Anomaly Detection Framework for Tracing Problems in Radio Networks

Jussi Turkka, Tapani Ristaniemi
Department of Mathematical Information Technology
P.O.BOX 35, Jyväskylä University
Jyväskylä, FI-40014, Finland
jussi.turkka@jyu.fi, tapani.ristaniemi@jyu.fi

Gil David, Amir Averbuch
School of Computer Science
Tel-Aviv University
Tel-Aviv, 69978, Israel
amir@math.tau.ac.il, gil.david@yale.edu

Abstract—**This paper shows a novel concept of using diffusion maps for dimensionality reduction when tracing problems in 3G radio networks. The main goal of the study is to identify abnormally behaving base station from a large set of data and find out reasons why the identified base stations behave differently. The paper describes an algorithm consisting of pre-processing, detection and analysis phases which were applied for RRC (Radio Resource Control) connection data gathered from the live radio networks. The results show that the proposed approach of using dimensionality reduction and anomaly detection techniques can be used to detect irregularly behaving base stations from a large set of data in a more self-organized manner.**

Keywords - radio network optimizatio; data mining; anomaly detection; self-organizing networks.

I. INTRODUCTION

Mobile phone service providers need to gather excessive amounts of information from the network to be able to optimize the network performance and solve several kinds of problems in the network operation. Data gathering and analysis can be a rather resource consuming task requiring a lot of manual work, specialized equipment and expertise. As the rapid evolution of the cellular networks and the increased capacity demands has led to a situation where the operators need to maintain several multi-vendor radio access networks (RAN) simultaneously, the burden of operating and maintaining such a complex network infrastructure has caused a need to develop more automated solutions for network deployment, operation and optimization.

Self-organizing networks (SON) and Minimization of Drive Tests (MDT) solutions are widely regarded as prominent approaches which would reduce the operational expenditures and at the same time improve the perceived end-user quality-of-service (QoS). Both approaches are also actively researched by the biggest network vendors and operators in the 3rd Generation Partnership Project (3GPP) making possible the solutions for network deployment automation.

The target of the MDT work item in 3GPP is to define a set of measurements and measurement reporting procedures which would help operators to gather more data from the network without excessive manual drive tests [1]. The MDT study targets to monitor and detect coverage problems in the network such as coverage holes, weak coverage, pilot pollution, overshoot coverage, coverage mapping and UL coverage, to name a few, as described in more details in [2].

On the other hand, the target of the SON work item in 3GPP is to define the necessary measurements, procedures and open interfaces to support self-configuring, self-optimization and self-healing use cases, which can dynamically affect the network operation, and therefore, improve the performance and reduce the manual operation efforts [3]. The SON use cases in [3] targets to coverage and capacity optimization, energy savings optimization, interference reduction, automatic configuration of physical cell identity, mobility robustness optimization, mobility load balancing optimization, random access channel optimization, automatic neighbor relations configuration, and inter-cell interference coordination.

Both SON and MDT use cases describe many measurement quantities. This creates a challenge, which is currently overlooked in the literature: how to effectively post-process the measurement data of *huge volumes*? In 3GPP this is left to be a vendor specific solution. Good introduction to the network monitoring and troubleshooting is found in [4] and comparison of different network monitoring tools is done in [5]. A simple and traditional method to monitor networks is to define a problem, a measurable performance indicator (KPI) and a pre-defined threshold, which indicates whether or not the problem exists [4]. However, there are several challenges if the network performance monitoring is done in this way. Firstly, the complexity of the network infrastructure results in many problems, and therefore, the number of performance indicators increases. This results in large databases and makes the identification of the fault/KPI-associations a much more complex task. Secondly, since the networks are complex and dynamic in nature, it is not straightforward at all to define what is a good threshold or what should be measured and how often. If one has several performance indicators, what is statistically the best one to reveal a certain feature in network behavior? Thirdly, due to the dynamic nature of the networks there is a strong need for predictability of the network behavior, which calls for advanced processing of the network performance data.

In this article we are addressing these challenges by proposing a novel algorithm to analyze a large and complex set of network performance data. The analyzed data is gathered from a live 3G network using trace functionality [6]

ISBN:978-1-61208-113-7

and it includes connection related measurement data. The subscriber and equipment traces provide very detailed information at call level on specific mobiles allowing advanced monitoring and optimization operations i.e., root cause analysis in troubleshooting, optimization of resource usage and QoS, and radio and core network end-to-end call procedure validation [6]. The proposed algorithm consists of four main phases, which are pre-processing, dimensionality reduction, detection of anomalies and post-processing the results for root cause analysis.

This paper is organized as follow. Section II describes the data, the goal of the data mining procedure and pre-processing of the data in more details. Section III describes the data mining techniques. Section IV summarizes the algorithm and shows the results of the analysis. Finally, sections V and VI discuss the future improvements of the algorithm and draw conclusions.

II. Pre-prosessing

A. *Motivation for Database Processing*

The goal of the study was to find a way to analyze the given data set and find a set of base stations which did not behave similarly compared with the regular behavior of the whole set of the base stations. It is worth of noting that no assumptions were made about the normal behavior. Instead, the measured statistics were used to classify the behavior of the base stations assuming that most of them worked as supposed. Moreover, besides finding the anomalous base stations, the target was to find out reasons, why certain base stations are different. Such an approach would make it possible to create a problem pattern data base which can be used to detect malfunctioning base stations in more self-organized manner. In the beginning, this requires manual efforts and experts knowledge in problem classification. However, when the problem pattern data base exists, it can make the problem detection less time and resource consuming in the long run. This makes the networks more profitable reducing the manual efforts of tracing the problems.

B. *Data Description*

The database was gathered from a foreign operators 3G network and it consists of call session data. Each sample consisted of 72 features with different kind of data types i.e., call duration, UE type and capability, connection related parameters, last visited sector and site. Samples were collected over the area of several radio network controllers (RNC) and 335 base station sectors. The total number of measured samples was 42215 but the samples were collected over a rather short time period of 8 minutes, which can have an effect to the statistical reliability of the analysis. Eventually, only the cells, which had a large number of connection attempts during the 8 minutes duration, were analyzed, and therefore, some of the cells were not included to the analysis at all.

The pre-processing phase of the dataset consisted of several steps. First, many of the original 72 features were filtered out from the dataset because there were many missing samples and it was unclear how to process the missing features. Second step was to choose a subset of features and convert the data to a form, which can be analyzed with the chosen data mining techniques. The chosen three nominal features were RRC establishment reason, RRC release cause and failure source. Each of the features were categorical consisting of several nominal attributes i.e., RRC establishment reason can be detach, registration and either originating or terminating certain connection type such as a low priority connection as seen in Table I. There were several other categorical features, which were not included to this study such as connection detail, UE capabilities, used release version or RRC success.

RRC success was left out because it plays a heavy role in the traditional problem solving strategy and one of the intentions was to find out whether or not the algorithm can provide similar and comparable results compared with the traditional root cause analysis. The process for the manual problem solving is explained briefly in the following.

First, an engineer filters the data and studies only the cells where the failure ratio exceeds the certain threshold e.g., the cells which have the poorest RRC connection success ratio. Next step is to choose the cell and study only the calls which failed. Based on the failure source, detail, establishment cause and release cause, the engineer gets an idea what might be wrong. Usually at this point, the engineer has knowledge about the cause of the error and they can access in more detailed log files to trace the error source. For the same reason, the above-mentioned features were chosen to be the most interesting for detecting anomalies in this study. In some cases, the mobile type is also a useful feature to trace the root cause of the problems i.e., the software related bugs in mobiles can cause problems and unwanted behavior that is not related to the problems in RAN elements at all.

C. *Data Pre-processing*

Data pre-processing phase consists of data selection, data filtering, and data transformation to a form which can be analyzed with the data mining algorithms. In the data selection, last cell id and three categorical features were selected for the analysis and rest of the 72 features was filtered out. Because the target was to detect cells which were anomalous and because the difficulties in analyzing the categorical attributes, the data was transformed to a numerical form which can be processed easily and effectively with the data mining algorithms. It is worth of noting that there are ways to analyze mixed data bases consisting of nominal and numerical features as explained in [7]. However, the numerical form of data was preferred in this study. The structure of the filtered database is shown in (1)

$$x_l=[\ a,\ b,\ c\ ], \qquad (1)$$

where the variable x_l is the lth RRC connection attempt recorded during the measurement. the variables a, b and c are categorical features indicating: the RRC connection *establishment reason* with 15 possible nominal attributes, the

ISBN:978-1-61208-113-7

RRC *release cause* with 35 possible nominal attributes and the *failure source* with 10 possible nominal attributes.

In data transmission from categorical data to numerical data, a vector for each feature was created with a length corresponding to the number of possible attributes. Furthermore, the three vectors were merged to one high dimensional vector x_{bs} which length is 60. The x_{bs} is the database sample for base station *bs* showing the count, how many times certain RRC connection establishment, released or failure attribute was measured during the 8 minutes measurement period. The structure of the x_{bs} is shown in (2)

$$x_{bs}=[\, a_1 \ldots a_i,\ b_1 \ldots b_j,\ c_1 \ldots c_k \,], \tag{2}$$

where the variable a_i is the count for how many times the *i*th categorical attribute was present during the measurement period for the categorical feature *a* e.g., RRC connection establishment reason. The variables b_j and c_k are the counts for how many times the *j*th and *k*th categorical features were present for the original categorical features *b* and *c*. The features are listed in Table I.

TABLE I
FEATURE LIST

RRC Establisment Reason	
a1) intr_rat_cell_re_select	b16) radio_link_failure
a2) orig_low_prior_signal	b17) synchronization_failure
a3) registration	b18) srnc_relocation
a4) term_low_prior_signal	b19) orig_background_call
a5) orig_background_call	b20) unspec_failure
a6) orig_high_prior_signal	b21) orig_streaming_call
a7) orig_conversational_call	b22) no_resp_from_rlc
a8) term_high_prior_signal	b23) orig_interactive_call
a9) term_conversational_call	b24) iuv_iu_rel_comm_received
a10) detach	b25) rrc_conn_req_nack
a11) orig_streaming_call	b26) call_re_establishment
a12) orig_interactive_call	b27) rrc_dir_sc_re_est
a13) call_re_establishment	b28) physical_channel_failure
a14) srnc_relocation	b29) no_resp_from_rrc_d
a15) orig_subscribed_tra_call	b30) no_resp_from_iuv
RRC Release Reason	b31) fail_in_r_if_proc
b1) pre_emption_failure	b32) rq_ci_ip_not_supp
b2) orig_low_prior_signal	b33) synchronazion_fail
b3) registration	b34) timer_expired
b4) intr_rat_cell_re_select	b35) orig_subscribed_tra_call
b5) term_low_prior_signal	**RRC Failure Source**
b6) nwk_optimisation	c1) bts
b7) orig_high_prior_signal	c2) default
b8) orig_conversational_call	c3) rnc_internal
b9) no_error	c4) cell_reselection
b10) serv_req_nack_from_rm2	c5) radio_interface
b11) rl_setup_failure	c6) iu
b12) detach	c7) transmissio
b13) term_conversational_call	c8) ms
b14) term_high_prior_signal	c9) frozen_bts
b15) inter_system_hard_ho	c10) ciphering

However, there is one very fundamental issue, which must be taken into account in the data mining and the knowledge pattern analysis. Always when one does the data selection and transformation, the information what the data includes can change and get biased depending on the actions what are done? This must be kept in mind whenever doing the data pre-processing and analyzing the results from reliability and validity point of view.

III. DATA MINING

A. Data Mining Principles

Data mining is a process for extracting interesting, previously unknown and potentially useful information patterns from large datasets [8]. The data mining process consists of several phases being data cleaning; data base integration; task relevant data selection; data mining; and pattern evaluation [8]. Data cleaning, integration and selection are data pre-processing phases where the data is prepared for the data mining [8]. The data mining consists of several functions such as classification of the data; association of data; clustering the data; dimensionality reduction and anomaly detection to mention a few [8]. In pattern evaluation phase, the information patterns are visualized and analyzed to see how usable, novel, valid and reliable the findings are i.e., even though something interesting is found, it doesn't mean that it is already usable or useful. In this study, dimensionality reduction and anomaly detection data mining techniques were used.

B. Dimensionality Reduction with Diffusion Maps

This section describes the diffusion maps framework that was introduced in more details earlier in [7][9][10][11]. Diffusion maps and diffusion distances provide a method for finding meaningful geometrical descriptions in high dimensional data sets consisting of points in R^n where n is large e.g., 60 in this study. The diffusion maps construct coordinates that parameterize the dataset and the diffusion distance provides a local preserving metric for this data [10]. The data is parameterized by using a graph G and weight function kernel W measuring the pairwise similarity of the training set of the points in R^n [10].

If a proper kernel is used, then W can be normalized into a Markov transition matrix P and the most significant eigenfunctions of the Markov matrices provide a good low dimensional geometric embedding in a way that the ordinary Euclidean distance in the embedding space measures the meaningful diffusion metrics of the data [10]. Moreover, the diffusion distance between two points, as described in [9], reflects the geometry of the dataset as well. The expression of the diffusion distance can be given as in [11]

$$D_t^2(x,y)=\sum_{k\geq 0}\lambda_k^{2t}\left(v_k(x)-v_k(y)\right)^2, \tag{3}$$

where k is the number of the most significant eigenvectors and variable λ is the eigenvalue of the kth eigenvector. The variables $v_k(x)$ and $v_k(y)$ are the kth right eigenvectors of transition matrix for points x and y. Euclidean distance between two points in the embedded space R^k represents the distances and similarity of the same points in high dimensional space. If two points are close then the diffusion distance in (3) is small. Furthermore, the diffusion distances and the density can be used to detect anomalous base stations as explained later.

C. *Anomaly Detection*

In anomaly detecting, samples can be either clustered to several categories as in [7][10] or to normal and abnormal samples. The diffusion distance metric in (3) and the diffusion coordinates can be used for the clustering since it reflects the similarity of the points in the original high dimensional space. If the density of a certain point is large in the embedded space then it has many neighbors nearby and it is considered to be regular. In contrary, irregular points have small density. For each point in the embedded space, a k-ball with certain radius is defined and the density is calculated by using any normalized density function based on the samples which lay inside the ball. Since the scale of the each diffusion coordinate in the embedded space is different, the ball radius is scaled as well. The density d_m of the mth point is defined in (4)

$$d_m=\frac{\eta_m}{\sum_{i=1}^{M}|\eta_i|}, \tag{4}$$

where η_m is the number of points inside the ball and the sum in the denominator is norm-1 over all M points.

In this study, the difference between the normal and the anomalous sample was based on the statistical properties of the density distribution of the dataset. A point was abnormal in case the density was smaller than the threshold shown in (5)

$$\mu_d-2\sigma_d, \tag{5}$$

where variables μ_d and σ_d are the median value and the standard deviation of the point density distribution.

IV. Results

A. *Summary of the Algorithm*

The used algorithm consists of the following steps:

1. *Dataset pre-processing*: The dataset feature selection and transformation resulted in an input matrix size of 129x60 showing the count of how many times nominal attributes were present for the selected three categorical features.

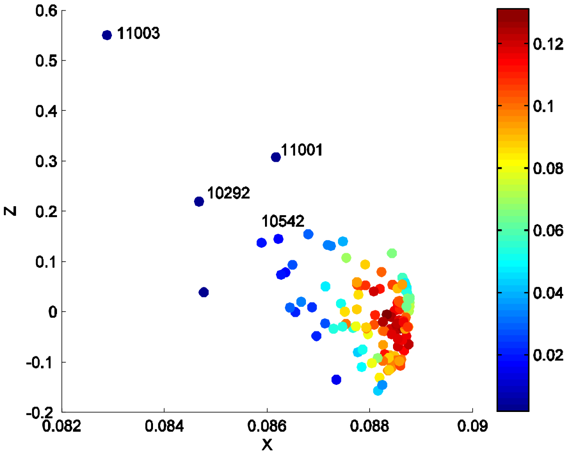

Figure 1. Base station dataset visualization in 3D embedded space.

2. *Dimensionality reduction*: Diffusion map framework was used to reduce the dimensions of the input matrix to size of 129x12 assuming that 12 eigenvectors represents the relevant data for 129 cells. The used pairwise distance metric was L2 and the diffusion epsilon factor 8. The decaying of eigenvalues proof that the findings can be visualized reliably in small dimensional space.
3. *Anomaly Detection*: Difference between normal and abnormal samples was made based on the point densities in the embedded low dimensional space according to (4).
4. *Root Cause Analysis*: An analysis of the reasons why the abnormal base stations are different from the regular ones was made based on the statistical properties of the abnormal samples in high dimensional space after the anomaly detection.

Figure 1 shows the visualization of the data in embedding space using the three most significant eigenvectors. The coloring of the points shows the measured density as described in (4). The dark blue points are irregular samples. There were totally 14 irregular base stations detected. The detection accuracy was rather high compared with the detection strategy descried in Section II which would be based on the observation of RRC connection failure ratio.

For four sectors out of 14, the RRC failure ratio was higher than 5% and this can already trigger an optimization in the network. On the other hand, not all the problems in the network are caused due to the unsuccessful calls, and therefore, it is interesting to study why some of the cells with good RRC success ratio are anomalous. It should be noted that the statistical reliability of the above mentioned observation may require more samples per base station than the calls recorded during the 8 minutes interval.

Moreover, what is the root cause of the problem and how to optimize the network based on the findings? The root cause analysis was done by comparing the 60 features of the irregular base stations one by one to the features of the regular base stations. The features with largest difference

TABLE II
ABNORMAL SECTOR PROPERTIES

Cell ID	11003	11001
Failure %	49.4%	43.8%
Calls	2040	937
Feature 1	a) intr_rat_cell_re_select	a) registration
Feature 2	a) registration	a) term_conversational_call
Feature 3	b) pre_emption_failure	b) pre_emption_failure
Feature 4	b) no_error	b) no_error
Feature 5	c) bts	c) bts
Feature 6	c) cell_reselection	c) cell_reselection
Cell ID	10292	10542
Failure %	6.5%	5.9%
Calls	107	153
Feature 1	b) no_resp_from_rlc	b) synchronization_failure
Feature 2	b) physical_channel_failure	b) rrc_conn_req_nack
Feature 3	c) radio_interface	c) frozen_bts
Feature 4	c) ms_c	-

compared with the regular behavior in all cells was used as an indicator to select which features are different i.e., possibly causing the anomalous behavior.

B. *Results of Base Station Detection*

Table II shows the results of four abnormal sectors with the high RRC failure rates. Sectors 11003 and 11001 had a critically bad RRC success ratios and a high number of call attempts during the 8 minutes period. Both cells seem to be rather congested and there might not be enough resources to create connections. The *pre-emption* property is related to releasing core network resources for creating radio access bearer connections [12]. Therefore, a large amount of pre-emption failures can indicate a poor capacity planning or lack of network resources. However, to find the root cause and a solution to the problem, a deeper analysis of the sector behavior would be needed to verify these assumptions.

Sector 10292 had as well a rather poor RRC success ratio. However, the number of call attempts was small and therefore the total number of failed calls was small. In this sector, the anomalies were due to the number of calls which had events *no_resp_from_rlc* and failures in *radio_interface*. In sector 10542, the anomalies were due to the number of calls which had events *rrc_conn_req_nack* and failures in *frozen_bts*. The details of the other irregular cells are not listed here since the three most abnormal cells already give the idea about the output of the results.

V. CONCLUSION AND FUTURE WORK

In this paper, a novel concept was introduced by using data mining techniques for tracing problems in radio networks. The data mining techniques were used to identify irregularly behaving cells from a large RRC connection dataset. The proposed algorithm consists of pre-processing; dimensionality reduction; anomaly detection; and root cause analysis, which were used with the real 3G network data. The algorithm detected base stations with high failure rate.

For future work, there are ways to improve the accuracy of the algorithm even more. Firstly, the data selection can be extended to include more features. The size of the input matrix affects to the measurement interval, and therefore, the usability of 8 minutes measurement period is to be clarified. Moreover, in this study, the anomaly detection was done in base station domain. On the other hand, the anomaly detection can be done in time domain as well. However, observing the behavior of a single base station over longer time period requires longer measurements. If longer measurements can be conducted, then the analysis can be extended to an automatic discrimination between good and bad base stations in a similar way the discrimination is done in [13]. Furthermore, the root cause analysis can be improved as well. In this paper a simple approach was chosen to indicate what high dimensional features are causing the irregularities. However, an efficient way to choose the meaningful features from the data causing the irregularities is still to be studied.

ACKNOWLEDGMENT

The author would like to thank colleagues from Nokia, Nokia Siemens Networks, Magister Solutions Ltd, Jyväskylä University and the Radio Network Group at Tampere University of Technology for their constructive criticism, comments and support with the work.

REFERENCES

[1] *3GPP TR 36.805, "*Study on minimization of drive-tests in Next Generation Networks*", ver. 9.0.0, December 2009.*

[2] *3GPP TS 37.320, "*Radio measurement collection for Minimization of Drive Tests*", ver. 0.7.0, June 2010.*

[3] *3GPP TR 36.902, "*Self-configuring and Self-optimization (SON) network use cases and solutions*", ver. 9.2.0, June 2010.*

[4] *R. Kreher, "*UMTS performance measurements. A Practical Guide to KPIs for the UTRAN Environment*", Wiley, 2006*

[5] *J. Laiho, A. Wacker and S Müller, "*Measurement Based Methods for WCDMA Radio Network Design Verification*", 10th Communications and Networking Simulation Symposium, March 2007.*

[6] *3GPP TS 32.421 "*Subscriber and equipment trace; Trace concepts and requirements*", ver. 9.1.0, June 2010.*

[7] *G. David and A. Averbuch, "*SpectralCAT: Categorical Spectral Clustering of Numerical and Nominal Data*", published in SampTA 2011, Singapore.*

[8] *J. Han and M. Kamber, "Data* Mining: Concepts and Techniques*", Morgan Kaufmann Publisher, 2000.*

[9] *S. Lafon, Y. Keller and R. Coifman, "*Data Fusion and Multicue Data Matching by Diffusion Maps*", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, No.11, November 2006.*

[10] *G. David, A. Averbuch and R. Coifman, "*Hierarchical Clustering via Localized Diffusion Folders*", Association for the Advancement of Artificial Intelligence (AAAI), November 11-13, Virginia, 2010, USA..*

[11] *N. Rabin and A. Averbuch, "*Hierarcial Data mining approach for detection and prediction of anomalies in dynamically evolving systems*", in review process, January 2010.*

[12] *Su. Kasera and N. Narang, "*3G networks: architecture, protocols and procedures: based on 3GPP specifications for UMTS WCDMA networks*", Tata McGraw-Hill, 2004.*

[13] *A. Averbuch et al., "*Automatic discrimination between bad and good laser machines*", Applied Materials for the IMG 4 consortium, January 2010.*