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. Interest in One-Class Classification methods has soared in recent years due to its wide applicability in many practical problems where 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 the embedding of 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 to its application on practical novelty detection problems.

1 Introduction

The problem of one-class classification can be stated as follows: using a data set which is known to have been 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 *n*ormal 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.
During the past years, a large number of different terms been used in the literature for this problem. The term one-class classification originates from the work in [5], but it is also named as outlier detection [8], novelty detection [6] or concept learning [2], 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 machinery fault detection [1][7] and document classification [9] their two more popular fields of application.
The number of methods in the literature specifically devised for one-class classification is scarce, although many algorithms in the realm of machine learning can be adapted for such task. Different approaches can be classified into: (a) Density

*This research was supported by the Spanish Ministry of Science and Innovation via a FPU doctoral grant to David Martinez-Rego.

methods: Gaussian model, Mixture of Gaussians, Parzen density estimation; (b) Boundary methods: K-centers, Nearest neighbor method, Support vector data description; (c) Reconstruction methods: k-means, LVQ, SOM, Principal Component Analysis, Mixtures of Principal Component Analyzers, Auto-Encoders and Diabolo networks. In [1] a thorough revision can be consulted.
In [4], 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 [3] is presented. It aims to exploit 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 a rather 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 [3] 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 density function $f(x)$ and 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 density $f(x)$ and distribution $F(x)$ functions. 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 $m(x)$ is the number of elements of the samples with a value $X_j \leq x$. 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. 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 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, $F_d$, $\alpha$)

foreach sample $s \in X$ do
    Calculate the vector 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 $\alpha$ leaving a $p$ fraction of data out of the support.

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

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

$$C(s) = I(P(D_k > d_k) - \alpha) = I\left(\prod_{i=1}^{k}(1 - F_{i:k}(d_k(i))) - \alpha\right) \tag{2}$$

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

$$C(s) = I(P(D_k > d_k) - \alpha) = I\left(\prod_{i=1}^{k}(1 - F_{i:k}(d_k(i))) - \alpha\right) \tag{3}$$

where I is the indicator function, d_k is the vector of distances to the k nearest neighbors of s in the normal state data set, and $F_{r:k}$ is the rth order statistics where we have plugged in the distribution function $F_d(x)$ of the distance to a neighbor and α is a threshold that controls under which level the data sample s is considered abnormal. Note that what the classification rule is monitoring is the probability of obtaining, under normal conditions, a set of k nearest neighbors more dissimilar that 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. In algorithm 1, the main steps of the proposed one-class classification model can be consulted.

3 Experimental results

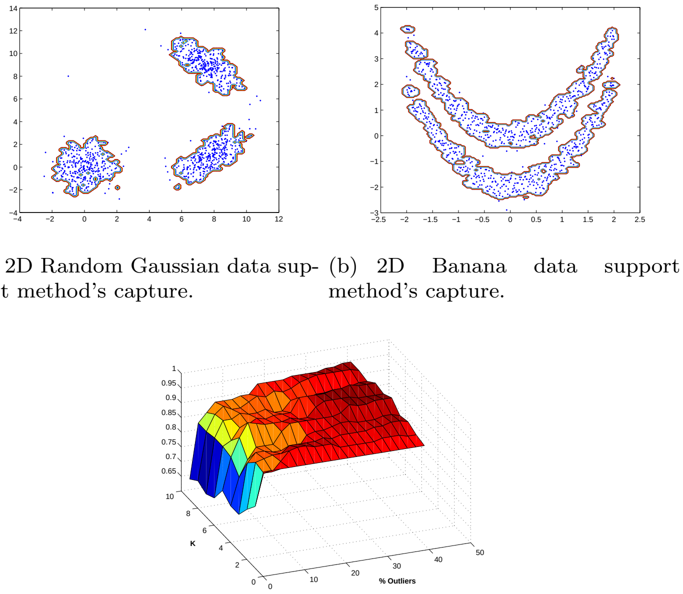
In this section we explore the features of the proposed method for both artificial and real one-class classification data sets. For the experiments presented hereafter, we adopt a parametric approach to model the distance between samples using as hypotesis the lognormal distribution.

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 a noise level of 6% 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 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 established methods in the field of one-class classification. Specifically, three data sets from the UCI Machine Learning Repository [10] are used. The first one, wine, was originally proposed as 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 on a collection of spam and non-spam e-mails. This data set is a good representation of 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 application. It consist in diagnose fetal cardiotocograms (CTGs), which were automatically processed and the respective diagnostic features extracted, between normal and pathological. For the case of one-class classification, we assume both suspect and pathologic as abnormal cardiotochograms. We compare the classification accuracy obtained by the proposed method with the one obtained by two most widespread used



(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.

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)
CardioTachography	80.38% (0.009)	76.71% (0.007)	75.44 % (0.03)

Table 1: Results for UCI data sets.

one-class classifiers: one-class ν-SVM [11] and Autoassociative Multilayer Perceptron [2]. For each data set, 30 random runs using 70% of normal class data were done, taking for each method the combination of hyperparameters giving the best restult. In table 1, the mean accuracy and its standard deviation is showed. 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 the proposed 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 the proposed model presents an stable behavior with accuracy above 92%.

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 R^n. Moreover, the proposed method has a reduced set of hyperparameters and presents a good performance compared to widespread used one-class classifiers. In this paper, proposed method's performance is also explored in problems settled in the Euclidean space, obtaining good results. Although it could be argued that as this is a memory-based learning method, additional effort could be made in reducing final model's complexity. Since the proposed method is only based on the distribution of distances between data under normal conditions, it is possible to trim the database used, provided that the distance distribution is updated accordingly.

References

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

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

[3] 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.

[4] 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.

[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. Bishop, Novelty detection and neural network validation. In *IEEE Proceedings on Vision, Image and Signal Processing. Special Issue on Applications of Neural Networks*, pages 217-222, 1994.

[7] D. Martinez-Rego, O. Fontenla-Romero and A. Alonso-Betanzos, Power Wind Mill Fault Detection via one-class nu-SVM Vibration Signal Analysis. In *Proceedings International Joint Conference on Neural Networks*, (IJCNN 2011), pages 511-518, 2011.

[8] G. Ritter, M. Gallegos, Outliers in statistical pattern recognition and an application to automatic chromosome classification. In *Pattern Recognition Letters*, 18:525-539, Elsevier, 1997.

[9] L.M. Manevitz, M. Yousef, One-Class SVMs for Document Classification. In *Journal of Machine Learning Research*, 2:139-154, Elsevier, 2001.

[10] 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

[11] 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.