Data Discretization Unification

Ruoming Jin Yuri Breitbart
Department of Computer Science
Kent State University, Kent, OH 44241
{jin,yuri}@cs.kent.edu

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 of merged data are asymptotically equivalent. Furthermore, we define a notion of generalized entropy and prove that discretization methods based on MDLP, Gini Index, AIC, BIC, and Pearson's X^2 and G^2 statistics are all derivable from the generalized entropy function. Finally, we design a dynamic programming algorithm that guarantees the best discretization based on the generalized entropy notion.

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 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 ten fold performance improvement for domains with a large number of continuous attributes with little or no loss of accuracy.

In this paper, we propose to treat the discretization of a single continuous attribute as a 1-*dimensional classification* problem. *Good* discretization may lead to new and more accurate knowledge. On the other hand, *bad* discretization leads to unnecessary loss of information or in some cases to false information with disastrous consequences. Any discretization process generally leads to a loss of information. The goal of the *good* discretization algorithm is to minimize such information loss. If discretization leads to an unreasonably small number of data intervals, then it may result in significant information loss. If a discretization method generates too many data intervals, it may lead to false information.

Discretization of continuous attributes has been extensively studied [5, 8, 9, 10, 13, 15, 24, 25]. There are a wide variety of discretization methods starting with the naive methods (often referred to as *unsupervised* methods) such as equal-width and equal-frequency [26], to much more sophisticated methods (often referred to as *supervised* methods) such as MDLP [15] and Pearson's X^2 or Wilks' G^2 statistics based discretization algorithms [18, 5]. 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.

Both unsupervised and supervised discretization methods can be further subdivided into *top-down* or *bottom-up* methods. A *top-down* method starts with a single interval that includes all data attribute values and then generates a set of intervals by splitting the initial interval into two or more intervals. A *bottom-up* method initially considers each data point as a separate interval. It then selects one or more adjacent data points merging them into a new interval. For instance, the methods based on statistical independence tests, such as Pearson's X^2 statistics [23, 27, 5], are examples of bottom-up methods. On the other hand, the method based on information theoretical measures, such as entropy and MDLP [25], is an example of the *top-down* method. Liu *et. al*. [26] introduce a nice categorization of a large number of existing discretization methods.

Regardless of the discretization method, a compromise must be found between the information quality of resulting intervals and their statistical quality. The former is generally achieved by considering a notion of entropy and a method's ability to find a set of intervals with a minimum of information loss. The latter, however is achieved by resorting to a specific statistic to evaluate the level of independence of merged data.

In spite of the wealth of literature on discretization, there are very few attempts to *analytically* compare different discretization methods. Typically, researchers compare the performance of different algorithms by providing experimental results of running these algorithms on publicly available data sets.

In [13], Dougherty *et al.* compare discretization results obtained by unsupervised discretization versus a supervised method proposed by [19] and the entropy based method proposed by [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 [15] is comparatively smaller than the number of errors generated by the discretization algorithm of [3]. 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 [38, 24, 5, 6]. The goal of the CAIM algorithm [24] is to find the minimum number of intervals that minimize the loss between class-attribute interdependency. The authors of [24] report that their algorithm generates fewer classification errors than two naive unsupervised methods (equal-width and equal-frequency) as well as four supervised methods (max entropy, Patterson-Niblett, IEM, and CADD).

Boulle [5] has proposed a new discretization method called $Khiops$. The method uses Pearson's X^2 statistic. It merges two intervals that maximize the value of X^2 and the two intervals are replaced with the result of their merge. He then shows that *Khiops* is as accurate as other methods based on X^2 statistics but performs much faster.

MODL is another latest discretization method proposed by Boulle [5]. 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. The experimental results show *MODL* can produce fewer intervals than other discretization methods with better or comparable accuracy.

Finally, Yang and Webb have studied discretization for naive-Bayes classifiers [38]. They have proposed a couple of methods, such as *proportional k-interval discretization* and *equal size discretization*, to manage the discretization *bias* and *variance*. They report their discretization can achieve lower classification error for the naive-Bayes classifiers than other alternative discretization methods.

To summarize, a comparison of different discretization methods that appeared so far have been done by running discretization methods on publicly available data sets and comparing certain performance criteria, such as the number of data intervals produced by a discretization method and the number of classification errors. 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? Can we analytically evaluate and compare these different discretization algorithms without resorting to experiments on different data sets? In other words, is there an objective function which can measure the goodness of different approaches? If so, how would this function look like? If such a function exists, what is the relationship between it and the existing discretization criteria? In this paper we provide a list of positive results toward answering these questions.

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 j-th class label falling in the i-th 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.

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
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$S_{I'}$	$c_{I'1}$	$c_{I'2}$	$\cdots$	$c_{I'J}$	$N_{I'}$
Column Sum	M_1	M_2	$\cdots$	M_J	N (Total)

Table 1: Notations for Contingency Table $C'_{I'\times J}$

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.

Thus, we define the discretization as a function g mapping each row of the new table to a set of rows in the original table, such that $g : \{1, 2, \cdots, I'\} \rightarrow 2^{\{1,2,\cdots I\}}$ with the following properties:

1. $\forall i, 1 \leq i \leq I', g(i) \neq \emptyset$;
2. $\forall i, 1 \leq i < I', g(i) = \{x, x+1, \cdots x+k\}$
3. $\cup_{i=1}^{I'} g(i) = \{1, 2, \cdots, I\}$.

In terms of the cell counts, for the i-th row in the new table, $c'_{i,j} = \sum_{y=x}^{x+k} c_{y,j}$, assuming $g(i) = \{x, x+1, \cdots, x+k\}$. Potentially, the number of valid g functions (the ways to do the discretization) is 2^{I-1}.

The discretization problem is defined then as selecting the $best$ discretization function. The quality of the discretization function is measured by a $goodness$ function we propose here that 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 function 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 [15, 16, 5, 23, 27, 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 $O(logN)$, 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. Finally, we design a dynamic programming algorithm that guarantees the best discretization based on a generalized goodness function.

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 (Subsection 2.1), the fitness of statistical models (Subsection 2.2), and the confidence level for statistical independence tests (Subsection 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 Minimal Description Length Principle (MDLP) is a commonly used approach for choosing the best classification model [31, 18]. 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 ($Cost_{MDLP}$) is derived as follows:

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

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

In the following, we 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. *[12] 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:

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

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

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

LEMMA 1. *There always exists information loss for the merged intervals: $H(S_1, S_2) \leq H(S_1 \cup S_2)$*

Proof: This can be easily proven by the concaveness of the entropy function. □

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 $Cost_{MDLP}$ becomes smaller.

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}.

$$\begin{aligned} &GF_{MDLP}(C) = Cost_{MDLP}(C^0) - Cost_{MDLP}(C) \\ &= N\times H(S_1\cup\cdots\cup S_{I'}) - N\times H(S_1,\cdots,S_{I'}) - \\ &\qquad ((I'-1)log\frac{N}{I'-1} + (I'-1)(J'-1)logJ) \end{aligned} \quad (2)$$

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 Selection (AIC and BIC):

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 [16]. 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:

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

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

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

According to [16], $S_L(\vec{\pi})$ 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 [16] can be used and it is defined as follows:

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

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 [16] is defined as follows:

$$Cost_{BIC} = 2S_L(D|\vec{\pi}) + (I' \times (J-1))logN \quad (4)$$

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

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.

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

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

2.3 Confidence Level from Independence Tests

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

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

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

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

We note that $1 - F_{\chi^2_{df}}(X^2)$ is essentially the P-value of the aforementioned statistical independence test [7]. The lower the P-value (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:

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

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

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

$$G^2 = -2log\Lambda = 2\sum_{i=1}^{I'}\sum_{j=1}^{j} c_{ij} log\frac{c_{ij}}{N_iM_j/N} \quad (8)$$

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

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

Indeed, this statistic has been applied in discretization (though not for the global goodness function), and is referred to as class-attributes interdependency information [37].

2.4 Properties of Proposed Goodness Functions

An important theoretical question we address is how these methods 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 now describe four basic principles we believe any goodness function for discretization must satisfy.

1. **Merging Principle (P1):** Let $S_i =< c_{i1}, \cdots, c_{iJ} >$, and $S_{i+1} =< c_{i+1,1}, \cdots, c_{i+1,J} >$ be two adjacent rows in the contingency table C. If $c_{ij}/N_i = c_{(i+1)j}/N_{i+1}, \forall j$, $1 \leq j \leq J$, 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 [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 Section 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 $C =< C_1, \ldots C_J >$ 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, $S_i =< c_{i1}, \cdots, c_{iJ} >$, $c_{ij}/N_i = M_j/N$, 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, $S_i =< c_{i1}, \cdots, c_{iJ} >$, 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: In Appendix. □

3. EQUIVALENCE OF GOODNESS FUNCTIONS

In this section, we analytically compare different discretization goodness functions introduced in Section 2. In particular, we find some rather surprising connection between these seemingly quite different approaches: the information theoretical complexity (Subsection 2.1), the statistical fitness (Subsection 2.2), and the statistical independence tests (Subsection 2.3). 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 $O(logN)$. 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 section 2 (Subsection 3.1). This is achieved by expressing G^2 statistics through information entropy (Theorem 3). Then, using a Wallace's result [35, 36] 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 (Subsection 3.3). Finally, we compare the methods based on the global independence tests, such as Khiops [5] (GF_{X^2}) and those based on local independence tests, such as ChiMerge [23] and Chi2 [27] (Subsection 3.4).

3.1 Unifying the Cost of Data ($cost(data|model)$) 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 $S_L(D|\vec{\pi})$ (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 $S_L(D|\vec{\pi})$, i.e.*

$$N \times H(S_1, \cdots, S_{I'}) = -logL(\vec{\pi})$$

Proof:

$$\begin{aligned} N \times H(S_1, \cdots, S_{I'}) &= -\sum_{i=1}^{I'} N_i \times N(S_i) \\ &= -\sum_{i=1}^{I'} N_i \times \sum_{j=1}^{J} \frac{c_{ij}}{N_i} log \frac{c_{ij}}{N_i} \\ &= -\sum_{i=1}^{I'} \sum_{j=1}^{J} c_{ij} log \frac{c_{ij}}{N_i} = -logL(\vec{\pi}) \end{aligned}$$

□

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*

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

Proof:

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

□

Theorem 3 can be generalized as follows.

THEOREM 4. *Assuming we have k consecutive rows, S_i,S_{i+1}, $\cdots$, S_{i+k-1}. Let G_k^2 be the likelihood independence test statistic for the k rows. Then, we have $G^2_{(i,i+k-1)}/2 =$*

$$N_{(i,i+k-1)}(H(S_i \cup \cdots \cup S_{i+k-1}) - H(S_i, \cdots, S_{i+k-1}))$$

Proof: Omit for simplicity. □

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

$$GF_{MDLP} =$$
$$G^2 - 2(I'-1)log\frac{N}{I'-1} - 2(I'-1)(J'-1)logJ \quad (11)$$
$$GF_{AIC} = G^2 - (I'-1)(J-1) \quad (12)$$
$$GF_{BIC} = G^2 - (I'-1)(J-1)logN/2 \quad (13)$$

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 5. *[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 Theorem 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 4 and 5 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. Subsection 3.2 presents our approach on tackling this issue.

3.2 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, $GF_1(C_1) \leq GF_1(C_2) \Longrightarrow GF_2(C_1) \leq GF_2(C_2)$ and $GF_2(C_1) \leq GF_2(C_2) \Longrightarrow GF_1(C_1) \leq GF_1(C_2)$.*

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 based on a statistical confidence and those that are based on MDLP, AIC, and BIC.

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 deviation corresponding to the chi-square distributed variable t. That is, the following equality holds:*

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

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

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

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 6. *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 [35, 36] to derive an asymptotically accurate closed form expression for a simple variant of GF'_{G^2}.

THEOREM 7. *[35, 36] For all $t > df$, all $df > .37$, and with $w(t) = [t - df - df\, log(t/df)]^{\frac{1}{2}}$,*

$$0 < w(t) \leq u(t) \leq w(t) + .60df^{-\frac{1}{2}}$$

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 7. This condition implies that $u(G^2) \geq 0$. [1] We show that under some conditions, $u^2(G^2)$ can be approximated as $w^2(G^2)$.

$$w^2(G^2) \leq u^2(G^2) \leq w^2(G^2) + \frac{0.36}{df} + 1.2\frac{w(G^2)}{\sqrt{df}}$$
$$1 \leq \frac{u^2(G^2)}{w^2(G^2)} \leq 1 + \frac{0.36}{w^2(G^2) \times df} + \frac{1.2}{w(G^2)\sqrt{df}}$$
$$If\ df \to \infty\ and\ w(t) >> 0,\ then$$
$$\frac{0.36}{w^2(G^2) \times df} \to 0\ and\ \frac{1.2}{w(G^2)\sqrt{df}} \to 0$$
$$Therefore,\ \frac{u^2(G^2)}{w^2(G^2)} \to 1$$

Thus, we can have the following goodness function:

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

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 16 indicate that all goodness functions introduced in section 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.

3.3 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 2 gives the upper bound.

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

Proof:

$$G^2 = 2N \times (H(S_1 \cup \cdots \cup S_I) - H(S_1, \cdots, S_I))$$
$$\leq 2N \times (-J \times (1/J \times log(1/J)) - 0)$$
$$\leq 2N \times log J$$

□

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

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

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 8. *Given an initial contingency table C with $log N \geq 2$ [2], 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 [16].

[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.

[2] The condition for the penalty of BIC is higher than AIC

3.4 Global and Local Independence Tests

This subsection compares the discretization methods based on global statistical independence tests versus local statistical independence tests. Note that the latter one does not have a global goodness function for the discretization. Instead, they treat each local statistical tests as an indication for merge action. The 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 invervals, they merge them into one row. Given such constraints, they usually try to find the best discretization with the minimal number of intervals. A natural question to ask is how such local independence test relates to global independence tests, as well as to the goodness functions GF_{X^2} and GF_{G^2}.

Formally, let $X^2_{(i,i+k)}$ be the Pearson's chi-square statistic for the $k+1$ consecutive rows (from i to $i+k$). Let $F_{\chi^2_{k\times(J-1)}}(X^2_{(i,i+k)})$ be the confidence level for these rows and their corresponding columns being statistically independent. If

$$F_{\chi^2_{k\times(J-1)}}(X^2_{(i,i+k)}) < \delta,$$

we can merge these $k+1$ rows into one row.

We only summarize our main results here. The detailed discussion is in Appendix. A typical local hypothesis test can be rewritten as follows:

$$G^2_{i,i+k-1} < df = (k \times (J-1))$$

In other words, as long as the above condition holds, we can merge these consecutive rows into one.

This suggests that the local condition essentially shares the penalty of the same order of magnitude as GF_{AIC}. In addition, we note that the penalty of $O(df log N)$ allows us to combine consecutive rows even if they are likely to be *statistically dependent* based on G^2 or X^2 statistic. In other words, the penalty of $O(df)$ in the goodness function is a stricter condition for merging consecutive rows than $O(df log N)$. Therefore, it would result in more intervals in the best discretization identified by the goodness using penalty of $O(df)$ than those identified by the goodness using penalty of $O(df log N)$. This essentially provides an intuitive argument for Theorem 8.

4. 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 section 2 are all can be derived from a generalized notion of entropy [29].

4.1 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]:

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

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 Appendix). 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:

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

4.2 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. *[32, 29] For a given interval S_i, the generalized entropy is defined as*

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

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

$$H_1(S_i) = \sum_{j=1}^{J}\frac{c_{ij}}{N_i}[1 - \frac{c_{ij}}{N_i}]/\beta = gini(S_i)$$

When $\beta \to 0$,

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

LEMMA 3. *$H_\beta[p_1, \cdots, p_J] = \sum_{j=1}^{J} p_i(1 - p_i^\beta)/\beta$ 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,

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

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

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

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

$$\sum_{i=1}^{I}\frac{N_i}{N}\times\sum_{j=1}^{J}\frac{c_{ij}}{N_i}[1-(\frac{c_{ij}}{N_i})^\beta]/\beta$$

Similarly, we have

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

THEOREM 9. *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. □

4.3 Parameterized Goodness Function

Based on the discussion in Section 3, 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 $O(dflogN)$.

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* $0<\beta\le 1$ *and* $0<\alpha$*, the parameterized goodness function for contingency table* C *is represented as*

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

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

THEOREM 10. *The parameter goodness function* $GF_{\alpha,\beta}$*, with* $\alpha>0$ *and* $0<\beta\le 1$*, satisfies all four principles,* **P1**, **P2**, **P3**, *and* **P4**.

Proof: In Appendix. □

By adjusting different parameter values, we show how goodness functions defined in section 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 $GF_{2(N-1)/(N(J-1)),1}=GF_{gini}$.
2. Let $\alpha=1/logN$ and $\beta\to 0$. Then $GF_{1/logN\beta\to 0}=GF_{AIC}$.
3. Let $\alpha=1/2$ and $\beta\to 0$. Then $GF_{1/2,\beta\to 0}=GF_{BIC}$.
4. Let $\alpha=const,\beta\to 0$ and $N>>I$. Then $GF_{const,\beta\to 0}=G^2-O(dflogN)=GF_{MDLP}$.
5. Let $\alpha=const,\beta\to 0$, and $G^2=O(NlogJ), N/(IJ)\to\infty$. Then $GF_{const,\beta\to 0}=G^2-O(dflogN)=GF''_{G^2}\approx GF''_{X^2}$.

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. Suppose we have a two dimension space where α is represented in the X-axis and β is represented in the Y-axis. Then, each point in the two-dimensional space for $\alpha>0$ and $0<\beta\le 1$ corresponds to a potential goodness function. The existing goodness functions corresponds to certain points in the two-dimensional space. These points are specified by the aforementioned parameter choices. Note that this treatment is in the same spirit of regularization theory developed in the statistical machine learning field [17, 34]. Secondly, finding the best discretization for different data mining tasks for a given dataset is transformed into a parameter selection problem. Ultimately, we would like to identify the parameter selection which optimizes the targeted data mining task. For instance, suppose we are discretizing a given dataset for a Naive Bayesian classifier. Clearly, a typical goal of the discretization is to build a Bayesian classifier with the minimal classification error. As described in regularization theory [17, 34], the methods based on cross-validation, can be applied here. 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 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.

4.4 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

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

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 $C[i:i+k]=\{S_i,\cdots,S_{i+k}\}$, 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:

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

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:

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

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)
```

5. 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 and Wilks' G^2 statistic as well as 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 to approach these difficulties and offer rationale as to why evaluation of different discretization methods for an arbitrary contingency table is difficult. Our generalized goodness function gives an affirmative answer to the question: is there an objective function to evaluate different discretization methods? Another contribution of this paper is to identify a dynamic programming algorithm that provides an optimal discretization which achieves the minimum of the generalized goodness function.

There are, however several questions that remain open. First of all, even if an objective goodness function exists, different parameter choices will result in different discretizations. Therefore, the question is for a particular set of applications, what are the best parameters for the discretization? Further, can we classify user-applications into different categories and identify the optimal parameters for each category? For example, considering medical applications; what is the best discretization function for them? Clearly for these applications, misclassification can be very costly. But the number of intervals generated by the discretization may not be that important. Pursuing these questions, we plan to conduct experimental studies to compare different goodness functions, and evaluate the effect of parameter selection for the generalized goodness function on discretization.

6. REFERENCES

[1] A. Agresti *Categorical Data Analysis*. Wiley, New York, 1990.

[2] H. Akaike. Information Theory and an Extension of the Maximum Likelihood Principle. In *Second International Symposium on Information Theory, 267-281*, Armenia, 1973.

[3] P. Auer, R. Holte, W. Maass. Theory and Applications of Agnostic Pac-Learning with Small Decision Trees. In *Machine Learning: Proceedings of the Twelth International Conference*, Morgan Kaufmann, 1995.

[4] L. Breiman, J. Friedman, R. Olshen, C. Stone *Classification and Regression Trees*. CRC Press, 1998.

[5] M. Boulle. Khiops: A Statistical Discretization Method of Continuous Attributes. *Machine Learning, 55, 53-69*, 2004.

[6] M. Boulle. MODL: A Bayes optimal discretization method for continuous attributes. *Mach. Learn. 65, 1 (Oct. 2006)*, 131-165.

[7] George Casella and Roger L. Berger. Statistical Inference (2nd Edition). *Duxbury Press*, 2001.

[8] J. Catlett. On Changing Continuous Attributes into Ordered Discrete Attributes. In *Proceedings of European Working Session on Learning, p. 164-178*, 1991.

[9] J. Y. Ching, A.K.C. Wong, K. C.C. Chan. Class-Dependent Discretization for Inductive Learning from Continuous and Mixed-Mode Data. *IEEE Transactions on Pattern Analysis and Machine Intelligence, V. 17, No. 7, 641-651*, 1995.

[10] M.R. Chmielewski, J.W. Grzymala-Busse. Global Discretization of Continuous Attributes as Preprocessing for Machine Learning. *International Journal of Approximate Reasoning, 15*, 1996.

[11] Y.S. Choi, B.R. Moon, S.Y. Seo. Genetic Fuzzy Discretization with Adaptive Intervals for Classification Problems. *Proceedings of 2005 Conference on Genetic and Evolutionary Computation, pp. 2037-2043*, 2005.

[12] Thomas M. Cover and Joy A. Thomas, Elements of Information Thoery, Second Edition. *Published by John Wiley & Sons, Inc.*, 2006.

[13] J. Dougherty, R. Kohavi, M. Sahavi. Supervised and Unsupervised Discretization of Continuous Attributes. *Proceedings of the 12th International Conference on Machine Learning, pp. 194-202*, 1995.

[14] Tapio Elomaa and Juho Rousu. Efficient Multisplitting Revisited: Optima-Preserving Elimination of Partition Candidates. *Data Mining and Knowledge Discovery, 8, 97-126*, 2004.

[15] U.M. Fayyad and K.B. Irani Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning. In *Proceedings of the 13th Joint Conference on Artificial Intelligence, 1022-1029*, 1993.

[16] David Hand, Heikki Mannila, Padhraic Smyth. *Principles of Data Mining* MIT Press, 2001.

[17] Federico Girosi, Michael Jones, and Tomaso Poggio. Regularization theory and neural networks architectures. In *Neural Computation, Volume 7 , Issue 2 (March 1995), Pages: 219 - 269.*

[18] M.H. Hansen, B. Yu. Model Selection and the Principle of Minimum Description Length. *Journal of the American Statistical Assciation, 96, p. 454*, 2001.

[19] R.C. Holte. Very Simple Calssification Rules Perform Well on Most Commonly Used Datasets. *Machine Learning, 11, pp. 63-90*, 1993.

[20] Janssens, D., Brijs, T., Vanhoof, K., and Wets, G. Evaluating the performance of cost-based discretization versus entropy-and error-based discretization. *Comput. Oper. Res. 33, 11 (Nov. 2006), 3107-3123.*

[21] N. Johnson, S. Kotz, N. Balakrishnan. Continuous Univariate Distributions, Second Edition. *John Wiley & Sons, INC.*, 1994.

[22] Ruoming Jin and Yuri Breitbart, Data Discretization Unification. *Technical Report (http://www.cs.kent.edu/research/techrpts.html), Department of Computer Science, Kent State Univeristy*, 2007.

[23] Randy Kerber. ChiMerge: Discretization of Numeric Attributes. *National Conference on Artificial Intelligence*, 1992.

[24] L.A. Kurgan, K.J. Cios CAIM Discretization Algorithm. *IEEE Transactions on Knowledge and Data Engineering, V. 16, No. 2, 145-153*, 2004.

[25] R. Kohavi,M. Sahami. Error-Based and Entropy-Based Discretization of Continuous Features. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, pp. 114-119*, Menlo Park CA, AAAI Press, 1996.

[26] Huan Liu, Farhad Hussain, Chew Lim Tan, Manoranjan Dash. Discretization: An Enabling Technique. *Data Mining and Knowledge Discovery, 6, 393-423*, 2002.

[27] H. Liu and R. Setiono. Chi2: Feature selection and discretization of numeric attributes. *Proceedings of 7th IEEE Int'l Conference on Tools with Artificial Intelligence*, 1995.

[28] X. Liu, H.Wang A Discretization Algorithm Based on a Heterogeneity Criterion. *IEEE Transaction on Knowledge and Data Engineering, v. 17, No. 9, 1166-1173*, 2005.

[29] S. Mussard, F. Seyte, M. Terraza. Decomposition of Gini and the generalized entropy inequality measures. *Economic Bulletin, Vol. 4, No. 7, 1-6*, 2003.

[30] B. Pfahringer. Supervised and Unsupervised Discretization of Continuous Features. *Proceedings of 12th International Conference on Machine Learning, pp. 456-463*, 1995.2003.

[31] J. Rissanen Modeling by shortest data description *Automatica, 14,pp. 465-471*, 1978.

[32] D.A. Simovici and S. Jaroszewicz An axiomatization of partition entropy *IEEE Transactions on Information Theory, Vol. 48, Issue:7, 2138-2142*, 2002.

[33] Robert A. Stine. Model Selection using Information Theory and the MDL Principle. In *Sociological Methods & Research, Vol. 33, No. 2, 230-260*, 2004.

[34] Trevor Hastie, Robert Tibshirani and Jerome Friedman. The Elements of Statistical Learning *Springer-Verlag*, 2001.

[35] David L. Wallace . Bounds on Normal Approximations to Student's and the Chi-Square Distributions. *The Annals of Mathematical Statistics, Vol. 30, No. 4, pp 1121-1130*, 1959.

[36] David L. Wallace . Correction to "Bounds on Normal Approximations to Student's and the Chi-Square Distributions". *The Annals of Mathematical Statistics, Vol. 31, No. 3, p. 810*, 1960.

[37] A.K.C. Wong, D.K.Y. Chiu. Synthesizing Statistical Knowledge from Incomplete Mixed-Mode Data. *IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 9, NNo. 6, pp. 796-805*, 1987.

[38] Ying Yang and Geoffrey I. Webb. Weighted Proportional k-Interval Discretization for Naive-Bayes Classifiers. In *Advances in Knowledge Discovery and Data Mining: 7th Pacific-Asia Conference, PAKDD, page 501-512*, 2003.

Appendix

Derivation of the Goodness Function based on 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.

$$cost_1(data|model) = N \times H(S_1, \cdots, S_{I'})$$
$$= N_1 \times H(S_1) + N_2 \times H(S_2) + \cdots + N_{I'} \times H(S_{I'})$$

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

$$L_1(I', N) = log_2(\tbinom{N+I'-1}{I'-1})$$
$$\approx (N + I' - 1)H(\frac{N}{N+I'-1}, \frac{I'-1}{N+I'-1})$$
$$= -(Nlog_2\frac{N}{N+I'-1} + (I'-1)log_2\frac{I'-1}{N+I'-1})$$
$$(log_2\frac{N}{N+I'-1} \to 0, N \to 0)$$
$$\approx (I'-1)log_2\frac{N+I'-1}{I'-1}$$
$$\approx (I'-1)log_2\frac{N}{I'-1}$$

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

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

Proof of Theorem 1

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, $S_i =< c_{i1}, \cdots, c_{iJ} >$, and $S_{i+1} =< c_{i+1,1}, \cdots, c_{i+1,J} >$, where $c_{ij} = c_{i+1,j}, \forall i, 1 \le j \le J$. Let C' be the resulting contingency table after we merge these two rows. Then we have

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

In addition, we have

$$(I-1)log_2\frac{N}{I-1} + (I-1) \times J \times log_2 J$$
$$-((I-2)log_2\frac{N}{I-2} + (I-2) \times J \times log_2 J)$$
$$= (I-1)log_2 N - (I-1)log_2(I-1) + (I-1) \times J \times log_2 J$$
$$-((I-2)log_2 NC(I-2)log_2(I-2) + (I-2) \times J \times log_2 J)$$
$$> log_2 NClog_2(I-1) + J \times log_2 J \quad (N \le I) > 0$$

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

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 $N \times H(S_1, \cdots, S_I)$.

$$N \times H(S_1, \cdots, S_I) \le N \times H(S_1 \cup \cdots \cup S_I)$$
$$N \times H(S_1, \cdots, S_I) = \sum_{k=1}^{I} N_k \times H(S_k)$$
$$= \sum_{k=1}^{I} N_k \times H(S_1 \cup \cdots \cup S_I)$$
$$= N \times H(S_1 \cup \cdots \cup S_I)$$

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 $N \times H(S_1, \cdots, S_J)$.

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

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, $S_i =< c_{i1}, \cdots, c_{iJ} >$, and $S_{i+1} =< c_{i+1,1}, \cdots, c_{i+1,J} >$, where $c_{ij} = c_{i+1,j}, \forall i, 1 \le j \le J$. Let C' be the resulting contin-

gency table after we merge these two rows. Then we have

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

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_{MDLP}: This can be directly derived from the symmetric property of X^2.

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 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_{MDLP}: 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.

$$X^2 = \sum\sum \frac{(c_{kj} - N_k \times M_j/N)^2}{N_k \times M_j/N}$$
$$= \sum\sum \frac{c^2_{kj} + (N_k \times M_j/N)^2 - 2 \times c_{kj} \times N_k \times M_j/N}{N_k \times M_j/N}$$
$$= \sum\sum (\frac{c^2_{kj}}{N_k \times M_j/N} + N_k \times M_j/N - 2 \times c_{kj})$$
$$= \sum\sum_{k=1}^{J} N/M_j \times [c_{kj} \frac{c_{kj}}{N_k}] + N - 2N \quad (\frac{c_{kj}}{N_k} \leq 1)$$
$$\leq \sum_{j=1}^{J} (N/M_j) \times \sum_{k=1}^{J} c_{kj} - N$$
$$= \sum_{j=1}^{J} (N/M_j) \times M_j - N$$
$$= (J-1) \times N$$

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}. □

Details of Global and Local Independence Tests

To facilitate our investigation, we first formally describe the local independence tests. Let $X^2_{(i,i+k)}$ be the Pearson's chi-square statistic for the $k+1$ consecutive rows (from i to $i+k$). Let $F_{\chi^2_{k\times(J-1)}}(X^2_{(i,i+k)})$ be the confidence level for these rows and their corresponding columns being statistical independent. The less the confidence level is, the harder we can not reject this hypothesis. Given this, assuming we have a user-specified threshold δ (usually less than 50%), if $F_{\chi^2_{k\times(J-1)}}(X^2_{(i,i+k)}) < \delta$, we can merge these $k+1$ rows into one row. Usually, the user defines a threshold for this purpose, such as 50%. If the confidence level derived from the independence test is lower than this threshold, that means we can not reject H_0, the independence hypothesis. Therefore, we treat them as statistically independent and allow to merge them.

Now, to relate the global goodness function to the local independence test, we map the global function to the local conditions. Considering we have two two contingency table C_1, and C_2, with the only difference between them is that the k consecutive rows in C_1 are merged into a single row in C_2. Given this, we can transform the global difference as the local difference:

$$Cost(C_1) - Cost(C_2) = -G^2_{i,i+k-1} + O((k-1)(J-1)logN)$$

where we assume the penalty for the global goodness function is O(df logN). Since the discretization reduces the value of the global goodness function, i.e. $Cost(C_1) > Cost(C_2)$, we need the local condition $-G^2_{i,i+k-1} + O((k-1)(J-1)logN) > 0$. In other words, as long as

$$G^2_{i,i+k-1} < O((k-1)(J-1)logN)$$

holds, we can combine the k rows into one.

Let us focus now on the local independence test, which has been used in the goodness function based on the number of intervals for the discretization. Note that we require

$$F_{X^2_{(k-1)\times(J-1)}}(G^2_{(i,i+k-1)}) < \delta$$

We would like the G^2 statistic for the k consecutive rows is as small as possible (as close as to 0). The δ usually choose less than or equal to 50%. This means we have at least more confidence to accept the independence hypothesis test than reject it. Based on the approximation results from Fisher and Wilson & Hilferty [21], the 50% percentile point of χ^2_{df} is

$$\chi^2_{df,0.5} \approx df = (k-1)(J-1)$$

Given this, we can rewrite our local hypothesis tests as

$$G^2_{i,i+k-1} < O(df)$$

In other words, as long as the above condition holds, we can merge these consecutive rows into one.

Derivation of Goodness Function based on Generalized Entropy (and Gini)

As discussed in Section 2, MDLP provides a general way to consider the trade-off between the complexity of the data given a model and the complexity of the model itself (or referred to as the

penalty of the model). Here, we apply the *generalized entropy* to describe both complexities and use their sum as the goodness function.

First, for a given interval S_i, we use $N_i \times H_\beta(S_i)$ to describe the labeling information of each point. Note that this is an analog to the traditional information theory, which states that the average length of the shortest message can not lower than $N_i \times H(S_i)$.

$$cost_\beta(data|model) = N \times H_\beta(S_1, \cdots, S_{I'})$$
$$= N_1 \times H_\beta(S_1) + N_2 \times H_\beta(S_2) + \cdots + N_{I'} \times H_\beta(S_{I'})$$

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. For transferring the interval information, we consider the message will be transferred is in the following format. Supposing there are I' intervals, we will first transfer N_1 zeros followed by a stop symbol 1, then, N_2 zeros, until the last $N_{I'}$ zeros. Once again, such information corresponds to the impurity measures of such message using the generalized entropy. Such cost is served as the penalty function for the model. The impurity of all the intervals, denoted as $L_{1\beta}(I', N)$, will be as follows.

$$L_{1\beta}(I', N) = (N + I' - 1)H_\beta(\frac{N}{N + I' - 1}, \frac{I' - 1}{N + I' - 1})$$
$$= (N + I' - 1) \times \{\frac{N}{N + I' - 1}[1 - (\frac{N}{N + I' - 1})^\beta]/\beta +$$
$$\frac{I - 1}{N + I' - 1}[1 - (\frac{I - 1}{N + I' - 1})^\beta]/\beta\}$$
$$= N \times [1 - (\frac{N}{N + I' - 1})^\beta]/\beta + (I - 1) \times [1 - (\frac{I - 1}{N + I' - 1})^\beta]/\beta$$

Clearly, when $\beta \to 0$, this is the traditional entropy measure L_1. When $\beta = 1$, this is the impurity measure based on gini. Next, we have to considering transfer the coding book for each interval. For a given interval S_i, each code will correspond to a class. Therefore, there will be a total of $J!$ ways of coding. Given this, the total cost for the coding book, denoted as $L_{2\beta}$, can be written as

$$L_{2\beta} = I' \times J! \times H_\beta(\frac{1}{J!}, \frac{J! - 1}{J!})$$
$$= I' \times J! \times \{\frac{1}{J!}[1 - (\frac{1}{J!})^\beta]/\beta + \frac{J! - 1}{J!}[1 - (\frac{J! - 1}{J!})^\beta]/\beta\}$$

Given this, the penalty of the discretization based on the generalized entropy is

$$penalty_\beta(model) = L_{1\beta}(I', N) + L_{2\beta}$$
$$\approx (I' - 1) \times [1 - (\frac{I' - 1}{N + I' - 1})^\beta]$$
$$+I'[1 - (\frac{1}{J!})^\beta]/\beta \approx (I' - 1) \times [1 - (\frac{I' - 1}{N})^\beta]/\beta$$
$$+I'[1 - (\frac{1}{J!})^\beta]/\beta$$

Note that when $\beta = 1$, we have $penalty_\beta(model) \approx 2I' - 1$. When $\beta \to 0$, we have $penalty_\beta(model) \approx (I'-1)log(N/(I'-1)) + I'(J-1)logJ$.

Put together, the cost of the discretization based on is

$$Cost_\beta = \sum_{i=1}^{I'} N_i H_\beta(S_i) + L_{1\beta}(I', N) + L_{2\beta} = \sum_{i=1}^{I'} N_i H_\beta(S_i)$$
$$+(I' - 1) \times [1 - (\frac{I' - 1}{N})^\beta]/\beta + I'[1 - (\frac{1}{J!})^\beta]/\beta$$

Similar to the treatment in Section 2, we define the generalized goodness function as the cost difference between contingency table C^0 and contingency table C.

$$GF_\beta(C) = Cost_\beta(C^0) - Cost_\beta(C)$$

Note that the goodness function based on Gini can be derived simply by fixing $\beta = 1$.

Proof of Theorem 10

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

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

In addition, we have

$$\alpha \times (I - 1)(J - 1)[1 - (\frac{1}{N})^\beta]/\beta$$
$$-\alpha \times (I - 2)(J - 1)[1 - (\frac{1}{N})^\beta]/\beta =$$
$$\alpha \times (J - 1)[1 - (\frac{1}{N})^\beta]/\beta > 0$$

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 $N \times H(S_1, \cdots, S_I)$. By the concaveness of the H_β (Theorem 9),

$$N \times H_\beta(S_1, \cdots, S_I) \le N \times H_\beta(S_1 \cup \cdots \cup S_I)$$
$$N \times H_\beta(S_1, \cdots, S_I) = \sum_{k=1}^{I} N_k \times H_\beta(S_k)$$
$$= \sum_{k=1}^{I} N_k \times H_\beta(S_1 \cup \cdots \cup S_I)$$
$$= N \times H_\beta(S_1 \cup \cdots \cup S_I)$$

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 $N \times H_\beta(S_1, \cdots, S_J)$.

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

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