Preprints, 10th IFAC International Symposium on
Advanced Control of Chemical Processes
Shenyang, Liaoning, China, July 25-27, 2018

A Kernel Connectivity-based Outlier Factor Algorithm for Rare Data Detection in a Baking Process ⋆

Yanxia Wang * **Kang Li** ** **Shaojun Gan** ***

* *School of Electronics, Electrical Engineering and Computer Science, Queens University Belfast, Belfast, BT9 5AH, U.K. (e-mail: yanxia.wang@qub.ac.uk).*
** *School of Electronics, Electrical Engineering and Computer Science, Queens University Belfast, Belfast, BT9 5AH, U.K. (e-mail: K.Li@qub.ac.uk)*
*** *School of Electronics, Electrical Engineering and Computer Science, Queens University Belfast, Belfast, BT9 5AH, U.K. (e-mail:S.Gan@qub.ac.uk)*

Abstract: Due to strict legislation on greenhouse gas emission reduction, energy intensive industries include the bakery industry are all under pressure to improve the energy efficiency in the manufacturing processes. In this paper, an energy monitoring system developed through the Point Energy Technology from the research group is first introduced for the data collection in a local bakery company. The outliers in the collected data may include valuable information about the status of machines, however, they also affect the data quality and the accuracy of the consequent data analysis. This paper discusses two algorithms for outlier detection, connectivity-based outlier factor (COF) and local outlier factor (LOF). For COF, the concept of connectivity-based outlier facto is adopted to identify whether an object is an outlier. For LOF, the local outlier factor based on a notion of local density represents the level of an object being an outlier. Experiments are conducted on the dataset from the oven in a production line to evaluate the effectiveness of three kernel functions, namely the Gaussian kernel, the Laplacian kernel and polynomial kernel. The experimental results show that the Gaussian-COF and the Laplacian-COF are more effective on valid oven data detection, which is significant for the further research work on energy management in the bakery company.

Keywords: Energy monitoring system, COF, LOF, kernel functions.

1. INTRODUCTION

The UK government has committed to reducing its greenhouse gas emissions (GHG) by 80% by 2050 compared to the 1990 level. Consuming 16% of total energy per year, the manufacturing industries are putting energy consumption optimization as a priority to meet the GHG reduction target[1]. The bakery industry, which produces fresh and frozen bread, cakes and other pastries to meed people's daily dietary demand, consumes a lot of energy from gas and electricity. It is therefore of significant importance to improve the energy efficiency in the baking processes. Modern bakeries are often equipped with automatic production lines [2-3]. Most bakery products have similar core manufacturing procedure with flour, water, and yeast. Minor ingredients such as fruits and nuts are used to increase the diversity and abundance of bread. Fig. 1 illustrates a generic production process for bread manufacturing. There are several main processes: mixing, dividing, proofing, baking, cooling, and slicing/packaging. The first stage is to incorporate the flour, water, and other ingredients in a big mixer, then kneading the dough by a motor. While the dough is prepared properly, it is then sliced to the expected size. Then, the divided pieces are sent to a prover, resting for a period of time before being sent into the oven. The length of proofing varies by size and species of dough. For baking, doughs are sent into the oven by a conveyor belt. When the bread is baked, they would be removed automatically from pans by depanner. The bread needs time to cool so that the moisture and carbon dioxide inside will dissipate. The last step is to slice the loaves with a machine and package them for convenience. To improve the energy efficiency, the energy consumption of the manufacturing processes needs to be known correctly and precisely. However, there are always outliers in a dataset which we need to investigate and check whether these outliers are caused by production error, measurement error or data recording error. In addition, the outliers may affect data quality and thus the quality of the follow-up data analysis procedure. Therefore, valid data detection about the outliers is an important stage in energy management of the bakery industry.

As described in [4], an outlier is an object behaving differently from the expected ones. There are a number of

⋆ The work presented in this paper is funded by EPSRC under grant EP/P004636/1.

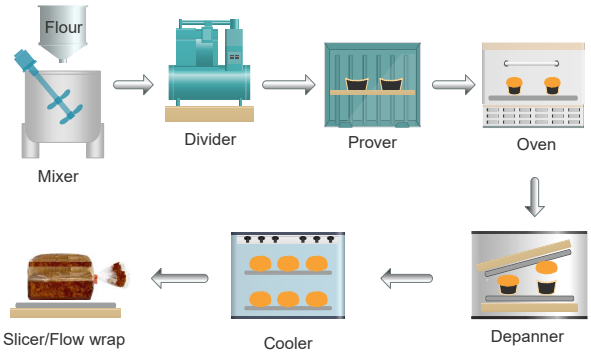

Fig. 1. Flow chart of bread production process

approaches for detecting outliers to acquire the embedded information and minimize their influence on the following statistic process for data analysis[5-6]. A distance-based method identifies outliers when their given radius area cannot include enough neighbors [7-8]. The drawback of this method is that it only considers the distance to the neighbors and ignores the information of closer objects [9]. A density-based approach provides a density measurement to identify that if a data point is an outlier or not. This algorithm can work well on the data set with unbalanced density regions. A connectivity-based scheme is proposed in [10-11], which considers an object and other connected objects with connectivity-based outlier factor (COF). The COF algorithm could not discern the difference between objects in small clusters which include fewer objects. The approaches based on density could handle the data set with different density areas, where the local outlier factor (LOF) is used to identify the degree of an object being an outlier [12-13]. It could not perform well when outliers are in the areas with various distributed densities [14]. Kernel functions map the initial data into a high-dimensional feature space, which could reflect the difference between objects better [15].

This paper firstly introduces the point energy monitoring system developed from the research team (www.pointenergy.org), is used in different industrial partners, including in a local bakery company to collect the voltage, current, power, power factor and frequency data from the baking process, one of the core production phases. Then we discuss data detection for the baking process with two algorithms, kernel connectivity-based outlier factor algorithm and kernel local outlier factor algorithm. Three kernel functions, namely the Gaussian function, the Laplacian function and Polynomial function, are used to assess the performance of the algorithm. The rest of this paper is organized as follows. The related work is introduced in section 2. In section 3, the experimental setup about the dataset and experimental procedure is described in detail, while section 4 presents the results and discussions. Finally, section 5 concludes this paper.

2. RELATED WORK

2.1 Point energy technology for energy monitoring

By working with a local bakery company which is eager to know how much energy they used daily and more specifically, how much the energy is consumed by each production line or even for each manufacturing process. Thus, an energy monitoring system, shown in Fig. 2, has been developed through the Point Energy Technology initiated from the research group (www.pointenergy.org). The system mainly contains two parts, energy data acquisition part and energy data analysis part. The energy data acquisition part is designed to install on site for obtaining the energy usage details along the whole production line at a component level. For energy consumed from the electricity, current transformers and intelligent power meters are deployed to measure the current, voltage of the production process. The active and the reactive power are also calculated from the acquired signals. The power factor is regarded as the indicator of the energy efficiency in the industry which is directly related to the electricity tariff, and low power factor not only leads to higher electricity tariff but also a potential large penalty. All these energy data are directly sent to a cloud server using raspberry pi boards which are single-board computers. The cloud server bridges the on-site data acquisition and remote data analysis, remote servers could gain access to the energy data stored to the cloud server by the MQTT protocol. Further, a WEB server is developed to display the real-time energy consumption and a data server is built to store the energy data locally for the preparation of more detailed data analysis.

2.2 Kernel connectivity-based outlier factor algorithm

In this section, the kernel connectivity-based outlier factor (COF) algorithm will be proposed. The core idea of this algorithm is to record each object the degree of being an outlier, which is called the connectivity-based outlier factor. The kernel COF algorithm can be formulated by the following steps:

(1) Map the initial data into a new feature using a kernel function, and the following steps are conducted in the new feature space.

(2) For each object x, find its k nearest neighbours. The set with point x and those neighbours is named as $N_k(x)$.

(3) Define a data set based on the nearest trail (SBN) from data point x, such that for all $1 \leq i \leq k-1$, x_{i+1} is the nearest neighbour point of set $\{x_1, \cdots, x_i\}$ in set $\{x_{i+1}, \cdots, x_k\}$.

(4) Let $e = \{e_1, \cdots e_k\}$, which is a sequence of edge points relating to the SBN path, that constitutes the consecutive nearest neighbours from point x in set $N_k(x)$. Each e_i is an edge point and $dist(e_i)$ means the distance between sets comprising an edge.

(5) Calculate the average chaining distance from x to $N_k(x) - \{x\}$, denoted by $dist_{N_k(x)}$ and defined as:

$$dist(x) = \sum_{i=1}^{k} \frac{2(k+1-i)}{k(k+1)} dist(e_i) \qquad (1)$$

$dist_{N_k(x)}$ can be viewed as the weighted distance in the cost description for the SBN path from point x.

(6) Compute the connectivity-based outlier factor (COF) at the data point x by its $k-$th neighbour using the following equation:

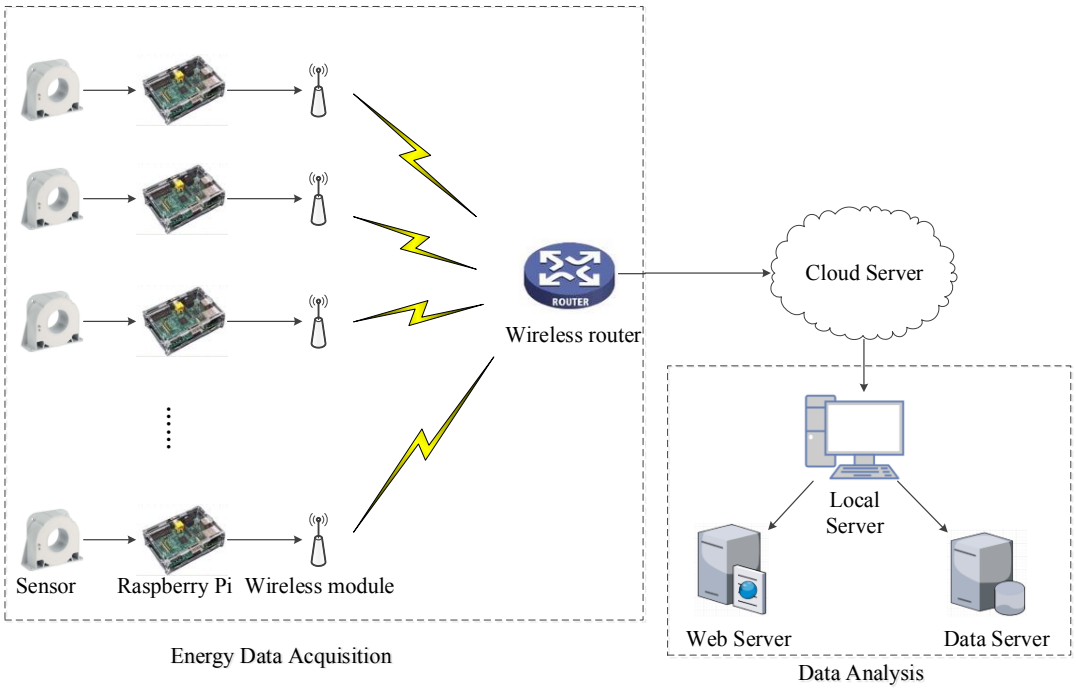

Fig. 2. Energy monitoring system

$$COF(x) = \frac{dist(x)}{\frac{1}{k}\sum_{o \in N_k(x)} dist(o)} \tag{2}$$

The COF of object x is the ratio of the average distance from x to $N_k(x)$ and the average distance of its neighbour records. It is easily inferred that the chance of an object being an outlier increases with the COF increases.

2.3 Kernel local outlier factor algorithm

In this section, the kernel local outlier factor (LOF) algorithm is provided. For this algorithm, the LOF of an object p is the average of the ratio of local reach-ability of p and those of p's k-neasrest neighbours. The kernel LOF algorithm could be described as folloes:

(1) Map the data into a feature space of higher dimensions.

(2) For each object, calculate all the distances between the point p and its k-th nearesrt neighbor $dist_k(p)$.

(3) Search all the objects in k-distance neighborhood of the object p, $N_k(p)$:

$$N_k(p) = \{p' | dist(p, p') dist_k(p)\} \tag{3}$$

(4) Calculate the local reach-ability density of the object p:

$$\begin{cases} Ird_k(p) = \dfrac{||N_k(p)||}{\sum_{p' \in N_k(p)} reachdist_k(p' \leftarrow p)} \\ reachdist_k(p' \leftarrow p) = max\{dist_k(p)), dist(p, p')\} \end{cases} \tag{4}$$

where $||N_k(p)||$ means the number of the objects in $N_k(p)$.

(5) For every object, calculate the local outlier factor (LOF) of object p.

$$LOF(p) = \frac{\sum_{p' \in N_k(p)} \frac{Ird_k(p')}{Ird_k(p)}}{||N_k(p)||} \tag{5}$$

(6) Sort the $LOF(p)$ for all objects, and the level of the object being an outlier becomes bigger as the value of LOF becomes bigger.

In this paper, three kernel functions are used in kernel COF and kernel LOF algorithms, the Gaussian kernel, polynomial kernel and Laplacian kernel.

The Gaussian function is a radial basis kernel, where α represents the width parameter.

$$K(x, y) = \exp(-||x - y||^2 / \alpha^2), \alpha > 0 \tag{6}$$

The polynomial function is a nonstationary kernel, where d is the order of polynomial.

$$K(x, y) = (x \cdot y + 1)^d, d > 0 \tag{7}$$

The Laplacian function is a radial basis kernel as well, with β being the width parameter.

$$K(x, y) = \exp(-\beta||x - y||), \beta > 0 \tag{8}$$

2.4 Evaluation criteria

In order to assess the performance of the two kernel algorithms, the precision, the recall and the rank power are used [16-17]. It is assumed that a data set $D = D_0 + D_n$, where D_0 denotes the set of all outliers and D_n represents the set of non-outliers. D_m is a dataset with outliers among the objects ranked in the top m positions returned by data detection algorithm, where $m \geq 1$. Let $|D_m|$ be the number

Table 1. Class distribution of oven data

	Case	Class label	Percentage of objects
Oven	Common	good	96.85
	Rare	bad	3.15

of data objects in $|D_m|$, and $|D_0|$ be the number of data objects in set $|D_0|$.

Precision is to represent the percentage of outliers among the top m ranked samples returned by the algorithm and defined as:

$$Precision = \frac{|O_m|}{m} \tag{9}$$

Recall is the ratio between detected outliers and all outliers in the data set, which can be defined as:

$$Recall = \frac{|O_m|}{|D_0|} \tag{10}$$

To calculate the position of the detected outlier, a rank power is introduced. It is assumed that m samples with position 1 to position m are returned by the data detection method, and there are Nr outliers in these m objects. For $1 < i < N_r$, let P_i denote the position of $i-$th outlier, then rank power being defined as:

$$RankPower = \frac{N_r(N_r+1)}{2\sum_{i=1}^{N_r} P_i} \tag{11}$$

It can be easily inferred that the rank power value is between 0 and 1, with 1 representing the best and 0 being the worst performance. Therefore, the efficiency of data detection algorithm can be judged by the above three variables. The performance could be better as the values of precision and recall become larger. For the same precision and recall, rank power should be used to judge the efficiency, and a larger value means a better efficiency.

3. EXPERIMENTAL SETUP

3.1 Data set

Since a large portion of electricity, about 30-35%, is used for the oven to bake the bread in the bakery company. In this section, we select the initial energy usage of the baking process from 00:00 to 24:00 on 02/02/2017. The following features are monitored at a 5-minute interval across all three phases: voltage, current, power, power factor, and frequency. The experiments are performed on oven data set, which has 286 samples with 11 attributes, two classes of good and bad. According to the methods used in [18], bad data samples are randomly generated to acquire an unbalanced distribution. As shown in table 1, the oven data set has 277 objects labeled as good and 9 objects labeled as bad.

3.2 Experimental procedure

To compare the effectiveness of COF and LOF algorithms, The experiment is conducted for the oven data set. Following the same value in the reference [19], k equals to 5% of the number of all objects in the data set. Let Nr be the number of rare objects detected. Moreover, m represents the number of top-ranked objects returned by the method. For these two algorithms, the parameters of the Gaussian, polynomial, and Laplacian kernel functions are denoted by α,d, and β respectively. For Gaussian function, the parameter is selected in the range of [0.1, 3] with a interval of 0.1; for Polynomial function, the scope of d is [1,30]; and β is defined from 0.01 to 0.05 with a interval of 0.005 for Laplacian function. The experiments are implemented by MATLAB 2017, and the computing environment is Windows 10 education, version 1703 for ×64-based system.

4. RESULTS AND DISCUSSION

As shown in Figures 3, the experiments are conducted for selecting the kernel parameters for kernel COF and kernel LOF. Precision and recall are considered first, then with the maximum values for the precision and recall, the kernel parameters with the most optimal rank power are determined. For the parameter selection of the kernel COF algorithm, α is selected as 0.3, d as 29, and β as 0.16. While for the parameter selection of the kernel LOF algorithm, α is selected as 0.4, d as 25, and β as 0.04.

The experiment results of kernel COF and kernel LOF are listed in table 2 and table 3 respectively. As is shown in table 2, The Gaussian-COF and the Laplacian-COF could detect more rare objects than the polynomial-COF for m being from 10 to 40, and the rank powers are also stronger. For all the kernel COF algorithms, with the increase of m from 10 to 30, the numbers of rare objects which are detected increase. For example, when the top 30 ranked objects are returned by the algorithms, the Gaussian-COF identified five records in the rare class, the Laplacian-COF detect six bad objects, while only four objects in the bad class are identified by the polynomial-COF. According to precision and recall, the Gaussian-COF and the Laplacian-COF have the same efficiency when the top 35 ranked records are returned. However, the rank power of the Gaussian-COF is 0.28, 17.6% lower than that of the Laplacian-COF. While for the polynomial-COF, the precision is the largest when the number of the ranked objects returned is 25, with four bad records detected, and the recall value is 0.44, much lower than those of the other two algorithms. When m varies from 25 to 40, the number of bad objects detected is keeping unvarying. For the kernel LOF algorithms, when m is 10 and 15, the number of rare objects detected by the Laplacian-LOF is 4, larger than that of Gaussian-LOF and polynomial-LOF. While the Gaussian-LOF could detect more bad objects than the other two algorithms when the top 20 ranked objects are returned. When m increase from 25 to 40, there are same numbers of rare objects detected for all kernel LOF algorithms. For Gaussian-LOF, the rank power is larger than that of Polynomial-LOF, but smaller than that of Laplacian-LOF. Therefore, the effectiveness of LOFs with different kernels would change along with the number of m changes. However, the maximum number of bad objects detected by kernel LOFs is 5, smaller than that of kernel COFs (Gaussian-COF and Laplacian-COF). Consequently, the experimental result shows that the Gaussian and Laplacian COFs perform much better on the oven data set.

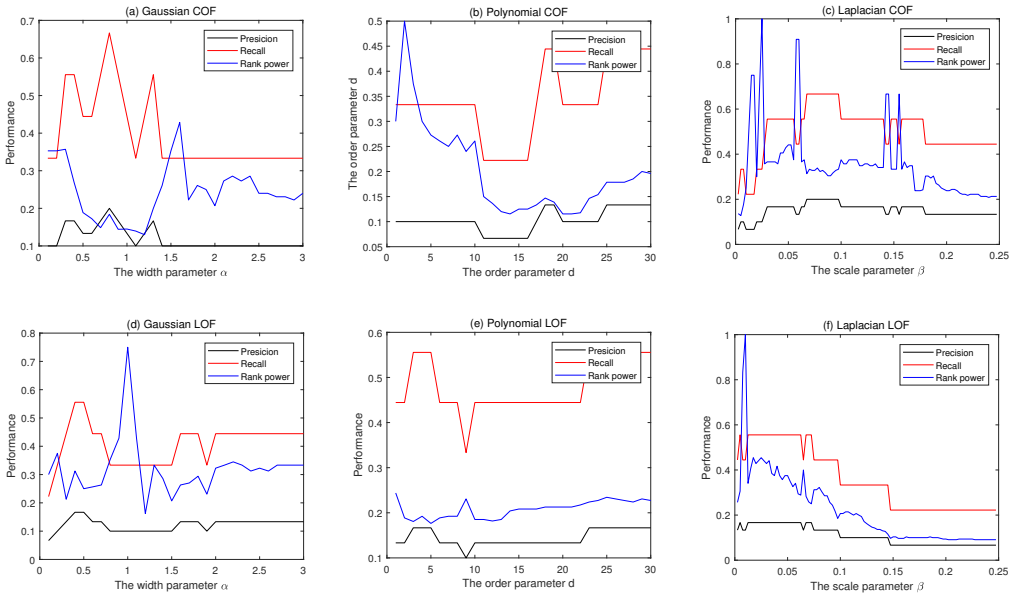

Fig. 3. Prameter selection of kernel function

Table 2. Kernel COF detect rare class

	Gaussian-COF				Polynomial-COF				Laplacian-COF			
m	{Nr}	precision	recall	Rank power	{Nr}	precision	recall	Rank power	{Nr}	precision	recall	Rank power
10	3	0.30	0.33	0.55	1	0.10	0.11	1.00	4	0.40	0.44	0.77
15	4	0.27	0.44	0.45	3	0.20	0.33	0.23	4	0.27	0.44	0.77
20	5	0.25	0.56	0.36	3	0.15	0.33	0.23	4	0.20	0.44	0.77
25	5	0.20	0.56	0.36	4	0.16	0.44	0.20	5	0.20	0.56	0.42
30	5	0.17	0.56	0.36	4	0.13	0.44	0.20	6	0.20	0.67	0.34
35	6	0.17	0.67	0.28	4	0.11	0.44	0.20	6	0.17	0.67	0.34
40	6	0.15	0.67	0.28	4	0.10	0.44	0.20	6	0.15	0.67	0.34

Table 3. Kernel LOF detect rare class

	Gaussian-LOF				Polynomial-LOF				Laplacian-LOF			
m	{Nr}	precision	recall	Rank power	{Nr}	precision	recall	Rank power	{Nr}	precision	recall	Rank power
10	3	0.30	0.33	0.50	2	0.20	0.22	0.75	4	0.40	0.44	0.83
15	3	0.20	0.33	0.50	3	0.20	0.33	0.32	4	0.27	0.44	0.83
20	5	0.25	0.56	0.31	4	0.20	0.44	0.26	4	0.20	0.44	0.83
25	5	0.20	0.56	0.31	5	0.20	0.56	0.23	5	0.20	0.56	0.45
30	5	0.17	0.56	0.31	5	0.17	0.56	0.23	5	0.17	0.56	0.45
35	5	0.14	0.56	0.31	5	0.14	0.56	0.23	5	0.14	0.56	0.45
40	5	0.13	0.56	0.31	5	0.13	0.56	0.23	5	0.13	0.56	0.45

5. CONCLUSION

To improve the energy efficiency in the baking industry, an energy monitoring system developed by the Point Energy Technology from the research group is first introduced to collect data from the production line. After data acquisition, the kernel connectivity-based outlier factor algorithm and kernel local outlier factor algorithm are proposed to detect rare objects. In experiments, an oven data set is used to verify the performance of kernel COFs and kernel LOFs, and the experimental results show that the Gaussian-COF and Laplacian-COF algorithms are effective for oven data detection. Once outliers are identified in the data set, we could check if a production error or data collection error had occurred by prior experience, which would guide the energy management in bread manufacturing processes. The outliers data detection is also useful in improving the data quality and the follow-up analysis accuracy. As a future work, incremental approaches (such as clustering, modelling, and optimisation) will be researched to improve the energy management in the bakery company.

6. REFERENCES

[1] WSP (2015). Industrial decarbonisation and energy efficiency roadmaps to 2050-cross sector summary. Report to DECC and BIS, 26.

[2] Arpita M., Datta A.K. (2008). Bread baking-a review. Journal of food engineering, 86, 465-474.

[3] Peter T., Eric M., and Ernst W. (2014). Energy efficiency opportunities in the U.S. commercial baking industry. Journal of Food Engineering, 130, 14-22. [4] Hawkins D. M. (1980), Identication of outliers–Monographs on

statistics and applied probability. Chapman and Hall Press, London.

[5] Hui C., Rui M., Hongliang R., Shuzhi Sam G.(2016). Data-defect inspection with kernel-neighbor-density-change outlier factor. IEEE transactions on automation science and engineering. DOI: 10.1109/TASE.2016.2603420.

[6] Hodge V. J., Austin J. (2004), A survey of outlier detection methodologies, Artificial intelligence review, 22(2), 85-126.

[7] Jin W., Tung A. K. H., and Han J. (2001), Mining top-n local outliers in large databases, In Proc. of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data mining, San Francisco, USA, 293298.

[8] Ramaswamy S., Rastogi R., and Shim K. (2000), Efcient algorithms for mining outliers from large data sets, Acm Sigmod Record Journal, 29, 427438.

[9] Radovanovic M., Nanopoulos A., and Ivanovic M. (2015), Reverse nearest neighbors in unsupervised distance-based outlier detection, IEEE Transactions on Knowledge and Data Engineering, 27(5), 13691382.

[10] Tang J., Chen Z., Fu A. W., and Cheung D. W. (2006), Capabilities of outlier detection schemes in large datasets, framework and methodologies, Knowledge and Information Systems, 11(1), 4584.

[11] Mansur M. O., Mohd. Noor Md. S. (2005), Outlier detection technique in data mining: a research perspective. Proceedings of the Postgraduate Annual Research Seminar, 23-31.

[12] Wang X., Wang X.L., Chen C., and Wilkes, D.M., Enhancing minimum spanning tree-based clustering by removing density-based outliers, Digit. Signal Process., vol. 23, pp. 15231538, Sep. 2013.

[13] Kim S. et al., Application of density-based outlier detection to database activity monitoring, Inf. Syst. Frontiers, vol. 15, pp. 5565, Mar. 2013.

[14] Breunig M.M., Kriegel H.P., Ng R.T., and Sander J., LOF: Identifying density-based local outliers, in Proc. Int. Conf. Manage. Data, Dallas, TX, USA, 2000, pp. 93104.

[15] Thomas H., Bernhard S., Alexander J. S. (2008). Kernel methods in machine learning. The Annals of Statistics, 36(3), 11711220.

[16] Baeza-Yates R. A., Ribeiro-Neto B. (1999), Modern information retrieval. Addison-Wesley, Boston.

[17] Meng X., Chen Z. (2004), On user-oriented measurements of effectiveness of Web information retrieval systems, International conference on internet computing, 527533.

[18] Ye M., Li X., and Orlowska M. E. (2009), Projected outlier detection in high-dimensional mixed-attributes data set, Expert Systems with Applications, 36, 71047113.

[19] Tang J., Chen Z., Fu A.W., and Cheung D.W. (2007), Capabilities of outlier detection schemes in large datasets, framework and methodologies, Knowl. Inf. Syst., 11(1), 4584.