Big Data and Information Analytics
©American Institute of Mathematical Sciences
Volume **1**, Number **2&3**, April & July **2016**
doi:10.3934/bdia.2016005
pp. **217–225**

FORWARD SUPERVISED DISCRETIZATION FOR MULTIVARIATE WITH CATEGORICAL RESPONSES

Wenxue Huang and Qitian Qiu

School of Mathematics and Information Science, Guangzhou University
Guangzhou, Guangdong 510006, China

(Communicated by Jianhong Wu)

Abstract. Given a data set with one categorical response variable and multiple categorical or continuous explanatory variables, it is required in some applications to discretize the continuous explanatory ones. A proper supervised discretization usually achieves a better result than the unsupervised ones. Rather than individually doing so as recently proposed by Huang, Pan and Wu in [12, 13], we suggest a forward supervised discretization algorithm to capture a higher association from the multiple explanatory variables to the response variable. Experiments with the GK-tau and the GK-lambda are presented to support the statement.

1. **Introduction.** In data analysis and mining, discretization algorithms are to turn continuous variables into certain levels. For instance, income, as a continuous variable, needs to be leveled as low, median or high for executable user profiling or for descriptive analysis. Many data mining technologies or methods also require categorical explanatory variables. For example, Bayesian classification[21] assumes that the explanatory variables are all categorical or discrete; decision trees[22] consists of tree node describing the condition that leads to the next node. The variables involved in each node have to be either categorical or described as a combination of intervals. One may see many continuous variables in many real data sets such as asset, income, debt, age, numerical measure of risk, etc.

A natural method of grouping distinct values in a continuous variable is to seek data-based cutting points that cut the whole range of data into intervals. There are two ways to identify the intervals: with or without a given response variable. An algorithm of discretizing continuous variables with a response variable (with a given criterion or objective function) is called supervised discretization; while the other (with no link to the response variable) is called unsupervised discretization [5]. The latter one usually is used in industrial data mining project to deal with multiple response variables at the same time for convenience, unification or for the purpose of saving production cost. One may use a normal distribution based unsupervised discretization to process the macro numerical consuming data because of the central limit theorem. Other unsupervised discretization methods include, but not limit to equal frequency intervals, equal width intervals[5] or more sophisticated ones associated with certain criterion measures (or objective functions), say, the information

2010 *Mathematics Subject Classification.* Primary: 62H20; Secondary: 62F07, 68T30.

Key words and phrases. Categorical data, the $GK-\lambda$, the $GK-\tau$, forward supervise discretization, independent supervised discretization.

theoretic entropy[4, 6]. The idea is to minimize or maximize the measures in each interval by adjusting the boundaries. Various clustering algorithms can also be used to accomplish an unsupervised discretization[7] to multi-dimensional cases, which compartmentalize a continuous multidimensional space into a given finite number of parts.

In contrast to the unsupervised one, supervised discretization algorithms aim to seek the boundaries by optimizing the intervals' coherence[16] associated with a target variable with an optimization criterion. In other words, an evaluation function is usually applied to measure the discretization algorithm's quality. Typical measures include conditional entropy, conditional Gini concentration or Chi-square. The Chi-square based methods include ChiMerge [15], Chi2 [17], Khiops [1], etc. One can refer to [2, 3, 10, 20] for the detailed discussion regarding the (conditional) entropy based methods. The conditional Gini concentration based methods can be found in [12, 13]. The choice of discretization method depends on the time computing complexity, the greediness for the accuracy, the interpretability [16] and/or how easy the result can be executed. Usually the unsupervised discretization methods are faster than the supervised ones but result in less accuracy in predicting the target variable. Dougherty et al.[5] prove that (conditional) entropy-based discretization methods perform quite well overall regarding the proportional association. Holte showed in [10] that even a one-dimensional supervised discretization system could yield similar classification results to a multiple dimensional ones if the system is carefully tuned.

Rather than using (conditional) entropy-based discretization method, Huang et al. propose two alternates, proportional and modal, association based independently (or individually) supervised discretization in [12, 13]. Beside the expected higher association with the responses and with limited higher computational complexity than an unsupervised one, they also argue that their measures are more interpretable than the conditional entropy ones.

Suppose we work with a categorical response variable and a number of explanatory continuous variables. We propose a forward supervised discretization algorithm to capture a better association with the response variable compared to the independently supervised discretization algorithm proposed by Huang et al. Experiments in the latter part of this article show remarkable improvements, with acceptable increase of computationally complexity, regarding the same association proposed in [12] and [13].

This article is organized as follows. Section 2 reassembles the concept of categorical variable association measures, especially, the GK-tau and GK-lambda. We also briefly recall Huang et al.'s independently supervised discretization algorithm in this section. The forward supervised discretization algorithm is presented in Section 3. Although the algorithms are introduced by GK-tau, it can be replaced by any other association measure including GK-lambda.

Section 4 show the supportive evidences by experiments with the GK-tau and GK-lambda. The major issues are summarized and commented in the last section.

2. **Categorical variable association measures and independently supervised discretization.**

Let X and Y be categorical variables with domains $\mathrm{Dmn}(X) = \{1, 2, \ldots, n_X\}$ and $\mathrm{Dmn}(Y) = \{1, 2, \ldots, n_Y\}$, respectively, where X is a explanatory variable and Y is a response (target) variable.

According to [18, p.70], an association measure (or degree) of Y on X is, in general, of the form

$$\delta^{Y|X} = \frac{V(Y) - E_X\{E_Y V(Y|X)]\}}{V(Y)},$$

where V is a given variance and E is the corresponding expectation. The choosing of a variance measure depends on the objective of the data analysis, the predictive model to be used, or even on the analyst's preference: the Gini, entropy and Chi-square are typical preferences. In this article, we choose the proportional and the modal prediction oriented variances, the GK-tau and GK-lambda, as in ([11, 12, 13, 8, 18], for their statistical interpretability.

The proportional association degree of Y on X, denoted by $\tau^{Y|X}$, referred also to as the GK-tau, is given by

$$\tau^{Y|X} = \frac{\sum_{j=1}^{n_Y}\sum_{i=1}^{n_X}\frac{p(X=i,Y=j)^2}{p(X=i)} - \sum_{j=1}^{n_Y} p(Y=j)^2}{1 - \sum_{j=1}^{n_Y} p(Y=j)^2},$$

where $p(\cdot)$ is the probability of an event. To help the reader better understand $\tau^{Y|X}$, we recall

(1) $\omega^{Y|X} = \sum_{j=1}^{n_Y}\sum_{i=1}^{n_X}\frac{p(X=i,Y=j)^2}{p(X=i)} = \sum_{j=1}^{n_Y}\sum_{i=1}^{n_X} p(X=i,Y=j)p(Y=j|X=i)$,

(2) $E(p(Y)) = \sum_{j=1}^{n_Y} p(Y=j)^2$, and

(3) $\text{Gini}(Y) = 1 - E(p(Y))$,

where $\omega^{Y|X}$ is the expected accuracy rate for predicting Y given X; $E(p(Y))$ is the expected accuracy rate for predicting Y by its own distribution; $\tau^{Y|X}$ is the error reduction rate of predicting Y given X over predicting Y on its own. Please refer to[12] and[8, 18] for more detailed discussions.

The optimal (or modal) association degree of Y on X, denoted by $\lambda^{Y|X}$ is given by

$$\lambda^{Y|X} = \frac{\sum_{j=1}^{n_X}\rho_{jm} - \rho_{.m}}{1 - \rho_{.m}}.$$

where $\rho(.)$ is probability of an event, and

$$\rho_{jm} = \max_{j\in\{1,2,\ldots,n_Y\}}\rho(X=i;Y=j),$$

where $\rho_{.m} = \max_{j\in\{1,2,\ldots,n_Y\}}\rho(Y=j)$ and $\rho_{.m}$ is the theoretical prediction accuracy rate for modally predicting target variable Y without information of variable X; ρ_{jm} is the accuracy rate for modally predicting target variable Y based on the information of variable X. Obviously, $\lambda^{Y|X}$ is the error reduction rate using information of X over using only the marginal information of Y by mode. [8, p.71] has detailed discussion in $\lambda^{Y|X}$.

Recall [12] again that for a continuous explanatory variable X and categorical nominal variable Y with n_Y different values from a given data set, assume $B_k = \{b_1, b_2, \ldots, b_k\}$ is a set of different values, where $b_1 < b_2 < \ldots < b_k$, and B_k can cut continue variable X into $k+1$ intervals: $(-\infty, b_1], (b_1, b_2], \ldots, (b_k, \infty)$. The

association measure $\tau^{Y|X}$ for a given cutting point set B_k can be defined as:

$$\tau^{Y|X(B_k)} = \frac{\sum_{j=1}^{n_Y}\sum_{i=1}^{k+1}\frac{p(b_{i-1}<X\le b_i, Y=j)^2}{p(b_{i-1}<X\le b_i)} - \sum_{j=1}^{n_Y} p(Y=j)^2}{1-\sum_{j=1}^{n_Y} p(Y=j)^2},$$

where $b_0 = -\infty$ and $b_{k+1} = \infty$.

Huang et al proposed an optimal splitting searching scheme in [12] to find these cutting points. This scheme can be briefly described as follows:

Step1. The first cutting point, denoted $r_1^*(X)$, can be found by
$r_1^*(X) = \text{argmax}_{\min(X)<r<\max(X)}\, \tau^{Y|((-\infty,r],(r,\infty))}$;

Step2. The second cutting point, denoted $r_2^*(X)$, can be found by

$$r_2^*(X) = \underset{\min(X)<r<\max(X)\setminus r_1^*(X)}{\text{argmax}} \tau^{Y|((-\infty,\min(r_1^*,r)],(\min(r_1^*,r),\max(r_1^*,r)],(\max(r,r_1^*),\infty))}$$

Step3. Continue the steps until a predefined number, say m , of intervals are found.

Please note that the number of values for a given variable in a given data set is always finite even this variable is defined as continuous. Thus not only the aforementioned $\min(X)$, $\max(X)$, but also the number of search steps are all finite. Besides, one can also discretize the continuous variable in a unsupervised manner to speed up the search.

Please also note that the previous searching schema can also applies to the case of $Gk - \lambda$ where the association is as follows.

$$\lambda(Y|X(B_K)) = \frac{\sum_{i=1}^{k+1}\max_{j\in\{1,2,\ldots,n_Y\}} p(b_{i-1}<X\le b_i, Y=j) - \max_{j\in\{1,2,\ldots,n_Y\}}\{p(Y=j)\}}{1-\max_{j\in\{1,2,\ldots,n_Y\}}\{p(Y=j)\}},$$

where $b_0 = -\infty$ and $b_{k+1} = \infty$.

3. **Forward supervised discretization.** For a data set with n continuous independent variables, $X_1, X_2, \ldots, X_n$, we first discretize them by independently supervised discretization approach introduced in the previous section. The discretized independent variables are denoted as $\text{id}X_i$, $i = 1, 2, \ldots, n$. We then identify the leading one , denoted as $\text{id}X_{i_0}$, that brings in the maximum association to the target, i.e., $\text{id}X_{i_0} = \text{argmax}_{i=1,2,\ldots,n}\, \tau^{Y|\text{id}X_i}$.

We assume without loss of generality that $i_0 = 1$. Then $\text{id}X_1$ is the base (categorical) variable for our forward supervised discretization, which means the subsequent discretization for any other X_i, $i \neq 1$ is based on $\text{id}X_1$. The searching for cutting points in X_i is similar to that in Section 2 as follows.

Step1. The first cutting point, denoted $\text{fd}^1 r_1^*(X_i)$, is determined by

$$\text{fd}^1 r_1^*(X_i) = \underset{\min(X_i)<r<\max(X_i)}{\text{argmax}} \tau^{Y|(\text{id}X_1,((-\infty,r],(r,\infty)))};$$

Step2. the second cutting point, denoted $\text{fd}^1 r_2^*(X_i)$, is determined by

$$\text{fd}^1 r_2^*(X_i) = \underset{\min(X_i)<r<\max(X_i)\setminus \text{fd}^1 r_1^*(X_i)}{\text{argmax}} \tau^{Y|(\text{id}X_1,\text{fd}^1 r_2(X_i))},$$

where

$$\text{fd}^1 r_2(X_i) = \{(-\infty, \min(\text{fd}^1 r_1^*, r)], (\min(\text{fd}^1 r_1^*, r), \max(\text{fd}^1 r_1^*, r)], (\max(\text{fd}^1 r_1^*, r), \infty);$$

Step3. continue the process in this fashion until a predefined number, say m , of intervals are found.

When all variables are discretized by this procedure, denoted as $\mathrm{fd}^1(X_i)$, $i = 2, \ldots, n$, we have two results for each X_i: one from independently supervised discretization and one from forward supervised discretization. One might ask which one brings in higher association when working together with $\mathrm{id}X_1$. We will show by experiments in the next section that the latter wins.

We further assuming without loss of generality that $\mathrm{fd}^1(X_2) = \mathrm{argmax}_{i=1,2,\ldots,n} \tau^{Y|(\mathrm{id}X_1, \mathrm{fd}^1(X_i)}$. We can then continue to discretize other variables following the similar pattern above by $\mathrm{id}X_1$ and $\mathrm{fd}^1(X_2)$.

Step1. The first cutting point, denoted $\mathrm{fd}^2 r_1^*(X_i)$, is determined by

$$\mathrm{fd}^2 r_1^*(X_i) = \underset{\min(X_i)<r<\max(X_i)}{\mathrm{argmax}} \tau^{Y|((\mathrm{id}X_1, \mathrm{fd}^1(X_2)),((-\infty,r],(r,\infty))};$$

Step2. The second cutting point, denoted $\mathrm{fd}^2 r_2^*(X_i)$, is determined by

$$\mathrm{fd}^2 r_2^*(X_i) = \underset{\min(X_i)<r<\max(X_i)\setminus \mathrm{fd}^2 r_1^*(X_i)}{\mathrm{argmax}} \tau^{Y|((\mathrm{id}X_1, \mathrm{fd}^1(X_2)), \mathrm{fd}^2 r_2(X_i))},$$

where

$$\mathrm{fd}^2 r_2(X_i) = \{(-\infty, \min(\mathrm{fd}^2 r_1^*, r)], (\min(\mathrm{fd}^2 r_1^*, r), \max(\mathrm{fd}^2 r_1^*, r)], (\max(\mathrm{fd}^2 r_1^*, r), \infty).$$

Step3. Continue the process until a predefined number, say m , of intervals are found.

We admit that the forward discretizing scheme in this article is a variation of stepwise feature selection procedure[9]. When to stop the discretizing loop depends on the condition to stop searching the next variable. The conditions include, but not limited to, reaching the maximum joint association, or reaching the predefined maximum number of variables.

Naturally, the computational expense for the forward supervised discretization is significantly higher than the individual ones. But the difference is still acceptable since our forward one is based on individual one; and the individual one finally chooses very few predictors. Given that this article is to recommend an alternative discretization procedure that can increase the association from the independent variables to the target, we believe it is worth the cost.

4. **Empirical experiment and discussion.**

Experiments. The purpose of this article's experiments is to show how the forward supervised discretization method improves the variable association in the multivariate case for both the GK-tau and the GK-lambda. Not only the associations but also the independent variables' domain size are evaluated under different circumstances to demonstrate the approximate reliability in statistical sense for each chosen variable set. In general, a variable set with a smaller domain size has higher confidence power. When two variable sets have the same association, the one with smaller domain size is preferred in most feature selection methods.

The data set in our experiment is The Survey of Family Expenditure conducted by Statistic Canada in 1996 (Famex96)[23]. It has $10,417$ rows with over 200 continuous and categorical variables. We use four of them as the continuous independent variables. They are Income Before Taxes(Inc-btax), Total Expenditure(Tot-expn),

Income before Taxes(Hh-incbt), Paper, Plastic And Foil Household Supplies(Hh-suply) and Age(age). Two target variables are manually picked from the data set to exemplify our statement. The first target variable is Class Of Tenure (Tenure) with four classes(1-4). The second target variable is the educational level(Edn) with six cases(1-6). Both variables are ordinal in nature, but we treat them as nominal for the purpose of simplicity. The maximal number of intervals for each continuous independent variable is set as $m = 3$ for the purpose of simplicity again.

Since the approach in this article is inspired and based by the independently supervised discretization algorithm, we are going to compare them in different scenarios in both experiments. Please note that ω, rather than τ, is chosen in the first experiment to explain the reason to select various models because not only because they are mathematically equivalent, but also because we believe ω has better interpretability. The same reason goes with the selection of ϕ over λ in the second experiment.

4.1. $GK - \tau$ **case by** *Tenure*. After the independent supervised discretization by $GK - \tau$, we get the independent variables' associations as in Table 1.

TABLE 1. ω and τ: the first round of discretizations

variable	Boundary1	Boundary2	$\omega^{Y\|X}$	$\tau^{Y\|X}$	$\mathrm{Dmn}(Y, X)$
Inc-btax	29875	44813	0.3456	0.0443	12
Tot-expn	21274	41948	0.3802	0.0946	12
Hh-incbt	19655	35337	0.3848	0.0694	12
Hh-suply	185	371	0.3376	0.0324	12
Age	34	54	0.3896	0.1084	12

Thus the best variable after the independent supervised discretization is Age with $\omega = 0.3896$. Denote each independently discretized variable as idInc-btax, idTot-expn, idHh-incbt, idHh-suply and idAge respectively, ωs for the 2-variable groups, (idAge, idInc-btax), (idAge, idTot-expn), (idAge, idHh-incbt) and (idAge, idHh-suply) are 0.4134, 0.4523, 0.4527, 0.4082 respectively. The best group then is (idAge, idHh-incbt) with $\omega = 0.4527$. Going on with the same process gives us the best 3-variable group as (idAge, idHh-incbt, idTotexpn) with $\omega = 0.4661$. Correspondingly, the domain size for the best 2-variable groups and 3-variable groups are 36 and 108.

The difference between the independently supervised discretization and the forward supervised discretization begins at the 2-variable groups. Using idAge as the first one for the subsequent forward supervised discretization process since it has the highest ω, we can discretize the other explanatory variables by the procedure introduced in the previous section. The discretized variables are denoted as $\mathrm{fd}^1 X_i$ where X_i are one of Inc-btax, Tot-expn, Hh-incbt and Hh-suply. Table 2 shows the detailed discretization result.

TABLE 2. the ω and τ: the second round of discretization

X	Bndry1	Bndry2	$\omega^{Y\|\mathrm{id}X_1,\mathrm{fd}^1X}$	$\tau^{Y\|\mathrm{id}X_1,\mathrm{fd}^1X}$	$\|(Y, \mathrm{id}X_1, \mathrm{fd}^1X)\|$
fd^1Inc-btax	14938	29875	0.4135	0.1433	35
fd^1Tot-expn	31611	52285	0.4506	0.1975	35
fd^1Hh-incbt	19887	35649	0.4587	0.2093	33
fd^1Hh-suply	93	185	0.4096	0.1377	35

Table 2 shows the best 2-variable group as (idAge, fd^1Hh-incbt) with $\omega = 0.4587$ and $\mathrm{Dmn}(Y, \mathrm{id}X_1, \mathrm{fd}^1X_2) = 35$.Since $\omega(Y|(\mathrm{id}X_1, \mathrm{fd}^1X_2)) \geq \omega(Y|(\mathrm{id}X_1, \mathrm{id}^1X_2))$ and

$\mathrm{Dmn}(Y, \mathrm{id}X_1, \mathrm{fd}^1X_2) \leq Domain(Y, \mathrm{id}X_1, \mathrm{id}^1X_2)$, the forward procedure is better than the individual one by both indicators.

Keep going with the same process, we have the discretization result for the third variable, denoted by fd^2X as Table 3

TABLE 3. ω and τ: the third round of discretization

X	Bndry1	Bndry2	$\omega^{Y\|\mathrm{id}X_1,\mathrm{fd}^1X_2,\mathrm{fd}^2X}$	$\tau^{Y\|\mathrm{id}X_1,\mathrm{fd}^1X_2,\mathrm{fd}^2X}$	$\|(Y, \mathrm{id}X_1, \mathrm{fd}^1X, \mathrm{fd}^2X)\|$
$\mathrm{fd}^2Inc-btax$	21249	42499	0.4818	0.2430	72
$\mathrm{fd}^2Tot-expn$	15305	44715	0.4776	0.2369	89
$\mathrm{fd}^2Hh-suply$	132	268	0.4721	0.2288	107

The best 3-variable group then is (Age,Hh-incbt,Inc-btax) with $\omega = 0.4818$ and its sample size is 72. Given that $\omega(Y|\mathrm{id}X_1, \mathrm{fd}^1X_2, \mathrm{fd}^2X_3) \geq \omega(Y|\mathrm{id}(X_1, X_2, X_3))$ and $\mathrm{Dmn}(Y, \mathrm{id}X_1, \mathrm{fd}^1X_2, \mathrm{fd}^2X_3) \leq \mathrm{Dmn}(Y, \mathrm{id}(X_1, X_2, X_3))$, the forward approach is still better than the individual one at the 3-variable level.

4.2. $GK-\lambda$ **case by** Edn**.** After individually discretized by $GK-\lambda$ with respect to Edn, we have the following initial result as follow Table 4.

TABLE 4. The ϕ and λ:: the first round of discretization

X	Bndry1	Bndry2	$\phi^{Y\|X}$	$\lambda^{Y\|X}$	$\|(Y, X)\|$
$Inc-Btax$	44813	59750	0.4331	0.0667	18
Tot-Expn	83297	93634	0.4213	0.0473	16
$Hh-Incbt$	82389	98073	0.4228	0.0497	17
Hh-Supply	2039	2966	0.4031	0.0272	17

The best variable regarding λ is then idInc-Btax. Calculation also shows that the next 2-variable groups (idInc-Btax, idTot-Expn), (idInc-Btax, idHh-Incbt), (idInc-Btax, idHh-Suply) have a list of ϕ as, respectively, 0.4346,0.4339,0.4335. It gives us the best 2-variable group as (idInc-Btax ,idTot-Expn). The remaining 3-variable groups, (idInc-Btax,idTot-Expn, idHh-Incbt) and (idInc-Btax,idTot-Expn,idHh-Suply) show ϕ as 0.4369 and 0.4347 respectively which gives us the better one as (idInc-Btax,idTot-Expn,idHh-Incbt). Let Inc-Btax,Tot-Expn,Hh-Incbt be denoted as X_1, X_2 and X_3, respectively. We also find out that $\omega(Y|\mathrm{id}X_1, \mathrm{id}X_2) = 0.4346$, $\mathrm{Dmn}(Y, \mathrm{id}X_1, \mathrm{id}X_2) = 46$; $\omega(Y|\mathrm{id}X_1, \mathrm{id}X_2, \mathrm{id}X_3) = 0.4369$, and Dmn $(Y, \mathrm{id}X_1, \mathrm{id}X_2, \mathrm{id}X_3) = 132$.

Using idInc-Btax as the leading variable in the subsequent forward supervised discretization process, we have the second discretization to other independent variables, denoted as (fd^1Inc-btax, fd^1Tot-expn, fd^1Hh-suply) or fd^1X_i as Table 5:

TABLE 5. The ϕ and λ: the second round of discretization

Variable	Bndry1	Bndry2	$\phi^{Y\|\mathrm{id}X_1,\mathrm{fd}^1X_i}$	$\lambda^{Y\|\mathrm{id}X_1,\mathrm{fd}^1X_i}$	$\|(Y, \mathrm{id}X_1, \mathrm{fd}^1X_i)\|$
fd^1Tot-Expn	113774	144803	0.4353	0.0702	40
fd^1Hh-Incbt	107890	184954	0.4358	0.0712	41
fd^1Hh-Suply	834	927	0.4333	0.0670	43

From Table 5, we see that (Inc-Btax, fd^1Hh-Incbt) has the highest ϕ as 0.4358 with domain size of 41. Hence

$$\phi^{Y|(\mathrm{id}X_1,\mathrm{fd}^1X_2)} \geq \phi^{Y|(\mathrm{id}X_1,\mathrm{id}X_2)}$$

with the additional advantage that

$$|\mathrm{Dmn}(Y, \mathrm{id}X_1, \mathrm{fd}^1X_2)| < |\mathrm{Dmn}(Y, \mathrm{id}X_1, \mathrm{id}X_2)|.$$

Table 6 shows the result of discretizing the third variable based on (idInc-Btax, fd^1Hh-incbt) from the previous step.

TABLE 6. the ϕ and λ result for 2 forwardly supervisedly discretized variables

X	Bndry1	Bndry2	$\phi^{Y\|\mathrm{id}X_1,\mathrm{fd}^1X_2,\mathrm{fd}^2X}$	$\lambda^{Y\|\mathrm{id}X_1,\mathrm{fd}^1X_2,\mathrm{fd}^2X}$	$\|(Y, \mathrm{id}X_1, \mathrm{fd}^2X_1, \mathrm{fd}^2X)\|$
fd^2Tot-Expn	47349	114874	0.438	0.0747	112
fd^2Hh-Supply	233	838	0.4375	0.0739	126

The 3-variable group with the highest accurate rate then goes to (idInc-Btax, fd^1Hh-Incbt, fd^2Tot-Expn). The corresponding

$$\phi^{Y|(\mathrm{id}X_1,\mathrm{fd}^1X_2,\mathrm{fd}^2X_3} = 0.438 \geq \phi^{Y|(\mathrm{id}X_1,\mathrm{id}X_2,\mathrm{id}X_3} = 0.4369$$

while

$$|\mathrm{Dmn}(Y, \mathrm{id}X_1, \mathrm{fd}^1X_2, \mathrm{fd}^2X_3)| = 112 < |\mathrm{Dmn}(Y, \mathrm{id}X_1, \mathrm{id}X_2, \mathrm{id}X_3)| = 132.$$

We may continue if the number of continuous independent variables is greater than 3 and that the sample size is large enough to ensure the reliability of information (or the confidence power of data). Till then both experiments show show improved performance by the forward supervised discretization than the individual one.

5. **Discussion and future work.** In this article, we propose a categorical variable association, e.g., the GK-tau or the GK-lambda, based forward supervised discretization method for multi-dimensional data set. This method is inspired and based on an individually supervised discretization proposed in [12, 13]. We demonstrate the new approach's advantage by two experiments. One is based on GK-tau or and another is based on GK-lambda. The experiments also have different target variables to show our approach's robustness. Admittedly, the new approach take more computational time. But we believe the cost is acceptable given the the improved performance including association.

Although the individual and the forward are applied to one single variable while the latter uses the information from the variable that are previously discretized by the same approach, it is natural to extend the case to compartmentalizing a multi-dimensional space using the same idea. A popular compartmentalization technology is clustering. Another interesting research is to find out the exact computational cost for the new approach.

REFERENCES

[1] M. Boulle, Khiops: A statistical discretization method of continuous attributes, *Machine Learning*, **55** (2004), 53–69.

[2] J. Catlett, On changing continuous attributes into ordered discrete attributes, In: *Machine LearningEWSL-91*, **482** (1991), 164–178.

[3] D. Chiu, B. Cheung and A. Wong, Information synthesis based on hierarchical maximum entropy discretization, *Journal of Experimental and Theoretical Artificial Intelligence*, **2** (1989), 117–129.

[4] M. Chmielewski and J. Grzymala-Busse, Global discretization of continuous attributes as preprocessing for machine learning, *International Journal of Approximate Reasoning*, **15** (1996), 319–331.
[5] J. Dougherty, R. Kohavi and M. Sahami, Supervised and unsupervised discretization of continuous features, In *Machine learning–International Workshop. Morgan Kaufmann Publishers*, **2** (1995), 194–202.
[6] U. Fayyad and K. Irani, Multi-interval discretization of continuous-valued attributes for classification learning, *Proceedings of the International Joint Conference on Uncertainty in AI*, **2** (1993), 1022–1027.
[7] G. Gan, C. Ma and J. Wu, Data clustering: Theory, algorithms, and applications(ASA-SIAM series on statistics and applied probability), *Society for Industrial and Applied Mathematics*, **20** (2007), xxii+466 pp.
[8] L. Goodman and W. Kruskal, Measures of association for cross classifications, *Journal of the American Statistical Association*, **49** (1954), 732–764.
[9] I. Guyon and A. Elisseeff, An Introduction to Variable and Feature Selection, *Applied Physics Letters*, **3** (2002), 1157–1182.
[10] R. Holte, Very sim1ple classification rules perform well on most commonly used datasets, *Machine Learning*, **11** (1993), 63–90.
[11] W. Huang and Y. Pan, On balalncing between optimal and proportional predictions, *Big Data and Information Analytics*, **1** (2016), 129–137.
[12] W. Huang, Y. Pan and J. Wu, Supervised discretization with $GK - \tau$, In *Procedia Computer Science*, **17** (2013), 114–120.
[13] W. Huang, Y. Pan and J. Wu, Supervised discretization with $GK - \lambda$, *Procedia Computer Science*, **30** (2014), 75–80.
[14] W. Huang, Y. Shi and X. Wang, A nomminal association matrix with feature selection for categorical data, *Communications in Statistics – Theory and Methods*, to appear.
[15] R. Kerber, Chimerge: Discretization of numeric attributes, In *Proceedings of the tenth national conference on Artificial intelligence.AAAI Press*, 1994, 123–128.
[16] S. Kotsiantis and D. Kanellopoulos, Discretization techniques: A recent survey, *GESTS International Transactions on Computer Science and Engineering*, **32** (2006), 47–58.
[17] H. Liu and R. Setiono, Chi2: Feature selection and discretization of numeric attributes, In: *Proceedings of the Seventh International Conference on Tools with Artificial Intelligence*, **55** (1995), 388–391.
[18] C. Lloyd, *Statistical Analysis with Missing Data*, John Wiley & Sons, Inc. 1987, New York, NY, USA.
[19] J. MacQueen, Some methods for classification and analysis of multivariate observations, *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, **1** (1967), 281–297.
[20] D. Olson and Y. Shi, Introduction to business data mining, *Knowledge and information systems*, 2007, McGraw-Hill/Irwin.
[21] I. Rish, An empirical study of the naive bayes classifier, *IJCAI 2001 workshop on empirical methods in artificial intelligence*, 2001, 41–46.
[22] S. Safavian and D. Landgrebe, A survey of decision tree classifier methodology, *IEEE Transactions on Systems, Man and Cybernetics*, **21** (1991), 660–674.
[23] STATCAN, Survey of Family Expenditures - 1996.
[24] K. Ting, *Discretization of Continuous-Valued Attributes and Instance-Based Learning*, Basser Department of Computer Science,University of Sydney, 1994.

Received April 2016; revised September 2016.

E-mail address: Wenxue Huang, whuang123@yahoo.com
E-mail address: Qitian Qiu, 18825058017@163.com