View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Universidad Carlos III de Madrid e-Archivo

Universidad Carlos III de Madrid

e-Archivo Institutional Repository

This is a postprint version of the following published document:

Martí L., Sanchez-Pi N., Molina J.M., García A.C.B. (2014) Combining Support Vector Machines and Segmentation Algorithms for Efficient Anomaly Detection: A Petroleum Industry Application. In: de la Puerta J. et al. (eds) International Joint Conference SOCO'14-CISIS'14-ICEUTE'14. Advances in Intelligent Systems and Computing, vol 299. Springer, Cham

DOI: 10.1007/978-3-319-07995-0_27

© Springer International Publishing Switzerland 2014

Combining Support Vector Machines and Segmentation Algorithms for Efficient Anomaly Detection: A Petroleum Industry Application

Luis Martí[1], Nayat Sanchez-Pi[2],
José Manuel Molina[3], and Ana Cristina Bicharra García[4]

[1] Dept. of Electrical Engineering,
Pontifícia Universidade Católica do Rio de Janeiro,
Rio de Janeiro (RJ), Brazil
lmarti@ele.puc-rio.br

[2] Instituto de Lógica, Filosofia e Teoria da Ciéncia (ILTC),
Niterói (RJ), Brazil
nayat@iltc.br

[3] Dept. of Informatics, Universidad Carlos III de Madrid,
Colmenarejo, Madrid, Spain
molina@ia.uc3m.es

[4] ADDLabs, Fluminense Federal University,
Niterói (RJ), Brazil
bicharra@ic.uff.br

Abstract. Anomaly detection is the problem of finding patterns in data that do not conform to expected behavior. Similarly, when patterns are numerically distant from the rest of sample, anomalies are indicated as outliers. Anomaly detection had recently attracted the attention of the research community for real-world applications. The petroleum industry is one of the application contexts where these problems are present. The correct detection of such types of unusual information empowers the decision maker with the capacity to act on the system in order to correctly avoid, correct, or react to the situations associated with them. In that sense, heavy extraction machines for pumping and generation operations like turbomachines are intensively monitored by hundreds of sensors each that send measurements with a high frequency for damage prevention. For dealing with this and with the lack of labeled data, in this paper we propose a combination of a fast and high quality segmentation algorithm with a one-class support vector machine approach for efficient anomaly detection in turbomachines. As result we perform empirical studies comparing our approach to other methods applied to benchmark problems and a real-life application related to oil platform turbomachinery anomaly detection.

Keywords: Anomaly detection, support vector machines, time series segmentation, oil industry application.

1 Introduction

The importance of anomaly detection is a consequence of the fact that anomalies in data translate to significant actionable information in a wide variety of application domains. The correct detection of such types of unusual information empowers the decision maker with the capacity to act on the system in order to correctly avoid, correct, or react to the situations associated with them. Anomaly detection has extensive use in a wide variety of applications such as fraud and intrusion detection [1], fault detection in safety critical systems [2], finance [3] or industrial systems (see [4,5] for surveys on this topic).

In the case of industrial anomaly detection, units suffer damage due to continuous usage and the normal wear and tear. Such damages need to be detected early to prevent further escalation and losses. The data in this domain is referred to as sensor data because it is recorded using different sensors and collected for analysis. The anomaly detection techniques in this domain monitor the performance of industrial components such as motors, turbines, oil flow in pipelines or other mechanical components and detect defects which might occur due to wear and tear or other unexpected circumstances. Data in this domain has a temporal aspect and time series analysis is also used in some works like: [6].

One example in industry applications is the detection of anomalies in turbomachinery installed in offshore petroleum extraction platforms. Recent history shows us how important a correct handling of this equipment is as failures in this industry have a dramatic economical, social and environmental impact.

Due to the lack of labeled data for training/validation of models in this paper we provide a solution for the detection of anomalies in turbomachinery, using a one-class SVM. This technique uses one class learning techniques for SVM [7] and learns a region that contains the training data instances (a boundary). Kernels, such as radial basis functions (RBF), can be used to learn complex regions. For each test instance, the basic technique determines if the test instance falls within the learnt region. If a test instance falls within the learnt region, it is declared as normal; else it is declared as anomalous. We combine this technique with a time series segmentation to prune noisy, unreliable and inconsistent data.

Therefore, the novelty of our approach is the combination of a fast and high quality segmentation algorithm with a one-class support vector machine approach for efficient anomaly detection. The remainder of this paper is organized as following. In the next section, we discuss some related work. Subsequently, we describe our proposal in detail. After that, we present a case study for offshore oil platform turbomachinery. This case study is used to compare our approach with alternatives methods of anomalies or outliers detection. Finally on section six, some conclusive remarks and directions for future work are presented.

2 Foundations

The preset work addresses the problem of anomaly detection by combining a one-class SVM classifier that has previously been used with success for anomaly detection with a novel and fast segmentation algorithm specially devised for this problem. In this section we present the theoretical pillars supporting the proposal.

2.1 Anomaly Detection

Fault and damage prevention is known as the problem of finding patterns in data that do not conform to an expected behavior [8]. Unexpected patterns are often referred as anomalies, outliers or faults, depending on the application domain. In broad terms, anomalies are patterns in data that do not conform to a well-defined normal behavior [8]. There are also extensive surveys of anomaly detection techniques.

Anomaly detection techniques have been proposed in literature, based on distribution, distance, density, clustering and classification. Their applications vary depending on the user, the problem domains, and even the dataset. In many cases the anomaly detection is related to outlier detection. In statistics, outliers are data instances that are deviate from given sample in which they occur. Grubbs in [9] defined *an outlying observation, or 'outlier', is one that appears to deviate markedly from other members of the sample in which it occurs.* Some of the anomaly detection techniques are:

- *Distribution-based approaches*: A given statistical distribution is used to model the data points. Then, points that deviate from the model are flagged as anomalies or outliers. These approaches are unsuitable for moderately high-dimensional datasets and require prior knowledge of the data distribution. They are also named as parametric and non-parametric statistical modeling [6].
- *Depth-based approaches*: This computes the different layers of convex hulls and flags objects in the outer layer as anomalies or outliers. It avoids the requirement of fitting a distribution to the data, but has a high computational complexity.
- *Clustering approaches*: Many clustering algorithms can detect anomalies or outliers as elements that do not belong —or are near— to any cluster.
- *Distance-based approaches*: Distance-based anomalies or outliers detection marks how distant is an element from a subset of the elements closest to it. It has been pointed out [10] that these methods cannot cope with datasets having both dense and sparse regions, an issue denominated *multi-density problem.*
- *Density-based approaches*: Density-based anomalies or outlier detection has been proposed to overcome the multi-density problem by means of the local outlier factor (LOF). LOF measures the degree of outlierness for each dataset element and depends on the local density of its neighborhood. This approach fails to deal correctly with another important issue: the *multi-granularity problem.* The local correlation integral (LOCI) method, and its outlier metric, the multi-granularity deviation factor (MDEF), were proposed with the purpose of correctly dealing with multi-density and multi-granularity [11].

- *Spectral decomposition*: Spectral decomposition is used to embed the data in lower dimensional subspace in which the data instances can be discriminated easily. Many techniques based on principal component analysis (PCA) have emerged [12]. Some of them decompose space to normal, anomaly and noise subspaces. The anomalies can be then detected in anomaly subspace [13].
- *Classification approaches*: In this case the problem is posed as the identification of which categories an observation belongs to. It operates in two phases: first it learns a model based on subset observations (training set) and second it infers a class for new observations (testing set) based on learnt model. This method operates under the assumption that a classifier distinguishes between normal and anomalous classes can be learnt in the given feature space. Based on the labels available for training phase, classification based anomaly detection techniques can be grouped into two broad categories: multi-class [14] and one-class anomaly detection techniques [15].

2.2 Time Series Segmentation

In the problem of finding frequent patterns, the primary purpose of time series segmentation is dimensionality reduction. For the anomalies detection problems in turbomachineries, it is essential to segment the dataset available in order to automatically discover the operational regime of the machine in the recent past. There is a vast work done in time series segmentation. Before start citing them, we state a segmentation definition and describe the available segmentation method classification.

A definition of a time series is a regular time series, where the amount of time between two consecutive pairs is constant [16].

Depending on the application, the goal of the segmentation is used to locate stable periods of time, to identify change points, or to simply compress the original time series into a more compact representation. Although in many real-life applications a lot of variables must be simultaneously tracked and monitored, most of the segmentation algorithms are used for the analysis of only one time-variant variable.

There is a vast literature about segmentation methods for different applications. Basically, there are mainly three categories of time series segmentation algorithms using dynamic programming. Firstly, sliding windows [17] top-down [18], and bottom-up [19] strategies. The sliding windows method is a purely implicit segmentation technique. It consists of a segment that is grown until it exceeds some error bound. This process is repeated with the next data point not included in the last created segment.

There are other novel methods for instance those using clustering for segmentation. The clustered segmentation problem is clearly related with the time series clustering problem [20] and there are also several definitions for time series [21]. One natural view of segmentation is the attempt to determine which components of a data set naturally "belong together".

3 Algorithm Proposal

As already hinted earlier in the paper our proposal combines a fast segmentation algorithm with a support vector machine one-class classifier. The segmentation algorithms takes care of identifying relatively homogeneous parts of the time series in order to focus the attention of the classifier to the most relevant portion of the time series. Therefore, parts of the time series that remain on the past can be safely disregarded.

3.1 Segmentation Algorithm

We devised a novel and fast algorithm for time series segmentation. Besides the obvious purpose of obtaining a segmentation method that produces low approximation errors another set of guidelines were observed while devising it. In particular we were interested in low computational impact and easy parameterization.

This yet another segmentation algorithm (YASA) is sketched in Fig. 1 in pseudocode form. It is best understood when presented in recursive form, as it goes by computing a linear regression with the time series passed as parameter. Segmentation procedure first checks if the current level of recursion is acceptable. After that it goes by fitting a linear regression to the time series data. If the regression passes the linearity statistical hypothesis test then the current time series is returned as a unique segment.

If the regression does not models correctly the data it means that it is necessary to partition the time series in at least two parts that should be further segmented. The last part of YASA is dedicated to this task. It locates the time instant where the regression had the larger error residuals.

3.2 One-Class Support Vector Machine

One-class classification based anomaly detection techniques assume that all training instances have only the same class label. Then, a machine learning algorithm is used to construct a discriminative boundary around the normal instances using a one-class classification algorithm. Any test instance that does not fall within the learned boundary is declared as anomalies. Support Vector Machines (SVMs) have been applied to anomaly detection in the one-class setting. One-class SVMs find a hyper-plane in feature space, which has maximal margin to the origin and a preset fraction of the training examples lay beyond it.

In this paper we have applied this approach combined with a evolutionary algorithm [22] for optimizing the maximal margin, as well as other SVM parameters, with respect to outlier detection accuracy.

4 Anomaly Detection in Offshore Oil Extraction Turbomachines

Equipment control automation that includes sensors for monitoring equipment behavior and remote controlled valves to act upon undesired events is nowadays

```
function SegmentData($\mathcal{S}^{(j)}_{t_{\max},t_0}$, $\rho_{\min}$, $l_{\max}$, $s_{\min}$, $l$)
    Parameters:
    ▷ $\mathcal{S}^{(j)}_{t_{\max},t_0}$, time series data of sensor $j$ corresponding to time interval $[t_0, t_{\max}]$.
    ▷ $\rho_{\min} \in [0, 1]$, minimum significance for statistical hypothesis test of linearity.
    ▷ $l_{\max} > 0$, maximum levels of recursive calls.
    ▷ $s_{\min} > 0$, minimum segment length.
    Returns:
    ▷ $\Phi := \{\phi_1, \ldots, \phi_m\}$, data segments.
    if $l = l_{\max}$ then
        return $\Phi = \left\{\mathcal{S}^{(j)}_{t_{\max},t_0}\right\}$
    Perform linear regression, $\{m, b\} \leftarrow$ LinearRegression($\mathcal{S}^{(j)}_{t_{\max},t_0}$).
    if LinearityTest($\mathcal{S}^{(j)}_{t_{\max},t_0}, m, b$) $> \rho_{\min}$ then
        return $\Phi = \left\{\mathcal{S}^{(j)}_{t_{\max},t_0}\right\}$.
    Calculate residual errors, $\{e_0, \ldots, e_{\max}\}$ = Residuals($\mathcal{S}^{(j)}_{t_{\max},t_0}, m, b$).
    $t_s \leftarrow t_0$.
    while $\max(\{e_0, \ldots, e_{\max}\}) > 0$ and $t_s \notin (t_0 + s_{\min}, t_{\max} - s_{\min})$ do
        Determine split point, $t_s = \arg\max_t \{e_t\}$.
    if $t_s \in (t_0 + s_{\min}, t_{\max} - s_{\min})$ then
        $\Phi_{\text{left}}$ = SegmentData($\mathcal{S}^{(j)}_{t_s,t_0}$, $\rho_{\min}$, $l_{\max}$, $s_{\min}$, $l + 1$).
        $\Phi_{\text{right}}$ = SegmentData($\mathcal{S}^{(j)}_{t_{\max},t_s}$, $\rho_{\min}$, $l_{\max}$, $s_{\min}$, $l + 1$).
        return $\Phi = \Phi_{\text{left}} \cup \Phi_{\text{right}}$.
    return $\Phi = \left\{\mathcal{S}^{(j)}_{t_{\max},t_0}\right\}$.
```

Fig. 1. Pseudocode of the proposed algorithm

a common scenario in the modern offshore oil platforms. Oil plant automation physically protects plant integrity. However, it acts reacting to anomalous conditions. Extracting information from the raw data generated by the sensors, is not a simple task when turbomachinery is involved.

Any devices that extract energy from or import energy to a continuously moving stream of fluid (liquid or gas) can be called a turbomachine. Elaborating, a turbomachine is a power or head generating machine which employs the dynamic action of a rotating element, the rotor; the action of the rotor changes the energy level of the continuously flowing fluid through the machine. Turbines, compressors and fans are all members of this family of machines. In contrast to Positive displacement machines especially of the reciprocating type which are low speed machines based on the mechanical and volumetric efficiency considerations, majority of turbomachines run at comparatively higher speeds without any mechanical problems and volumetric efficiency close to hundred per cent.

The approach described in Section 3 was prompted by the complexity and requirements of the task of early detection of behaviors that could potentially lead to machine or platform failures in the application context of interest.

In order to experimentally study and validate our approach we carried out an study involving a real-world test case. In this case in particular we deal with a dataset of measurements taken with a five minutes frequency obtained during the first half of year 2012 from 64 sensors connected to an operational turbomachine. An initial analysis of the data yields that there are different profiles or patterns that are shared by different sensors. This is somewhat expected as sensors with similar purposes or supervising similar physical properties should have similar readings characteristics.

There are at least three time series profiles in the dataset. On one hand, we have smooth homogeneous time series that are generally associated with slow-changing physical properties. Secondly, we found fast changing/unstable sensor readings that could be a result of sensor noise or unstable physical quantity. There is a third class of time series which exhibit a clear change in operating profile attributable to different usage regimes of the machine or the overall extraction/processing process.

In order to provide a valid ground for comparison we tested the method currently used by the platform operator, which is based on statistical confidence intervals [23], a one-class support vector machine-based classifier —as described earlier in this work— and our proposal. Problem data was transformed as to detect an anomaly based on consecutive sensor measurements in one hour.

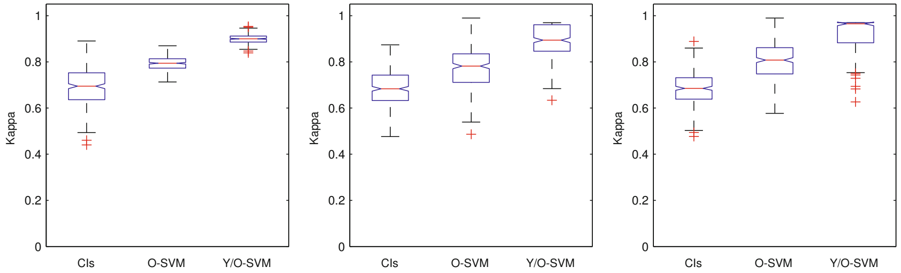
The approach in current use was not (and can not be) fully disclosed, as it is business sensitive information. However, in broad terms, for each sensor, this method receives a sample data chunk, which has been selected by an expert as a valid one. It filters out outlier elements and computes the confidence intervals at a predefined percent of the resulting dataset. A possible failure is detected when a given set of sensor measurements are consistently outside such interval.

All of these approaches can be said to be of an unsupervised learning nature, as they do not require to have labeled data. However, in order to evaluate the quality of the methods in anomaly detection it was necessary to prepare a test dataset that contain regular and anomalous data. We carried out that task by creating a test data set, which contained 20 anomaly instances extracted from each of the 64 time series and 20 regular or non-anomalous situations.

The need for comparing the performance of the algorithms when confronted with the different sensor data prompts the use of statistical tools in order to reach a valid judgement regarding the quality of the solutions, how different algorithms compare with each other and their computational resource requirements. Box plots [24] are one of such representations and have been repeatedly applied in our context. Although box plots allows a visual comparison of the results and, in principle, some conclusions could be deduced out of them.

Figure 2 shows the quality of the results in terms of the Kappa statistic [25] obtained from each algorithm in the form of box plots. We have grouped the results according to the class of sensor data for the sake of a more valuable presentation of results.

The statistical validity of the judgment of the results calls for the application of statistical hypothesis tests [26]. The McNemar test [27] is particularly suited for the assessment of classification problem results, like ones addressed here. This test is a normal approximation used on paired nominal data. It is applied to 2×2 contingency tables with a dichotomous trait, with matched pairs of subjects, to



(a) Errors for homogeneous series.

(b) Errors for multi-modal series.

(c) Errors for noisy series.

Fig. 2. Box plots of the Kappa statistic yielded by each class of dataset

Table 1. Results of the McNemar statistical hypothesis tests. Green cells (+) denote cases where the algorithm in the row statistically was better than the one in the row. Cells marked in red (−) are cases where the method in the column yielded statistically better results when compared to the method in the row. Finally, cells in blue (∼) denote cases where results from both methods were statistically indistinguishable.

	Y+OSVM	OSVM	CIs
Homogeneous series			
YASA + One-class SVM (Y+OSVM)	·	∼	+
One-class SVM (OSVM)		·	−
Confidence intervals (CIs)			·
Multi-modal series			
YASA + One-class SVM (Y+OSVM)	·	+	+
One-class SVM (OSVM)		·	∼
Confidence intervals (CIs)			·
Noisy series			
YASA + One-class SVM (Y+OSVM)	·	+	+
One-class SVM (OSVM)		·	∼
Confidence intervals (CIs)			·
All data			
YASA + One-class SVM (Y+OSVM)	·	+	+
One-class SVM (O-SVM)		·	+
Confidence intervals (CIs)			·

determine whether the row and column marginal frequencies are equal. In our case, we applied the test using to the confusion matrices performing pair-wise tests on the significance of the difference of the indicator values yielded by the executions of the algorithms. A significance level, α, of 0.05 was used for all tests.

Table 1 contains the results of the statistical analysis which confirm the judgements put forward before.

5 Final Remarks

In this work we combined a novel online segmentation method specially devised to deal with massive or big data problems with a one-class support vector machine in order to effectively detect anomalies. We have applied this algorithm to the segmentation sensor measurements of turbomachines used as part of offshore oil extraction and processing plants. In the problem under study, our approach was able to outperform the current approach used in the production system as well as the traditional formulation of a one-class SVM.

A computational system —whose essential formulation is the method described in this paper— is currently deployed by a major petroleum industry conglomerate of Brazil and is to be presented as a whole in a forthcoming paper.

Further work in this direction is called for and is currently being carried out. An important direction is the formal understanding of the computational complexity of the proposal. We also intend to extend the context of application to other big data application contexts.

Acknowledgement. This work was partially funded by CNPq BJT Project 407851/2012-7 and CNPq PVE Project 314017/2013-5.

References

1. Eskin, E., Arnold, A., Prerau, M., Portnoy, L., Stolfo, S.: A geometric framework for unsupervised anomaly detection. In: Applications of Data Mining in Computer Security, pp. 77–101. Springer (2002)
2. King, S., King, D., Astley, K., Tarassenko, L., Hayton, P., Utete, S.: The use of novelty detection techniques for monitoring high-integrity plant. In: Proceedings of the 2002 International Conference on Control Applications, vol. 1, pp. 221–226. IEEE (2002)
3. Borrajo, M.L., Baruque, B., Corchado, E., Bajo, J., Corchado, J.M.: Hybrid neural intelligent system to predict business failure in small-to-medium-size enterprises. International Journal of Neural Systems 21(4), 277–296 (2011)
4. Woźniak, M., Graña, M., Corchado, E.: A survey of multiple classifier systems as hybrid systems. Information Fusion 16, 3–17 (2014)
5. Calvo-Rolle, J.L., Corchado, E.: A bio-inspired knowledge system for improving combined cycle plant control tuning. Neurocomputing 126, 95–105 (2014)
6. Keogh, E., Lonardi, S., Chiu, B.: c.: Finding surprising patterns in a time series database in linear time and space. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 550–556. ACM (2002)
7. Ratsch, G., Mika, S., Scholkopf, B., Muller, K.: Constructing boosting algorithms from svms: an application to one-class classification. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(9), 1184–1199 (2002)

8. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: A survey. ACM Computing Surveys (CSUR) 41(3), 15 (2009)
9. Grubbs, F.E.: Procedures for detecting outlying observations in samples. Technometrics 11(1), 1–21 (1969)
10. Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J.: LOF: Identifying density-based local outliers. In: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, SIGMOD 2000, pp. 93–104. ACM, New York (2000)
11. Papadimitriou, S., Kitagawa, H., Gibbons, P., Faloutsos, C.: LOCI: Fast outlier detection using the local correlation integral. In: Proceedings 19th International Conference on Data Engineering (ICDE 2003), pp. 315–326. IEEE Press (2003)
12. Ringberg, H., Soule, A., Rexford, J., Diot, C.: Sensitivity of pca for traffic anomaly detection. In: ACM SIGMETRICS Performance Evaluation Review, vol. 35, pp. 109–120. ACM (2007)
13. Fujimaki, R., Yairi, T., Machida, K.: An approach to spacecraft anomaly detection problem using kernel feature space. In: Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, pp. 401–410. ACM (2005)
14. Barbara, D., Wu, N., Jajodia, S.: Detecting novel network intrusions using Bayes estimators. In: First SIAM Conference on Data Mining. SIAM (2001)
15. Roth, V.: Outlier detection with one-class kernel Fisher discriminants. In: Advances in Neural Information Processing Systems, vol. 17, pp. 1169–1176. MIT Press (2005)
16. Bouchard, D.: Automated time series segmentation for human motion analysis. Center for Human Modeling and Simulation, University of Pennsylvania (2006)
17. Bingham, E., Gionis, A., Haiminen, N., Hiisilä, H., Mannila, H., Terzi, E.: Segmentation and dimensionality reduction. In: SDM. SIAM (2006)
18. Lemire, D.: A better alternative to piecewise linear time series segmentation. In: SDM. SIAM (2007)
19. Hunter, J., McIntosh, N.: Knowledge-based event detection in complex time series data. In: Horn, W., Shahar, Y., Lindberg, G., Andreassen, S., Wyatt, J.C. (eds.) AIMDM 1999. LNCS (LNAI), vol. 1620, pp. 271–280. Springer, Heidelberg (1999)
20. Vlachos, M., Lin, J., Keogh, E., Gunopulos, D.: A wavelet-based anytime algorithm for k-means clustering of time series. In: Proc. Workshop on Clustering High Dimensionality Data and Its Applications. Citeseer (2003)
21. Bollobás, B., Das, G., Gunopulos, D., Mannila, H.: Time-series similarity problems and well-separated geometric sets. In: Proceedings of the Thirteenth Annual Symposium on Computational Geometry, pp. 454–456. ACM (1997)
22. Martí, L.: Scalable Multi-Objective Optimization. PhD thesis, Departmento de Informtica, Universidad Carlos III de Madrid, Colmenarejo, Spain (2011)
23. Neyman, J.: Outline of a theory of statistical estimation based on the classical theory of probability. Philosophical Transactions of the Royal Society A 236, 333–380 (1937)
24. Chambers, J., Cleveland, W., Kleiner, B., Tukey, P.: Graphical Methods for Data Analysis, Wadsworth, Belmont (1983)
25. Di Eugenio, B., Glass, M.: The Kappa statistic: A second look. Computational Linguistics 30(1), 95–101 (2004)
26. Salzberg, S.L.: On comparing classifiers: Pitfalls to avoid and a recommended approach. Data Mining and Knowledge Discovery 1(3), 317–328 (1997)
27. McNemar, Q.: Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 12(2), 153–157 (1947)