Partially supervised anomaly detection using convex hulls on a 2D parameter space

Gabriel B. P. Costa[1], Moacir Ponti[1], and Alejandro C. Frery[2]

[1] Instituto de Ciências Matemáticas e de Computação — Universidade de São Paulo
13566-590 São Carlos, SP, Brazil
moacir@icmc.usp.br, gpbcosta@icmc.usp.br
[2] Universidade de Alagoas — Maceió, AL, Brazil

Abstract. Anomaly detection is the problem of identifying objects appearing to be inconstistent with the remainder of that set of data. Detecting such samples is useful on various applications such as fault detection, fraud detection and diagnostic systems. Partially supervised methods for anomaly detection are interesting because they only need data labeled as one of the classes (normal or abnormal). In this paper, we propose a partially supervised framework for anomaly detection based on convex hulls in a parameter space, assuming a given probability distribution. It can be considered a framework since it support any distribution or combination of distribution, modelling only normal samples. We investigated an algorithm based on this framework, assuming the normal distribution for the not anomalous data and compared the results with statistical algorithms, the One-class SVM and Naive Bayes classifiers. The proposed method performed well and showed results comparable or better than the competing methods. Furthermore, this approach can handle any probability distribution or mixture of distributions, allowing the user to choose a parameter space that better models the problem of finding anomalies.

Keywords: Anomaly, outlier, semi-supervised learning

1 Introduction

Anomaly detection is the problem of finding patterns with an unexpected behaviour. Barnett and Lewis [1] defined outlier (anomaly) as an observation (or subset of observations) which appears to be inconstistent with the remainder of that set of data.

Due to the nature of the problem, anomalies are often rare and dealing with it can help on applications such as fault detection, fraud detection, network intrusion, diagnostic systems, medical condition monitoring. An anomaly detection method take as input a sample or set of samples, and indentify whether those samples are "normal" or "abnormal", according to what is expected to be found. On most applications, "normal" samples are widely available, while anomalies are scarse or frequently not available at all [2].

Detecting anomalies can be considered critical in many systems, as they indicate an abnormal condition. Some examples are given by Hodge and Austin [6]: an engine rotation defect, a flow problem on a pipeline, an intruder inside a system with malicious intentions, a fault on a factory production, a disease or a dangerous medical condition, among others.

There are three fundamental approaches to detect anomalies: 1) unsupervised, 2) supervised and 3) partially supervised. In the first approach, when there is no prior knowledge of the data, unsupervised learning algorithms such as clustering are used. The general idea is to identify outliers, observations that appear not to belong to any of the detected groups. The supervised approach models both normality and abnormality, requiring samples labeled as both normal and abnormal. In this paper we focus on a semi-supervised (or partially supervised) method, which models only normality (in a very few cases model only abnormality [4]).

According to Hodge and Austing [6], the advantages of using a partially supervised method to detect anomalies are: a) it only needs data labeled as normal, b) it is suitable for static or dynamic data, as it only learns one class, c) most method are incremental, d) it does not assume any distribution for the abnormal data.

In a review paper, Chandola et al. [2] pointed out that nearest neighbor and clustering-based method can suffer with high-dimension datasets, since distance measures are not able to differentiate between normal and anomalous instances. Classification-based techniques can be a better choice in this scenario but require sufficent labeled data from both normal and abnormal classes. Statistical techniques are effective when the dimensionality of data is low and statistical assumptions hold.

In this paper we propose a framework for anomaly detection based on convex hulls in a distribution parameter space (instead of a feature space), assuming a probability distribution with 2 parameters, in this case a Gaussian distribution. The convex hull computed over a set of points representing normal instances. Each axis of the parameter space represents one parameter of the probability distribution assumed. It is possible to use any distribution and high-dimensional data.

2 Anomaly detection based on convex hulls on a (μ, σ) parameter space

The training stage of our method computes a set of pairwise parameter points $\hat{\theta}$ of a given distribution. The points are computed using pairs of randomly selected normal instances. For instance, assuming a Gaussian distribution, $\hat{\theta} = (\hat{\mu}, \hat{\sigma})$, and a 2-dimensional parameter space is formed by the points.

In order to obtain the points inside the parameter space, every pair of feature vectors are concatenated (Fig. 1), so that a single convex hull is obtained, and the parameters estimated with the whole concatenated feature vector. After that, a convex hull H_N is computed over the set of points, obtaining a geometric interpretation of the normal data.

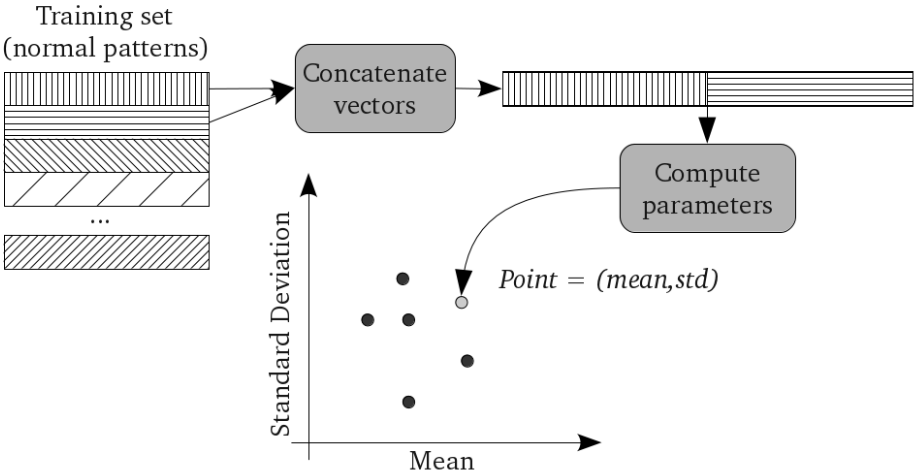


Fig. 1. Pairwise parameter estimation by concatenation of pairs of feature vectors

New instances are classified by computing a new set of pairwise parameter points, this time the points are obtained by pairing the new unknown instance with each point inside the normal convex hull, H_N. A new convex hull H_U is computed over this new set of points.

Assuming that the new pattern comes from the same law that the normal set of data, the detection algorithm uses the intersection $H_N \cap H_U$. The intersection is expected to be high if an anomaly is observed, as H_N was formed by the set of points obtained by pairs of normal observations, while H_U was obtained by merging the new pattern with only normal observations.

The algorithm is defined as follows:

- **I – Normal class parameter estimation**, estimate $\hat{\Theta}_N$, which is the set of parameters for normal data pairs:
  1. selects two normal instances, a and b,
  2. concatenate the values of a and b, and compute all parameters $\hat{\theta}_N$, a representation of this procedured for one point generation is shown in Figure 1,
  3. compute the convex hull H_N using as input the set of parameter points $\hat{\Theta}_N$.
- **II – New observation parameter estimation**, estimate $\hat{\Theta}_U$, which is the set of parameters for normal data pairs computed with the unknown data:
  1. given a new output x
  2. for each point $y \in H_N$, concatenate y and x, and compute the parameters $\hat{\Theta}_U$.
  3. compute the convex hull H_U using as input the set of parameter points $\hat{\Theta}_U$.
- **III – Convex hull intersection**, compute $I = H_N \cap H_U$.
- **IV – Detection**: use the difference of the size of the intersection I in comparison to both convex hulls, H_N and H_U:
  1. obtain $d = (H_N - I) + (H_U - I)$
  2. compare the value of d with a threshold obtained through a validation set composed by normal samples. The higher d value obtained is considered the threshold.
  3. x is considered an anomaly if its value of d is higher then the threshold obtained in the previous step.

3 Experiments

In this paper we assumed the normal distribution for the “normal” data (no assumption is made by our algorithm for the anomalies). Therefore, our parameter space is 2-D (considering mean and standard deviation parameters).

We used a repeated random subsampling validation, each experiment was repeated 10 times, using 70% for training, 15% for validation and 15% for testing. The average and standard deviation were computed by these repetitions. The evaluation was based on an balanced accuracy value that takes into account the balance between the classes:

$$\text{Acc} = 1 - \frac{\sum_{i=1}^{c} E(i)}{2c},$$

where c is the number of classes, and $E(i) = e_{i,1} + e_{i,2}$ is the partial error of c, computed by:

$$e_{i,1} = \frac{FP(i)}{N - N(i)} \text{ and } e_{i,2} = \frac{FN(i)}{N(i)}, i = 1, ..., c,$$

where $FN(i)$ (false negatives) is the number of samples belonging to i incorrectly classified as belonging to other classes, and $FP(i)$ (false positives) the samples $j \neq i$ that were assigned to i [10].

In order to find a threshold for d that is able to verify if a new observation is considered an anomaly, we used a validation set composed by 15% of samples. Those normal validation samples were not used neither in the training nor the test stages. However, the value found for the thresold was used further to test the unseen instances.

The methods used to compare against our results were: the normal univariate and multivariate statistical algorithms [1], and the normal Bayes classifier [3]. All competing methods assume a Gaussian distribution of the data, except for the One-class SVM that fits a hyper-sphere to the data in order to describe it in a high dimensional space.

The experimental settings for each algorithm was:

- **Normal univariate and multivariate algorithms**: were trained with 70% of the normal data, and the threshold for the probability obtained by using 15% of both normal and anomalous data. The test step used 15% of the dataset, including both normal and anomalous samples.
- **Naive Bayes classifier**: was trained with 70% of the normal data and 5% of the anomalous data, as this is the proportion of available abnormal samples often found on datasets [2]. The test step used 15% of the dataset, including both normal and anomalous samples.
- **One-class SVM**: setting similar to Naive Bayes. A grid search with step 0.25 was performed to tune the parameters, using an evaluation dataset with 10% of the samples.
- **Convex Hull on parameter space method**: trained with 70% of the normal data, and the threshold for d (difference between convex hulls), obtained by using 15% of normal samples. The test step used 15% of the dataset, including both normal and anomalous samples.

The proposed detector is denoted Conv.Hull-PS, where the feature vectors from the paired samples are concatenated in order to compute the parameters.

Table 1. Dataset characteristics

Dataset	Type	#Samples	#Features	Anomaly rate
Normal-vs-2	synthetic	112	2	10.7%
Ionosphere	real	193	34	23.8%
Parkinsons	real	351	23	36.0%
BreastR	real	147	12	17.0%
BreastW	real	699	9	34.5%

Detailed information about the datasets used in the experiments are shown in Table 1, including synthetic and real data:

- **Normal vs 2 distributions**: a normally distributed class was randomly generated to be the "normal" class, with 100 samples. Another 12 samples were generated using a Lithuanian distribution (6 samples) and a Banana-shaped distribution (6 samples), to be used as anomalies.
- **Ionosphere**: Ionosphere data from UCI Machine Learning Depository [5]. Consists of a phased array of 16 high-frequency antennas that trasmitted around 6.4 kilowatts. Are considered "normal" radars those that showed evidence of some type of structure in the ionosphere. "anomalous" radars are those whose signals pass through the ionosphere and therefore do not show [11]. This dataset is composed by 351 instances of which 225 are considered good and 126 bad. Every instace has 34 continuous attributes.
- **Parkinson**: dataset from UCI Machine Learning Depository [5]. Created by Max Little of the University of Oxford, in collaboration with the National Centre for Voice and Speech, Denver, Colorado, who recorded speech signals. The original study published the feature extraction methods for general voice disorders. It is composed of a range of biomedical voice measurements from 31 people, 23 with Parkinson's disease [7].
- **BreastR**: breast cancer fom University of Roma "Tor Vergata". This dataset consists of 127 cases of benign tumors and 25 cases of malignant tumors. Each instance has 12 attributes that were obtained through the use of Gabor Wavelets [9].

- **BreastW**: Wisconsin Breast Cancer data from UCI Machine Learning Depository [5]. This dataset was obtained from the University of Wisconsin Hospitals by Dr. William H. Wolberg [8]. It has 699 instances each one with with 9 integer attributes. 458 of this instances are classified as benign and 241 as malignant.

4 Results and Discussion

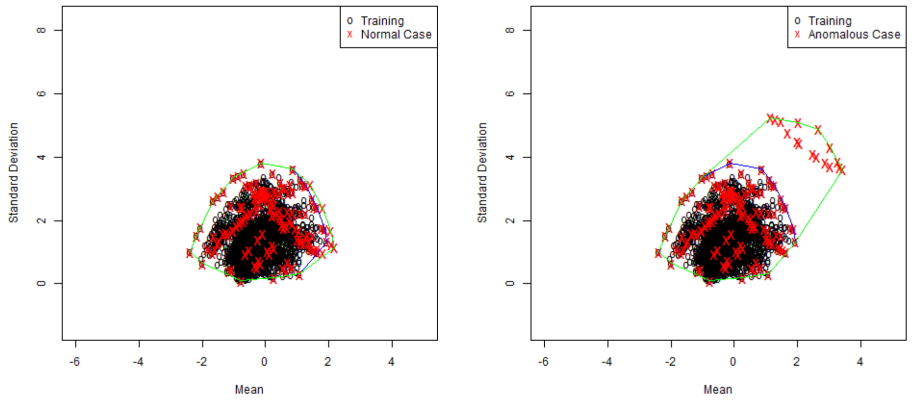
The average balanced accuracies and standard deviation values are shown in Table 2. The boldfaced values represents the best results by comparing mean and standard deviation. Besides, an example of parameter space with pairwise estimated points, and convex hulls computed from normal and abnormal samples are shown in Figure 2.

Table 2. Results

Dataset	Normal-Univ	Normal-Multiv	Bayes Class.	One-class SVM	Conv.Hull-PS
Normal-vs-2	**93.8 ± 4.9**%	**95.2 ± 1.4**%	50.0 ± 0.0%	80.1 ± 0.1%	**93.9 ± 1.5**%
Ionosphere	71.2 ± 5.2%	**78.2 ± 2.4**%	**78.5 ± 8.6**%	72.0 ± 0.1%	**77.2 ± 2.7**%
Parkinsons	62.9 ± 6.5%	63.9 ± 5.2%	54.9 ± 6.4%	**67.3 ± 0.2**%	**65.5 ± 5.6**%
BreastR	67.2 ± 7.5%	**70.1 ± 6.6**%	50.0 ± 0.0%	**70.6 ± 0.7**%	52.1 ± 2.9%
BreastW	**94.1 ± 2.3**%	92.6 ± 2.0%	85.1 ± 1.6%	90.8 ± 0.5%	**93.1 ± 1.8**%

The proposed method achieved results comparable or better than the competing methods, except for the BreastR dataset. The average performance for this database is probably due to the high overlap rate presented by the normal and anomalous cases, which cannot be captured by the distribution model. This overlap rate makes it harder to identify the anomalies since they are spacially mixed with the normal cases. The One-class SVM was better probably because it fits a hypersphere in a higher dimensional space. However, it's possible that by changing the parameter space used to classify the instances, the results presented by the proposed method can significantly improve. Moreover, the results presented by the other datasets show that using a parameter space can help on an anomaly detection task when only normal samples are available.

The Naive Bayes classifier suffered from the scarse anomaly data available, misclassifying most anomalies. The statistical methods showed results comparable with our method, but could not deal well higher dimensional datasets, while our algorithm was more robust in such cases.



(a) (b)

Fig. 2. Examples of convex hulls on parameter spaces using the dataset Normal-vs-2. The green line shows H_U and the blue line H_N. A normal sample detection is shown in (a) and an anomaly detection shown in (b).

The results supports this claim, specially by the results obtained with Parkinsons and Ionosphere datasets.

5 Conclusions

This paper reports results of a new anomaly detection framework based only on normal class samples. The interesting feature of this framework is that it can handle any probability distribution or mixture of distributions. It also include all advantages of partially supervised algorithms. Besides it is possible to include information about anomalies in the validation step. The parameter space allow one to specify parameters that better models the problem of finding anomalies. The use of a convex hull make it possible to draw the boundary between normal and abnormal data behaviour. Future works can explore variations by including the use of multiple parameters, exploring both the use of anomalous data in the training step.

Acknowledgment

This work was supported by FAPESP (grants n. 2011/16411-4 and n. 2012/12524-1).

References

1. Barnett, V., Lewis, T.: Outliers in statistical data. John Wiley & Sons (1994)
2. Chandola, V., Banerjee, A., Kumar, A.: Anomaly detection: a survey. ACM Computing Surveys 41(3), 15 (2009)
3. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. Wiley (2000)
4. Fawcett, T., Provost, F.J.: Activity monitoring: noticing interesting changes in behavior. In: Proc. 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 53–62 (1999)
5. Frank, A., Asuncion, A.: UCI machine learning repository (2010), http://archive.ics.uci.edu/ml
6. Hodge, V.J., Austin, J.: A survey of outlier detection methodologies. Artificial Intelligence Review 22(2), 85–126 (2004)
7. Little, M., McSharry, P., Roberts, S., Costello, D., Moroz, I.: Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. BioMedical Engineering OnLine 6, 23 (2007)
8. Mangasarian, O.L., Wolberg, W.H.: Cancer diagnosis via linear programming. SIAM News 23(5), 1–18 (1990)
9. Mencattini, A., Salmeri, M., Casti, P., Pepe, M.L., Mangieri, F., Ancona, A.: Local active contour models and gabor wavelets for an optimal breast region segmentation. Computer Assisted Radiology and Surgey (CARS 12) (2012)
10. Ponti, M.P.: Segmentation of low-cost remote sensing images combining vegetation indices and mean shift. Geoscience and Remote Sensing Letters, IEEE 10(1), 67–70 (2013)
11. Sigillito, V.G., Wing, S.P., Hutton, L.V., Baker, K.B.: Classification of radar returns from the ionosphere using neural networks. Johns Hopkins APL Technical Digest 10, 262–266 (1989)