Türkiye Klinikleri Biyoistatistik Dergisi
*Turkiye Klinikleri Journal of Biostatistics* Turkiye Klinikleri J Biostat. 2024;16(1):38-46

ORIGINAL RESEARCH ORİJİNAL ARAŞTIRMA DOI: 10.5336/biostatic.2023-100313

# Continuous Variable Issue in the Association Rule Mining: A Simulation Study

## Birliktelik Kuralı Madenciliğinde Sürekli Değişken Sorunu: Bir Simülasyon Çalışması

Damla Hazal SUCU[a], Bahar TAŞDELEN[a], Asena Ayça ÖZDEMİR[b]

[a]Mersin University Faculty of Medicine, Department of Biostatistics and Medical Informatics, Mersin, Türkiye
[b]Mersin University Faculty of Medicine, Department of Medical Education, Mersin, Türkiye

**ABSTRACT Objective:** In order to apply association rule mining to data sets with continuous variables, it is necessary to convert the variables into categorical structure. Therefore, we aim to compare the results obtained by categorizing continuous variables using different methods and analyzing them with association rule mining. **Material and Methods:** In this study, ChiMerge, clustering, Minimum Description Length Principle, equal interval, equal frequency methods were used to transform continuous variables into categorical structure by discretizing them. Various datasets were generated in the R with sample sizes of 100, 200, 500 and one binary dependent variable was created in each dataset. Additionally, 2, 3, 4, 5, 6, 7 continuous variables with a standard normal distribution were generated and various methods for transforming variables into categorical format were applied. A support of 10%, a confidence level of 80% were used. The number of rules varies based on the number of variables and the number of categories. **Results:** The study's results compared the descriptive statistics of the number of rules and the lift values. It can be said that high lift values are observed in scenarios with high levels of correlation and a higher number of variables and increasing the sample size can reduce the lift values of rules. **Conclusion:** In terms of the number of rules, the ChiMerge is the most affected by increasing the sample size. Furthermore, the ChiMerge yields stricter and higher lift values compared to other methods. While using association analysis, data type and multi-level associations should be considered.

**Keywords:** Association rule mining; lift ratio; continuous variable

**ÖZET Amaç:** Birliktelik kuralı madenciliğinin uygulanabilmesi için verilerin kategorik yapıda bulunması gerekmektedir. Bu amaçla, sürekli yapıdaki değişkenlerin farklı yöntemlerle kategorize edilerek, birliktelik kuralı madenciliği ile analiz edilerek elde edilen sonuçların karşılaştırılması hedeflenmiştir. **Gereç ve Yöntemler:** Sürekli değişkenlerin bulunduğu veri setlerine birliktelik kuralı madenciliği uygulayabilmek için değişkenleri kategorik yapıya dönüştürmek gereklidir. Bu çalışmada, sürekli değişkenleri denetimli ve denetimsiz biçimlerde ayrıklaştırarak kategorik yapıya dönüştürmek için Ki-birleştirme, kümeleme, Minimum Açıklama Uzunluğu Prensibi, eşit aralık, eşit frekans yöntemleri kullanılmıştır. Bu amaçla R programında çeşitli senaryolarda veri setleri üretilmiştir. Örneklem 100, 200, 500 olarak alınmış ve her veri setinde bir adet binary yapıda bağımlı değişken oluşturulmuştur. Bağımlı değişkene ek, birbiri ile %60, %70 ve %80 düzeylerinde ilişkili 2, 3, 4, 5, 6 ve 7 adet olmak üzere standart normal dağılıma sahip sürekli değişkenler üretilmiş ve değişkenlere kategorik yapıya dönüştürme yöntemleri uygulanarak birliktelik analizi sonuçları kaydedilmiştir. Destek değeri %10 ve güven değeri %80 alınmıştır. Kural sayısı, tüm kurgularda %10 destek ve %80 güven değerleri sabit olmak şartı ile değişken sayısına ve değişkenlerin kaç kategoriden oluştuğuna göre değişim göstermektedir. **Bulgular:** Çalışma sonuçlarında yöntemlerin ürettiği kural sayısı ve kurallara ait kaldıraç değerlerinin tanımlayıcı istatistikleri karşılaştırılmıştır. Yüksek kaldıraç değerlerinin, korelasyon düzeyinin ve değişken sayısının fazla olduğu senaryolarda görüldüğü ve örneklemi artırmanın birliktelik kurallarına ait kaldıraç değerlerini düşürdüğü gözlenmiştir. **Sonuç:** Kural sayısı bakımından, örnek genişliği artışından en çok etkilenen yöntem Ki-birleştirme yöntemidir. Ayrıca Ki-birleştirme yönteminde diğer yöntemlere göre daha katı ve daha yüksek kaldıraç değerleri elde edilmiştir. Birliktelik analizi uygulanırken verideki çok düzeyli birliktelikler göz önünde bulundurulmalıdır.

**Anahtar kelimeler:** Birliktelik kuralı madenciliği; kaldıraç oranı; sürekli değişken

**TO CITE THIS ARTICLE:**
Sucu DH, Taşdelen B, Özdemir AA. Continuous variable issue in the association rule mining: A simulation study. Turkiye Klinikleri J Biostat. 2024;16(1):38-46.

**Correspondence:** Damla Hazal SUCU
Mersin University Faculty of Medicine, Department of Biostatistics and Medical Informatics, Mersin, Türkiye
**E-mail:** dmlhzl5@gmail.com

Peer review under responsibility of Turkiye Klinikleri Journal of Biostatistics.

***Received:*** 15 Nov 2023 ***Received in revised form:*** 09 Feb 2024 ***Accepted:*** 09 Feb 2024 ***Available online:*** 28 Feb 2024

Association rule mining is a method used to identify relationships within the dataset by using data mining and statistical analysis techniques together, to reveal the relationships between risk factors and protective factors that play a role in preventing the disease. This method discovers patterns within the data set by identifying relationships between variables. That is, it measures the probability that a variable will occur together with other variables and utilizes the correlation technique.[1,2] Association rule mining can also detect the impact of specific treatment protocol or factors on survival rates. This information can be used to provide patients with better treatment options and increase survival times.[3,4]

In the literature, one of the most commonly used methods for association analysis is the Apriori algorithm. Apriori is a valid algorithm for Boolean association rule mining and operates iteratively based on the "prior" step in each knowledge request.[5] In this algorithm, probability calculations are based on the co-occurrence frequencies of variables in the dataset.[6] The algorithm produces outputs by considering specific support, confidence and lift values. Support represents the total frequency of a variable's occurrence in the association analysis. Confidence indicates the reliability of relationships in the association analysis. A rule's support and confidence values must be equal to or greater than a minimum threshold. When setting the minimum support threshold, it should not be too high to avoid producing a small number of rules, but it should not be too low to prevent excessive and unnecessary rules from being generated.[7,8] The lift value is calculated based on the support values of association rules and indicates how strong the association rules are. A lift ratio of 1 means that the two variables are independent of each other. A lift ratio greater than 1 indicates a positive relationship between two variables. If the lift ratio is less than 1, it is said that there is a negative relationship between the two variables. In other words, the probability of one variable occurring is considered lower than the probability of the other variable occurring.[9] The lift value is an important measure in association analysis as it helps determine how meaningful the association rules are and how powerful reflect real relationships. It is one of the most important criteria for selecting and interpreting association rules.

In this study, we compared the descriptive statistics of the number of rules generated by different supervised and unsupervised transformation methods when discretizing continuous variables using a simulation. The study considered different sample sizes, various numbers of continuous variables and different levels of correlation, aiming to evaluate the number of rules generated and the lift values associated with these rules.

## MATERIAL AND METHODS

To apply association rule mining, variables need to be in a categorical format.[10] For continuous variables in the dataset, certain pre-processing steps are required before conducting association analysis. One of them is discretization.[11] Continuous data can be categorized in a supervised or unsupervised manner. In supervised discretization, the class information of the data is considered, while unsupervised discretization does not rely on class information.[12] This study compares supervised discretization methods, such as ChiMerge and the Minimum Description Length Principle (MDLP), and unsupervised methods, including clustering (k-means), equal interval, and equal frequency.

The ChiMerge method combines similar data points into fewer categories by hierarchically sorting the data and selecting cut-off points to divide the data into specific category intervals. The cut-off points vary based on the data distribution and the purpose of the analysis. For each category interval, the chi-square statistic is computed. The results are compared, and categories with similar statistics are merged. The MDLP method uses the minimum description length and a stopping criterion known as the entropy criterion to group continuous attributes of the data matrix. The k-means clustering method groups data points based on a similarity measure and calculates a center point for each cluster. K-means clustering performs clustering analysis on the data and generates clustering results based on the similarities of continuous variables. In equal interval and equal frequency methods, data is grouped into categories with equal intervals or equal frequencies.[13,14]

Simulations were conducted using R version 4.2.2. the R program (R Foundation for Statistical Computing, Vienna, Austria), along with the arules, arulesCBA and faux libraries, was used for data generation, method application and result extraction.[15-17] In each scenario, the number of repetitions (loops) was set to 1,000. Sample sizes of 100, 200 and 500 were considered. Each dataset included a binary variable coded as 0-1 as the dependent variable. To predict the dependent variable, various simulations generated 2, 3, 4, 5, 6 and 7 continuous variables with standard normal distribution, each with moderate to high levels of association (60%, 70%, and 80%). These generated variables were converted to categorical structure. The average values of 1,000 repetitions were given in tables and the number of rules was approximated to an integer. When converting variables from continuous to categorical format, the number of categories for each variable can be determined manually or automatically. In this study, they were determined as automatically by the method used. In all simulations, association rule mining was performed using the Apriori method with recommended minimum support level of 10% and a minimum confidence level of 80%.[5] The number of rules varies depending on the number of variables and the number of categories.

## RESULTS

When the sample size was 100, the highest (2.15) and lowest (1.51) lift values were observed in a scenario with 60% correlation and 6 continuous variables. When comparing the methods, it was found that the ChiMerge method produced higher lift values and fewer rules in each correlation level (Table 1).

When the sample size was doubled (200), the widest range of variation in lift values was observed when the correlation level was 80% and there were 7 continuous variables (1.54 and 2.09). Additionally, rules with higher lift values were generated in scenarios with a correlation level above 70% and more than 4 continuous variables (Table 2).

When the sample size was increased to 500, it was observed that the ChiMerge method did not produce rules with low correlation level and the continuous variables smaller than 7 (Table 3). According to general results, as the sample size increased, faster decrease was observed in the number of rules with the ChiMerge method than the others.

In all sample sizes, lift values varied depending on the number of variables and the level of correlation. Higher lift values were obtained with more variables and higher correlation, indicating the presence of stronger and more meaningful rules. Moreover, it can be said that scenarios with high lift values were observed in cases with the high correlation level and the large number of variables.

As the sample size increased, the number of rules and lift values tended to decrease. Each method was highly sensitive to even small changes in the number of variables. In detail, it can be said that methods were more influenced by changes in the level of correlation than changes in the number of variables. As the number of variables and correlation level increased, the generated number of rules increased. The ChiMerge method generally produced fewer rules, while MDLP, equal frequency and clustering methods produced more rules.

As a result of the discretization process, the number of categories (k) for variables converted into categorical (nominal) format was sensitive to sample size ($n \cong 10k$). When the ChiMerge method was applied, variables had the highest number of categories. In other methods, it was observed that the average number of categories was three and not sensitive to sample size (Table 4).

**TABLE 1:** Number of rules and lift values of methods based on different number of variables and levels of correlation (sample size=100).

| Number of continuous variables | Methods | ρ=0.8 | | | ρ=0.7 | | | ρ=0.6 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Rules (n) | X±SD | Minimum-maximum | Rules (n) | X±SD | Minimum-maximum | Rules (n) | X±SD | Minimum-maximum |
| 7 | ChiMerge | 105 | 1.97±0.12 | 1.59-2.11 | 37 | 1.93±0.14 | 1.60-2.06 | 23 | 1.87±0.18 | 1.57-2.09 |
| | Cluster | 240 | 1.94±0.14 | 1.56-2.11 | 170 | 1.88±0.13 | 1.57-2.06 | 104 | 1.82±0.15 | 1.54-2.09 |
| | MDLP | 320 | 1.92±0.15 | 1.54-2.11 | 276 | 1.90±0.15 | 1.56-2.06 | 235 | 1.84±0.14 | 1.53-2.04 |
| | Interval | 177 | 1.91±0.15 | 1.57-2.11 | 101 | 1.88±0.14 | 1.56-2.06 | 73 | 1.79±0.15 | 1.53-2.07 |
| | Frequency | 256 | 1.95±0.13 | 1.62-2.11 | 231 | 1.88±0.12 | 1.58-2.04 | 169 | 1.85±0.14 | 1.54-2.07 |
| 6 | ChiMerge | 62 | 1.99±0.12 | 1.76-2.14 | 29 | 1.95±0.15 | 1.62-2.11 | 18 | 1.88±0.19 | 1.57-2.15 |
| | Cluster | 131 | 1.94±0.15 | 1.55-2.12 | 107 | 1.86±0.17 | 1.53-2.11 | 73 | 1.85±0.18 | 1.51-2.15 |
| | MDLP | 158 | 1.91±0.17 | 1.54-2.13 | 130 | 1.87±0.16 | 1.53-2.07 | 118 | 1.86±0.19 | 1.51-2.11 |
| | Interval | 91 | 1.94±0.16 | 1.57-2.14 | 66 | 1.83±0.15 | 1.55-2.10 | 45 | 1.80±0.18 | 1.51-2.11 |
| | Frequency | 129 | 1.93±0.13 | 1.60-2.12 | 121 | 1.88±0.15 | 1.56-2.11 | 100 | 1.84±0.17 | 1.52-2.10 |
| 5 | ChiMerge | 41 | 1.98±0.08 | 1.72-2.06 | 24 | 1.93±0.14 | 1.62-2.06 | 15 | 1.90±0.17 | 1.60-2.12 |
| | Cluster | 65 | 1.93±0.13 | 1.60-2.06 | 52 | 1.86±0.13 | 1.58-2.03 | 40 | 1.81±0.15 | 1.55-2.10 |
| | MDLP | 73 | 1.91±0.13 | 1.57-2.06 | 60 | 1.86±0.13 | 1.59-2.03 | 55 | 1.86±0.15 | 1.56-2.08 |
| | Interval | 46 | 1.92±0.13 | 1.63-2.06 | 36 | 1.88±0.13 | 1.61-2.06 | 33 | 1.82±0.16 | 1.57-2.11 |
| | Frequency | 64 | 1.92±0.11 | 1.64-2.06 | 57 | 1.85±0.12 | 1.60-2.02 | 50 | 1.81±0.15 | 1.56-2.06 |
| 4 | ChiMerge | 22 | 1.97±0.14 | 1.77-2.13 | 12 | 1.95±0.13 | 1.70-2.08 | 10 | 1.87±0.18 | 1.63-2.08 |
| | Cluster | 32 | 1.89±0.15 | 1.59-2.10 | 30 | 1.84±0.15 | 1.57-2.07 | 20 | 1.78±0.15 | 1.57-2.02 |
| | MDLP | 31 | 1.88±0.18 | 1.54-2.11 | 28 | 1.84±0.15 | 1.58-2.04 | 21 | 1.84±0.16 | 1.59-2.03 |
| | Interval | 26 | 1.90±0.16 | 1.63-2.13 | 20 | 1.84±0.13 | 1.60-2.06 | 15 | 1.77±0.15 | 1.56-2.02 |
| | Frequency | 30 | 1.87±0.15 | 1.62-2.09 | 27 | 1.82±0.13 | 1.58-2.02 | 22 | 1.78±0.13 | 1.57-2.00 |
| 3 | ChiMerge | 16 | 1.97±0.17 | 1.66-2.14 | 10 | 1.94±0.17 | 1.68-2.13 | 5 | 1.91±0.17 | 1.73-2.07 |
| | Cluster | 16 | 1.91±0.18 | 1.63-2.13 | 16 | 1.84±0.17 | 1.59-2.11 | 9 | 1.78±0.14 | 1.62-2.02 |
| | MDLP | 15 | 1.88±0.19 | 1.58-2.12 | 12 | 1.87±0.18 | 1.58-2.09 | 8 | 1.84±0.12 | 1.67-1.96 |
| | Interval | 14 | 1.92±0.17 | 1.67-2.12 | 12 | 1.88±0.17 | 1.61-2.11 | 7 | 1.77±0.14 | 1.61-1.99 |
| | Frequency | 15 | 1.88±0.16 | 1.63-2.11 | 14 | 1.82±0.18 | 1.57-2.11 | 8 | 1.75±0.11 | 1.59-1.89 |
| 2 | ChiMerge | 7 | 1.96±0.17 | 1.71-2.14 | 5 | 1.97±0.11 | 1.84-2.08 | 4 | 1.88±0.17 | 1.72-2.06 |
| | Cluster | 7 | 1.88±0.18 | 1.65-2.11 | 5 | 1.82±0.11 | 1.69-1.95 | 5 | 1.76±0.18 | 1.58-1.97 |
| | MDLP | 6 | 1.82±0.19 | 1.60-2.06 | 4 | 1.82±0.12 | 1.66-1.94 | 4 | 1.82±0.14 | 1.70-1.96 |
| | Interval | 5 | 1.85±0.18 | 1.66-2.08 | 6 | 1.82±0.11 | 1.68-1.96 | 4 | 1.76±0.13 | 1.63-1.92 |
| | Frequency | 6 | 1.81±0.15 | 1.66-2.03 | 5 | 1.75±0.10 | 1.63-1.88 | 3 | 1.72±0.12 | 1.62-1.86 |

SD: Standard deviation; MDLP: Minimum Description Length Principle.

**TABLE 2:** Number of rules and lift values of methods based on different numbers of variables and levels of correlation (sample size=200).

| Number of continuous variables | Methods | ρ=0.8 | | | ρ=0.7 | | | ρ=0.6 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Rules (n) | X±SD | Minimum-maximum | Rules (n) | X±SD | Minimum-maximum | Rules (n) | X±SD | Minimum-maximum |
| 7 | ChiMerge | 31 | 1.98±0.13 | 1.67-2.09 | 12 | 1.93±0.11 | 1.74-2.03 | 6 | 1.83±0.15 | 1.64-2.03 |
| | Cluster | 252 | 1.95±0.13 | 1.56-2.09 | 167 | 1.89±0.10 | 1.57-2.05 | 88 | 1.79±0.13 | 1.54-2.05 |
| | MDLP | 288 | 1.93±0.15 | 1.54-2.09 | 248 | 1.90±0.13 | 1.58-2.05 | 213 | 1.79±0.13 | 1.54-2.03 |
| | Interval | 110 | 1.91±0.15 | 1.57-2.09 | 73 | 1.85±0.12 | 1.59-2.04 | 36 | 1.76±0.12 | 1.56-2.00 |
| | Frequency | 254 | 1.95±0.11 | 1.65-2.08 | 242 | 1.90±0.10 | 1.60-2.04 | 177 | 1.80±0.12 | 1.55-2.03 |
| 6 | ChiMerge | 27 | 1.99±0.10 | 1.73-2.07 | 9 | 1.90±0.17 | 1.66-2.08 | 5 | 1.82±0.11 | 1.65-2.01 |
| | Cluster | 125 | 1.95±0.11 | 1.60-2.07 | 94 | 1.87±0.14 | 1.55-2.07 | 58 | 1.81±0.12 | 1.58-2.02 |
| | MDLP | 145 | 1.92±0.13 | 1.57-2.07 | 127 | 1.87±0.15 | 1.54-2.09 | 97 | 1.83±0.12 | 1.58-2.01 |
| | Interval | 75 | 1.92±0.12 | 1.60-2.07 | 49 | 1.83±0.14 | 1.57-2.06 | 27 | 1.77±0.12 | 1.58-1.99 |
| | Frequency | 127 | 1.94±0.10 | 1.66-2.07 | 120 | 1.87±0.13 | 1.56-2.07 | 94 | 1.82±0.11 | 1.57-2.01 |
| 5 | ChiMerge | 18 | 1.97±0.10 | 1.76-2.05 | 6 | 1.94±0.13 | 1.76-2.06 | 3 | 1.81±0.12 | 1.68-1.93 |
| | Cluster | 66 | 1.93±0.11 | 1.63-2.05 | 53 | 1.85±0.13 | 1.56-2.06 | 34 | 1.78±0.12 | 1.57-2.01 |
| | MDLP | 75 | 1.91±0.13 | 1.58-2.05 | 62 | 1.88±0.14 | 1.56-2.05 | 46 | 1.82±0.13 | 1.57-2.03 |
| | Interval | 41 | 1.91±0.10 | 1.66-2.04 | 32 | 1.82±0.12 | 1.57-2.03 | 23 | 1.75±0.11 | 1.59-1.96 |
| | Frequency | 62 | 1.91±0.09 | 1.68-2.03 | 60 | 1.86±0.12 | 1.58-2.04 | 48 | 1.80±0.12 | 1.59-2.02 |
| 4 | ChiMerge | 13 | 1.97±0.12 | 1.72-2.08 | 5 | 1.91±0.15 | 1.70-2.04 | 3 | 1.73±0.11 | 1.64-1.83 |
| | Cluster | 32 | 1.91±0.12 | 1.61-2.06 | 29 | 1.83±0.13 | 1.57-2.02 | 20 | 1.76±0.10 | 1.60-1.94 |
| | MDLP | 36 | 1.90±0.15 | 1.59-2.08 | 29 | 1.85±0.14 | 1.58-2.03 | 23 | 1.78±0.11 | 1.60-1.96 |
| | Interval | 23 | 1.89±0.12 | 1.65-2.08 | 19 | 1.85±0.12 | 1.62-2.01 | 12 | 1.73±0.10 | 1.60-1.90 |
| | Frequency | 30 | 1.88±0.11 | 1.65-2.04 | 29 | 1.82±0.12 | 1.59-1.99 | 23 | 1.76±0.09 | 1.59-1.91 |
| 3 | ChiMerge | 8 | 1.99±0.08 | 1.90-2.07 | 3 | 1.97±0.10 | 1.87-2.04 | 2 | 1.86±0.10 | 1.79-1.92 |
| | Cluster | 16 | 1.90±0.11 | 1.70-2.04 | 14 | 1.82±0.12 | 1.62-2.00 | 10 | 1.74±0.10 | 1.59-1.90 |
| | MDLP | 16 | 1.90±0.14 | 1.64-2.07 | 12 | 1.83±0.13 | 1.60-2.00 | 9 | 1.81±0.15 | 1.61-2.00 |
| | Interval | 12 | 0.89±0.11 | 1.71-2.02 | 11 | 1.83±0.11 | 1.67-2.00 | 7 | 1.73±0.11 | 1.59-1.87 |
| | Frequency | 14 | 1.86±0.10 | 1.69-2.01 | 12 | 1.77±0.11 | 1.60-1.95 | 10 | 1.72±0.10 | 1.57-1.86 |
| 2 | ChiMerge | 4 | 1.98±0.12 | 1.85-2.08 | 3 | 1.87±0.12 | 1.79-1.97 | 0 | - | - |
| | Cluster | 6 | 1.87±0.12 | 1.71-2.03 | 6 | 1.80±0.13 | 1.64-1.97 | 4 | 1.74±0.13 | 1.60-1.89 |
| | MDLP | 6 | 1.87±0.16 | 1.67-2.05 | 5 | 1.83±0.13 | 1.70-1.98 | 3 | 1.85±0.11 | 1.75-1.93 |
| | Interval | 5 | 1.86±0.13 | 1.72-2.02 | 5 | 1.78±0.11 | 1.64-1.89 | 3 | 1.73±0.12 | 1.67-1.80 |
| | Frequency | 6 | 1.82±0.11 | 1.69-1.97 | 6 | 1.73±0.11 | 1.61-1.90 | 3 | 1.71±0.10 | 1.63-1.81 |

SD: Standard deviation; MDLP: Minimum Description Length Principle.

**TABLE 3:** Number of rules and lift values of methods based on different numbers of variables and levels of correlation (sample size=500).

| Number of continuous variables | Methods | ρ=0.8 | | | ρ=0.7 | | | ρ=0.6 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Rules (n) | X̄±SD | Minimum-maximum | Rules (n) | X̄±SD | Minimum-maximum | Rules (n) | X̄±SD | Minimum-maximum |
| 7 | ChiMerge | 8 | 1.99±0.09 | 1.85-2.06 | 2 | 1.89±0.09 | 1.84-1.94 | 1 | 1.72 | - |
| | Cluster | 239 | 1.96±0.08 | 1.63-2.05 | 153 | 1.90±0.09 | 1.60-2.03 | 73 | 1.81±0.11 | 1.56-2.02 |
| | MDLP | 167 | 1.96±0.10 | 1.61-2.06 | 89 | 1.87±0.11 | 1.59-2.03 | 73 | 1.79±0.12 | 1.56-2.05 |
| | Interval | 66 | 1.89±0.10 | 1.62-2.04 | 39 | 1.81±0.10 | 1.61-1.99 | 23 | 1.73±0.11 | 1.58-1.96 |
| | Frequency | 254 | 1.94±0.07 | 1.70-2.04 | 247 | 1.89±0.09 | 1.60-2.02 | 167 | 1.83±0.10 | 1.57-2.02 |
| 6 | ChiMerge | 6 | 1.99±0.06 | 1.94-2.05 | 2 | 1.90±0.03 | 1.87-1.93 | 0 | - | - |
| | Cluster | 122 | 1.94±0.08 | 1.65-2.04 | 86 | 188±0.10 | 1.60-2.02 | 48 | 1.77±0.10 | 1.57-1.95 |
| | MDLP | 88 | 1.93±0.10 | 1.62-2.05 | 61 | 1.86±0.11 | 1.60-2.02 | 46 | 1.77±0.11 | 1.57-1.97 |
| | Interval | 42 | 1.93±0.09 | 1.68-2.03 | 30 | 1.84±0.10 | 1.64-2.00 | 16 | 1.71±0.09 | 1.57-1.86 |
| | Frequency | 126 | 1.92±0.08 | 1.69-2.04 | 123 | 1.87±0.10 | 1.58-2.01 | 89 | 1.78±0.09 | 1.58-1.97 |
| 5 | ChiMerge | 5 | 1.98±0.06 | 1.90-2.04 | 2 | 1.92±0.09 | 1.98-1.83 | 0 | - | - |
| | Cluster | 64 | 1.93±0.10 | 1.63-2.04 | 53 | 1.85±0.10 | 1.60-2.00 | 32 | 1.78±0.10 | 1.58-1.94 |
| | MDLP | 61 | 1.93±0.11 | 1.60-2.06 | 41 | 1.85±0.12 | 1.58-2.03 | 28 | 1.77±0.11 | 1.59-1.95 |
| | Interval | 26 | 1.90±0.08 | 1.72-2.04 | 18 | 1.81±0.10 | 1.62-1.99 | 12 | 1.74±0.09 | 1.61-1.89 |
| | Frequency | 62 | 1.91±0.09 | 1.70-2.03 | 60 | 1.83±0.10 | 1.59-1.98 | 49 | 1.78±0.09 | 1.58-1.96 |
| 4 | ChiMerge | 4 | 1.99±0.09 | 1.91-2.04 | 1 | 1.84 | - | 0 | - | - |
| | Cluster | 32 | 1.91±0.10 | 1.68-2.04 | 28 | 1.83±0.12 | 1.60-2.01 | 20 | 1.77±0.10 | 1.58-1.94 |
| | MDLP | 30 | 1.91±0.11 | 1.66-2.05 | 22 | 1.84±0.13 | 1.61-2.03 | 19 | 1.77±0.11 | 1.59-1.96 |
| | Interval | 18 | 1.91±0.08 | 1.74-2.03 | 13 | 1.82±0.11 | 1.66-1.98 | 10 | 1.72±0.10 | 1.60-1.88 |
| | Frequency | 30 | 1.88±0.09 | 1.68-2.02 | 28 | 1.81±0.11 | 1.60-1.98 | 22 | 1.77±0.10 | 1.60-1.94 |
| 3 | ChiMerge | 3 | 1.99±0.10 | 1.94-2.02 | 0 | - | - | 0 | - | - |
| | Cluster | 15 | 1.88±0.10 | 1.70-2.01 | 14 | 1.81±0.12 | 1.62-1.97 | 10 | 1.74±0.10 | 1.61-1.88 |
| | MDLP | 15 | 1.90±0.11 | 1.69-2.03 | 11 | 1.82±0.12 | 1.64-1.98 | 8 | 1.78±0.10 | 1.64-1.92 |
| | Interval | 11 | 1.87±0.10 | 1.70-1.99 | 8 | 1.80±0.11 | 1.65-1.94 | 6 | 1.74±0.09 | 1.63-1.87 |
| | Frequency | 14 | 1.84±0.09 | 1.70-1.98 | 13 | 1.76±0.10 | 1.60-1.92 | 9 | 1.73±0.08 | 1.60-1.84 |
| 2 | ChiMerge | 2 | 2.00±0.09 | 1.97-2.03 | 0 | - | - | 0 | - | - |
| | Cluster | 6 | 1.88±0.10 | 1.76-2.01 | 6 | 1.80±0.11 | 1.65-1.97 | 5 | 1.74±0.09 | 1.64-1.87 |
| | MDLP | 7 | 1.90±0.11 | 1.72-2.03 | 6 | 1.84±0.13 | 1.69-2.00 | 4 | 1.77±0.12 | 1.63-1.90 |
| | Interval | 5 | 1.88±0.10 | 1.76-2.00 | 4 | 1.85±0.11 | 1.71-1.97 | 4 | 1.74±0.11 | 1.65-1.86 |
| | Frequency | 6 | 1.83±0.09 | 1.73-1.96 | 6 | 1.73±0.10 | 1.63-1.88 | 3 | 1.70±0.09 | 1.62-1.79 |

**TABLE 4:** Average number of categories of discretized variables.

| | | n=100 | | | n=200 | | | n=500 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Number of continuous variables** | **Methods** | **ρ=0.8** | **ρ=0.7** | **ρ=0.6** | **ρ=0.8** | **ρ=0.7** | **ρ=0.6** | **ρ=0.8** | **ρ=0.7** | **ρ=0.6** |
| | | **(n)** | **(n)** | **(n)** | **(n)** | **(n)** | **(n)** | **(n)** | **(n)** | **(n)** |
| 7 | ChiMerge | 11 | 11 | 12 | 21 | 23 | 23 | 50 | 56 | 55 |
| | Cluster | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| | MDLP | 2 | 2 | 2 | 3 | 3 | 2 | 4 | 4 | 3 |
| | Interval | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| | Frequency | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 6 | ChiMerge | 11 | 13 | 13 | 21 | 22 | 23 | 49 | 57 | 58 |
| | Cluster | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| | MDLP | 2 | 2 | 2 | 3 | 3 | 2 | 4 | 4 | 3 |
| | Interval | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| | Frequency | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 5 | ChiMerge | 10 | 11 | 12 | 22 | 22 | 23 | 52 | 56 | 55 |
| | Cluster | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| | MDLP | 2 | 2 | 2 | 3 | 3 | 2 | 4 | 4 | 3 |
| | Interval | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| | Frequency | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 4 | ChiMerge | 11 | 12 | 11 | 19 | 23 | 23 | 53 | 56 | 58 |
| | Cluster | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| | MDLP | 2 | 2 | 2 | 3 | 3 | 2 | 4 | 4 | 3 |
| | Interval | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| | Frequency | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 3 | ChiMerge | 10 | 11 | 12 | 20 | 23 | 25 | 46 | 53 | 56 |
| | Cluster | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| | MDLP | 2 | 2 | 2 | 3 | 3 | 2 | 4 | 4 | 3 |
| | Interval | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| | Frequency | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 2 | ChiMerge | 10 | 12 | 11 | 20 | 25 | 24 | 51 | 55 | 57 |
| | Cluster | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| | MDLP | 2 | 2 | 2 | 3 | 3 | 2 | 4 | 4 | 3 |
| | Interval | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| | Frequency | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |

## DISCUSSION

In all scenarios, it has been observed that higher lift values and more meaningful, fewer, and stronger rules are obtained by the supervised ChiMerge method. In the simulation study conducted by Moreno et al., it was observed that strong rules were produced in small numbers, which supports the results of our study.[18] It was determined that the lowest category numbers of the variables converted to a categorical structure were 2 and the highest was 58. While the highest number of categories created was in the ChiMerge method, other methods produced a similar number of variables with an average of 3 categories. It is thought that the ChiMerge method produces stronger rules because it divides continuous variables into more homogeneous subgroups by increasing the number of categories. Based on lift ratios, it can be said that equal interval and equal frequency methods which are unsupervised methods are weaker than other methods. Mitov et al. compared ChiMerge, equal interval and equal frequency methods and reported that ChiMerge method obtained lower classification error than other discretization methods.[19]

Overall, scenarios with high lift values are observed in situations where the correlation level and the number of variables is high. In terms of the number of rules, the ChiMerge method is the most sensitive to sample size. In some scenarios, with a sample size of 500, this method does not generate any rules.

All simulations include association rule mining results of data sets generated from the standard normal distribution. The strengths and weaknesses of the methods used may vary depending on the distribution types of the variables in the data set. Dash et al. stated that different discretization methods may yield varying results when applied to different data sets.[20] Therefore, it is crucial to choose appropriate methods based on the data sets and learning context.

We believe that this study is important in terms of elucidating the continuous variable problem in association rule mining at various correlation levels, with various numbers of variables, using supervised and unsupervised discretization methods, with a detailed simulation setup.

## CONCLUSION

When examined by number of rules, generalization of analysis results as “better” or “worse” for a small or large number of rules depends on the purposes of analysis, and fewer or more rules may have advantages and disadvantages. Fewer rules can make results more understandable and interpretable. It can reduce the complexity of the analysis. More rules may provide opportunities to explore more associations. For example, in the field of medicine, especially in diagnostic and treatment studies, a small number of rules may be more effective, while in the field of marketing, a larger number of rules may be more effective. As a result, to obtain the optimum rule, items such as the research topic, the area where association analysis is used, data type, multicollinearity level should be taken into consideration and the analysis results should be evaluated depending on specific targets.

In this study, it should be considered that all inferences are made for a dataset with correlated standard normal distribution. Under these conditions, it can be said that increasing the sample size reduces the lift values of association rules and ChiMerge method provides more stringent and higher lift values than the other methods.

***Source of Finance***

*During this study, no financial or spiritual support was received neither from any pharmaceutical company that has a direct connection with the research subject, nor from a company that provides or produces medical instruments and materials which may negatively affect the evaluation process of this study.*

### *Conflict of Interest*

*No conflicts of interest between the authors and/or family members of the scientific and medical committee members or members of the potential conflicts of interest, counseling, expertise, working conditions, share holding and similar situations in any firm.*

### *Authorship Contributions*

***Idea/Concept:*** *Damla Hazal Sucu, Bahar Taşdelen, Asena Ayça Özdemir;* ***Design:*** *Damla Hazal Sucu, Asena Ayça Özdemir;* ***Control/Supervision:*** *Damla Hazal Sucu, Bahar Taşdelen, Asena Ayça Özdemir;* ***Data Collection and/or Processing:*** *Damla Hazal Sucu;* ***Analysis and/or Interpretation:*** *Damla Hazal Sucu, Bahar Taşdelen, Asena Ayça Özdemir;* ***Literature Review:*** *Damla Hazal Sucu;* ***Writing the Article:*** *Damla Hazal Sucu;* ***Critical Review:*** *Damla Hazal Sucu, Bahar Taşdelen, Asena Ayça Özdemir.*

## REFERENCES


1. Flank A. Multirelational Association Rule Mining. 2004. [Link]
2. Jain D, Gautam S. Implementation of apriori algorithm in health care sector: a survey. International Journal of Computer Science and Communication Engineering. 2013;2(4):22-8. [Link]
3. Köse A. Sağlık göstergelerinin birliktelik kuralları ile analizi [Analysis of health indicators by association rules]. Eskişehir Türk Dünyası Uygulama ve Araştırma Merkezi Bilişim Dergisi. 2022;3(2):31-7. [Crossref]
4. Abdullah U, Ahmad J, Ahmed A. Analysis of effectiveness of apriori algorithm in medical billing data mining. 4th International Conference on Emerging Technologies. IEEE. 2008. p.327-31. [Crossref] [PubMed]
5. Han JK, Kamber M, Pei J. Data Mining: Concepts and Techniques. 3rd ed. USA: Elsevier Inc; 2001.
6. Agrawal R, Srikant R. Fast algorithms for mining association rules. Proc. 20th int. conf. very large data bases, VLDB. 1994;1215:487-99. [Link]
7. Awadalla MH, El-Far SG. Aggregate function based enhanced apriori algorithm for mining association rules. International Journal of Computer Science Issues. 2012;9(3):277-87. [Link]
8. Yalçın A, Karabatak M. Nicel birliktelik kuralları için çoklu minimum destek değeri [Multiple minimum support value for quantitative association rules]. Fırat Üniversitesi Mühendislik Bilimleri Dergisi. 2017;29(2):57-65. [Link]
9. Maimon O, Rokach L. Data mining and Knowledge Discovery Handbook. Vol. 2. 1st ed. New York: Springer; 2005. [Crossref]
10. Rastogi R, Shim K. Mining optimized association rules with categorical and numeric attributes. IEEE Transactions on Knowledge and Data Engineering. 2002;14(1):29-50. [Crossref]
11. Yıldız F. Tarımsal veri madenciliğinde apriori birliktelik kuralının uygulanması [Doktora tezi]. Adana: Çukurova Üniversitesi; 2018.
12. Peker N. Geliştirilmiş ki-birleştirme algoritması ile ayrıklaştırılan verinin veri madenciliği yöntemleri ile sınıflandırılması [Doktora tezi]. Sakarya: Sakarya Üniversitesi; 2021. [Link]
13. Li W, Han J, Pei J. CMAR: Accurate and efficient classification based on multiple class-association rules. Proceedings 2001 IEEE international conference on data mining. 2001. p.369-76. [Link]
14. Yin X, Han J. CPAR: Classification based on predictive association rules. Proceedings of the 2003 SIAM international conference on data mining. Society for Industrial and Applied Mathematics. 2003. p.331-5. [Crossref]
15. Hahsler M, Grün B, Hornik K. arules - a computational environment for mining association rules and frequent item sets. Journal of Statistical Software. 2015;14(15):1-25. [Crossref]
16. DeBruine L. faux: Simulation for Factorial Designs (1.2.0). Zenodo. 2023. [Link]
17. Hahsler M, Johnson I, Kliegr T, Kucha J. Associative Classification in R: arc, arulesCBA, and rCBA. R Journal. 2019;11(2):254-67. [Crossref]
18. Moreno MN, Segrera S, López VF, Polo MJ. A method for mining quantitative association rules. Proc. of the 6th WSEAS International Conference on Simulation, Modelling and Optimization. 2006. p.173-8. [Link]
19. Mitov I, Ivanova K, Markov K, Velychko V, Stanchev P, Vanhoof K. Comparison of discretization methods for preprocessing data for pyramidal growing network classification method. New trends in intelligent technologies. Sofia. 2009;31-9. [Link]
20. Dash R, Paramguru RL, Dash R. Comparative analysis of supervised and unsupervised discretization techniques. International Journal of Advances in Science and Technology. 2011;2(3):29-37. [Link]