Turkish Journal of Agriculture - Food Science and Technology

Available online, ISSN: 2148-127X
www.agrifoodscience.com,
Turkish Science and Technology

Unsupervised Discretization of Continuous Variables in a Chicken Egg Quality Traits Dataset

Zeynel Cebeci*, Figen Yildiz

Department of Biometry and Genetics, Faculty of Agriculture, Çukurova University, 01330 Adana, Turkey

ARTICLE INFO

Research articles

Received 26 October 2016
Accepted 14 March 2017

Keywords:
Data preprocessing
Discretization
Unsupervised discretization
Egg quality traits
Classification trees

*Corresponding Author:
E-mail: zcebeci@cu.edu.tr

ABSTRACT

Discretization is a data pre-processing task transforming continuous variables into discrete ones in order to apply some data mining algorithms such as association rules extraction and classification trees. In this study we empirically compared the performances of equal width intervals (EWI), equal frequency intervals (EFI) and K-means clustering (KMC) methods to discretize 14 continuous variables in a chicken egg quality traits dataset. We revealed that these unsupervised discretization methods can decrease the training error rates and increase the test accuracies of the classification tree models. By comparing the training errors and test accuracies of the model applied with C5.0 classification tree algorithm we also found that EWI, EFI and KMC methods produced the more or less similar results. Among the rules used for estimating the number of intervals, the Rice rule gave the best result with EWI but not with EFI. It was also found that Freedman-Diaconis rule with EFI and Doane rule with EFI and EWI slightly performed better than the other rules.

Türk Tarım – Gıda Bilim ve Teknoloji Dergisi, 5(4): 315-320, 2017

Tavuk Yumurtası Kalite Özellikleri Veri Setindeki Sürekli Değişkenlerin Yönetimsiz Ayrıklaştırılması

MAKALE BİLGİSİ

Araştırma makalesi

Geliş 26 Ekim 2016
Kabul 14 Mart 2017

Anahtar Kelimeler:
Veri önişleme
Ayrıklaştırma
Yönetimsiz ayrıklaştırma
Yumurta kalite özellikleri
Sınıflama ağaçları

*Sorumlu Yazar:
E-mail: zcebeci@cu.edu.tr

ÖZET

Ayrıklaştırma, sınıflama ağaçları ve birliktelik kuralları çıkarma gibi bazı veri madenciliği algoritmalarında sürekli değişkenleri kesikli değişkenlere dönüştüren bir veri önişleme adımıdır. Bu çalışmada eşit genişlikli aralıklar (EWI), eşit frekanslı aralıklar (EFI) ve K-ortalamalar kümelemesi (KMC) yöntemleri, bir tavuk yumurtası kalite özellikleri veri setinde 14 sürekli değişkenin ayrıklaştırmasındaki performansları bakımından deneysel olarak karşılaştırılmıştır. Bu yönetimsiz ayrıklaştırma yönteminin sınıflama ağacı modelleri için öğrenme hatalarını düşürdüğü ve doğruluğu yükselttiği belirlenmiştir. C5.0 sınıflama ağacı algoritması kullanılarak uygulanan modelin öğrenme hatası ve test doğruluğu kullanılarak yapılan karşılaştırmalara göre EWI, EFI ve KMC yöntemlerinin birbirine yakın sonuçlar verdikleri görülmüştür. Yöntemlerde aralık sayısını hesaplamak için kullanılan kurallar arasında, Rice kuralı EFI'de olmamakla birlikte EWI ile en iyi sonucu üretmiştir. Ayrıca EWI ile Freedman-Diaconis kuralının ve EFI ve EWI'nin her ikisinde ise Doane kuralının diğer kurallardan kısmen daha iyi oldukları saptanmıştır.

Introduction

Discretization is a data pre-processing task transforming a continuous variable to a discrete one by splitting the range of values into a finite number of subranges called intervals, buckets or bins. For example, chicken egg weight (gram) as an example of continuous variables can be transformed into 4 categories as: (1) small: -53, (2) medium: 53-62, (3) large: 63-73, (4) very large: 73- . Like the above given example, discretization divides a continuous data range into a finite number of intervals, and labels them as categories or classes. Discretization is frequently required by data mining applications because the most algorithms for feature selection, classification and association rules extraction generally handle discrete variables.

Although the domain knowledge and experience should be consulted for increasing the success in discretization, we mostly work with the data-driven discretization algorithms or methods because the prior knowledge is often unavailable (Muhlenbach and Rakotomalala, 2005). According to the surveys by Dougherty et al. (1995), Liu et al. (2002), Kotsiantis and Kanellopoulos (2006), García et al. (2013) and recently an advanced review by Ramírez-Gallego et al. (2015), many discretization algorithms have been proposed in the last two decades because of the increasing demand by some popular data mining applications. In general the discretization algorithms are categorized as the supervised and the unsupervised algorithms. However the supervised algorithms use the prior information about datasets the unsupervised algorithms do not use such kind of information. Although the choice of a discretization algorithm largely depends on the user needs as well as on the structure of data to be discretized (Dash et al., 2011), the unsupervised methods are commonly in use because of their simplicity.

In animal science, the real data set usually consists of the continuous variables that measured in the interval or ratio scales. On the other hand, the research works are limited for benefitting from data mining applications on the datasets gathered in animal production environments. In this study, it was aimed to compare the performances of three common unsupervised methods in order to use in a forthcoming study to be conducted for the association rules mining between the chicken egg quality traits.

Materials and Methods

Material

This study used 4320 eggs obtained from white layer hens at the Poultry Research and Application Farm of Animal Science Department of the Faculty of Agriculture, Cukurova University. 600 layers from 3 successive rearing groups were used. One of the groups was white native strain Atabey (A) and two of them were commercial white layers Decalp (D) and Nick (N). They were raised at apartment cage systems at three floors. Every genotype group consists of 200 layer hens. There were 6 replicates for all genotype groups. Totally 150 cages were used and 4 layer hens were allocated to each cage. At the end of each week the eggs from cages were collected and labeled for measuring the quality traits listed in Table 1.

In the analyzed dataset, there was 1 class variable (genotype / line) and 14 continuous variables to be discretized as shown in Table 2. Before starting to discretization, the dataset pre-processed for the missing values and outliers, and data size was reduced from 4320 to 3493 after this preprocessing. The number of instances in three classes of the class variable was 1146 for Line A, 1187 for Line D and 1146 for Line N after deletion of the outliers.

Unsupervised Discretization Methods and Rules for Interval Numbers Calculation

Although some more sophisticated unsupervised methods such as the novel method using kernel density estimation by Biba et al. (2007) have been proposed in recent years, equal width intervals, equal frequency intervals and clustering are the common unsupervised methods in data mining.

The Equal Width Intervals (EWI) discretization is one of the most common and simplest methods in discretization of continuous features to discrete ones. EWI is based on to split the range ($R=max - min$) of continuous data sorted in ascending order into k equal width intervals with $k-1$ cut-points ($c_1, c_2, \ldots, c_{(k-1)}$) as seen in Equation 1.

$$c_1 = \min + i \cdot h, \quad i = 1, \ldots, k-1 \tag{1}$$

As shown in Equation 2, the width of intervals (h) is simply computed by dividing the range into the number intervals.

$$h = R/k \tag{2}$$

The problem with EWI is that some intervals may be empty or contain more observations than the others if a variable has outliers or extreme values. The Equal Frequency Intervals (EFI) discretization is an unsupervised method overcoming the problem of EWI for the outliers. For this purpose, EFI allocates the equal number of instances of a sorted variable into k intervals or bins. In this way the frequencies of intervals become the same with (n/k) equal frequencies. The disadvantage of this method is that two or more adjacent intervals may contain the values of the same magnitude, and the original density function is lost after discretization.

Clustering is an alternative unsupervised approach to discretization of continuous variables. The divisive and agglomerative hierarchical clustering methods and K-means clustering as a partition clustering method can be also benefitted in discretization of continuous variables. The K-means is one of the most widely used clustering methods that partitions the sample instances into k clusters in where the within cluster variance is as small as possible and between cluster variance is as large as possible.

The above unsupervised discretization methods require a right number of intervals parameter (k) as a user-supplied input. Several rules as listed in Table 1 have been proposed to calculate it. However, recommending an appropriate k value is not an easy task since a certain amount of information may be lost with small k values. On the other hand interpreting the results may be very difficult with big k values.

Table 1 The rules/methods for estimation of number of intervals

Rules	Name of rules	Formula for calculation of k	Author
R1	Square root	$\lfloor n^{1/2} \rfloor$	Davies and Goldsmith (1980)
R2	Sturge	$\lceil 1 + log_2 n \rceil \cong$	Sturges (1926)
	Huntsberger	$\lceil 1 + 3.3\ log_{10} n \rceil$	Doran and Hodson (1975)
R3	Brooks-Carruthers	$\lceil 5\ log_{10} n \rceil$	Brooks and Carruthers (1953)
R4	Cencov	$\lceil n^{1/3} \rceil$	Cencov (1962)
R5	Rice	$\lceil 2\ n^{1/3} \rceil$	Lane et al. (2016)
R6	Terrell-Scott	$\lceil (2n)^{1/3} \rceil$	Terrell and Scott (1985)
R7	Scott	$\lceil R\ /\ 3.5\ \hat{\sigma}\ n^{-1/3} \rceil$	Scott (1979)
R8	Freedman-Diaconis	$\lceil R\ /\ IQR\ n^{-1/3} \rceil$	Freedman and Diaconis (1981)
R9	Doane	$1 + log_2 n + log_2 \left(1 + \frac{\lvert g_1 \rvert}{\sigma_{g_1}}\right)$; $\sigma_{g_1} = \left(\frac{6(n-2)}{(n+1)(n+3)}\right)^{1/2}$	Doane (1976)
R10	K-means clustering	*f(K)* algorithm defined in Pham et al. (2005)	Pham et al. (2005)

Table 2 Descriptive statistics for the continuous variables in the egg quality traits dataset

Variables	Description of variables	Mean	SD	Min	Max	IQR	Skewness
V1	Egg weight (g)	66.16	4.91	47.68	74.72	6.74	0.26
V2	Egg width (mm)	43.19	1.22	42.38	46.67	1.59	0.04
V3	Egg length (mm)	56.93	2.23	50.43	63.52	3.12	0.28
V4	Egg pH	8.46	0.20	7.89	9.04	0.28	-0.13
V5	Shell breaking strength	4.68	1.05	1.70	7.65	1.44	-0.07
V6	Shell thickness (µm)	366.40	22.64	303.33	429.43	31.33	0.02
V7	Shell weight (g)	6.80	0.64	4.99	8.66	0.90	0.19
V8	Yolk weight (g)	16.08	1.90	11.03	21.23	2.49	0.01
V9	Yolk height (mm)	18.36	1.07	15.43	21.26	1.44	-0.16
V10	Yolk width (mm)	39.92	2.60	32.55	47.41	3.66	-0.11
V11	Yolk color index (E)	81.77	5.36	66.29	97.42	7.54	-0.22
V12	White height (mm)	8.64	1.15	5.32	11.78	1.60	-0.12
V13	White width (mm)	64.85	5.53	50.04	80.18	7.28	0.37
V14	White length (mm)	85.42	7.03	66.77	104.61	9.75	0.26
CL	Genotype of chicken	Class variable has three levels: A, D, N					

According to Hyndman (1995) Sturges rule was the first rule and most statistical packages use it for selecting the number of classes in constructing histograms. Brooks and Carruthers (1953) proposed a rule using log_{10} instead of log_2 giving always larger *k* when compared to Sturges rule. The rule by Huntsberger (1962) gives nearly equal results to Sturges rule. These two rules work well if *n* is less than 200 but problematic with large number of *n*. Scott (1992) argued that Sturges rule leads to generate oversmoothed histograms in case of large number of *n.* In his rule Cencov (1962) used the cube root of *n* simply. This rule was followed by its extension such as Rice, and Terrell & Scott with the formulas shown in Table 1. However the square root of *n* produces larger *k* when compared to the others, it has been suggested in Davies and Goldsmith (1980) because of its simplicity.

As seen in Table 1 the rules mostly include *n* only. On the other hand, the methods or algorithms using the measures about the variation and shape of data distributions could provide more optimal *k* values. For instance, Doane (1976) extended the Sturges formula by adding standardized skewness in order to overcome the problem with non-normal distributions need more bins. Scott (1979) used the standard deviation in order to estimate optimal *k* values. Freedman and Diaconis (1981) proposed to use the interquartile range (IQR) statistic which is less sensitive to outliers than the standard deviation.

Computational Tools and Data Analysis

The variables in the dataset were discretized by using the discretize function of the arules library (Hashler et al., 2016) in the R statistical computing environment (R Core Team, 2016). In order to determine the interval numbers via clustering we used the kselection package (Rodriguez, 2016). In order to evaluate the discretization performances of the methods, we compared the classification training error rates and test accuracies calculated with the C5.0 Decision Tree Algorithm. The C5.0 function of C50 library was run on each discretized dataset obtained with the discretization processes. We built the classification tree model by using all the variables (*V1* to *V14* in Table 1) as the predictors (X) and the genotype of chickens (*CL*) as the class variable (Y), and ran the model with 10 iterations with boosting option. We randomly sampled 80% of the data points (n=2750) as the training dataset (trainY and trainX) and the remaining 20% (n=743) as the test dataset (testY and testX). The applied model was C5.0 (trainY ~., data = trainX, trials = 10).

Results and Discussion

As seen in Table 3, the first 6 rules produced the same results for all the variables in the dataset because they do not take into account the variable specific statistics for calculating *k*. The highest *k* was obtained as 59 from the square root rule (R1) while the smallest was 13 from Sturges rule and its counterpart Huntsberger rule (R2). Among the cube root based rules, Cencov rule (R4) and Terrell and Scott rule (R6) gave *k* value as 18 and 20 respectively while Rice rule (R5) produced a bigger *k* value of 30.

Scott rule (R7) using the standard deviation for each variable resulted with the *k* values between 24 and 26, and it was mostly equal to 25. Freedman-Diaconis rule (R8) gave the *k* as 31 for the majority of variables, and changed between 29 and 33. The results from this rule were nearly equal to the results from Rice rule (R5). The results of Doane rule (R9) changed between 13 and 16, and were closer to the results from R2 and R6. The *k* obtained with K-means clustering (R10) was the smallest for the variable yolk width, and highest for the variable yolk color index.

The training error rate of the classification tree model computed on the original dataset containing the continuous values of the variables was found 5.0%. For all the discretized datasets, the training error rates of the model varied between 0.0% and 0.8%. They were smaller than those obtained on the original dataset. This finding revealed that the model worked better on the discretized datasets. According to paired t-test analysis, there was no significant difference between the error rates obtained from EWI and EFI discretized data sets for all the rules ($t=1.8932$, $P>0.05$).

Table 3 Number of intervals (*k*) by the rules

Rules	Variables													
	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14
R1	59	59	59	59	59	59	59	59	59	59	59	59	59	59
R2	13	13	13	13	13	13	13	13	13	13	13	13	13	13
R3	18	18	18	18	18	18	18	18	18	18	18	18	18	18
R4	16	16	16	16	16	16	16	16	16	16	16	16	16	16
R5	30	30	30	30	30	30	30	30	30	30	30	30	30	30
R6	20	20	20	20	20	20	20	20	20	20	20	20	20	20
R7	24	25	26	25	26	25	25	25	24	24	25	25	24	24
R8	30	33	32	31	31	31	31	31	31	31	31	31	31	29
R9	16	14	16	15	15	14	13	15	13	15	15	15	16	16
R10	32	32	32	20	26	33	29	37	37	23	38	36	34	34

Table 4 Model training errors and test accuracies by the discretization methods

Datasets	Training Error (%)	Test Accuracy (%)
Continuous	5.0	53.24
Equal Width Intervals (EWI)		
EWI-R1	0.2	51.32
EWI-R2	0.1	51.44
EWI-R3	0.1	49.34
EWI-R4	0.1	51.54
EWI-R5	0.0	54.43
EWI-R6	0.1	52.10
EWI-R7	0.3	51.55
EWI-R8	0.1	52.32
EWI-R9	0.8	53.10
EWI-R10	0.3	50.55
Equal Frequency Intervals (EFI)		
EFI-R1	0.0	51.88
EFI-R2	0.1	52.32
EFI-R3	0.0	52.10
EFI-R4	0.0	52.21
EFI-R5	0.3	52.21
EFI-R6	0.0	50.33
EFI-R7	0.1	52.88
EFI-R8	0.0	53.20
EFI-R9	0.1	53.20
EFI-R10	0.0	50.33
K-Means Clustering (KMC)		
KMC	0.2	52.54

The test accuracy of the classification model computed from the original undiscretized dataset was found as 53.24%. As seen from Table 4, the test accuracies of the model changed between 49.34% and 54.43%, and were nearly equal for all the discretization methods. The highest accuracy was computed as 54.43% from the equal width intervals method using *k* value of Rice rule (EWI-R5). This was followed by EFI with Freedman-Diaconis rule (EFI-R8) and Doane rule (EFI-R9), and EWI with Doane rule (EWI-R9) again with the test accuracies of 53.20%, 53.20% and 53.10% respectively. The smallest accuracy was found as 49.24% from EWI with Brooks & Carruthers rule (EWI-R3). The remaining rules with EFI and EWI, and K-means clustering performed more or less similar. According to the pairwise t-test there was no significant difference between EWI and EFI methods ($t = -0.6475$; $P > 0.05$).

Conclusions

In this study we empirically compared the performances of EWI, EFI and K-means clustering methods to discretize the 14 continuous features in a chicken egg quality traits dataset. We revealed that discretization can slightly decrease the training error rates and increase the accuracies of classification tree models. By comparing the training errors and test accuracies of the model applied with C5.0 classification tree algorithm we also found that there were no significant differences between the EWI, EFI and K-means clustering methods. According to the findings Rice rule gave the best result with EWI but not with EFI. Following this Freedman-Diaconis rule with EFI and Doane rule with EFI and EWI slightly performed better than the other rules.

According to the results obtained in this study, we propose to use any of unsupervised methods with Freedman-Diaconis rule and Doane rule in discretization of the continuous variables in the chicken egg quality traits datasets. We nevertheless need to make these findings more conclusive by using the more sophisticated unsupervised methods on the other datasets. The supervised methods were reported to be better than the unsupervised ones in some literature (Dougherty et al., 1995) while the contradicting results were obtained by some others (Cantú-Paz, 2001). Therefore we need further empirical comparisons of the unsupervised methods versus some of the common supervised methods. 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.

Acknowledgement

We gratefully thank to Assoc. Prof. Dr. Mikail Baylan and his colleagues at the Cukurova University for their permission to use the dataset analyzed in this study.

References

Biba M, Esposito F, Ferilli S, Mauro ND, Basile TMA. 2007. Unsupervised. Discretization using kernel density estimation. Proc. of the 20th Int. Conf. on AI, Hyderabad, India, p. 696–701.

Brooks CEP, Carruthers N. 1953. Handbook of statistical methods in meteorology. H M Stationery Office, London.

Cantú-Paz E. 2001. Supervised and unsupervised discretization methods for evolutionary algorithms. In proc. Of the genetic and evolutionary computation conference (GECCO-2001), p. 213-216.

Cebeci Z, Yıldız F, Kayaalp GT. 2015. K-ortalamalar kümelemesinde optimum K değeri seçilmesi. 2. Ulusal Yönetim Bilişim Sistemleri Kongresi. Erzurum, 8-10 Ekim 2015. Bildiriler Kitabı (Ed: Ü. Özen ve ark.), p. 231-242.

Cencov NN. 1962. Evaluation of an unknown distribution density from observations. Soviet Mathematics, 3: 1559–1562.

Doane DP. 1976. Aesthetic frequency classification. American Statistician, 30 (4): 81-183.

Dash R, Paramguru RL, Lochan RR, Dash. 2011. Comparative analysis of supervised and unsupervised discretization techniques. Int. J. of Advances in Science and Technology, 2(3): 29-37.

Davies OL, Goldsmith PL .1980. Statistical methods in research and production. 4th edn longman London, p. 478.

Doran JE, Hodson FR. 1975. Mathematics and computers in archaeology. Massachusetts: Harvard Univ. Press Cambridge, p. 381.

Dougherty J, Kohavi R, Sahami M. 1995. Supervised and unsupervised discretization of continuous feature. In proc. Of the 12th Int. Conf. on Machine Learning, p. 194-202.

Freedman D, Diaconis P. 1981. On the histogram as a density estimator: L_2 theory. Zeit. Wahr. ver. Geb. 57(4): 453–476.

García S, Luengo J, Sáez J A, López V, Herrera F. 2013. A survey of discretization techniques, taxonomy and empirical analysis in supervised learning. IEEE Transactions on Knowledge and Data Engineering, 25(4): 734-750.

Hahsler M, Buchta C, Gruen B, Hornik K. 2016. Arules: mining association rules and frequent itemsets. R package version 1.4-1. https://CRAN.R-project.org/package=arules [Accessed on 26.07.2016].

Hemada B, Lakshmi KSV. 2013. A study on discretization techniques. Int. J. of Engineering Research & Technology, 2(8): 1887-1892.

Huntsberger DV. 1962. Elements of statistical inference. London: prentice-hall.

Hyndman RJ. 1995. The problem with Sturges' rule for constructing histograms.
URL: http://robjhyndman.com/papers/sturges.pdf [Accessed on 26.07.2016]

Kotsiantis S, Kanellopoulos D. 2006. Discretization techniques: a recent survey. GESTS International Transactions on Computer Science and Engineering, 32 (1): 47-58.

Kuhn M, Weston S, Coulter N, Clup M. 2016. C50: C5.0 decision trees and rule-based models. R package version 0.1.0-24 (C code for C5.0 by R. Quinlan License: GPL-3) (https://cran.r-project.org/web/packages/C50/) [Accessed on 26.07.2016]

Lane DM, Scott D, Hebl M, Guerra R, Osherson D, Zimmer H. 2016. Introduction tostatistics: a multimedia course of study (http://onlinestatbook.com/) [Accessed on 26.07.2016]

Liu H, Hussain F, Tan C L, Dash M. 2002. Discretization: an enabling technique. Data Mining and Knowledge Discovery, 6(4): 393-423.

Muhlenbach F, Rakotomalala R. 2005. Discretization of continuous attributes. In Encyclopedia of Data Warehousing and Mining (Ed. J. Wang), p: 397–402.

Pham DT, Dimov SS, Nguyen CD. 2005. Selection of K in K-means clustering. Journal of Mechanical Engineering Science, 219: 103 -119.

R Development Core Team 2016. A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/

Ramírez-Gallego S, García S, Mouriño-Talín H, Martínez-Rego D, Bolón-Canedo V, Alonso-Betanzos A, Benítez JM, Herrera F. 2015. Data discretization: taxonomy and big data challenge. WIREs Data Mining Knowledge Discovery, 6(1): 5- 21.

Rodriguez G. 2016. Kselection: selection of K in K-means clustering. R package version 0.2.0. http://CRAN.R-project.org/package=kselection [Accessed on 26.07.2016].

Scott DW. 1979. On optimal and data-based histograms. Biometrika, 66(3): 605–610.

Scott DW. 1992. Multivariate density estimation: theory, practice and visualization. New York: John Wiley & Sons.

Sturges H (1926). The choice of a class-interval. J. Amer. Statist. Assoc. 21(153): 65– 66.

Terrell GR, Scott DW. 1985. Oversmoothed Nonparametric Density Estimates. Journal of the American Statistical Association, 80(389): 209–214.