Analyzing 3G Quality Distribution Data with Fuzzy Rules and Fuzzy Clustering

Pekka Kumpulainen[1], Mika Särkioja[2], Mikko Kylväjä[3], Kimmo Hätönen[4]

[1] Tampere University of Technology, Department of Automation Science and Engineering,
Korkeakoulunkatu 3, 33720 Tampere, Finland
[2] Nokia Siemens Networks, BSO OSS Radio Network Optimizer, Espoo, Finland
[3] Aditro Software, Espoo, Finland
[4] Nokia Siemens Networks, CTO Research, Espoo, Finland
pekka.kumpulainen@tut.fi, {mika.1.sarkioja,
kimmo.hatonen}@nsn.com, mikko.kylvaja@aditro.com

Abstract. The amount of data collected from telecommunications networks has increased significantly during the last decade. In comparison to the earlier networks, present-day 3G networks are able to provide more complex and detailed data, such as distributions of quality indicators. However, the operators lack proper tools to efficiently utilize these data in monitoring and analyzing the networks. Classification of the network elements (cells) into groups of similar behavior provides valuable information for a radio expert, who is responsible of hundreds or thousands of elements.

In this paper we propose fuzzy methods applied to 3G network channel quality distributions for analyzing the network performance. We introduce a traditional fuzzy inference system based on features extracted from the distributional data. We provide interpretation of the resulting classes to demonstrate their usability on network monitoring. Constructing and maintaining fuzzy rule sets are laborious tasks, therefore there is a demand for data driven methods that can provide similar information to the experts. We apply fuzzy C-means clustering to create performance classes. Finally, we introduce further analysis on how the performance of individual network elements varies between the classes in the course of time.

Keywords: 3G mobile network, quality variable distribution, channel quality, fuzzy clustering, fuzzy rules, unsupervised classification

1 Introduction

The amount of data collected from telecommunications networks has increased during last decade significantly [6]. The data available today is richer in detail and complexity, which creates more requirements also for automated data analysis tools [2, 10]. In this research we propose unsupervised classification methods for analyzing the quality of cells. A cell is a geographical area that one sector of a base station covers. For classification we use daily performance metric data from 3G network [4].

The channel quality distributions describe the quality of radio conditions of HSPA channel in 3G network cells. The target is to provide methods that can support radio experts in monitoring the network and to detecting problematic behavior.

In the following section we first introduce the data with examples of their characteristics. Next, we introduce a fuzzy rule bases system for classification. In section 4 we present fuzzy clustering and interpretation of the clusters as quality classes. In section 5 we propose methods to monitor the variation of the behavior of individual cells. We summarize the results in the last section.

2 Quality Variable Distributions

The data used in this study are 3G network channel quality indicator (CQI) measurements [3]. The data consist of distributions from high speed packet access (HSPA) connections in 3G network. The collected channel quality measurement samples are grouped into 31 classes, each of which represents a different quality category. These reported measurement samples are collected from every HSPA connection when mobile terminal sends channel quality report indicators, that takes place constantly during the active connection. Class 1 contains the samples that have the worst radio channel quality, providing the worst radio interface throughput for mobile. Class 31 contains samples that have the best radio quality. The measurements are used during high speed packet call for deciding the optimal packet size and modulation coding scheme. Naturally, more samples in higher quality range in the CQI statistics of a 3G cell implies higher throughput for users of the cell. Throughput is one of the most significant contributors to user experience in packet data transfer.

Counting the samples in each quality class during one day constitutes a quality histogram. Dividing the counts by the total number of samples results in proportions of the samples in each quality class; producing a daily quality distribution.

The data set consists of daily CQI measurement distributions of 771 3G cells over a time period of 15 days. Most of the cells (747) have complete data for all days. The rest have smaller subset of CQI distribution profiles available. However, all cells are included in the analysis. The total amount of daily distributions is 11441.

The original data consists of absolute amount of reported measurement samples in each CQI bin. The data is scaled so that sum of samples for a cell CQI distribution profile is one. To illustrate the differences in CQI distribution profiles, some handpicked examples of distributions are shown in Fig1.

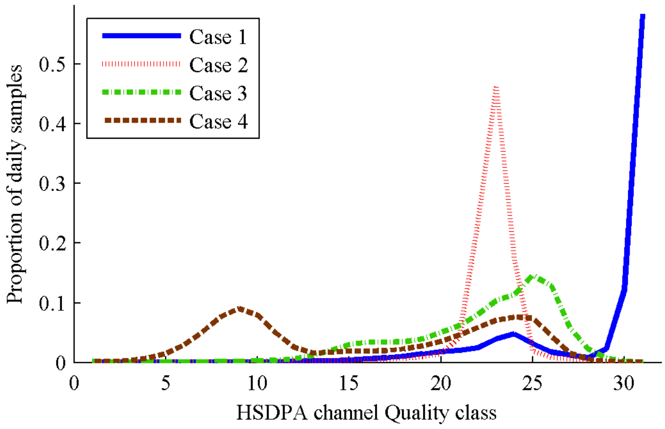

Fig. 1. Example distributions.

Case 1 is an example of a cell, which has very good radio channel conditions. Case 2 is a cell that provides satisfactory radio conditions. CQI distribution profile of case 3 is distributed to wide quality range, which most probably means that the cell is providing service for users in different channel conditions. In case 4 the users are probably located in two different locations, one of which provides good channel conditions and the other one much worse. These examples present extreme behavior and especially cases 1 and 4 are very rare in practice.

3 Quality Classification with Fuzzy Rules

Fuzzy sets were first introduced by Zadeh in 1965 [9]. Fuzzy logic has been successfully used in a wide variety of application areas ever since.

We implement a Mamdani type fuzzy interference system [6] to classify the cells according to their CQI distributions. Using the full 31 dimensional distributions as inputs would end up in very high numbers of membership functions and fuzzy rules. Therefore we extract 3 features from the distributions to present the most interesting aspects of the distributions: the average (or most common) quality and the variation of the quality during the day. The features are named: *MaxPosition*, *MaxValue* and *NLowValues*.

MaxPosition is the number of the quality class that contains maximum number of samples, thus representing the average quality of the day. *MaxValue* is the maximum value of the distribution, the proportion of all samples in the most common quality class. It represents the width of the peak in the distribution and can be thought as a confidence weight of the average quality. It is also related to the variation of the quality. *NLowValues* describes the variation of quality. It is the number of values below 1/31 in the distribution. Lowest possible value, 0 would mean equal distribution, thus maximum possible variation of the quality. The other extreme value

30 would imply that all the samples are in one single class, thus representing minimum possible quality variation.

The membership functions of the inputs are presented in Fig. 2.

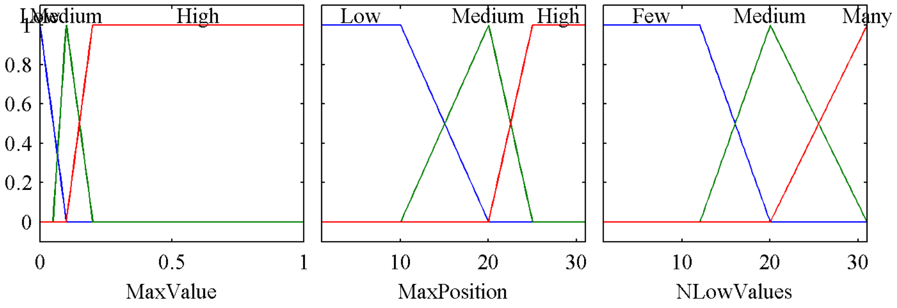

Fig. 2. Membership functions of the inputs.

We have two outputs: Quality and Variation. Quality describes the average quality of the distribution and Variation represents the stability of the quality. The Quality has four fuzzy values: Bad, Medium, Good and Excellent. The variation has outputs: Low, Medium and High. An ideally performing cell would have excellent Quality and low Variation. The membership functions of the outputs are presented in Fig. 3.

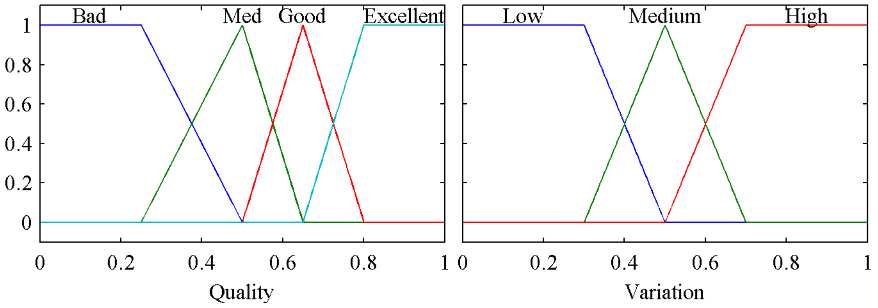

Fig. 3. Membership functions of the outputs.

We formed a total of 15 rules to map the inputs to appropriate fuzzy outputs. Some of the rules are given below as examples.

```
1. If (MaxValue is High) and (MaxPosition is High) then
(Quality is Excellent) (1)
2. If (MaxValue is High) and (MaxPosition is Low) then
(Quality is bad) (1)
5. If (MaxValue is Low) or (NLowValues is Few) then
(Variation is High) (1)
7. If (MaxValue is High) or (NLowValues is Many) then
(Variation is Low) (1)
```

```
15. If (MaxValue is Medium) and (MaxPosition is Low)
then (Quality is bad) (1)
```

The resulting classes of the outputs of the 11441 daily distributions are collected in Table 1.:

Table 1. Number of cells in the fuzzy performance classes

Quality \ Variation	Low	Medium	High
Bad	77 (0.7%)	268 (2.3%)	300 (2.6%)
Medium	981 (8.6%)	2312 (20.2%)	956 (8.4%)
Good	2923 (25.5%)	1516 (13.3%)	169 (1.5%)
Excellent	1870 (16.3%)	69 (0.6%)	0 (0%)

Most of the data have either good or excellent quality. 12.5% of the daily distributions have high variation. 0.7% have bad quality and low variation which means constant poor quality.

We present examples of the behavior in four classes in Fig 4. The selected classes are Bad and Excellent Quality with Low Variation as well as Bad and Good Quality with High Variation. All the distributions in these classes are shown as boxplots [5].

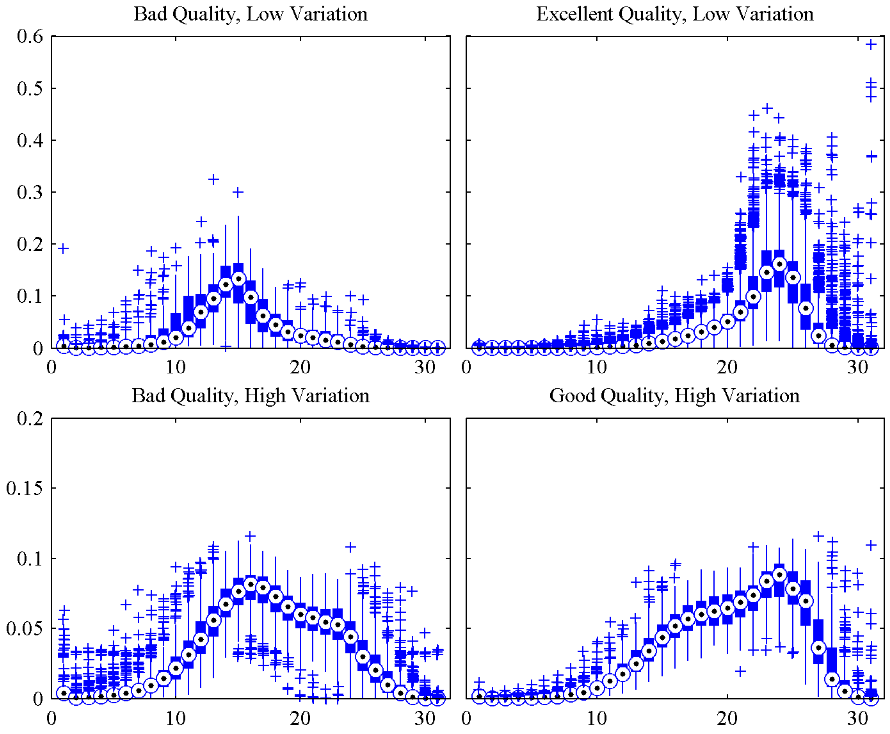

Fig. 4. Boxplots of the data in 4 classes.

These examples of four CQI cluster classes describe cluster channel quality behavior and expected end user throughput: In [Bad Quality, Low Variation] – cluster large numbers of bad samples represent low channel quality. When a cell is in this cluster end users are not able to get a high throughput. In contrary [Excellent Quality, Low Variation] - cluster channel conditions are constantly excellent. If a cell is in this class the users are likely to experience good throughput. In [Bad Quality, Low Variation] – cluster users are experiencing both bad and good channel conditions. This means that radio conditions in cells in this class extend over several quality classes. This is likely to take place. In [Good Quality, High Variation] most users are experiencing good channel conditions. Behavior is similar to previous example, but most users are having good channel conditions.

4 Quality Classification with Fuzzy Clustering

Clustering is a term for unsupervised methods that discover groups of similar observations in multivariate data [1]. Each observation is assigned into one of the discovered groups. Thus, clustering is unsupervised classification, where the classes are not known beforehand, but identified from the data. Instead of a strict assignment into one group only, fuzzy clustering allows each observation to be a member of all clusters with varying degree of membership [7]. Fuzzy C-means (FCM) clustering identifies a predefined number, C, of cluster centers so that the fuzzified sum of Euclidean distances from observations is minimized [8]. FCM produces the centers of fuzzy clusters and for each observation the degree of membership in the C clusters. In case the observations need to be assigned to one cluster, the ones where the observations have the highest degree of membership are selected. On the other hand, if the maximum degree of membership is low and does not differ significantly from the others, it is a sign that the observation does not belong well to any of the clusters and is most likely an unusual case or an outlier.

In this study we apply FCM to the 31-dimensional data of CQI distributions. After testing several values of C, the number of clusters, we decided to use 8 clusters. That is still a reasonable number of classes for the radio expert to interpret and assign labels to. Yet, that is sufficient to cover the most of the variations in the behavior. Centers of the 8 clusters are presented in Fig. 5. The legend also shows the number of daily distributions that have the maximum membership in each cluster. The data are relatively evenly distributed across the clusters.

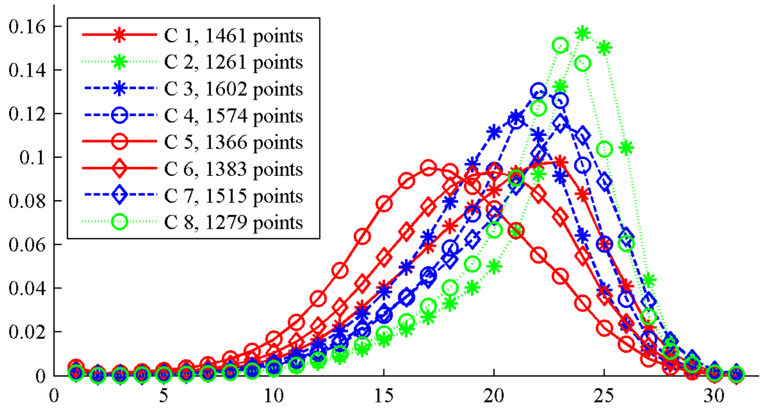

Fig. 5. Centers of 8 fuzzy clusters.

It is easy to see from the centers that the main differences between the clusters are the position of the peak value, which is the most common quality class, and the width, which describes the variation of the quality in the cluster. These two features of CQI distribution are the most relevant factors for radio expert to define the behavior of the cells in the cluster.

Clusters 1 and 4 have peak values close to each other. However, cluster 4 represents better behavior, since the variation of the quality (width of the distribution) is smaller. Cluster 1 has more samples in the range that represents poor quality classes.

Clusters 1, 5 and 6 (solid red lines) have similar shape of distribution, representing similar level of variation of the quality. However, as the location of the peak differs, they represent cells that have different quality conditions. A radio expert can tell that at least most of the variation is caused by mobiles that are located in wide range of distances around the cell.

Clusters 3, 4 and 7 (dashed blue lines) provide better channel quality for users. In other words, most probably the customers are often close to the cell.

Clusters 2 and 8 (dotted green lines) represent very well behaving cells. In these clusters it is typical to have the quality distribution biased to high quality range. Also the variation is low. This behavior is common in indoor cells. They do not experience much interference outside and users are mostly located nearby.

Altogether, it is possible for radio experts to introduce descriptions to the clusters that are similar to those generated by the fuzzy rules.

5 Tracking the Changes in Quality Behavior

Tracking how the cells change clusters in the course of time provides valuable information for analyzing their behavior. Here we assign each distribution to the cluster of the maximum membership. Some cells behave constantly, and are a member of the same behavioral cluster most of the time. Some cells are more restless and visit multiple clusters.

First we introduce statistics about the most common clusters of each cell. We count the assignments of each cell in the clusters. The one with the maximum count is the most common cluster for the cells. The optimal situation is if the cell stays constantly in a high quality cluster. Even if a cell is constantly in cluster of low quality, that is reasonably good situation, because it is easier to optimize elements that are behaving constantly. These cells are also most probably not causing much random interference to neighboring cells.

Histogram of the maximum counts is presented in Fig. 6. As the monitoring period in these data is 15 days, the maximum value of 15 means that the cell has stayed in the same cluster for the whole period. That is the case for 25 cells as depicted in Fig. 7. There are 21 cells that have the maximum value 14, thus they stay in the same cluster except for one day.

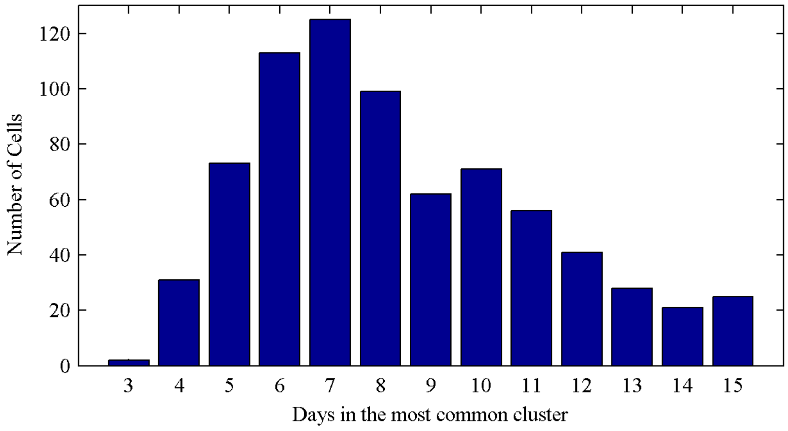

Fig. 6. Histogram of days the cells spend in their most common cluster.

Table 2 shows which clusters are the most common ones. Column 1 displays the maximum count, the days spent in the most common cluster. The rest of the columns present the number of cells that have the corresponding cluster as the most common one.

Table 2. Deviation of the cells in the fuzzy clusters.

Max value	C 1	C 2	C 3	C 4	C 5	C 6	C 7	C 8
15	1	5	1	0	13	0	3	2
14	4	7	3	2	10	7	6	3
13	4	6	9	6	15	14	7	6
12	12	19	15	15	10	14	15	16
6 - 10	287	144	260	285	143	245	249	201
3 – 5	91	50	84	88	44	75	75	64

The first row contains the cells that spend all 15 days in one cluster. Most of them (13) stay in cluster 5, which presents the poorest quality. Also most of the cells that spend 13 or 14 days in one cluster, stay in cluster 5. This is an indication that those cells have mostly poor performance and they should be checked for possible improvements.

On the other hand, the cells that have a lot of variation in their behavior, (two bottom rows) visit cluster 5 less often than other clusters. Majority of the cells, 495 never visit cluster 5.

Clusters 2 and 8 are the ones that have the best quality. There are 64 cells that stay in either of those clusters at least 12 days out of the 15. All these are performing well and require no further attention.

Further statistics on the variation among the clusters is presented in Fig. 7. It shows a histogram of how many clusters the cells visited during monitoring period.

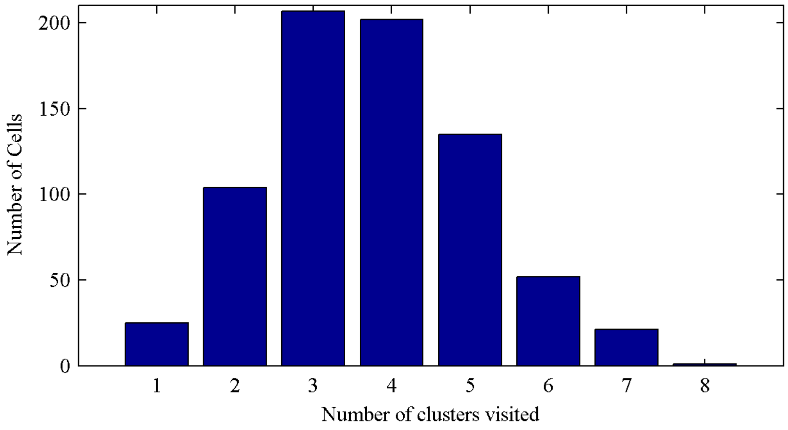

Fig. 7. Histogram of the clusters covered by cells.

The same 25 cells that stay all the days in one cluster are the ones that cove only one cluster. 104 cells split between two clusters. 21 of them are divided to 14 days in one cluster and one day in another cluster, as seen in Table 2. The rest have more even distribution across the two clusters. Only one cell spreads its behavior in all 8 clusters and 21 cells in 7 clusters.

6 Conclusion

CQI distribution measures channel conditions that change according to traffic volume and user positions. Radio experts are working with thousands of elements, where channel conditions change rapidly and traffic patterns change. Identifying behavioral patterns of cell radio interface can involve a very large number of quality samples. Data abstraction and summarization are powerful ways of supporting the expert's decision making.

Fuzzy rules and fuzzy clustering can be used to create behavioral classes. Labels of these classes can be assigned to the network elements. Fuzzy rules were used to classify the distribution data based on the quality and the daily variation of the quality. Weakness of rule based analysis is that the rules need to be designed and maintaining the rule base and membership functions is laborious thus raising a demand for automated data driven methods. We used fuzzy C-means to create similar behavioral classes by clustering the distributional data. Now, the radio experts have only a handful of clusters to interpret and associate with an appropriate label. We believe that if radio experts can identify the typical behavior of a cell, it helps them to optimize or configure the cell. The radio expert can apply similar improvements methods for cells that share similar performance behavior.

We also proposed procedures to track the stability of the behavior of the cells which is important in quality monitoring. This is based on observing how often a cell is changing from one behavioral class to another and which classes the cell is visiting.

Now we used all the data in the classification training. For further research we need longer periods of data in order to acquire more information about the behavior variations of cells. Another direction in future is to use more than one distribution type indicator. Propagation delay distribution, for example, should contain interesting information to combine with the CQI distribution.

References

1. Everitt, B., Landau, S., Leese, M.: Cluster analysis. Arnold, London (2001)
2. Guerzoni, R., Zanier, P., Suman, R., DiSarno, Fabio.: Signaling-based radio optimization in WCDMA networks. In: IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks, pp. 1—6. IEEE (2007)
3. Johnson, C: Radio Access Networks for UMTS: Principles and Practice. John Wiley & Sons (2008)
4. Kaaranen, H., Ahtiainen, A., Laitinen, L., Naghian, S., Niemi, V.: UMTS Networks. John Wiley & Sons, Chichester (2009)
5. McGill, R., Tukey, J.W., Larsen, W.A.: Variations of Boxplots. The American Statistician, 32, 12-16 (1978)
6. Mishra, A. R.: Advanced Cellular Network Planning and Optimisation: 2G/2.5G/3G…Evolution to 4G. John Wiley & Sons (2006)
7. Ross, T. J.: Fuzzy Logic With Engineering Applications. McGraw-Hill, New York (1995)
8. Xu, R., Wunsch II, D.C.: Clustering. John Wiley & Sons, New Jersey (2009)
9. Zadeh, L.A.: Fuzzy sets. Information and Control, 338—353 (1965)
10. Zanier, P., Guerzoni, R., Soldani, D.: Detection of Interference, Dominance and Coverage Problems in WCDMA Networks. In: IEEE 17th International Symposium on Personal, Indoor and Mobile Radio Communications, pp. 1—5. IEEE (2006)