A novel method for anomaly detection using beta Hebbian learning and principal component analysis

FRANCISCO ZAYAS-GATO, *Department of Industrial Engineering, University of A Coruña, CTC, Avda. 19 de Febrero s/n, 15405, Ferrol, A Coruña, Spain.*

ÁLVARO MICHELENA, *CITIC Research, University of A Coruña, Elviña Campus s/n, 15008, A Coruña, Spain.*

HÉCTOR QUINTIÁN, *Department of Industrial Engineering, University of A Coruña, CTC, Avda. 19 de Febrero s/n, 15405, Ferrol, A Coruña, Spain.*

ESTEBAN JOVE*, *Department of Industrial Engineering, University of A Coruña, CTC, Avda. 19 de Febrero s/n, 15405, Ferrol, A Coruña, Spain. CITIC Research, University of A Coruña, Elviña Campus s/n, 15008, A Coruña, Spain.*

JOSÉ-LUIS CASTELEIRO-ROCA, *Department of Industrial Engineering, University of A Coruña, CTC, Avda. 19 de Febrero s/n, 15405, Ferrol, A Coruña, Spain.*

PAULO LEITÃO, *CeDRI - Research Centre in Digitalization and Intelligent Robotics, Polytechnic Institute of Bragança and INESC TEC, Porto, Portugal.*

JOSÉ LUIS CALVO-ROLLE, *Department of Industrial Engineering, University of A Coruña, CTC, Avda. 19 de Febrero s/n, 15405, Ferrol, A Coruña, Spain. CITIC Research, University of A Coruña, Elviña Campus s/n, 15008, A Coruña, Spain.*

Abstract

In this research work a novel two-step system for anomaly detection is presented and tested over several real datasets. In the first step the novel Exploratory Projection Pursuit, Beta Hebbian Learning algorithm, is applied over each dataset, either to reduce the dimensionality of the original dataset or to face nonlinear datasets by generating a new subspace of the original dataset with lower, or even higher, dimensionality selecting the right activation function. Finally, in the second step Principal Component Analysis anomaly detection is applied to the new subspace to detect the anomalies and improve its classification capabilities. This new approach has been tested over several different real datasets, in terms of number of variables, number of samples and number of anomalies. In almost all cases, the novel approach obtained better results in terms of area under the curve with similar standard deviation values. In case of computational cost, this improvement is only remarkable when complexity of the dataset in terms of number of variables is high.

Keywords: One-class, dimensional reduction, BHL, PCA.

*E-mail: esteban.jove@udc.es

Vol. 31, No. 2, © The Author(s) 2022. Published by Oxford University Press.

Advance Access published 21 February 2022 https://doi.org/10.1093/jigpal/jzac026

1 Introduction

In recent decades, the world is facing a significant technology development in terms of communication, digitalization and instrumentation, which have changed the operation environment and internal functioning of industries and companies [12]. Overlooking these breakthroughs could result in competitiveness problems, especially in the current global market. In fact, according to different authors [4, 17], the digitalization will lead to a higher impact than the industrial revolution that took place in 18th century.

One of immediate consequences of these advances is the availability of detailed information about all processes involved in an activity. The possibility of registering a wide number of variables helps to monitor the current operation of a system regardless the field: medicine, industry or cyber security, among others [2, 7]. The information gathered plays a key role to supervise the correct performance and detect any kind deviation from the expected operation, since it contributes to a desired optimization [11]. An example of this circumstance could be the fraud detection on credit cards transactions or the breast cancer diagnosis [3].

Due to the reasons explained above, the use of anomaly detection technique has been the focus of attention of the scientific community in recent years. Depending on the features of the data available, three types of approaches can be considered to detect anomalies [3]. First, if the registered dataset presents instances labelled as correct and anomalous, supervised classifiers are proposed, which are trained to separate two known classes. In some cases, the nature of the dataset is not available and, hence, unsupervised techniques are applied to identify the label of each of the instances without previous knowledge. However, the most common scenario is presented when only information about correct operation is available and the potential anomalies are limited or unknown. In this case, the use of one-class classifiers are taken into consideration to determine if new test instances belong to the correct operation class, also known as target class. Samples outside this class are considered anomalous [10].

One of the most used approaches to tackle the one-class classification problem is based on the use of reconstruction methods [16]. They are trained to model the system behaviour and calculate the residuals between the input and the output, also known as reconstruction error. To achieve this approach, the Principal Component Analysis (PCA) and Autoencoder are two of the most used techniques, offering great performance [7].

On the other hand, the use of boundary techniques to establish the limits around the target class can also be considered. In [16], the Support Vector Data Description is proposed. This method maps the training data into a high-dimensional space to determine the boundaries of the target class. On the contrary, other techniques such as Approximate Polytope Ensemble [1] or Non-Convex Boundary Over Projections [8], aim to determine the limits of the training set by means of 2D projections. These techniques are especially interesting when the dataset is high dimensional.

This structure of this paper is as follows: after this section, the motivation of the presented contribution is described. Then, next section details the proposal. After Section 4, the following section describes the experiments and the results carried out to validate the proposed method. Finally, the conclusions and future works are presented.

2 Motivation

The use of PCA has been widely applied for dimensional reduction tasks and one-class classification purposes in many different fields [7, 15]. The main idea of this technique is to reduce the dimensionality of a training set using a linear transformation looking for the minimum loss of

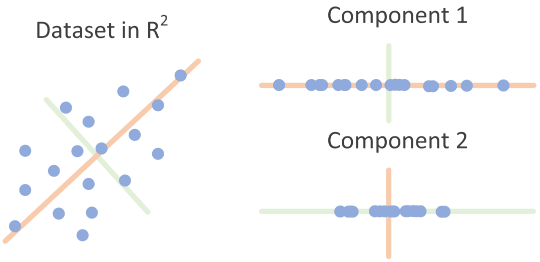

FIGURE 1. An example of dimensional reduction using PCA.

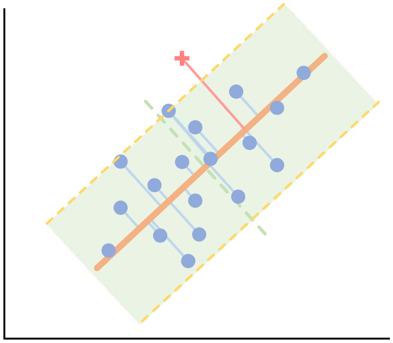

FIGURE 2. Training set, PCA components and anomaly detection.

information [16]. This is achieved through the identification of the directions where the data presents higher variation. Each direction is called principal component [15].

PCA finds new features by means of linear combinations of the original set. This subspace mapping is the result of the calculation of the covariance matrix eigeinvalues being the principal components of the eigeinvectors with the highest eigenvalues [16].

To illustrate the main goal of this technique, an example of a 2D dataset is presented in Figure 1. In the left side, the two principal components are depicted in brown and green traces and represent the two main direction variabilities of the dataset. In the right side, the dataset is projected from $\mathbb{R}^2$ to $\mathbb{R}^1$.

This technique has been widely used for many different tasks, such as face recognition [5] or denoising data [9], among data. Furthermore, the use of principal components can be considered to achieve a one-class approach. Following this idea, the training data is supposed to belong to the target class and the directions with the highest variability are computed. Then, the distances from the training points to its projections represent a measure to determine the appearance of anomalies. An example of this operating basis can be found in Figure 2, where the blue dots are the training points and the distances to the first principal component are depicted. The shaded area indicates the maximum distance computed on the training set. The figure includes the appearance of a test point labelled as anomalous, since the distance to its projection is higher than the calculated during the training stage.

This simple but effective method has led to interesting results if the subspace is clearly linear [16]. Then, it is commonly applied to solve a wide range of applications, being considered as one of

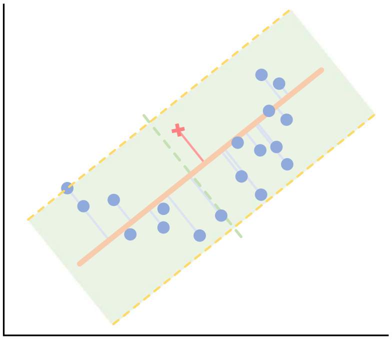

FIGURE 3. Training set, PCA components and misclassification.

the most used one-class techniques. Besides its good performance, a significant advantage of this approach is the low computation times, especially compared with similar approaches, such as the Autoencoder technique [7].

Despite the remarkable strengths of PCA, it presents a weakness caused by the fact that its use over nonlinear sets may result in low classifying performance. An illustration of this situation is depicted in Figure 3, where the principal components are not suitable to determine the appearance of an anomaly. In this case, the shape of the training set leads to misclassification when the anomalies are inside the convex area of the dataset. This weakness represents the main motivation of this work and reflects the need of adding a previous step to avoid the misclassification caused by nonlinear data shapes. This should be done without undermining the positive features of PCA in terms of computation time and performance over linear sets.

3 Proposal

Based on previous motivation and in order to improve the capability of PCA to detect anomalies, a novel hybrid method defined by a two-step process is proposed in this research.

In the first step a recent nonlinear Exploratory Projection Pursuit (EPP) method called Beta Hebbian Learning (BHL) [13], based on a new family of learning rules of negative feedback network, is applied either to reduce the dimensionality of the original dataset or to face to nonlinear datasets by generating a new subspace of the original dataset with lower, or even higher, dimensionality selecting the right activation function. In the case of PCA, it is based on first statistical moment, the variance, in order to project the original dataset in a new subspace that maximize the variance by some scalar projection of the data. However, in case of BHL, it provides a new subspace that can maximize other statistical moments such as kurtosis and skewness based on the right combination of parameters, providing in this way a better representation of the original dataset and allowing to find hiding structures in the original dataset not discovered by PCA.

With this first step it is possible to get two improvements at once, first the reduction of dimension will reduce the computational cost of the next step (PCA anomaly detection) and also prepare the dataset to allow PCA anomaly detection to improve the classifying performance when it deals with linear and nonlinear datasets.

In the second step PCA anomaly detection is applied to the new subspace to detect the anomalies and improve its classification capabilities. In order to illustrate this process, Figure 4 shows the described hybrid method.

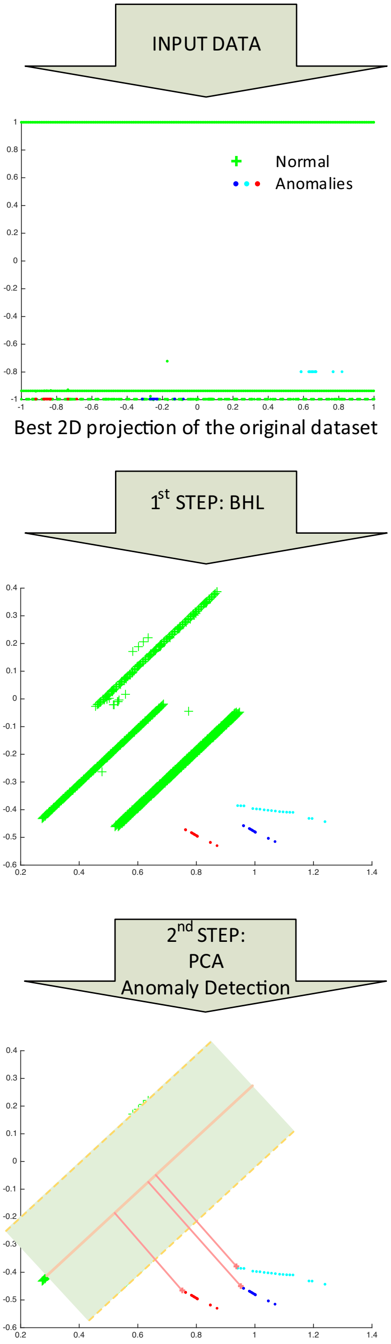

FIGURE 4. Two steps of the developed model approach.

As can be seen in this figure, the original dataset presents normal samples (green crosses) and anomalies (blue, cyan and red dots) mixed, therefore application of PCA anomaly detection technique will not provide satisfactory results. However, after applying BHL (first step) normal samples and anomalies are now easily separable, and therefore PCA anomaly detection can settle the right boundaries to discriminate normal samples from anomalies, as can be seen in the final step of Figure 4.

3.1 Beta Hebbian Learning

BHL is an EPP algorithm based on a novel family of learning rules derived from the PDF of the residual based on Beta distribution to extract information from high-dimensional datasets by projecting the data onto low-dimensional (typically 2D) subspaces and providing a clear representation of the data's internal structure.

Therefore, applying the Beta PDF over the residual of the neural network ($e = Wy$), it obtained the following PDF:

$$p(e) = e^{\alpha-1}(1-e)^{\beta-1} = (x - Wy)^{\alpha-1}(1 - x + Wy)^{\beta-1}, \tag{1}$$

where α and β define the shape of Beta PDF, e is the residual, x is the input of the network, W is the weight matrix and y is the output of the network.

If the gradient descend is applied to maximize the likelihood of the data respect to weights, the weights update rule can be defined through Equation 3.

$$Feedforward : y_i = \sum_{j=1}^{N} f(W_{ij}x_j), \forall i \tag{2}$$

$$Feedback : e_j = x_j - \sum_{i=1}^{M} W_{ij}y_i \tag{3}$$

$$Weightsupdate : \varDelta W_{ij} = \eta(e_j^{\alpha-2}(1-e_j)^{\beta-2}(1-\alpha+e_j(\alpha+\beta-2)))y_i \tag{4}$$

Once the training process have finished, the original dataset is projected onto a new subspace by applying the feed-forward step and all new components are projected by means of a scatter plot matrix.

3.2 PCA anomaly detection

PCA for anomaly detection is based on the reconstruction error of PCA [ref]. Therefore, from the original data a new subspace is obtained by means of the eigenvectors of the data covariance matrix, from a geometrical point of view, such operation consist on a rotation of the original axes to a new ones ordered in terms of variance, which can be expressed by equation 5.

$$y_i = W_i^T x, \tag{5}$$

where x^d in an N-dimensional space onto vectors in an M-dimensional space $(x_1...x_N)$, where $M \leq N$, Wi are the N eigenvectors of the covariance matrix and y are the projected original data onto the new output M-dimensional subspace $(y_1...y_N)$.

TABLE 1. Main characteristics of the 10 datasets used to validate the model.

Dataset	N° of samples	N° of variables	N° of anomalies
BreastW	683	6	239
Cardio	1831	21	176
Letter	1600	32	100
Vowels	1456	12	50
Wine	129	13	10
Annthyroid	7200	6	534
Mammography	11183	6	260
Musk	3062	166	97
Speech	3686	400	61
WBC	378	30	21

Once the original data is projected onto the new axes, the reconstruction error is computed as the difference between the original dataset and its projection over the new subspace, computed in the original data. Such projection can be expressed as

$$x_{proj} = W(W^T W)^{-1} W^T x \tag{6}$$

and finally the reconstruction error is obtained through the following measure:

$$f(x) = ||x - x_{proj}||^2. \tag{7}$$

In order to determine if a new sample is considered as an anomaly, during training process two criteria are established:

1. if the percentage of anomalies present in the original data is 0%, then the reconstruction error limit for a new sample is set as the maximum reconstruction error obtained during the training process;
2. if the percentage of anomalies present in the original data is higher than 10%, then to establish the new limit the 10% of higher reconstruction errors are discarded and the maximum of the remaining values is selected.

4 Experiments and results

In this section, datasets used to validate the novel approach are described, experiments developed are detailed and finally results obtained are presented.

In order to validate the proposed model, several real datasets from OODS dataset repository [14], with different complexity in terms of number of samples, number of variables and number of anomalies have been tested to guarantee the performance of the model with a wide range of scenarios. Table 1 summarizes the main characteristics of the used datasets.

The experimental part of this research can be divided in two parts. The first step (BHL) consists on applying BHL over original dataset with a combination of α and β parameters that generate a PDF

TABLE 2. Results obtained for each dataset.

Dataset	Dimensions	Components	Best AUC	*k–fold* cariance	Cost ms
BreastW orginal	9	1	94.35	1.23	18.17
BreastW BHL	6	1	95.21	1.15	30.35
Cardio orginal	21	4	89	1.14	21.82
Cardio BHL	10	3	91.94	0.82	34.62
Letter orginal	32	25	74.75	1.22	25.5
Letter BHL	20	14	78.03	1.82	32.92
Vowels orginal	12	4	87.96	1.63	21.47
Vowels BHL	6	5	89.62	1.53	39.22
Wine orginal	13	4	93.14	3.9	40.17
Wine BHL	8	7	95.18	3.89	36.72
Annthyroid orginal	6	4	90.69	0.59	41.83
Annthyroid BHL	3	4	95.06	0.63	45.37
Mammography orginal	6	2	83.42	0.71	59.77
Mammography BHL	4	1	85.5	1	29.44
Musk orginal	166	26	100	0	77.6
Musk BHL	6	5	99.97	0.08	46.01
Speech orginal	400	385	57.38	2.25	190.41
Speech BHL	12	284	57.9	2.41	113.02
WBC orginal	30	4	90.86	1.78	29.03
WBC BHL	6	6	93.14	3.26	36.99

similar to the original dataset samples distribution [13] and searching the most clear projections in the new subspace by projecting all combinations in a scatter plot matrix.

In the second step, a wide range of parameters are tested. It is combined three types of data pre-processing and for each type it is generated a new experiment with an incremental number of components. First, the data is normalized using a zero to one normalization over each variable. The second pre-processing is the z-score, that measures how many standard deviations a point is away from the mean. Finally, the raw data is introduced to the algorithm. Then, for instance, if original dataset has 6 variables, it is generated a total of 18 experiments (3 pre-processing * 6 components). Finally, for each experiment generated, in order to validate the PCA anomaly detection performance, a *k-fold* cross-validation is proposed. It is important to emphasize that, in this case, only target objects are considered to train the classifier.

In all datasets, PCA anomaly detection has been tested with the original dataset against new subspace generated in previous step by BHL.

For all experiments 10folds have been used, and mean and variance of the area under the curve (AUC) measure (related to Receiving Operating Characteristic (ROC) curve [6]) has been recorded.

Final results obtained for each dataset are presented in Table 2, where a comparison between PCA anomaly detection over the original dataset and over new subspace generated by BHL can be seen. For all dataset, the number of dimensions of the dataset, optimal number of components used for PCA anomaly detection, the best AUC among all experiments performed, the variance for this AUC in the *k-fold* and the computational cost of PCA anomaly detection training process are detailed.

In almost all cases the new proposal obtains better results in terms of AUC with similar standard deviation values. These improvements go from 0.52% to 4.37% and only in one case that BHL model is not able to improve the results; however, in this case the AUC difference is less than 0.03%. In terms of computational cost, differences are especially remarkable when dimensions of the original dataset are high, as in the case of *Musk* and *Speech* datasets, when dimensions of the dataset are smaller, results are not conclusive as in some cases BHL improves this time, as in case of *Mammography* dataset (reducing the cost in a 50%), and in other cases BHL takes more time, as in case of *Letter* dataset (increasing the cost in a 13%).

5 Conclusions and future works

In the present work, a novel approach consisting a two-step system for PCA anomaly detection has been tested over several real datasets and compared with the simple version of PCA anomaly detection method, in terms of accuracy and computational cost.

Based on the research performed in this study, it can be concluded that the use of new techniques for dimensionality reduction and exploratory projection pursuit models as an initial step can lead to an effective improvement of the performance of anomaly detection in terms of AUC. Especially when dimensionality of the datasets is high, the computational cost is also improved by the novel approach presented.

Future works will include testing the approach with other well-known EPP algorithms such as Maximum Likelihood Hebbian Learniing and also over other anomaly detection techniques.

Funding

Funding for open access charge: Universidade da Coruña/CISUG.

Acknowledgements

CITIC, as a Research Center of the University System of Galicia, is funded by Consellería de Educación, Universidade e Formación Profesional of the Xunta de Galicia through the European Regional Development Fund and the Secretaría Xeral de Universidades (ref. ED431G 2019/01).

References

[1] P. Casale, O. Pujol and P. Radeva. Approximate polytope ensemble for one-class classification. *Pattern Recognition*, **47**, 854–864, 2014.

[2] J.-L. Casteleiro-Roca, E. Jove, J. M. Gonzalez-Cava, J. A. M. Pérez, J. L. Calvo-Rolle and F. B. Alvarez. Hybrid model for the ANI index prediction using remifentanil drug and EMG signal. *Neural Computing and Applications*, **32**, 1249–1258, 2018.

[3] V. Chandola, A. Banerjee and V. Kumar. Anomaly detection: a survey. *ACM Computing Surveys (CSUR)*, **41**, 15, 2009.

[4] C. Degryse. Digitalisation of the economy and its impact on labour markets. *ETUI Research Paper-Working Paper*, 2016.

[5] B. A. Draper, K. Baek, M. S. Bartlett and J. R. Beveridge. Recognizing faces with PCA and ICA. *Computer Vision and Image Understanding*, **91**, 115–137, 2003.

[6] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, **27**, 861–874, 2006.

[7] E. Jove, J. Casteleiro-Roca, H. Quintián, J. A. Méndez-Pérez and J. L. Calvo-Rolle. Anomaly detection based on intelligent techniques over a bicomponent production plant used on wind generator blades manufacturing. *Revista Iberoamericana de Automática e Informática Industrial*, **17**, 84–93, 2020.

[8] E. Jove, J.-L. Casteleiro-Roca, H. Quintián, J.-A. Méndez-Pérez and J. L. Calvo-Rolle. A new method for anomaly detection based on non-convex boundaries with random two-dimensional projections. *Information Fusion*, **65**, 50–57, 2020.

[9] S. Mika, B. Schölkopf, A. J. Smola, K.-R. Müller, M. Scholz and G. Rätsch. Kernel PCA and de-noising in feature spaces. In *Advances in Neural Information Processing Systems*, pp. 536–542, 1999.

[10] D. Miljković. Fault detection methods: a literature survey. In *The 2011 Proceedings of the 34th International Convention MIPRO*, pp. 750–755. IEEE, 2011.

[11] Ian Parmee and Prabhat Hajela. Optimization in industry. Springer Science & Business Media, 2012.

[12] P. Parviainen, M. Tihinen, J. Kääriäinen and S. Teppola. Tackling the digitalization challenge: how to benefit from digitalization in practice. *International Journal of Information Systems and Project Management*, **5**, 63–77, 2017.

[13] H. Quintián and E. Corchado. Beta hebbian learning as a new method for exploratory projection pursuit. *International Journal of Neural Systems*, **27**, 1–16, 2017.

[14] S. Rayana. *Odds library*, 2016.

[15] M. Ringnér. What is principal component analysis? *Nature Biotechnology*, **26**, 303, 2008.

[16] D. M. J. Tax. *One-Class Classification: Concept-Learning in the Absence of Counter-Examples*. PhD Thesis], Delft University of Technology, 2001.

[17] M. Tihinen, M. Iivari, H. Ailisto, M. Komi, J. Kääriäinen and I. Peltomaa. An exploratory method to clarify business potential in the context of industrial internet–a case study. In *Working Conference on Virtual Enterprises*, pp. 469–478. Springer, 2016.

Received 20 February 2021