ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

One-class classifier based on extreme value statistics

David Martinez-Rego[2], Evan Kriminger[1], Jose C. Principe[1], Oscar Fontenla-Romero[2] and Amparo Alonso-Betanzos[2] *

1- Computational NeuroEgineering Lab.
University of Florida, Gainesville, FL 32611,
evankriminger@gmail.com, principe@cnel.ufl.edu

2- LIDIA Group - Dept. of Computer Science
Campus de Elvina, s/n, 15071, A Coruna - Spain
{dmartinez, ofontenla, ciamparo}@udc.es

Abstract. In recent years, interest in one-class classification methods has soared due to its wide applicability in many practical problems in which classification in the absence of counterexamples is needed. In this paper, a new one class classification rule based on order statistics is presented. It only relies on embedding the classification problem into a metric space, so it is suitable for Euclidean or other structured mappings. The suitability of the proposed method is assessed through a comparison both for artificial and real life data sets. The good results obtained pave the road for its application on practical novelty detection problems.

1 Introduction

One-class classification problem can be stated as follows: using a data set which is known to have originated under well defined conditions, the 'normal' regime, build a classifier able to detect whether new data samples have been normally generated or, otherwise, generated under different conditions or corrupted by noise. The rationale is that, under what shall be called *normal conditions*, data samples are similar to each other covering an area of input space called support of normal data, whilst when the conditions change, the generated data shall start to fill in different areas of the input space.
The term one-class classification originates from the work in [5], but it has also been called outlier detection, novelty detection, anomaly detetion or concept learning, all stemming from the different applications to which one-class classifiers can be applied to. In terms of real life problems, one-class classifiers are encountered in Cyber-Intrusion Detection, Fraud Detection, Medical Anomaly Detection, Industrial Damage Detection and Textual Anomaly Detection. A great deal of early methods in the literature devised for one-class classification were based on the adaptation of algorithms in the realm of machine learning for such task and in recent years, techniques specifically devised for this problem

*Research supported by the Spanish Ministry of Education via an FPU doctoral grant to David Martinez-Rego and by the Secretaría de Estado de Investigación of the Spanish Goverment (TIN2009-10748) and the European Union by FEDER funds.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

have appeared [1] [2]. Different one-class classifiers can be grouped depending on the assumptions they make for detecting anomalies: Classification Based, Clustering Based, Nearest Neighbor Based, Statistical, Information Theoretic and Spectral. In [3], the fact that distances between the samples of different classes can be exploited to better model their support in a multiple class classification problem is explored and is a proof that cumulative distances are a source of information able to improve classification accuracy results. Based on this evidence, in this paper a new classification rule for one-class scenarios based on extreme value statistics [4] is presented. It aims at exploiting distance information under normal conditions to solve one-class classification problems. Specifically, the probability distribution function of the distance of a sample, under normal conditions, to its close neighbors is modeled parametrically or non-parametrically. Afterwards, using a result from extreme value statistics, when a new sample s is presented, the probability of obtaining a more dissimilar sample than s under normal conditions is obtained and eventually used as an indicative of an anomaly. Thanks to the specific modeling of the distribution of distances to close samples under normal conditions, we shall be able to capture the support of normal data while we neglect possible spurious data in the normal state data set, as these data are typically characterized as being disperse and far from the normal state support. In the experimental section, it can be noticed that exploiting distance information in this manner leads to an accurate one-class classifier.

2 Method description

In this section the principal results and rationale under the proposed Extreme Value One-Class Classifier (EVOC) method are presented. Firstly, a theorem rooted in Extreme Value Statistics field [4] on which the proposed method is largely based on is presented:

Theorem (Distribution Function of any order statistics). *The probability distribution function $F_{r:n}$ of any order statistics r of a sample of n values of a random variable with distribution function $F(x)$ is:*

$$F_{r:n} = B_{F(x)}(r, n-r+1) \tag{1}$$

where B is the regularized incomplete beta function.

This result is based on treating the sampling process as a multinomial distribution and using the probabilities extracted from the original distribution function $F(x)$. It can be derived as follows:

$$\begin{aligned} F_{r:n}(x) &= P(X_{r:n} \leq x) = 1 - F_{m_n(x)}(r-1) = \sum_{k=r}^{n} \binom{n}{r} F^k(x)[1-F(x)]^{n-k} \\ &= r\binom{n}{r} \int_0^{F(x)} u^{r-1}(1-u)^{n-r} du = B_{F(x)}(r, n-r+1) \end{aligned}$$

where $m_n(x)$ is the number of elements of the samples with a value $X_j \leq x$ and $F_{m_n(x)}(k)$ with $k \in (0, n)$ represents the probability that the number of samples

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

below x is less than or equal to k. Based on this result, when a new sample s is presented, it is possible to model the probability of obtaining a sample more discrepant or abnormal than s. In order to do this, consider a metric space M where the data we want to classify belongs to. First, in the training phase the probability distribution function $F_d(x)$ of the distance of each sample to its k nearest neighbors is modeled based on data drawn from only one class, which shall be called *Normal state data*. In order to do this, for each data sample in the normal state data set, we search its k nearest neighbors and use those distances to model $F_d(x)$. It is important to remark that this step relies only on the fact that the data is embedded into a metric space where we have a distance function d, so it is possible to use other data encodings apart form the Euclidean space $\mathbb{R}^n$. For the probability distribution function estimation, both parametric and non-parametric methods are available. In this work we will adopt a parametric approach.

Algorithm 1: Proposed EVOC classification method

Training Stage
Input: Normal State data X, number of neighbors k, estimated fraction of outliers p
Output: Classifier (X, F_d, α)

foreach *sample* $s \in X$ **do**
  Calculate the set of distances d_s of s to its k-nearest neighbors in X.
  Add the distances in d_s to the set D.
Estimate F_d based on the values in D.
Set α leaving a p fraction of data out of the support.

Classification Stage
Input: Classifier (D, F_d, α), and a new sample s
Output: Classification result $C(s)$, {1 - Normal State, 0 - Novelty}

Calculate the set of distances d_s of s to its k-nearest neighbors in X.
Classify s following the rule:

$$C(s) = I(P(D_k > d_s) - \alpha) = I\left(\prod_{i=1}^{k}(1 - F^d_{i:k}(d_s(i))) - \alpha\right) \quad (2)$$

Subsequently, when a new data point s is to be classified, the following rule is used:

$$C(s) = I(P(D_k > d_s) - \alpha) = I\left(\prod_{i=1}^{k}(1 - F^d_{i:k}(d_s(i))) - \alpha\right) \quad (3)$$

where I is the heaviside function, d_s is the set of distances to the k nearest neighbors of s in the normal state data set, D is the multiset of all the dis-

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

tances to nearest neighbors in the normal state data set, $d_s(i)$ is the distance to the i-th closest pattern to s in the normal state data set, D_k is a random variable that represents the distance to the k nearest neighbors ($D_k > d_k$ if $\forall i \in [1,k], D_k(i) > d_k(i)$), $F^d_{r:k}$ is the r-th order statistics formula of theorem 1 in which the estimated distribution function $F_d(x)$ of the distance to a neighbor is plugged in, and α is a threshold that controls under which level the data sample s is considered abnormal (in this rule, logarithms can be taken to prevent underflow). Note that what the classification rule is monitoring if the probability of obtaining, under normal conditions, a set of k nearest neighbors more dissimilar than the ones we have found for s. If this probability falls, it means that the normal state hypothesis for s has been violated and so it is classified as abnormal or as a counterexample (see algorithm 1).
When applied to high dimensional data sets, negative effects due to curse of dimensionality can appear. Depending on the used metric, the notion of proximity can become meaningless [6] degrading the contrast between sparse and close neighbors in which the proposed model is based. In this situation, careful selection of distances [6] and dimensionality reduction techniques should be considered.

3 Experimental results

In this section we explore the features of the proposed method for both artificial and real one-class classification data sets. For the experiments presented hereafter, we adopt a parametric approach to model the distance between samples using as hypotesis the lognormal distribution.

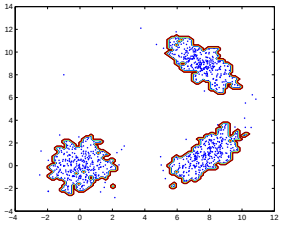
3.1 Artificial data sets

In this section, the ability of the proposed method to capture the support of normal data is depicted for two 2D artificial data sets. Specifically, we tested it with a multimodal gaussian data set and the well known banana data set. The results for number of neighbors $k = 4$ and $p = 0.06$ are depicted in figures 1(a) and 1(b). As it can be observed in the figures, the exploiting of nearest neighbors distances through order statistics, makes it possible to automatically detect far outliers while still capturing the main support of the normal data.

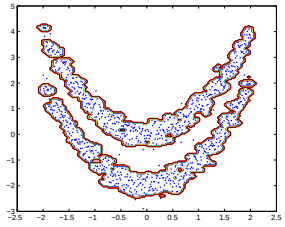
3.2 Real data sets

In this section we explore the applicability of the proposed method and compare it with well-established methods in the field of one-class classification. Specifically, three data sets from the UCI Machine Learning Repository [8] are used. The first one, Wine, was originally proposed as a binary classification problem and in this case it will be casted into a one-class classificacion following a one vs the rest approach. It is a well posed classification problem so it shall be treated as first benchmark under good conditions. The second one, Spambase, consists of a collection of spam and non-spam e-mails. This data set is a

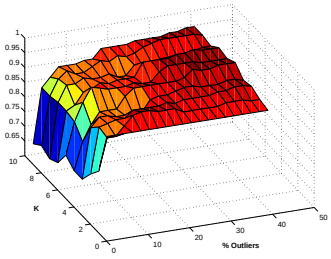
ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.



(a) 2D Random Gaussian data support method's capture.



(b) 2D Banana data support method's capture.



(c) Accuracy for Wine data set when changing hyperparameters.

good representation of the proposed method's applicability in abnormality detection from normal data (non-spam e-mail) in harder scenarios. The third one, Cardiotocography, exemplifies the applicability of this method to biomedical applications. This dataset consists of fetal cardiotocograms (CTGs), which were automatically processed to extract diagnostic features, and the diagnosis label of 'normal', 'suspect' and 'pathological'. For our case of one-class classification, we assume both suspect and pathologic as abnormal cardiotocograms. We compare the classification accuracy obtained by the proposed method with the one obtained by two most widespread used one-class classifiers: one-class ν-SVM [9] and Autoassociative Multilayer Perceptron [7].

For each data set, 30 random runs using 70% of normal class data as training set were done. In table 1, the mean accuracy of each method and its standard deviation is shown (best combination of hyperparameters along the random runs). As it can be observed, the proposed method obtains equal or better accuracy than the other two well established one-class classifiers. In addition, for the Wine data set, we tested the ability of the three methods to tackle noise samples in the normal state data set introducing in the training set a 10% of abnormal samples. It can be noticed that EVOC still maintains better accuracy than the other two tested methods. Moreover, in figure 1(c) the variability of the accuracy obtained by the proposed method is experimentally studied. It can be observed that for a large range of combinations of k and p, EVOC presents an stable behavior with accuracies above 92%.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0. Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

Data set	EVOC	One-class ν-SVM	Autoass-MLP
Wine	94.70% (0.001)	92.45% (0.004)	94.30 % (0.013)
Wine (10%)	93.25% (0.015)	91.70% (0.0034)	91.93 % (0.008)
Spambase	86.33% (0.003)	86.26% (0.002)	79.52 % (9.81e-4)
Cardiotocography	80.38% (0.009)	76.71% (0.007)	75.44 % (0.03)

Table 1: Results for UCI data sets.

4 Conclusion and future work

In this paper, a one-class classifier based on extreme value statistics is presented. The proposed methodology relies only in the existence of a measure of dissimilarity or metric between samples, so it can be extended to other spaces different from $\mathbb{R}^n$. Moreover, the proposed EVOC method has a reduced set of hyperparameters and presents a good performance compared to other frequently used one-class classifiers. In this paper, proposed method's performance is also explored in problems settled in the Euclidean space, obtaining good results. As it is a memory-based learning method, it can suffer when applied to large data set. Additional effort could be made in the future aiming at reducing final model's complexity using efficient nearest neighbor methods and trimming the normal state data set.

References

[1] V. Chandola, A. Banerjee, V. Kumar, Anomaly detection: a survey, In *ACM Computing Surveys*, vol. 41, no. 3, art. 15, 2009.

[2] D. M. Johannes, *One-class classification. Concept learning in the absence of counter-examples* , Delft University Phd. Thesis, 2001.

[3] E. Kriminger, C. Lakshminarayan, and Jose C. Principe, Nearest Neighbor distributions for imbalanced classification. In *IEEE International Conference on Acoustics, Speech ad Signal Processing* (ICASSP 2012), In Press.

[4] E. Castillo, A. S. Hadi, N. Balakrishnan, and J. M. Sarabia, *Extreme Value and Related Models with Applications in Engineering and Science*, Wiley Series in Probability and Statistics, 2005.

[5] M. Moya, M. Koch, and L. Hostetler, One-class classifier networks for target recognition applications. In *Proceedings of INNS World Congress on Neural Networks*, pages 797-801, 1993.

[6] C.C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim, On the Surprising Behavior of Distance Metrics in High Dimensional Space, In *Lecture Notes in Computer Science (ICDT '01)*, pages 420-434, 2001.

[7] N. Japkowicz, *Concept-Learning in the absence of counterexamples: an autoassociation-based approach to classification* , New Jersey State University Phd. Thesis, 1999.

[8] A. Frank, A. Asuncion, UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. University of California, School of Information and Computer Science, Irvine, CA, 2010

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