Finding Association Rules by Direct Estimation of Likelihood Ratios

Kento Kawakami*, Masato Kikuchi*, Mitsuo Yoshida*, Eiko Yamamoto†, Kyoji Umemura*
*Department of Computer Science and Engineering
Toyonashi University of Technology, Toyonashi 441–8580, Japan
{k131820@edu.m143313@edu,yoshida@cs}.tut.ac.jp, umemura@tut.jp
†Department of Economics and Information
Gifu Shotoku Gakuen University, Gifu 500–8288, Japan
eiko@gifu.shotoku.ac.jp

Abstract—In this paper, we propose a cost function that corresponds to the mean square errors between estimated values and true values of conditional probability in a discrete distribution. We then obtain the values that minimize the cost function. This minimization approach can be regarded as the direct estimation of likelihood ratios because the estimation of conditional probability can be regarded as the estimation of likelihood ratio by the definition of conditional probability. When we use the estimated value as the strength of association rules for data mining, we find that it outperforms a well-used method called Apriori.

Index Terms—Data Mining, Association Rule, Likelihood Ratio, Apriori

I. INTRODUCTION

Finding association rules between items is a fundamental operation in data mining [1], [2]. Apriori [3] is an effective and efficient method for this operation. This method uses the maximum likelihood estimator (MLE) of conditional probability \( P(x \in R_c \mid y \in R_c) \) where \( x \) and \( y \) are variables for items, and \( R_c \) is a random variable corresponding to a drawn record in a sampling space. The event \( x \in R_c \) indicates that drawn record \( r_c \) (i.e., the sample of \( R_c \)) contains item \( x \). Let \( C(x, y) \) be the number of records that contain both item \( x \) and item \( y \). This indicates that \( C(x, y) \) is the frequency of co-occurrence of item \( x \) and item \( y \). Let \( C(x) \) be equivalent to \( C(x, x) \). The MLE of \( P(x \in R_c \mid y \in R_c) \) becomes \( C(x, y)/C(y) \). It is well known that the value of MLE is unstable and may contain large amount of errors when \( C(x, y) \) is small. Apriori notices this problem and considers the strength of rules as zero when \( C(x, y) \) is smaller than a certain threshold. This threshold, called minimum support, is a type of parameter in Apriori, which should be tuned according to the data.

Although ignoring infrequently co-occurring items would be a natural, simple, and effective option for Apriori to find the strong association rules, the true strength of association for these items is actually not zero even when they co-occur in only a few records. In addition, changing the treatment of the co-occurred item pairs on the boundary of the threshold may not be a fair treatment for all item pairs.

For this reason, we propose to use a cost function of errors, and the cost function reflects the mean square errors between true values and estimated values of conditional probability \( P(x \in R_c \mid y \in R_c) \). Moreover, we propose to use a normalized term in this cost function to overcome the instability of the estimation for some item pairs. We then decide the estimated value that minimizes the cost function and consider the value as the strength of association of item pairs.

This method uses the framework of the direct estimation of likelihood ratios [4], which is typically used for a continuous distribution. This method requires the selection of kernel functions whose linear combination is the estimated result. This selection is application dependent, and our selection for association rule mining is unique because our distribution is discrete.

We compared the effectiveness of the proposed method with that of Apriori, and found that the proposed method outperformed Apriori with statistical significance. In addition to Apriori, we compared the proposed estimation with a smoothed estimator of probability. This is because our analytical solution of cost minimization slightly resembles the formula of additive smoothing [5], which contains Laplace smoothing in a special case. We found that the smoothed estimator is worse than Apriori. This suggests that the proposed method is novel and does not belong to this smoothed estimator.

II. RELATED WORK

Apriori [3] is a standard method for finding the association rules. Apriori uses MLE with a threshold called “minimum support”, as explained in the introduction. Although the value of Apriori is famous for its efficient implementation that utilizes the threshold, we first focus our attention toward the effectiveness rather than efficiency. Apriori achieves the effectiveness by ignoring the rules that may not exhibit sufficient evidence.

Predictive Apriori [6] obtains its effectiveness by using the expected value instead of MLE. In this framework, the current database is considered as the sample. First, Predictive Apriori learns the prior distribution of \( C(x, y)/C(y) \) over the choice of samples. This prior distribution should be of the same shape for all \( x \) and all \( y \). Predictive Apriori then estimates the posterior distribution of ratio \( C(x, y)/C(y) \). The posterior
distribution should be affected by the \( C(x, y) \) and \( C(y) \). Predictive Apriori computes the expected value of the ratio over the choice of samples in the posterior distribution. As it is hard or unreasonable to assume some parametric distribution for \( C(x, y)/C(y) \), a histogram is suggested for expressing the prior distribution. As a histogram consists of many numbers in its shape, Predictive Apriori needs many parameters. We may interpret that Predictive Apriori obtains its effectiveness by tuning many parameters.

The direct estimation of likelihood ratios [4] is reported to be more accurate than obtaining two estimations of probability and then utilizing its ratio. It forms the cost function of estimation error of ratio rather than forming two cost functions for each probability. Then, the cost function is empirically expressed by the observed data. By assuming that the estimation of ratio is expressed by the linear combination of functions (called kernel functions), the solution of cost minimization problem provides the weights on each kernel function and the shape of the likelihood ratio. The primary concern here is how to decide the set of kernel functions. The suggested functions [7], [8] are a family of Gaussian functions, which is only significant for a continuous distribution.

There is another approach for this task. Kikuchi et al. [9] argue that the cost where false association rules are selected is considerably higher than the cost where true association rules are not selected. Rather than estimating the true value, they proposed to form a confidence interval of estimation of rules. Rather than estimating the true value, they proposed to form a confidence interval of estimation of rules. In this paper, we use \( f(x) := g(x) \) to define \( f \). To explain the proposed method, we define the following functions.

By the definition of conditional probability, we have the following formula as [7].

\[
 r(x, y) = \frac{p(x, y)}{p(y)}. \tag{4}
\]

Then, we introduce a model of \( \hat{r}_\alpha(x, y) \) for the estimation of \( r(x, y) \). This is a linear combination of kernel functions \( \phi_{ij}(x, y) \), where the variables \( i \) and \( j \) move in \( S \) from 1 to \( v(= |S|) \), where \( \phi_{ij} \) is some non-negative real number and \( \alpha \) is a vector whose \( (i \times v + j) \)-th element is \( \alpha_{ij} \) and \( \phi(x, y) \) is a vector whose \( (i \times v + j) \)-th element is \( \phi_{ij}(x, y) \).

\[
 \hat{r}_\alpha(x, y) := \alpha^T \phi(x, y) \tag{5}
\]

These kernel functions are different from [7]. We select the kernel function that can provide an independent value for each item. Item \( w_i \) is the \( i \)-th element of \( S \), and item \( w_j \) is the \( j \)-th element of \( S \).

\[
 \phi_{ij}(x, y) := \begin{cases} 1 & (x = w_i, y = w_j), \\ 0 & \text{otherwise}. \end{cases} \tag{7}
\]

By the definition of \( \phi_{ij}(x, y) \), we have the following formula.

\[
 \hat{r}_\alpha(w_i, w_j) = \sum_{i' = 1}^{v} \sum_{j' = 1}^{v} \alpha_{i'j'} \phi_{i'j'}(w_i, w_j) \tag{8}
\]

Now, we can define the cost function \( J_0(\alpha) \), which corresponds to some of the square errors for all pairs of items.

\[
 J_0(\alpha) := \frac{1}{2} \sum_{i = 1}^{v} \sum_{j = 1}^{v} (\hat{r}_\alpha(w_i, w_j) - r(w_i, w_j))^2 p(w_j) \tag{10}
\]

\[
 - \frac{1}{2} \sum_{i = 1}^{v} \sum_{j = 1}^{v} \hat{r}_\alpha(w_i, w_j)^2 p(w_j) 
 - \sum_{i = 1}^{v} \sum_{j = 1}^{v} \hat{r}_\alpha(w_i, w_j) p(w_i, w_j) + C, \tag{11}
\]

where

\[
 C := \frac{1}{2} \sum_{i = 1}^{v} \sum_{j = 1}^{v} r(w_i, w_j) p(w_i, w_j). \tag{12}
\]

As \( C \) is a constant to \( \alpha \), the minimization of \( J(\alpha) \) provides an identical solution with \( J_0(\alpha) \).

\[
 J(\alpha) := J_0(\alpha) - C \tag{13}
\]

\[
 = \frac{1}{2} \sum_{i = 1}^{v} \sum_{j = 1}^{v} \hat{r}_\alpha(w_i, w_j)^2 p(w_j) 
 - \sum_{i = 1}^{v} \sum_{j = 1}^{v} \hat{r}_\alpha(w_i, w_j) p(w_i, w_j). \tag{14}
\]

III. FORMALIZING THE PROBLEM

Suppose there is a database, whose record is a set of finite items. Our sampling space is records including future data. The number of such records is infinite. Our sample is the record in the database at present. The number of this record is finite. Items are defined to be the labels in a record. We will use the random variable for a record in the sampling space, and let \( x \) and \( y \) be the specified items. Our objective is to estimate \( P(x \in R_c | y \in R_c) \) from the sample and consider the estimated value as the strength of association rule \( x \leftarrow y \).

IV. PROPOSED METHOD

In this paper, we use \( f(x) := g(x) \) to define \( f \). To explain the proposed method, we define the following functions.
From the observations (or current database), we can estimate \( \hat{J}(\alpha) \) as follows.

\[
\hat{J}(\alpha) := \frac{1}{2} \sum_{i=1}^{u} \sum_{j=1}^{v} \hat{r}_{ij}(w_{ij}, w_{ij}) \frac{C(w_{ij})}{N} - \sum_{i=1}^{u} \sum_{j=1}^{v} \hat{r}_{ij}(w_{ij}, w_{ij}) \frac{C(w_{ij}, w_{ij})}{N},
\]

where \( N \) is the number of records in the observation, \( C(w_{ij}) \) is the number of records that contain \( w_{ij} \), and \( C(w_{ij}, w_{ij}) \) is the number of records that contain both \( w_{ij} \) and \( w_{ij} \).

Our optimization problem is to minimize \( \hat{J}(\alpha) \) with the L2 normalization term. In the following formula, \( \lambda \) should be a positive number. Considering that \( \hat{J}(\alpha) \) becomes small according to \( 1/N \), we consider the coefficients as \( \lambda/N \), and the value of \( \lambda \) will be a constant. The reason for introducing the normalization term is to avoid overfitting to the observed data. The problem is expressed as follows.

\[
\min_{\alpha \in \mathbb{R}^{u \times v}} \left[ \hat{J}(\alpha) + \frac{\lambda}{2N} \alpha^T \alpha \right] \tag{16}
\]

subject to \( \forall i, \forall j \, \alpha_{ij} \geq 0 \).

Fortunately, this problem exhibits an analytical solution because the solution with constraints is the same as the solution without constraints.

To obtain the solution without constraints, we use the following formula.

\[
\frac{\partial}{\partial \alpha_{ij}} \left( \hat{J}(\alpha) + \frac{\lambda}{2N} \alpha^T \alpha \right) = 0.
\]

(17)

It provides the solution without constraints.

\[
\alpha_{ij} \left( \frac{C(w_{ij})}{N} + \frac{\lambda}{N} \right) - \frac{C(w_{ij}, w_{ij})}{N} = 0,
\]

\[
\alpha_{ij} = \frac{C(w_{ij}, w_{ij})}{C(w_{ij}) + \lambda \cdot N},
\]

\[
\alpha_{ij} = \frac{C(w_{ij}, w_{ij})}{C(w_{ij}) + \lambda}.
\]

This solution satisfies the constraints. Therefore, this is the solution with constraints. In summary, we estimate the strength by using the following formula. The value of \( \lambda \) should be tuned using the learning dataset. It should also be noted that if \( \lambda = 0 \) then the system becomes MLE.

\[
\hat{r}_{\text{proposed}}(x, y) = \frac{C(x, y)}{C(y) + \lambda}.
\]

V. EXPERIMENTAL SETTING AND EVALUATION METHOD

For the experiment and evaluation, we selected \( S \) as a set of city and prefecture names. Each record of database corresponds to the symbols in \( S \) in a newspaper article of seven years (1991 – 1997). We constructed 14 databases that correspond to newspaper articles of each half year, and we labeled each database by its year and an additional character. TABLE I shows each period, number of the articles, number of the true rules, and number of the all of the pairs in each dataset. We assume that each newspaper article includes the description of several places of interest; thus, the name of the city or prefecture appears in the article with some probability. It should also be noted that many labels of personal names could also be the labels of places. This makes the finding association rules difficult or interesting. Moreover, as we can judge whether the observed city is in the observed prefecture, it is possible to judge whether the mined association rule is correct or not. As we have consensus for the relationship, the judgement agrees among all people.

As we do not have the correct answer for the relations among cities or among prefectures, we focus on the relation between a city and a prefecture. In other words, we compute \( \hat{r}(x, y) \), in the case where \( x \) is the city name, and \( y \) is the prefecture name. The output of the system is the ranked list \( (x, y) \) by \( \hat{r}(x, y) \). We eliminated the result in other cases, because we cannot determine whether the pair is correct or not. By sorting according to the estimated value, the system outputs the ranked list of relations. At each rank of the output, we compute the recall rate by a given value of rank. This metrics is also reference [9].

\[
\text{Recall} = \frac{\text{Number of true positive}}{\text{Number of true rules}}.
\]

As there are no systems that can detect the relationship for an unseen city and an unseen prefecture, the total number of true rules is the number of distinct cities times the number of distinct prefecture in the database used. By plotting the recall rate for each rank, we can obtain the graph in Fig. 1 which will be explained later. Moreover, we can compare the value of precision using the same graph. As precision is proportional to the slope toward the origination point, considering the number of total output is the definition of rank. This metrics is also reference [9].

\[
\text{Precision} = \frac{\text{Number of true positive}}{\text{Number of total output}}.
\]

| TABLE I | PROPERTIES OF THE DATASETS |
|---------|---------------------------|
| each period | articles | true rules | all of the pairs |
| 91(a) | 1/1/1991–6/30/1991 | 25,510 | 3,398 | 187,994 |
| 91(b) | 7/1/1991–12/31/1991 | 26,622 | 3,349 | 198,314 |
| 92(a) | 1/1/1992–6/30/1992 | 26,415 | 2,827 | 98,884 |
| 92(b) | 7/1/1992–12/31/1992 | 30,172 | 3,294 | 202,157 |
| 93(a) | 1/1/1993–6/30/1993 | 26,904 | 2,797 | 96,700 |
| 93(b) | 7/1/1993–12/31/1993 | 26,327 | 3,251 | 117,970 |
| 94(a) | 1/1/1994–6/30/1994 | 26,530 | 1,183 | 106,660 |
| 94(b) | 7/1/1994–12/31/1994 | 33,592 | 3,685 | 214,703 |
| 95(a) | 1/1/1995–6/30/1995 | 37,856 | 3,256 | 148,010 |
| 95(b) | 7/1/1995–12/31/1995 | 38,707 | 3,206 | 148,201 |
| 96(a) | 1/1/1996–6/30/1996 | 37,683 | 3,218 | 119,299 |
| 96(b) | 7/1/1996–12/31/1996 | 20,834 | 2,721 | 88,143 |
| 97(a) | 1/1/1997–6/30/1997 | 42,688 | 3,127 | 112,795 |
| 97(b) | 7/1/1997–12/31/1997 | 29,298 | 2,699 | 96,464 |
When the plots line of two systems cross each other, we compare the system by the recall rate at a certain rank, for example, 1000, 4000 or 12000, where we can observe the difference among the systems in our experiment.

We compared three systems: Apriori, additive smoothing, and the proposed method. Each system contains one parameter to tune. In this experiment, as we have several datasets, the parameter is decided to be tuned in a certain year. Moreover, the value is used for obtaining the recall rate of the database of another year.

The estimator used by Apriori [3] is as follows. In this formula, is the parameter of Apriori.

\[
\hat{r}_{apriori}(x, y) = \begin{cases} 
\frac{C(x, y)}{C(y)} & (C(x, y) > \theta), \\
0 & \text{otherwise.}
\end{cases}
\]

The estimator used by additive smoothing [5] is as follows. In this formula, \( \mu \) is the parameter and \( B \) is number of classes in the classification problem. Our experiment contains two classes: one is the case where the pair is a true rule and the other is the case where the pair is a false rule. Therefore, we set 2 to \( B \). In theory, additive smoothing assumes that the prior distribution is the uniform distribution. The parameter \( \mu \) corresponds to the confidence in the prior distribution. The value is the expected value of the posterior distribution. Thus, if \( \mu = 0 \), the system becomes Laplace smoothing [5].

\[
\hat{r}_{additive\_smoothing}(x, y) = \frac{(C(x, y) + \mu)}{(C(y) + \mu B)}.
\]

### A. Parameter Learning

We tested three systems: Apriori, additive smoothing, and the proposed method, that is, the direct estimator of the likelihood ratio. Let us call the proposed method “Direct”. As every system contains the same number of parameters, it is fair to compare with the same size of dataset.

We use the recall rate at rank 4000 for tuning the parameter. We call this condition as TOP-4000. Please note that this rate is also proportional to the precision rate at rank 4000 because both rates are proportional to the number of true positives, and the denominators of both ratios are constants at a fixed rank. The number 4000 is selected for the condition that the results reveal the modest but the largest recall rate.

TABLE II shows the values of the parameter for each system and each dataset. We noticed that the dataset “95(a)” shows different values from other datasets for all the three systems. This indicates that the distribution of the true rules in “95(a)” is different from other datasets. As it affects all the three systems, we require the care for the treatment of database “95(a)”.

It should be noted that the obtained parameter is not zero for all the cases. This indicates that all the three systems outperform the simple MLE, because all the three systems become MLE when the parameter is equal to zero.

### B. Evaluation Analysis

First, we computed the recall rates at TOP-1000 (small recall rate condition), TOP-4000 (modest recall rate condition), and TOP-12000 (recall rate-oriented condition), by learning the parameters from the dataset of the previous period. TABLE II shows the result. For each dataset, the underlined numbers show the best result system among proposed method, Apriori and additive smoothing. Among the three systems, the proposed method shows the best result. Compared with Apriori using one sided paired t-test, the \( \alpha \)-level of significance is 0.0002 in the TOP-4000 condition, and 0.002 in the TOP-12000 condition. Compared with additive smoothing, the \( \alpha \)-level of significance is extremely small in both the TOP-4000 and TOP-12000 conditions. In the TOP-1000 condition, the difference is not apparent because all the three systems behave similar to MLE, when \( C(x, y) \) is large.

Fig. 1 is the plot of recall rate at rank, which is a learned parameter in 91(a), and a tested parameter in 91(b). As the curves do not cross each other, the proposed method is recommended to be used for all range of ranks.

### C. Discussion

It is difficult to compare the proposed system with Predictive Apriori owing to the difference in the number of parameters. If we need to learn the prior distribution in the shape of a histogram, we require at least ten or more parameters. In
In this paper, we proposed a cost function corresponding to the mean square errors between estimated values and true values of conditional probability in a discrete distribution. We then obtained the values that minimized the cost function. This minimization approach can be considered as a direct estimation of likelihood ratios.

We compared the proposed method with Apriori using 14 datasets. By comparing the recall rates at TOP-4000, we can observe the modest but largest recall rate. We measured the recall rates at TOP-1000, TOP-4000, and TOP-12000 by using the previous year database for learning the parameter of each system. The average of recall rates at these conditions reveals that the proposed method outperforms Apriori. We can observe the statistical significance by using the one-sided t-test.

The optimal parameter value for dataset “95(a)” requires more examination. This is because it may provide a hint to reveal the hidden problem of our framework, and this is our future investigation. Moreover, we endeavor to use another method of smoothing for estimating this conditional probability in our future investigation.

### References

[1] Y. Nong, The Handbook of Data Mining. CRC Press, 2004.

[2] W. H. Ian, F. Eibe, and H. A. Mark, Data Mining: Practical Machine Learning Tools and Techniques, Third Edition. Morgan Kaufmann, 2011.

[3] R. Agrawal and R. Srikant, “Fast Algorithms for Mining Association Rules,” in Proceedings of the 20th International Conference on Very Large Data Bases, 1994, pp. 487–499.

[4] M. Kikuchi, E. Yamamoto, M. Yoshida, M. Okabe, and K. Umemura, “Using conservative estimation for conditional probability instead of direct importance estimation,” Journal of Machine Learning Research, vol. 9, no. 4, pp. 381–395, 2008.

[5] C. D. Manning and H. Schütze, Foundations of Statistical Natural Language Processing, Cambridge, MA: The MIT Press, 1999.

[6] T. Scheffer, “Finding association rules that trade support optimally against confidence,” Intelligent Data Analysis, vol. 9, no. 4, pp. 381–395, 2005.

[7] M. Sugiyama, I. Takeuchi, T. Suzuki, T. Kanamori, H. Hachiya, and D. Okanohara, “Least-squares conditional density estimation,” IEICE Transactions on Information and Systems, vol. E93-D, no. 3, pp. 583–594, 2010.

[8] M. Sugiyama, T. Suzuki, and T. Kanamori, Density Ratio Estimation in Machine Learning. Cambridge, Cambridge University Press, 2012.

[9] M. Kikuchi, E. Yamamoto, M. Yoshida, M. Okabe, and K. Umemura, “Using conservative estimation for conditional probability instead of ignoring infrequent case,” in Proceedings of the 3rd International Conference On Advanced Informatics: Concepts, Theory And Application (ICAICTA2016), 2016, pp. 1–5.

[10] L. Wasserman, All of Statistics: a concise course in statistical inference. Verlag, NY: Springer, 2004.