EMPIRICAL COMPARISON OF METHODS OF DATA DISCRETIZATION IN LEARNING PROBABILISTIC MODELS

Michał Wójciak, Anna Łupińska–Dubicka

Faculty of Computer Science, Bialystok University of Technology, Białystok, Poland

Abstract: Very often statistical method or machine learning algorithms can handle discrete attributes only. And that is why discretization of numerical data is an important part of the pre–processing. This paper presents the results of the problem of data discretization in learning quantitative part of probabilistic models. Four data sets taken from UCI Machine Learning Repository were used to learn the quantitative part of the Bayesian networks. The continuous variables were discretized using two supervised and two unsupervised discretization methods. The main goal of this paper was to study whether method of data discretization in given data set has an influence on model's reliability. The accuracy was defined as the percentage of correctly classified records.

Keywords: discretization, continuous feature, probabilistic models, Bayesian networks, classification

1. Introduction

Often data are given in the form of continuous values. If their number is huge, building a proper model for such data can be difficult. Moreover, many data mining algorithms operate only in discrete variable space. For instance, probabilistic models such Bayesian networks, require discrete values for their nodes. In addition, discretization also can work as a variable (feature) selection method that can significantly impact the performance of classification algorithms used in the analysis of high–dimensional data.

This paper presents the results of data discretization in problem of learning quantitative part of probabilistic models, in particular one of their prominent members – Bayesian networks. One of the most important features of Bayesian networks is the

Advances in Computer Science Research, vol. 14, pp. 177-192, 2018. DOI: 10.24427/acsr-2018-vol14-0011

fact that they provide an elegant mathematical structure for modeling complicated relationships among random variables while keeping a relatively simple visualization of these relationships.

The experiments involved learning the conditional probability distribution of models created on the basis of four data set taken from UCI Machine Learning Repository [23]: *Banknote authentication*, *Heart disease*, *Image segmentation* and *Abalone*. The main purpose of this article was to study whether method of data discretization in given data set has an influence on model's reliability. The accuracy was defined as the percentage of correctly classified records.

The remainder of this paper is structured as follows. Section 2. explains the problem of data discretization and shortly outlines the methods of it. Section 3. explains the basic concepts of Bayesian networks. Section 4. introduces selected data sets and presents created Bayesian network models. Section 5. presents the results of experiments conducted on data sets with implemented methods of data discretization. Section 6. concludes the paper and indicates possible directions for further research.

2. Data discretization

Discretization of numerical data is an important part of the pre–processing, necessary in typical processes of knowledge discovery and data mining. Transforming continuous attribute values into their discrete counterparts enables further analysis using data mining algorithms, such as learning parameters of probabilistic models (existing algorithms mainly assume discrete variables for nodes). Even in the absence of such a requirement, discretization allows accelerating the process of data mining and increasing the accuracy (accuracy) of predictions (classification) [3].

According to the surveys [3,6,8,11] and finally the advanced review [17] many different discretization algorithms have been proposed in the last two decades. Their authors used a different approach, derived from statistics, machine learning, information theory and logic. In order to be able to better understand these issues, the comparative criteria used should be taken into account [8,17]:
Local or *global* discretization – in the case of global discretization, the entire problem space is considered at the same time. Local discretization at the moment solves only a selected subproblem. The division is made on the basis of a limited amount of information.
Supervised or *unsupervised* – during supervised discretization the decision (class) of each of the objects is taken into account. The main premise of supervision is to separate instances having different decisions from each other. If the method does not use information given by classes, it is known as unsupervised. The advantage of unsu-

pervised methods is the ability to use them to discretize databases that do not have a decision attribute.

Static or *dynamic* discretization – the static method of attribute dependence is not taken into account. During one discretization cycle, the maximum number of intervals for a given attribute is obtained, regardless of the others. Dynamic methods simultaneously consider cutting for many features, which allows the use of high–level dependencies.

Due to the multitude of existing discretization methods, there is a need to introduce quality assessment criteria [8]:

Number of intervals – the fewer intervals, the simpler the result table. It can be seen that the problem of minimizing the number of intervals is synonymous with minimizing the number of cuts.

Number of inconsistencies – it would be best if discretization did not introduce additional inconsistencies over those contained in the input database. Otherwise significant information can be lost.

Accuracy of predictions – defines how discretization helps to improve predictions. It should be emphasized that this criterion depends on the classification method and the procedure of conducting the experiment. It should also be emphasized that only the first two criteria are directly measurable. The accuracy of predictions is a function of both discretization and the classification algorithm. These criteria do not indicate unambiguously which of the tested methods is the best. Depending on the chosen base and the expected results, the weight of each criterion may fluctuate. What's more, there is no discretization method that would have an advantage over all criteria at the same time.

As mentioned before, there are several method to discretize continuous variables. Below short description of four of them, used in this paper, is presented.

OneR algorithm [3,6] is a supervised method of discretization, using information about the class. Values that have been previously sorted are divided into intervals whose limits are set based on both continuous values and class labels. There is an assumption that each of the intervals must contain a minimum number of examples equals to k, where k is usually set to six. This assumption does not apply to the last range, which contains other, ungrouped examples. The exception occurs when the next attribute has the same class as most examples in a given range.

Chi merge algorithm [6,12,13,17] is a simple, supervised algorithm that uses the χ^2 statistic to discretize numeric attributes. It checks each pair of adjacent rows in order to determine if the class frequencies of the two intervals are significantly different. It tests the hypothesis that the two adjacent intervals are independent. If the hypothesis is confirmed the intervals are merged into a single interval, if not, they remain separated.

Equal–Width Discretization (EWD) algorithm [6,7,9,14,17,18] belongs to class unsupervised methods. The main assumption of this algorithm is to divide the data set into k intervals determined by the user of the algorithm. It first finds the minimum and maximum values of every variable, X_i, and then divides this range into a number, k, of user–specified, equal–width intervals. The discussed algorithm has one fundamental disadvantage – in most cases all elements of the data set will be unevenly distributed in groups. In extreme cases, even empty sets may be created or one set having more elements than all the others combined. Therefore, it is very important to properly adjust the k parameter to minimize this span.

Equal–Frequency Discretization (EFD) algorithm [1,6,7,9,14,17], like **EWD**, is a representative of the unsupervised discretization methods. It determines the minimum and maximum values of the variable X_i, sorts all values in ascending order, and divides the range into a user–defined number of intervals k, in such a way that every interval contains the equal number of sorted values. Each of these intervals contains N/k elements, where N means the total number of X_i variable values. This method eliminates the possibility of disproportionate intervals because the entire interval $< X_{min}; X_{max} >$ containing specific values is divided into compartments in terms of a specific number of elements, not on the basis of ranges of values.

3. Bayesian Networks

Bayesian networks (also knows as belief networks or causal networks, BNs) [16] are a special case of probabilistic models. They have found many practical application over the years, among them the best known and probably the most successful are decision support systems. Bayesian networks offer natural mechanism for reasoning under uncertainty, when we do not have access to the full knowledge of the analyzed phenomenon. They allow for easy and readable representations of the actual relationships, which makes it easier to apply the real relationships. Furthermore, Bayesian networks enable a combination of *a priori* knowledge and collected data.

Formally, a Bayesian network $\mathcal{B}$ is a pair $<\mathcal{G},\Theta>$, where $\mathcal{G}$ is an acyclic directed graph in which nodes represent random variables $X_1,\ldots,X_n$ and edges represent direct dependencies between pairs of variables [16]. Θ represents the set of parameters that describes the probability distribution for each node X_i in $\mathcal{G}$, conditional on its parents in $\mathcal{G}$, i.e., $P(X_i|Pa(X_i))$. Often, the structure of the graph is given as a causal interpretation, convenient from the point of view of knowledge engineering and user interfaces. BNs allow for computing probability distributions over subsets of their variables conditional on other subsets of observed variables. The joint probability distribution is represented as follows:

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{N} P(X_i | Pa(X_i)) \tag{1}$$

where $Pa(X_i)$ represents set of parents of X_i.

Using Equation 1 the occurrence of a specific state of all network variables can be determined, knowing only their local conditional probabilities. Knowing the values of the variable that do not have parents in the graph (the root cause), the expected value of the other nodes can be calculated, since each variable in the network depends on them either directly or indirectly.

Note that in the Equation 1, probability of a random variable X_i depends only on the states of its parents. This simplification resulting from the assumption of conditional independence of variable, allows to represent the joint probability distribution more compactly. This is particularly significant in the case of large–scale networks with a large number of variables. If a network consists of n binary nodes, then the full joint probability distribution would require storing 2^n values. Using the factored form would require $n2^k$, where k is the maximum number of parents of a node.

4. Data sets and Models

For the purpose of this work, the UCI Machine Learning Repository [23] has been searched and four data set containing continuous attributes were chosen: *Banknote authentication*, *Hearth disease*, *Statlog (image segmentation)*, and *Abalone*. Then, for particular data set the probabilistic model were constructed. The graphical structure of a Bayesian network represents a set of domain variables and relationships among them.

4.1 Banknote authentication

The *Banknote authentication* [24] data set is a collection of data extracted from images of original and fake banknotes. The images were created using an industrial camera used to control the print quality. The resulting images are 400 x 400 pixels and 660 dpi. To extract interesting data from these images, a wavelet transformation was used. The data set contains four continuous variable and one decision class. The data set contains 1372 objects, however, some of them were removed due to the fact that they contained missing elements, which could significantly lead to incorrect results of classification quality. Figure 1 presents the Bayesian network created on the basis of *Banknote authentication* data set.

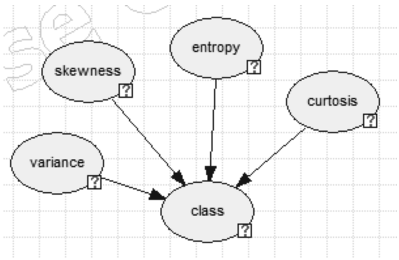

Fig. 1. A Bayesian network model of *Banknote authentication* data set.

4.2 Hearth disease

The *Hearth disease* [25] is a data set presenting knowledge about the diagnosis of a patient's heart disease. The measurements were carried out in four locations around the world: in Cleveland (United States, Ohio), in Budapest (Hungary), in Zurich (Switzerland) and Long Beach (United States, California). For the purposes of the work, only data collected by the clinic in Cleveland was used, as it was the only one that was processed. The data does not contain real information about the personal data of each of the patients examined. Data for the analysis of knowledge offered by the *Hearth disease* collection has thirteen attributes (including four continuous ones) and fourteenth, which is a decision class. The collection contains only 303 objects. Figure 2 presents the Bayesian network created on the basis of *Hearth disease* data set.

4.3 Statlog (image segmentation)

The *Image segmentation* [26] data set presents data extracted from seven pictures presenting brick, sky, vegetation, cement, window, path and grass. These images were adapted to analyze each pixel. All observations included in this set are presented for nine–pixel blocks (3x3). The data was presented using 19 attributes, where 18 of them were continuous attributes, one is a constant attribute and the twentieth attribute was a decision class. The data set contains 2 310 observations. There are no missing values in it, therefore all objects have been included in the research. Figure 3 presents the Bayesian network created on the basis of *Image segmentation* data set.

4.4 Abalone

Abalone [27] is a data set presenting a few basic physical data of abalone – an edible mollusc of warm seas, with a shallow ear-shaped shell lined with mother–of–pearl

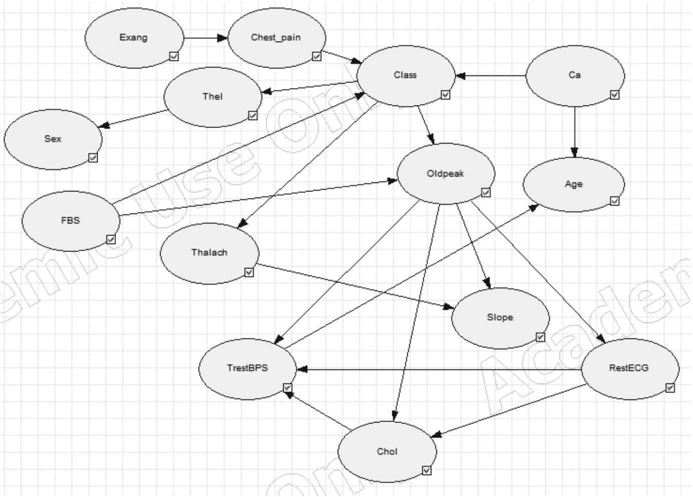

Fig. 2. A Bayesian network model of *Hearth disease* data set.

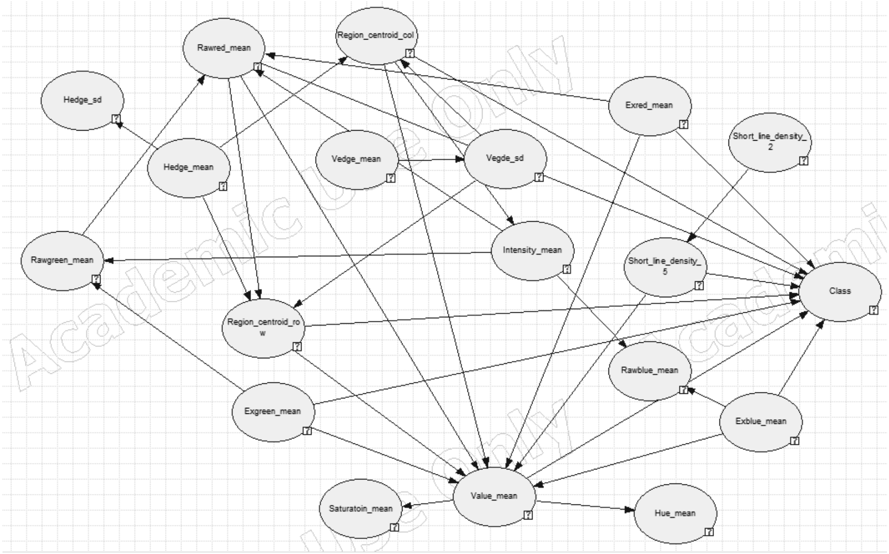

Fig. 3. A Bayesian network model of *Statlog (image segmentation)* data set.

and pierced with a line of respiratory holes. Based on these parameters, the age of the abalone is determined. The age of the snail is determined by counting the number of rings on the body using a microscope, but it is a very arduous and time–consuming process. The purpose of the collection is to determine the age of this creature without using a microscope. The set is presented by one nominal attribute, seven continuous attributes and ninth, which is a decision class. The above set contains 4177 observations and contains no missing data. Figure 4 presents the Bayesian network created on the basis of *Abalone* data set.

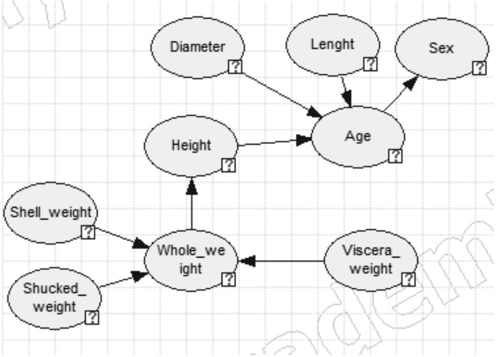

Fig. 4. A Bayesian network model of *Abalone* data set.

5. Experiments and Results

The main goal of the conducted experiments was to examine how particular methods of data discretization affect the quality of classification of created Bayesian network models, which were learned using discretized data. The quality assessment was determined by means of 10–fold cross–validation. Parameters of discretization methods for research purposes were selected as follows:

- The OneR method as a supervised method does not require specification of the interval length parameter because it is set to the value 6 in advance. However, it has also been decided to test additional values: 7 and 8.
- The Chi Merge method requires the value of parameter χ^2. During the research confidence coefficients of 0.1, 0.2, 0.3 and 0.4 were used. The degree of freedom was determined based on the number of classes in the classification attribute. For the *Banknote authentication* set it was the value of 1, for *Hearth disease* the value of 4, for *Image segmentation* the value of 6 and for the set of *Abalone* – 27.
- In case of EWD and EFD methods, the number of intervals k was set from 2 to 12.

The empirical part of the paper was performed using SMILE, an inference engine, and GeNIe Modeler, a development environment for reasoning in graphical probabilistic models, both developed at BayesFusion LLC, and available at [22].

5.1 OneR method

Table 1 shows the results obtained for the OneR method. OneR as a supervised method should not take parameters and its task is to classify based on the default value of the minimum interval length equal to $k = 6$. However, as part of the research, it has been decided to evaluate the quality of the network classification based on the discretized sets of variables not only for the minimum number of elements 6 but also for the minimum number of elements equal to 7 and 8. In the case of the *Hearth disease* and *Image segmentation* data sets, this modification brought a positive result, as the quality assessment increased relative to the result for the default parameter. In the case of the hearth disease data set, an increase of over 12 percentage points was achieved (for $k = 8$) and the best quality result for this set was obtained from among all the methods studied. *Image segmentation* achieved a slight improvement of about 2.8% for intervals with $k = 7$.

Table 1. The classification accuracy for the OneR discretization method for different intervals length.

	k=6	k=7	k=8
Banknote	**91.62%**	91.18%	89.80%
H. disease	63.37%	74.59%	**75.58%**
Image seg.	74.51%	**77.23%**	76.17%
Abalone	**45.70%**	41.90%	42.40%

5.2 Chi Merge method

Table 6.4 presents the classification results of the network for sets discretized using the Chi Merge method. The best results were obtained for the smallest value of a confidence test of 0.1. For the *Abalone* and *Image segmentation* data sets the best results for values of 0.2 and 0.4 were obtained respectively. In addition, it can be observed that with the increase of the confidence value, the overall quality of the classification decreased. The exception is the set of *Abalone* for which the percentage of accurately classified objects grew with the increase of this coefficient value reaching 56.33%,

Table 2. The classification accuracy for the Chi Merge discretization method for different values of χ^2.

	$\chi^2 = 0.1$	$\chi^2 = 0.2$	$\chi^2 = 0.3$	$\chi^2 = 0.4$
Banknote	**96.23%**	96.04%	94.54%	92.28%
H. disease	**61.22%**	60.87%	59.79%	58.12%
Image seg.	69.43%	**69.53%**	68.71%	66.82%
Abalone	53.68%	52.91%	55.17%	**65.33%**

which is the highest quality value measured for this set among all the analyzed algorithms. However, these fluctuations are not big for any of the data sets – the difference between the maximum and minimum closing in around 3-4 percentage points.

5.3 EWD method

Table 3 presents the results obtained for the EWD method, taking into account the value of parameter *k* (length of the interval). It can be noticed that the overall accuracy of the classification for the *Abalone* set is very low. The most probable reason is that the decision class attribute contains as many as 28 decision classes that include very diverse number of objects assigned to them. For the *Image segmentation* data set, empty values mean that calculations were impossible due to hardware limitations and the complexity of the Bayesian network. The EWD method worked well in the case of the *Banknote authentication* data set, where the results are very high, mostly exceeding 90%. The biggest differences between the maximum and minimum values occur for the *Image segmentation* data set – the difference between the maximum and minimum values is about 15%. The lowest range occurs for the *Abalone* data set, which is around 3.6%.

Table 3. The classification accuracy for the EWD discretization method for different intervals length.

	k=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10	k=11	k=12
Banknote	85.35%	94.1%	95.63%	97.67%	97.08%	**98.03%**	97.45%	95.77%	92.71%	94.9%	94.75%
H. disease	57.76%	57.43%	56.77%	56.11%	58.09%	58.75%	56.44%	57.10%	**59.41%**	55.45%	57.10%
Image seg.	68.87%	77.66%	79.57%	79.26%	80.74%	83.27%	**83.94%**	–	–	–	–
Abalone	22.60%	22.07%	24.28%	24.37%	25.19%	23.63%	**26.22%**	24.66%	24.83%	24.92%	23.94%

5.4 EFD method

Table 4 shows the results obtained for EFD method taking into account the value of parameter *k*, i.e. the length of the interval. The first conclusion is that the obtained results for each data set are weaker than in case of EWD method. In the case of the *Image segmentation* collection, this result is definitely weaker and the difference was around 15 percentage points. In the case of other collections, they are about 1–2 percentage points. Again, a very poor classification result was achieved for the *Abalone* data set – about 23% on an average level. In turn, the best results were obtained by the *Banknote authentication* data set, with the difference that for the higher values of parameter *k*, the classification quality for this set began to decrease. When comparing the maximum quality results of the network classification, it can be observed that for EFD method they are higher for the *Banknote authentication* and *Hearth disease* data sets, and lower for the *Image segmentation* and *Abalone*. Very clear difference in the quality of EWD and EFD methods can be seen in the case of the *Image segmentation* data set.

Table 4. The classification accuracy for the EFD discretization method for different intervals length.

	k=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10	k=11	k=12
Banknote	88.12%	92.06%	97.89%	**98.40%**	98.25%	97.38%	94.9%	93.44%	92.27%	88.34%	86.88%
H. disease	58.42%	54.79%	58.42%	**60.07%**	55.45%	57.10%	55.78%	54.13%	57.43%	53.14%	55.12%
Image seg.	63.85%	**78.87%**	70.87%	60.3%	50.74%	54.23%	57.82	–	–	–	–
Abalone	21.43%	21.33%	23.13%	**24.83%**	24.44%	24.37%	23.65%	21.45%	21.52%	22.62%	22.53%

5.5 Methods comparison

Figure 5 presents the comparison of classification accuracy of different discretization methods for all data sets. For each analyzed data sets the best result for each particular algorithm was chosen and presented in the chart. As can be noticed each of the data sets received the best result for a different method. For the *Hearth disease* data set (75.58%) the OneR method (for the interval length $k = 8$) turned out to be the best. The Chi Merge method achieved the highest classification result (56.33%) for the *Abalone* data set (with the confidence coefficient $\chi^2 = 0.1$). The EWD method with number of intervals $k = 5$ proved to be the best for the *Banknote authentication* data set (98.40%). On the other hand, the EFD method achieved the best result for the *Image segmentation* data set equal to 83.94% for the length of the interval $k = 8$. At this point it should be mentioned that the supervised methods present generally higher quality than the unsupervised ones. However, this trend is not clearly visible in obtained results. In the case of the analyzed data sets, the proportions were divided in half.

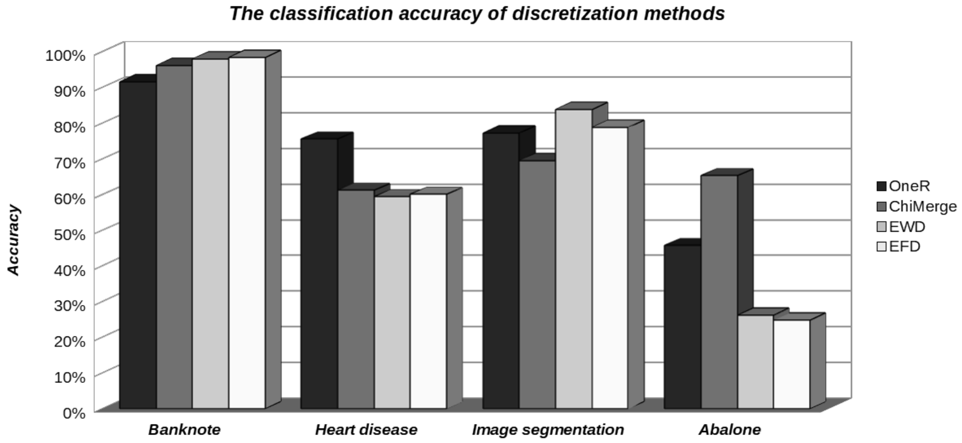

Fig. 5. The comparison of classification accuracy of different discretization methods for all data sets.

Regardless to the discretization method, the best results were achieved for *Banknote authentication* data set – each method's classification accuracy was above 90%. The probabilistic model created for this set was the only one in the form of a naïve Bayesian network. Taking into consideration the fact, that this data set contained only about 1 000 objects, such result can confirm the hypothesis stated in [3] that the accuracy of classification can depend on the complexity of created model and almost any discretization method results in significant performance gains for naïve Bayes networks.

The overall accuracy of the classification for the *Abalone* data set was very low – below 50% in most cases. The most probable reason is that the decision class attribute contains as many as 28 decision classes that include very uneven number of objects assigned to them. In such case, the Chi Merge method proved to be the best with the highest result about 65%.

6. Conclusion

The conducted research confirmed the belief that there is no universal discretization method, which gives the best result in every data set. Therefore, it is very important to carefully analyze the data on which the tests will be carried out. In order to choose the most effective method, it is worth conducting an experiment using few discretization methods. Basing on such experiment, the appropriate method should be chosen for the given data set.

The *Banknote authentication* data set, regardless of the method used, offers the results of measured quality above 90%. Such result can be the basis to hypothesise that the quality of classification does not only depend on the number and quantity of the observations examined, but also on the designed network model and its complexity. However, further experiments should be carried out using only naíve Bayes network models to check if they produce similar results or not.

Research has also shown that models created for data sets such as *Abalone*, which after the discretization process have many decision classes, achieve very poor classification results regardless of the chosen method. For such type of data sets, the Chi Merge method seemed to be a more universal method that produces good results, regardless of the type or size of data input, relative to other methods of discretization of sets. This does not mean, however, that it always achieved the best results. In some literature [15,20], Chi Merge method is reported to achieve lower classification error than those trained on data pre–processed by the other discretization methods. However, further experiments would be advisable to confirm its effectiveness in the case of data sets with a large number of attributes in the decision class.

At this point, it is also worth adding that in some literature [3] the supervised methods were reported to achieve better results than the unsupervised ones while the contradicting results were obtained by some others [2]. Also the results obtained in this article (as well as in work [21]) do not confirm the superiority of supervised method over unsupervised and vice versa. Therefore, further experimental comparison of of the unsupervised methods versus some of the common supervised methods should be carried out. However, the unsupervised methods will still remain as the only discretization option when we do not have prior known class labels required by the supervised methods.

References

[1] R. Abraham, J. B. Simha, S. S. Iyengar, A comparative analysis of discretization methods for Medical Datamining with NaĂŻve Bayesian classifier, Information Technology, 2006.

[2] E. Cantú–Paz, Supervised and unsupervised discretization methods for evolutionary algorithms, In proc. Of the genetic and evolutionary computation conference, pp. 213–216, 2001.

[3] J. Dougherty, R. Kohavi, M. Sahami, Supervised and Unsupervised Discretization of Continuous Features, Machine Learning: Proceedings of the Twelfth International Conference, 1995.

[4] A. Ekbal, Improvement of Prediction Accuracy Using Discretization and Voting Classifier, The 18th International Conference on Pattern Recognition, IEEE, 2006.

[5] N. Friedman, D. Geiger, M. Goldszmidt, Bayesian network classifiers, Machine Learning 29 (1997), 131–163.

[6] S. García, J. Luengo, J. A. Sáez, V. López, F. Herrera, Survey of discretization techniques, Taxonomy and empirical analysis in supervised learning, IEEE Transactions on Knowledge and Data Engineering, vol. 25(4), pp. 734–750, 2013.

[7] M. Hacibeyoğlu, M. H. Ibrahim, Comparison of the effect of unsupervised and supervised discretization methods on classification process, International Journal of Intelligent Systems and Applications in Engineering, vol. 4(1), pp. 105–108, 2016.

[8] F. Hussain H. Liu C. L. Tan M.Dash, Discretization: An enabling technique, Data Mining and Knowledge Discovery (2002) 6: 393.

[9] , F. Kaya, Discretizing Continuous Features for Naive Bayes and C4. 5 Classifiers, University of Maryland publications: College Park, MD, USA, 2008.

[10] R. Kerber., Chi Merge: Discretization of numeric attributes, In Proc. Tenth National Conference on Artificial Intelligence, pp. 123–128. MIT Press 1992.

[11] S. Kotsiantis, D. Kanellopoulos, Discretization Techniques: A recent survey, GESTS International Transactions on Computer Science and Engineering, vol.32 (1), 2006.

[12] K. Lavangnananda, S. Chattanachot, Study of discretization methods in classification, 9th International Conference on Knowledge and Smart Technology, pp. 50–55, IEEE, 2017.

[13] P. Lehtinen, M.i Saarel, T. Elomaa, Online Chi Merge Algorithm, Springer, 2012.

[14] D. M. Maslove, T. Podchiyska, H. J. Lowe Discretization of continuous features in clinical datasets J Am Med Inform Assoc., vol. 20(3), pp. 544–553, 2013.

[15] I. Mitov, I. Krassimira, M. Krassimir, V. Velychko, P. Stanchev, K. Vanhoof Comparison of discretization methods for preprocessing data for pyramidal growing network classification method, Information Science & Computing, International Book Series, Number 14, pp. 31–39, 2009.

[16] J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Morgan Kaufmann PUBLISHERs, Inc., San Mateo, CA, 1988.

[17] S. Ramírez–Gallego, S. García, H. Mouriño–Talín, D. Martínez–Rego, V. Bolón–Canedo, A. Alonso–Betanzos, J. M. Benítez, F. Herrera Data discretization: taxonomy and big data challenge, WIREs Data Mining Knowledge Discovery, 2015.

[18] A. Rayner, Discretization Numerical Data for Relational Data with One-to-Many Relations

[19] P. Spirtes, C. Glymour, R. Scheines, Causation Prediction and Search, Springer-Verlag, New York, 1993.

[20] C. Zeynel, Y. Figen, Comparison of Chi-square based algorithms for discretization of continuous chicken egg quality traits, Journal of Agricultural Informatics, vol. 8, pp. 13–22, 2017.

[21] C. Zeynel, Y. Figen, Unsupervised Discretization of Continuous Variables in a Chicken Egg Quality Traits Dataset, Turkish Journal of Agriculture – Food Science and Technology, vil. 5, pp. 315–320, 2017.

[22] BayesFusion, LLC, `[https://www.bayesfusion.com/]`, Accessed 19-08-2017.

[23] UCI Repository of machine learning databases, `[http://archive.ics.uci.edu/ml/datasets.html]`, Accessed 05-04-2017,

[24] Volker Lohweg University of Applied Sciences, Ostwestfalen-Lippe `[https://archive.ics.uci.edu/ml/datasets/Banknote+authentication]`, Accessed 03-07-2017.

[25] Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D., University Hospital, Zurich, Switzerland: William Steinbrunn, M.D., University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D., V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D, `[https://archive.ics.uci.edu/ml/datasets/heart+disease]`, Accessed 10-07-2017.

[26] Vision Group, University of Massachusetts, `[https://archive.ics.uci.edu/ml/datasets/Statlog+(Image+segmentation)]`, Accessed 01-06-2017.

[27] W. J. Nash, T. .L Sellers, S. R. Talbot, Andrew J. Cawthorn, W. B. Ford, The Population Biology of Abalone (_Haliotis_ species) in Tasmania. I. Blacklip Abalone (_H. rubra_) from the North Coast and Islands of Bass Strait, Sea Fisheries Division, Technical Report No. 48 (ISSN 1034-3288), 1994. `[https://archive.ics.uci.edu/ml/datasets/abalone]`, Accessed 12-07-2017.

PORÓWNANIE METOD DYSKRETYZACJI DANYCH W UCZENIU MODELI PROBABILISTYCZNYCH

Streszczenie Bardzo często algorytmy uczenia maszynowego są nie są przystosowane do korzystania ze zmiennych ciągłych. Z tego powodu dyskretyzacja danych jest istotną częścią wstępnego przetwarzania. W artykule przedstawiono wyniki prac nad problemem dyskretyzacji danych w uczeniu modeli probabilistycznych. Cztery zestawy danych pobrane z repozytorium uczenia maszynowego UCI zostały wykorzystane do nauczenia parametrów ilościowej części sieci bayesowskich. Występujące w wybranych zbiorach zmienne ciągłe były dyskretyzowane przy użyciu dwóch metod nadzorowanych i dwóch nienadzorowanych. Głównym celem tego artykułu było zbadanie, czy metoda dyskretyzacji danych w danym zbiorze ma wpływ na niezawodność modelu. Dokładność metod była definiowana jako odsetek poprawnie sklasyfikowanych rekordów.

Słowa kluczowe: dyskretyzacja, zmienne typu ciągłego, modele probabilistyczne, sieci Bayesa, klasyfikacja

Artykuł częściowo zrealizowano w ramach pracy badawczej S/WI/2/2018.