Big Data and Information Analytics
©American Institute of Mathematical Sciences
doi:10.3934/bdia.2017004

PROPORTIONAL ASSOCIATION BASED ROI MODEL

Wenxue Huang

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

Yuanyi Pan

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

Lihong Zheng*

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

(Communicated by Yong Shi)

Abstract. Based on a local-to-global proportional association measure proposed by Huang, Shi and Wang [9], with cost and revenue information known, an association measure is proposed to maximize the expected RoI. A descriptive experiment with a synthetical data set is presented.

1. **Introduction.** Multi-nominal data are common in scientific and engineering research such as biomedical research, customer behavior analysis, network analysis, search engine marketing optimization, web mining etc. When the response variable has more than two levels, the principle of mode-based or distribution-based proportional prediction can be used to construct nonparametric nominal association measure. For example, Goodman and Kruskal [3, 4] and others proposed some local-to-global association measures towards optimal predictions. Both Monte Carlo and discrete Markov chain methods are conceptually based on the proportional associations. The association matrix, association vector and association measure were proposed by the thought of proportional associations in [9]. If there is no ordering to the response variable's categories, or the ordering is not of interest, they will be regarded as nominal in the proportional prediction model and the other association statistics.

But in reality, different categories in the same response variable often are of different values, sometimes much different. When selecting a model or selecting explanatory variables, we want to choose the ones that can enhance the total revenue, not just the accuracy rate. Similarly, when the explanatory variables with

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

Key words and phrases. Dimensioncost, proportional association matrix, return-on-investment, revenue.

The first named author is partially supported by grant GZhu_HWX1-2001.

*Corresponding authors: Wenxue Huang and Lihong Zheng.

cost weight vector, they should be considered in the model too. The association measure in [9], $\omega^{Y|X}$, doesn't consider the revenue weight vector in the response variable, nor the cost weight in the explanatory variables, which may lead to less profit in total. Thus certain adjustments must be made for a better decisionning.

To implement the previous adjustments, we need the following assumptions:

- X and Y are both multi-categorical variables where X is the explanatory variable with domain $\{1, 2, ..., \alpha\}$ and Y is the response variable with domain $\{1, 2, ..., \beta\}$ respectively;
- the amount of data collected in this article is large enough to represent the real distribution;
- the model in the article mainly is based on the proportional prediction;
- the relationship between X and Y is asymmetric;

It needs to be addressed that the second assumption is probably not always the case. The law of large number suggests that the larger the sample size is, the closer the expected value of a distribution is to the real value. The study of this subject has been conducted for hundreds of years including how large the sample size is enough to simulate the real distribution. Yet it is not the major subject of this article. The purpose of this assumption is nothing but a simplification to a more complicated discussion.

The article is organized as follows. Section 2 discusses the adjustment to the association measure when the response variable has a revenue weight; section 3 considers the case where both the explanatory and the response variable have weights; how the adjusted measure changes the existing feature selection framework is presented in section 4. Conclusion and future works will be briefly discussed in the last section.

2. **Response variable with revenue weight vector.** Let's first recall the association matrix $\{\gamma^{s,t}(Y|X)\}$ and the association measure $\omega^{Y|X}$ and $\tau^{Y|X}$.

$$\begin{aligned}
\gamma^{s,t}(Y|X) &= \frac{E(p(Y=s|X)p(Y=t|X))}{p(Y=s)} \\
&= \sum_{i=1}^{\alpha} p(X=i|Y=s)p(Y=t|X=i); s,t=1,2,..,\beta \\
\tau^{Y|X} &= \frac{\omega^{Y|X} - Ep(Y)}{1 - Ep(Y)} \\
\omega^{Y|X} &= E_X(E_Y(p(Y|X))) \\
&= \sum_{s=1}^{\beta}\sum_{i=1}^{\alpha} p(Y=s|X=i)^2 p(X=i) \qquad (1) \\
&= \sum_{s=1}^{\beta} \gamma^{ss} p(Y=s)
\end{aligned}$$

$\gamma^{st}(Y|X)$ is the (s,t)-entry of the association matrix $\gamma(Y|X)$ representing the probability of assigning or predicting $Y=t$ while the true value is in fact $Y=s$. Given a representative train set, the diagonal entries, γ^{ss}, are the expected accuracy rates while the off-diagonal entries of each row are the expected first type error rates. $\omega^{Y|X}$ is the association measure from the explanatory variable X to the response variable Y without a standardization. Further discussions to these metrics can be found in [9].

Our discussion begins with only one response variable with revenue weight and one explanatory variable without cost weight. Let $R = (r_1, r_2, ..., r_\beta)$ to be the revenue weight vector where r_s is the possible revenue for $Y = s$. A model with highest revenue in total is then the ideal solution in reality, not just a model with highest accuracy. Therefore comes the extended form of $\omega^{Y|X}$ with weight in Y as in 2:

Definition 2.1.

$$\begin{aligned} \widehat{\omega}^{Y|X} &= \sum_{s=1}^{\beta}\sum_{i=1}^{\alpha} p(Y=s|X=i)^2 r_s p(X=i) \\ &= \sum_{s=1}^{\beta} \gamma^{ss} p(Y=s) r_s \\ r_s &> 0, s = 1, 2, 3..., \beta \end{aligned} \tag{2}$$

Please note that $\omega^{Y|X}$ is equivalent to $\tau^{Y|X}$ for given X and Y in a given data set. Thus the statistics of $\tau^{Y|X}$ will not be discussed in this article.

It is easy to see that $\widehat{\omega}^{Y|X}$ is the expected total revenue for correctly predicting Y. Therefore one explanatory variable X_1 with $\widehat{\omega}^{Y|X_1}$ is preferred than another X_2 if $\widehat{\omega}^{Y|X_1} \geq \widehat{\omega}^{Y|X_2}$. It is worth mentioning that $\widehat{\omega}^{Y|X}$ is asymmetric,i.e., $\widehat{\omega}^{Y|X} \neq \widehat{\omega}^{X|Y}$ and that $\omega^{Y|X} = \widehat{\omega}^{Y|X}$ if $r_1 = r_2 = ... = r_\beta = 1$.

Example. Consider a simulated data motivated by a real situation. Suppose that variable Y is the response variable indicating the different computer brands that the customers bought; X_1, as one explanatory variable, shows the customers' career and X_2, as another explanatory variable, tells the customers' age group. We want to find a better explanatory variable to generate higher revenue by correctly predicting the purchased computer's brand. We further assume that X_1 and X_2 both contain 5 categories, Y has 4 brands and the total number of rows is 9150. The contingency table is presented in 1.

Table 1. Contingency tables:X_1 vs Y and X_2 vs Y

$X_1\|Y$	y_1	y_2	y_3	y_4	$X_2\|Y$	y_1	y_2	y_3	y_4
x_{1_1}	1000	100	500	400	x_{2_1}	500	300	200	1500
x_{1_2}	200	1500	500	300	x_{2_2}	500	400	400	50
x_{1_3}	400	50	500	500	x_{2_3}	500	500	300	700
x_{1_4}	300	700	500	400	x_{2_4}	500	400	1000	100
x_{1_5}	200	500	400	200	x_{2_5}	200	400	500	200

Let us first consider the association matrix $\{\gamma^{Y|X}\}$. Predicting Y with the information of X_1, or X_2 is given by the association matrix $\gamma(Y|X_1)$, or $\gamma(Y|X_2)$ as in Table 2.

Please note that Y contains the true values while $\hat{Y}$ is the guessed one. One can see from this table that the accuracy rate of predicting y_1 and y_2 by X_1 on the left

TABLE 2. Association matrices:X_1 vs Y and X_2 vs Y

$Y\|\hat{Y}$	$\hat{y_1}\|X_1$	$\hat{y_2}\|X_1$	$\hat{y_3}\|X_1$	$\hat{y_4}\|X_1$	$Y\|\hat{Y}$	$\hat{y_1}\|X_2$	$\hat{y_2}\|X_2$	$\hat{y_3}\|X_2$	$\hat{y_4}X_2$
y_1	0.34	0.18	0.27	0.22	y_1	0.26	0.22	0.27	0.25
y_2	0.13	0.48	0.24	0.15	y_2	0.25	0.24	0.29	0.23
y_3	0.24	0.28	0.27	0.21	y_3	0.25	0.24	0.36	0.15
y_4	0.25	0.25	0.28	0.22	y_4	0.22	0.18	0.14	0.46

are larger than that on the right. The cases of y_3 and y_4, on the other hand, are opposite.

The correct prediction contingency tables of X_1 and Y, denoted as W_1, plus that of X_2 and Y, denoted as W_2, can be simulated through Monte Carlo simulation as in Table 3.

TABLE 3. Contingency table for correct predictions: W_1 and W_2

$X_1\|Y$	y_1	y_2	y_3	y_4	$X_2\|Y$	y_1	y_2	y_3	y_4
x_{1_1}	471	6	121	83	x_{2_1}	98	34	19	926
x_{1_2}	101	746	159	107	x_{2_2}	177	114	113	1
x_{1_3}	130	1	167	157	x_{2_3}	114	124	42	256
x_{1_4}	44	243	145	85	x_{2_4}	109	81	489	6
x_{1_5}	21	210	114	32	x_{2_5}	36	119	206	28

The total number of the correct predictions by X_1 is 3142 while it is 3092 by X_2, meaning the model with X_1 is better than X_2 in terms of accurate prediction. But it maybe not the case if each target class has different revenues. Assuming the revenue weight vector of Y is $R = (1, 1, 2, 2)$, we have the association measure of $\omega^{Y|X}$, and $\widehat{\omega}^{Y|X}$ as in Table 4:

TABLE 4. Association measures: $\omega^{Y|X}$, and $\widehat{\omega}^{Y|X}$

X	$\omega^{Y\|X}$	$\widehat{\omega}^{Y\|X}$	total revenue	average revenue
X_1	0.3406	0.456	4313	0.4714
X_2	0.3391	0.564	5178	0.5659

Given that $revenue = \sum_{i,s} W_k^{i,s} r_s, i = 1, 2, ..., \alpha, s = 1, 2, ..., \beta, k = 1, 2$, we have the revenue for W_1 as 4313, and that for W_2 as 5178. Divide the revenue by the total sample size, 9150, we can obtain 0.4714 and 0.5659 respectively. Contrasting these to $\widehat{\omega}^{Y|X_1}$ and $\widehat{\omega}^{Y|X_2}$ above, we believe that they are similar, which means then $\widehat{\omega}^{Y|X}$ is truly the expected revenue.

In summary, it is possible for an explanatory variable X with bigger $\widehat{\omega}^{Y|X}$, i.e, the larger revenue, but with smaller $\omega^{Y|X}$, i.e., the smaller association. When the total revenue is of the interest, it should be the better variable to be selected, not the one with larger association.

3. **Explanatory variable with cost weight and response variable with revenue weight.** Let us further discuss the case with cost weight vector in predictors in addition to the revenue weight vector in the dependent variable. The goal is to

find a predictor with bigger profit in total. We hence define the new association measure as in 3.

Definition 3.1.

$$\bar{\omega}^{Y|X} = \sum_{i=1}^{\alpha}\sum_{s=1}^{\beta} p(Y = s|X = i)^2 \frac{r_s}{c_i} p(X = i) \tag{3}$$

$c_i > 0, i = 1, 2, 3, ..., \alpha$, and $r_s > 0, s = 1, 2, ..., \beta$.

c_i indicates the cost weight of the ith category in the predictor and r_s means the same as in the previous section. $\bar{\omega}^{Y|X}$ is then the expected ratio of revenue and cost, namely RoI. Thus a larger $\bar{\omega}^{Y|X}$ means a bigger profit in total. A better variable to be selected then is the one with bigger $\bar{\omega}^{Y|X}$. We can see that $\bar{\omega}^{Y|X}$ is an asymmetric measure, meaning $\bar{\omega}^{Y|X} \neq \bar{\omega}^{Y|X}$. When $c_1 = c_2 = ... = c_\alpha = 1$, Equation 3 is exactly Equation 2; when $c_1 = c_2 = ... = c_\alpha = 1$ and $r_1 = r_2 = ... = r_\beta = 1$, equation 3 becomes the original equation 1.

Example. We first continue the example in the previous section with new cost weight vectors for X_1 and X_2 respectively. Assuming $C_1 = (0.5, 0.4, 0.3, 0.2, 0.1)$, $C_2 = (0.1, 0.2, 0.3, 0.4, 0.5)$ and $R = (1, 1, 1, 1)$, we have the associations in Table 5.

Table 5. Association with/without cost vectors: X_1 and X_2

X	$\omega^{Y\|X}$	$\widehat{\omega}^{Y\|X}$	$\bar{\omega}^{Y\|X}$	total profit	average profit
X_1	0.3406	0.3406	1.3057	12016.17	1.3132
X_2	0.3391	0.3391	1.8546	17072.17	1.8658

By $profit = \sum_{i,s} W_k^{i,s} \frac{r_s}{C_{k_i}}, i = 1, 2, .., \alpha; s = 1, 2, .., \beta$ and $k = 1, 2$ where W_k is the corresponding prediction contingency table, we have the profit for X_1 as 12016.17 and that of X_2 as 17072.17. When both divided by the total sample size 9150, they change to 1.3132 and 1.8658, similar to $\bar{\omega}(Y|X_1)$ and $\bar{\omega}(Y|X_2)$. It indicates that $\bar{\omega}^{Y|X}$ is the expected RoI. In this example, X_2 is the better variable given the cost and the revenue vectors are of interest.

We then investigate how the change of cost weight affect the result. Suppose the new weight vectors are: $R = (1, 1, 1, 1)$, $C_1 = (0.1, 0.2, 0.3, 0.4, 0.5)$ and $C_2 = (0.5, 0.4, 0.3, 0.2, 0.1)$, we have the new associations in Table 6.

Table 6. Association with/without new cost vectors: X_1 and X_2

X	$\omega^{Y\|X}$	$\widehat{\omega}^{Y\|X}$	$\bar{\omega}^{Y\|X}$	total profit	average profit
X_1	0.3406	0.3406	1.7420	15938.17	1.7419
X_2	0.3391	0.3391	1.3424	12268.17	1.3408

Hence $\bar{\omega}^{Y|X_1} > \bar{\omega}^{Y|X_2}$, on the contrary to the example with the old weight vectors. Thus the right amount of weight is critical to define the better variable regarding the profit in total.

4. **The impact on feature selection.**

By the updated association defined in the previous section, we present the feature selection result in this section to a given data set S with explanatory categorical variables $V_1, V_2, .., V_n$ and a response variable Y. The feature selection steps can be found in [9].

At first, consider a synthetic data set simulating the contribution factors to the sales of certain commodity. In general, lots of factors could contribute differently to the commodity sales: age, career, time, income, personal preference, credit, etc. Each factor could have different cost vectors, each class in a variable could have different cost as well. For example, collecting income information might be more difficult than to know the customer's career; determining a dinner waitress' purchase preference is easier than that of a high income lawyer. Therefore we just assume that there are four potential predictors, V_1, V_2, V_3, V_4 within the data set with a sample size of 10000 and get a feature selection result by monte carlo simulation in Table 7.

Table 7. Simulated feature selection: one variable

X	$\|Dmn(X)\|$	$\omega^{Y\|X}$	$\bar{\omega}^{Y\|X}$	total profit	average profit
V_1	7	0.3906	3.5381	35390	3.5390
V_2	4	0.3882	3.8433	38771	3.8771
V_3	4	0.3250	4.8986	48678	4.8678
V_4	8	0.3274	3.7050	36889	3.6889

The first variable to be selected is V_1 using $\omega^{Y|X}$ as the criteria according to [9]. But it is V_3 that needs to be selected as previously discussed if the total profit is of interest. Further we assume that the two variable combinations satisfy the numbers in Table 8 by, again, monte carlo simulation.

Table 8. Simulated feature selection: two variables

X_1, X_2	$\|Dmn(X_1, X_2)\|$	$\omega^{Y\|(X_1,X_2)}$	$\bar{\omega}^{Y\|(X_1,X_2)}$	total profit	average profit
V_1, V_2	28	0.4367	1.8682	18971	1.8971
V_1, V_3	28	0.4025	2.1106	20746	2.0746
V_1, V_4	56	0.4055	1.8055	17915	1.7915
V_3, V_2	16	0.4055	2.3585	24404	2.4404
V_3, V_4	32	0.3385	2.0145	19903	1.9903

As we can see, all $\omega^{Y|(X_1,X_2)} \geq \omega^{Y|X_1}$, but it is not case for $\bar{\omega}^{Y|(X_1,X_2)}$ since the cost gets larger with two variables thus the profit drops down. As in one variable scenario, the better two variable combination with respect to $\omega^{Y|(X_1,X_2)}$ is (V_1, V_2) while $\bar{\omega}^{Y|(X_1,X_2)}$ suggests $(V_3\ ,V_2)$ is the better choice.

In summary, the updated association with cost and revenue vector not only changes the feature selection result by different profit expectations, it also reflects a practical reality that collecting information for more variables costs more thus reduces the overall profit, meaning more variables is not necessarily better on a Return-Over-Invest basis.

5. **Conclusions and remarks.**

We propose a new metrics, $\omega^{\bar{Y}|X}$ in this article to improve the proportional prediction based association measure, $\omega^{Y|X}$, to analyze the cost and revenue factors in the categorical data. It provides a description to the global-to-global association with practical RoI concerns, especially in a case where response variables are multi-categorical.

The presented framework can also be applied to high dimensional cases as in national survey, misclassification costs, association matrix and association vector [9]. It should be more helpful to identify the predictors' quality with various response variables.

Given the distinct character of this new statistics, we believe it brings us more opportunities to further studies of finding the better decision for categorical data. We are currently investigating the asymptotic properties of the proposed measures and it also can be extended to symmetrical situation. Of course, the synthetical nature of the experiments in this article brings also the question of how it affects a real data set/application. It is also arguable that the improvements introduced by the new measures probably come from the randomness. Thus we can use k-fold cross-validation method to better support our argument in the future.

REFERENCES

[1] C. Cornforth, What makes boards effctive? an examination of the relationships between board inputs, structures, processes and effctiveness in non-profit organisations, *Corporate Governance: An International Review,* **9** (2011), 217–227.

[2] L. L. Fong, M. S. Squillante and R. E. Hough, Computer resource proportional utilization and response time scheduling, *US Patent*, **6** (2001), 263–359.

[3] L. A. Goodman, A single general method for the analysis of cross-classifed data: Reconciliation, and synthesis of some methods of pearson, yule, and fisher, and also some methods of correspondence analysis and association analysis, *Journal of the American Statistical Association*, **91** (1996), 408–428.

[4] L. A. Goodman and W. H. Kruskal, *Measures of Association for Cross Classifications*, Springer, 1979.

[5] M. F. Gregor, L. Yang, E. Fabbrini, B. S. Mohammed, J. C. Eagon, G. S. Hotamisligil and S. Klein, Endoplasmic reticulum stress is reduced in tissues of obese subjects after weight loss, *Diabetes*, **58** (2009), 693–700.

[6] W. Huang and Y. Pan, On balancing between optimal and proportional categorical predictions, *Big Data and Information Analytics*, **1** (2016), 129–137.

[7] W. Huang, Y. Pan and J. Wu, Supervised discretization with GK-τ, *Procedia Computer Science*, **17** (2013), 114–120.

[8] W. Huang, Y. Pan and J. Wu, Performance measures of rare events targeting, *International Journal of Data Analysis Techniques and Strategies*, **6** (2014), 105–120.

[9] W. Huang, Y. Shi and X. Wang, A nominal association matrix with feature selection for categorical data, *Comunications in Statistic - Theory and Methods*, **46** (2017), 7798–7819.

[10] H. Hwang, T. Jung and E. Suh, An ltv model and customer segmentation based on customer value: A case study on the wireless telecommunication industry, *Expert Systems with Applications*, **26** (2004), 181–188.

[11] T. Lin, Y. Yang and H. T. Shiau, A work weighted state vector control method for geometrically nonlinear analysis, *Computers and Structures*, **46** (1993), 689–694.

[12] C. X. Ling and C. Li, Data mining for direct marketing: Problems and solutions, in *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98)*, AAAI Press, 1998, 73–79.

[13] J. R. Quinlan, Induction of decision trees, *Machine Learning*, **1** (1986), 81–106.

E-mail address: whuang123@yahoo.com
E-mail address: Yuanyi.Pan@gmail.com
E-mail address: zhenglihongmila@qq.com