IOP Conference Series: Earth and Environmental Science

PAPER • OPEN ACCESS

An Improvement Method to Mining Outliers Based on Social and Spatial Information

To cite this article: Ce Zhou and Xiao Luo 2019 *IOP Conf. Ser.: Earth Environ. Sci.* **234** 012085

An Improvement Method to Mining Outliers Based on Social and Spatial Information

Ce Zhou[1] and Xiao Luo[2]

[1]College of Computer Science and Technology, Tsinghua University, Beijing 10084
[2]College of Computer Science and Technology, Jilin University, Changchun 130012, Changchun
Email: zhouce319@163.com; 630267916@qq.com

Abstract. In this paper, a method for finding outliers is proposed for communication data. By extracting the social and spatial information as the eigenvalues from the user's call list. We use clustering method to aggregate users with the same attributes, and then analyze the clusters to find outliers. We use the real communication data of a communication company as the data set, and the experimental results show that our method can quickly and accurately find the abnormal communication user.

1. Introduction

With the rapid development of Internet technology and computer science, many of our daily behaviors will be stored in data way. With the continuous growth of data, the research on the hidden characteristics of these data has become a hot topic at present. With the continuous improvement of data mining, outlier detection as an important area of data mining has also been valued. Outliers are also called anormaly points, which are obviously different from most data in a set of data sets. [1] The outlier is different from the noise. It has no commendatory or derogatory meaning, but it has many important meanings behind it. The analysis of abnormal points can bring a lot of application to people's work and life. [2]

Outlier detection has been widely used in many fields. E.M.Knorr and R.T.NG find the star players of the whole game by analyzing the competition data of the International Ice Hockey League players. K.Yamanishi and J.Takeuchi [3] first applied outlier detection to market analysis in the financial field. At the same time, the research of anomaly detection has also become the main object of data analysis in the fields of network intrusion detection, disease diagnosis, emergency detection and spatiotemporal data anomaly, and its value is becoming more and more indispensable [4-9].

In this paper, a clustering method is proposed to find the outlier communication data. By using social and spatial information as the eigenvalues, we distinguish normal users and abnormal users. The experimental results show that our method can quickly and accurately find the outliers.

The rest of this paper is organized as follows. Section 2 introduce the model of the outlier mining based on social and spatial information. Section 3 shows the result of our model and compares with other methods to prove the efficiency and accuracy of our method. Finally, section 4 concludes this paper.

2. Model And Algorithm

2.1Social relationship feature selection

The social relationship of the user can be distinguished according to the age distribution and number of the contacts. Through the network library of python, we can get the following social relationships of users from different ages.

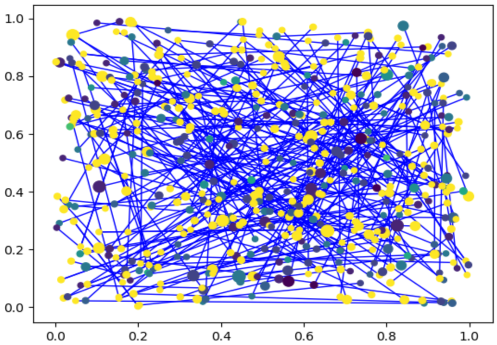

Figure 1: Social relationship map of the age of 10-20

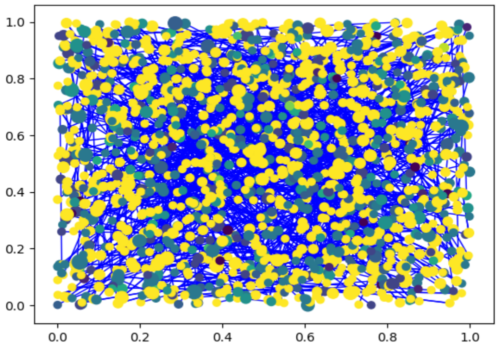

Figure 2: Social relationship map of the age of 40-50

According to the above picture, a 40-50 year old person has a much greater frequency of communication than a 10-20 year old, and the age range of the 40-50 year old user with a close contact is also different with the 10-20 year old.

We define user set U = {u1, u2, ...,un} which n=|U|. So we can define user's social relationship eigenvalue Vui as follows:

$$\mathrm{V}_{\mathrm{u_i}} = \left[\mathrm{c}_{0\mathrm{u_i}}, \mathrm{c}_{1\mathrm{u_i}}, \ldots, \mathrm{c}_{9\mathrm{u_i}}\right] \tag{1}$$

Where$C_{\mathrm{ku}_i}(k \in [0,9])$presents the frequency of user u_i make phone call to different age.The frequency of communication between users on age group K is as follows:

$$\mathrm{C}_{\mathrm{ku_i}} = \frac{\left|\mathrm{N}_{\mathrm{ku_i}}\right|}{\sum_{\mathrm{k=0}}^{9}\left|\mathrm{N}_{\mathrm{ku_i}}\right|} \tag{2}$$

Where N_{ku_i} is user u_i make phone call to the users in age group k.Then we can get all the users eigenvalue as follows:

$$\mathrm{V} = \begin{bmatrix} \mathrm{V}_{\mathrm{u_1}} \\ \mathrm{V}_{\mathrm{u_2}} \\ \cdots \\ \mathrm{V}_{\mathrm{u_n}} \end{bmatrix} = \begin{bmatrix} \mathrm{c}_{0\mathrm{u_1}} & \mathrm{c}_{1\mathrm{u_1}} & \cdots & \mathrm{c}_{9\mathrm{u_1}} \\ \mathrm{c}_{0\mathrm{u_2}} & \mathrm{c}_{0\mathrm{u_2}} & \cdots & \mathrm{c}_{0\mathrm{u_2}} \\ \vdots & \vdots & \vdots & \vdots \\ \mathrm{c}_{0\mathrm{u_n}} & \mathrm{c}_{0\mathrm{u_n}} & \cdots & \mathrm{c}_{0\mathrm{u_n}} \end{bmatrix} \tag{3}$$

2.2Spatial features selection

In a period of time, the spatial information of a class of users will follow certain rules.In this paper, the base station connected by the user for a period of time is approximated as the spatial information of the user, and the user's trajectory is generated by integrating all the users' spatial information in a period of time.

```
Algorithm 1 Gird clustering
Start form a level
For each cell at this level, we compute the query related attribute values.
From the computed attribute values and constraints, we mark each cell as relevant or not. (The unrelated cell is no longer considered, the next lower level processing only checks the remaining related units).
If this layer is bottom, turn (6), otherwise turn (5).
We move from the hierarchical structure to the next level, according to step 2.
The query results are satisfied and go to step 8, otherwise go to step 7
Restore data to related cells for further processing to obtain satisfactory results, and move to step (8).
End
```

In previous research, we use DBSCAN to cluster our base station information to cluster [10], and use it as the basis of user geographic location information. When the density of spatial clustering is not uniform or the difference of the cluster spacing is very large, the parameter MinPts and Eps are difficult to select, and the quality of clustering is not good. Therefore, in this paper, we use grid clustering to do a more detailed division of spatial information. The algorithm is as follows.

2.3Data Process

In order to make the eigenvalues smoother, we use linear functions to standardize data as follows:

$$y = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (4)$$

Where x presents standardized data,x_{max}presents the maximum value of the data,x_{min} presents the minimum value of the data,y is the value after data standardization.

The eigenvalue attribute is a high dimensional matrix.If there are few social activities between users, the matrix will be sparse,which may affect the accuracy of clustering results.In this paper, based on the data reduction, the high dimension eigenvalue is projected to the two-dimensional feature space without losing the key attributes of the user to ensure the analysis of the data. At the same time, the data reduction will reduce the computational complexity of the distance between objects and improve the clustering speed.

2.4Algorithm

The main step of outliers mining based on social and spatial information are shown in the follow fig 3:

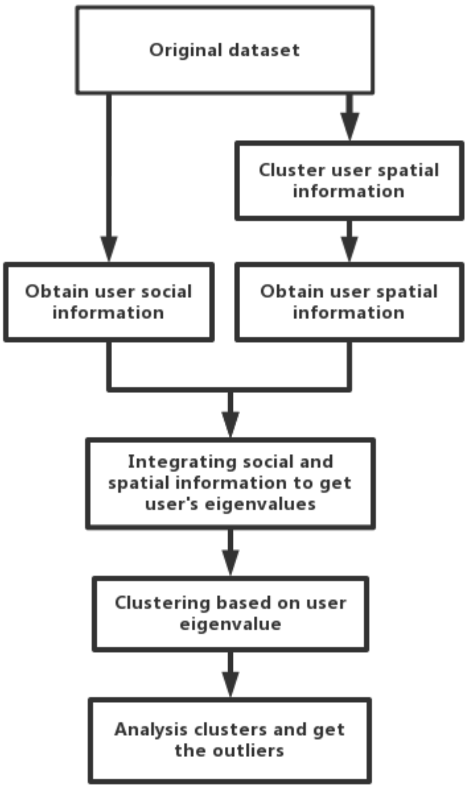
Original dataset
Cluster user spatial information
Obtain user social information
Obtain user spatial information
Integrating social and spatial information to get user's eigenvalues
Clustering based on user eigenvalue
Analysis clusters and get the outliers

Figure 3: Algorithm of outliers mining based on social and spatial information

3. Result

3.1 Analysis of experimental results

In this paper, we use a user's monthly data from a communication company as the original data set of the experiment. After extracting the user's social and spatial attributes as the user's eigenvalue, we cluster the data with KMEANS, with the parameter K=15. After the clustering results are obtained, we reverse query the age of the user in each cluster. The distribution results are shown as follows.

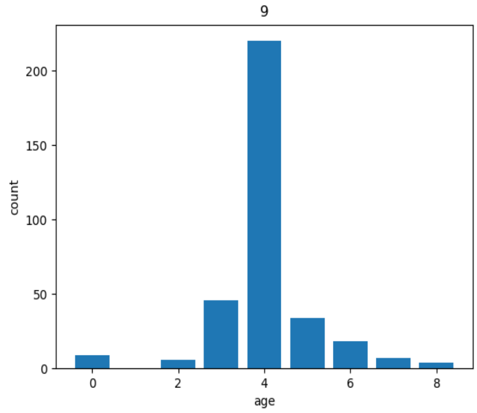

Figure 4: Age distribution in cluster 9

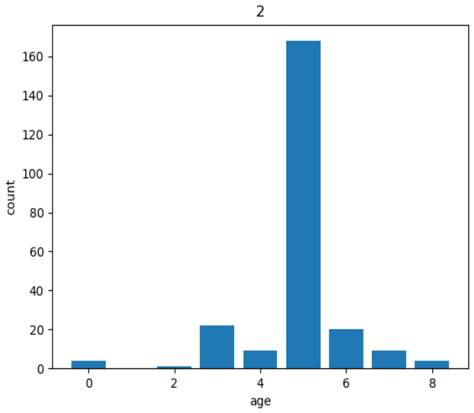

Figure 5: Age distribution in cluster 2

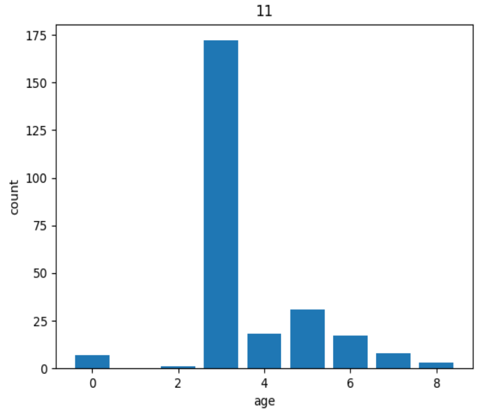

Figure 6: Age distribution in cluster 11

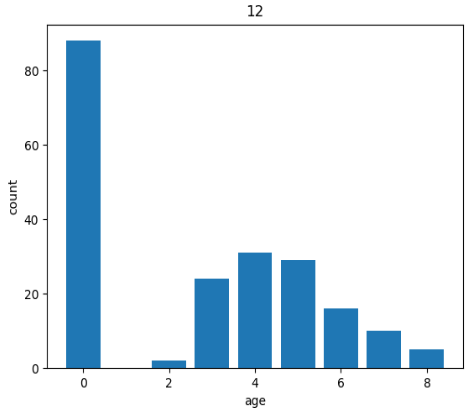

Figure 7: Age distribution in cluster 12

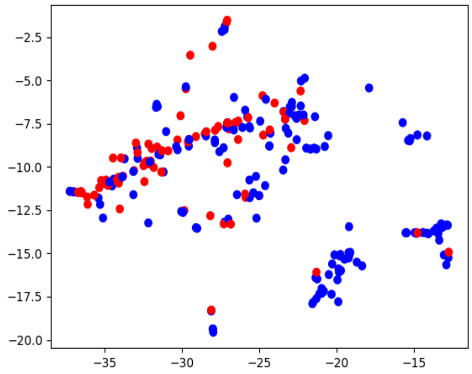

Figure 8: User scatter graph in abnormal cluster

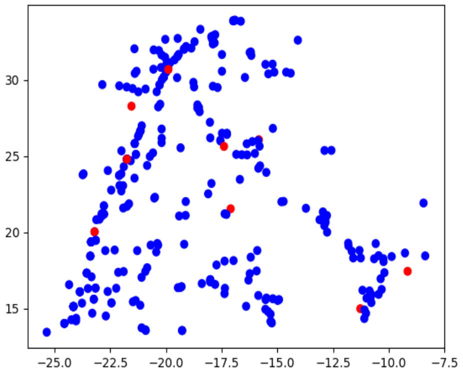

Figure 9: User scatter graph in cluster 9

In fig 4, 5, 6, the proportion of users of the same age group is very large. By querying users within these clusters, we find that user in the same cluster have the similar social behavior, so we can think that these clusters represent a social feature of a class of users in a class of age groups.Fig 7 presents the age distribution in an abnormal class of clusters, it can be seen that the proportion of users with zero age in this cluster is very large, and there are some other age groups in the cluster.

According to fig 5 and 7, we extract the distribution of the exception class cluster and the class cluster 9 such as fig 8 and 9 .The red point in the graph is user with the age attribute of zero, and the blue point is the user with normal age attribute. By contrast, it is found that the users with normal age attributes in fig 8 have common characteristics with the abnormal users of zero age, so the social relationships of the users in these ages are abnormal and are the anomaly points to be detected.

3.2 Comparison result

In this paper, we use recall and precision to verify the accuracy of outlier mining algorithm.which are given as follow:

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{5}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{6}$$

TP represents the number of abnormal users judged correctly in the test. FN represents the number of abnormal users in the test being judged to be the normal point in the test. FP represents the number of normal users in the test as the number of abnormal points, and TN is judged to be the number of different users in the test.

According to the above formula, we compare the following table of three evaluation indexes using KMEANS based on social relationships(KS),using KMEANS based social and spatial location(KSS), and using DBSCAN based on social and spatial location(DSS) as follows.

Table 1: Performance Static of KS, KSG, DSG

Experiment result	Recall	Precision
KS	22.64%	92.45%
KSS	98.53%	96.18%
DSS	37.73%	83.33%

4. Conclusions

In this paper, we propose an outliers mining method based on social and spatial information. The method increases the sample space by increasing the dimension of the feature matrix of the data set, so that the object can be accurately divided in the clustering process. By using the user's one week call record as the data set, we model the two dimensions of the user's social and spatial information, and use the KMEANS method to cluster the data. After clustering, the user composition of each cluster is analyzed, and the abnormal class clusters are found according to the user with zero age and the final outliers are determined. By comparing with different method, we prove that the algorithm is feasible to judge the outlier under the condition of sufficient data. The experimental results show that we can not only get the outliers in the dataset, but also judge the significance behind these outliers. It can also be used to determine the characteristics of a group of people in the data set to determine the missing attributes of the outliers, and to complete a precise personalized service for the user.

5. References

[1] Chandola V,Banerjee A,Kumar V,Anomaly detection: A survey[J],ACM Computing Surveys,2009,41(3) : 1-58.

[2] Gogoi P,Bhattacharyya D K,Borah B,et al. A survey of outlier detection methods in network anomaly identification[J]. Computer Journal, 2011,54(4) : 570-588.

[3] Yamanishi K,Takeuchi J I, Williams G, et al. On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms.[J]. Data Mining & Knowledge Discovery, 2004, 8(3):275-300.

[4] Gogoi P,Bhattacharyya D K,Borah B,et al. A survey of outlier detection methods in network anomaly identification[J]. Computer Journal, 2011,54(4) : 570-588.

[5] Kaur G,Saxena V,Gupta J P. Anomaly detection in network traffic and role of wavelets[C]//International Conference on Computer Engineering and Technology. Chengdu: IEEE, 2010: 46-51.

[6] Phua C,Lee V,Smith K,et al. A comprehensive survey of data mining-based fraud detection research[J]. Artificial Intelligence Review,2010, 21(3) : 60-74.

[7] Gupta M,Gao J,Aggarwal C C,et al. Outlier detection for temporal data: A survey[J]. IEEE Transactions on Knowledge and Data Engineering, 2014,26(9) : 2250-2267.

[8] Zimek A, Schubert E,Kriegel H P. A survey on unsupervised outlier detection in high-dimensional numerical data[J]. Statistical Analysis and Data Mining the Asa Data Science Journal,2012,5(5) : 363-387.

[9] Aggarwal C C, Outlier analysis[M], New York: Springer,2013: 50-89.

[10] X. Luo,Detection of Outliers Based on Social Relations and Geographical Location,Master,Jilin University,Changchun,Jilin,China,2018.