Clustering-based KPI Data Association Analysis Method in Cellular Networks

Xingyu Guo, Peng Yu, Wenjing Li and Xuesong Qiu
State Key Laboratory of Networking and Switching Technology,
Beijing University of Posts and Telecommunications,
Beijing, 100876, P. R. China
Email: guoxingyu@bupt.edu.cn

***Abstract*— With the rapid development of cellular network systems, the operators need more experience to deal with complicated network management system and wide range of Key Performance Indicators (KPIs). There are many indicators related to each other due to the definition or communication process. But several implicit associations still exist among these KPIs. This paper proposes an approach to figure out the implicit linear relationship among indicators clearly in which a new clustering technique is used for distinguishing different relationships. Data analysis using real network data shows that the approach can well divide data into clusters, and each cluster can effectively reflect the relationship between indicators.**

Keywords—Key Performance Indicators; multi-linear relation; data analysis; clustering;

I. Introduction

In order to achieve high throughput, seamless coverage and quality of service, modern cellular communication system is becoming increasingly complex and heterogeneous [1]. And the scales of network Key Performance Indicators (KPIs) data which reflect the network performance and status intuitively increase as well. It makes the management and optimization of cellular network more complex and difficult as well. Therefore, deep mining the KPI data and figuring out the implicit association among indicators has reference value for operators to optimize the network [2].

KPIs are used to evaluate the network performance and network operation quality. There are many indicators defined by ratio. This part of KPIs is non-neutral and most of these represent success or failure rate. For instance, successful RRC connection setup rate is defined as (successful RRC connection setup frequency) / (RRC connection setup request frequency)*100%, and successful wireless connection setup rate is defined as (successful RRC connection setup rate)*(successful RAB establishment rate). Here RRC means Radio Resource Control and RAB means Radio Access Bearer. We can intuitively discover their association by definition. But the implicit associations of some indicators are caused by communication setup procedures or environmental factors. Thus, the implicit association is one of research hotspot in cellular network KPI data analysis.

Recently, there are many literatures carried out in the area of cellular network operating data analysis. For instance, the work based on the classification [3], clustering [4] and anomaly detection method [5] are focused on self-healing especially for Cell Outage Detection (COD), which is a function of Self-Organizing Networks (SON). But they are mainly using the data in measurement reports which are reported to NodeB or eNodeB by user-side. These data cannot be accessed by network management system and hard to be collected. In [6], the researcher proposed an ensemble method for anomaly detection in cellular network, which is KPI-based and use many time-series analysis method. But the method does not consider the association information for indicators, which may impact the result. Still, associations among KPIs are always linear which is easier to be analyzed.

The purpose of this paper is to help cellular network operators to figure out the linear relationship among indicators and identify the possible reason of network degradation. One indicator deteriorating may be caused by other indicator, which possibly arising from the user-side problems, or the network environment. A density Clustering algorithm named CFSFDP (Clustering by Fast Search and Find of Density Peaks) [7] is used to find different models of the association among indicators, but it’s not used in cellular networks for KPI analysis.

To resolving above problems, the work of this paper is described as follows.

a) We analyze the cellular network KPI data and summarize their characteristics, which are the theoretical supports for our following work.

b) We distinguish the different characteristics of the data with the CFSFDP algorithm. In each cluster, it may contain the implied association between indicators.

c) We use regression method to fit each cluster data with a linear model to obtain the relationships among indicators.

The rest of the paper is organized as follows. Section II introduces the characteristics of the cellular network KPI data, and discusses why other method do not fit the KPI data; section III presents the clustering-based approach to distinguish between different data association using CFSFDP; the result of our approach is then discussed in section IV; section V summarizes this paper and describes future work.

II. The Characteristics of KPI data

A. Time-Series analysis

KPI data are real-time collected as ordered sequences by network management system. Typically, the indicators include connection success rate, call drop rate, congestion rate,

978-1-5090-0223-8/16/$31.00 © 2016 IEEE

handover success rate and other indicators. They have the typical characteristics of time series and spatial information as each KPI record containing a record time stamp and its base station cell.

As a time series, the KPI data may be analyzed with known methods for time-series analysis. But they may not be effective. As we have said in Section I, many of indicators are defined by ratio. Each formula element has the feature of time-series, but their ratio may not have such feature.

Fig. 1 shows two days data which Y-axis is the successful PS (Packet Switch) connection setup rate. From Fig. 1, we can see that its values are highly random. Thus, the KPI data do not suit for time series analysis.

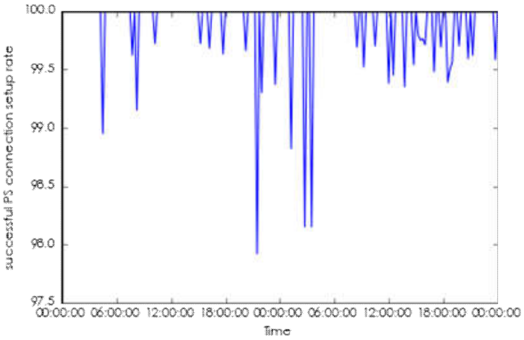

Fig. 1 the KPI data's time-series distribution

B. The correlation among indicators

In statistics, the Pearson product-moment correlation coefficient (sometimes referred to as the PPMCC or PCC or Pearson's r) is a measure of the linear correlation between two variables X and Y, giving a value between +1 and −1 inclusive, where 1 is total positive correlation, 0 is no correlation, and −1 is total negative correlation. It is defined as follow.

$$\rho_{X,Y} = \frac{\mathrm{cov}(\mathrm{X},\mathrm{Y})}{\sigma_X \sigma_Y} = \frac{E[(\mathrm{X}-\mu_X)(\mathrm{Y}-\mu_Y)]}{\sigma_X \sigma_Y} \quad (1)$$

This coefficient can find some obvious linear relationships among KPIs. But calculating the correlation coefficient directly is still a problem when linear relationships are not clear.

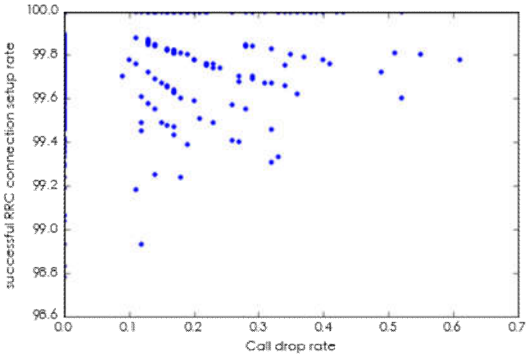

Fig. 2 Scatter plot of Call drop rate-successful RRC connection setup rate scatter plot

Fig.2 and Fig. 3 show scatter plots of different indicators'. In the two figures, part of data is strong correlative. But when calculating their correlation coefficient, we find that the correlation of indicators in Fig. 2 is -0.02619 and the correlation of indicators in Fig. 3 is 0.5836.

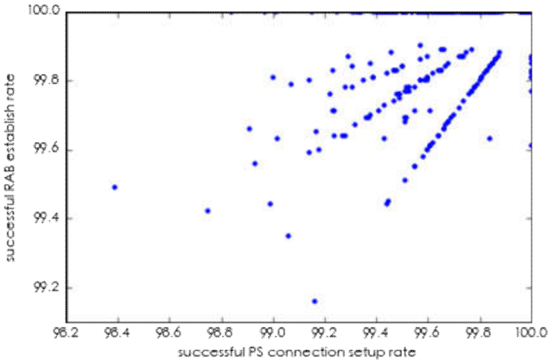

Fig. 3 Scatter plot of successful PS connection setup rate- successful RAB establish rate

III. Algorithm Description

According to the analysis above, we propose a clustering-based approach using CFSFDP to divide the data into different clusters and in each cluster the data point is strongly correlative. Then using regression method, we fit a linear equation within the same cluster to clear the relationship between indicators.

A. Clustering the data

Considering a strong linear association between the indicators, the data points are congregated in the vicinity of a linear function. And these data points have high density. Thus we use density clustering method to discover each linear relation among the indicators. Classical density-based clustering methods such as DBSCAN[8] are highly sensitive to the user-defined parameters. Slight different parameters may lead to different results, and selection of parameters has no rules to follow. Thus they can only be determined empirically.

CFSFDP [7] does not require iteration, and the method uses a graph named Decision Graph and dependent on the data point itself, which is suitable for KPI analysis.

The algorithm has its basis in the assumptions as:

- ➢ Cluster centers are surrounded by neighbors with lower local density;
- ➢ Cluster centers are at a relatively large distance from any points with a higher local density;

For Fig.3 and Fig. 4 above, we can derive that the high local density points have a close relationship.

For each data point x_i, we compute two quantities: 1) its local density ρ_i and 2) its distance δ_i from points with higher density. Both these quantities only depend on the distances d_i between data points, which are assumed to satisfy the triangular inequality.

We consider $S = \{x_i\}^N$ as the data set and N is the number of data; $I_S = \{1, 2, \cdots, N\}$ as the index of data set; $d_{ij} = dist(x_i, x_j)$ as the distances between x_i and x_j.

The local density ρ_i is defined as follows.

$$\rho_i = \sum_{j\in I_S\setminus\{\mathrm{i}\}} e^{-(\frac{d_{ij}}{d_c})^2} \tag{2}$$

The variable $d_i > 0$ is the only parameters to be set, which is named cutoff distance. It is specified by user. In order to avoid the sensitive on parameters, we use percentage instead of the specific value. In this paper, d_i is set to 3%.

The distance δ_i is defined as follows.

$$\delta_i = \begin{cases} \min\limits_{j\in I_S^i}(d_{ij}), i \geq 2 \\ \max\limits_{j\in I_S}(d_{ij}), i = 1 \end{cases} \tag{3}$$

When the x_i has maximum local density, δ_i represents the greatest distance between x_i and others data points in data set *S*; otherwise δ_i represents the minimum distance between x_i and others data points that have greater local density than x_i.

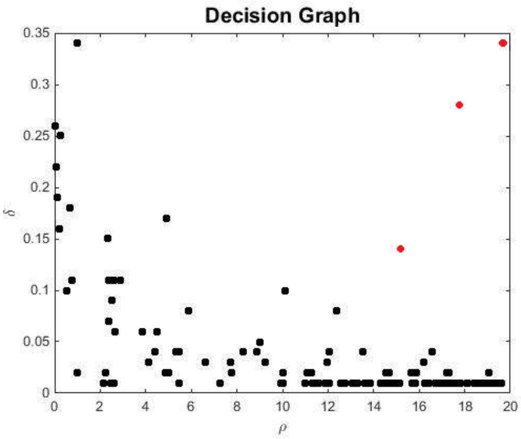

Fig. 4 Decision Graph based on local density and distance

For each data point x_i, we compute $(\rho_i, \delta_i), i \in I_S$. Based on the definitions above, we can obtain decision graph as Fig .4.

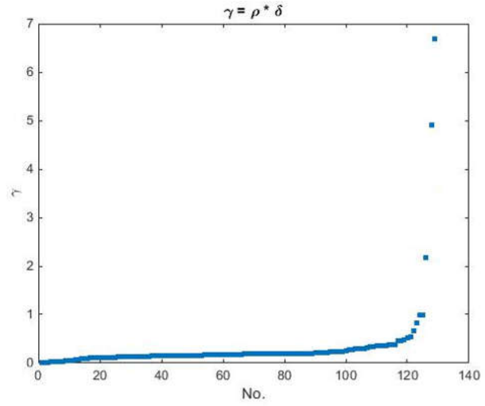

Fig. 5 Decision Graph sorted by γ_i

According to the assumption of this section beginning, on the right and up points (like the red points) of the decision graph are chosen as the cluster center. But because this choice is subjective, we use a variable γ_i to determine which data point may be cluster center. It's defined as follows.

$$\gamma_i = \rho_i \delta_i, i \in I_S \tag{4}$$

Obviously, greater value of γ_i is more likely become cluster center. Thus, descending γ_i, we can find that its value has a significant jump. We use this feature to determine the number of cluster and cluster center point. The sorted value of γ_i is showed in Fig. 5. According to sorted value of γ_i, the number of cluster should be 3 or 4.

After determining the cluster centers, the other data point which is non-cluster center is assigned to one of cluster. On the basis of descending traverse local density ρ_i, we assign each data point to a cluster center which has greater local density and nearest distance. In this way, clusters expand layer by layer from the cluster center.

B. Fit the data in same cluster

After clustering, we divide the data into clusters. As our assumptions, the data in same cluster distribute in the vicinity of a line. Using the linear regression method, we fit the data and find out a clear linear relationship.

Assumed the fitting equation is, we use linear least squares to calculate the parameters as $y = bx + a$.

$$\begin{cases} b = \dfrac{\sum\limits_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum\limits_{i=1}^{n}(x_i - \overline{x})^2} = \dfrac{\sum\limits_{i=1}^{n} x_i y_i - n\overline{x}\overline{y}}{\sum\limits_{i=1}^{n}(x_i^2 - n\overline{x}^2)} \\ a = \overline{y} - b\overline{x} \end{cases} \tag{5}$$

The results are analyzed in the next section.

IV. Result Analysis

Our experimental corpus consisted of a KPI dataset containing data from a TD-SCDMA network. For each cell, 103 KPIs were collected every 15 minutes for a week, from 04/01/2015 to 04/07/2015. We use data in the cell level to discover the relevance. Here we analyze the data in one cell every single time in order to avoid the differences between different cells. These data together cause an increase of randomness, thereby adding errors. Here, we show three common association in our analysis.

A. Several relations

For cell drop rate and successful RRC connection setup rate, we do not consider them exist relationship subjectively. However, according to the data, we can figure out the relation as shown in Fig 6.

The colors of points represent the clusters they belong to. The lines represent the fitting results in each cluster. As Fig. 7 shown, linear relationship in green and yellow points indicates that the successful RRC connection setup rate and call drop rate is not related to each other; the relationship of red and blue points can be obtained by fitting a significant linear function.

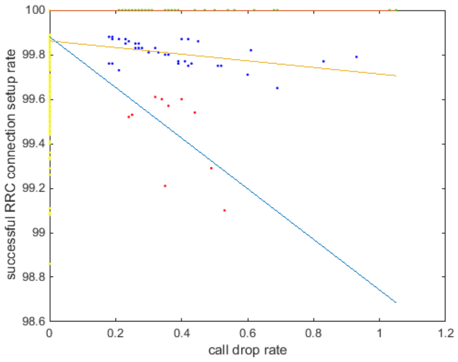

Fig. 6 The relationship between call drop rate and successful connection rate

Typically, the red points corresponding fitting equation are similar to the relation $y = -1x + 100$. Thus there is an inverse relation between dropped calls drop rate and successful RRC connection setup rate. The blue points is obvious different with the red point. They show different network status and dividing the data also means dividing the network status that is help for the analysis.

B. Single relation

For HSDPA channel abnormal release rate and HSDPA drop rate, we believe that there is relevance between them to their communicate meaning: when HSDPA channel abnormal release, it leads to lose HSDPA connection. We can also find this apparent rule through the data in Fig. 7. The slope of the red line equation is 1.358, greater than 1, which means HSDPA drop rate is higher than the abnormal release rate. We conclude that may be other causes lead HSDPA connection dropped besides abnormal release.

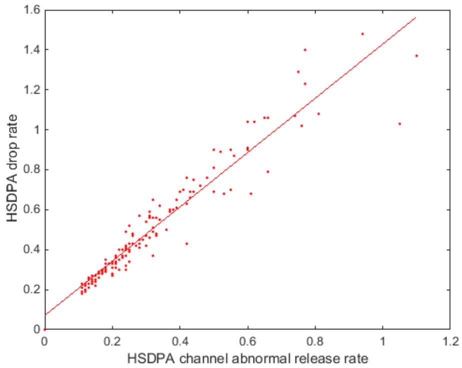

Fig. 7 The relationship between HSDPA channel abnormal release rate and HSDPA drop rate

C. Non-relation

For PS drop rate and PS retransmission rate, we believe that there may be an association between them. Because when there are dropped PS connections, there is necessity for re-sending the data. However, the real data shows that their associated is small in Fig. 8: the blue points are fitted by line approximately parallel to the x-axis, which means that the relation may not exist.

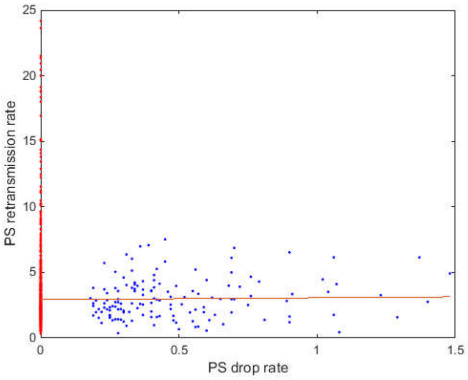

Fig. 8 The relationship between PS drop rate and PS retransmission rate

V. CONCLUTIONS AND FUTURE WORK

This paper proposes an approach for clustering the cellular network KPI data for finding implied multiple relations with CFSFDP. Real KPI data analysis results verify the effectiveness of the approach.

By data validation above, we find that some of these relations have significant meaning in the communication priori knowledge. But others have unexpected association according their meaning. Figuring out the implicit associations in data are helpful to locate the root cause of network degradation.

We are currently working towards experimental evaluation of our approach to deal with high-dimensional data and non-linear associations. In the future, additional work is needed to preprocess the data to avoid the curse of dimensionality.

ACKNOWLEDGMENT

This work was supported by 863 Program under Grant No.2015AA01A705 and the National Natural Science Foundation of China under Grants No. 61271187.

REFERENCES

[1] Yu, Peng, et al. "Dynamic multi-stage Energy-Saving Management mechanism based on Base Station cooperation." In IEEE 2013 9th International Conference on Network and Service Management (CNSM), 2013, pp. 193-197.

[2] Aliu, Osianoh Glenn, et al. "A survey of self organisation in future cellular networks." IEEE Communications Surveys & Tutorials, 2013, 15(1): 336-361.

[3] Xue, Wenqian, et al. "Classification-based approach for cell outage detection in self-healing heterogeneous networks." In IEEE 2014 Wireless Communications and Networking Conference (WCNC), 2014, pp. 2822-2826

[4] Ma, Yu, et al. "A dynamic affinity propagation clustering algorithm for cell outage detection in self-healing networks." In IEEE 2013 Wireless Communications and Networking Conference (WCNC), 2013, pp. 2266-2270.

[5] Zoha, Ahmed, et al. "Data-driven analytics for automated cell outage detection in Self-Organizing Networks." In IEEE 2015 11th International Conference on the Design of Reliable Communication Networks (DRCN), 2015, pp. 203-210.

[6] Ciocarlie, Gabriela F., et al. "Detecting anomalies in cellular networks using an ensemble method." In IEEE 2013 9th International Conference on Network and Service Management (CNSM), 2013, pp. 171-174.

[7] Rodriguez, Alex, and A. Laio. "Clustering by fast search and find of density peaks." Science 344.6191, 2014:1492-149.

[8] Ester, Martin, et al. "A density-based algorithm for discovering clusters in large spatial databases with noise." In KDD. Vol. 96. No. 34. 1996