

# EVALUATION OF ASSOCIATION RULES BASED ON CERTAINTY FACTOR: AN APPLICATION ON DIABETES DATA SET


M. Kivrak, F.B. Akcesme, and C. Colak


**Abstract**— Data mining is the process of discovering useful information that has not been previously revealed from large amounts of data. Association rules mining is one of the most important techniques used in data mining and artificial intelligence. The first research in the association rules was to find relationships between different products in the customer transaction database and customer purchase models. Based on these relationships, researchers have begun to expand the field of data mining. One of these areas is the application of the rules of association in the field of medicine. Thus, through these applications, the relationship of various features in medical data can be discovered, and the findings obtained can aid medical diagnosis. Support and confidence are the two primary measures employed in the evaluation of association rules. The rules obtained with these two values are often correct; however, they are not strong rules. For this reason, there are many interestingness measures proposed to achieve stronger rules. Most of the rules, especially with a high support value, are misleading. For this reason, there are many interestingness measures proposed to achieve stronger rules. This study aims to establish strong association rules with variables in the open-sourced diabetes data set. In the current study, the Apriori algorithm was used to obtain the rules. As a result of the analysis, only 52 confidence and support criteria were taken into consideration. For more powerful rules, certainty factor was used as one of the interestingness measures proposed in the literature, and it was concluded that only 39 of these rules were strong as a result of the analysis.


**Keywords**— *Machine learning, classification, artificial neural network, support vector machines, decision tree, logistic regression, linear discriminant, nearest neighbor.*

## 1. INTRODUCTION

THE data mining is an important process of extracting previously unknown and useful information from data in databases [1]. Data mining techniques include classification, prediction, associations, and clustering. One of the most important data mining applications is that of mining association rules [2]. Initial research in association rules is about finding relationships between different products in the customer transaction database, as well as customer purchase models.

**Mehmet Kivrak**, Dept. of Biostatistics and Medical Informatics Faculty of Medicine, Inonu University, Malatya, Turkey, (e-mail: mehmetkivrak83@gmail.com) 

**F. Berat Akcesme**, Dept. of Biostatistics and Medical Informatics Faculty of Medicine, University of Health Sciences, Istanbul, Turkey, (e-mail: farukberat.akcesme@sbu.edu.tr) 

**Cemil Colak**, Dept. of Biostatistics and Medical Informatics Faculty of Medicine, Inonu University, Malatya, Turkey, (e-mail: cemil.colak@inonu.edu.tr) 

Based on these relationships, researchers had begun to expand the field of data mining.

One of these areas is the application of association rules in the medical field. Thus, through these applications, the association of various attributes can be discovered in medical data, and this can help medical diagnosis [3]. This study aims to establish strong association rules with variables in the open-sourced diabetes data set.

## 2. MATERIAL AND METHODS

### 2.1. Data Set

The dataset used for the analysis was obtained from the website (<http://datahub.i.o>machine-learning>diabetes>) [4]. The data set contains 768 samples and nine variables. These variables are age, pregnancies, PG concentration, diastolic BP, tri-fold thick, resting electrocardiographic results, serum ins, BMI, DP function, and diabetes. The detailed explanation of the variables are given in Table I. Data analysis was performed by using RStudio Version 1.1.463 programming language.

### 2.2. Knowledge Discovery in Databases (KDD)

In the process of KDD, data selection (diabetes dataset), data preprocessing (extreme and missing value analyses), data transformation (normalization, etc.), data mining and evaluation, and interpretation of the results were performed.

### 2.3. Association Rules Mining

Association rules mining is a very common technique that can define various rules or relationships between variables. [5]. These association rules are composed of two item sets, the antecedent (left-hand side) and consequent (right-hand side), the expression of the form is  $X \Rightarrow Y$ , where X and Y are called antecedent and consequent of the rule respectively [6,7]. There are many algorithms used in association rules such as the Apriori algorithm, Eclat algorithm, and FP-growth algorithm. The most classic and frequently used algorithm is the Apriori algorithm, and it used to find all frequent item sets in a given database [8].

#### 2.3.1. Basic Measures

The support and confidence are basic measures in building strong association rules from the frequent item sets [9].

Support is defined as the probability that transactions in the database contains items both the antecedent and the consequent of the rule, as follows.

TABLE I.  
THE DETAILED EXPLANATION OF THE VARIABLES  
DATA SET

| Abbreviation     | Explanation  |
|------------------|--|
| Age              | Age  |
| Pregnancies      | Number of pregnancy                                      |
| PG Concentration | Plasma glucose in 2 hours in oral glucose tolerance test |
| Diastolic BP     | Diastolic blood pressure (mm Hg)                         |
| Tri-Fold Thick   | Three-layer binding layer thickness (mm)                 |
| Serum Ins        | 2 Hour serum insulin (mu U / ml)                         |
| BMI              | (weight in kg / kg (height in m) ^ 2)                    |
| DP Function      | Diabetes family tree function                            |
| Diabetes         | Diabetes (1 = yes; 0 = no)                               |

Support ( $X \Rightarrow Y$ ) = (Transactions containing both X and Y items) / Total number of transactions.

Confidence is a measure that reveals the association between antecedent and consequent of a rule.

Confidence ( $X \Rightarrow Y$ ): Total number of transactions containing items X and Y divided by the total number of transactions containing item X.

Confidence ( $X \Rightarrow Y$ ) = Support (X, Y) / Support(Y) [5].

If a minimum threshold for support is chosen low, large numbers of rules are created, and assessment of such rules is rather complex and time-consuming. Also, selecting the minimum threshold value high causes some rules to be skipped [5]. Therefore, some interestingness measures (IM) are developed to solve this problem [9].

### 2.3.2. Interestingness Measure (IM)

Certainty Factor (CF) is interpreted as a measure of the variation of the probability that Y is in a transaction when we consider only those transactions where X is defined as follows [7].

$$CF(X \Rightarrow Y) = (\text{Conf}(X \Rightarrow Y) - \text{Supp}(Y)) / (1 - \text{Supp}(Y))$$

If  $\text{Conf}(X \Rightarrow Y) > \text{Supp}(Y)$ ,

and,

$$CF(X \Rightarrow Y) = (\text{Conf}(X \Rightarrow Y) - \text{Supp}(Y)) / \text{Supp}(Y)$$

If  $\text{Conf}(X \Rightarrow Y) < \text{Supp}(Y)$ , and 0 otherwise.

## 1. EXPERIMENTAL RESULTS

Since the variables of age, pregnancies, PG concentration, diastolic BP, tri-fold thick, serum ins, BMI, and DP function were continuous variables on the data set, they were transformed into categorical variables by using the program.

For the experimental results, the minimum support value was 1.5 percent (1.5%), and the confidence value was 80 percent (80.0%).

As a result of the analysis, it was found that 52 rules were formed respectively, which consisted of triple, quadruple, and

quintile association rules, which were observed the most as of confidence and support values. It is necessary to take care of the high confidence and support values, which are among the measures used in the interpretation of the obtained rules [10].

Interpretation of the rules obtained only with the measures of confidence and support led to the obtaining of correct but not strong rules. There are many interestingness measures used in the literature to obtain stronger rules. In this context, in addition to the confidence and support measures, the certainty factor was used. Certainty factor approaching 1 indicates that the rules are identified with high accuracy [7]. For this reason, it was concluded that only 39 of the 52 rules that could provide a relationship between the diagnosis and other variables of the patient were strong. The rules obtained are given in Table II. Also, explanations of the variables used in Table II are given in Table I.

The three examples of the interpretation of the rules in Table II are given below:

- Rule 1: Patients with a Pregnancies [10], not having a tri-fold thick, do not have serum insulin with a probability of 100 %.
- Rule 2: Patients with a tri-fold thick [0] and between the age of [27] do not have serum insulin with a probability of 100 %.
- Rule 7: Patients with a Pregnancies[1] and between the age of [27] do not have diabetes with a probability of 0.96%.

Other rules achieved from the association rules mining are similarly interpreted in the related table.

## 2. DISCUSSION

Association rule mining is to discover association rules that satisfy a given database with predefined minimum support and confidence. The problem is usually broken down into two sub problems. One is to find those item sets whose occurrences in the database surpass a predefined threshold; such item sets are called regular or large item sets. The second problem is to create association rules with the minimum confidence constraints from those broad item sets [11].

Interpreting association rules by using only confidence and support measures might create many disadvantages; therefore, it will be more accurate to assess it with certainty factor, proposed by Shortliffe and Buchanan (1984) [12]. Similar to this study, Berzal et al. (2002) found that there were some disadvantages of evaluating association rules just with the measures of confidence and support; and also suggested that especially items with a very high support value might cause misleading rules [13].

There are many alternative interest measures to reach stronger rules in association rules mining. In this study, the Apriori algorithm was used to obtain association rules with open-sourced diabetes data set. In the application, the confidence factor and the rules with high support values are obtained on the right side of the rule. For that reason, in addition to the confidence and support measures, the certainty factor, which is one of the interestingness measures, was used to define the interesting rules in this research.

TABLE II.  
THE GENERATED ASSOCIATION RULE

| Row num. | Rule num. | Association Rule (X ⇒Y)   | Confidence | Support | Certainty Factor |
|----------|-----------|---|------------|---------|------------------|
| 1        | 1         | {Pregnancies=10,Tri.Fold.Thick=0} => {Serum.Ins=0}              | 1          | 0.0156  | 1                |
| 2        | 2         | {Tri.Fold.Thick=0,Age=27} => {Serum.Ins=0}                      | 1          | 0.0156  | 1                |
| 3        | 3         | {Diastolic.BP=0,Tri.Fold.Thick=0} => {Serum.Ins=0}              | 0.937      | 0.0195  | 0.911            |
| 4        | 4         | {Diastolic.BP=0,Diabetes=yes} => {Serum.Ins=0}                  | 1          | 0.0429  | 1                |
| 5        | 5         | {Diastolic.BP=0,Diabetes=no} => {Serum.Ins=0}                   | 0.942      | 0.0429  | 0.918            |
| 6        | 6         | {Pregnancies=8,Tri.Fold.Thick=0} => {Serum.Ins=0}               | 0.947      | 0.0234  | 0.925            |
| 7        | 7         | {Pregnancies=1,Age=23} => {Diabetes=no}                         | 0.965      | 0.0208  | 1                |
| 8        | 8         | {Diastolic.BP=76,Tri.Fold.Thick=0} => {Serum.Ins=0}             | 1          | 0.0247  | 1                |
| 9        | 10        | {Diastolic.BP=78,Tri.Fold.Thick=0} => {Serum.Ins=0}             | 1          | 0.0234  | 1                |
| 10       | 12        | {Pregnancies=6,Tri.Fold.Thick=0} => {Serum.Ins=0}               | 1          | 0.0169  | 1                |
| 11       | 13        | {Diastolic.BP=74,Tri.Fold.Thick=0} => {Serum.Ins=0}             | 1          | 0.0182  | 1                |
| 12       | 14        | {Diastolic.BP=70,Tri.Fold.Thick=0} => {Serum.Ins=0}             | 1          | 0.0195  | 1                |
| 13       | 15        | {Pregnancies=5,Tri.Fold.Thick=0} => {Serum.Ins=0}               | 1          | 0.0182  | 1                |
| 14       | 17        | {Pregnancies=4,Tri.Fold.Thick=0} => {Serum.Ins=0}               | 1          | 0.0208  | 1                |
| 15       | 21        | {Pregnancies=2,Tri.Fold.Thick=0} => {Serum.Ins=0}               | 1          | 0.0273  | 1                |
| 16       | 22        | {Pregnancies=0,Tri.Fold.Thick=0} => {Serum.Ins=0}               | 1          | 0.0182  | 1                |
| 17       | 23        | {Pregnancies=1,Tri.Fold.Thick=0} => {Serum.Ins=0}               | 1          | 0.0169  | 1                |
| 18       | 24        | {Tri.Fold.Thick=0,Diabetes=yes} => {Serum.Ins=0}                | 1          | 0.0325  | 1                |
| 19       | 28        | {Pregnancies=6,Tri.Fold.Thick=0,Diabetes=no} => {Serum.Ins=0}   | 1          | 0.0156  | 1                |
| 20       | 30        | {Pregnancies=4,Tri.Fold.Thick=0,Diabetes=yes} => {Serum.Ins=0}  | 1          | 0.0351  | 1                |
| 21       | 31        | {Pregnancies=4,Tri.Fold.Thick=0,Diabetes=no} => {Serum.Ins=0}   | 1          | 0.0234  | 1                |
| 22       | 33        | {Pregnancies=2,Tri.Fold.Thick=0,Diabetes=no} => {Serum.Ins=0}   | 1          | 0.0156  | 1                |
| 23       | 34        | {Pregnancies=0,Tri.Fold.Thick=0,Diabetes=no} => {Serum.Ins=0}   | 0.967      | 0.0234  | 1                |
| 24       | 35        | {Pregnancies=1,Age=21} => {Diabetes=no}                         | 1          | 0.0299  | 1                |
| 25       | 37        | {Diastolic.BP=0,Serum.Ins=0,Diabetes=no} => {Tri.Fold.Thick=0}  | 1          | 0.0416  | 1                |
| 26       | 38        | {Diastolic.BP=0,Serum.Ins=0} => {Tri.Fold.Thick=0}              | 1          | 0.0208  | 1                |
| 27       | 39        | {Diastolic.BP=0,Diabetes=yes} => {Tri.Fold.Thick=0}             | 1          | 0.1145  | 1                |
| 28       | 40        | {Diastolic.BP=0,Serum.Ins=0,Diabetes=yes} => {Tri.Fold.Thick=0} | 1          | 0.1809  | 1                |
| 29       | 41        | {Pregnancies=1,Age=24} => {Diabetes=no}                         | 1          | 0.0195  | 1                |
| 30       | 42        | {Pregnancies=2,Age=21} => {Diabetes=no}                         | 0.937      | 0.0195  | 0.911            |
| 31       | 43        | {Pregnancies=1,Age=22} => {Diabetes=no}                         | 1          | 0.0234  | 1                |
| 32       | 44        | {Serum.Ins=0,Age=21} => {Diabetes=no}                           | 0.947      | 0.0234  | 0.925            |
| 33       | 45        | {Serum.Ins=0,Age=24} => {Diabetes=no}                           | 1          | 0.0169  | 1                |
| 34       | 46        | {Diastolic.BP=68,Serum.Ins=0} => {Diabetes=no}                  | 1          | 0.0208  | 1                |
| 35       | 47        | {Pregnancies=2,Age=25} => {Diabetes=no}                         | 1          | 0.0182  | 1                |
| 36       | 49        | {Diastolic.BP=60,Serum.Ins=0} => {Diabetes=no}                  | 1          | 0.0169  | 1                |
| 37       | 50        | {Pregnancies=4,Serum.Ins=0,Diabetes=yes} => {Tri.Fold.Thick=0}  | 1          | 0.0169  | 1                |
| 38       | 51        | {Pregnancies=2,Serum.Ins=0} => {Diabetes=no}                    | 1          | 0.0234  | 1                |
| 39       | 52        | {Pregnancies=8,Diabetes=no} => {Serum.Ins=0}                    | 1          | 0.0273  | 1                |

### 3. CONCLUSION

As a result of this research, in addition to the 52 rules obtained with confidence and support measures, in order to eliminate the misleading rules and to obtain stronger rules, 26 rules were obtained by using certainty factors. In addition to the basic measures, it is recommended that different interestingness measures should also be used to reach more accurate results.

### ACKNOWLEDGMENT

The study was reported as oral presentation in 1st International Data Science Congress in Health on 05-06 December 2019.

### REFERENCES

- [1] M.-S. Chen, J. Han, and P. S. Yu, "Data mining: an overview from a database perspective," *IEEE Transactions on Knowledge and data Engineering*, vol. 8, no. 6, pp. 866-883, 1996.
- [2] M. Ilayaraja and T. Meyyappan, "Mining medical data to identify frequent diseases using Apriori algorithm," in *2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering*, 2013, pp. 194-199: IEEE.
- [3] W.-J. Zhang, D.-L. Ma, and B. Dong, "The automatic diagnosis system of breast cancer based on the improved Apriori algorithm," in *2012 International Conference on Machine Learning and Cybernetics*, 2012, vol. 1, pp. 63-66: IEEE.
- [4] D. Dua and C. J. C. a. C. U. Graff, "UCI machine learning repository [<http://archive.ics.uci.edu/ml>]. <https://archive.ics.uci.edu/ml/datasets>," 2019.
- [5] S. Kumar and N. Joshi, "Rule power factor: a new interest measure in associative classification," *Procedia Computer Science*, vol. 93, pp. 12-18, 2016.
- [6] S. Rao and P. Gupta, "Implementing Improved Algorithm Over APRIORI Data Mining Association Rule Algorithm 1," 2012.
- [7] F. Berzal, I. Blanco, D. Sánchez, and M.-A. Vila, "Measuring the accuracy and interest of association rules: A new framework," *Intelligent Data Analysis*, vol. 6, no. 3, pp. 221-235, 2002.
- [8] D. Jain and S. Gautam, "Implementation of apriori algorithm in health care sector: a survey," *International Journal of Computer Science and Communication Engineering*, vol. 2, no. 4, pp. 22-8, 2013.
- [9] J. Manimaran and T. Velmurugan, "Analysing the quality of association rules by computing an interestingness measures," *Indian Journal of Science and Technology*, vol. 8, no. 15, pp. 1-12, 2015.
- [10] O. Başak, B. Uğur, and M. K. SAMUR, "Kulak Burun Boğaz Epikriz Notlarından Birlikte Kurallarının Çıkarılması," 2009.
- [11] S. Kotsiantis, D. J. G. I. T. o. C. S. Kanellopoulos, and Engineering, "Association rules mining: A recent overview," vol. 32, no. 1, pp. 71-82, 2006.
- [12] W. van Melle, E. H. Shortliffe, and B. G. J. R.-b. e. s. T. M. e. o. t. S. H. P. P. Buchanan, "EMYCIN: A knowledge engineer's tool for constructing rule-based expert systems," pp. 302-313, 1984.

- [13] F. Berzal, I. Blanco, D. Sánchez, and M.-A. J. I. D. A. Vila, "Measuring the accuracy and interest of association rules: A new framework," vol. 6, no. 3, pp. 221-235, 2002.

### BIOGRAPHIES

**Mehmet Kıvrak** obtained his BSc degree in statistics from Dokuz Eylül University (DEU) in 2001. He received the BSc. and MSc. Diploma in Statistics from Dokuz Eylül University in 2001 and 2006 respectively, and Ph.D. degrees in the Graduate Department of Biostatistics and Medical Informatics of Inonu University in 2017. He accepted as an expert statistician in the Turkish Statistical Institute in 2009. His research interests are data mining, cognitive systems, reliability and genetics and bioengineering, and signal processing. His current research interests are genetics, bio engineering and data mining.

**F. Berat Akçeşme** obtained his BSc degree in Biological Sciences and Bioengineering from the International University of Sarajevo (IUS) in 2009. He pursued his master study at Mediterranean Agronomic Institute of China in the field of Horticultural Genetics and Biotechnology. In 2012, he got accepted to the Ph.D program in Genetics and Bioengineering at IUS where he was working as a research assistant at the same department. He obtained his PhD degree in 2016. He continued to work at IUS as an assistant professor until the end of 2017. His research interests are cognitive systems, bioinformatics, structural bioinformatics. In 2017, he joined the Department of Biostatistics and Medical Informatics at the Faculty of Medicine, University of Health Sciences as an assistant professor. He is active in teaching and research in the genetics and bioengineering. Beside, he is director of Bioinformatics and Biostatistics Application and Research Center at University of Health Sciences.

**Cemil Çolak** obtained his BSc. Degree in statistics from Ondokuz Mayıs University in 1999. He received MSc. Diploma in statistics from the Inonu University in 2001, and Ph.D. degrees in the Graduate Department of Biostatistics and Medical Informatics of Ankara University in 2005. He accepted as a Postdoctoral Researcher by the department of biostatistics and medical informatics of Inonu University in 2007. His research interests are cognitive systems, data mining, reliability, and biomedical system, and genetics, and bioengineering. In 2016, he joined the Department of Biostatistics and Medical Informatics at Inonu University as a professor, where he is presently a professor. He is active in teaching and research in the general image processing and data mining modeling, analysis.