A Survey of Recent Trends in One Class Classification

Shehroz S. Khan and Michael G. Madden

College of Engineering and Informatics,
National University of Ireland Galway, Ireland
{shehroz.khan,michael.madden}@nuigalway.ie

Abstract. The One Class Classification (OCC) problem is different from the conventional binary/multi-class classification problem in the sense that in OCC, the negative class is either not present or not properly sampled. The problem of classifying positive (or target) cases in the absence of appropriately-characterized negative cases (or outliers) has gained increasing attention in recent years. Researchers have addressed the task of OCC by using different methodologies in a variety of application domains. In this paper we formulate a taxonomy with three main categories based on the way OCC has been envisaged, implemented and applied by various researchers in different application domains. We also present a survey of current state-of-the-art OCC algorithms, their importance, applications and limitations.

Keywords: One Class Classification, Outlier Detection, Support Vector Machines, Positive and Unlabeled Data.

1 Introduction

Conventional multi-class classification algorithms aim to classify an unknown object into one of several pre-defined categories. A problem arises when the unknown object does not belong to any of those categories. In one-class classification [1][2], one of the classes (referred to as the positive class or target class) is well characterized by instances in the training data. For the other class (non-target), it has either no instances at all, very few of them, or they do not form a statistically-representative sample of the negative concept.

To motivate the importance of one-class classification, let us consider some scenarios. One-class classification can be relevant in detecting machine faults, for instance. A classifier should detect when the machine is showing abnormal/faulty behaviour. Measurements on the normal operation of the machine (positive class training data) are easy to obtain. On the other hand, most faults will not have occurred so one will have little or no training data for the negative class. As another example, a traditional binary classifier for text or web pages requires arduous pre-processing to collect negative training examples. For example, in order to construct a homepage classifier [3], collecting sample of homepages

L. Coyle and J. Freyne (Eds.): AICS 2009, LNAI 6206, pp. 188–197, 2010.

(positive training examples) is relatively easy, however collecting samples of non-homepages (negative training examples) is very challenging because it may not represent the negative concept uniformly and may involve human bias.

The outline of the paper is as follows. In Section 2 we compare OCC and multi-class classification problems. In Section 3 we propose a taxonomy for the study of OCC and present current state-of-the-art survey of some of the major research contributions under the proposed taxonomy. Section 4 concludes our presentation with summary of the research work in OCC and guidelines for future research.

2 OCC vs. Multi-class Classification

In a conventional multi-class classification problem, data from two (or more) classes are available and the decision boundary is supported by the presence of example samples from each class. Moya et al. [4] originate the term One-Class Classification in their research work. Different researchers have used other terms to present similar concepts such as Outlier Detection [5], Novelty Detection [6] or Concept Learning [7]. These terms originate as a result of different applications to which OCC has been applied.

The drawbacks that are encountered in the conventional classification problems, such as the estimation of error rates, measuring the complexity of a solution, the curse of dimensionality, the generalization of the method, and so on, also appear in OCC, and sometimes become even more prominent.

As stated earlier, in OCC tasks, the negative data is either absent or limited in its distribution, so only one side of the classification boundary can be determined definitively by using the data. This makes problem of one-class classification harder than the problem of conventional multi-class / binary classification. The task in OCC is to define a classification boundary around the positive (or target) class, such that it accepts as many objects as possible from the positive class, while it minimizes the chance of accepting non-positive (or outlier) objects. Since only one side of the boundary can be determined, in OCC, it is hard to decide, on the basis of just one class how tightly the boundary should fit in each of the directions around the data. It is also harder to decide which attributes should be used to find the best separation of the positive and non-positive class objects. In particular, when the boundary of the data is long and non-convex, the required number of training objects might be very high. Hence it is to be expected that one-class classification algorithms will require a larger number training instances relative to conventional multi-class classification algorithms [2].

3 Taxonomy and Review of OCC Work

Based on reviewing past research that has been carried out in the field of OCC by using different algorithms, methodologies and application domains, we propose a taxonomy with three broad categories for the study of OCC problems. The taxonomy can be summarized as (see Fig. 1):

(a) *Availability of Training Data:* Learning with positive data only (or with a limited amount of negative samples) or learning with positive and unlabeled data
(b) *Methodology Used:* Algorithms based on One Class Support Vector Machines (OSVMs) or methodologies based on algorithms other than OSVMs
(c) *Application Domain Applied:* OCC applied in the field of text/document classification or in other application domains

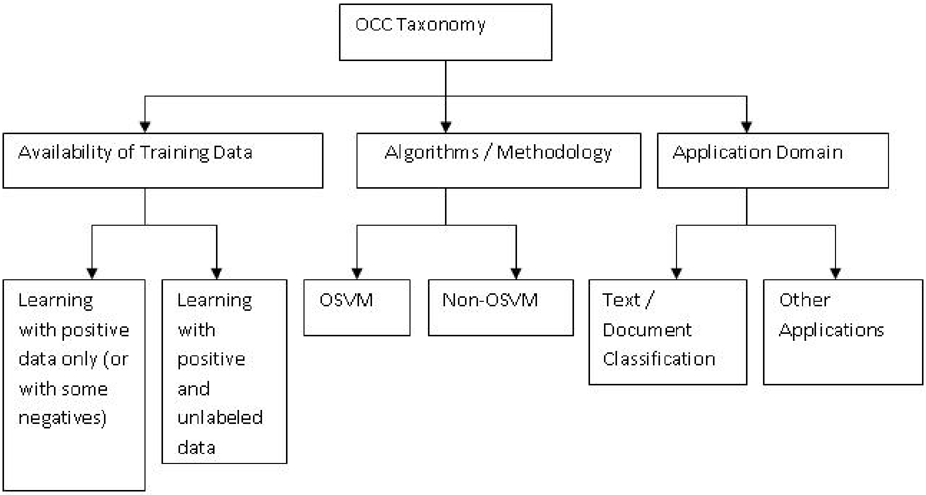
OCC Taxonomy
Availability of Training Data
Algorithms / Methodology
Application Domain
Learning with positive data only (or with some negatives)
Learning with positive and unlabeled data
OSVM
Non-OSVM
Text / Document Classification
Other Applications

Fig. 1. Proposed Taxonomy for the Study of OCC Techniques

The proposed categories are not mutually exclusive, so there may be some overlapping among the research carried out in each of these categories. However, they cover almost all of the major research conducted using the concept of OCC in various contexts and application domains. The key contributions in most OCC research fall into one of the above-mentioned categories.

3.1 Availability of Training Data

OCC problems have been studied extensively under three broad frameworks:

1. Learning with positive examples only
2. Learning with positive examples and some amount of poorly distributed negative examples
3. Learning with positive and unlabeled data

The last category has recieved much research interest among the text/document classification community [8][9][10] that will be discussed in detail below in Section 3.3.

Tax and Duin [11][12] and Scholkopf et al.[13] have developed algorithms based on support vector machines to tackle the problem of OCC using positive examples only (refer to Section 3.2). The main idea behind these strategies is to construct a decision boundary around the positive data so as to differentiate the outliers (non-positives) from the positive data.

For many learning tasks, labeled examples are rare, whereas numerous unlabeled examples are easily available. The problem of learning with the help of unlabeled data given a small set of labeled examples is studied by Blum and Mitchell [14] by using the concept of co-training for text classification. The co-training setting can be applied when a data set has natural separation of its features. Co-training algorithms incrementally build basic classifiers over each of these feature sets. They show that under the assumptions that each set of features is sufficient for classification, and the feature sets of each instance are conditionally independent, given the class, PAC (Probably Approximately Correct) learning [15] guarantees on learning from labeled and unlabeled data. Muggleton [16] presents a theoretical study in the Bayesian framework where the distribution of functions and examples are assumed to be known. Skabar [17] describes the use of feed-forward neural network to learn a classifier from a data set consisting of labeled positive examples along with a corpus of unlabeled examples containing positive and negative samples.

3.2 Algorithm Used

OSVM. The one-class classification problem is often solved by estimating the target density [4], or by fitting a model to the data support vector classifier [18]. Tax and Duin [11][12] seek to solve the problem of OCC by distinguishing the positive class from all other possible patterns in the pattern space. Instead of using a hyper-plane to distinguish between two classes, a hyper-sphere is found around the positive class data that encompasses almost all points in the data set with the minimum radius. This method is called the Support Vector Data Description (SVDD). Thus training this model has the possibility of rejecting some fraction of the positively-labeled training objects, when this sufficiently decreases the volume of the hyper-sphere. Furthermore, the hyper-sphere model of the SVDD can be made more flexible by introducing kernel functions. Tax [2] considers a Polynomial and a Gaussian kernel and found that the Gaussian kernel works better for most data sets. A drawback of this technique is that they often require a large data set; in particular, in high dimensional feature spaces, it becomes very inefficient. Also, problems may arise when large differences in density exist. Objects in low-density areas will be rejected although they are legitimate objects.

Scholkopf et al. [13][19] present an alternative approach to the above mentioned work of Tax and Duin on OCC using a separating hyper-plane. The difference between theirs and Tax and Duins approach is that instead of trying to find a hyper-sphere with minimal radius to fit the data, they try to separate the surface region containing data from the region containing no data. This is achieved by constructing a hyper-plane which is maximally distant from origin, with all data points lying on the opposite side from the origin and such that the margin is positive. Their paper proposes an algorithm that computes a binary function that returns +1 in small regions (subspaces) that contain data and -1 elsewhere. The data is mapped into the feature space corresponding to the kernel and is separated from the origin with maximum margin. They evaluate the efficacy of their method on the US Postal Services data set of handwritten digits and show that the algorithm is able to extract patterns which are very hard to assign to their respective classes and a number of outliers were identified.

Manevitz and Yousef [20] propose a different version of the one class SVM which is based on identifying outlier data as representative of the second class. The idea of this methodology is to work first in the feature space, and assume that not only is the origin the second class, but also that all data points close enough to the origin are to be considered as noise or outliers. The vectors lying on standard sub-spaces of small dimension (i.e. axes, faces, etc.) are treated as outliers. They evaluate their results on Reuters Data set[1] and the results are worse than the OSVM algorithm presented by Scholkopf et al. [19].

Classifiers are commonly ensembled to provide a combined decision by averaging the estimated posterior probabilities. When Bayes theorem is used for the combination of different classifiers, under the assumption of independence, a product combination rule can be used to create classifier ensemble. The outputs of the individual classifiers are multiplied and then normalized (also called the logarithmic opinion pool [21]). In OCC, as the information on the non-positive data is not available, in most cases, the outliers are assumed to be uniformly distributed and the posterior probability can be estimated. Tax [2] mentions that in some OCC methods, distance is estimated instead of probability for one class classifier ensembling. Tax observes that the use of ensembles in OCC improves performance, especially when the product rule is used to combine the probability estimates.

Yu [22] proposes an OCC algorithm with SVMs using positive and unlabeled data, and without labeled negative data, and discusses some of the limitations of other OCC algorithms [1][3][13][20]. Yu comments that in the absence of negative examples, OSVM requires a much larger amount of positive training data to induce an accurate class boundary.

Non-OSVMs. Ridder et al. [23] conduct an experimental comparison of various OCC algorithms, including: (a) Global Gaussian approximation; (b) Parzen density estimation; (c) 1-Nearest Neighbor method; and (d) Gaussian approximation (combines aspects of (a) and (b)). Manevitz and Yousef [24] trained a simple neural network to filter documents when only positive information is available. To incorporate the restriction of availability of positive examples only, they used a three-level feed forward network with a "bottleneck".

DeComite et al. [25] modify the C4.5 decision tree algorithm [26] to get an algorithm that takes as input a set of labeled examples, a set of positive examples, and a set of unlabeled data, and then use these three sets to construct the decision tree. Letouzey et al. [27] design an algorithm which is based on positive statistical queries (estimates for probabilities over the set of positive instances)

[1] http://www.daviddlewis.com/resources/testcollections/reuters21578

and instance statistical queries (estimates for probabilities over the instance space). They design a decision tree induction algorithm, called POSC4.5, using only positive and unlabeled data. They present experimental results on UCI data sets[2] that are comparable to the C4.5 algorithm.

Wang et al. [28] investigate several one-class classification methods in the context of Human-Robot interaction for face and non-face classification. Some of the noteworthy methods used in their study are: (a) SVDD; (b) Gaussian data description; (c) KMEANS-DD; (d) Principal Component Analysis-DD. In their experimentation, they observe that SVDD attains better performance than the other OCC methods they studied.

3.3 Application Domain Used

Text Classification. Traditional text classification techniques require an appropriate distribution of positive and negative examples to build a classifier; thus they are not suitable for this problem of OCC. It is of course possible to manually label some negative examples, though that it is labour-intensive and a time consuming task. However the core problem remains, that it is difficult or impossible to compile a set of negative samples that provides a comprehensive characterization of everything that is ‘not’ the target concept, as is assumed by a conventional binary classifier.

The ability to build classifiers without negative training data is useful in a scenario when one needs to extract positive documents from many text collections or sources. Liu et al. [29] propose a method (called Spy EM) to solve this problem in the text domain. It is based on Naïve Bayesian classification (NB) and the Expectation Maximization (EM) algorithm [30]. The main idea of the method is to first use a technique to identify some reliable / strong negative documents from the unlabeled set. It then runs EM to build the final classifier. Yu et al. [3][31] propose an SVM-based technique called PEBL (Positive Example Based Learning) to classify Web pages with positive and unlabeled pages. Once a set of strong negative documents is identified, SVM is applied iteratively to build a classifier. PEBL is sensitive to the number of positive examples and gives poor results when they are small in number. Li and Liu [9] propose an alternative method that extracts negative data from the unlabeled set using the Rocchio method [32] . Although the second step also runs SVM iteratively to build a classifier, there is a key difference in selection of the final classifier. Their technique selects a “good” classifier from a set of classifiers built by SVM, while PEBL does not. It is shown theoretically by Liu et al. [29] that if the sample size is large enough, maximizing the number of unlabeled examples classified as negative while constraining the positive examples to be correctly classified will produce a good classifier. Liu et al. [8] develop a benchmarking system called LPU (Learning from Positive and Unlabeled data)[3] and also propose an approach based on a biased formulation of SVM that allows noise (or error) in

[2] http://www.ics.uci.edu/ mlearn/MLRepository.html

[3] http://www.cs.uic.edu/ liub/LPU/LPU-download.html

positive examples. They experiment on Reuters data set and Usenet articles by Lang [33] and conclude that biased-SVM approach outperforms other existing two-step techniques.

Peng et al. [34] present a text classifier from positive and unlabeled documents based on Genetic Algorithms (GA). They perform experiments on the Reuters data set and compare their results against PEBL [31] and OSVM, and show that their GA based classification performs better. Koppel et al. [35] study the "Authorship Verification" problem where only examples of writings of a single author are given and the task is to determine if a given piece of text is or is not written by this author. They test their algorithm on a collection of twenty-one 19th century English books written by 10 different authors and spanning a variety of genres. They obtain overall accuracy of 95.7% with errors almost equally distributed between false positives and false negatives.

Denis et al. [36] introduce a Naïve Bayes algorithm and shows its feasibility for learning from positive and unlabeled documents. The key step in their method is in estimating word probabilities for the negative class because negative examples were not available. This limitation can be overcome by assuming an estimate of the positive class probability (the ratio of positive documents in the set of all documents). In practical situations, the positive class probability can be empirically estimated or provided by domain knowledge. Their results on the WebKB data set[4] show that error rates of Naïve Bayes classifiers obtained from positive examples trained with enough unlabeled examples are lower than error rates of Naïve Bayes classifiers obtained from labeled documents.

Other Application Domains. OSVMs have been successfully applied in a wide variety of application domains such as Handwritten Digit Recognition [1][2][19], Information Retrieval [20], Face Recognition Applications [28][37], Medical Analysis [38], Bioinformatics [39][40], Spam Detection [41], Anomaly Detection [42][43], Machine Fault Detection [44]. Compression neural networks for one-sided classification have been used in many application areas, including detecting Mineral Deposits [17]. Wang and Stolfo use one-class Naïve Bayes to detect Masquerade Detection [45] in a network and show that less effort in data collection is required with comparable performance as that of a multi-class classifier. Munroe and Madden [46] present a one class k-nearest neighbor approach for vehicle recognition from images and showed that the results are comparable to that of standard multi-class classifiers.

4 Conclusions and Future Work

The goal of One-Class Classification is to induce classifiers when only one class (the positive class) is well characterized by the training data. In this paper, we have presented a survey of current state-of-the-art research work using OCC. We observe that the research carried out in OCC can be broadly presented by three different categories or areas of study, which depends upon the availability

[4] http://www-2.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/

of training data, classification algorithms used and the application domain investigated. Under each of these categories, we further provide details of commonly used OCC algorithms. Although the OCC field is becoming mature, still there are several fundamental problems that are open for research, not only in describing and training classifiers, but also in scaling, controlling errors, handling outliers, using non-representative sets of negative examples, combining classifiers and reducing dimensionality.

Classifier ensembles have not been exploited very much for OCC problems, and techniques such as boosting and bagging deserve further attention. Another point to note here is that in OSVMs, the kernels that have been used mostly are Linear, Polynomial, Gaussian or Sigmoidal. We suggest it would be fruitful to investigate some more innovative forms of kernel, for example Genetic Kernels [47], that have shown greater potential in standard SVM classification. In the case where abundant unlabeled examples and some positive examples are available, researchers have used many two-step algorithms (as have been discussed in Section 3.3). We believe that a Bayesian Network approach to such OCC problems would be an interesting exercise.

This survey provides a broad insight into the study of the discipline of OCC. Depending upon the data availability, algorithm use and application, appropriate OCC techniques can be applied and improved upon. We hope that this survey will provide researchers with a direction to formulate future novel work in this field.

References

1. Tax, D., Duin, R.: Uniform object generation for optimizing one-class classifiers. J. Machine Learning Research 2, 155–173 (2001)
2. Tax, D.: One Class Classification. PhD thesis, Delft University of Technology (2001)
3. Yu, H., Han, J., Chang, K.C.C.: Positive-example based learning for web page classification using svm. In: Proc. Eighth International Conference on Knowledge Discovery and Data Mining (KDD 2002), pp. 239–248 (2002)
4. Moya, M., Koch, M., Hostetler, L.: One-class classifier networks for target recognition applications. In: Proceedings World Congress on Neural Networks, pp. 797–801 (1993)
5. Ritter, G., Gallegos, M.: Outliers in statistical pattern recognition and an application to automatic chromosome classification. Pattern Recognition Letters 18, 525–539 (1997)
6. Bishop, C.: Novelty detection and neural network validation. IEEE Proceedings on Vision, Image and Signal Processing, Special Issue on Applications of Neural Networks 141(4), 217–222 (1994)
7. Japkowicz, N.: Concept-Learning in the absence of counterexamples: An autoassociation-based approach to classification. PhD thesis, New Brunswick Rutgers, The State University of New Jersey (1999)
8. Liu, B., Dai, Y., Li, X., Lee, W., Yu, P.: Building text classifiers using positive and unlabeled examples. In: Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM 2003) (2003)

9. Li, X., Liu, B.: Learning to classify texts using positive and unlabeled data. In: 18th International Joint Conf. on Artificial Intelligence (IJCAI 2003), pp. 587–594 (2003)
10. Lee, W., Liu, B.: Learning with positive and unlabeled examples using weighted logistic regression. In: Proceedings of the 20th International Conference on Machine Learning (ICML 2003) (2003)
11. Tax, D., Duin, R.: Data domain description using support vectors. In: Proc. ESAN 1999, Brussels, pp. 251–256 (1999)
12. Tax, D., Duin, R.: Support vector domain description. Pattern Recognition Letters 20, 1191–1199 (1999)
13. Scholkopf, B., Williamson, R., Smola, A., Taylor, J., Platt, J.: Support vector method for novelty detection. In: Solla, S.A., Leen, T., Muller, K. (eds.) Neural Information Processing Systems, pp. 582–588 (2000)
14. Blum Mitchell, T.: Combining labeled and unlabeled data with co-training. In: Proceedings of 11th Annual conference on Computation Learning Theory, pp. 92–100. ACM Press, New York (1998)
15. Valiant, L.: Theory of the learnable. ACM 27(11), 1134–1142 (1984)
16. Muggleton, S.: Learning from the positive data. Machine Learning (2001)
17. Skabar, A.: Single-class classifier learning using neural networks: An application to the prediction of mineral deposits. In: Proceedings of the Second International Conference on Machine Learning and Cybernetics, vol. 4, pp. 2127–2132 (2003)
18. Burges, C.J.C.: A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery 2(2), 1–47 (1998)
19. Scholkopf, B., Williamson, R., Smola, A., Taylor, J.: Sv estimation of a distributions support. In: Advances in Neural Information Processing Systems (1999)
20. Manevitz, L.M., Yousef, M.: One-class svms for document classification. Journal of Machine Learning Research 2, 139–154 (2001)
21. Benediktsson, J., Swain, P.: Consensus theoretic classification methods. IEEE Transactions on Systems, Man and Cybernetics 22(4), 688–704 (1992)
22. Yu, H.: Single-class classification with mapping convergence. Machine Learning 61(1), 49–69 (2005)
23. de Ridder, D., Tax, D., Duin, R.: An experimental comparison of one-class classification methods. In: Proceedings of the 4th Annual Conference of the Advacned School for Computing and Imaging, Delft (1998)
24. Manevitz, L., Yousef, M.: Document classification on neural networks using only positive examples. In: Proc. of 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 304–306 (2000)
25. De Comite, F., Denis, F., Gillerson, R., Letouzey, F.: Positive and unlabeled examples help learning. In: Watanabe, O., Yokomori, T. (eds.) ALT 1999. LNCS (LNAI), vol. 1720, pp. 219–230. Springer, Heidelberg (1999)
26. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo (1993)
27. Letouzey, F., Denis, F., Gilleron, R.: Learning from positive and unlabeled examples. In: Algorithmic Learning Theory, 11th International Conference, Sydney, Australia (2000)
28. Wang, Q., Lopes, L.S., Tax, D.J.: Visual object recognition through one-class learning. In: International Conference on Image Analysis and Recognition, pp. 463–470 (2004)
29. Liu, B., Lee, W., Yu, P., Li, X.: Partially supervised classification of text documents. In: Proc. of ICML (2002)

30. Dempster, A.P., Laird, N.M., Rubin, D.: Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society (1977)
31. Yu, H., Han, J., Chang, K.: PEBL: Web page classification without negative examples. IEEE Transactions on Knowledge and Data Engineering 16(1) (2004)
32. Rocchio, J.: Relevant feedback in information retrieval. In: Salton, G. (ed.) The SMART retrieval system- experiments in automatic document processing, NJ, Englewood Cliffs (1971)
33. Lang, K.: Newsweeder: Learning to filter netnews. In: ICML 1995 (1995)
34. Peng, T., Zuo, W., He, F.: Text classification from positive and unlabeled documents based on ga. In: Proc. of VECPAR 2006, Brazil (2006)
35. Koppel, M., Schler, J.: Authorship verification as a one-class classification problem. In: Proceedings of the twenty-first International Conference on Machine learning, Alberta, Canada, vol. 69 (2004)
36. Denis, F., Gilleron, R., Tommasi, M.: Text classification from positive and unlabeled examples. In: 9th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (2002)
37. Zeng, Z., Fu, Y., Roisman, G.I., Wen, Z., Hu, Y., Huang, T.S.: One-class classification for spontaneous facial expression analysis. In: Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition, pp. 281–286 (2006)
38. Gardner, B., Krieger, A.M., Vachtsevanos, G., Litt, B.: One-class novelty detection for seizure analysis from intracranial eeg. Journal of Machine Learning Research 7, 1025–1044 (2006)
39. Spinosa, E.J., de Carvalho, A.C.P.L.F.: SVMs for novel class detection in bioinformatics. In: Brazilian Workshop on Bioinformatics, pp. 81–88 (2004)
40. Alashwal, H.T., Deris, S., Othman, R.M.: One-class support vector machines for protein-protein interactions prediction. International Journal Biomedical Sciences 1(2), 120–127 (2006)
41. Sun, D., Tran, Q., Duan, H., Zhang, G.: A novel method for chinese spam detection based on one-class support vector machine. Journal of Information and Computational Science 2(1), 109–114 (2005)
42. Li, K., Huang, H., Tian, S., Xu, W.: Improving one-class svm for anomaly detection. In: Proceedings of the second international conference on machine learning and cybernetics, pp. 2–5 (November 2003)
43. Perdisci, R., Gu, G., Lee, W.: Using an ensemble of one-class svm classifiers to harden payload-based anomaly detection systems. In: Sixth International Conference on Data Mining, pp. 488–498 (2006)
44. Shin, H.J., Eom, D.W., Kim, S.S.: One-class support vector machines: an application in machine fault detection and classification. Computers and Industrial Engineering 48(2), 395–408 (2005)
45. Wang, K., Stolfo, S.J.: One class training for masquerade detection. In: ICDM Workshop on Data Mining for Computer Security (2003)
46. Munroe, D.T., Madden, M.G.: Multi-class and single-class classification approaches to vehicle model recognition from images. In: Proc. AICS 2005: Irish Conference on Artificial Intelligence and Cognitive Science, Portstewart (2005)
47. Howley, T., Madden, M.G.: An evolutionary approach to automatic kernel construction. In: Kollias, S.D., Stafylopatis, A., Duch, W., Oja, E. (eds.) ICANN 2006. LNCS, vol. 4132, pp. 417–426. Springer, Heidelberg (2006)