Descriptive Data Mining Modeling in Telecom Systems

Ivo Pejaković[1], Zoran Skočir[1], Damir Medved[2]

[1] Faculty of Electrical Engineering and Computing, University of Zagreb
Unska 3, HR-10000 Zagreb, Croatia
Tel: +385 1 6129 763; +385 1 6129 831; Fax: +385 1 6129 616
E-mails: ivo.pejakovic@fer.hr, zoran.skocir@fer.hr

[2] Croatian Telecom Inc.
Hebrangova 32-34, HR-10001 Zagreb, Croatia
Tel: +385 1 4911 140; Fax: +385 1 4911 055
E-mail: damir.medved@ht.hr

Abstract: Croatian Telecom possesses huge amounts of data which has been collected for various purposes. The pilot project was deployed to get new insights about the distribution of errors and interferences in Croatian Telecom's network. This paper describes the usage of descriptive data mining modeling for getting valuable insights from the observational data, the results of the analysis and possible benefits from these results. The special emphasis was given to the usage of probabilistic model-based clustering and on EM algorithm which is used for estimation of parameters of the model.

1. INTRODUCTION

Telecom operators are among the biggest generators of data and the adoption of new ways of exploring and analyzing data in telecom systems is very important. These new ways of analyzing captured and stored data should provide new insights about various problems which can be found in telecom operators' business. Data mining analysis begins to play a significant role in the analysis of telecom's data and there are numerous emerging data mining applications in telecommunications [1].

Data mining can be defined as the analysis of (often very large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner. Over the time, numerous data mining techniques have been developed for analyzing data which can be grouped as follows: exploratory data analysis (EDA), descriptive modeling, predictive modeling, finding patterns and rules, retrieval by content [2].

Croatian Telecom has been collecting various data for a long time and for various purposes. Today it possesses huge amounts of data which has been originally collected for purposes other than data mining analysis. This paper describes pilot project of using selected number of data mining techniques for the analysis of the telecom's data and possible benefits of their usage. The importance of this project is that this is one of the first data mining projects in Croatian Telecom which can pave the way to the other similar projects.

The data that was used for the analysis was the data about the errors and interferences in Croatian Telecom's network. The goal wanted to be achieved was getting new insights about the distribution of errors and interferences in the part of the telecom's network. Due to the nature of the analyzed data and the goal wanted to be achieved we used descriptive data mining modeling and EDA techniques for finding interesting patterns and natural groupings within the data. The paper describes the analysis of the observational telecom's data based on probabilistic model-based clustering using mixture models and especially, models based on EM (Expectation-Maximization) algorithm.

This paper is structured as follows: Section 2 describes descriptive modeling and EDA techniques, Section 3 describes the idea of mixture models and EM algorithm that is used for modeling, Section 4 describes the data used for the analysis, Section 5 describes the analysis and results of the analysis and conclusion in Section 6 is followed by the reference list.

2. DESCRIPTIVE MODELING AND EDA

The goal of descriptive modeling is to build a model which describes all of the data. Examples of such descriptions include models for the overall probability distribution of the data (*density estimation*), partitioning of the *p*-dimensional space into groups (*cluster analysis and segmentation*) and models describing the relationships between variables (*dependency modeling*) [2]. In cluster analysis the aim is to discover "natural" groups in data, for example, in telecom operators' databases or scientific databases. In cluster analysis and segmentation there is no "right" number of clusters or segments. It is usually chosen by researcher on the basis of some objective or subjective criterion.

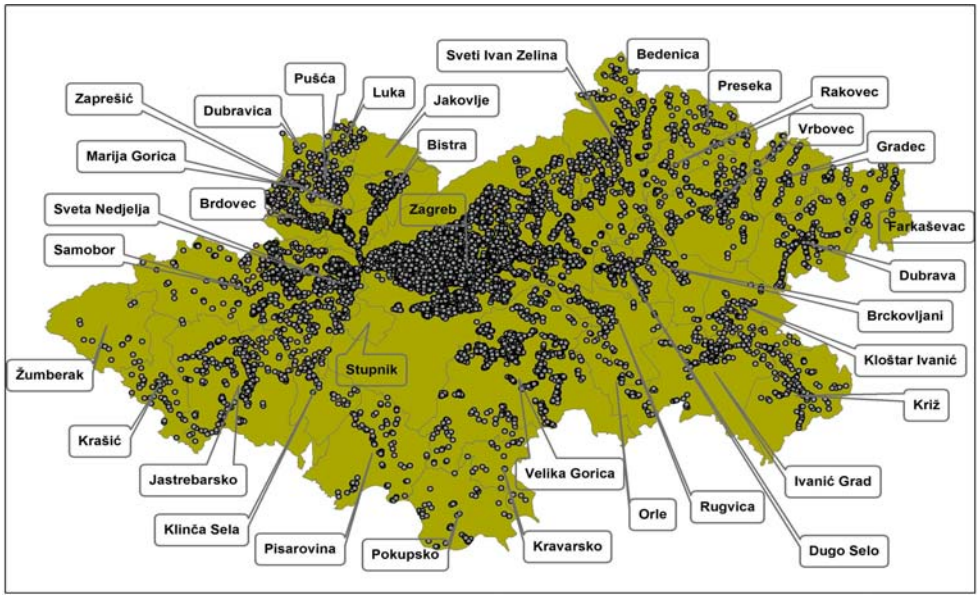

Figure 1 – The map of the analyzed area

Exploratory data analysis (EDA) is the name for group of techniques where the goal is simply to explore data without any clear idea what we are looking for. Typically, EDA techniques are interactive and visual. There are many effective graphical display methods for relatively small, low-dimensional data sets. As the dimensionality (number of variables, p) increases, it becomes much more difficult to visualize the cloud of points in p-space. For dimensions above 3 or 4 EDA techniques are not very useful. The examples of EDA techniques are various histograms, scatterplots, contour plots etc.

3. EM ALGORITHM FOR MIXTURE MODELS

A multivariate random variable $\mathbf{X}$ is a set $X_1,\ldots,X_p$ of p random variables and $\mathbf{x} = \{x_1,\ldots,x_p\}$ denotes a set of values for $\mathbf{X}$. The general form of a mixture distribution for multivariate $\mathbf{x}$ is given with [2]:

$$f(x) = \sum_{k=1}^{K} \pi_k f_k(x;\theta_k) \qquad (1)$$

Where π_k is the probability that an observation will come from the k-th component, K is the number of components, $f_k(\mathbf{x};\theta_k)$ is the distribution of the k-th component and θ_k is the vector of parameters describing the k-th component.

The mixture models provide a general framework for clustering in a probabilistic context. In the literature this is referred to as probabilistic model-based clustering since there is an assumed probability model for each component cluster. In this framework it is assumed that data come from a multivariate finite mixture model that is described with (1) and every cluster is described by one component of a mixture model.

Roughly speaking, the general procedure is as follows: given an observational data set of n observations $D = \{\mathbf{x}(1),\ldots,\mathbf{x}(n)\}$, determine how many clusters K we want to fit to the data, choose parametric models for each of this K clusters (for example multivariate Normal distributions, exponential distributions etc.) and then use EM algorithm to determine the component parameters θ_k and component probabilities π_k from the data. Once mixture decomposition has been found, the data can be assigned to the clusters by assigning each point to the cluster from which it is most likely to have come.

The properties of the EM algorithm are described in details in [2], [3], [4], [5], [6], [7]. EM algorithm is an iterative algorithm for estimating the parameters of the parametric models as is the case with mixture models. It consists of two steps and generally can be described as follows:

E-step – calculation of the expected cluster probabilities. This is the first step and it is called *Expectation*.

M-step – calculation of the distribution parameters through maximization of the likelihood of the distributions given the data. This step is called *Maximization*.

Considering the observational data set D, independently sampled from the same distribution $f(\mathbf{x}|\theta)$ where θ is a set of model parameters, we can define the likelihood function $L(\theta|D)$ as the probability that data would have arisen, for a given value of θ, regarded as a function of θ [2]:

$$L(\theta \mid D) = L(\theta \mid x(1),...,x(n)) = \\ = p(x(1),...,x(n) \mid \theta) = \prod_{i=1}^{n} f(x(i) \mid \theta) \quad (2)$$

The likelihood function is used as a measure of how well some parametric model fits the data. The likelihood computation is simply the multiplication of the sum of the probabilities for each of the instances. For example, with two clusters A and B containing instances from *D* with cluster probabilities π_A and π_B the computation of *L(θ|D)* is:

$$L(\theta \mid D) = [\pi_A p(x_1 \mid A) + \pi_B p(x_1 \mid B)] * \\ *[\pi_A p(x_2 \mid A) + \pi_B p(x_2 \mid B)] *...* \quad (3) \\ *[\pi_A p(x_n \mid A) + \pi_B p(x_n \mid B)]$$

4. UNDERSTANDING THE DATA

The analysis was done on the data about errors and interferences in the part of Croatian Telecom's network that covers area of TK center Zagreb (Telecommunication center Zagreb). The analyzed data was collected by Croatian Telecom from December 1999 to the end of May 2003. The data set contained information about more than 300000 registrated errors and interferences. The collected data contained various attributes that describe errors and interferences entered in the database. Among these various attributes we were especially interested in the attributes that describe geographical location of the errors and interferences and the attribute that describes the nature (or kind) of the errors and interferences.

Figure 1 shows the map of the whole analyzed area. The spots on the map represent places of registered interferences and errors.

The attributes that describe geographical location of the errors are Gauss-Kruger coordinates. These attributes are labeled with X and Y and represent distance in meters related to the origin for this zone (the fifth zone) of Gauss-Kruger projection system [8]. The attribute X corresponds to the longitude and attribute Y corresponds to the latitude. The values of these attributes are represented as real numbers. Figures 2 and 3 show the distribution of values for the attributes X and Y respectively.

The attribute that describes the nature of errors and interferences is labeled with MSMLO. It is categorical attribute and may have 66 different values that describe the nature or kind of an error or interference. For example, it may be an error on the cable of specific kind (optical, shielded, unshielded), on PCM (Pulse Code Modulation), on telephone line etc.

Therefore, an error or interference is represented as set [X,Y,MSMLO].

The preparation of data for modeling included removing of noise and inconsistencies from the data. The data contained 0.2% of noise and wrong inputs, and they were dropt from the data set for building a model. The whole process of preparing the data for the analysis is described in [9].

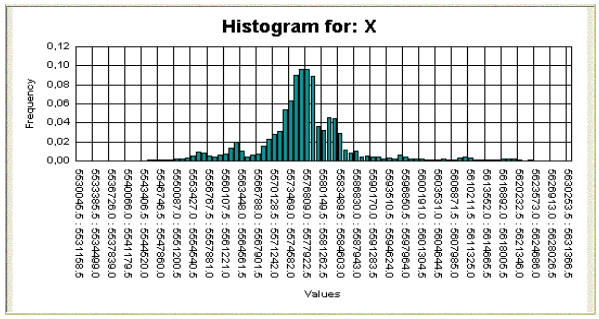

Figure 2 - Distribution of values for attribute X

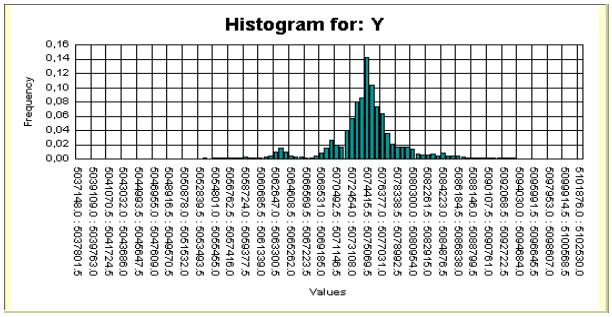

Figure 3 - Distribution of values for attribute Y

5. THE ANALYSIS

The analysis was aimed at discovering natural groups within collected data. Considering the distribution of the coordinates of errors that is shown in Figures 2 and 3, it can be concluded that it can be best described as the mixture model with multivariate Normal distributions. Actually, during the analysis we have tried few different models based on different clustering algorithms (for example k-means [2], Oracle's o-cluster [10] etc.) and it has been shown that the model based on mixture model with multivariate Normal distribution fits the observed data the best [9].

For modeling it has been used the implementation of EM algorithm contained in Weka open-source data mining tool [11]. This implementation of EM algorithm handles both numerical and categorical attributes. Supposed distributions of numerical attributes are mixture Normal distributions.

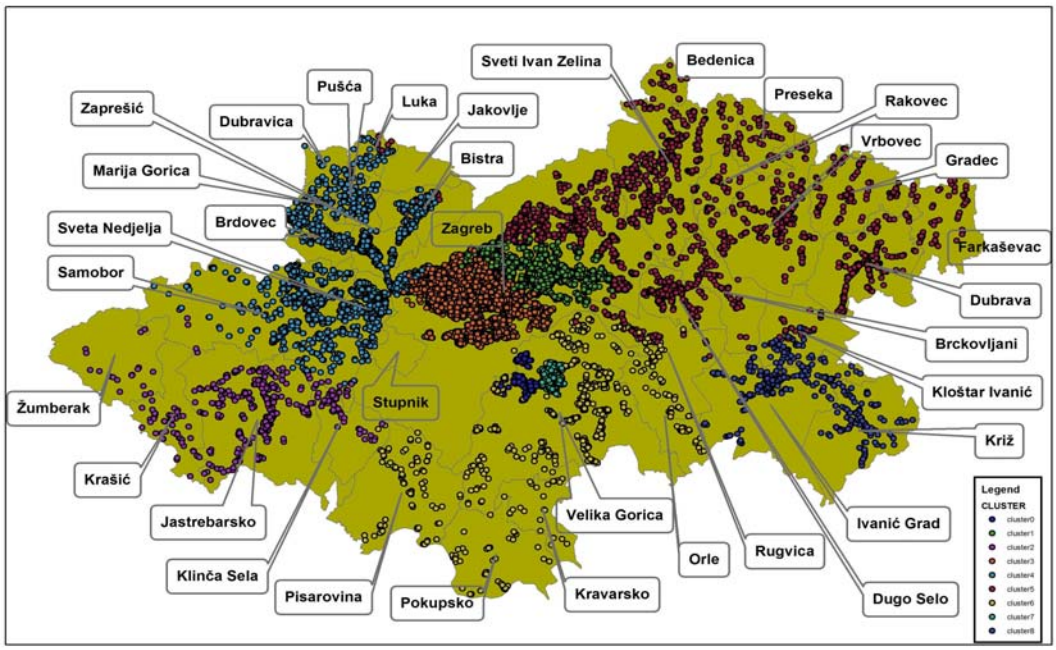

Figure 4 – Distribution of clusters

After the preparing process it has been used modeling with EM algorithm for finding the model's parameters. The model has been built on the values of set [X,Y,MSMLO]. So, resulting clusters from model have three dimensions. Dimensions X and Y determine geographical location of the clusters while dimension MSMLO gives distribution of the errors and interferences.

The final number of clusters was chosen comparing the values of the likelihood function (2) for the different number of clusters *K*. The likelihood function reached the maximum for the number of 9 clusters. The parameters that had needed to be estimated were means, component (cluster) probabilities and standard deviations. The values of these parameters can be found in [9]. Every cluster corresponds to one component of the given mixture model.

Figure 4 shows geographic representation of the clusters (only geographic dimensions of the clusters are shown). Every cluster is represented with spots of a different color and every cluster has its number. The clusters have various size and shape, probability (that is proportional to the number of errors and interferences covered by them) and standard deviation. For example cluster 3, which covers the central part of Zagreb town, has the highest cluster probability 58%, which means that it covers 58% of all registered errors etc. [9].

Distribution of the attribute MSMLO is very different for each of these clusters even for very near clusters (in geographical sense). Comparing the distributions of the attribute MSMLO for different clusters very interesting patterns can be noticed. For example, Figures 5 and 6 show the distributions of the attribute MSMLO of the clusters that cover area of Velika Gorica, clusters 0 and 7 in Figure 5.

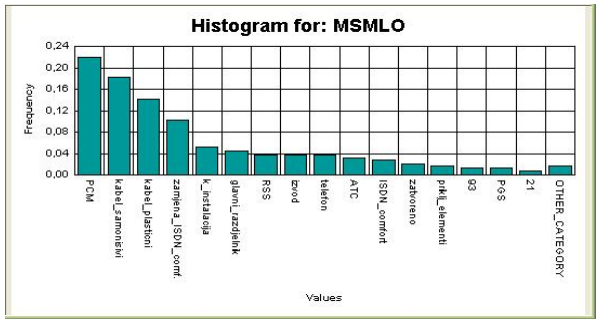

Figure 5 – Distribution of attribute MSMLO for cluster 0

Although these two clusters are very near to each other, we can see significant differences in distribution of errors and interferences in these two clusters. For example, the second and third bar from the left in Figure 5 represent frequencies of errors and interferences on two kinds of cables. Together, they take more than 30% of all errors and interferences in the cluster 0. In Figure 6 it can be noticed that frequency of errors on the cables in the cluster 7 takes only 15% of all errors and interferences. It could be very interesting to see what is the reason for that. Deeper understanding and finding out the reasons of these patterns is not the task of descriptive modeling and it is usually on the domain experts (in this case, engineers responsible for network management and network control).

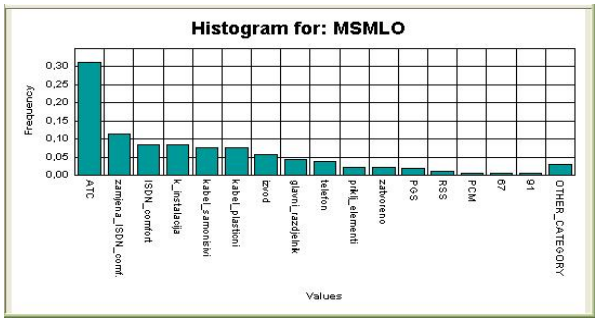

Figure 6 – Distribution of attribute MSMLO for cluster 7

The same observations can be made for any pair of clusters and any type of errors or interferences. Furthermore, knowing these distributions of clusters, analysis can be repeated on the part of the analyzed area that is covered with only few of the clusters. These results give more detailed picture of what is going on in the part of the analyzed area that cannot be seen in the analysis of the whole area. For example, if we were interested in what is going on in Zagreb town the analysis should be repeated on the area covered with clusters 1 and 3, which cover the region of Zagreb town and surrounding places. The number of clusters in that area can be also estimated considering the values of the likelihood function for the estimated parameters of the mixture model and number of clusters. By doing that we can make a hierarchy of clusters each having its view on the problem. Clusters on the top of hierarchy have more general view of what is going on than clusters below them.

The overall probability of the occurrence of the errors and interferences is governed by many factors: density of the telecom infrastructure, kind of the equipment, number of users, external interferences etc. So it is quite possible that some areas will have very different distributions of errors and interferences due to some specific factor (or factors). Exploring these distributions could result in discovering valuable information. Finding areas with the unusual distributions of errors and interferences may lead to further investigation of the reasons their existence. This can be accomplished by investigation on the ground and/or using some other techniques of data mining analysis (for example predictive modeling). Using predictive modeling techniques requires more detailed data about telecom infrastructure in specific area and external interferences. The final result can be the discovery of reasons of some types of errors and interferences. Once the reasons of errors and interferences are known, some of them can be removed.

6. CONCLUSION

In descriptive modeling the validation of the results is always on the side of the domain experts. Discovered patterns and natural groupings within the analyzed data reveals a lot information about distributions of errors and interferences of specific kind in the areas related to the specific clusters. Some of the distributions that have been found may already be known to the domain experts but some of them may not. Discovering these patterns in the analyzed data can improve the quality of service that operator offers to its customers (for example, by eliminating some reasons of the interferences on the telephone lines etc.).

Descriptive data mining modeling techniques have been used to make the model which describes the data. EDA techniques have been used to visualize the results of modeling.

Using data mining techniques offers the possibility of finding the new and valuable information that may otherwise be lost. This information can be used for various purposes in telecommunications: improving the quality of service, Business Intelligence systems, CRM (Customer Relationship Management) applications [12] etc.

REFERENCES

[1] Jaiwei Han, Russ B. Altman, Vipin Kumar, Heikki Mannila, Daryl Pregibon: “Emerging Scientific Applications in Data Mining”, Communications of ACM 45, 8 (August 2002), 54-58.

[2] David Hand, Heikki Mannila, Padhraic Smyth: “Principles of Data Mining”, MIT Press, 2001.

[3] A.P.Dempster, N.M. Laird, and D.B. Rubin: “Maximum-likelihood from incomplete data via the em algorithm”, J. Royal Statist. Soc. Ser. B., 39, 1977.

[4] R. Redner and H. Walker: “Mixture densities, maximum likelihood and the em algorithm”, SIAM Review, 26(2), 1984.

[5] Z. Ghahramami and M. Jordan: “Learning from incomplete data”, Technical Report AI Lab Memo No. 1509, CBCL Paper No. 108, MIT AI Lab, August 1995.

[6] M. Jordan and R. Jacobs: “Hierarchical mixtures of experts and the em algorithm”, Neural Computation, 6:181–214, 1994.

[7] C.F.J. Wu: “On the convergence properties of the em algorithm”, The Annals of Statistics, 11(1):95–103, 1983.

[8] URL: http://exchange.manifold.net/manifold/manuals/5_userman/mfd50Projections_Tutorial.htm

[9] Pejakovic, Ivo: “Data Mining Methods in Business Intelligence Systems” (in Croatian), Master thesis, Faculty of Electrical Engineering and Computing, Zagreb, 2004.

[10] URL:http://www.oracle.com

[11] URL: http://www.cs.waikato.ac.nz/ml/weka/

[12] Michael J. A. Berry, Gordon S. Linoff: “Mastering Data Mining – The Art and Sience of Customer Relationship Management”, John Wiley & Sons Inc., 2000.