Outlier detection of time series with a novel hybrid method in cloud computing

Qi Liu*

Jiangsu Collaborative Innovation Center of Atmospheric Environment
and Equipment Technology (CICAEET),
Nanjing University of Information Science and Technology,
Nanjing, 210044, China
and
School of Computing,
Edinburgh Napier University,
Edinburgh, Scotland, UK
Email: q.liu@napier.ac.uk
*Corresponding author

Zhen Wang

School of Computer and Software,
Nanjing University of Information Science and Technology,
Nanjing, China
Email: wangzznj@163.com

Xiaodong Liu

School of Computing,
Edinburgh Napier University,
10 Colinton Road, Edinburgh EH10 5DT, UK
Email: x.liu@napier.ac.uk

Nigel Linge

School of Computing, Science and Engineering,
The University of Salford,
Salford, Greater Manchester, M5 4WT, UK
Email: n.linge@salford.ac.uk

Abstract: In the wake of the development in science and technology, cloud computing has obtained more attention in different field. Meanwhile, outlier detection for data mining in cloud computing is playing significant role in different research domains and massive research works have devoted to outlier detection. However, the existing available methods spend high computation time. Therefore, the improved algorithm of outlier detection, which has higher performance to detect outlier, is presented. In this paper, the proposed method, which is an improved spectral clustering algorithm (SKM++), is fit for handling outliers. Then, pruning data can reduce computational complexity and combine distance-based method Manhattan distance ($dist_m$) to obtain outlier score. Finally, the method confirms the outlier by extreme analysis. This paper validates the presented method by experiments with a real collected data by sensors and comparison against the existing approaches, the experimental results turn out that our proposed method precedes the existing.

Keywords: cloud computing; data mining; outlier detection; spectral clustering; Manhattan distance.

Reference to this paper should be made as follows: Liu, Q., Wang, Z., Liu, X. and Linge, N. (xxxx) 'Outlier detection of time series with a novel hybrid method in cloud computing', *Int. J. High Performance Computing and Networking*, Vol. X, No. Y, pp.xxx–xxx.

Biographical notes: Qi Liu (M'11) received his BSc in Computer Science and Technology from the Zhuzhou Institute of Technology, China in 2003, and MSc and PhD in Data Telecommunications and Networks from the University of Salford, UK in 2006 and 2010. His research interests include context awareness, data communication in MANET and WSN, and smart grid. His recent research work focuses on intelligent agriculture and meteorological observation systems based on WSN.

Zhen Wang received her Bachelor's in Computer Science and Technology from the Nanjing University of Information, Science and Technology in 2015, and she is currently pursuing a Master's in Computer Science and Technology at the Nanjing University of Information Science and Technology. Her research interests include clustering algorithm and household user power analysis.

Xiaodong Liu is a reader and Director of Centre for Information and Software Systems in School of Computing at the Edinburgh Napier University. His research interests include context-aware adaptive services, service evolution, mobile clouds, pervasive computing, software reuse, and green software engineering. He is a member of IEEE Computer Society and British Computer Society. He received his Bachelor's in Computer Science and Technology from the Nanjing University of Information, Science and Technology in 2015, and he is currently pursuing a Master's in Computer Science and Technology at the Nanjing University of Information Science and Technology. His research interests include household user power anomaly analysis.

Nigel Linge received his BSc in Electronics from the University of Salford, UK in 1983, and his PhD in Computer Networks from the University of Salford, UK, in 1987. He was promoted as a Professor of Telecommunications at the University of Salford, UK in 1997. His research interests include location-based and context aware information systems, protocols, mobile systems and applications of networking technology in areas such as energy and building monitoring.

This paper is a revised and expanded version of a paper entitled 'Outlier detection of power data with an improved method in cloud computing' presented at The 3rd International Conference on Cloud Computing and Security, Nanjing, China, 16–18 June 2017.

1 Introduction

Along with the advanced development of cloud computing, more and more fields began to adopt it. Cloud computing not only provides a storage environment, but also provides a computation framework (Chen et al., 2016). Meanwhile, facing with such enormous data volume, cloud computing offer an efficient processing environment for data mining. Recently, data mining (Wu et al., 2014; Yang et al., 2016) also draws more and more researchers' attention, because it can process different types of data, such as time series, spatial sequence and spatial-time sequence (Zhang et al., 2016; Fu et al., 2015). Data mining generally refers to the process of hiding information from a large number of data by algorithm searching. Data mining is usually associated with computer science, statistics, and achieves above objectives through online analysis processing, information retrieval, machine learning, expert system (depending on the old rules of thumb) and pattern recognition and many other methods to achieve the above objectives (Huang et al., 2016).

With the fast-paced development of power systems and cloud computing, a comprehensive survey states three different areas such as energy management, information management and security with different cloud computing applications for the smart grid architecture (Bera et al., 2015). Meanwhile, cloud computing is crucial to the future development of the smart grid. Cloud computing is also used in many fields, such as image processing (Puthal et al., 2015), intelligent transportation (Bitam and Mellouk, 2013) and intelligent medical (Bhavani et al., 2015), data mining (Nekvapil, 2015). In addition, the distributed system is also becoming increasingly important, which is suitable for large amount of data calculation, and the calculation speed is greatly improved.

Data mining tasks include two kinds: descriptive tasks and predictive tasks. The descriptive tasks include clustering, association analysis, sequence, anomaly detection. The predictive tasks include regression and classification. Clustering is an important component of data mining description task. Mining knowledge is based on the intrinsic characteristics and the observation of data. Data mining observes the quality of data by dividing data into different clusters on the basis of clustering (Kumar, 2015). Data mining always has various exceptions, and does not exclude the exception of the data itself (Wang et al., 2015).

With the rapid development of large data mining and data mining, more and more researchers begin to pay attention to some practical problems reflected by data anomalies and outliers in the process of research. Through the analysis of the abnormal worth, it can not only retrieve the abnormal data, but also cause abnormal and abnormal value of electricity through finding abnormal position or abnormal behaviour of users, it is of great significance for the research in electric home users. Outlier detection (Alves et al., 2017) is becoming a significant research point in many information technology fields, such as machine learning, artificial intelligence, information security field and data mining (Xia et al., 2016; Fu et al., 2016). For outlier detection, the pre-sent work is to detect outlier from the collected data with appropriate algorithm. At the same time, outlier analysis is on purpose of detecting and identifying outlier, which is also an essential process in data mining and plays an important role in different application areas and different types of data.

On top of the typical Markov chain model, a feasible multi-order Markov chain based framework for anomaly detection was proposed in (Fu et al., 2017), which was described by multi-order Markov chain and multivariate time series composed of the proposed algorithm with the

framework of statistical learning in the course of training form. In order to reduce the time complexity and space complexity, the design and implementation of the algorithm of numerical and nonzero value tables based on initial and transition matrix. A method to detect the time delay of the whole handoff scheme based on Bayesian mixture model was proposed (Hu et al., 2016). In order to solve the problem of outliers, a new kind of threshold and sliding window is proposed to modify the class set and the model parameters.

For the sake of managing outlier detection with clustering-based methods, we present an improved outlier detection method to solve this problem for real power dataset, which can effectively achieve the recognition of abnormal power data. The method presented in this paper is primarily split into three steps. The first step is to adopt improved hybrid clustering algorithm (SKM++) to realise data pruning, which can reduce data complexity and improve calculation efficiency. The second step is to utilise distance-based Manhattan distance ($dist_m$) to compute outlier score, which measures the $dist_m$ between outlier data point and the closest cluster. The last step is to apply extreme analysis to confirm the outlier, and then the outlier can be detected.

The rest of this paper is structured as follows: Section 2 summarises research methods to the outlier detection. The presented method in this paper is introduced in details in Section 3. In Section 4, the experimental studies and evaluation of methods are reported, while conclusion and future work are covered in Section 5.

2 Related work

We overview available and typical methods for clustering-based outlier detection and distance-based outlier detection. The approaches may have more high computational complexity and do not confirm the outlier through different ways. Recently, with the development of hardware and software technology, there has been a great deal of work and research on time series outlier from a computational point of view in computer science. Clustering-based algorithms concentrate mainly on clusters and then consider outlier detection, which may have the order reversed (Gupta et al., 2014). Meanwhile, these methods do not have better performance for high dimensional dataset.

Outlier detection (Yu et al., 2017) has achieved more and more focus in machine learning, artificial intelligence and data mining. For outlier detection, the present work is to detect the outlier with appropriate algorithms for the collected data. Meanwhile, it is also pivotal to analyse the detected outlier and identify abnormal types of this outlier reflects. Outlier detection of data processing is an indispensable process in data mining, which plays an important role in different application areas and the different types of data. Firstly, outlier detection has been applied to many different fields, such as internet of things (IoT) dataset, the outlier detection techniques is proposed based on one-class quarter-sphere support vector machine, which combine with the spatial and temporal characteristics of sensor data, the proposed method can reduce the rate of data transmission misinformation (Zhang et al., 2010). Such as medicine, a method of outlier detection of biomedical data based on fuzzy logic, which is used in detect abnormal point during the insertion of data. The anomaly of medical data in transmission is detected by two angles, horizontal and vertical (Yong et al., 2014). Such as physical distribution, an extended clustering algorithm – package balancing *k*-means is proposed and load balancing is achieved by adding weight values on the basis of standard *k*-means algorithm. The experiment shows that this method is less expensive, and the calculation time and convergence speed are improved (Dai et al., 2017).

The existing outlier detection methods, mainly divided into three categories, the distance-based outlier detection, the density-based outlier detection and the clustering-based outlier detection (Dargahi et al., 2016; Gupta and Badve, 2015). The distance-based outlier detection algorithm shows that a point far away from most of the points, then the point is outlier. The reverse nearest neighbour algorithm was presented, which was applied to the unsupervised outlier detection problem and improved *k*-nearest neighbour algorithm that based on the distance. It is appropriate to dispose the outlier detection problems of high-dimensional data processing (Radovanovic et al., 2015). The distance-based outlier detection is closely related to the density-based outlier detection, which are usually defined by using the proximity. Local anomaly detection was proposed, which based on local density data examples for outlier detection (Breunig et al., 2000). Clustering analysis is always applied to a strong correlation, while outlier detection usually is used in correlation with other objects are little strong with the data object. The CBOD algorithm for outlier detection was proposed, which was based on clustering and divided into two steps, the first is the primary application of clustering algorithm is used to cluster the existing data, and then through the abnormal factors actualise evaluation of outlier detection (Jiang and An, 2008).

In Salehi et al. (2016), the traditional outlier detection techniques can no longer assume that all data can be stored and processed. While the well-known local outlier factor (LOF) algorithm is an incremental version, it assumes an infinite memory to store all previous data points, an efficient incremental algorithm for detecting local anomalies of data streams was presented. A more flexible version (MILOF_F), which has an accuracy close to the incremental LOF but at a fixed memory limit.

Outlier detection method KPCA-based Mahalanobis kernel was proposed (Lavoie and Merlo, 2012). The method consists of three phases: the first phase is training phase, the original model is used to detect the existing data and then receive new data constantly. The second phase is detection phase, in which the mapping distance of each data vector in the subspace defined by the global principal component is calculated. The last phase is Mahalanobis Kernel, which is combined with Manhattan distance to isolate outliers from

normal data distribution patterns. Such a scale combines the increasingly complex systems of components and interactions, providing critical challenges for highly reliable cloud computing, as well as anomaly detection and resource management.

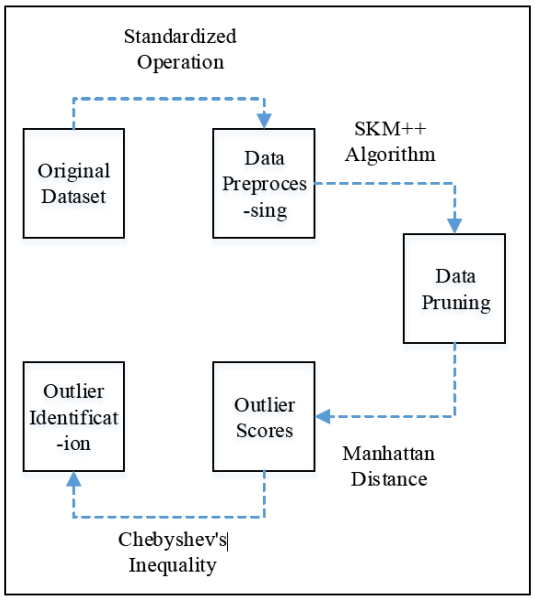
A hybrid self-evolving anomaly detection framework using one-class and two-class support vector machines was presented (Fu et al., 2012), which adopts a hybrid outlier detection mechanism based on the first level and two level support vector machine.

In this paper, we employ SKM++ algorithm instead of the traditional *k*-means with *k*-means++ for the spectral clustering algorithm. Meanwhile, SKM++ algorithm achieves the pruning data and reduces the computational complexity and computation time of calculating the outlier score.

3 The proposed method

In this section, we mainly introduce the method proposed in this paper. The first part is data pre-processing. In this part, standardised operation of data, which applies Z-score standardisation, was implemented. The SKM++ algorithm and the interrelated concepts are described in details in the second part. Then the Manhattan distance ($dist_m$) and extreme analysis are recommended in the third part. The overall design of the proposed method is shown in Figure 1.

Figure 1 The overall design of the presented method (see online version for colours)



3.1 Data pre-processing

The main tasks of data pre-processing are as follows:

- data cleaning: fill in vacancy values, smooth noise data, identify, delete outliers and resolve inconsistencies
- data integration: integrating multiple databases, data cubes, files
- data transformation: eliminating redundant attributes and data aggregation, projecting data from a larger subspace to a smaller subspace
- data reduction: the compressed representation of the data set is small, but can obtain similar or identical results
- data discretisation: a part of data protocol, which is used to define data by concept hierarchy and discretisation of data, which is important to digital data.

We handle original dataset to accommodate our proposed method before we apply it. By intercepting partial sample data from original data set, we can remove the noise and employ standardised processing of sample data set. In this paper, we use Z-score standardisation to process sample data. The quantity of the Z value represents the distance between the original fraction and the parent mean, and is calculated in standard deviation. When the original score is below the average, Z is negative, whereas the positive number is positive. Z-score standardisation, the processed data is in accord with the standard normal distribution, that is, the mean value of 0, the standard deviation of 1, the conversion function is defined as follows:

$$X = \frac{x - \mu}{\sigma} \tag{1}$$

μ is mean of all the sample data and σ is the standard deviation of all samples data.

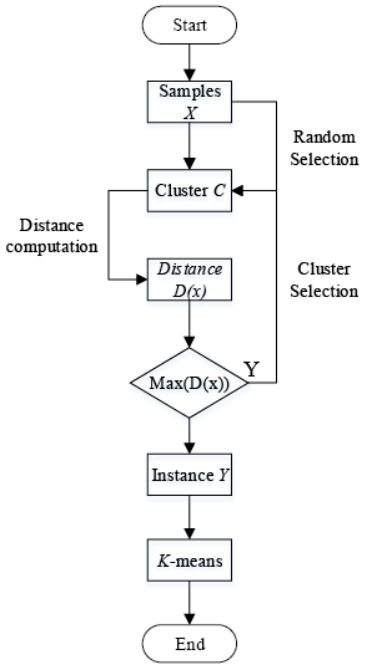
3.2 SKM++ algorithm

K-means++: the basic idea of the initial seeds selection is that the distance between the initial cluster centres is as far away as possible. The number of cluster centres k does not need to be given in advance which solves the problem that the selection of the k value is very difficult to estimate in practice. The flowchart of k-means++ is shown in Figure 2.

K-means++ is a clustering model which can keep the clustering accuracy as much as possible. The detailed steps of the algorithm are as follows:

- randomly select a point from the set of input data points as the first clustering centre
- for each point in the dataset X, calculate the distance between X and the nearest cluster centre (the chosen cluster centre)
- choose a new data point as a new clustering centre, the selection principle is: the points is corresponding to that $D(x)$ is larger, the probability of being selected as centre is larger than the others
- repeat step 2 and step 3 until the k cluster centres are selected
- use the k initial clustering centres to run standard *k*-means algorithms.

Figure 2 The flowchart of k-means++



Spectral clustering algorithm: it is based on spectral graph theory. Compared with the traditional clustering algorithm, it has many advantages in the sample space of any shape clustering and convergence to the global optimum solution. In this paper, we improve spectral clustering algorithm (SKM++) by replacing the k-means algorithm with *k*-means++. Next, we describe the SKM++ in detail. Assume that the standardised power dataset is X, which has n instances $\{S1, S2, \ldots, Sn\}$ and each instance has m properties, for $i, j \in [1, n]$, the related.

- Similarity matrix (Ws): computing the similarity degree of all instances. indicates the similarity of i^{th} instance and the j^{th} instance and the calculation formula as follows:

$$w_{ij} = \exp\left(\frac{-\|S_i - S_j\|^2}{2\sigma^2}\right) \tag{2}$$

- Degree matrix (D_d): computing sum of the similarity of i^{th} instance other instances. In the above, we provide the method of calculation with W, then we compute the degree matrix based on W.

$$d_i = w_{i1} + w_{i2} + \ldots + w_{in} = \sum_{j=1}^{m} w_{ij} \tag{3}$$

- Random walking Laplacian matrix (L_{rw}): the degree matrix and similarity matrix do subtraction operation and then multiply the inverse of D_d. Then the eigenvalues and eigenvectors and are obtained by dealing with L_{rw}.

$$L_{rw} = D_d^{-1}(D_d - W_s) \tag{4}$$

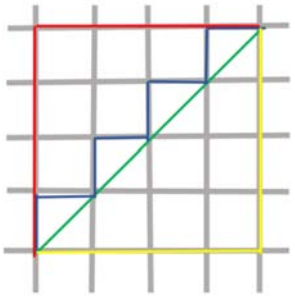
Finally, selecting the k^{th} eigenvectors makes up a new feature vector matrix and cluster the new matrix by *k*-means++ with threshold f. The algorithm 1 describes the implementation process of SKM++ algorithm.

3.3 Manhattan distance

In the last part, we have introduced SKM++ to prune dataset, and then we dispose the pruned dataset, that is we measure the Manhattan distance ($dist_m$) (Gupta and Bhatia, 2015) of candidate outlier data points to the nearest cluster for receiving outlier score. The calculation formula is as follows:

$$dist_m = \sum_{i=1}^{n} |x_i - y_i| \tag{5}$$

Figure 3 Manhattan distance figure (see online version for colours)



In Figure 3, the red, blue and yellow lines represent the distance between the two black spots of Manhattan.

Algorithm 1 SKM++ algorithm

Input Dataset $X = \{S_1, S_2, \ldots, S_m\}$ the number of clusters k
Output Clusters C_m
Let $X \neq \phi$, $S_n \neq \phi$
while $S_n \in X$ **do**
  calculate W_s by Gauss similarity w_{ij}
  if $i \neq j$
    $w_{ij} = \exp\left(\frac{-\|S_i - S_j\|^2}{2\sigma^2}\right)$
  else if $i = j$, **then**
    $w_{ij} = 0$
  eliminate their own similarity, reduce the amount of calculation
    $D_d(i,i) = \sum_{j=1}^{n} S_{ij}$
  standardised D_d in order for a node to be kicked out
  end if
  if $W_s \neq \mathbf{0}$, $D_d \neq \mathbf{0}$, **then**
  calculate L_{rw}
    $L_{rw} = D_d^{-1}(D_d - W_s)$
  end if

16 calculate the eigenvalue $\{\lambda_1, \lambda_2, \ldots, \lambda_k\}$ and eigenvector and $\{\eta 1, \eta 2, \ldots, \eta_k\}$ select the k-th eigenvectors of L_{rw}, form a new matrix, then cluster k eigenvectors

17 **end if**

18 **end while**

3.4 Chebyshev's inequality

In order to determine the final outliers, we introduce Chebyshev's inequality (Du et al., 2016) as shown in (6). Assuming that the mathematical expectation and the variance of the random variable X are existent, then the arbitrary constant $\varepsilon > 0$

$$P(|x-\mu| < \varepsilon) \geq 1 - \frac{\sigma}{\varepsilon^2} \tag{6}$$

in which x represents the variable, μ is the mean and σ is the standard deviation and k is the number of standard deviation from the mean.

4 Experiments and evaluation

4.1 Experiments analysis

We download the power dataset in http://iawe.github.io/, which was collected about 73 days of power and water consumption, but we choose the part of power dataset and the mains data, which includes 4,391,060 instances in seconds. Then we dispose the power mains data in hours and label 200 outlier records. Secondly, we standardise it with Z-score standardisation and the standardised dataset tends to standard normal distribution. The distribution of the standardised dataset is as shown in Figure 4. The fluorescent green dots express the data instances and we cannot distinguish the outlier from the figure.

We cannot make a decision whether the candidate outliers simply from the figure. In Figure 5, some candidate outliers are encircled by red square, they are close to the cluster centre, but they are recognised as candidate outliers. Thus, we need to confirm them with Manhattan distance and Chebyshev's inequality.

Figure 4 The distribution of the standardised power dataset (see online version for colours)

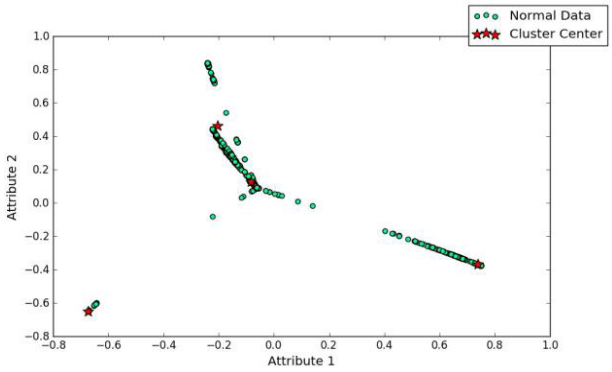


Figure 5 The clustered power dataset by SKM++ (see online version for colours)

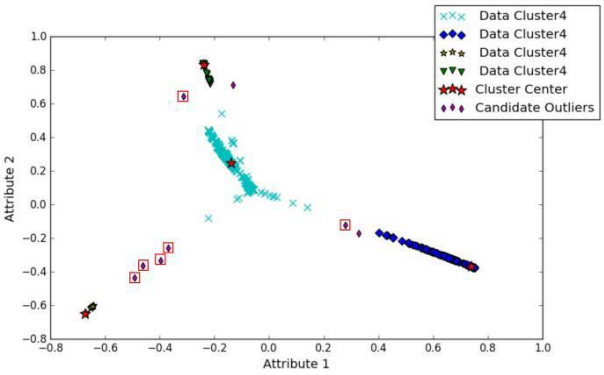


Figure 6 The four clusters for dataset 1, (a) the cluster 1 (b) the cluster 2 (c) the cluster 3 (d) the cluster 4 (see online version for colours)

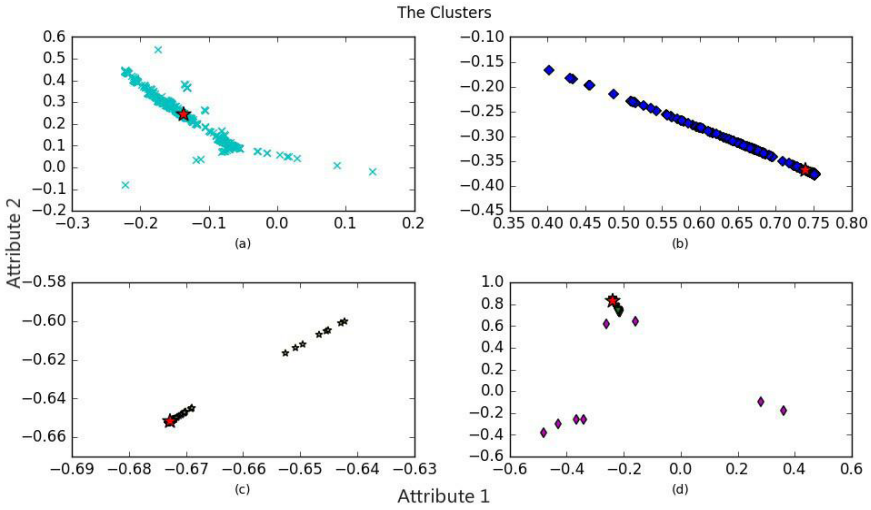


Figure 6 is the local details figure of Figure 3, which depicts the four clusters of dataset 1 respectively,

a the data cluster 1

b the data cluster 2

c the data cluster 3

d the data cluster 4.

The red star indicates the cluster centre the purple diamond indicates the candidate outlier and the other indicate normal data. In Figure 5, we can ascertain that each cluster have candidate outliers.

Here, we apply the SKM++ to prune data for reducing amount of calculation. Figure 6 describes the distribution of pruned data and candidate outliers, the blue dots indicate the pruned data and the violet diamond indicate candidate outliers. By making comparison between Figures 7 and 8, we can conclude that the candidate outliers are not always the final outliers. In Figure 8, the final outliers are described by employing the Chebyshev's inequality.

Figure 7 The pruned dataset (see online version for colours)

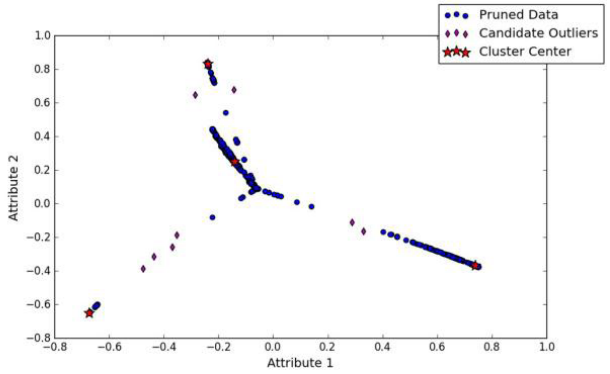
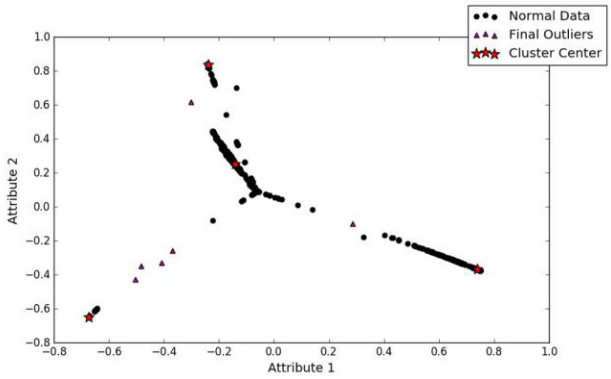


Figure 8 Outlier results analysis (see online version for colours)



4.2 Performance evaluation

In Section 4, we have labelled 200 outliers for assessing accuracy of detecting outliers with different algorithms. We apply the purity method, which is a very simple clustering evaluation method. It only needs to calculate the proportion of the number of samples in the correct cluster, the formula is as follow:

$$purity_{SKM++}(X, C) = \frac{1}{N}\max_{j}\left|\omega_k \cap c_j\right| \tag{7}$$

X is the set of samples, C = $\{\omega_1, \omega_2, \ldots, \omega_k\}$ is the set of clusters, x_j is the j^{th} sample, N is the total number of sample.

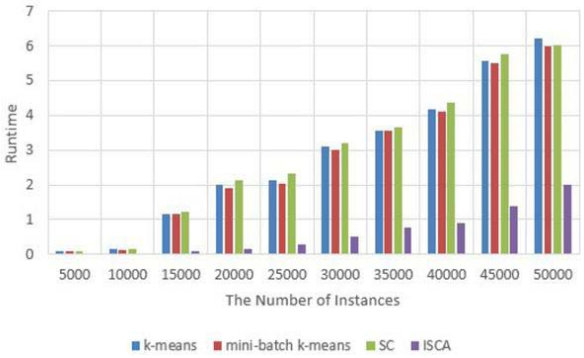
The advantage of the purity method (Aguena and Lima, 2016) is that it is convenient to compute, the value is between 0–1, and the completely wrong clustering method has a value of 0, and the correct method value is 1. At the same time, purity is also a very obvious shortcoming. I t is impossible for the degradation of clustering method in evaluation, if each document clustering algorithm clustered into one class, then the algorithm that all documents are correctly classified, then the purity value is 1.

Table 1 The accuracy of four algorithms

Algorithms	*Accuracy*
k-means	87.55%
Mini-batch k-means	88.50%
SC	86.50%
SKM++	92.50%

The results are shown in Table 1 we can reach a conclusion that the proposed method SKM++ 92.50% is more accurate than the others.

Figure 9 Runtime of the related algorithms (see online version for colours)



According to experiments, we make a comparison with the other three related algorithm in Figure 9. With the growth of the number of data, the runtime of the four algorithms is increasing. From Figure 9, we can conclude that the SKM++ has better performance than the other three algorithms. The method we proposed can reduce run time and have lower computational complexity, which also detect the outliers effectively.

5 Conclusions and future work

An improved SKM++ method was proposed, which was applied in the outlier detection. We performed some evaluations with real datasets and the experimental results show that the proposed method SKM++ outperforms the existing clustering methods based on clustering with high processing efficiency for high-dimensional data, the accuracy of the algorithm is also improved. However, there are still some deficiencies in our approach, to deal with large amounts of data should be improved. When the amount of power data is increasing, the graph based algorithm in this paper will consume more memory and CPU during computing dataset matrix, and the computation time will also increase exponentially with the increase of data. With the growth of data, the outlier data also increases, and the accuracy of the improved algorithm will not be guaranteed. In addition, k-means algorithm involves k, and is usually selected through continuous iteration until the optimal centre point is obtained, but it is easy to cause the algorithm to fall into the local optimal solution. With the growth of data, the exception data also increases, and the accuracy of the improved algorithm will not be guaranteed.

In the future, for the accuracy of the spectral clustering algorithm, we would apply the new outlier detection approach and improve the existing method to abnormal power dataset, The introduction of genetic simulated annealing algorithm based on spectral clustering algorithm, which has the three main benefits:

1 to avoid the iterative value for K, save some computing resources

2 can prevent the algorithm into a local optimal solution

3 improve the accuracy of the algorithm in anomaly detection.

For memory and CPU, a distributed computing framework, Hadoop or Spark, will be used to deal with massive power data sets through the idea of 'divide and conquer'. The problem of insufficient resources will be solved when the amount of data is huge. The MapReduce of Hadoop is mainly divided into two stages, the Map stage and the Reduce stage. In the map stage, the data is piecewise calculated, and the reduce stage converged the results of the map stage calculation. Spark abstracts data into an elastic dataset and calculates through memory, and the intermediate result data is still stored in memory. These two architectures are well suited to dealing with massive data sets and frequent computing operations.

Acknowledgements

This work is supported by Marie Curie Fellowship (701697-CAR-MSCA-IF-EF-ST), the NSFC (61300238 and 61672295), the 2014 Project of six personnel in Jiangsu Province under Grant No. 2014-WLW-013, and the PAPD fund.

References

Aguena, M. and Lima, M. (2016) 'Effects of completeness and purity on cluster dark energy constraints', *Cosmology and Nongalactic Astrophysics*, (astro-ph.CO), ArXiv e-prints [arXiv:1611.05468].".

Alves, W., Martins, D., Bezerra, U. and Klautau, A. (2017) 'A hybrid approach for big data outlier detection from electric power SCADA system', *IEEE Latin America Transactions*, Vol. 15, No. 1, pp.57–64.

Bera, S., Misra, S. and Rodrigues, J.J.P.C. (2015) 'Cloud computing applications for smart grid: a survey', *Parallel and Distributed Systems IEEE Transactions on*, Vol. 26, No. 5, pp.1477–1494.

Bhavani, S.R., Senthilkumar, J., Manjula, D., Krishnamoorthy, R. and Kannan, A. (2015) 'CIMIDx: prototype for a cloud-based system to support intelligent medical image diagnosis with efficiency', *JMIR Medical Informatics*, Vol. 3, No. 1, p.e12.

Bitam, S. and Mellouk, A. (2013) 'ITS-cloud: cloud computing for intelligent transportation system', *Global Communications Conference*, pp.2054–2059.

Breunig, M.M., Kriegel, H.P., Ng, R.T. and Sander, J. (2000) 'LOF: identifying density-based local outliers', in *ACM SIGMOD International Conference on Management of Data*, 16–18 May, Vol. 29, pp.93–104, Dallas, Texas, USA.

Chen, J., Li, K., Tang, Z., Bilal, K., Yu, S., Weng C. and Li K. (2016) 'A parallel random forest algorithm for big data in a spark cloud computing environment', *IEEE Transactions on Parallel and Distributed Systems*, Vol. 28, No. 4, pp.919–933.

Dai, Y., Yang, W., Wang, G. (2017) 'Package balancing k-means algorithm for physical distribution', *International Journal of Computational Science and Engineering*, Vol. 14, No. 4, p.349.

Dargahi, T., Javadi, H.H.S., Shafiei, H. et al. (2016) 'Detection and mitigation of pulse-delay attacks in pairwise-secured wireless sensor networks', *International Journal of High Performance Computing and Networking*.

Du, H., Zhao, S., Zhang, D. and Wu, J. (2016) 'Novel clustering-based approach for local outlier detection', *Computer Communications Workshops*.

Fu, S., Liu, J. and Pannu, H. (2012) 'A hybrid anomaly detection framework in cloud computing using one-class and two-class support vector machines', *Advanced Data Mining and Applications*, Vol. 7716, pp.726–738, Springer, Berlin Heidelberg.

Fu, Z., Sun, X., Liu, Q., Zhou, L. and Shu, J. (2015) 'achieving efficient cloud search services: multi-keyword ranked search over encrypted cloud data supporting parallel computing', *IEICE Transactions on Communications*, Vol. 98, No. 1, pp.190–200.

Fu, Z., Wu, X., Guan, C., Sun, X. and Ren, K. (2017) 'Statistical learning for outlier detection in cloud server systems: a multi-order markov chain framework', *IEEE Transactions on Cloud Computing*. Vol. 11, No. 12, pp.2706–2716.

Fu, Z.J., Wu, X.L., Guan, C.W., Sun, X.M. and Ren, K. (2016) 'Toward efficient multi-keyword fuzzy search over encrypted outsourced data with accuracy improvement', in *IEEE Transactions on Information Forensics and Security*, Vol. 11, No.12, pp.2706–2716.

Gupta, B.B. and Badve, O.P. (2015) 'GARCH and ANN-based DDoS detection and filtering in cloud computing environment', *International Journal of Embedded Systems*, Vol. 9, No. 5, 10.1504/IJES.2017.10007720.

Gupta, M., Gao, J., Aggarwal, C. and Han, J.W. (2014) 'Outlier detection for temporal data: a survey', in *IEEE Transactions on Knowledge and Data Engineering*, Vol. 26, No. 9, pp.2250–2267.

Gupta, S. and Bhatia, V. (2015) 'A Manhattan distance approach for energy optimization in wireless sensor network', *International Conference on Next Generation Computing Technologies*, pp.203–206.

Hu, H., Yuan, D., Liao, M. and Liu, Y. (2016) 'Packet cache-forward method based on improved Bayesian outlier detection for mobile handover in satellite networks', *China Communications*, Vol. 13, No. 6, pp.167–177.

Huang, W., Meng, L., Zhang, D. and Zhang, W. (2016) 'In-memory parallel processing of massive remotely sensed data using an apache spark on Hadoop YARN model', *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Vol. 10, No. 1, pp.3–19.

Jiang, S.Y. and An, Q.B. (2008) 'Clustering-based outlier detection method', in *International Conference on Fuzzy Systems and Knowledge Discovery*, pp.429–433.

Kumar, G. (2015) *A Survey on Clustering – A Data Mining Technique*.

Lavoie, T. and Merlo, E. (2012) 'An accurate estimation of the Levenshtein distance using metric trees and Manhattan distance', *International Workshop on Software Clones*, pp.1–7.

Nekvapil, V. (2015) 'Cloud computing in data mining – a survey', *Journal of Systems Integration*, Vol. 6, No. 1, pp.12–23.

Puthal, D., Sahoo, B.P.S., Mishra, S. and Swain, S. (2015) 'Cloud computing features, issues, and challenges: a big picture', *International Conference on Computational Intelligence and Networks*, Bhubaneshwar, pp.116–123.

Radovanovic, M., Nanopoulos, A. and Ivanovic, M. (2015) 'Reverse nearest neighbors in unsupervised distance-based outlier detection', *IEEE Transactions on Knowledge and Data Engineering*, Vol. 27, No. 5, pp.1369–1382.

Salehi, M., Leckie, C., Bezdek, J.C., Vaithianathan, T. and Zhang, X. (2016) 'Fast memory efficient local outlier detection in data streams', *IEEE Transactions on Knowledge and Data Engineering*, Vol. 28, No. 12, pp.3246–3260.

Wang, L.J., Zheng-Wei, H.E. and Feng, P.X. (2015) 'Study of outlier data mining algorithm based on ICA', *Journal of University of Electronic Science and Technology of China*, Vol. 44, No. 2, pp.211–214.

Wu, X., Zhu, X., Wu, G.Q. and Ding, W. (2014) 'Data mining with big data', *IEEE Transactions on Knowledge and Data Engineering*, Vol. 26, No. 1, pp.97–107.

Xia, Z., Wang, X., Sun, X. and Wang, Q. (2016) 'A secure and dynamic multi-keyword ranked search scheme over encrypted cloud data', *IEEE Transactions on Parallel and Distributed Systems*, Vol. 27, No. 2, pp.340–352.

Yang, C., Yu, M., Hu, F., Jiang, Y. and Li, Y. (2016) 'Utilizing cloud computing to address big geospatial data challenges', *Computers Environment and Urban Systems*, pp.120–128.

Yang, Y. and Wang, H. (2018) 'Multi-view clustering: a survey', *Big Data Mining and Analytics*, Vol. 1, No. 2, pp.83–107.

Yong, K.K., Sang, Y.L., Seo, S. et al. (2014) 'Fuzzy logic-based outlier detection for bio-medical data', *International Conference on Fuzzy Theory and ITS Applications*, IEEE, pp.117–121.

Yu, Z., Yang, Z., Su, X. et al. (2017) 'Evaluation and comparison of ten data race detection techniques', *International Journal of High Performance Computing and Networking*, Vol. 10, No. 4, p.279.

Zhang, C.K., Sun, Y.Q., Guo, J.W. and Xiong, T.F. (2016 'Mining dynamic association rules from multiple time-series data based on data of power plant', *IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*, IEEE, pp.1968–1972.

Zhang, Y., Meratnia, N. and Havinga, P.J.M. (2010) 'Ensuring high sensor data quality through use of online outlier detection techniques', *International Journal of Sensor Networks*, Vol. 7, No. 3, pp.141–151.