

Detection of Outliers in an Unsupervised Environment

M. Ashwini Kumari, M.S. Bhargavi and Sahana D. Gowda

Abstract Outliers are exceptions when compared with the rest of the data. Outliers do not have a clear distinction with respect to regular samples in the dataset. Analysis and knowledge extraction from data with outliers lead to ambiguity and confused conclusions. Therefore, there is a need for detection of outliers as a pre-processing stage for data mining. In a multidimensional perspective, outlier detection is a challenging issue as an object may deviate in one subspace and may appear perfectly regular in another subspace. In this paper, an ensemble meta-algorithm has been proposed to analyze and vote the samples for outlier identification in multidimensional subspaces. Cook's distance, a regression based model has been applied to detect the outliers voted by the ensemble meta-algorithm. Extensive experimentation on real datasets demonstrates the efficiency of the proposed system in detecting outliers.

Keywords Outlier detection · Outlier ensemble · Multidimensional subspace analysis · Cook's distance

1 Introduction

Data mining task aims to find the general patterns applicable to majority of the data, but due to occurrence of abnormal behaviour of the data, the valuable knowledge hidden behind may be suppressed and aid in ambiguous or confused conclusions.

M. Ashwini Kumari (✉) · M.S. Bhargavi · S.D. Gowda
Department of Computer Science and Engineering, BNM Institute of Technology,
Bangalore, India
e-mail: mk_ashwini@yahoo.com

M.S. Bhargavi
e-mail: ms.bhargavi@gmail.com

S.D. Gowda
e-mail: sahanagowda@rediffmail.com

Abnormal behaviour of data are outliers. In practise, however there is not always a clear distinction between outlier and regular data as data have different roles with respect to different attribute sets in a multidimensional space. Therefore, outlier detection in a multidimensional subspace is a challenging issue.

Detection of outliers in an unsupervised environment makes the problem more complex as the basic knowledge about the number of clusters and the behaviour would not be available. Thus, to ensure robustness and generality of outlier detection across variety of data, an outlier ensemble with diverse outlier factor analysis is required.

In this paper, an outlier detection method has been proposed by devising an ensemble meta-algorithm. The ensemble meta-algorithm is based on three diverse factors. Each factor analyses the data in different perspective in a sub-space. Multidimensional subspace is a permutable combination of attribute set. The number of sub-spaces depends on the number of the attributes considered for analysis. The three factors considered are: Distribution based factor, depth based factor and proximity based factor. Distribution based factor uses the Mahalanobis distance measure to model the data distribution and portray the occurrence of the samples to the centre of the distribution. Depth based factor utilizes the Mahalanobis depth function which gives the outward ordering of the data points from the deepest point among the data samples. Proximity based factor utilizes the k-nearest neighbour distances that shows the proximity of the samples in the local neighbourhood.

Based on the analysis of the dataset by every factor, each sample is scored by 1 or 0. Score 1 indicates the rise in voting level where as 0 indicates no change. The cumulative scores indicate the number of votes gained by every sample. This may not suffice in declaring the samples as outliers as it is just the cumulative rise in number. To assimilate the voted samples of subspace to declare outliers into multidimensional level, a regression based Cook's distance has been adapted to analyze the slope deviation of the regression due to change in membership occurrence of voted samples as outliers. If the slope deviation is within the specified threshold computed dynamically, the sample is not an outlier else it is considered as outlier.

The rest of this paper is organized as follows. Section 2 signifies the state-of-art, in Sect. 3, a detailed description of the proposed methodology is given. Section 4 depicts experimental analysis and results. This paper is concluded in Sect. 5.

2 State-of-Art

Detecting and eliminating outliers is of great significance to knowledge extraction. Various techniques have been proposed for outlier detection. Available techniques vary from single conceptual model [1–10] to multiple conceptual models [11–17]. Single conceptual models build detection strategies based on single outlier factor for analysis whereas multiple conceptual models build the detection strategies with a set of outlier factors.

In literature, outlier detection methods with single outlier factors can be classified as statistical [1–3, 7, 10] and non-statistical [4–6, 8, 9] methods. Statistical methods depend on statistical reasoning of the data distribution and prominent categories include distribution-based [1, 10, 18] and depth-based [2, 19, 20] approaches. Distribution-based approaches indicate those observations that are located relatively far from the center of the data distribution as outliers. Depth-based approaches are non-parametric and do not assume any underlying distribution for the data. They indicate that an observation is an outlier based on its “center-outward” ordering in the data. Non-statistical or proximity-based methods focus on detecting outliers based on compactness of the samples in the local neighborhood. Prominent approaches for non-statistical outlier detection include distance-based [4, 5], density-based [14, 15] and clustering-based [6, 9] approaches.

Single conceptual models for outlier detection make specific assumptions about the data to define what constitutes an outlier. They may not be effective if those assumptions do not hold for the given data. Multiple conceptual models, also called outlier ensembles have been proposed to overcome this problem [11–13, 16, 17, 21–25]. Outlier ensembles can be classified into those that involve multiple detection factors [12, 13, 17], those with multiple executions of a single outlier factor using different parametric values each time [14, 15] and those which perform single outlier factors in multiple feature subspaces [11, 22–24]. The ensembles that explore single factors in multiple subspaces are also called data-centered ensembles [26].

Data-centered outlier ensembles analyze the outliers from various perspectives. These methods select random subspaces [11, 23] or use statistical methods for selection of relevant subspaces [22, 24, 25] for outlier analysis. OutRank [21] is one technique which performs outlier detection in all subspaces. Majority of the work that use feature bagging, use single factor for analysis. There is very limited work in literature that uses a combination of model centered and data centered models [12, 13].

In any form of outlier ensemble, an important aspect is the combination and interpretation of results from different runs. Prominent approaches for combining the scores are model averaging [22], best fit [14], aggregation [11] and product of scores [12]. For final interpretation of outlier scores, user specified information such as number of outliers L [12] or threshold [13] is the usually adopted method. There are certain methods proposed to overcome such static cut-off. Gao [17] use calibration approaches to fit outlier scores provided by different detectors into probability values. The probability estimates are then used to select the appropriate threshold for declaring outliers using a Bayesian risk model. Papadimitriou et al. [15] introduce a new definition of density-based outliers. The outlier score of each data sample is used to compare against the normalized deviation of its neighborhood’s scores and standard-deviation is employed in the outlier detection. However, above techniques are limited to specific outlier ensembles.

It is evident from survey that for a robust detection of outliers, outlier ensembles have been widely used. The existing methods for outlier ensembles can be classified into data-centered methods and model-centered methods. Data-centered methods

use the concept of feature bagging and model-centered methods use different detection strategies in full data space. The methods that use both the concepts, concentrate on the selection of relevant subspaces for outlier detection. But these are considered weak guesses and the true subsets of attributes relevant for outlier analysis in an unsupervised environment may not be accurately identified [26]. Therefore, there is a need for an ensemble meta-algorithm with outlier factors at multidimensional subspace level.

3 Proposed Methodology

Multidimensional subspace analysis explores the data in different subspaces. All permutable combinations of attributes are considered to analyze the abnormal behavior of outliers. The abnormality of a sample is due to its occurrence, neighborhood relationship and nearness to the deepest point in the data clutter. To measure all these factors, an ensemble meta-algorithm is devised with three factors such as distribution based outlier factor to measure the occurrence based on distribution, depth based outlier factor to measure the nearness and proximity based outlier factor to measure the nearest neighborhood relationship.

3.1 Distribution Based Outlier Factor

Distribution based factor uses Mahalanobis distance measure to model the data distribution and portray the occurrence of the samples to the center of the distribution. The samples that are highly deviating from the center of the data distribution are outliers. For multidimensional data, the distance of sample from the center or the mean of the data can be obtained by using multivariate distance measures. Mahalanobis distance measure is a multivariate distance measure which has been used for the distribution based factor in the proposed ensemble meta-algorithm. Mahalanobis distance calculates the distance of a sample from its mean considering the variance and the covariance of the features.

Mahalanobis distance [1] is defined as follows. Consider a d-dimensional multivariate dataset with n samples, $x_1, \dots, x_n \in R^d$. The Mahalanobis distance for each multivariate sample $x_i = (x_{i1}, \dots, x_{id})^T$ is given by

$$MD_i = \sqrt{(x_i - \mu)^T C^{-1} (x_i - \mu)} \quad (1)$$

where, $\mu = (\mu_1, \dots, \mu_d)^T$ denotes the overall sample mean and C = variance-covariance matrix given by,

$$C = \begin{bmatrix} c11 \dots c1d \\ \vdots \\ cd1 \dots cdd \end{bmatrix} \tag{2}$$

and each term is a covariance between two features A and B given by

$$CAB = \frac{\sum_{i=1}^n (xiA - \mu A)(xiB - \mu B)}{(n - 1)} \tag{3}$$

Distances obtained follow the Chi-squared distribution with d-degrees of freedom. Cut-off value for the distribution is obtained from the Chi-squared distribution table at 97.5th percentile. Distance values above the cutoff are selected as outliers.

3.2 Depth Based Outlier Factor

The depth factor in the ensemble provides an outward ordering of the data points or samples from the deepest point in the data. This ordering of the points does not depend on the distribution of the data. Mahalanobis depth function [2] is used and is defined as follows.

Consider a d-dimensional dataset with n samples $x_1, \dots, x_n \in R^d$. The depth function for each multivariate sample $xi = (x1i, \dots, xidi)^T$ is given by

$$MDEi = 1/(1 + SMDi) \tag{4}$$

where, SMD_i = squared Mahalanobis distance of the sample given by

$$SMDi = (xi - \mu)^T C^{-1}(xi - \mu) \tag{5}$$

where, μ and C are the overall sample mean and the variance-covariance matrix respectively.

The depth values (distances) obtained shows bigger values for samples nearer to the deepest point in the data clutter and smaller values for the samples far away. Samples with smaller values are considered outliers. Through trial and learning process, a threshold of 0.15 is set to identify the outliers.

3.3 Proximity Based Outlier Factor

The proximity based outlier function is the distance between a sample and its k-nearest neighbors [5]. For every sample, the average distance is computed to the corresponding k-nearest neighbors considered.

For a d-dimensional multivariate dataset with n samples $x_1, \dots, x_n \in \mathbb{R}^d$, the distance function for a sample $x_i = (x_{i1}, \dots, x_{id})^T$ is given by,

$$AvgDist_i = \sum_{j=1}^k Euclidean_Dist_i, k/k \quad (6)$$

where, k = number of neighbors,

$Euclidean_Dist_{i,k}$ = Euclidean distance between sample x_i and its j th neighbor is given by,

$$ED_{ij} = \sqrt{(x_{ij} - x_{j1})^2 + \dots + (x_{id} - x_{jd})^2} \quad (7)$$

The distance of the samples obtained is sorted and the samples with bigger values are considered outliers. The threshold to declare outliers is based on the average of the number of outlier samples obtained in the above two factors.

3.4 Interpretation of Outlier Scores

The outliers identified in each factor vary depending on the metric utilized. Due to the abnormal behavior, a sample may be voted as outlier by a factor in a subspace and may not be true in another factor in the same space. If there are n dimensional attributes, then the total number of subspaces considered are $2^n - n - 1$, where n is the number of single feature subspaces and 1 is an empty subspace. Single feature subspaces are eliminated which drastically reduces the computational complexity by n without eliminating any permutable combination of attribute sets.

To assimilate the analyses of every subspace of every factor, cumulative voting record is generated. In the cumulative record, there is a rise in value by 1 if any sample is voted as outlier. The cumulative record is the additive value of every sample in the dataset voted outlier in a subspace by a factor. There is a rise in value by 1 if the sample is voted as outlier else the numeric value remains unchanged for non-outliers. The samples which are never voted always have the value zero.

The comprehensive outlier scores obtained for the data samples do not provide a direct indication of whether the samples are outliers in the cumulative record. Outlier scores that not only have high values but also have a large deviation from the rest of the scores are to be identified dynamically. To analyze the voted samples on a single plot, the cumulative records of all the samples are assimilated by a regression based model called Cook's distance [27]. Based on the observation, when the outlier scores are plotted, a majority of the outlier scores (representing the weak outliers) form a linear band and can be represented by a linear model. Those samples with high scores affect the linearity of the data. To identify such scores, regression analysis using Cook's distance has been adapted.

For a regression Y on the outlier scores (y_1, y_2, \dots, y_n) , the Cook's distance for a point y_i is given by,

$$D_i = \sum_{j=1}^n (\hat{y}_j - \hat{y}_j(i))^2 / (k + 1)s^2 \tag{8}$$

where,

\hat{y}_j is the predicted value of the j th observation.
 $\hat{y}_j(i)$ is the predicted value of the j th observation with the i th point removed and s^2 is the mean squared error or the variance from the fit based on all the observations (variance is the squared difference between the predicted value \hat{y} and observed value y).

The predicted value of y from a regression i.e. \hat{y} is given by

$$\hat{y} = Hy \tag{9}$$

where, H is called the Hat matrix [28] given by,

$$H = X(X'X)^{-1}X' \tag{10}$$

Here, X is the predictor variable for the response variable Y (vector of outlier scores).

The point i is treated as an outlier if

$$D_i \geq 4 / (n - (k + 1)) \tag{11}$$

where, k is the number of predictor variables which is equal to 1 as the outlier score is a single vector.

4 Experimental Analysis and Results

In this section, an experimental study is presented to evaluate the efficacy of the proposed method in detecting true outliers. Experiments are conducted on four real world datasets from the UCI Machine Repository [29]. The real world datasets used for experimental analysis are Wisconsin breast cancer dataset, New Thyroid dataset 1 (with class 2 as rare class), New Thyroid dataset 2 (with class 3 as rare class) and Pima Indian diabetes dataset. The datasets selected are binary class sets with the larger class being the normal data and the smaller class being the rare class data. Datasets with multiple classes are converted to binary class sets.

In the following section, a brief overview of all the datasets used for experimental analysis is provided along with the results achieved using the proposed method. The details of the datasets are as follows:

Wisconsin breast cancer dataset: The original Wisconsin breast cancer data consists of 699 records with nine attributes each. The records are labeled either benign (458 records, 65.5 %) or malignant (241 records, 34.5 %). For experimental purpose, 444 benign records and 39 malignant records are chosen. This leads to 92 % normal records and 8 % abnormal or rare class records.

New Thyroid dataset 1: The New thyroid dataset gives information about the thyroid disease. The task is to detect whether a given patient is normal or suffers from hyperthyroidism or hypothyroidism. The original dataset consists of 215 records out of which 150 records are normal, 35 records denote hyperthyroidism and 30 records denote hypothyroidism. For experimental purpose, 150 normal records and 15 records from class 2 are considered to form 90 % normal records and 10 % abnormal or rare class records.

New Thyroid dataset 2: The New thyroid dataset is used with another combination of binary class set. 150 normal records and 15 records from class 3 are considered to form 90 % normal records and 10 % abnormal or rare class records.

Pima Indian diabetes dataset: The original Pima Indian Diabetes dataset consists of 500 ‘tested negative’ records and 268 ‘tested positive’ records. For experimental purpose, 150 normal records and 15 records from class 2 are considered to form 90 % normal records and 10 % abnormal or rare class records.

Precision and Recall are used for measuring the quality of outlier detection by the proposed method. Precision, also called the Positive predictive rate, is the percentage of the reported outliers, which turn to be true outliers. Recall, also called True Positive Rate (TPR), is the percentage of the true outliers that have been reported as outliers at a given threshold. False Positive Rate (FPR) is also shown, which is the percentage of falsely reported outliers out of the true inliers.

From Table 1, it is evident that the recall values for the majority of the datasets are high indicating that the proposed system can detect outliers effectively (high percentage of outliers are detected). High values of precision and low values of FPR indicate that true outliers are more likely to be detected as compared to false outliers. Hence the quality of outlier detection is high for the proposed system.

The overall expected performance of the proposed method is evaluated using the well-established Receiver Operating Characteristics (ROC) curve. ROC curve is obtained using the plot of TPR against the FPR rate. The ROC curve characterizes the trade-off between the TPR and FPR values. It is preferred that the outlier detection method has high TPR and low FPR values. This will have an ROC curve

Table 1 Precision, recall and FPR metrics for the proposed meta-algorithm

Dataset	Precision (%)	Recall (TPR) (%)	FPR (%)
Wisconsin breast cancer data	74	94.87	2.92
New thyroid data 1	91.67	73.3	0.67
New thyroid data 2	92.3	80	0.67
Pima Indian Diabetes data	27.3	24	6.4

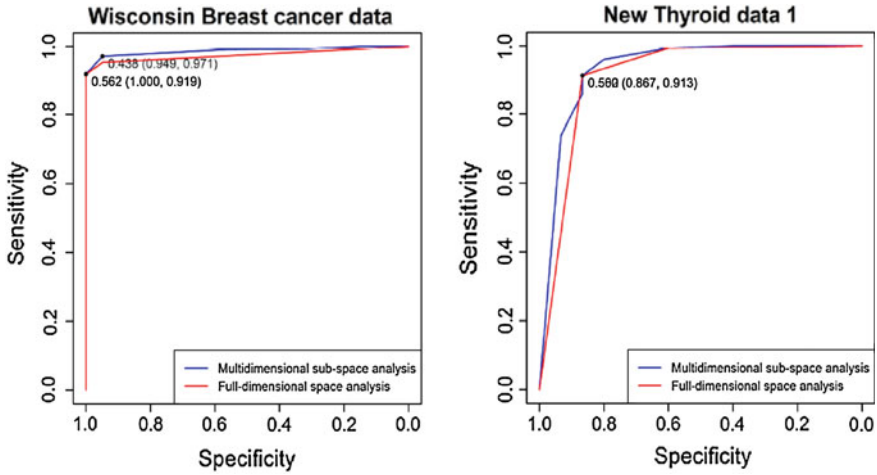


Fig. 1 ROC curves for wisconsin breast cancer and new thyroid 1 datasets using the proposed meta-algorithm as against full-space outlier detection

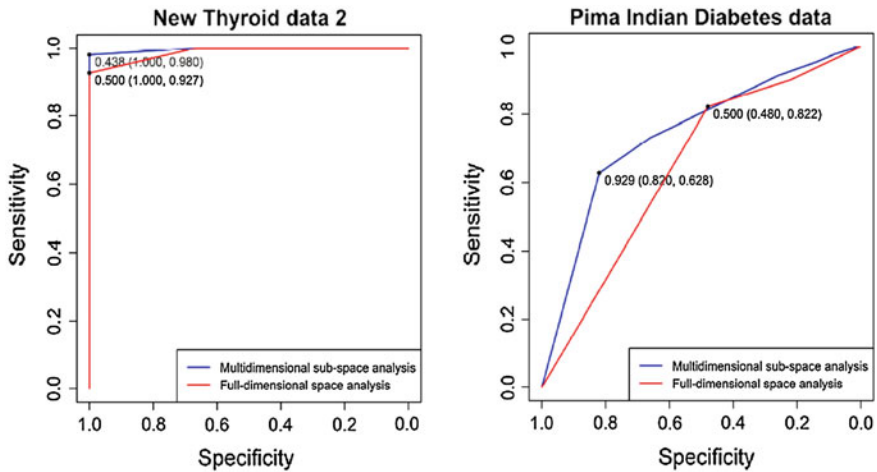


Fig. 2 ROC curves for new thyroid 2 and pima Indian diabetes datasets using the proposed meta-algorithm as against full-space outlier detection

that is closer to the upper left corner of the graph indicating high TPR and low FPR. The ROC curves for the datasets are shown in the Figs. 1 and 2. Results obtained for the proposed model with multidimensional subspaces are compared with full dimensional space with the same factors. The ROC curves in the plots demonstrate the improvement in performance due to multi-dimensional subspace analysis.

Table 2 Results on mentioned datasets

Dataset	AUC (multidimensional subspace)	AUC (full dimensional space)
Wisconsin breast cancer data	0.98	0.97
New thyroid data 1	0.94	0.91
New thyroid data 2	0.99	0.98
Pima Indian diabetes data	0.74	0.65

It is evident from the ROC curves that that overall expected performance of the proposed system is high compared to full-dimensional space analysis.

The area under the ROC curve (ROC AUC) is a summary statistic used to describe the overall expected performance numerically. AUC values closer to one indicate better outlier detection. The ROC AUC values for the datasets for the proposed system and ensemble with full-dimensional space analysis are given in Table 2. It can be seen that the ROC AUC values for the proposed multi-dimensional subspace analysis performs better than full-space analysis.

5 Conclusion

Outlier detection is a preprocessing stage for knowledge extraction. Analysis of data which is free of outliers reduces ambiguity and fuzziness in the conclusion. The proposed ensemble meta-algorithm with diverse factors, aims at identifying outliers in unsupervised environment. Compared to the existing outlier detection methods, the proposed model identifies outliers more accurately. Experimental analysis with real world datasets demonstrates the efficacy of the proposed method in detecting true outliers. An enhancement to the proposed model would be to dynamically set the threshold for outlier identification in every factor. Presently, the threshold is predefined by the knowledge of literature and training.

References

1. Mahalanobis, P.C.: On the generalized distance in statistics. Proc. Natl. Inst. Sci. India **12**, 49–55 (1936)
2. Liu, R.Y., Singh, K.: A quality index based on data depth and multivariate rank tests. J. Am. Stat. Assoc. **88**, 252–260 (1993)
3. Arning, A., Agrawal, R., Raghavan, P.: A linear method for deviation detection in large databases. In: Proceedings of Data Mining and Knowledge Discovery, pp. 164–169. Portland, Oregon (1996)
4. Knorr, E., Ng. R.: Algorithms for mining distance-based outliers in large datasets. In: Proceedings of 24th International Conference on Very Large Data Bases (VLDB), pp. 392–403, 24–27 (1998)

5. Ramaswamy S., Rastogi, R., Shim, K.: Efficient algorithms for mining outliers from large data sets. In: Proceedings of the ACM SIGMOD International Conference on Management of Data. Dallas, TX (2000)
6. Jiang, M.F., Tseng, S.S., Su, C.M.: Two-phase clustering process for outliers detection. *Pattern Recogn. Lett.* **22**(6), 691–700 (2001)
7. Filzmoser, P.: A multivariate outlier detection method (2004)
8. Hawkins, S., et al.: Outlier detection using replicator neural networks. *Data Warehousing and Knowledge Discovery*, pp. 170–180. Springer, Berlin (2002)
9. He, Z., Xu, X., Deng, S.: Discovering cluster-based local outliers. *Pattern Recogn. Lett.* **24**(9), 1641–1650 (2003)
10. Rousseeuw, P.J., Van Zomeren, B.C.: Unmasking multivariate outliers and leverage points. *J. Am. Stat. Assoc.* **85**(411), 633–639 (1990)
11. Lazarevic, A., Kumar, V.: Feature bagging for outlier detection. In: Proceedings of KDD. pp. 157–166 (2005)
12. He, Z., Deng, S., Xu, X.: A unified subspace outlier ensemble framework for outlier detection. In: Fan, W., Wu, Z., Yang, J. (eds.). LNCS, vol. 3739 pp. 632–637. Springer, Heidelberg, (2005)
13. Nguyen, H.V., Ang, H.H., Gopalkrishnan V: Mining outliers with ensemble of heterogeneous detectors on random subspaces. *Database Systems for Advanced Applications. Springer Berlin Heidelberg* (2010)
14. Breunig, M., Kriegel, H.P., Ng, R., Sander, J.: LOF: Identifying Density-based Local Outliers. *ACM SIGMOD Conference* (2000)
15. Papadimitriou, S., et al.: Loci: Fast outlier detection using the local correlation integral. In: Proceedings of 19th International Conference on Data Engineering. IEEE (2003)
16. Zimek, A., et al.: Subsampling for efficient and effective unsupervised outlier detection ensembles. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM (2013)
17. Gao, J., Tan, P.N.: Converting output scores from outlier detection algorithms into probability estimates. In: Sixth International Conference on Data Mining. IEEE (2006)
18. Rousseeuw, P.J., Leroy, A.M.: *Robust Regression and Outlier Detection*. Wiley (1987)
19. Tukey, J.: *Exploratory Data Analysis*. Addison-Wesley (1977)
20. Ruts, I., Rousseeuw, P.J.: Computing depth contours of bivariate point clouds. *Comput. Stat. Data Anal.* **23**, 153–168 (1996)
21. Müller, E., et al.: Outlier Ranking via Subspace Analysis in Multiple Views of the Data. *ICDM* (2012)
22. Keller, F., Muller, E., Bohm, K.: HiCS: high contrast subspaces for density-based outlier ranking. In: IEEE 28th International Conference on Data Engineering (ICDE) (2012)
23. Foss, A., Zaïane, O.R.: Class separation through variance: a new application of outlier detection. *Knowl. Inf. Syst.* **29**(3), 565–596 (2011)
24. Nguyen, H.V., et al.: CMI: an information-theoretic contrast measure for enhancing subspace cluster and outlier detection. *SDM* (2013)
25. Aggarwal, C.C., Yu, P.S.: Outlier detection for high dimensional data. *ACM Sigmod Record*, vol. 30. No. 2, ACM, New York (2001)
26. Aggarwal, C.C.: Outlier ensembles. Position paper. *ACM SIGKDD Explorations Newsletter*. pp. 49–58, (2013)
27. Cook, R.: Detection of influential observations in linear regression. *Technometrics* **19**, 15–18 (1977)
28. Hoaglin, D., Welsch, R.: The hat matrix in regression and anova. *Am. Stat.* **32**, 17–22 (1978)
29. Bache, K., Lichman, M.: UCI machine learning repository. <http://archive.ics.uci.edu/ml>