

Using Component Extraction Association Rules for Sensor Data

Di Dong and Rafael A. Calvo

School of Electrical and Information Engineering,
The University of Sydney
Sydney, Australia
<http://weg.ee.usyd.edu.au/index.html>

Abstract. Wearable sensors are being used to collect biomedical data in order to monitor human health or activities. Association rules represent a powerful technique used to discover hidden regularities among very large amounts of data. However, the number of discovered rules can be very large, especially for data sets whose attribute items are highly correlated and highly dimensioned, as is common with biomedical sensor data. In this paper, we present a novel method that we refer to as component extraction association rules (CEAR) to tackle this problem. Our evaluation results show that the final set of rules produced by CEAR is typically small enough to be analyzed manually by a human user. Besides generating rules, CEAR also provides Artificial Summary Factors, which are useful for interpreting the original data set.

Key words: Biomedical Sensors, Numerical Association Rule, Rule Filtering, Principal Component Analysis, Linear Regression

1 Introduction

A wearable biomedical sensor is a sensor device that is continually attached to a person in order to monitor and observe their physiological information. Currently, these devices are widely used in various contexts, such as ambulatory care, sports training, research and support for first responders. The data generated by these biomedical sensor systems normally has two characteristics. First, these data are usually highly dimensioned since human physiological activities are so complex that many attributes need to be recorded and monitored. Second, the attribute items of these data are highly autocorrelated as well as correlated among themselves, as is typical with physiological signals.

Association rules mining was introduced in [1]. An association rule is a relationship between attributes in the form $C_1 \Rightarrow C_2$, where C_1 and C_2 are a pair of conjunctions in the way $A = v$ if it is a categorical attribute or $A \in [v_1, v_2]$ if the attribute is numerical. The two most commonly used measures that filter only the most significant rules in a data set are defined as:

1. Support: A rule $C_1 \Rightarrow C_2$ has a support s , if an $s\%$ of the records in the data set contain both C_1 and C_2 .

2. Confidence: A rule $C_1 \Rightarrow C_2$ has a confidence c , if the $c\%$ of the records that contain C_1 in the data set also contains C_2 .

Currently, most association rule mining algorithms employ this support-confidence framework. With this framework, association rule mining has the ability to discover the set of associations that exists within the data [2]. As a result, association rules is an ideal tool to make biomedical sensor data more clear and interpretable to end users. However, due to the challenges described in the next section, association rule mining techniques must still be improved upon in order to be practical in this context.

In this paper, we present a novel Component Extraction Association Rule (CEAR) mining process that aims at solving these challenges. The rest of this paper is organized as follows: the second section reviews the current problems in association rule mining and the related research. In section 3, the details of the novel CEAR approach are described. Section 4 describes a practical application of the CEAR approach in the sports sciences context, which shows the feasibility of using the new method for a realistic application. Finally, section 5 concludes with a discussion of the results and of future work.

2 Background

The strength of association rules mentioned in the Introduction section comes with a major drawback. Normally, the number of discovered rules can be excessively large; easily in the thousands or tens of thousands. This problem is particularly true for biomedical sensor data sets because of their high dimensionality and highly correlated attributes. Obviously, finding interesting regularities from all rules holding in a data set manually, (e.g., with statistical methods) is time consuming if not impossible [3]. This situation massively increases the difficulty of presenting discovered regularities and using them to solve real-world problems. Research illustrates that the confidence of a rule is only an estimate of the conditional probability of item set B given item set A , as it does not measure the real strength (or lack of strength) of the implication. There is no support, from the statistical perspective, for the rules discovered by such a framework. So the support-confidence framework may identify a rule as interesting when, in fact, the occurrence of antecedent does not imply the occurrence of consequence.

Soon after the introduction of association rules, the issue of interestingness of discovered knowledge was put forward by Piatetsky-Shapiro [4]. The Hoschka and Klosgen [5] approach is based on a few fixed statement types and partial ordering of attributes, called templates. Influenced by Hoschka and Klosgen [3] introduced templates to a set of discovered rules. In general, a template is a form of pattern expressions that describe a set of rules by specifying what attributes occur in the antecedent and those attributes that are the consequent. So templates can be used to describe the form of interesting rules, and also to specify which rules are not interesting. In a formal definition, a template is an expression $A_1, A_2, \dots, A_k \Rightarrow A_{k+1}$, where each A_i is either an attribute item, a class name, or an expression C^+ or C^* , where C is a class name. Here C^+

and C^* corresponds to one or more and zero or more instances of the class C , respectively.

An association rule matches the pattern specified by a template, if the rule can be considered to be an instance of the pattern. With a template, users can explicitly specify both what is interesting and what is not. This method, however, does not prune those insignificant rules and does not provide a summary of the discovered rules. In subjective interestingness research in data mining, several researchers [6][7][8] have proposed several of methods for finding unexpected rules. Instead of asking the user to specify what he/she wants to see by using a template, these approaches ask a user to specify his/her existing knowledge about the domain. The system then finds those unexpected rules by comparing the user's knowledge with the discovered rules.

Srikant et al.[9] and Lakshmanan et al. [10] investigated using item constraints specified by users in rule mining to generate only the relevant rules. Essentially, the constraints restrict the items or combination of items allowed to participate in mined rules. This approach also does not prune those insignificant rules or summarize the rules not pruned.

Toivonen et. al [11] introduced the concept of a 'cover', which is basically a subset of the discovered associations that can cover the database. The number of rules in a cover can be quite small. A greedy algorithm is proposed to find a good cover and the remaining rules are pruned. The problem is that the advantage of association rules (completeness) is lost.

There are also a number of other methods in classification research for rule pruning such as pessimistic error rate and minimum description length-based pruning [12]. Another category of approaches employ the widely used chi-square test statistics for rule pruning. Correlation rule mining presented in [13] uses a chi-square test to measure the correlation. In [2], statistical correlation obtained by using the Chi-square test rather than minimum confidence was used as the basis for finding rules that represent the fundamental relations of the domain, as minimum confidence is not reflected in the data [14].

3 Component Extraction Association Rules

In figure 1, the entire process of the Component Extraction Association Rules (CEAR) mining algorithm is presented step by step.

3.1 Encode User's Interests in a template

Templates are employed to allow end-users to specify their interests in the consequent part of a produced rule. A template is defined by an expression:

$$A_1, A_2, \dots, A_k \Rightarrow Indicator$$

where *Indicator* is the interesting attribute specified by end-users and A_i is one or more of the rest of the attributes that are contained in the input data set. A rule that matches such an expression is presented to users.

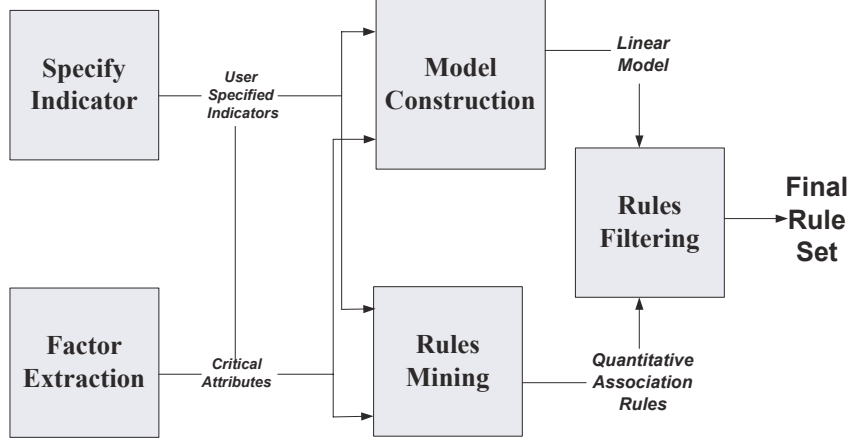


Fig. 1. CEAR Process

After obtaining the users' interests, the original data set needs to be split into two parts, indicator data set and attribute data set. The indicator data set contains only the indicators (dependent variables) defined by end-users through templates, while the attribute data set contains the rest of the attributes (independent variables) in the original data set.

3.2 Summary Factor & Critical Attributes Generation

We are going to summarize the attribute data set (independent variables) obtained in the last step with several of the most significant principal components (summary factor), so Principal Component Analysis (PCA) [21] is employed here. In mathematical terms, n correlated random variables are transformed into a set of $d < n$ uncorrelated variables. These uncorrelated variables, called principal components, are linear combinations of the original variables and can be used to express the data in a reduced form, and then find critical attribute items that are original variables with significantly large loadings in principal components. All the variables are standardized so all column means will equal zero and standard deviations will equal one.

We use the classical matrix PCA approaches [15] [16] to extract the summary factors. The goal of such methods is to find the eigenvectors $\{\alpha_i | i \in [1, n]\}$ of the original data set's correlation matrix. These eigenvectors stand for the directions of the principal components, and their statistical significance is described by the corresponding eigenvalue λ_i . These eigenvectors are then sorted according to their corresponding eigenvalue, from highest to lowest. The most obvious criterion for choosing d , is to select a cumulative percentage of total variance which is desired that the selected principal components should contribute. The required number of new components is then the smallest value of d for which this

chosen percentage has then exceeded. The portion of variance that is explained by the first d principal components, EV_d is

$$EV_d = \frac{\sum_{i=1}^d (\lambda_i)}{Tr(C)} = \frac{\sum_{i=1}^d (\lambda_i)}{\sum_{i=1}^n (\lambda_i)}$$

3.3 Linear Model Construction

At this stage, we try to construct a linear model that can summarize the quantitative relationship between user-specified indicators (dependent variables, Y) and summary factors obtained from the attribute data set (independent variables) in the last phase.

We use standard multiple linear regression (MLR) analysis [17], which finds a set of partial regression coefficients b_j such that the dependent variable Y can be approximated as well as possible by a linear combination of the independent variables $X_1, \dots, X_k, \dots, X_K$. The partial regression coefficients are found by using Ordinary Least Squares (OLS) [19]. Normally, the MLR equation is expressed with matrix notation. The dependent and independent variables are denoted with the matrices in Figure 2.

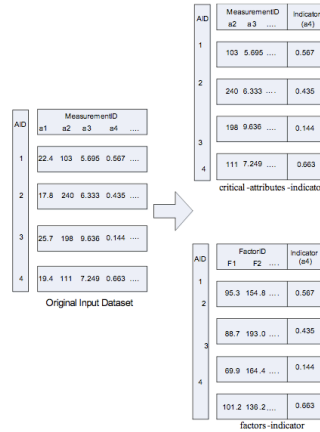


Fig. 2. Structure of X and Y matrices

So the predicted values of the dependent variable are collected in a vector denoted y' , which can be calculated using MLR as:

$$Y' = Xb \quad \text{with} \quad b = (X^T X)^{-1} X^T Y$$

Since the regression model is built from samples, we conduct hypothesis tests to assess whether the independent variables help explain the dependent variable. To address this question, we test the null hypothesis that the entire slope coefficients b_1, \dots, b_K are simultaneously equal to 0. This is because if none of the

independent variables in a regression model help explain the dependent variable, the slope coefficients b_k should all be equal to 0. Since the individual tests do not account for the effects of interactions among the independent variables, we can not test the null hypothesis that all slope coefficients equal to 0 based only on t -tests. To test the null hypothesis ($H_0 : b_1 = b_2 = \dots = b_k = 0$) against the alternative hypothesis that at least one slope coefficient is not equal to 0, we conduct an F -test, which is the ratio of the mean regression sum of squares to the mean squared error, which are the mean value of regression sum of squares (RSS) and squared error (SSE). For RSS, the degree of freedom is equal to the number of slope coefficient estimated, K ; while for SSE, it is equal to the number of observations, n , minus $(K + 1)$, as there are a total of $k + 1$ coefficients (k slope coefficient plus one intercept, b_0) needed to be estimated during the construction of the result model. So the F -statistic is constructed as follows:

$$F = \frac{MSR}{MSE} = \frac{\frac{RSS}{K}}{\frac{SSE}{n-(k+1)}}$$

where $RSS = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$ and $SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n \hat{\varepsilon}_i^2$. The statistical value calculated using the formula above needs to be compared with the critical value of F -distribution at the certain significance level, p . Here we set the significance level for this test to $p = 0.05$, rejecting the null hypothesis if the calculated value of F is greater than the upper 0.05 critical value of the F distribution, with the specified numerator and denominator degrees of freedom k and $n - (k + 1)$. The reason that we only use a one-tailed F -test here is because MSR necessarily increases relative to MSE as the explanatory power of the regression increases. We only use the linear model once it passes the F -test.

Like the other statistical techniques, MLR relies upon certain assumptions about the variables used in the analysis. In order to make a valid inference from a multiple linear regression model, the following five assumptions [20] need to be made:

1. The relationship between the dependent variable, Y , and the independent variables, X_1, X_2, \dots, X_k , is linear.
2. There is no exact linear relation existing between two or more of the independent variables X_1, X_2, \dots, X_k (No multicollinearity).
3. The expected value of the error term, conditioned on the independent variables, is 0: $E(\varepsilon | X_1, X_2, \dots, X_k) = 0$.
4. The variance (standard deviation) of the error term is constant for all observations: $E(\varepsilon_i^2) = \sigma_\varepsilon^2$.
5. The error term is normally distributed.

In our case, the K is equal to the number of summary factors, d . As a result, the linear relationships that are contained in the original input data set will be

concluded in a formula as follows:

$$\begin{aligned} f(x) &= b_0 + b_1PC_1 + b_2PC_2 + \dots + b_dPC_d \\ &= b_0 + b_1(W_{11}A_1 + W_{12}A_2 + \dots + W_{1n}A_n) \\ &\quad + b_2(W_{21}A_1 + W_{22}A_2 + \dots + W_{2n}A_n) + \dots \\ &\quad + b_d(W_{d1}A_1 + W_{d2}A_2 + \dots + W_{dn}A_n) \end{aligned}$$

3.4 Mining Rules

Instead of mining association rules from the original input data set, we find the rules that satisfy both of the following two requirements:

1. Only include the user-specified indicators, Y , in the consequent part of the rule.
2. Only include either the critical attribute items or combinations.

According to the approaches of finding quantitative rules (described in [18]) the mining process consists of three steps:

1. Disperse original continuous numerical value domain into equal-length value domains according to the formula as follows:

$$NumberOfIntervals = \frac{2 \times n}{m \times (K - 1)}$$

where $n = NumberOfQuantitativeAttributes$, $m = MinimumSupport$ and $K = PartialCompletenessLevel$.

2. Map original attributes' numeric value to the discontinuous value domains found in the previous step.
3. Use normal nominal association rule mining approaches to generate rules.

3.5 Filtering Rules using a Model

In the last step, we construct the filtering criterion based on the quantitative linear model that was built in the model construction phase. The filtering process consists of two steps. First, the general trend of result rules will be compared to the linear relationship (positive or negative) shown by the regression model. If their trend doesn't match the linear relationship, the result rules will be discarded. For example, if two rules show that user-specified indicators decrease with the increase of some critical attribute or their combination, but the linear relationship indicated by the result regression model is positive, then these two rules will be removed from the result set. In the second part, we calculate a regression confidence:

$$RegressionConfidence(X \Rightarrow Y) = Support(X \cup Y') / Support(X)$$

Where Y' stands for the predicated values produced by the regression model, which falls in the same value range of the original Y . The regression-confidence

accounts for the probability of finding the consequent part of a rule in the regression predicted values under the condition that the antecedent part of a rule has been satisfied. Then we compare the confidence of a rule with its regression-confidence. If the regression-confidence is larger than the normal confidence, it means that the probability that the consequent part will be true given the condition that the antecedent part of the rule has been satisfied can be fully supported by the regression model. On the other hand, if the regression-confidence is less than the normal confidence, it means the regression model cannot approve all the probabilities we just mentioned. The extra probability beyond the regression-confidence may be produced by noise or outliers. So we only present a rule that has a larger regression-confidence value than its normal confidence values in the final result rule set.

4 Evaluation

The CEAR mechanism was evaluated using a set of rowing performance data from the NSW Institute of Sport, Australia. This set of data was collected by a wearable biomedical sensor system, named RowSys2.

We evaluated the first point by conducting a discussion on the final rule set with the domain experts. Interesting interpretations have been made on several result rules from the perspectives of biomedical and biomechanics, however, these interpretations are beyond the scope of this paper.

Beyond looking at the value change of support and confidence, we focus on the effect of the required accumulated variance explained by principal components to the final result rules. Generally, with the increase of the required accumulated variance, the number of principal components required will increase as will the number of critical attribute items. This increases the search space, generating frequent item sets as more attributes are involved. As a result, the size of the final rule set may increase as well. However, by selecting a different value for the required accumulated variance, we found that the change of this value doesn't have much effect on increasing the size of the result rule set, as shown in Figure 3.

This is because as more principal components (summary factors) are involved, the explanation power of the regression model increases. As a result, the regression model may be more powerful when identifying temporary associations. So the increase in result rules size caused by more critical attributes involved may be offset, as more temporary rules will be filtered by the regression model.

5 Conclusion and Future Work

In this paper, we present the component extraction association rule mining algorithm, which aims to deal with the large-size problem of final rule sets when finding association rules in biomedical sensor data sets. Compared to the common association rule mining mechanisms, such an approach has the following advantages:

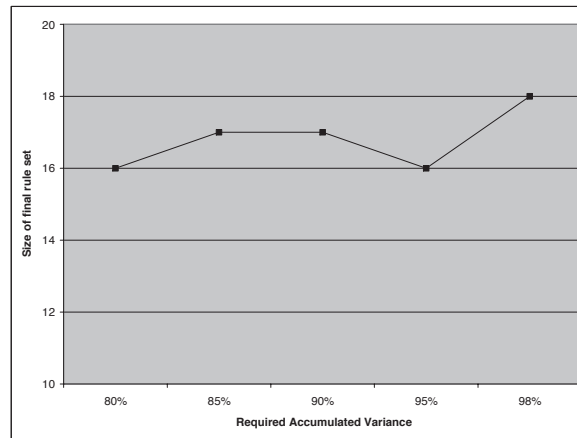


Fig. 3. Structure of X and Y matrices

1. Reduces the computational and memory resource required to generate a frequent item set, as the attributes involved only contain user-specified indicators and critical attributes, rather than all attributes in the original data set.
2. Improves the interestingness of the discovered rules with user-specified interests and statistical significance.
3. Provides other patterns that may help to illustrate the regularities hiding in the original data set, such as Summary Factor and Linear Quantitative Model.

References

1. Agrawal, R., Imielinski, T., Swami, A.: Mining association rules between sets of items in large databases. In: ACM SIGMOD International Conference on Management of Data, pp. 207–216. ACM Press, Washington DC, USA (1993)
2. Liu, b., Hsu, W., Ma, Y.: Pruning and Summarizing the Discovered Associations. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 125–124. ACM Press, San Diego, USA (1999)
3. Klemettinen, M., Mannila, H., Ronkainen, P., Toivonen, H., Verkamo, A. I.: Finding interesting rules from large sets of discovered association rules. In: Conference on Information and Knowledge Management, ACM Press, Maryland, USA (1994)
4. Piatesky-Shapiro, G.: Discovery, analysis, and presentation of strong rules. In: Piatesky-Shapiro, G., Frawley, W. (eds) Knowledge Discovery in Databases. pp. 229–248. AAAI Press/The MIT Press (1991)
5. Hoschka, S., Klosgen, W. : A support system for interpreting statistical data. In: Piatesky-Shapiro, G., Frawley, W. (eds) Knowledge Discovery in Databases. pp.325–345. AAAI Press/The MIT Press (1991)
6. Silberschatz, A., Tuzhilin, A.: What makes patterns interesting in knowledge discovery systems. IEEE Transactions on Knowledge and Data Engineering. vol. 8, No. 6, 970–974 (1996)

7. Liu, B., Hsu, W. : Post-analysis of learned rules. In: 13th National Conference on Artificial Intelligence, pp. 828–834. AAAI Press, Oregon, USA (1996)
8. Liu, B., Hsu, W., Chen, S.: Using general impressions to analyze discovered classification rules. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 234–243. ACM Press, California, USA (1997)
9. Srikant, R., Vu, Q., Agrawal, R.: Mining association rules with item constraints. In: ACM SIGKDD International Conference on Knowledge Discovery and Data mining, pp. 453–462. ACM Press, California, USA (1997)
10. Lakshmanan, Ng. R. T., Han, J.: Exploratory mining and pruning optimizations of constrained association rules. In: ACM SIGMOD International Conference on Management of data, pp. 541–549. ACM Press, Seattle, USA (1998)
11. Toivonen, H., Klemetinen, M., Ronkainen, P., Hatonen, K., Mannila, H.: Pruning and grouping discovered association rules. In: Mlnet Workshop on Statistics, Machine Learning, and Discovery in Databases, pp. 47–52. (1995)
12. Mahta, M., Agrawal, R., Rissanen, J.: SLIQ: A fast scalable classifier for data mining. In: International Conference on Extending Database Technology, pp. 278–285. ACM Press, Avignon, France (1996)
13. Brin, S. Motwani, R., Silverstein, R. (1997): Beyond market basket: generalizing association rules to correlations. In: ACM SIGMOD International Conference on Management of Data, pp. 265–276. ACM Press, Arizona, USA (1997)
14. Bayardo, R., Agrawal, R, Gunopulos, D.: Constraint-based rule mining in large, dense databases. In: IEEE International Conference on Data Engineering, IEEE Press, Sydney, Australia (1999)
15. Rencher, A. C.: Methods of multivariate analysis. Wiley, New York (1995)
16. Tabachnick, B. G., Fidell, L. S.: Using Multivariate Statistics. Allyn & Bacon, New York (2000)
17. Darlington, D.B.: Regression and linear models. Wiley, New York (1990)
18. Srikant, R., Agrawal, R.: Mining quantitative association rules in large relational tables. In: ACM SIGMOD International Conference on Management of Data, pp. 665–664. ACM Press, Montreal, Canada (1996)
19. Rao, C.R., Toutenburg, H., Fieger, A., Heumann, C., Nittner, T., Scheid, S.: Linear Models: Least Squares and Alternatives. Springer Series in Statistics. Springer, New York, USA (1999)
20. Draper, N.R., Smith, H.: Applied Regression Analysis. Wiley Series in Probability and Statistics. (1998)
21. Jolliffe, I.T.: Principal Component Analysis. Springer Series in Statistics, Second Edition. Springer, New York, USA (2002)