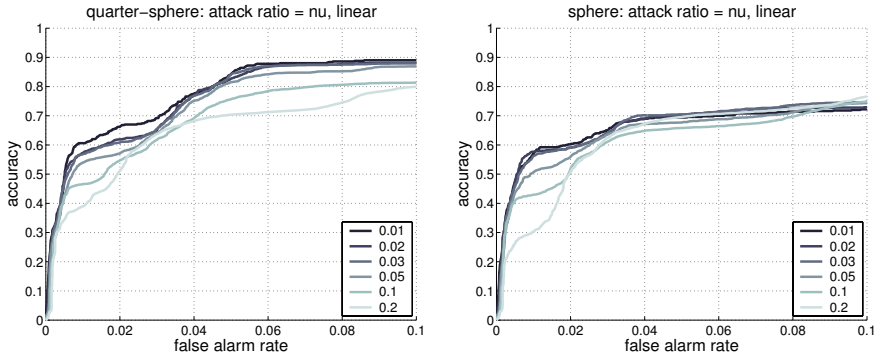

Figure 6: Impact of the anomaly ratio on the accuracy of the sphere and quarter-sphere SVM: anomaly ratio is equal to ν.

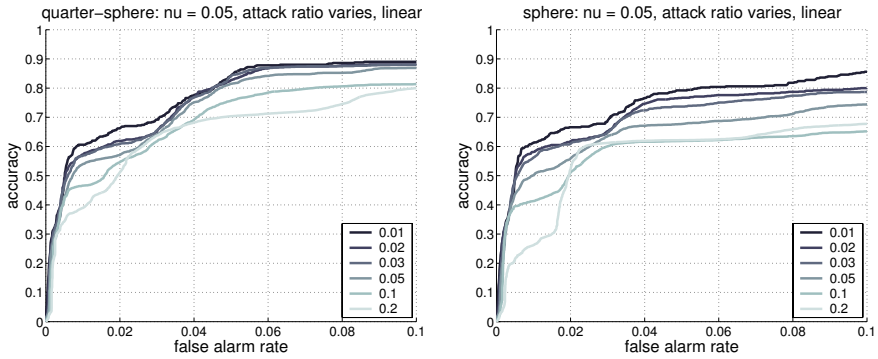

Figure 7: Impact of the anomaly ratio on the accuracy of the sphere and quarter-sphere SVM: ν is fixed at 0.05, anomaly ratio varies.

achieved on the higher values. *The parameter ν does not have any impact on the accuracy of the quarter-sphere one-class SVM*.

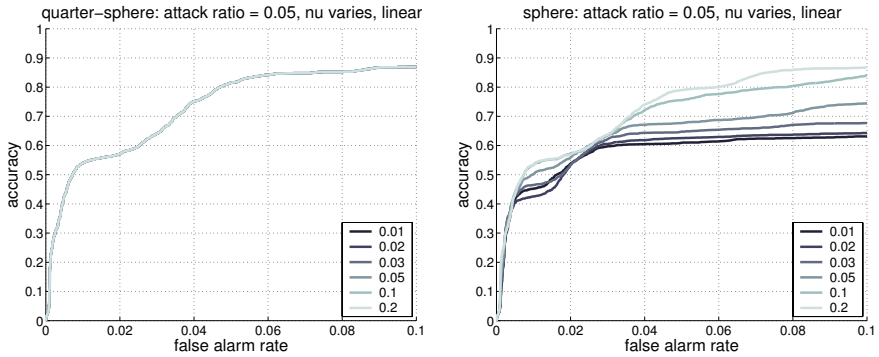

Figure 8: Impact of the anomaly ratio on the accuracy of the sphere and quarter-sphere SVM: anomaly ratio is fixed at 5%, ν varies.

4 Conclusions and future work

We have presented a novel one-class SVM formulation, the quarter-sphere SVM, that is optimized for non-negative attributes with one-sided distribution. Such data is frequently used in intrusion detection systems. The one-class SVM formulations previously applied in the context of unsupervised anomaly detection do not account for non-negativity and one-sidedness; as a result, they can potentially detect very common patterns, their attributes close to the origin, as anomalies. The quarter-sphere SVM avoids this problem by aligning the center of the sphere fitted to the data with the "center of mass" of the data in feature space.

Our experiments conducted on the KDDCup 1999 dataset demonstrate significantly better accuracy of the quarter-sphere SVM in comparison with the previous, sphere or plane, formulations. Especially noteworthy is the advantage of the new algorithm at low false alarm rates.

We have also investigated the behavior of one-class SVM as a function of attack rate. It is shown that the accuracy of all three formulations of one-class SVM considered here degrades with the growing percentage of attacks, contrary to the expectation that the parameter ν of one-class SVM, if properly set, should tune it to the required anomaly rate. We have found that the performance degradation with the perfectly set tuning parameters is essentially the same as when the parameter is set to some arbitrary value. We believe that performance of anomaly detection algorithms on higher anomaly rates should be given special attention in the future work, especially with respect to the data normalization techniques.

Acknowledgements

The authors gratefully acknowledge the funding from the *Bundesministerium für Bildung und Forschung* under the project MIND (FKZ 01-SC40A). We also thank Klaus-Robert Müller and Stefan Harmeling for valuable suggestions and discussions.

References

[BCJ+01] Barbará, D., Couto, J., Jajodia, S., Popyack, L., und Wu, N.: ADAM: Detecting intrusions by data mining. In: *Proc. IEEE Workshop on Information Assurance and Security*. S. 11–16. 2001.

[De87] Denning, D.: An intrusion-detection model. *IEEE Transactions on Software Engineering*. 13:222–232. 1987.

[DR90] Dowell, C. und Ramstedt, P.: The ComputerWatch data reduction tool. In: *Proc. 13th National Computer Security Conference,*. S. 99–108. 1990.

[EAP+02] Eskin, E., Arnold, A., Prerau, M., Portnoy, L., und Stolfo, S.: *Applications of Data Mining in Computer Security*. chapter A geometric framework for unsupervised anomaly detection: detecting intrusions in unlabeled data. Kluwer. 2002.

[JLA+93] Jagannathan, R., Lunt, T. F., Anderson, D., Dodd, C., Gilham, F., Jalali, C., Javitz, H. S., Neumann, P. G., Tamaru, A., und Valdes, A.: Next-generation intrusion detection expert system (NIDES). Technical report. Computer Science Laboratory, SRI International. 1993.

[LEK+03] Lazarevic, A., Ertoz, L., Kumar, V., Ozgur, A., und Srivastava, J.: A comparative study of anomaly detection schemes in network intrusion detection,. In: *Proc. SIAM Conf. Data Mining*. 2003.

[LV92] Liepins, G. und Vaccaro, H.: Intrusion detection: its role and validation. *Computers and Security,*. 11(4):347–355. 1992.

[NWY02] Noel, S., Wijesekera, D., und Youman, C.: *Applications of Data Mining in Computer Security*. chapter Modern intrusion detection, data mining, and degrees of attack guilt. Kluwer. 2002.

[PES01] Portnoy, L., Eskin, E., und Stolfo, S.: Intrusion detection with unlabeled data using clustering. In: *Proc. ACM CSS Workshop on Data Mining Applied to Security*. 2001.

[PN97] Porras, P. A. und Neumann, P. G.: Emerald: event monitoring enabling responses to anomalous live disturbances. In: *Proc. National Information Systems Security Conference*. S. 353–365. 1997.

[SMB+99] Schölkopf, B., Mika, S., Burges, C., Knirsch, P., Müller, K.-R., Rätsch, G., und Smola, A.: Input space vs. feature space in kernel-based methods. *IEEE Transactions on Neural Networks*. 10(5):1000–1017. September 1999.

[SPST+01] Schölkopf, B., Platt, J., Shawe-Taylor, J., Smola, A., und Williamson, R.: Estimating the support of a high-dimensional distribution. *Neural Computation*. 13(7):1443–1471. 2001.

[SS02] Schölkopf, B. und Smola, A.: *Learning with Kernels*. MIT Press. Cambridge, MA. 2002.

[SSM98] Schölkopf, B., Smola, A., und Müller, K.-R.: Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*. 10:1299–1319. 1998.

[TD99] Tax, D. und Duin, R.: Data domain description by support vectors. In: Verleysen, M. (Hrsg.), *Proc. ESANN*. S. 251–256. Brussels. 1999. D. Facto Press.

[Va95] Vapnik, V.: *The nature of statistical learning theory*. Springer Verlag. New York. 1995.

[Va98] Vapnik, V.: *Statistical Learning Theory*. Wiley. New York. 1998.

[VS00] Valdes, A. und Skinner, K.: Adaptive, model-based monitoring for cyber attack detection. In: *Proc. RAID 2000*. S. 80–92. 2000.

[WFP99] Warrender, C., Forrest, S., und Perlmutter, B.: Detecting intrusions using system calls: alternative data methods. In: *Proc. IEEE Symposium on Security and Privacy*. S. 133–145. 1999.