

*XVII IMEKO World Congress
Metrology in the 3rd Millennium
June 22–27, 2003, Dubrovnik, Croatia*

AUTOMATIC DISCRETIZATION IN PREPROCESSING FOR DATA ANALYSIS IN MOBILE NETWORK

*Pekka Kumpulainen**, *Kimmo Hätönen***, and *Pekko Vehviläinen**

*Measurement and Information Technology, Tampere University of Technology, Tampere, Finland

**Nokia Research Center, Helsinki, Finland

Abstract – Several data mining methods require discretization of the data, rough sets and association rules for example. In this paper we present simple discretization methods for two types of variables common in telecommunications network. The variables are first classified in order to apply the proper discretization method. These methods are based on the amplitude distribution of the variables. Methods are unsupervised, i.e. they do not require class information for the samples.

Keywords: preprocessing, discretization, telecommunications network.

1. INTRODUCTION

Thousands of variables are recorded in modern mobile radio access network environment. They are used to monitor and manage the network. To reduce the number of variables the measurements are aggregated into Key Performance Indicators (KPIs) [1]. Even less in number than the original variables, there are still dozens of KPIs to deal with. Therefore data mining or knowledge discovery methods are required to reveal the valid information from the enormous data flow.

There are lot of data mining methods to choose from [2, 3]. Each method has its own characteristics and requirements on the data presented to it. Preprocessing of the data is an essential part in any data mining process. Some of the methods require the input data to be discrete, taken rough sets and association rules, for example [4]. That is why a new task in preprocessing is needed: discretization.

An expert who is familiar with the process would most probably do the best discretization. Unfortunately, the task is huge because of the amount of data and resources are limited so that usually there are not enough experts to do the discretization. Therefore automatic discretization is required. Usually there is no expert classification information available for the samples and the discretization method has to be unsupervised.

The data sets can have very different characteristics, including value ranges and distributions. This makes it impossible in practise to have only one method for discretization. Different variables have to be classified and a suitable discretization method has to be developed for each class of variables.

The goal of our research was to develop unsupervised methods, which are not necessarily optimal, but usable and robust instead. In this paper we present a simple variable classification method. Also two discretization methods are presented. Based on the classification of variable a suitable discretization method is selected.

2. CLASSIFICATION OF VARIABLES

Different types of KPIs have to be treated separately to get the best possible results out from the use of analysis methods. This applies to various preprocessing tasks, such as discretizing, scaling and selecting indicators for analysis methods.

There are two basic groups of KPIs by nature. Quantity and quality related KPIs [1, 5]. Quantity KPIs are typically related to the amount of traffic or other cumulative counters. Quality KPIs are often relative counters and scaled to percentages.

Before the appropriate discretization method can be applied to a variable it has to be classified.

Classification is based on the distributions of the variables. For each variable feature vectors describing the variable's distribution are calculated. They are compared to the predefined reference features and based on the comparison the most similar reference class is assigned to the variables. Similarity is measured by distance between these feature vectors. Minkowski metrics has been used with $m=1$ and $m=2$, which correspond to City Block and Euclidean distances [6].

Examples of distributions of indicators belonging to different classes are given in Fig 1. Classes *A* and *D* are typical quality KPIs. Class *B* is a typical quantity KPI. Class *C* is a cumulative failure counter. Its distribution is a mirror image of class *A*.

The feature vectors used in the classification consist of five values derived from the Propability Density Function (PDF) estimate. These values are the proportions of the total value range of the variable covered by a given percentile of the data. Percentiles 10%, 25%, 50%, 75%, 90% are used. This is equivalent to scaling the data between 0 and 1 and then finding the percentiles.

The predefined reference feature vectors have been manually selected. Following reference features were found to give suitable classification for classes presented in Fig 1:

Class A: [0.7 0.9 1 1 1]
 Class B: [0.1 0.2 0.5 0.6 0.7]
 Class C: [0 0 0 0.01 0.05]
 Class D: [0.4 0.4 0.6 0.7 0.8]

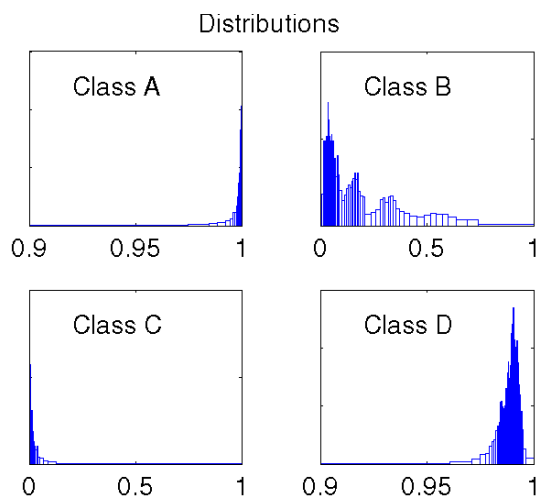


Fig. 1 PDF estimates of indicators in different classes

3. DISCRETIZATION

Equal Width and *Equal Frequency Binning* are mentioned as unsupervised methods [7]. Unsupervised methods are most frequently used in image processing [8 - 10]. Unsupervised discretization procedure is called *thresholding* and is used for image segmentation.

The two suggested discretization methods are based on the PDF estimates (histograms) of the variables. Equal frequency bin estimate is used in both methods, which both are unsupervised.

The results are compared to those given by an expert who is competent in managing and optimizing GSM networks.

3.1. Discretization method 1

The first method is *hierarcical valley detection*. It is suitable for variables that have more or less multinormal distribution. This method is applied to variables belonging to Class B. An example is shown in Fig 3.

Only the number of discretization limits to detect is required as user input. The limits are detected as the local minima i.e. valleys in the PDF estimate. The algorithm starts with a coarse estimate using 3 bins. This is a minimum number of bins, where a single local minimum can be found. Example is given in Fig. 2. A PDF estimate of a sinusoid is calculated using 3 bins. The number of bins in the PDF estimate is increased by two in each step thus keeping odd number of bins. This is continued until the required number of limits is found.

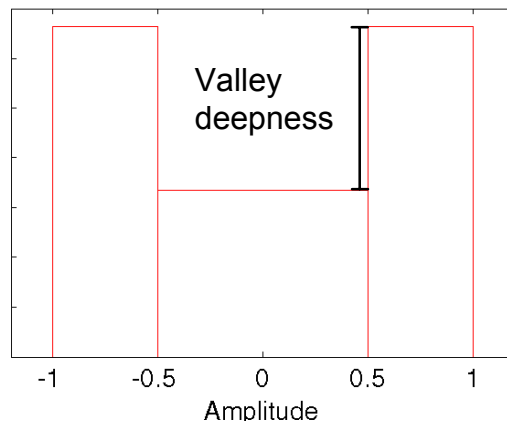


Fig. 2. 3-bin PDF estimate of a sinusoid

It is possible that after increasing the number of bins there are more local minima than required. In this case a minimum deepness is calculated for each of the valleys. This is the minimum height difference between the valley and the surrounding local maxima in the PDF estimate as shown in Fig. 2.

An example of limits given by this discretization method is given in Fig. 3. The value range is normalized between 0 and 1.

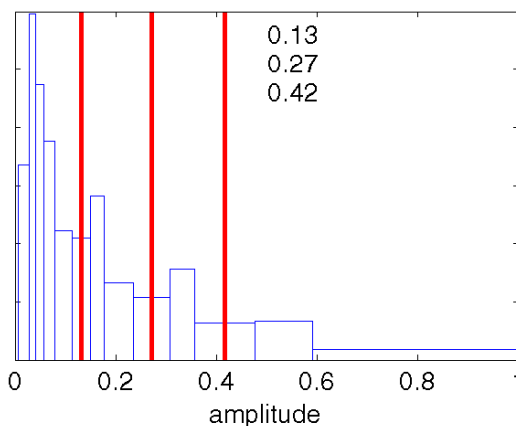


Fig. 3. Example of a PDF and detected discretization limits

This method gives discretization limits that are very similar to those given by human experts. The local minimum values in the distribution are natural limits to human. The results are compared to the corresponding limits given by an expert in TABLE I.

TABLE I. Automatic and manual discretization limits

	Limit 1	Limit 2	Limit 3
Automatic	0.13	0.27	0.42
Human expert	0.11	0.22	0.47

3.2. Discretization method 2

The second algorithm is designed for typical quality type variables. This type of variable, often scaled to percentages, is common as a quality measurement in telecommunication networks. The distribution is heavily skewed towards 100 %. Classes A and D in Fig.1 are examples of distributions of this type of variables.

For skewed variables we want to find two limits:

- *Normal*: the limit between *normal* and *noticeably decreased* performance
- *Minimum*: the limit of *heavily decreased* performance

This algorithm takes the advantage on the skewness of the distribution.

Samples above the median are discarded and a PDF estimate is calculated from the remaining samples. The widths of the bins in the PDF estimate are used to select discretization limits.

To detect the *normal* limit we take into account only those values that are smaller than the median. The median is the 50 % percentile of the distribution. Some other predefined percentile can be chosen as well. The PDF estimate is then calculated for this data using proportional bins. The number of bins is the square root of the number of samples in use - in this case the number of samples that are smaller than the median. Starting from the left (smaller values) the width of each bin is compared to the median of the widths of bins remaining on the right (larger values). The limit is set at the right side of the first bin that is smaller than the median of the rest.

To detect the minimum limit we take into account only values smaller than previously detected normal cut. The PDF estimate is calculated for the remaining data using proportional intervals.

Find the first bin from the left that is narrower than half (or other portion) of the range of the data to the right (larger values). Right edge of the bin is set to the minimum limit.

Parameters:

- percentile limit in the data selection
- number of bins to use in the PDF estimates
- portion of the data range to which the bin widths are compared

This method is designed to Variables in classes A and D, which are skewed to the right, thus they have negative skewness. This method is also suitable to variables in class C if the values of the variable are negated, which reverses the distribution.

An example of limits given by this discretization method for a variable in class A is presented in Fig 4. The PDF estimate shown is calculated from all the samples available.

The results are compared to the corresponding limits given by an expert in TABLE II.

TABLE II. Automatic and manual discretization limits

	<i>Minimum</i> limit	<i>Normal</i> limit
Automatic	95.57 %	98.79 %
Human expert	98 %	99 %

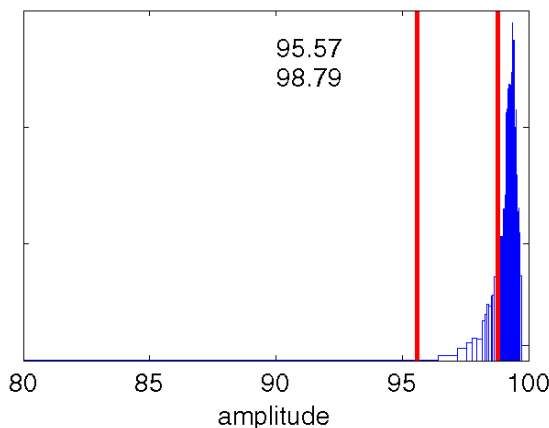


Fig. 4 Example of a PDF and detected discretization limits

This method performs reasonably well for all the data we have been using. Since the total value range of the variable is used, this method is sensitive to outliers. Therefore it should be used only after outlier detection. One topic of future research is integrating outlier detection with discretization.

4. RESULTS AND CONCLUSIONS

The presented method for variable classification meets the objectives of being usable for preparing the data for the analysis methods and being robust at the same time. The suggested discretization method is also able to classify variables in telecommunication network. The presented method is based on the use of the characteristics of variable distributions. The benefit of the method is that the characteristics are very simple to extract automatically from the input data. The simplicity and the requirement of only few parameters make the method easily applicable to many platforms, including the network management system that the network operating personnel uses in daily bases.

The observed KPIs can automatically be discretized to value ranges that are understandable and meaningful for the operating personnel or other analysts that need the information for example network optimisation and planning tasks.

Our future research of automatic discretization focuses on the development of the method to be even more general and robust. As performance data from real third generation network (based on the Universal Mobile Telecommunications System, UMTS, standard) becomes available the necessary adjustments for the method are also studied.

REFERENCES

[1] Suutarinen, J., Performance Measurements of GSM Base Station System. Thesis (Lic.Tech.) Tampere University of Technology. 1994

[2] K. Cios, W. Pedrycz, R. Swiniarski, "Data Mining Methods for Knowledge Discovery", *Kluwer Academic Publishers*, 1998.

- [3] J. Han, M. Kamber, "Data Mining: Concepts and Techniques", *Morgan Kaufman Publishers*, San Francisco, CA, 2001.
- [4] Vehviläinen P., Hätönen, K., Kumpulainen, P., "Data mining in quality analysis of mobile Digital telecommunications network". *XVII IMEKO World Congress Dubrovnik, Croatia*, June 22-27, 2003.
- [5] Lempiläinen, J. Manninen, M., "Radio Interface System Planning for GSM/GPRS/UMTS". *Dordrecht, The Netherlands: Kluwer Academic Publishers*. 2001
- [6] R.A. Johnson, D.W. Wichern, "Applied Multivariate Statistical Analysis", *Prentice Hall, Inc.* New Jersey, 1998.
- [7] J. Dougherty, R. Kohavi, M. Sahami, "Supervised and Unsupervised Discretization of Continuous Features", In A. Prieditis, S. Russel, eds., *Machine Learning: Proceedings of the Twelfth International Conference*, Morgan Kaufman Publishers, San Francisco, CA. 1995.
- [8] D.M. Tsai, "A fast thresholding selection procedure for multimodal and unimodal histograms", *Pattern Recognition Letters* 16(6), 653-666. 1995.
- [9] P.L. Rosin, "Unimodal thresholding", *Proc. Scandanavian Conference on Image Analysis*, pp. 585-592, Kangerlusuaq, Greenland, 1999.
- [10] C. Chang, L. Wang, "A fast multilevel thresholding method based on lowpass and highpass filtering," *Pattern Recognition Letters*, 18(14), 1469-1478, 1997.

AUTHORS: Pekka Kumpulainen, Tampere University of Technology, Measurement and Information Technology, Box 692, 33101, Tampere, Finland, phone +358 40 8490930, Fax +358 3 31152171, email pekka.kumpulainen@tut.fi.
Kimmo Hätönen, Nokia Research Center, P.O.Box 407, FIN-00045, Nokia Group, Finland, email kimmo.hatonen@nokia.com.
Pekko Vehviläinen, Tampere University of Technology, Measurement and Information Technology, Box 692, 33101, Tampere, Finland, phone +358 3 31153572, Fax +358 3 31152171, email pekko.vehvilainen@tut.fi.