

# Quantization of continuous input variables for binary classification

Michał Skubacz<sup>1</sup> and Jaakko Hollmén<sup>2</sup>

<sup>1</sup> Siemens Corporate Technology, Information and Communications, Neural Computation, 81730 Munich, Germany, [Michal.Skubacz@mchp.siemens.de](mailto:Michal.Skubacz@mchp.siemens.de)

<sup>2</sup> Helsinki University of Technology, Laboratory of Computer and Information Science, P.O. Box 5400, 02015 HUT, Finland, [Jaakko.Hollmen@hut.fi](mailto:Jaakko.Hollmen@hut.fi)

**Abstract** Quantization of continuous variables is important in data analysis, especially for some model classes such as Bayesian networks and decision trees, which use discrete variables. Often, the discretization is based on the distribution of the input variables only whereas additional information, for example in form of class membership is frequently present and could be used to improve the quality of the results. In this paper, quantization methods based on equal width interval, maximum entropy, maximum mutual information and the novel approach based on maximum mutual information combined with entropy are considered. The two former approaches do not take the class membership into account whereas the two latter approaches do. The relative merits of each method are compared in an empirical setting, where results are shown for two data sets in a direct marketing problem, and the quality of quantization is measured by mutual information and the performance of Naive Bayes and C5 decision tree classifiers.

## 1 Introduction

Whereas measurements in many real-world problems are continuous, it may be desirable to represent the data as discrete variables. The discretization simplifies the data representation, improves interpretability of results, and makes data accessible to more data mining methods [6]. In decision trees, quantization as a pre-processing step is preferable to local quantization process as part of the decision tree building algorithm [1,4]. In this paper, quantization of continuous variables is considered in a binary classification problem. Three standard quantization approaches are compared to the novel approach, which attempts to balance the quality of input representation (measured by entropy) and the class separation (measured by mutual information).

The comparison of the four approaches to quantization is performed on two data sets from a direct marketing problem. Mutual information, Naive Bayes classifier, and C5 decision tree [8] are used in measuring the quality of the quantizations.

## 2 Quantization

Quantization, also called discretization, is the process of converting a continuous variable into a discrete variable. The discretized variable has a finite number of values ( $J$ ), the number usually being considerably smaller than the number of possible values in the empirical data set. In the binary classification problem, a data sample  $(\mathbf{x}_i, y_i)_{i=1}^N$  and  $y_i \in \{0, 1\}$  is available. Variable  $\mathbf{x}_i \in \mathbb{R}^k$  is a vector of variables on a continuous scale. In the quantization process, the component  $k$  of the  $x_i$  later denoted by  $x_{ik}$ , is mapped to the discrete counterpart  $x'_{ik}$  when the original variable  $x_{ik}$  belongs to the interval defined by the lower and upper bounds of the bin. The number of data falling into a bin  $j$  is defined as  $n_{kj}$  and the probability of a bin as  $p_{kj} = \frac{n_{kj}}{N}$ .

One could approach the discretization process in many different ways, starting for example from naive testing of random configurations and selecting the best one for a particular problem. More structured approaches may consider discretizing all variables at the same time (global), or each one separately (local). The methods may use all of the available data at every step in the process (global) or to concentrate on a subset of data (local) according to the current level of discretization. Decision trees, for instance, are usually local in both senses. Furthermore, two following search procedures could be employed. The top-down approach [4] starts with a small number of bins, which are iteratively split further. The bottom-up approach [5], on the other hand, starts with a large number of narrow bins which are iteratively merged. In both cases, a particular split or merge operation is based on a defined performance criterion, which can be global (defined for all bins) or local (defined for two adjacent bins only). An example of a local criteria was presented in [5].

In this paper, a globally defined performance criterion is optimized using a greedy algorithm. In each iteration of the one-directional greedy algorithm, a most favorable action at the time is chosen. In the initial configuration one allocates a large number of bins to a variable and starts merging two adjacent bins by choosing the most favorable merge operation. The approaches used in this paper are local in the sense that variables are discretized separately and global in the sense that all the available data are used in every step of the quantization process. Discretizing variables separately assumes independence between them, an assumption which is usually violated in practice. However, this simplifies the algorithms and makes them scalable to large data sets with many variables. In contemporary data mining problems, these attributes become especially important. In a real situations, one particular value on the continuous scale may occur very frequently overwhelming the entire distribution of the variable. For example, the field "total length of the international telephone calls" for a particular private customer is likely to be predominately filled with zeros. This situation corresponds to a peak in the probability density function and can lead to the deterioration of the quantization process. If this is detected, for example by checking if a given value appears in more than 60% of the samples, a dedicated interval should be allocated and these samples removed from the discretization process.

*Equal Width Interval* By far the simplest and most frequently applied method of discretization is to divide the range of data to a predetermined number of bins [6]. Each bin is by construction equally wide, but the probabilities of the bins may vary according to the data. In classification problems, this approach ignores the information about the class membership of data assigned to each bin.

*Maximum Entropy* An alternative method is to create bins so that each bin equally contributes to the representation of the input data. In other words, probability of each bin for the data should be approximately equal. In fact, this is achieved by maximizing the entropy of the binned data. The entropy for the binned variables may be defined as  $H_k = \sum_{j=1}^J p_{kj} \log p_{kj}$ , where the sum is over all bins. Entropy has been used in context of discretizing variables in [9].

*Maximum Mutual Information* In classification problems, it is important to optimize the quantized representation with regard to the distribution of the output variable. In order to measure information about the output preserved in the discretized variable, mutual information may be employed [3]. Mutual information was used in the discretization process of the decision tree construction algorithm (ID3) in [7]. Mutual information is measured in terms of quantized variables as

$$I_k = \sum_{j=1}^J P(b_{kj}, y = 1) \log \frac{P(b_{kj}, y = 1)}{P(b_{kj})P(y = 1)} + P(b_{kj}, y = 0) \log \frac{P(b_{kj}, y = 0)}{P(b_{kj})P(y = 0)}$$

*Maximum Mutual Information with Entropy* By combining the maximum entropy and the mutual information approaches, one hopes to obtain a solution with the merits of both. This should strike a balance between the representation of the input and the knowledge of the output variable at the same time. In other words, one would like to retain balanced bins that turn out to be more reliable (prevent overfitting in this context) but simultaneously to optimize the binning for classification. Our greedy algorithm is based on a criterion function which is the product of the mutual information and the maximum entropy as

$$G_k = H_k I_k.$$

The greedy algorithm approximates the gradient ascent optimization. Writing the gradient of the product of two functions as  $\frac{\partial f(x;\theta)g(x;\theta)}{\partial \theta} = f'(x;\theta)g(x;\theta) + g'(x;\theta)f(x;\theta)$ , we note that the search direction is driven by the balance of the two factors subject to constraints imposed by data. A similar measure involving mutual information divided by entropy was proposed in the context of discretization in [8]. However, the measure was used for the problem of binary discretization in splitting operation. Our novel approach assumes discretization into several bins and the comparison is done among all merging operations.

### 3 Experiments

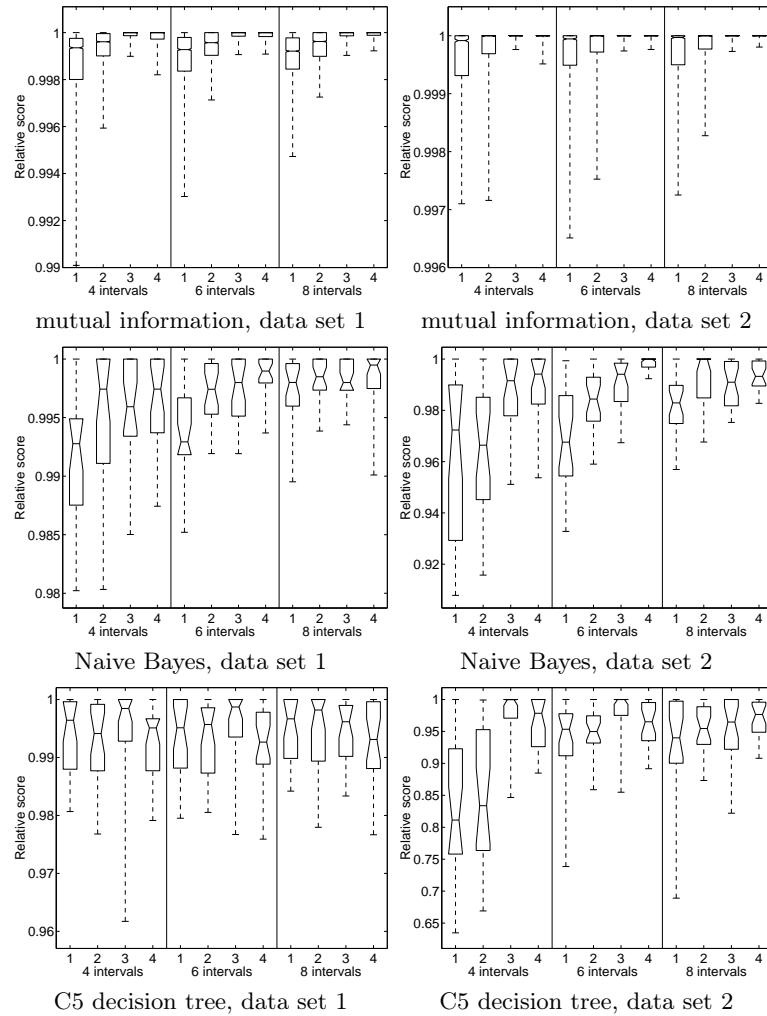
Two data sets were used in the evaluation. Both of them were collected and used in direct marketing campaigns. The input variables represented customer

information and the output was the customer's binary response. The data set 1 consisted of 144 input variables and 12496 samples whereas the data set 2 had 75 input variables and 35102 samples. The first data set was artificially balanced to contain an equal number of positive and negative responses, in the second data set only one tenth of the samples belonged to the positive response class as in usually strongly imbalanced direct marketing problems. The evaluation criteria used for measuring the influence of the discretization procedure on the classification problem were mutual information, predicted 50 % response rate based on Naive Bayes, and classification accuracy of C5 classifier. Each experiment was conducted with a randomly selected training set and a testing set of the same size, the results shown are based on the testing set. All the experiments were repeated 25 times. In the case of mutual information, all the variables of each data set were discretized and the mutual information of the discretized variable and the output variable were measured on the test data. From each data set, 10 most relevant variables were chosen and in order to create different subproblems randomly selected subsets of four variables were used for building classifiers. Using response rate together with the Naive Bayes, the possibly imbalanced class priors present in the data do not have any effect. In C5 classifier, a fixed cost matrix was given to flatten out the imbalanced class distribution. All the experiments were repeated with the goal of discretizing the continuous variables to 4, 6, and 8 bins. The results are shown in terms of relative performance in Fig 1.

## 4 Discussion

Measuring the relative scores by mutual information, the approaches that take into account the class membership of data prove to be superior. Ranking of the methods remains the same in both the balanced and the imbalanced data sets. In general, the addition of bins improves the performance of the discretization methods. Moreover, the mutual information approach is better than the novel method in case of low number of bins, whereas the novel method was superior when the number of bins was bigger, even though mutual information is used as the assessment measure. The importance of the entropy term in the novel method increases along the number of bins. Of the simple methods, which ignore the available output information in the classification problem, the entropy-based method is better than the equal width interval method.

Using 50 % response rate based on Naive Bayes classifier, results are somewhat more difficult to interpret. In this case it is important to note that each variable is treated separately, which is likely to increase the independence of the discretized variables compared with the original ones. It seems that the novel method is superior to all other methods, although the large variance on the estimates makes this subject to a debate. For example, in the case of data set 1 and the experiment with eight intervals, the median of the novel method is the best, 75 % confidence interval is similar to others, and finally the 95 % confidence limits are much worse than in the case of mutual information. On the other hand, the median performance of the novel method proves to be the best in most cases.



**Figure 1.** Relative scores of the discretization methods measured mutual information are shown for the data set 1 (first row, left panel) and data set 2 (first row, right panel). Relative scores of the discretization methods measured by 50 % response rate of Naive Bayes are shown for the data set 1 (second row, left panel) and data set 2 (second row, right panel). Relative scores of the discretization methods measured by classification performance achieved with C5 classifier are shown for the data set 1 (third row, left panel) and data set 2 (third row, right panel). In all figures, the horizontal axis is divided to three sections for experiments with four, six and eight bins. The order of discretization methods in each section is equal width interval (1), maximum entropy (2), maximum mutual information (3), and maximum mutual information with entropy (4). The performance of repeated experiments are visualized with median, 25 % and 75 % percentiles. In addition, 95 % confidence interval is shown with dashed lines.

In case of a C5 classifier, none of the methods outperforms others, especially on the first data set. The variance of the estimates is also relatively large as to make accurate judgments. In the second data set, however, equal width interval approach is clearly worse than the other presented methods. One possible reason for the questionable performance of the tree classifier could be that our discretization works for each input variable separately, whereas optimal creation of the decision tree would take into account the interdependencies between variables. Using the novel discretization method, these interdependencies are essentially ignored and the solution is likely to weaken the interdependencies between discretized input variables. Taking all the variables into account at the same time may be seen beneficial in this context as proposed in [2].

## 5 Summary

Methods for quantizing continuous input variables in classification problems were presented. Relative merits of the equal width interval, maximum entropy, maximum mutual entropy and the novel maximum mutual information with entropy approaches were compared with two data sets from direct marketing problems using three criteria. Concluding, none of the tested approaches would be preferred over others whenever the C5 decision tree is to be used for modeling. On the other hand, the novel method proposed in this paper would be recommended for Naive Bayes classifiers where it may lead to performance improvements.

## References

1. J. Catlett. On changing continuous attributes into ordered discrete attributes. In *Machine Learning - EWSL-91*, volume 482, pages 164–178. Springer, March 1991.
2. Michał R. Chmielewski and Jerzy W. Grzynala-Busse. Global discretization of continuous attributes as preprocessing for machine learning. *International Journal of Approximate Reasoning*, 15:319–331, 1996.
3. Thomas M. Cover and Joy A. Thomas. *Elements of Information theory*. Wiley Series in telecommunications. John Wiley & Sons, 1991.
4. Usama M. Fayyad and Keki B. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In *Proceedings of IJCAI-93*, volume 2, pages 1022–1027. Morgan Kaufmann Publishers, August/September 1993.
5. R. Kerber. Chimerge: Discretization of numeric attributes. In *Proceedings of the Tenth National Conference on Artificial Intelligence*, pages 123–128. AAAI Press/MIT Press, 1992.
6. Huan Liu and Hiroshi Motoda. *Feature selection for knowledge discovery and data mining*. Kluwer International Series in Engineering and Computer Science, Secs 454. Kluwer Academic Publishers, 1998.
7. J. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.
8. J. Ross Quinlan. *C4.5 : Programs for Machine Learning*. Morgan Kaufmann Series in Machine Learning. Kluwer Academic Publishers, 1993.
9. A.K.C. Wong and D.K.Y. Chiu. Synthesizing statistical knowledge from incomplete mixed-mode data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9(6):796–805, 1987.