Big Data and Information Analytics
©American Institute of Mathematical Sciences
Volume **1**, Number **4**, October **2016**
doi:10.3934/bdia.2016014
pp. **341–347**

INCREASE STATISTICAL RELIABILITY WITHOUT LOSING PREDICTIVE POWER BY MERGING CLASSES AND ADDING VARIABLES

Wenxue Huang* and Xiaofeng Li*

School of Mathematics and Information Sciences, Guangzhou University
Guangzhou, 510006, China

Yuanyi Pan

Clearpier Inc., 1300-121 Richmond St.W.
Toronto, Ontario
Canada M5H 2K1

(Communicated by Zhen Mei)

Abstract. It is usually true that adding explanatory variables into a probability model increases association degree yet risks losing statistical reliability. In this article, we propose an approach to merge classes within the categorical explanatory variables before the addition so as to keep the statistical reliability while increase the predictive power step by step.

1. **Introduction.** In any applications where feature selection or dimension reduction is required, a key question to be answered is how many variables or features are enough. More variables may increase data based association degree but may also result in explanatory information reliability reduction or model over fitting. It is particularly important for a stepwise forward feature selection procedure [8] to decide when to stop the variable aggregation. It can be stopped when the maximum joint association or the predefined maximum number of variables is reached. More discussions about this subject can be found in [2].

The prediction accuracy naturally attracts most of the attention and has been studied for hundreds of years. Categorical data analysis alone has the rate of point-hit accuracy, of distribution bias and of the balanced one between them [9]. Huang, Shi and Wang [12] suggested that the measure of association is fundamental to obtain the prediction accuracy rate and that this measure will increase as more explanatory variables added in that probabilistic model [12].

The risk of model failure, or the model's reliability, is usually related to the average number of categories in a categorical predictive model. Guttman [7] presented methods to estimate the upper and lower bounds to a categorical data set's reliability. These estimates are functions of the number of categories available and the proportion of instance from which the model response is chosen. Probably the most

2010 *Mathematics Subject Classification.* Primary: 62H20, 62F07; Secondary: 68T30, 58F17.
Key words and phrases. Association, categorical data, category merging, statistical reliability, predictive power.
The first named author is partially supported by grant GZhU_HWX1-2001.
* Corresponding authors: Wenxue Huang and Xiaofeng Li.

generally applicable and widely used method for estimating the reliability of rating or judgment is with the intra-class correlation, or some variation of it [3]. However, none of these methods reflect the response variable's distribution.

We hence introduce a new measure, denoted as $E(\text{Gini}(X|Y))$, to measure the reliability. It is based on the classical measurement theory and the Gini coefficient [13]. $E(\text{Gini}(X|Y))$ measures the independent variable X's concentration degree given the dependent variable Y. This measure ensures that the reliability will always increase when two categories in X are merged, meaning these two categories are treated as one.

We also prove that the association between the merged independent variable and the target variable keeps exactly the same after the merge if the merged independent classes have the same condition probabilities. Thus, we believe that the solution to the dilemma of the association increase and the reliability decrease along the feature selection process is to merge categories with similar conditional probabilities before adding new variables.

This article is organized as follows. Section 2 presents the definitions for the association and the reliability measures; section 3 discusses how and why the independent classes are merged; two supportive experiments are analyzed in section 4; the last section is a brief summarization and discussion to the future work.

2. Association, reliability and the comparison matrix Φ.

2.1. The association measures.

Given a nominal categorical data set with one independent variable X and one dependent variable Y, the following two association measures are of our interest in this article to address the predicting accuracy issue. Both measures were further discussed in [6]. The first one is a measure based on modal (or optimal) prediction, the Goodman-Kruskal λ (denoted as λ thereafter).

$$\lambda = \frac{\sum_x \rho_{xm} - \rho_{\cdot m}}{1 - \rho_{\cdot m}},$$

where

$$\rho_{\cdot m} = \max_y \rho_{\cdot y} = \max_y p(Y = y), \quad \rho_{xm} = \max_y \rho_{xy} = \max_y p(X = x; Y = y).$$

Please note that $p(\cdot)$ is the probability of a statistical event. One can see that λ is the relative decrease rate of predicting errors as we go from predicting Y with X to that without X. The other association measure is the Goodman-Kruskal τ (denoted as τ thereafter). It is based on the proportional prediction and defined as follows.

$$\tau = \frac{\sum_x \sum_y {\rho_{xy}}^2 / \rho_{x\cdot} - \sum_y {\rho_{\cdot y}}^2}{1 - \sum_y {\rho_{\cdot y}}^2},$$

where

$$\rho_{x\cdot} = p(X = x).$$

τ calculates the relative decrease of long-run proportion of predictions accuracy from predicting Y with X to that without X. Both measures are used in many applications including supervised discretization [11, 10]. Both can be used to measure the prediction errors: the first one aims to maximize the point-to-point accuracy and the second one wants to keep the same distribution between the predicted and the real target variable.

2.2. **Reliability measure.** Going by "precision" in some publications, reliability may be ambiguous in certain cases [17]. But in our context, it is how much a probability model built upon a given nominal categorical data set may fail in predicting the unknowns hence the number of classes of the independent variable, or the expected one, approximately shows the model's reliability. The expected number of classes in a variable X is a variation of the well-known Gini index defined as follows.

$$E_p(X) = \sum_x {\rho_{x\cdot}}^2$$

Roughly speaking, the more independent classes a predictive model has, the less support each conditional probability has in a given data set with limited size hence the less reliable the constructed model is. However, this measure does not count adequately the target variable's distribution. We believe it is more appropriate to construct one that considers the concentration of the independent values in each dependent class. Here then comes our proposed measure of reliability of explanatory information

$$E(\text{Gini}(X|Y)) = 1 - \sum_x \sum_y \frac{{\rho_{xy}}^2}{\rho_{\cdot y}} = 1 - \sum_x \sum_y p(X = x, Y = y)p(X = x|Y = y),$$

which is nothing but the average number of independent classes within each dependent class.

It is easy to see that the value of $E(\text{Gini}(X|Y))$ is within $[0, 1]$, that it reaches the minimum when and only there is one independent category in each dependent class and that it reaches the maximum when and only when all independent classes equally distributed within each dependent class. Given $\text{Dmn}(X) = \{1, 2, \cdots, n_x\}$, we can further conclude that $E(\text{Gini}(X|Y))$ is within $[0, 1 - \frac{1}{n_x}]$; and the smaller $E(\text{Gini}(X|Y))$ is, the more reliable the predictor information is.

3. **Comparison matrix Φ and merging process.** To decide which independent classes to be merged, a category-to-variable measure is required to estimate each independent class' overall predictive power to the target variable. This new measure to all element pairs in X, denoted as $\Phi(Y|X)$, is a matrix given by

$$\Phi(Y|X) = (\phi^{st}(Y|X)),$$

where

$$\phi^{st}(Y|X) = \sum_y (\frac{\rho_{sy}}{\rho_{s\cdot}} - \frac{\rho_{ty}}{\rho_{t\cdot}})^2 \rho_{\cdot y}; s, t \in Dmn(X).$$

Thus, the (s,t)-entry in $\Phi(Y|X)$ is the weighted difference between the conditional probabilities of $X = s$ and of $X = t$. It can be seen that this comparison matrix has the following properties:

1. $\Phi(Y|X)$ is symmetrical.
2. The value in the diagonal entries is zero.
3. The smaller $\phi^{st}(Y|X)$ is, the more similar the two conditional distributions are. When $\phi^{st}(Y|X) = 0$, $X = s$ and $X = t$ have the exactly the same conditional distributions.

In addition, when we want to merge the classes in two independent variables, denoted as X_1 and X_2 with domains $Dmn(X_1) = \{1, 2, \cdots, n_{x_1}\}$ and $Dmn(X_2) =$

$\{1, 2, \cdots, n_{x_2}\}$, respectively, the extended expression of $\phi^{st}(Y|X)$ is

$$\phi^{ijkl}(Y|X_1, X_2) = \Sigma_y(\frac{\rho_{ijy}}{\rho_{ij\cdot}} - \frac{\rho_{kly}}{\rho_{kl\cdot}})^2\rho_{\cdot y} \quad for \quad i, k = 1, 2, \cdots, n_{x_1}, \quad and \quad for \quad j, l = 1, 2, \cdots, n_{x_2}.$$

Thus, the measure of association, Φ, can be applied to multi-dimensional or even high dimensional models.

3.1. **Merging the classes and enhancing the measure of associations.** The proportional prediction based association measure τ can also be rewritten as follows

$$\tau = \frac{\omega^{Y|X} - E_p(Y)}{1 - E_p(Y)},$$

where

$$\omega^{Y|X} = \sum_x \sum_y \frac{{\rho_{xy}}^2}{\rho_{x\cdot}} \quad and \quad E_p(Y) = \sum_y {\rho_{\cdot y}}^2.$$

Thus, $\omega^{Y|X}$ is equivalent to τ when it goes to evaluate the associations in a given data set before and after the independent classes are merged.

We also have the following theorem to explain why merging the nominal classes works.

Theorem 3.1. *If the conditional probabilities of $X = s$ and $X = t$ are all equal, i.e.,*

$$\frac{\rho_{sy}}{\rho_{s\cdot}} = \frac{\rho_{ty}}{\rho_{t\cdot}} = a_y, \ for\ y = 1, 2, \ldots, n_y,$$

then merging the classes $X = s$ and $X = t$, and labelling the merged variable X as X' gives us

$$\omega^{Y|X} = \omega^{Y|X'}.$$

Proof. Let

$$\omega^{Y|X'} = \sum_{x \neq s,t} \sum_y \frac{\rho_{xy}^2}{\rho_{x\cdot}} + \sum_y \frac{\rho_{my}^2}{\rho_{m\cdot}},$$

where m is the merged class of s and t

Because

$$\begin{aligned} \sum_y \frac{\rho_{my}^2}{\rho_{m\cdot}} &= \sum_y \frac{(\rho_{sy} + \rho_{ty})^2}{\rho_{s\cdot} + \rho_{t\cdot}} = \sum_y \frac{(a_y\rho_{s\cdot} + a_y\rho_{t\cdot})^2}{\rho_{s\cdot} + \rho_{t\cdot}} = \sum_y a_y^2(\rho_{s\cdot} + \rho_{t\cdot}) \\ &= \sum_y a_y^2\rho_{s\cdot} + \sum_y a_y^2\rho_{t\cdot} = \sum_y \frac{\rho_{sy}^2}{\rho_{s\cdot}^2}\rho_{s\cdot} + \sum_y \frac{\rho_{ty}^2}{\rho_{t\cdot}^2}\rho_{t\cdot} = \sum_y \frac{\rho_{sy}^2}{\rho_{s\cdot}} + \sum_y \frac{\rho_{ty}^2}{\rho_{t\cdot}}, \end{aligned}$$

we have

$$\sum_{x \neq s,t} \sum_y \frac{\rho_{xy}^2}{\rho_{x\cdot}} + \sum_y \frac{\rho_{my}^2}{\rho_{m\cdot}} = \sum_{x \neq s,t} \sum_y \frac{\rho_{xy}^2}{\rho_{x\cdot}} + \sum_y (\frac{\rho_{sy}^2}{\rho_{s\cdot}} + \frac{\rho_{ty}^2}{\rho_{t\cdot}}),$$

that is

$$\omega^{Y|X} = \omega^{Y|X'}.$$

□

Thus, $\tau(Y|X') = \tau(Y|X)$, where the X' represents the variable X with $X = s$ and $X = t$ merged, when the conditional probabilities are the same for $X = s$ and $X = t$.

On the other hand, when the conditional probabilities of $X = s$ and $X = t$ are extremely similar with each other, i.e., $\phi^{st}(Y|X)$ is very small, $\tau(Y|X')$ should be very close to $\tau(Y|X)$. It is then practically very possible to find another variable Z or merged Z' in an usual high dimensional data set such that $\tau(Y|X', Z) > \tau(Y|X)$, since it is almost certain that the added variable Z satisfies $\tau(Y|X, Z) > \tau(Y|X, S)$, where S is any other predictor besides X and Z. Meanwhile, a smaller $E(\text{Gini}(X|Y))$ ensures a better reliability. Thus, ideally, the added variables Z also satisfies that $E(Gini(X', Z|Y)) \leq E(\text{Gini}(X|Y))$. In the next section, we are going to show two examples to support the previous statements. Please note that the nominal classes to be merged don't need to have exactly the same, but the sufficiently close conditional probabilities. $\tau^{Y|X}$, $\lambda^{Y|X}$ and $E(Gini(X|Y)$ are all going to be investigated to evaluate the goodness of the merge.

4. Experiments.

Both experiments use the 1996 Survey of Family Expenditure administrated by The Statistics Canada [16]. It has $10,417$ rows with over 200 continuous and categorical variables but we are only going to use some of them as the supportive evidences.

4.1. Occupation, sex, age group and education.

The first result shows how the reliability and the association degrees are changed when Sex is added to Age group with Occupation as the target variable. The result briefly demonstrate how a regular feature selection process without merging works. It is also going to be used as the baseline to evaluate the performance after the merge.

TABLE 1. Feature selection without merging: Occupation

X	$\tau^{Y\|X}$	$\lambda^{Y\|X}$	$E(\text{Gini}(X\|Y))$
Age group	0.1344	0.0311	0.8773
Age group + Sex	0.1511	0.0476	0.9228

As discussed above, the added variable Sex increases the association, measured by τ or λ, but reduces the reliability.

Knowing that Age group has 13 categories and the $E(\text{Gini}(X|Y))$ is 0.8773, we choose $\phi^{st}(Y|X) \leq 0.003$ as the criteria to merge class 2 to class 7 and class 11 to 13. Treating merged Age group, denoted as Age group', and Sex as a single variable, we can merge it again using $\phi^{ijkl}(Y|X_1, X_2)$ and the same threshold. Table 2 shows the computation result.

TABLE 2. Feature selection with merging: Occupation

X	$\tau_b^{(Y\|X)}$	$\lambda^{(Y\|X)}$	$E(\text{Gini}(X\|Y))$
Agegroup'+Sex	0.1484	0.0375	0.6688
(Age group'+Sex)'+Education'	0.1542	0.0447	0.6620

Table 2 tells us that the merged Age Group, combined with Sex is better in reliability given the smaller $E(Gini(X|Y)$ with worse association both in τ and λ, compared with that without merging in Table 1. However, if we merge the merged

then add *Education* into the variable list, we have a better association AND better reliability, which was impossible in the old feature selection process without merging.

It is clear that the merging threshold determines how many classes will be merged therefore affects the quality. Table 3 shows some simple analysis.

TABLE 3. Compare different merging threshold:Occupation

X	$\phi^{st}(Y\|X)$	$\lambda^{(Y\|X)}$	$\tau^{(Y\|X)}$	$E(Gini(X,Y))$
Age group	-	0.0311	0.1344	0.8773
$Agegroup'$+Sex	0.0005	0.0414	0.1493	0.9222
$Agegroup'$+Sex	0.0030	0.0375	0.1484	0.6688
$Agegroup'$+Sex	0.0100	0.0000	0.0209	0.2710

As Table 3 suggests, the bigger the merging threshold is, the more classes are merged then the higher the reliability is while the lower the association is. One can tune this parameter to achieve the needed result given certain trade-off considerations. The chosen ones in this article come from practical considerations than theoretical optima.

4.2. **House type, rooms, bedroom and tenure.** Following similar steps to those the previous section presents with different variable sets, we consider House type as the target variable and investigate the effect of merging. Please note that the threshold is still $\phi^{st}(Y|X) \leq 0.003$.

TABLE 4. Compare different merging threshold

X	$\lambda^{(Y\|X)}$	$\tau^{(Y\|X)}$	$E(\text{Gini}(X\|Y))$
Rooms	0.3443598	0.3004656	0.8200656
$Rooms'$+$Tenure'$	0.4255117	0.3583277	0.7911177
$(Rooms'$+$Tenure')' + bedroom'$	0.4381247	0.3901767	0.7165204

Table 4 also shows us an example of not only the better reliability but also the higher association after two merged variables are combined.

5. **Conclusion.** Based on the theory of association measure and the Gini coefficient, we take $E(\text{Gini}(X|Y))$ to measure the statistical reliability. A category-to-variable comparison matrix $\Phi(Y|X)$ is proposed to represent the conditional probability differences between the explanatory variable's classes. We are going to implement both $E(\text{Gini}(X|Y))$ and $\Phi(Y|X)$ in an improved feature selection process in the future. Generally, this improved process will merge classes in the candidate variables before adding one of them into the selected independent variable list. By doing so, the selected features will keep reliability high while the associations increased step by step, as shown in the above experiments.

REFERENCES

[1] H. L. Costner, Criteria for measure of association, *American Sociology Review,* **30** (1965), 341–353.
[2] M. Dash and H. Liu, Feature selection for classification, *Intell. Data. Anal.,* **1** (1997), 131–156.
[3] R. L. Ebel, Estimation of the reliability of ratings, *Psychomereika,* **16** (1951), 407–424.
[4] G. S. Fisher, *Monte Carlo: Concepts, Algorithms, and Applications*, Springer-Verlag, 1996.

[5] P. Glasserman, *Monte Carlo Method in Financial Engineering*, (Stochastic Modelling and Applied Probability) (V. 53), Spinger, 2004.
[6] L. A. Goodman and W. H. Kruskal, *Measures of Associations for Cross Classification*, With a foreword by Stephen E. Fienberg. Springer Series in Statistics, 1. Springer-Verlag, New York-Berlin, 1979.
[7] L. Guttman, The test-retest reliability of qualitative data, *Psychometrika,* **11** (1946), 81–95.
[8] I. Guyon and A. Elisseeff, An introduction to variable and feature selection, *J. Mach. Learn. Res.,* **3** (2003), 1157–1182.
[9] W. Huang and Y. Pan, On balancing between optimal and proportional categorical predictions, *Big Data and Info. Anal.,* **1** (2016), 129–137.
[10] W. Huang, Y. Pan and J. Wu, Supervised Discretization with $GK - \tau$, *Proc. Comp. Sci.,* **17** (2013), 114–120.
[11] W. Huang, Y. Pan and J. Wu, Supervised discretization for optimal prediction, *Proc. Comp. Sci.,* **30** (2014), 75–80.
[12] W. Huang, Y. Shi and X. Wang, A nominal association matrix with feature selection for categorical data, *Communications in Statistics -Theory and Methods,* 2017.
[13] M. G. Kendall, *The Advanced Theory of Statistics*, London, Charles Griffin and Co., Ltd, 1946.
[14] C. J. Lloyd, *Statistical Analysis of Categorical Data*, John Wiley Sons, 1999.
[15] K. Pearson and D. Heron, On Theories of association, *Biometrika,* **9** (1913), 159–315.
[16] STATCAN, *Survey of Family Expenditures - 1996.* (1998)
[17] D. L. Streiner and G. R. Norman, "Precision" and "accuracy": Two terms that are neither, *J. of Cli. Epid.,* **59** (2006), 327–330.

E-mail address: whuang123@yahoo.com
E-mail address: xiaofeng-li@foxmail.com
E-mail address: Yuanyi.Pan@gmail.com