Knowl Inf Syst
DOI 10.1007/s10115-008-0142-6

REGULAR PAPER

Data discretization unification

Ruoming Jin · Yuri Breitbart · Chibuike Muoh

Received: 29 October 2007 / Revised: 18 December 2007 / Accepted: 29 January 2008

Abstract Data discretization is defined as a process of converting continuous data attribute values into a finite set of intervals with minimal loss of information. In this paper, we prove that discretization methods based on informational theoretical complexity and the methods based on statistical measures of data dependency are asymptotically equivalent. Furthermore, we define a notion of generalized entropy and prove that discretization methods based on Minimal description length principle, Gini index, AIC, BIC, and Pearson's X^2 and G^2 statistics are all derivable from the generalized entropy function. We design a dynamic programming algorithm that guarantees the best discretization based on the generalized entropy notion. Furthermore, we conducted an extensive performance evaluation of our method for several publicly available data sets. Our results show that our method delivers on the average 31% less classification errors than many previously known discretization methods.

Keywords Discretization · Entropy · Gini index · MDLP · Chi-square test · G^2 test

1 Introduction

Many real-world data mining tasks involve continuous attributes. However, many of the existing data mining systems cannot handle such attributes. Furthermore, even if a data mining task can handle a continuous attribute, its performance can be significantly improved by replacing a continuous attribute with its discretized values. Data discretization is defined as a process of converting continuous data attribute values into a finite set of intervals and

This research in part is supported by Lady Davis Fellowship, Haifa, Israel.

R. Jin (✉) · Y. Breitbart · C. Muoh
Department of Computer Science, Kent State University, Kent, OH 44241, USA
e-mail: jin@cs.kent.edu

Y. Breitbart
e-mail: yuri@cs.kent.edu

C. Muoh
e-mail: cmuoh@cs.kent.edu

associating with each interval some specific data value. There are no restrictions on discrete values associated with a given data interval except that these values must induce some ordering on the discretized attribute domain. Discretization significantly improves the quality of discovered knowledge [8,30] and also reduces the running time of various data mining tasks such as association rule discovery, classification, and prediction. Catlett in [8] reported tenfold performance improvement for domains with a large number of continuous attributes with little or no loss of accuracy.

However, any discretization process generally leads to a loss of information. Thus, the goal of the *good* discretization algorithm is to minimize such information loss.

Discretization of continuous attributes has been extensively studied [5,8–10,12,15,24,25]. There is a wide variety of discretization methods starting with naive (often referred to as *unsupervised*) methods such as equal-width and equal-frequency [26], to much more sophisticated (often referred to as *supervised*) methods such as *Entropy* [15] and Pearson's X^2 or Wilks' G^2 statistics based discretization algorithms [5,18]. Unsupervised discretization methods are not provided with class label information whereas supervised discretization methods are supplied with a class label for each data item value. Liu et. al. [26] introduce a nice categorization of a large number of existing discretization methods.

In spite of the wealth of literature on discretization methods, there are very few attempts to *analytically* compare them. Typically, researchers compare the performance of different algorithms by providing experimental results of running these algorithms on publicly available data sets. In [12], Dougherty et al. compare discretization results obtained by unsupervised discretization versus a supervised method proposed by Holte [20] and the entropy based method proposed by Fayyad and Irani [15]. They conclude that supervised methods are better than unsupervised discretization method in that they generate fewer classification errors. In [25], Kohavi and Sahami report that the number of classification errors generated by the discretization method of Fayyad and Irani [15] is comparatively smaller than the number of errors generated by the discretization algorithm of Auer et al. [2]. They conclude that entropy based discretization methods are usually better than other supervised discretization algorithms.

Recently, many researchers have concentrated on the generation of new discretization algorithms [5,6,24,28,35]. The goal of the CAIM algorithm ptoposed by Kurgan and Klas [24] is to find the minimum number of intervals that minimize the loss between class-attribute interdependency. Boulle [5] has proposed a new discretization method called $Khiops$, which uses Pearson's X^2 statistic to merge consecutive intervals in order to improve the global dependence measure. MODL is another latest discretization method proposed by Boulle [7]. This method builds an optimal criteria based on a Bayesian model. A dynamic programming approach and a greedy heuristic approach are developed to find the optimal criteria. Yang and Webb [36] have studied discretization for naive-Bayes classifiers. They have proposed a couple of methods, such as *proportional k-interval discretization* and *equal size discretization*, to manage the discretization *bias* and *variance*. Bay [3] has studied multivariate discretization instead of the single variate discretization using a multivariate test of differences. All these algorithms have shown certain advantages, such as improving classification accuracy and/or complexity.

Several fundamental questions of discretization, however, remain to be answered. How these different methods are related to each other and how different or how similar are they? Is there an objective function which can measure the goodness of different approaches? If so, how would this function look like? In this paper we provide a set of positive results toward answering these questions.

Table 1 Notations for contingency table C

Intervals	Class 1	Class 2	$\cdots$	Class J	Row sum
S_1	c_{11}	c_{12}	$\cdots$	c_{1J}	N_1
S_2	c_{21}	c_{22}	$\cdots$	c_{2J}	N_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
S_I	c_{I1}	c_{I2}	$\cdots$	c_{IJ}	N_I
Column sum	M_1	M_2	$\cdots$	M_J	N (Total)

1.1 Problem statement

For the purpose of discretization, the entire dataset is projected onto the targeted continuous attribute. The result of such a projection is a two dimensional *contingency table*, C with I rows and J columns. Each row corresponds to either a point in the continuous domain, or an initial data interval. We treat each row as an atomic unit which cannot be further subdivided. Each column corresponds to a different class and we assume that the dataset has a total of J classes. A cell c_{ij} represents the number of points with jth class label falling in the ith point (or interval) in the targeted continuous domain. Table 1 lists the basic notations for the contingency table C.

In the most straightforward way, each continuous point (or initial data interval) corresponds to a row of a contingency table. Generally, in the initially given set of intervals each interval contains points from different classes and thus, c_{ij} may be more than zero for several columns in the same row.

The goal of a discretization method is to find another contingency table, C', with $I' << I$, where each row in the new table C' is the combination of several consecutive rows in the original C table, and each row in the original table is covered by exactly one row in the new table.

The quality of the discretization function is measured by a *goodness* function, which depends on two parameters. The first parameter (termed $cost(data)$) reflects the number of classification errors generated by the discretization function, whereas the second one (termed $penalty(model)$) is the complexity of the discretization which reflects the number of discretization intervals generated by the discretization function. Clearly, the more discretization intervals created, the fewer the number of classification errors, and thus the cost of the data is lower. That is, if one is interested only in minimizing the number of classification errors, the *best* discretization function would generate I intervals—the number of data points in the initial contingency table. Conversely, if one is only interested in minimizing the number of intervals (and therefore reducing the penalty of the model), then the *best* discretization function would generate a single interval by merging all data points into one interval. Thus, finding the *best* discretization is to find the best trade-off between the $cost(data)$ and the $penalty(model)$.

1.2 Our contribution

Our results can be summarized as follows:

1. We demonstrate a somewhat unexpected connection between discretization methods based on information theoretical complexity, on one hand, and the methods which are based on statistical measures of the data dependency of the contingency table, such as

Pearson's X^2 or G^2 statistics on the other hand. Namely, we prove that each goodness function defined in Fayyad and Irani [15], Hand et al. [17], Boulle [5], Kerber [23], Liu and Setiono [27], Breiman et al. [4], is a combination of G^2 defined by Wilks' statistic [1] and degrees of freedom of the contingency table multiplied by a function that is bounded by $O(\log N)$, where N is the number of data samples in the contingency table.
2. We define a notion of generalized entropy and introduce a notion of generalized goodness function. We prove that goodness functions for discretization methods based on MDLP, Gini index, AIC, BIC, Pearson's X^2, and G^2 statistic are all derivable from the generalized goodness function.
3. We design a dynamic programming algorithm that guarantees the best discretization based on a generalized goodness function.
4. We conduct an extensive performance evaluation of our discretization method and demonstrate that our method on the average outperforms by 31% the prior discretization methods for all publicly available data sets we tried.

1.3 Paper outline

Section 2 discusses goodness functions based on the MDLP [31], AIC, BIC [17], and goodness functions based on χ^2 and G^2 statistics [1]. Section 3 formulates criteria that we believe any *good* discretization function should satisfy, and we prove that all the functions introduced in Sect. 2 satisfy these goodness function properties. Section 4 contains the main result of the paper. We prove there that each goodness function defined in Sect. 2 is a sum of G^2 defined by Wilks' statistic [1] and degree of freedom of the contingency table multiplied by a function that is bounded by $O(\log\ N)$, where N is the number of data samples in the contingency table. Section 5 describes a generalized entropy function and formulates ageneralized notion of the goodness function. We prove then that any goodness function discussed in Section 2 is derivable from the generalized goodness function. Sect. 6 experimentally compare the new appoarach which utilizes the generalized goodness function with existing well-known discretization approaches, including *Entropy* and *Chi-Merge*, among others. Section 7 concludes the paper.

2 Goodness functions

In this section, we introduce a list of goodness functions which are used to evaluate different discretization for numerical attributes. These goodness functions intend to measure three different qualities of a contingency table: the information quality (Sect. 2.1), the fitness of statistical models (Sect. 2.2), and the confidence level for statistical independence tests (Sect. 2.3).

2.1 Information theoretical approach and MDLP

In the information theoretical approach, we treat discretization of a single continuous attribute as a 1-*dimension classification* problem. The MDLP is a commonly used approach for choosing the best classification model [18,31]. It considers two factors: how good the discretization fit the data, and the penalty for the discretization which is based on the complexity of discretization. Formally, MDLP associates a cost with each discretization, which has the

following form:

$$cost(model) = cost(data|model) + penalty(model)$$

where both terms correspond to these two factors, respectively. Intuitively, when a classification error increases, the penalty decreases and vice versa.

In MDLP, the cost of discretization ($cost(model)$) is calculated under the assumption that there are a sender and a receiver. Each of them knows all continuous points, but the receiver is without the knowledge of their labels. The cost of using a discretization model to describe the available data is then equal to the length of the shortest message to transfer the label of each continuous point. Thus, the first term ($cost(data|model)$) corresponds to the shortest message to transfer the label of all data points of each interval of a given discretization. The second term $penalty(model)$ corresponds to the coding book and delimiters to identify and translate the message for each interval at the receiver site. Given this, the cost of discretization based on MDLP ($Cost_{\text{MDLP}}$) is derived as follows:

$$\sum_{i=1}^{I'} N_i H(S_i) + (I' - 1)\log_2 \frac{N}{I' - 1} + I'(J - 1)\log_2 J \tag{1}$$

where $H(S_i)$ is the *entropy* of interval S_i, the first term corresponds to $cost(data|model)$, and the rest corresponds to $penalty(model)$.

Before we provide the detailed derivation for $Cost_{\text{MDLP}}$, we first formally introduce a notion of *entropy* and show how a merge of some adjacent data intervals results in information loss as well as in the increase of the $cost(data|model)$.

Definition 1 [11]. The entropy of an ensemble X is defined to be the average Shannon information content of an outcome:

$$H(X) = \sum_{x \in \mathcal{A}_x} P(x)\log_2 \frac{1}{P(x)}$$

where $\mathcal{A}_x$ is the possible outcome of x.

Let the i-th interval be S_i, which corresponds to the i-th row in the contingency table C. For simplicity, consider that we have only two intervals S_1 and S_2 in the contingency table, then the entropies for each individual interval is defined as follows:

$$H(S_1) = -\sum_{j=1}^{J} \frac{c_{1j}}{N_1}\log_2 \frac{c_{1j}}{N_1}, \quad H(S_2) = -\sum_{j=1}^{J} \frac{c_{2j}}{N_2}\log_2 \frac{c_{2j}}{N_2}$$

If we merge these intervals into a single interval (denoted by $S_1 \cup S_2$) following the same rule, we have the entropy as follows:

$$H(S_1 \cup S_2) = -\sum_{j=1}^{k} \frac{c_{1j} + c_{2j}}{N}\log_2 \frac{c_{1j} + c_{2j}}{N}$$

Further, if we treat each interval independently (without merging), the total entropy of these two intervals is expressed as $H(S_1, S_2)$, which is the weighted average of both individual entropies. Formally, we have

$$H(S_1, S_2) = \frac{N_1}{N} H(S_1) + \frac{N_2}{N} H(S_2)$$

From concaveness of the entropy function, it follows that $H(S_1, S_2) \leq H(S_1 \cup S_2)$. Thus, every merge operation leads to information loss. The entropy gives the lower bound of the cost to transfer the label per data point. This means that it takes a longer message to send all data points in these two intervals if they are merged ($N \times H(S_1 \cup S_2)$) than sending both intervals independently ($N \times H(S_1, S_2)$. However, after we merge, the number of intervals is reduced. Therefore, the discretization becomes simpler and the penalty of the model in $Cost_{\text{MDLP}}$ becomes smaller.

Derivation of $Cost_{\text{MDLP}}$:

For an interval S_1, the best way to transfer the labeling information of each point in the interval is bounded by a fundamental theorem in information theory, stating that the average length of the shortest message is higher than $N_1 \times H(S_1)$. Though we can apply the Hoffman coding to get the optimal coding for each interval, we are not interested in the absolute minimal coding. Therefore, we will apply the above formula as the cost to transfer each interval. Given this, we can easily derive the total cost to transfer all the I' intervals as follows.

$$cost_1(data|model) = N \times H(S_1, \ldots, S_{I'}) = N_1 \times H(S_1) + \cdots + N_{I'} \times H(S_{I'})$$

In the meantime, we have to transfer the model itself, which includes all the intervals and the coding book for transferring the point labels for each interval. The length of the message to transferring the model is served as the penalty function for the model. The cost to transfer all the intervals will require a $\log_2\binom{N+I'-1}{I'-1}$-bit message. This cost, denoted as $L_1(I', N)$, can be approximated as

$$\begin{aligned} L_1(I', N) &= \log_2\binom{N+I'-1}{I'-1} \approx (N+I'-1)H\left(\frac{N}{N+I'-1}, \frac{I'-1}{N+I'-1}\right) \\ &= -\left(N\log_2\frac{N}{N+I'-1} + (I'-1)\log_2\frac{I'-1}{N+I'-1}\right) \\ &\approx (I'-1)\log_2\frac{N+I'-1}{I'-1} \approx (I'-1)\log_2\frac{N}{I'-1} \\ &\quad \times \left(\log_2\frac{N}{N+I'-1} \to 0, \quad N \to \infty\right) \end{aligned}$$

Next, we have to consider the transfer of the coding book for each interval. For a given interval S_i, each code will correspond to a class, which can be coded in $\log_2 J$ bits. We need to transfer such codes at most $J-1$ times for each interval, since after knowing $J-1$ classes, the remaining class can be inferred. Therefore, the total cost for the coding book, denoted as L_2, can be written as

$$L_2 = I' \times (J-1) \times \log_2 J$$

Given this, the penalty of the discretization based on the theoretical viewpoint is

$$penalty_1(model) = L_1(I', N) + L_2 = (I'-1)\log_2\frac{N}{I'-1} + I' \times (J-1) \times \log_2 J$$

Put together, the cost of the discretization based on MDLP is

$$Cost_{\text{MDLP}} = \sum_{i=1}^{I'} N_i H(S_i) + (I'-1)\log_2\frac{N}{I'-1} + I'(J-1)\log_2 J$$

Goodness function based on MDLP: To facilitate the comparison with other cost functions, we formally define a goodness function of a MDLP based discretization method applied to contingency table C to be the difference between the cost of C^0, which is the resulting table after merging all the rows of C into a single row, and the cost of C. We will also use *natural log* instead of the $\log_2$ function. Formally, we denote the goodness function based on MDLP as GF_{MDLP}.

$$GF_{\text{MDLP}}(C) = Cost_{\text{MDLP}}(C^0) - Cost_{\text{MDLP}}(C) = N \times H(S_1 \cup \cdots \cup S_{I'}) \\ -N \times H(S_1, \ldots, S_{I'}) - \left((I'-1)\log\frac{N}{I'-1} + (I'-1)(J-1)\log J\right) \quad (2)$$

Note that for a discretization problem, any discretization method shares the same C^0. Thus, the least cost of transferring a contingency table corresponds to the maximum of the goodness function.

2.2 Statistical model based goodness functions

A different way to look at a contingency table is to assume that all data points are generated from certain distributions (models) with unknown parameters. Given a distribution, the maximal likelihood principle (MLP) can help us to find the best parameters to fit the data [17]. However, to provide a better data fitting, more expensive models (including more parameters) are needed. Statistical model selection tries to find the right balance between the complexity of a model corresponding to the number of parameters, and the fitness of the data to the selected model, which corresponds to the likelihood of the data being generated by the given model.

In statistics, the multinomial distribution is commonly used to model a contingency table. Here, we assume the data in each interval (or row) of the contingency table are independent and all intervals are independent. Thus, the kernel of the likelihood function for the entire contingency table is:

$$L(\vec{\pi}) = \prod_{i=1}^{I'}\left(\prod_{j=1}^{J}\pi_{j|i}^{c_{ij}}\right)$$

where $\vec{\pi} = (\pi_{1|1}, \pi_{2|1}, \ldots, \pi_{J|1}, \ldots, \pi_{J|I'})$ are the unknown parameters. Applying the *maximal likelihood* principle, we identify the best fitting parameters as $\pi_{j|i} = c_{ij}/N_i$, $1 \le i \le I'$, $1 \le j \le J$. We commonly transform the likelihood to the log-likelihood as follow:

$$S_L(D|\vec{\pi}) = -\log L(\vec{\pi}) = -\sum_{i=1}^{I'}\sum_{j=1}^{J} c_{ij}\log\frac{c_{ij}}{N_i}$$

According to Hand et al. [17], $S_L(\vec{\pi})$ is treated as a type of *entropy* term that measures how well the parameters $\vec{\pi}$ can compress (or predict) the training data.

Clearly, different discretizations correspond to different multinomial distributions (models). For choosing the best discretization model, the *Akaike information criterion* or AIC [17] can be used and it is defined as follows:

$$Cost_{\text{AIC}} = 2S_L(D|\vec{\pi}) + 2(I' \times (J-1)) \quad (3)$$

where, the first term corresponds to the fitness of the data given the discretization model, and the second term corresponds to the complexity of the model. Note that in this model for each

row we have the constraint $\pi_{1|i} + \cdots + \pi_{J|i} = 1$. Therefore, the number of parameters for each row is $J - 1$.

Alternatively, for choosing the best discretization model based on Bayesian arguments that take into account the size of the training set N is also frequently used. The *Bayesian information criterion* or BIC [17] is defined as follows:

$$Cost_{\text{BIC}} = 2S_L(D|\vec{\pi}) + (I' \times (J-1))\log N \tag{4}$$

In the BIC definition, the penalty of the model is higher than the one in the AIC by a factor of $1/2\log N$.

Goodness function based on AIC and BIC: For the same reason as MDLP, we denote the goodness function of a given contingency table based on AIC and BIC as the difference between the cost of C^0 (the resulting table after merging all the rows of C into a single row), and the cost of of C.

$$GF_{\text{AIC}}(C) = Cost_{\text{AIC}}(C^0) - Cost_{\text{AIC}}(C) \tag{5}$$

$$GF_{\text{BIC}}(C) = Cost_{\text{BIC}}(C^0) - Cost_{\text{BIC}}(C) \tag{6}$$

2.3 Confidence level based goodness functions

Another way to treat discretization is to merge intervals so that the rows (intervals) and columns (classes) of the entire contingency table become more statistically *dependent*. In other words, the goodness function of a contingency table measures its statistical quality in terms of independence tests.

Pearson's X^2: In the existing discretization approaches, the Pearson statistic X^2 [1] is commonly used to test the statistical independence. The X^2 statistic is as follows:

$$X^2 = \sum\sum \frac{(c_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}}$$

where, $\hat{m}_{ij} = N(N_i/N)(M_j/N)$ is the expected frequencies. It is well known that Pearson X^2 statistic has an asymptotic χ^2 distribution with degrees of freedom $df = (I'-1)(J-1)$, where I' is the total number of rows. Consider a null hypothesis H_0 (the rows and columns are statistically independent) against an alternative hypothesis H_a. Consequently, we obtain the confidence level of the statistical test to reject the independence hypothesis (H_0). The confidence level is calculated as

$$F_{\chi^2_{df}}(X^2) = \frac{1}{2^{df/2}\Gamma\left(\frac{df}{2}\right)} \int_0^t s^{\frac{df}{2}-1} \mathrm{e}^{-s/2} \mathrm{d}s$$

where, $F_{\chi^2_{df}}$ is the cumulative χ^2 distribution function. We use the calculated confidence level as our goodness function to compare different discretization methods that use Pearson's X^2 statistic. Our goodness function is formally defined as

$$GF_{X^2}(C) = F_{\chi^2_{df}}(X^2) \tag{7}$$

We note that $1 - F_{\chi^2_{df}}(X^2)$ is essentially the Pvalue of the aforementioned statistical independence test [7]. The lower the Pvalue (or equivalently, the higher the goodness), with more confidence we can reject the independence hypothesis (H_0). This approach has been used in Khiops [5], which describes a heuristic algorithm to perform discretization.

Wilks' G^2: In addition to Pearson's chi-square statistic, another statistic called likelihood-ratio χ^2 statistic, or Wilks' statistic [1], is used for the independence test. This statistic is derived from the likelihood-ratio test, which is a general-purpose way of testing a null hypothesis H_0 against an alternative hypothesis H_a. In this case we treat both intervals (rows) and the classes (columns) equally as two *categorical variables*, denoted as X and Y. Given this, the null hypothesis of statistical independence is $H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$ for all row i and column j, where $\{\pi_{ij}\}$ is the joint distribution of X and Y, and π_{i+} and π_{+j} are the marginal distributions for the row i and column j, respectively.

Based on the multinomial sampling assumption (a common assumption in a contingency table) and the maximal likelihood principle, these parameters can be estimated as $\hat{\pi}_{i+} = N_i/N$, $\hat{\pi}_{+j} = M_j/N$, and $\hat{\pi}_{ij} = N_i \times M_j/N^2$ (under H_0). In the general case under H_a, the likelihood is maximized when $\hat{\pi}_{ij} = c_{ij}/N$. Thus the statistical independence between the rows and the columns of a contingency table can be expressed as the ratio of the likelihoods:

$$\Lambda = \frac{\prod_{i=1}^{I'} \prod_{j=1}^{J} (N_i M_j/N^2)^{c_{ij}}}{\prod_{i=1}^{I'} \prod_{j=1}^{J} (c_{ij}/N)^{c_{ij}}}$$

where the denominator corresponds to the likelihood under H_a, and the nominator corresponds to the likelihood under H_0.

Wilks has shown that $-2\log\Lambda$, denoted by G^2, has a limiting null chi-squared distribution, as $N \to \infty$.

$$G^2 = -2\log\Lambda = 2\sum_{i=1}^{I'}\sum_{j=1}^{j} c_{ij}\log\frac{c_{ij}}{N_i M_j/N} \tag{8}$$

For large samples, G^2 has a chi-squared null distribution with degrees of freedom equal to $(I'-1)(J-1)$. Clearly, we can use G^2 to replace X^2 for calculating the confidence level of the entire contingency table, which serves as our goodness function

$$GF_{G^2}(C) = F_{\chi^2_{df}}(G^2) \tag{9}$$

Indeed, this statistic has been applied in discretization (though not for the global goodness function), and is referred to as class-attributes interdependency information [35].

3 Properties of goodness functions

An important theoretical question we address is how these methods introduced in Sect. 2 are related to each other and how different they are. Answering these questions helps to understand the scope of these approaches and shed light on the ultimate goal: for a given dataset, automatically find the best discretization method.

We first investigate some simple properties shared by the aforementioned goodness functions (Theorem 1). We describe four basic principles we believe any goodness function for discretization must satisfy.

1. **Merging principle (P1):** Let $S_i = \langle c_{i1}, \ldots, c_{iJ}\rangle$, and $S_{i+1} = \langle c_{i+1,1}, \ldots, c_{i+1,J}\rangle$ be two adjacent rows in the contingency table C. If $c_{ij}/N_i = c_{(i+1)j}/N_{i+1}, \forall j, 1 \le j \le J$, then $GF(C') > GF(C)$, where N_i and N_{i+1} are the row sums, GF is a goodness function and C' is the resulting contingency table after we merge these rows.
   Intuitively, this principle reflects a main goal of discretization, which is to transform the continuous attribute into a *compact* interval-based representation with minimal loss

of information. As we discussed before, a good discretization reduces the number of intervals without generating too many classification errors. Clearly, if two consecutive intervals have exactly the same data distribution, we cannot differentiate between them. In other words, we can merge them without information loss. Therefore, any goodness function should prefer to merge such consecutive intervals. We note that the merging principle (P1) echoes the cut point candidate pruning techniques for discretization which have been studied by Fayyand and Irani [15] and Elomaa and Rousu [13,14]. However, they did not explicitly define a global goodness function for discretization. Instead, their focus is either on evaluating the goodness of each single cut or on the goodness when the total target of intervals for discretization is given. As we mentioned in Sect. 1, the goodness function discussed in this paper is to capture the tradeoff between the information/statistical quality and the complexity of the discretization. In addition, this principle can be directly applied to reduce the size of the original contingency table since we can simply merge the consecutive rows with the same class distribution.

2. **Symmetric principle (P2):** Let C_j be the j-th column of contingency table C. $GF(C) = GF(C')$, where $C = \langle C_1, \ldots C_J \rangle$ and C' is obtained from C by an arbitrary permutation of C's columns.
   This principle asserts that the order of class labels should not impact the goodness function that measures the quality of the discretization. Discretization results must be the same for both tables.
3. **MIN principle (P3):** Consider all contingency tables C which have I rows and J columns, and the same marginal distribution for classes (columns). If for any row S_i in C, $S_i = \langle c_{i1}, \ldots, c_{iJ} \rangle$, $c_{ij}/N_i = M_j/N$, then the contingency table C reaches the minimum for any goodness function.
   This principle determines what is the worst possible discretization for any contingency table. This is the case when each row shares exactly the same class distribution in a contingency table, and thus the entire table has the maximal redundancy.
4. **MAX principle (P4):** Consider all the contingency tables C which have I rows and J columns. If for any row S_i in C, $S_i = \langle c_{i1}, \ldots, c_{iJ} \rangle$, there exists one cell count such that $c_{ij} \neq 0$, and others $c_{ik}, k \neq j, c_{ik} = 0$, then the contingency table C achieves the maximum in terms of a goodness function for any $I \times J$ contingency table.
   This principle determines what is the best possible discretization when the number of intervals is fixed. Clearly, the best discretization is achieved if we have the maximal discriminating power in each interval. This is the case where all the data points in each interval belong to only one class.

The following theorem states that all aforementioned goodness functions satisfy these four principles.

Theorem 1 *GF_{MDLP}, GF_{AIC}, GF_{BIC}, GF_{X^2}, GF_{G^2} satisfy all four principles,* **P1**, **P2**, **P3**, *and* **P4**.

Proof We will first focus on proving for GF_{MDLP}. The proof for GF_{AIC} and GF_{BIC} can be derived similarly.

Merging principle (P1) for GF_{MDLP}: Assuming we have two consecutive rows i and $i+1$ in the contingency table C, $S_i = \langle c_{i1}, \ldots, c_{iJ} \rangle$, and $S_{i+1} = \langle c_{i+1,1}, \ldots, c_{i+1,J} \rangle$, where $c_{ij} = c_{i+1,j}, \forall i, 1 \leq j \leq J$. Let C' be the resulting contingency table after we merge these two rows. Then we have

$$\begin{aligned}
&\sum_{k=1}^{I} N_k \times H(S_k) \\
&= \sum_{k=1}^{i-1} N_k \times H(S_k) + N_i \times H(S_i) + N_{i+1} \times H(S_{i+1}) + \sum_{k=i+2}^{I} N_k \times H(S_k) \\
&= \sum_{k=1}^{i-1} N_k \times H(S_k) + (N_i + N_{i+1}) \times H(S_i) + \sum_{k=i+2}^{I} N_k \times H(S_k) = \sum_{k=1}^{I-1} N'_k H(S_k)'
\end{aligned}$$

In addition, let

$$\begin{aligned}
\delta &= (I-1)\log_2 \frac{N}{I-1} + (I-1) \times J \times \log_2 J - ((I-2)\log_2 \frac{N}{I-2} + (I-2) \times J \times \log_2 J) \\
&= (I-1)\log_2 N/(I-1) - (I-2)\log_2 N/(I-2) + J \times \log_2 J
\end{aligned}$$

For $I = 2$, $\delta = \log_2 N + J \times \log_2 J > 0$.
For $I \geq 3$, we need more detailed analysis. Let $f = N/(I-1) \geq 1$. Given this, we have

$$(I-1)\log_2 \frac{N}{I-1} - (I-2)\log_2 \frac{N}{I-2} = \log_2 f - (I-2)\log_2 \frac{I-1}{I-2}$$

Based on the first order derivative of and second order derivative analysis, we can see that $(I-2)\log_2 \frac{I-1}{I-2}$ is strictly increasing when $I \geq 3$. It will converge to $\log_2 e$ as $I \to \infty$. Thus, we have

$$\delta = J\log_2 J + \log_2 f - (I-2)\log_2 \frac{I-1}{I-2} \geq J\log_2 J - \log_2 e > 0 \quad (J \geq 2)$$

Adding together, we have $Cost_{\text{MDLP}}(C) > Cost_{\text{MDLP}}(C')$, and $GF_{\text{MDLP}}(C) < GF_{\text{MDLP}}(C')$.

Symmetric principle (P2) for GF_{MDLP}: This can be directly derived from the symmetric property of entropy.

MIN principle (P3) for GF_{MDLP}: Since the number of rows (I), the number of samples (N), and the number of classes (J) are fixed, we only need to maximize $N \times H(S_1, \ldots, S_I)$.

$$\begin{aligned}
&N \times H(S_1, \ldots, S_I) \leq N \times H(S_1 \cup \cdots \cup S_I) \\
&N \times H(S_1, \ldots, S_I) = \sum_{k=1}^{I} N_k \times H(S_k) = N \times H(S_1 \cup \cdots \cup S_I)
\end{aligned}$$

MAX principle (P4) for GF_{MDLP}: Since the number of rows (I), the number of samples (N), and the number of classes (J) are fixed, we only need to minimize $N \times H(S_1, \ldots, S_J)$.

$$N \times H(S_1, \ldots, S_J) = \sum_{k=1}^{J} N_k \times H(S_k) \geq \sum_{k=1}^{J} N_k \times (\log_2 1) \geq 0$$

Now, we prove the four properties for GF_{X^2}.

Merging principle (P1) for GF_{X^2}: Assuming we have two consecutive rows i and $i+1$ in the contingency table C, $S_i = \langle c_{i1}, \ldots, c_{iJ} \rangle$, and $S_{i+1} = \langle c_{i+1,1}, \ldots, c_{i+1,J} \rangle$, where $c_{ij} = c_{i+1,j}, \forall i, 1 \leq j \leq J$. Let C' be the resulting contingency table after we merge these two rows. Then we have

$$
\begin{aligned}
X_C^2 - X_{C'}^2 &= \sum\sum \frac{(c_{kj} - N_i \times M_j/N)^2}{N_k \times M_j/N} - \sum\sum \frac{(c'_{kj} - N'_i \times M_j/N)^2}{N'_k \times M_j/N} \\
&= \sum_{j=1}^{J} \frac{(c_{ij} - N_i \times M_j/N)^2}{N_i \times M_j/N} + \sum_{j=1}^{J} \frac{(c_{i+1,j} - N_{i+1} \times M_j/N)^2}{N_{i+1} \times M_j/N} \\
&\quad - \sum_{j=1}^{J} \frac{((c_{ij} + c_{i+1,j}) - (N_i + N_{i+1}) \times M_j/N)^2}{(N_i + N_{i+1}) \times M_j/N} \\
&= 2 \times \sum_{j=1}^{J} \frac{(c_{ij} - N_i \times M_j/N)^2}{N_i \times M_j/N} - \sum_{j=1}^{J} \frac{(2c_{ij} - 2N_i \times M_j/N)^2}{2N_i \times M_j/N} \\
&= 2 \times \sum_{j=1}^{J} \frac{(c_{ij} - N_i \times M_j/N)^2}{N_i \times M_j/N} - \sum_{j=1}^{J} \frac{4 \times (c_{ij} - N_i \times M_j/N)^2}{2N_i \times M_j/N} = 0
\end{aligned}
$$

We note that the degree of freedom in the original contingency table is $(I-1)(J-1)$ and the second one is $(I-2)(J-1)$. In addition, we have for any $t > 0$, $F_{\chi^2_{(I-1)(J-1)}}(t) < F_{\chi^2_{(I-2)(J-1)}}(t)$. Therefore, the second table is better than the first one.

Symmetric principle (P2) for GF_{X^2}: This can be directly derived from the symmetric property of X^2.

MIN principle (P3) for GF_{X^2}: Since the number of rows (I), the number of samples (N), and the number of classes (J) are fixed, we only need to minimize X^2. Since $c_{kj} = 1/J \times N_i$, we can see that $M_j = N/J$.

$$
X^2 = \sum\sum \frac{(c_{kj} - N_k \times M_j/N)^2}{N_k \times M_j/N} = \sum\sum \frac{(M_j/N \times N_k - N_k \times M_j/N)^2}{N_k \times M_j/J} = 0
$$

Since $X^2 \geq 0$, we achieve the minimal of X^2.

MAX principle (P4) for GF_{X^2}: Since the number of rows (J), the number of samples (N), and the number of classes (J) are fixed, we only need to maximize X^2.

$$
\begin{aligned}
X^2 &= \sum\sum \frac{(c_{kj} - N_k \times M_j/N)^2}{N_k \times M_j/N} \\
&= \sum\sum \frac{c_{kj}^2 + (N_k \times M_j/N)^2 - 2 \times c_{kj} \times N_k \times M_j/N}{N_k \times M_j/N} \\
&= \sum\sum \left(\frac{c_{kj}^2}{N_k \times M_j/N} + N_k \times M_j/N - 2 \times c_{kj} \right) \\
&= \sum\sum_{k=1}^{J} N/M_j \times \left[c_{kj} \frac{c_{kj}}{N_k} \right] + N - 2N \qquad \left(\frac{c_{kj}}{N_k} \leq 1 \right) \\
&\leq \sum_{j=1}^{J} (N/M_j) \times \sum_{k=1}^{J} c_{kj} - N = \sum_{j=1}^{J} (N/M_j) \times M_j - N = (J-1) \times N
\end{aligned}
$$

Note that this bound can be achieved in our condition. Basically, in any row k, we will have one cell $c_{kj} = N_k$. Therefore, $F_{\chi^2_{(I-1)(J-1)}}(X^2)$ is maximized. In other words, we have the best possible discretization given existing conditions.

The proof for GF_{G^2} can be derived similarly from GF_{MDLP} and GF_{X^2}. □

4 Equivalence of goodness functions

In this section, we analytically compare different discretization goodness functions introduced in Sect. 2. In particular, we find some rather surprising connection between these seemingly quite different approaches: the information theoretical complexity, the statistical fitness, and the statistical independence tests. We basically prove that all these functions can be expressed in a uniform format as follows:

$$GF = G^2 - df \times f(G^2, N, I, J) \tag{10}$$

where, df is a degree of freedom of the contingency table, N is the number of data points, I is the number of data rows in the contingency table, J is the number of class labels and f is bounded by $O(\log N)$. The first term G^2 corresponds to the cost of the data given a discretization model ($cost(data|model)$), and the second corresponds to the penalty or the complexity of the model ($penalty(model)$).

To derive this expression, we first derive an expression for the cost of the data for different goodness functions discussed in Sect. 2 (Sect. 4.1). This is achieved by expressing G^2 statistics through information entropy (Theorem 3). Then, using a Wallace's result [33,34] on approximating χ^2 distribution with a normal distribution, we transform the goodness function based on statistical independence tests into the format of Formula 10. Further, a detailed analysis of function f reveals a deeper relationship shared by these different goodness functions (Sect. 4.3).

4.1 Unifying the cost of data to G^2

In the following, we establish the relationship among *entropy*, *log-likelihood* and G^2. This is the first step for an analytical comparison of goodness functions based on the information theoretical, the statistical model selection, and the statistical independence test approaches.

First, it is easy to see that for a given contingency table, the cost of the data transfer ($cost(data|model)$, a key term in the information theoretical approach) is equivalent to the log likelihood S_L (used in the statistical model selection approach) as the following theorem asserts.

Theorem 2 *For a given contingency table $C_{I' \times J}$, the cost of data transfer ($cost_1(data|model)$) is equal to the log-likelihood S_L, i.e.*

$$N \times H(S_1, \ldots, S_{I'}) = S_L = -\sum_{i=1}^{I'} \sum_{j=1}^{J} c_{ij} \log \frac{c_{ij}}{N_i}$$

Proof

$$N \times H(S_1, \ldots, S_{I'}) = -\sum_{i=1}^{I'} N_i \times H(S_i) = -\sum_{i=1}^{I'} N_i \times \sum_{j=1}^{J} \frac{c_{ij}}{N_i} \log \frac{c_{ij}}{N_i} = -\log L(\vec{\pi})$$

□

The next theorem establishes a relationship between entropy criteria and the likelihood independence testing statistics G^2. This is the key to discover the connection between the information theoretical and the statistical independence test approaches.

Theorem 3 *Let C be a contingency table. Then*

$$G^2/2 = N \times H(S_1 \cup \cdots \cup S_{I'}) - N \times H(S_1, \ldots, S_{I'})$$

Proof

$$\begin{aligned} G^2/2 = -\log\Lambda &= \sum_{i=1}^{I'}\sum_{j=1}^{J} c_{ij}\log\frac{c_{ij}}{N_i M_j/N} = \sum_{i=1}^{I'}\sum_{j=1}^{J}\left(c_{ij}\log\frac{c_{ij}}{N_i} + c_{ij}\log\frac{N}{M_j}\right) \\ &= \sum_{i=1}^{I'}\sum_{j=1}^{J} c_{ij}\log\frac{c_{ij}}{N_i} - \sum_{j=1}^{J}\log\frac{M_j}{N} \times \sum_{i=1}^{I'} c_{ij} = \sum_{i=1}^{I'}\sum_{j=1}^{J} c_{ij}\log\frac{c_{ij}}{N_i} - \sum_{j=1}^{J} M_j\log\frac{M_j}{N} \\ &= -N \times (H(S_1, \ldots, S_{I'}) \ + \ H(S_1 \cup \cdots \cup S_{I'})) \end{aligned}$$

□

Consequently, we rewrite the goodness functions GF_{MDLP}, GF_{AIC} and GF_{BIC} as follows.

$$GF_{\text{MDLP}} = G^2 - 2(I'-1)\log\frac{N}{I'-1} - 2(I'-1)(J'-1)\log J \tag{11}$$

$$GF_{\text{AIC}} = G^2 - (I'-1)(J-1) \tag{12}$$

$$GF_{\text{BIC}} = G^2 - (I'-1)(J-1)\log N/2 \tag{13}$$

For the rest of the paper we use the above formulas for GF_{MDLP}, GF_{AIC} and GF_{BIC}.

It has long been known that they are asymptotically equivalent. The next theorem provides tool to connect the information theoretical approach and the statistical independence test approach based on Pearson's chi-square (X^2) statistic.

Theorem 4 [1]. *Let N be the total number of data values in the contingency table T of $I \times J$ dimensions. If the rows (columns) of contingency table are independent, then probability of $X^2 - G^2 = 0$ converges to one as $N \to \infty$.*

In the following, we mainly focus on the asymptotic properties shared by X^2 and G^2 based cost functions. Thus, our further discussions on G^2 can also be applied to X^2.

Note that Theorems 2 and 3 basically establish the basis for Formula 10 of goodness functions based on the information theoretical approach and statistical model selection approaches. Even though Theorems 3 and 4 relate the information theoretical approach (based on entropy) to the statistical independence test approach (based on G^2 and X^2), it is still unclear how to compare them directly since the goodness function of the former one is based on the total *cost* of transferring the data and the goodness function of the latter one is the *confidence level* for a hypothesis test. Section 4.2 presents our approach on tackling this issue.

4.2 Unifying statistical independence tests

In order to compare the quality of different goodness functions, we introduce a notion of *equivalent* goodness functions. Intuitively, the equivalence between goodness functions means that these functions rank different discretization of the same contingency table identically.

Definition 2 Let C be a contingency table and $GF_1(C)$, $GF_2(C)$ be two different goodness functions. GF_1 and GF_2 are equivalent if and only if for any two contingency tables C_1 and C_2, $GF_1(C_1) \leq GF_1(C_2) \Longleftrightarrow GF_2(C_1) \leq GF_2(C_2)$.

Using the equivalence notion, we transform goodness functions to different scales and/or to different formats. In the sequel, we apply this notion to compare seemingly different goodness functions.

The relationship between the G^2 and the confidence level is rather complicated. It is clearly not a simple one-to-one mapping as the same G^2 may correspond to very different confidence level depending on degrees of freedom of the χ^2 distribution and, vice versa the same confidence level may correspond to very different G^2 values. Interestingly enough, such many-to-many mapping actually holds the key for the aforementioned transformation. Intuitively, we have to transform the confidence interval to a scale of entropy or G^2 parameterized by the degree of freedom for the χ^2 distribution.

Our proposed transformation is as follows.

Definition 3 Let $u(t)$ be the normal deviate of the chi-square distributed variable t [21]. That is, the following equality holds:

$$F_{\chi^2_{df}}(t) = \Phi(u(t))$$

where, $F_{\chi^2_{df}}$ is the χ^2 distribution function with df degrees of freedom, and Φ is the normal distribution function. For a given contingency table C, which has the log likelihood ratio G^2, we define

$$GF'_{G^2} = u(G^2) \tag{14}$$

as a new goodness function for C.

The next theorem establishes the equivalence between a goodness functions GF_{G^2} and GF'_{G^2}.

Theorem 5 *The goodness function $GF'_{G^2} = u(G^2)$ is equivalent to the goodness function $GF_{G^2} = F_{\chi^2_{df}}(G^2)$.*

Proof Assuming we have two contingency tables C_1 and C_2 with degree of freedom df_1 and df_2, respectively. Their respective G^2 statistics are denoted as G_1^2 and G_2^2. Clearly, we have

$$F_{\chi^2_{df_1}}(G_1^2) \leq F_{\chi^2_{df_2}}(G_2^2) \Longleftrightarrow \Phi(u(G_1^2)) \leq \Phi(u(G_2^2)) \Longleftrightarrow u(G_1^2) \leq u(G_2^2)$$

This basically establishes the equivalence of these two goodness functions. □

The newly introduced goodness function GF'_{G^2} is rather complicated and it is hard to find for it a closed form expression. In the following, we use a theorem from Wallace [33,34] to derive an asymptotically accurate closed form expression for a simple variant of GF'_{G^2}.

Theorem 6 [33,34]. *For all $t > df$, all $df > 0.37$, and with $w(t) = [t - df - df \log(t/df)]^{\frac{1}{2}}$,*

$$0 < w(t) \leq u(t) \leq w(t) + 0.60 df^{-\frac{1}{2}}$$

Note that if $u(G^2) \geq 0$, then, $u^2(G^2)$ is *equivalent* to $u(G^2)$. Here, we limit our attention only to the case when $G^2 > df$, which is the condition for Theorem 6. This condition implies that $u(G^2) \geq 0$.[1] We show that under some conditions, $u^2(G^2)$ can be approximated as $w^2(G^2)$.

$$w^2(G^2) \leq u^2(G^2) \leq w^2(G^2) + \frac{0.36}{df} + 1.2\frac{w(G^2)}{\sqrt{df}}$$

$$1 \leq \frac{u^2(G^2)}{w^2(G^2)} \leq 1 + \frac{0.36}{w^2(G^2) \times df} + \frac{1.2}{w(G^2)\sqrt{df}}$$

If $df \to \infty$ and $w(t) >> 0$, then

$$\frac{0.36}{w^2(G^2) \times df} \to 0 \quad \text{and} \quad \frac{1.2}{w(G^2)\sqrt{df}} \to 0$$

Therefore,

$$\frac{u^2(G^2)}{w^2(G^2)} \to 1$$

Thus, we can have the following goodness function:

$$GF''_{G^2} = u^2(G^2) = G^2 - df\left(1 + \log\left(\frac{G^2}{df}\right)\right) \tag{15}$$

Similarly, function GF''_{χ^2} is obtained from GF''_{G^2} by replacing in the GF''_{G^2} expression G^2 with X^2. Formulas 11, 12, 13 and 15 indicate that all goodness functions introduced in Sect. 2 can be (asymptotically) expressed in the same closed form (Formula 10). Specifically, all of them can be decomposed into two parts. The first part contains G^2, which corresponds to the cost of transferring the data using information theoretical view. The second part is a linear function of degrees of freedom, and can be treated as the penalty of the model using the same view.

4.3 Penalty analysis

In this section, we perform a detailed analysis of the relationship between penalty functions of these different goodness functions . Our analysis reveals a deeper similarity shared by these functions and at the same time reveals differences between them.

Simply put, the penalties of these goodness functions are essentially bounded by two extremes. On the lower end, which is represented by AIC, the penalty is on the order of degree of freedom, $O(df)$. On the higher end, which is represented by BIC, the penalty is $O(df \log N)$.

Penalty of GF''_{G^2} (Formula 15): The penalty of our new goodness function $GF''_{G^2} = u^2(G^2)$ is between $O(df)$ and $O(df \log N)$. The lower bound is achieved, provided that G^2 being strictly higher than df ($G^2 > df$). Lemma 1 gives the upper bound.

Lemma 1 *G^2 is bounded by $2N\log J$ ($G^2 \leq 2N\log J$).*

[1] If $u(G^2) < 0$, it becomes very hard to reject the hypothesis that the entire table is statistically independent. Here, we basically focus on the cases where this hypothesis is likely to be reject.

Proof

$$\begin{aligned} G^2 &= 2N \times (H(S_1 \cup \cdots \cup S_I) - H(S_1, \ldots, S_I)) \\ &\leq 2N \times (-J \times (1/J \times \log(1/J)) - 0) \leq 2N \times \log J \end{aligned}$$

□

In the following, we consider two cases for the penalty $GF''_{G^2} = u^2(G^2)$. Note that these two cases corresponding to the lower bound and upper bound of G^2, respectively.

1. if $G^2 = c_1 \times df$, where $c_1 > 1$, the penalty of this goodness function is $(1 + \log c_1)df$, which is $O(df)$.
2. if $G^2 = c_2 \times N \log J$, where $c_2 \leq 2$ and $c_2 >> 0$, the penalty of the goodness function is $(1 + \log(c_2 N \log J/df))$.

The second case is further subdivided into two subcases.

1. If $N/df \approx N/(IJ) = c$, where c is some constant, the penalty is $O(df)$.
2. If $N \to \infty$ and $N/df \approx N/(IJ) \to \infty$, the penalty is

$$df(1 + \log(c_2 N \log J/df) \approx df(1 + \log N/df) \approx df(\log N)$$

Penalty of GF_{MDLP} (Formula 11): The penalty function f derived in the goodness function based on the information theoretical approach can be written as

$$\frac{df}{J-1} \log \frac{N}{I-1} + df \log J = df \left(\log \frac{N}{I-1} / (J-1) + \log J \right)$$

Here, we again consider two cases:

1. If $N/(I-1) = c$, where c is some constant, we have the penalty of MDLP is $O(df)$.
2. If $N >> I$ and $N \to \infty$, we have the penalty of MDLP is $O(df \log N)$.

Note that in the first case, the contingency table is very sparse ($N/(IJ)$ is small). In the second case, the contingency table is very dense ($N/(IJ)$ is very large).

To summarize, the penalty can be represented in a generic form as $df \times f(G^2, N, I, J)$ (Formula 10). This function f is bounded by $O(\log N)$. Finally, we observe that different penalty clearly results in different discretization. The higher penalty in the goodness function results in the less number of intervals in the discretization results. For instance, we can state the following theorem.

Theorem 7 *Given an initial contingency table C with* $\log N \geq 2$ (*the condition for the penalty of BIC is higher than the penalty of AIC*), *let I_{AIC} be the number of intervals of the discretization generated by using GF_{AIC} and I_{BIC} be the number of intervals of the discretization generated by using GF_{BIC}. Then $I_{AIC} \geq I_{BIC}$.*

Note that this is essentially a direct application of the well-known facts from statistical machine learning research: higher penalty will result in more concise models [17].

Finally, we note that several well-known discretization algorithms based on local independence test include ChiMerge [23] and Chi2 [27], etc. Specifically, for consecutive intervals, these algorithms perform a statistical independence test based on Pearson's X^2 or G^2. If they could not reject the independence hypothesis for those intervals, they merge them into one row. A simple analysis in [22] suggests that the local merge condition essentially shares the penalty in the same order of magnitude as GF_{AIC}. Interested readers can refer [22] for detailed discussion.

5 Parametrized goodness function

The goodness functions discussed so far are either entropy or χ^2 or G^2 statistics based. In this section we introduce a new goodness function which is based on *gini* index [4]. *Gini* index based goodness function is strikingly different from goodness functions introduced so far. In this section we show that a newly introduced goodness function GF_{Gini} along with the goodness functions discussed in Sect. 2 are all can be derived from a generalized notion of entropy [29].

5.1 Gini based goodness function

Let S_i be a row in contingency table C. *Gini* index of row S_i is defined as follows [4]:

$$Gini(S_i) = \sum_{j=1}^{J} \frac{c_{ij}}{N_i}\left[1 - \frac{c_{ij}}{N_i}\right]$$

and $Cost_{Gini}(C) = \sum_{i=1}^{I'} N_i \times Gini(S_i)$.

The penalty of the model based on gini index can be approximated as $2I' - 1$ (see detailed derivation in the technical report [22]). The basic idea is to apply a generalized MDLP principle in such a way so that the cost of transferring the data ($cost(data|model)$) and the cost of transferring the coding book as well as necessary delimiters ($penalty(model)$) are treated as the *complexity* measure. Therefore, the gini index can be utilized to provide such a measure. Thus, the goodness function based on gini index is as follows:

$$GF_{\text{Gini}}(C) = -\sum_{i=1}^{I'}\sum_{j=1}^{J} \frac{c_{ij}^2}{N_i} + \sum_{j=1}^{J} \frac{M_j^2}{N} + 2(I' - 1) \tag{16}$$

5.2 Generalized entropy

In this subsection, we introduce a notion of *generalized* entropy, which is used to uniformly represent a variety of *complexity* measures, including both information entropy and gini index by assigning different values to the parameters of the generalized entropy expression. Thus, it serves as the basis to derive the parameterized goodness function which represents all the aforementioned goodness functions, such as GF_{MDLP}, GF_{AIC}, GF_{BIC}, GF_{G^2}, and GF_{Gini}, in a closed form.

Definition 4 [29,32]. For a given interval S_i, the generalized entropy is defined as

$$H_\beta(S_i) = \sum_{j=1}^{J} \frac{c_{ij}}{N_i}\left[1 - \left(\frac{c_{ij}}{N_i}\right)^\beta\right]/\beta, \quad \beta > 0$$

When $\beta = 1$, we can see that

$$H_1(S_i) = \sum_{j=1}^{J} \frac{c_{ij}}{N_i}\left[1 - \frac{c_{ij}}{N_i}\right] = Gini(S_i)$$

When $\beta \to 0$,

$$H_{\beta\to 0}(S_i) = -\sum_{j=1}^{J} \frac{c_{ij}}{N_1}\log\frac{c_{ij}}{N_1} = H(S_i)$$

Let $C_{I\times J}$ be a contingency table. We define the generalized entropy for C as follows.

$$H_\beta(S_1,\ldots,S_I)=\sum_{i=1}^{I}\frac{N_i}{N}H_\beta(S_i)$$

$$H_\beta(S_1\cup\cdots\cup S_I)=\sum_{j=1}^{J}\frac{M_j}{N}\left[1-\left(\frac{M_j}{N}\right)^\beta\right]/\beta$$

Lemma 2 *$H_\beta[p_1,\ldots,p_J]=\sum_{j=1}^{J}p_i(1-p_i^\beta)/\beta$ is concave when $\beta>0$.*

Proof

$$\frac{\partial H_\beta}{\partial p_i}=(1-(1+\beta)p_i^\beta/\beta$$

$$\frac{\partial^2 H_\beta}{\partial^2 p_i}=-(1+\beta)p_i^{\beta-1}/\beta<0$$

$$\frac{\partial^2 H_\beta}{\partial p_i\partial p_j}=0$$

Thus,

$$\nabla^2 H_\beta[p_1,\ldots,p_J]=\begin{bmatrix}\frac{\partial^2 H_\beta}{\partial^2 p_1} & \cdots & \frac{\partial^2 H_\beta}{\partial p_1\partial p_J}\\ \vdots & \vdots & \vdots\\ \frac{\partial^2 H_\beta}{\partial p_1\partial p_J} & \cdots & \frac{\partial^2 H_\beta}{\partial^2 p_J}\end{bmatrix}$$

Clearly, $\nabla^2 H_\beta[p_1,\ldots,p_J]$ is negative definite. Therefore, $H_\beta[p_1,\ldots,p_J]$ is concave. □

Let $C_{I\times J}$ be a contingency table., We define the generalized entropy for C as follows.

$$H_\beta(S_1,\ldots,S_I)=\sum_{i=1}^{I}\frac{N_i}{N}H_\beta(S_i)=\sum_{i=1}^{I}\frac{N_i}{N}\times\sum_{j=1}^{J}\frac{c_{ij}}{N_i}\left[1-\left(\frac{c_{ij}}{N_i}\right)^\beta\right]/\beta$$

Similarly, we have

$$H_\beta(S_1\cup\cdots\cup S_I)=\sum_{j=1}^{J}\frac{M_j}{N}\left[1-\left(\frac{M_j}{N}\right)^\beta\right]/\beta$$

Theorem 8 *There always exists information loss for the merged intervals*: $H_\beta(S_1,S_2)\le H_\beta(S_1\cup S_2)$

Proof This is the direct application of the concaveness of the generalized entropy. □

5.3 Parameterized goodness function

Based on the discussion in Sect. 4, we derive that different goodness functions basically can be decomposed into two parts. The first part is for G^2, which corresponds to the information theoretical difference between the contingency table under consideration and the marginal distribution along classes. The second part is the penalty which counts the difference of complexity for the model between the contingency table under consideration and the one-row

contingency table. The different goodness functions essentially have different penalties ranging from $O(df)$ to $O(df \log N)$.

In the following, we propose a parameterized goodness function which treats all the aforementioned goodness functions in a uniform way.

Definition 5 Given two parameters, α and β, where $0 < \beta \leq 1$ and $0 < \alpha$, the parameterized goodness function for contingency table C is represented as

$$GF_{\alpha,\beta}(C) = N \times H_\beta(S_1 \cup \cdots \cup S_{I'}) - \sum_{i=1}^{I'} N_i \times H_\beta(S_i)$$
$$-\alpha \times (I' - 1)(J - 1)\left[1 - \left(\frac{1}{N}\right)^\beta\right]/\beta \tag{17}$$

The following theorem states the basic properties of the parameterized goodness function.

Theorem 9 *The parameter goodness function $GF_{\alpha,\beta}$, with $\alpha > 0$ and $0 < \beta \leq 1$, satisfies all four principles,* **P1**, **P2**, **P3**, *and* **P4**.

Proof **Merging principle (P1) for $GF_{\alpha,\beta}$:** Assuming we have two consecutive rows i and $i+1$ in the contingency table C, $S_i = \langle c_{i1}, \ldots, c_{iJ}\rangle$, and $S_{i+1} = \langle c_{i+1,1}, \ldots, c_{i+1,J}\rangle$, where $c_{ij} = c_{i+1,j}, \forall i, 1 \leq j \leq J$. Let C' be the resulting contingency table after we merge these two rows. Then we have

$$\begin{aligned}
&\sum_{k=1}^{I} N_k \times H_\beta(S_k) \\
&= \sum_{k=1}^{i-1} N_k \times H_\beta(S_k) + N_i \times H_\beta(S_i) + N_{i+1} \times H_\beta(S_{i+1}) + \sum_{k=i+2}^{I} N_k \times H_\beta(S_k) \\
&= \sum_{k=1}^{i-1} N_k \times H_\beta(S_k) + (N_i + N_{i+1}) \times H_\beta(S_i) + \sum_{k=i+2}^{I} N_k \times H_\beta(S_k) = \sum_{k=1}^{I-1} N'_k H_\beta(S_k)'
\end{aligned}$$

In addition, we have

$$\begin{aligned}
&\alpha \times (I-1)(J-1)\left[1 - \left(\frac{1}{N}\right)^\beta\right]/\beta - \alpha \times (I-2)(J-1)\left[1 - \left(\frac{1}{N}\right)^\beta\right]/\beta \\
&= \alpha \times (J-1)\left[1 - \left(\frac{1}{N}\right)^\beta\right]/\beta > 0
\end{aligned}$$

Thus, we have $GF_{\alpha,\beta}(C) < GF_{\alpha,\beta}(C')$.

Symmetric principle (P2) for $GF_{\alpha,\beta}$: This can be directly derived from the symmetric property of entropy.

MIN principle (P3) for $GF_{\alpha,\beta}$: Since the number of rows (I), the number of samples (N), and the number of classes (J) are fixed, we only need to maximize $N \times H(S_1, \ldots, S_I)$. By the concaveness of the H_β (Theorem 8),

$$\begin{aligned}
N \times H_\beta(S_1, \ldots, S_I) &\leq N \times H_\beta(S_1 \cup \cdots \cup S_I) \\
N \times H_\beta(S_1, \ldots, S_I) &= \sum_{k=1}^{I} N_k \times H_\beta(S_k) = \sum_{k=1}^{I} N_k \times H_\beta(S_1 \cup \cdots \cup S_I) \\
&= N \times H_\beta(S_1 \cup \cdots \cup S_I)
\end{aligned}$$

MAX principle (P4) for $GF_{\alpha,\beta}$: Since the number of rows (I), the number of samples (N), and the number of classes (J) are fixed, we only need to minimize $N \times H_\beta(S_1, \ldots, S_J)$.

$$N \times H_\beta(S_1, \ldots, S_J) = \sum_{k=1}^{J} N_k \times H_\beta(S_k) \geq \sum_{k=1}^{J} N_k \times 0 \geq 0$$

Note that the proof of $GF_{\alpha,\beta}$ immediately implies that the four principles hold for GF_{AIC} and GF_{BIC}. □

By adjusting different parameter values, we show how goodness functions defined in Sect. 2 can be obtained from the parametrized goodness function. We consider several cases:

1. Let $\beta = 1$ and $\alpha = 2(N-1)/(N(J-1))$. Then $GF_{2(N-1)/(N(J-1)),1} = GF_{\text{Gini}}$.
2. Let $\alpha = 1/\log N$ and $\beta \to 0$. Then $GF_{1/\log N\beta \to 0} = GF_{\text{AIC}}$.
3. Let $\alpha = 1/2$ and $\beta \to 0$. Then $GF_{1/2,\beta\to 0} = GF_{\text{BIC}}$.
4. Let $\alpha = \text{const}$, $\beta \to 0$ and $N >> I$. Then $GF_{\text{const},\beta\to 0} = G^2 - O(df \log N) = GF_{\text{MDLP}}$.
5. Let $\alpha = \text{const}$, $\beta \to 0$, and $G^2 = O(N \log J)$, $N/(IJ) \to \infty$. Then $GF_{\text{const},\beta\to 0} = G^2 - O(df \log N) = GF''_{G^2} \approx GF''_{X^2}$.

The parameterized goodness function not only allows us to represent the existing goodness functions in a closed uniform form, but, more importantly, it provides a new way to understand and handle discretization. First, the parameterized approach provides a flexible framework to access a large collection (potentially infinite) of goodness functions. Any valid pair of α and β corresponds to a potential goodness function. Note that this treatment is in the same spirit of regularization theory developed in the statistical machine learning field [16,19].

Secondly, finding the best discretization for different data mining tasks for a given dataset is transformed into a parameter selection problem. However, it is an open problem how we may automatically select the parameters without running the targeted data mining task. In other words, can we analytically determine the best discretization for different data mining tasks for a given dataset? This problem is beyond the scope of this paper and we plan to investigate it in future work.

Finally, the unification of goodness functions allows to develop efficient algorithms to discretize the continuous attributes with respect to different parameters in a uniform way. This is the topic of the next subsection.

5.4 Dynamic programming for discretization

This section presents a dynamic programming approach to find the best discretization function to maximize the parameterized goodness function. Note that the dynamic programming has been used in discretization before [14]. However, the existing approaches do not have a global goodness function to optimize, and almost all of them have to require the knowledge of targeted number of intervals. In other words, the user has to define the number of intervals for discretization. Thus, the existing approaches can not be directly applied to discretization for maximizing the parameterized goodness function.

In the following, we introduce our dynamic programming approach for discretization. To facilitate our discussion, we use GF for $GF_{\alpha,\beta}$, and we simplify the GF formula as follows. Since a given table C, $N \times H_\beta(S_1 \cup \cdots \cup S_I)$ (the first term in GF, Formula 17) is fixed,

we define

$$F(C) = N \times H_\beta(S_1 \cup \cdots \cup S_I) - GF(C)$$
$$= \sum_{i=1}^{I'} N_i \times H_\beta(S_i) + \alpha \times (I' - 1)(J - 1)\left[1 - \left(\frac{1}{N}\right)^\beta\right]/\beta$$

Clearly, the minimization of the new function F is equivalent to maximizing GF. In the following, we will focus on finding the best discretization to minimize F. First, we define a sub-contingency table of C as $C[i : i + k] = \{S_i, \ldots, S_{i+k}\}$, and let $C^0[i : i + k] = S_i \cup \cdots \cup S_{i+k}$ be the merged column sum for the sub-contingency table $C[i : i + k]$. Thus, the new function F of the row $C^0[i : i + k]$ is:

$$F(C^0[i : i + k]) = \left(\sum_{r=i}^{i+k} N_r\right) \times H_\beta(S_i \cup \cdots \cup S_{i+k})$$

Let C be the input contingency table for discretization. Let $Opt(i, i + k)$ be the minimum of the F function from the partial contingency table from row i to $i + k$, $k > 1$. The optimum which corresponds to the best discretization can be calculated recursively as follows:

$$Opt(i, i + k) = \min\Big(F(C^0[i : i + k]),$$
$$\min_{1 \le l \le k-1}\left(Opt(i, i + l) + Opt(i + l + 1, i + k) + \alpha \times (J - 1)\left[1 - \left(\frac{1}{N}\right)^\beta\right]/\beta\right)\Big)$$

where $k > 0$ and $Opt(i, i) = F(C^0[i : i])$. Given this, we can apply the dynamic programming to find the discretization with the minimum of the goodness function, which are described in Algorithm 1. The complexity of the algorithm is $O(I^3)$, where I is the number of intervals of the input contingency table C.

Algorithm 1 Discretization (Contingency table $C_{I \times J}$)

```
for i = 1 to I do
  for j = i downto 1 do
    Opt(j, i) = F(C^0[j : i])
    for k = j to i − 1 do
      Opt(j, i) = min(Opt(j, i), Opt(j, k)+
        Opt(k + 1, i) + α(J − 1)[1 − (1/N)^β]/β)
    end for
  end for
end for
return Opt(1, I)
```

6 Experimental results

The major goal of our experimental evaluation is to demonstrate that the dynamic programming approach with appropriate parameters can significantly reduce the classification errors compared with the existing discretization approaches.

We chose 12 datasets from the UCI machine learning repository [37]. Most of the datasets have been used in the previous experimental evaluation for discretization study [12,26].

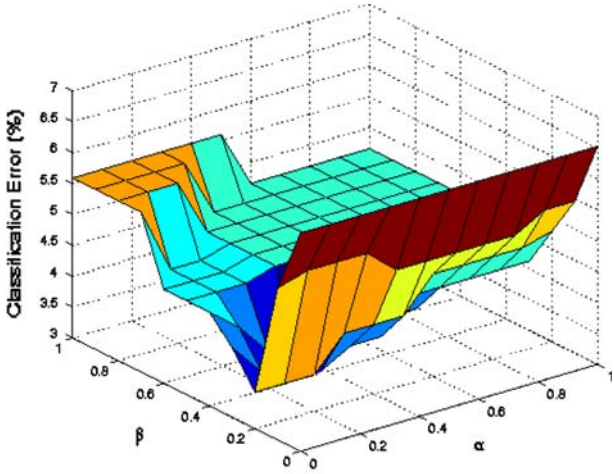
Table 2 Summary of dataset

Dataset	Instances	Continuous feature	Nominal feature
Anneal	898	6	32
Australian	690	6	8
Diabetes	768	8	0
Glass	214	9	0
Heart	270	13	0
Hepatitis	155	6	13
Hypothyroid	3,168	7	18
Iris	150	4	0
Labor	57	8	8
Liver	345	6	0
Sick-euthyroid	3,163	7	18
Vehicle	846	18	0

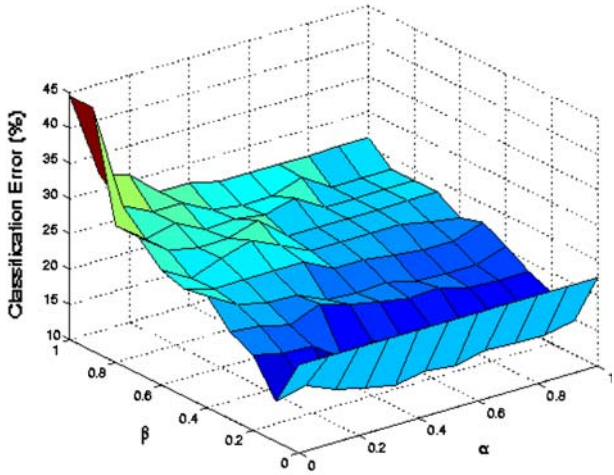
Table 2 describes the size and the number of continuous and nominal features of each dataset. We apply discretization as a preprocessing step for two well-known classification methods: the C4.5 decision tree and Naive Bayes classifier [17]. For comparison purpose, we apply four discretization methods: *equal-width* (EQW), *equal-frequency* (EQF), *Entropy* [15], and *ChiMerge* [23]. The first two are unsupervised approaches and the last two are supervised approaches. We set the number of discretization intervals to be 10 for the first two. All their implementations are from Weka 3 [38].

Our dynamic programming approach for discretization (referred to as *Unification* in the experimental results) depends on two parameters, α and β. How to analytically determine the best parameters which can result in the minimal classification error is still an open question and beyond the scope of this paper. Here, we apply an experimental-validation approach to choose the optimal parameters α and β. For a given dataset and the data mining task, we create a 10×10 uniform grid for $0 \leq \alpha \leq 1$ and $0 \leq \beta \leq 1$ In addition, we use a value 10^{-5} to replace 0 for β since it cannot be equal to 0. Then we apply the dynamic programming at each grid point to discretize the dataset. We score each point using the mean classification error based on a five-trial fivefold cross-validation on the discretized data. Figure 1a shows the surface of the classification error rate of $C4.5$ running on the discretized *iris* dataset [37] using the unification approach with parameters from the 10×10 grid points. Figure 1b illustrates the surface of the classification error rate of Naive Bayes classifier running on the discretized *glass* dataset [37]. Clearly, we can see that different α and β parameters can result in very different classification error rates. Given this, we choose the α and β pair which achieves the minimal classification error rate as the selected unification parameters for discretization. For instance, in these two figures, we choose $\alpha = 0.3$ and $\beta = 0.3$ as the parameters to discretize *iris* for $C4.5$, and choose $\alpha = 0.4$ and $\beta = 0.1$ to discretize *glass* for Naive Bayes classifier. Note that the objective of using five trials instead of only one is to choose parameters in a more robust fashion to avoid outliers.

Finally, for each of the discretization method (our *Unification* method with the best predicated parameter), we run a five-trial five-fold cross-validation, and report their mean and standard deviation of the cross-validation. Note that here each trial will re-shuffle the dataset and is different from the trials in the parameter selection process.



(a) Iris+C4.5



(b) glass+Naive Bayes classifier

Fig. 1 The surface of classification error rate using parameters from 10×10 grid

Tables 3 and 4 show the experimental results for C4.5 and Naive Bayes Classifier, respectively. In the left part of each table, we show the mean classification error and standard deviation using different discretization methods (the first one, *Continuous* corresponding to no-discretization). The right part of each table shows the percentage differences between two leading discretization approaches, *Entropy* and *ChiMerge*, with our new approach *Unification*. The last column chooses the minimal classification errors from all five existing approaches to compare with the unification approach.

We can see that the unification approach performs significantly better than the existing approaches. First, based on the average classification error for all the 12 datasets, the unification is the best among all these approaches (14.45% error rate for C4.5 and 10.60% for Naive Bayes classifier). For C4.5, it reduces the error rate on an average of 19.40% compared

Table 3 C4.5 results

Dataset	Experimental results 5 × 5 validation						Comparision with unification		
	Continuous	EQW	EQF	Entropy	ChiMerge	Unification	Entropy	ChiMerge	Min
Anneal	8.62 ± 2.16	9.84 ± 2.07	9.40 ± 1.95	8.58 ± 1.49	7.32 ± 1.19	7.33 ± 1.40	17.05	−0.14	−0.14
Australian	14.34 ± 2.55	15.10 ± 2.99	12.83 ± 3.51	13.70 ± 2.94	14.64 ± 3.11	12.46 ± 2.78	9.95	17.50	2.97
Diabetes	26.07 ± 3.07	25.84 ± 2.83	26.02 ± 3.11	22.75 ± 3.13	26.05 ± 3.02	22.34 ± 2.35	1.84	16.61	1.84
Glass	33.44 ± 6.33	44.29 ± 6.09	42.05 ± 4.78	26.72 ± 7.52	27.23 ± 5.39	25.15 ± 4.64	6.24	8.27	6.24
Heart	20.30 ± 5.18	21.42 ± 5.21	22.01 ± 4.62	16.52 ± 4.01	20.60 ± 5.30	16.52 ± 4.01	0.00	24.70	0.00
Hepatitis	18.60 ± 5.85	16.92 ± 4.96	15.88 ± 5.06	19.36 ± 5.38	17.81 ± 5.19	15.36 ± 5.03	26.04	15.95	3.39
Hypothyroid	0.78 ± 0.26	2.69 ± 0.54	1.73 ± 0.39	0.76 ± 0.30	1.69 ± 0.48	0.76 ± 0.30	0.00	122.37	0.00
Iris	5.60 ± 3.33	4.00 ± 2.98	6.01 ± 2.04	5.46 ± 3.28	3.34 ± 3.13	3.06 ± 3.43	78.43	9.15	9.15
Labor	15.77 ± 7.75	28.84 ± 7.06	28.05 ± 6.76	14.37 ± 10.84	9.09 ± 8.10	9.82 ± 9.33	46.33	−7.43	−7.43
Liver	34.54 ± 5.05	39.18 ± 5.44	42.77 ± 5.60	36.80 ± 5.23	34.32 ± 3.72	30.54 ± 4.93	20.50	12.38	12.38
Sick-euthyroid	2.09 ± 0.53	3.94 ± 0.62	4.96 ± 0.67	2.49 ± 0.64	4.11 ± 0.86	2.09 ± 0.52	19.14	96.65	0.00
Vehicle	27.64 ± 3.41	30.62 ± 2.50	34.51 ± 3.10	30.00 ± 1.94	29.03 ± 3.51	27.97 ± 3.53	7.26	3.79	−1.18
Average	17.32	20.22	20.52	16.46	16.27	14.45	19.40	26.65	2.27

Table 4 Naive Bayes results

Dataset	Experimental results 5 × 5 validation						Comparisons with unification		
	Continuous	EQW	EQF	Entropy	ChiMerge	Unification	Entropy	ChiMerge	Min
Anneal	20 ± 2.15	6.1 ± 1.55	3.07 ± 1.4	3.67 ± 1.45	2.36 ± 1.25	2.18 ± 1.15	68.35	8.26	8.26
Australian	22.55 ± 2.6	14.81 ± 3	13.48 ± 2.7	14.29 ± 3.1	10.52 ± 2.25	10.11 ± 1.95	41.28	4.00	4.00
Diabetes	24.45 ± 2.85	24.40 ± 3.6	25.36 ± 3.8	22.03 ± 2.75	15.26 ± 2.35	13.13 ± 2.6	67.82	16.26	16.26
Glass	53.74 ± 7	42.06 ± 5.6	28.59 ± 5.35	25.8 ± 4	20 ± 4.6	13.36 ± 4.45	93.04	49.64	49.64
Heart	16 ± 5.7	15.70 ± 3.3	16.96 ± 2.7	16.3 ± 4.55	12.885 ± 3.45	12.74 ± 3.05	27.94	1.14	1.14
Hepatitis	16.12 ± 5.6	16.12 ± 4.05	17.42 ± 5.7	14.32 ± 3.95	9.93 ± 3.35	10.32 ± 3.9	38.76	−3.73	−3.73
Hypothyroid	2.11 ± 0.45	2.99 ± 0.45	2.835 ± 0.75	1.36 ± 0.5	1.23 ± 0.5	0.99 ± 0.45	37.37	24.75	24.75
Iris	4.26 ± 4.15	5.2 ± 3.65	7.33 ± 4.55	5.73 ± 4.55	4.93 ± 4.15	2.8 ± 3.95	104.64	76.07	52.32
Labor	8.03 ± 6.25	8.48 ± 8.7	9.42 ± 9.65	6.36 ± 6.85	5.3 ± 5.35	2.88 ± 4.4	120.83	84.03	84.03
Liver	44.87 ± 7.6	36.23 ± 3.8	37.97 ± 5.2	36.81 ± 5.3	20.465 ± 4.25	20.23 ± 4.15	81.96	1.16	1.16
Sick-euthyroid	15.63 ± 2.15	6.375 ± 0.85	5.855 ± 1.05	3.8 ± 0.8	3.32 ± 0.65	3.26 ± 0.7	16.56	1.84	1.84
Vehicle	55.05 ± 3.1	39.52 ± 3.2	37.28 ± 2.6	37.49 ± 3.5	30.47 ± 2.95	35.27 ± 2.9	6.31	−13.60	−13.60
Average	23.57	18.17	17.13	15.66	11.39	10.61	58.74	20.82	18.84

with *Entropy*, and reduces the error rate on average of 26.65% compared with *ChiMerge*. For Naive Bayes classifier, it reduces the error rate on an average of 58.74% compared with *Entropy*, and reduces the error rate on an average of 20.82% compared with *ChiMerge*. The overall improvement is on an average of 31% in terms of classification error rate. Finally, in 9 out of 12 datasets for C4.5, the unification approach shows better or equal performance with the best existing approach. In other 3 datasets, the performance are fairly close to the minimal error rate as well. For Naive Bayes classifier, the unification method perform the best in 10 out of 12 datasets and the second for the other 2 datasets.

7 Conclusions

In this paper, we introduced a generalized goodness function to evaluate the quality of a discretization method. We have shown that seemingly disparate goodness functions based on entropy, AIC, BIC, Pearson's X^2, Wilks' G^2, and Gini index are all derivable from our generalized goodness function. Furthermore, the choice of different parameters for the generalized goodness function explains why there is a wide variety of discretization methods. Indeed, difficulties in comparing different discretization methods were widely known. Our results provide a theoretical foundation in understanding these difficulties and offer rationale as to why evaluation of different discretization methods for an arbitrary contingency table is difficult. We have designed a dynamic programming algorithm that for given set of parameters of a generalized goodness function provides an optimal discretization which achieves the minimum of the generalized goodness function. We have conducted an extensive performance tests for a set of publicly available data sets. Our experimental results demonstrate that our discretization method consistently outperforms the existing discretization metods on the average by 31%. These results clearly validate our approach and open a new way of tackling discretization problems.

References

1. Agresti A (1990) Categorical data analysis. Wiley, New York
2. Auer P, Holte R, Maass W (1995) Theory and applications of agnostic pac-learning with small decision trees. In: Machine learning: proceedings of the twelth international conference. Morgan Kaufmann
3. Bay SD (2001) Multivariate discretization for set mining. Knowl Inf Syst 3(4):491–512
4. Breiman L, Friedman J, Olshen R, Stone C (1998) Classification and regression trees. CRC Press
5. Boulle M (2004) Khiops: a statistical discretization method of continuous attributes. Mach Learn 55: 53–69
6. Boulle M (2006) MODL: a Bayes optimal discretization method for continuous attributes. Mach Learn 65(1):131–165
7. Casella G, Berger RL (2001) Statistical inference, 2nd edn. Duxbury Press
8. Catlett J (1991) On changing continuous attributes into ordered discrete attributes. In: Proceedings of European working session on learning, pp 164–178
9. Ching JY, Wong AKC, Chan KCC (1995) Class-dependent discretization for inductive learning from continuous and mixed-mode data. IEEE Trans Pattern Anal Mach Intell 17(7):641–651
10. Chmielewski MR, Grzymala-Busse JW (1996) Global discretization of continuous attributes as preprocessing for machine learning. Int J Approx Reason 15
11. Cover TM, Thomas JA (2006) Elements of information thoery, 2nd edn. Wiley, New York
12. Dougherty J, Kohavi R, Sahavi M (1995) Supervised and unsupervised discretization of continuous attributes. In: Proceedings of the 12th international conference on machine learning, pp 194–202
13. Elomaa T, Rousu J (2003) Necessary and sufficient pre-processing in numerical range discretization. Knowl Inf Syst 5(2):162–182

14. Elomaa T, Rousu J (2004) Efficient multisplitting revisited: optima-preserving elimination of partition candidates. Data Mining Knowl Discovery 8:97–126
15. Fayyad UM, Irani KB (1993) Multi-interval discretization of continuous-valued attributes for classification learning. In: Proceedings of the 13th joint conference on artificial intelligence, pp 1022–1029
16. Girosi F, Jones M, Poggio T (1995) Regularization theory and neural networks architectures. Neural Comput 7(2):219–269
17. Hand D, Mannila H, Smyth P (2001) Principles of data mining. MIT Press
18. Hansen MH, Yu B (2001) Model selection and the principle of minimum description length. J Am Statist Assci 96:454
19. Hastie T, Tibshirani R, Friedman J (2001) The elements of statistical learning. Springer, Heidelberg
20. Holte RC (1993) Very simple calssification rules perform well on most commonly used datasets. Mach Learn 11:63–90
21. Johnson N, Kotz S, Balakrishnan N (1994) Continuous univariate distributions, 2nd edn. Wiley, New York
22. Jin R, Breitbart Y (2007) Data discretization unification. Technical Report, Department of Computer Science, Kent State University. http://www.cs.kent.edu/research/techrpts.html
23. Kerber R (1992) ChiMerge: discretization of numeric attributes. In: National conference on artificial intelligence
24. Kurgan LA, Cios KJ (2004) CAIM discretization algorithm. IEEE Trans Knowl Data Eng 16(2):145–153
25. Kohavi R, Sahami M (1996) Error-based and entropy-based discretization of continuous features. In: Proceedings of the second international conference on knowledge discovery and data mining. Menlo Park. AAAI Press, pp 114–119
26. Liu H, Hussain F, Tan CL, Dash M (2002) Discretization: an enabling technique. Data Mining Knowl Discovery 6:393–423
27. Liu H, Setiono R (1995) Chi2: feature selection and discretization of numeric attributes. In: Proceedings of 7th IEEE int'l conference on tools with artificial intelligence
28. Liu X, Wang H (2005) A discretization algorithm based on a heterogeneity criterion. IEEE Trans Knowl Data Eng 17(9):1166–1173
29. Mussard S, Seyte F, Terraza M (2003) Decomposition of Gini and the generalized entropy inequality measures. Econ Bull 4(7):1–6
30. Pfahringer B (1995) Supervised and unsupervised discretization of continuous features. In: Proceedings of 12th international conference on machine learning, pp 456–463
31. Rissanen J (1978) Modeling by shortest data description. Automatica 14:465–471
32. Simovici DA, Jaroszewicz S (2002) An axiomatization of partition entropy. IEEE Trans Inf Theory 48(7):2138–2142
33. Wallace DL (1959) Bounds on normal approximations to Student's and the Chi-square distributions. Ann Mathe Stat 30(4):1121–1130
34. Wallace DL (1960) Correction to "Bounds on Normal Approximations to Student's and the Chi-Square Distributions". Ann Math Statist 31(3):810
35. Wong AKC, Chiu DKY (1987) Synthesizing statistical knowledge from incomplete mixed-mode data. IEEE Trans Pattern Anal Mach Intell 9(6):796–805
36. Yang Y, Webb GI (2003) Weighted proportional k-interval discretization for naive–Bayes classifiers. In: Advances in knowledge discovery and data mining: 7th Pacific-Asia Conference, PAKDD, pp 501–512
37. UCI Machine Learning Repository (2007) http://www.ics.uci.edu/mlearn/ML.Repository.html
38. Weka 3 (2007) Data mining software in Java. http://www.cs.waikato.ac.nz/ml/weka

Author's biography

Ruoming Jin is an Assistant Professor in the Department of Computer Science at Kent State University, Ohio. He received the BE and ME degrees in the computer engineering from Beijing University of Aeronautics and Astronautics, China in 1996, 1999, respectively. He received the MS degree in computer science from the University of Delaware in 2001 and the Ph.D. degree in computer science from the Ohio State University in 2005. His research interest includes system support and algorithm design for scalable data mining, data stream processing, massive graph mining, databases and bioinformatics. He has published over 40 research papers in these areas.

Yuri Breitbart is an Ohio Board of Regents Distinguished Professor of Computer Science in the Department of Computer Science at Kent State University. Prior to that he was a Member (and later a Distinguished Member) of Technical Staff at the Bell Laboratories at Murray Hill, New Jersey. From 1986 to 1996 he was a Professor in the Department of Computer Science at University of Kentucky and was the Department Chair there for the first seven years. Prior to 1986, he was leading the Database Research group first at ITT Research Center and then at Amoco Production Company Research Center. Dr. Breitbart has held visiting positions as Guest Professor at Swiss Technological Institute Institute(ETH) in Zurich, as a Lady Davis Visiting Professor at Israel Technological Institute (Technion), and as a Member of Technical Staff at HP Research Center in Palo Alto. He has been consulting for numerous companies and among them IBM, Boeing, Amoco, HP, and Bell Labs. His research is in the areas of distributed information system, network management systems and data mining. He received DSc degree in Computer Science from Israel Institute of Technology (Technion). He is a Fellow of the ACM and member of IEEE Computer Society and SIGMOD. He has served on numerous program committees and NSF panels.

Chibuike Muoh is a Research Assistant at the Knowledge Discovery and Database (KDDB) lab in the Computer Science department of Kent State University. His current research interest include Data mining, social networks, bioinformatics and P2P systems. Chibuike Muoh recieved his B.S. degree cum laude in Computer Science with a minor in Business from Kent State University where he is also currently a Masters of Science candidate. His hobbies include soccer and reading.