SİİRT ÜNİVERSİTESİ
Bilimin Işığında

Türkiye Tarımsal Araştırmalar Dergisi
dergipark.org.tr/tutad

Research Article

Turk J Agric Res
2024, 11(3): 296-304
© TÜTAD
ISSN: 2148-2306
e-ISSN: 2528-858X
doi: 10.19159/tutad.1518371

Association Rules Analysis for Continuous Chicken Egg Traits Dataset*

Figen CERİTOĞLU[1], Zeynel CEBECİ[2]**

[1]*Siirt University, Faculty of Agriculture, Department of Animal Science, Siirt, TÜRKİYE*
[2]*Çukurova University, Faculty of Agriculture, Department of Animal Science, Adana, TÜRKİYE*

Received: 21.07.2024 **Accepted:** 07.11.2024

ORCID ID (By author order)
orcid.org/0000-0003-4016-0394 orcid.org/0000-0002-7641-7094

**Corresponding Author: figenyildiz@siirt.edu.tr

Abstract: This study aims to apply the Apriori association rule algorithm on 14 continuous egg quality traits recorded from 4320 eggs of three commercial white-laying chicken lines. In the study all the continuous data were discretized using Equal-Width-Interval method based the number of intervals obtained with Rice formula. Association rules analysis on the discretized dataset resulted with a total of 349 rules consists of 3 and 4 items. According to the top five rules by support and confidence, some important associations were obtained between the certain value ranges of the traits egg weight, egg width, egg length, shell thickness, and shell breaking strength when compared to the others. The appropriate biological and economic interpretations of the obtained rules may contribute to the poultry industry in practice.

Keywords: Data mining, association rules, discretization, Apriori algorithm, egg traits

1. Introduction

In the last decade, rapid advances in information and communication technologies have led to the emergence of big data and, in parallel, data mining has become a popular field of study. Data mining is a field of data science that encompasses techniques and methods used to discover and extract interesting, important and useful information and patterns which are hidden in large relational and transactional databases (Han and Kamber, 2001). Data mining studies are commonly examined under two main groups: descriptive and predictive data mining. Association Rules (AR) analysis is a descriptive data mining process aiming to find the associations and co-occurrences between the items in data sets, especially in the large transactions data sets (Han and Kamber, 2001; Patel and Patel, 2014). Market basket analysis, which is done to determine which goods are purchased together in retail stores, is a typical common example of AR. As it is well understood from this classical example, ARs are "if-then" statements, showing the probability of associations between items. So, the rules can be defined as the antecedent itemset implies the consequent itemset. Since AR aims to discover such rules, it is also called to as association rule mining or mining associations.

Association rules has been originally introduced by Agrawal, Imielinski, Swami using an algorithm called AIS (Agrawal, 1993; Kumbhare and Chobe, 2014). Following the AIS algorithm many AR algorithms have been developed such as the Apriori (Agrawal and Srikant, 1994), Partition (Savesere et al., 1995), SET-oriented Mining of association rules (SETM) (Houtsma and Swami, 1995) and Closed Association Rule Mining (CHARM) (Zaki and Hsiao, 2012). The Apriori association rule algorithm that is the most well-known and widely used among them.

Association rule algorithms have been defined specifically for binary and multi-categorical data in the databases. For instances the Apriori association rule algorithm works with binary data. However, in agriculture as well as in many fields, databases consist of both categorical and continuous data. A typical poultry dataset contains mixed data types

*: This study is derived from the PhD dissertation titled “Use of Apriori Algorithm for Discovering Association Rules in Agricultural Data Mining”, which was accepted by Çukurova University, Institute of Science and Techno logy, and authored by the first author.

such as egg weight, shell color, albumen index, shell shape, etc., in egg quality parameters. This poses a challenge for association rule algorithms to apply AR with continuous variables. Fortunately, it is possible to convert continuous variables into categorical variables using discretization methods and apply existing AR algorithms. In order to address this issue, Dougherty et al. (1995), Liu et al. (2002), Kotsiantis and Kanellopoulos (2006), García et al. (2013), and more recently Ramírez-Gallego et al. (2015) have suggested various discretization methods for applying AR algorithms with quantitative data. Cebeci and Yildiz (2017a) compared the performances of well-known EWI, EFI and K-Means Clsutering (KMC) discretization methods.

Although it has been applied especially in the fields of marketing and medicine, AR studies are not often encountered in the field of agriculture and livestock farming. Although some studies have been done on rule mining in crop production, there are very few studies in the field of animal husbandry. In recent years, Nyambo et al. (2019), Niu et al. (2020), Balhara et al. (2021) and Patil (2021) have addressed association rule mining in livestock datasets in their studies. Additionally, Hahsler and Karpienko (2017), Nyambo et al. (2019), Wang et al. (2019), Niu et al. (2020) and Mehta and Bura (2020) have utilized advanced visualization tools within R statistical software to enhance the understanding of associations. Moreover, these previous studies were not conducted on continuous variables. Thus, in this study, a dataset of egg quality data of three white egg-laying chicken breeds raised at in an institutional poultry house was used to extract the rules between quantitative data. The parameters determining egg quality are examined in two categories as internal and external characteristics. While external quality traits include egg weight, shell thickness, shape index, shell color, and shell thickness while the internal quality traits are albumen index, yolk index, yolk weight, and yolk color (Molnar and Szöllösi, 2020; Okon et al., 2020). Out of them, egg weight is one of the most important quality traits. Researchers indicate that moderate egg weight is better for incubation hatchability (Wilson, 1991; Narushin and Romanov, 2002) due to positive correlation between hatching and egg weight. Simultaneously, shell thickness also affects hatchability. It has been observed that hatchability in eggs with thick shells is 30% higher compared to thin shells. Shell thickness is crucial for table eggs as well during transportation, eggs with thin shells are more likely to break. Another parameter affecting hatchability is egg shape. Successful hatchability is achieved with normal-shaped eggs, particularly when the egg shape index is between 72-76% (Narushin and Romanov, 2002; Elibol, 2009). For consumer perception, not only the shape of the egg but also the yolk color is important, as many consumers believe that eggs with a dark yolk color and shell color are preferable (Align et al., 2023). Therefore, a good understanding of both internal and external egg characteristics provides the opportunity to optimize production programs, adjust feeding schedules, and regulate environmental conditions effectively. Egg quality traits s also play a pivotal role in influencing the reproductive function of poultry and are essential for ensuring the production of healthy chicks (Durmus, 2014). Although some egg quality characteristics may not directly impact yield, they are economically significant depending on the preferences. Therefore, insufficient egg quality characteristics can have negative effects on both egg quality and chick hatchability, posing a significant economic challenge for poultry farmers and egg producers. Egg quality characteristics are crucial factors utilized to achieve a healthy chick hatchability and impact the reproductive function in poultry (Durmuş, 2014). While some egg quality characteristics may not directly affect yield, they are still significant factors influencing economic profitability due to producer preferences. Eggshell thickness is highly important for the extended storage, preservation, and packaging of eggs (Gül et al., 2021). Therefore, insufficient egg quality characteristics adversely affect both egg quality and chick hatchability. This leads to a significant economic problem for poultry farmers.

This study aims to apply AR to obtain the rules between egg quality traits as an example of AR studies in animal production. The goal of the study is to determine associations between the egg quality traits in addition to demonstrate how the Apriori algorithm can be applied to quantitative data.

2. Materials and Methods

2.1. Study data

A dataset consists of 15 traits measured on 4320 eggs of three commercial lines (Atabey, Nick and Decalp) at the Poultry Research and Application Farm in the Faculty of Agriculture, Çukurova University in Adana, Türkiye were used in this study. The dataset used comprises the line as a categorical variable and 14 continuous traits as given in Table 1. In order to ease to refer and analyze, the traits are renamed as short variable names from V1 to V14 as shown in the Table 1.

Table 1. The descriptive statistics and test results for the egg traits

Traits	TN	Mean	SD	Min.	Max.	CV (%)	IQR	ADT(p)	Outliers
Egg weight (g)	V1	60.16	4.91	47.68	74.72	8.19	6.74	4.28e-10***	57
Egg width (mm)	V2	43.19	1.22	42.38	46.67	2.83	1.59	0.02*	62
Egg length (mm)	V3	56.93	2.23	50.43	63.52	3.92	3.12	1.04e-14***	44
Egg pH	V4	8.46	0.20	7.89	9.04	2.38	0.28	8.28e-06***	31
Yolk color index	V5	81.77	5.36	66.29	97.42	6.55	7.54	0.003***	74
Shell breaking strength (n cm^{-2})	V6	4.68	1.05	1.70	7.65	22.44	1.44	1.20e-11***	119
Shell thickness (µm)	V7	366.4	22.64	303.3	429.4	6.18	31.3	0.0014**	42
Shell weight (g)	V8	6.80	0.64	4.99	8.66	9.48	0.90	0.173^{ns}	45
Yolk weight (g)	V9	16.08	1.90	11.03	21.23	11.80	2.49	2.80e-06***	42
Yolk height (mm)	V10	18.36	1.07	15.43	21.26	5.84	1.44	0.056^{ns}	37
Yolk width (mm)	V11	39.92	2.60	32.55	47.41	6.53	3.66	1.04e-06***	62
White height (mm)	V12	8.64	1.15	5.32	11.78	13.32	1.60	0.0011**	39
White width (mm)	V13	64.85	5.53	50.04	80.18	8.53	7.28	3.70e-24***	94
White length (mm)	V14	85.42	7.03	66.77	104.61	8.23	9.75	2.07e-10***	31
Chicken line	CL	The class variable has three levels: A, D, N							

TN: Analysis name of trait, SD: Standart deviation, Min.: Minimum, Max.: Maximum, CV: Coefficient of variation, IQR: Inter Quartile Range, ADT(p): Anderson-Darling Test, *: p<0.05, **: p<0.01, ***: p<0.001, ns: Not significant

2.2. Data preprocessing

Before starting AR analysis for continuous variables, it is important to perform two basic data preprocessing tasks such as detection and removal of outliers, and imputation of the missing values. The R, a language and environment for statistical computing (Anonymous, 2023) was used in all of the steps of data preprocessing and AR analysis in this study.

As the first data preprocessing task, the missing values were counted and imputed in the dataset. Although there are various methods available for imputation of missing values, one commonly efficient method is the Predictive Mean Matching (PMM) to apply on multivariate data (Little, 1988). The Predictive Mean Matching involves predicting missing values based on available data and matching them with observed values to create a complete dataset. It provides indispensable advantages such as avoiding potential loss of information and maintaining accuracy with complete data in subsequent analyses. An implementation of the PMM included in the "mice" package (Van Buuren and Groothuis-Oudshoorn, 2011) of R statistical software was used to impute the missing values of the traits in the study dataset.

For outlier detection, various statistics-based, distance-based, density-based, cluster-based and graph-based methods are available in the literature. Although the motivation to these methods is beyond the scope of this study, the boxplot method based on Tukey's interquartile range (IQR= Q3-Q1) statistics is one of the simplest methods that can be used successfully on many data. The boxplot method allows identifying potential outliers lying beyond the quartiles Q1 and Q3 in addition to visually inspection of data distribution. As described by Wang et al. (2019), the values below Q1-1.5 IQR and above Q3+1.5 IQR are evaluated as outliers with this method. The outliers for each trait were detected using boxplot function of R based on Tukey's IQR method, and they were removed from the dataset using a custom R script. The numbers of outliers detected and removed from the dataset are given in Table 1 according to the traits.

2.3. Normality tests and normalization

After missing value imputation and outliers removal in the dataset, a normality check was applied using Anderson Darling Test (ADT) in order to decide which discretization method will be employed in the study. In addition to the descriptive statistics of the traits, the normality test results were also provided in the Table 1. According the results of ADT, since all of the traits except yolk height were normal, only a normalization was applied for the trait of yolk height. In this way, deviations that may arise from different distributions of the examined traits were avoided.

2.4. Discretization of continuous data

Various unsupervised and supervised learning methods can be used for discretization of continuous variables for AR analysis. The supervised discretization, mainly Chi-square based, methods use the class information for setting discretization intervals which is mostly missing in practice (Cebeci and Yildiz, 2017b). On the other hand, the unsupervised methods need not class information and thus directly applied on the study data (Cebeci and Yildiz, 2017a). In this study, two unsupervised discretization methods called Equal Width Interval Discretization (EWI) and Equal Frequency Interval Discretization (EFI) were applied to discretize the continuous traits.

Equal Width Interval Discretization method is an effective and straightforward discretization method that transforms continuous features into discrete ones. This method divides the continuous data range into a fixed number of k intervals, as determined by the user. The width of each interval is obtained by dividing the data range by the number of intervals. In other words, the width of each interval is calculated using the Equation 1.

$$\text{Width}= [\max(X)-\min(X)]/k \qquad (1)$$

Where, X is a continuous data set and k is the number of intervals.

Equal Frequency Interval Discretization method is a widely used and effective discretization technique. In this method, the continuous dataset is transformed into categorical data by dividing it into intervals of equal frequency. The width of each interval is obtained by dividing the number of continuous values in the dataset by the number of intervals. The width of each interval is calculated as shown in Equation 2.

$$\text{Frequency}= n/k \qquad (2)$$

Where, n is the number of observations in the dataset X and k is the number of intervals (Hacibeyoglu and Ibrahim, 2018; Putri et al., 2023). The number of intervals for discretization were computed using ten different methods listed in Table 2.

The function "discretize" from the R package "arules" (Hahsler et al., 2016) was used to discretize the continuous traits on the filtered or cleaned dataset from previously described data preprocessing stage.

2.5. Association rules and Apriori algorithm

Association rule analysis, is a rapidly emerging field of study in data mining which is used to discover the interesting associations as the rules in large data sets. Consider a set of items $I = \{I_1, I_2, \dots, I_m\}$. Let T represent transactions containing sets of items, where $T \subseteq I$. If A is a set, then in transaction $T_i, A \subseteq T_i$. Association rules are represented as $A \Rightarrow B$; where A is termed the antecedent and B is the consequent of the rule. An $A \Rightarrow B$ association rule indicates that $A \subset I$ and $B \subset I$, with $A \cap B = \emptyset$ (Qiao et al., 2017).

The performance of ARs is evaluated through some metrics such as support, confidence. Support (s) quantifies the usefulness of the rule, while the confidence value (c) signifies the strength of the rule (Bhatia and Gupta, 2014). Let D be a database with different transaction records, and A, B be items within that database. The support value of the association rule $A \Rightarrow B$, denoted as $s(A \cup B)$, expresses the proportion of transactions involving $A \cup B$ to the total transactions (N) in D. It is computed as a ratio in Equation 3.

$$s(\mathrm{A} \Rightarrow \mathrm{B}) = \frac{s(A \cup B)}{N} \qquad (3)$$

Similarly, the confidence value (*c)* of association rule $A \Rightarrow B$ is denoted by $c\ (A \cup B)$, and it expresses the ratio of transactions involving $A \cup B$ to transactions containing A. $c\ (A \cup B)$ as seen in Equation 4.

$$c(A \Rightarrow B) = \frac{s(A \cup B)}{s(A)} \qquad (4)$$

In this study the Apriori algorithm which has been proposed by Agrawal and Srikant (1994) was

Table 2. Methods for calculating the numbers of classes/intervals to discretize the continuous data

Method	Rule	Formula	Reference
M1	Square root	$\lfloor n^{1/2} \rfloor$	Davies and Goldsmith (1980)
M2	Sturges Huntsberger	$\lceil 1 + log_2 n \rceil \cong$ $\lceil 1 + 3.3\ log_{10} n \rceil$	Sturges (1926) Doran and Hodson (1975)
M3	Brooks-Carruthers	$\lceil 5\ log_{10} n \rceil$	Brooks and Carruthers (1953)
M4	Cencov	$\lceil n^{1/3} \rceil$	Cencov (1962)
M5	Rice	$\lceil 2\ n^{1/3} \rceil$	Lane et al. (2016)
M6	Terrell-Scott	$\lceil (2n)^{1/3} \rceil$	Terrell and Scott (1985)
M7	Scott	$\lceil R\ /\ 3.5\ \hat{\sigma}\ n^{-1/3} \rceil$	Scott (1979)
M8	Freedman-Diaconis	$\lceil R\ /\ 2IQR\ n^{-1/3} \rceil$	Freedman and Diaconis (1981)
M9	Doane	$1 + \ log_2 n + \ log_2 \left(1 + \frac{\lvert g_1 \rvert}{\sigma_{g_1}}\right);$ $\sigma_{g_1} = \left(\frac{6(n-2)}{(n+1)(n+3)}\right)^{1/2}$	Doane (1976)
M10	K-means clustering	The *f(K)* algorithm defined by Pham et al. (2005)	Pham et al. (2005)

n: Number of observations, R: Max-min, IQR= Q3-Q1, log2: 2-base logarithm, 1og10: 10-based logarithm used as one of the well-known association rule algorithms. The Apriori is an algorithm which has two main steps consisting of pruning and joining as shown in the pseudo-code of the algorithm in Table 3 (Raj et al., 2021).

In each iteration, the algorithm creates a set of candidate items from the frequent itemsets found in the previous iteration as it is seen in Table 3. The pruning step is performed to create potential candidate itemsets. Ultimately, the dataset is scanned to count the support of candidate itemsets. If the support for the candidate item is greater than the user-defined threshold, this set of candidate items is called a frequent item. The support value refers to the frequency with which an item occurs, in other words, it refers to the number of transactions containing its item. However, in the Apriori algorithm, for an item set to be expressed as a frequent set, all its subsets must also be frequent. As the name of algorithm, Apriori suggests, it uses prior knowledge.

An implementation of the Apriori algorithm in the "arules" R package was run with the dataset consists of the discrete values of traits. The R packages "arules" and "arulesViz" (Hahsler and Chelluboina, 2011; Shin et al., 2015) were used to visualize the associations obtained from AR analysis. For this purpose, the apriori function of the "arules" package was run with the following parameter setting. In order to derive meaningful associations from the dataset, it is essential to establish the minimum support and confidence as the necessary parameters for the algorithm before conducting the analysis. In our analysis, we set minimum support and confidence to 0.001 and 0.8 respectively. Additionally, we configured the minimum and maximum length of the rules as 2 and 10 respectively.

Table 3. Pseudo code of Apriori algorithm

```
Algorithm: Apriori algorithm
Input:      D: Input Dataset
            minSup: minimum support threshold
Output:     All 2 to k-frequent itemsets
    $L_1$= {1-frequent itemset}// found separately
    For (k=2; $L_{k-1}$≠φ; k++)
        $C_k$= apriori_gen($L_{k-1}$)// finds k-candidate itemsets by joining and pruning $L_{k-1}$ with itself
        for each transaction t in D
            $C_1$=subset ($C_k$, t)// finds k-candidate itemsets in t
            for each c in $C_t$
                c.count++
            end for each
        end for each
        $L_k$= {c ∈ $C_k$ | c.count≥ minSup}
    end for
    Return $\cup_k L_k$
```

3. Results and Discussion

The number of intervals proposed by ten estimation methods (listed in Table 2) were given by the traits in Table 4. According to the results, M1 proposed the highest number of intervals while M2 proposed the least. On the other hand, M3, M4 and M9 estimated the number of intervals between 16 and 18. The remaining methods found the moderate values between 20 and 30. At this stage, it is critical to decide on the optimal number of intervals for discretization. A high number of intervals may lead to too much rule extractions and longer processing times, while working with a small number of intervals may lead to some valuable or rare rules to remain undiscovered. In this study, to decide the optimal number of intervals, the classification performance was calculated using the C5.0 classification algorithm (Kuhn and Quinlan, 2023) on the datasets discretized with methods EWI and EFI for each of the suggested interval numbers in Table 4. The test accuracy performance of the trained models on the test data was examined. Running the C5.0 algorithm we built the models using the independent discrete values for each trait and chicken line as independent or class variable as explained in Pandya and Pandya (2015) and Cebeci and Yildiz (2017a). For this purpose, 80% of the data, randomly selected, was used as training data and the remaining 20% as test data.

Table 5 shows the test accuracies obtained with the C5.0 algorithm. According to the results the test accuracies were computed around 50% almost for all of the discretization methods. Although these test accuracies for each dataset are at a moderate level, they are valuable in terms of revealing the relative performance of the discretization methods.

Table 4. Number of intervals to be used in discretization

Methods	Discretization methods	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14
M1	EWI	59	59	59	59	59	59	59	59	59	59	59	59	59	59
	EFI	59	59	59	59	59	59	59	59	59	59	59	59	59	59
M2	EWI	13	13	13	13	13	13	13	13	13	13	13	13	13	13
	EFI	13	13	13	13	13	13	13	13	13	13	13	13	13	13
M3	EWI	18	18	18	18	18	18	18	18	18	18	18	18	18	18
	EFI	18	18	18	18	18	18	18	18	18	18	18	18	18	18
M4	EWI	16	16	16	16	16	16	16	16	16	16	16	16	16	16
	EFI	16	16	16	16	16	16	16	16	16	16	16	16	16	16
M5	EWI	30	30	30	30	30	30	30	30	30	30	30	30	30	30
	EFI	30	30	30	30	30	30	30	30	30	30	30	30	30	30
M6	EWI	20	20	20	20	20	20	20	20	20	20	20	20	20	20
	EFI	20	20	20	20	20	20	20	20	20	20	20	20	20	20
M7	EWI	24	25	26	25	26	25	25	25	24	24	25	25	25	24
	EFI	24	25	26	25	26	25	25	25	24	24	25	25	25	24
M8	EWI	30	33	32	31	31	32	31	31	31	30	31	30	32	29
	EFI	30	33	32	31	31	32	31	31	31	30	31	30	32	29
M9	EWI	16	14	16	15	15	14	13	15	13	15	15	15	16	16
	EFI	16	14	16	15	15	14	13	15	13	15	15	15	16	16
M10	EWI	34	28	35	30	27	33	24	31	39	33	37	33	23	34
	EFI	34	28	35	30	27	33	24	31	39	33	37	33	23	34

EWI: Equal Width Interval, EFI: Equal Frequency Interval

Table 5. The test accuracies by the discretization and the interval number estimation methods

Ds	Continuous data	EWI	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
Ta (%)	48		50	51	51	50	52	51	50	50	51	48
Ds	Continuous data	EFI	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
Ta (%)	48		50	51	49	51	52	51	53	51	50	49

Ds: Dataset, Ta: Test accuracies

However, since M5-EWI, M5-EFI and M7-EFI resulted with slightly higher test accuracy when compared to the remaining, it was deemed appropriate to use one of them in discretization process. So, the dataset populated with the discrete values from EWI discretization based on the number of intervals from M5 (Rice method) was used in AR analysis in this study. As a result of this decision the rules were inferred for 30 classes/intervals for each trait.

As a result of AR analysis on the selected discretized dataset, totally 349 rules were obtained, of these rules, 134 were 3-items and 215 were 4-items. The top five rules with the highest support values are given in Table 6. Similarly, the top five rules having the highest confidence values are listed in Table 7. In these tables the left hand side (LHS) and right hand sides (RHS) represent the antecedent and consequent of the obtained rules. The rules plot and parallel coordinates plot of the obtained rules are also shown in Figure 1 and Figure 2. Figure 1 illustrates the composition of rules in terms of their elements and visually depicts which elements are shared among rules. In parallel coordinate plots, the intensity of colors represents confidence, while the width of the arrows indicates support.

In Table 6, the first rule is V1[60.3,61.2), V2[43.5,43.7), V6[5.27,5.47) ⇒ V3[56.5,57.0). This rule indicates if egg weight (V1), egg width (V2) and shell resistance (V6) are in the ranges of [60.3,61.2) and [43.5,43.7) respectively, the then egg length (V3) will be within a range of

Table 6. The top 5 rules with high support values

No	LHS						RHS			Support
	Item 1		Item 2		Item 3		Item 4			
1	V1	[60.3,61.2)	V2	[43.5,43.7)	V6	[5.27,5.47)	⇒	V3	[56.5,57.0)	0.0020
2	V1	[51.3,52.2)	V3	[56.5,57.0)	V6	[5.27,5.47)	⇒	V2	[43.5,43.7)	0.0020
3	V2	[40.9,41.2)	V7	[371,375)			⇒	V1	[51.3,52.2)	0.0017
4	V2	[42.3,42.5)	V3	[53.0,53.5)			⇒	V1	[53.1,54.0)	0.0017
5	V2	[41.6,41.8)	V14	[71.8,73.1)			⇒	V1	[52.2,53.1)	0.0017

LHS: Left hand side, RHS: Right hand side

Table 7. The top 5 rules with high confidence values

No	LHS					RHS		Confidence
	Item 1		Item 2			Item 3		
1	V1	[47.7,48.6)	V3	[52.6,53.0)	⇒	V2	[40.0,40.2)	1
2	V5	[87.0,88.1)	V7	[312,316)	⇒	V8	[5.60,5.72)	1
3	V5	[70.4,71.5)	V14	[71.8,73.1)	⇒	V6	[5.27,5.47)	1
4	V3	[59.2,59.6)	V5	[80.8,81.9)	⇒	V1	[66.6,67.)	1
5	V10	[18.5,18.7)	V11	[34.5,35.0)	⇒	V14	[79.4,80.6)	1

LHS: Left hand side, RHS: Right hand side

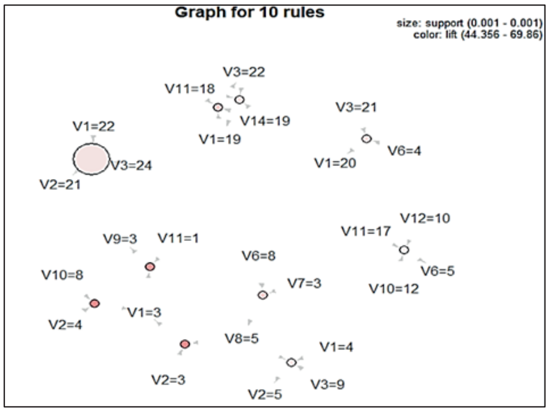

Figure 1. Rules plot for ten top rules

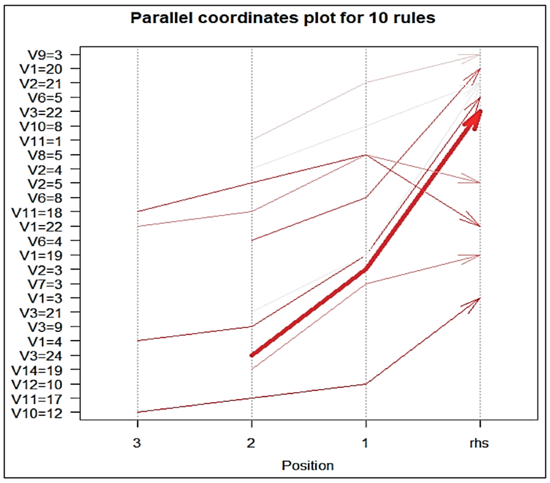

Figure 2. Parallel coordinates plot for ten top rules

[56.5,57.0). Similarly, in Table 7 the first rule V1[47.7,48.6), V3[52.6,53.0) ⇒ V2[40.0,40.2). This rule implies if the traits egg weight (V1) and egg length (V3) are in the ranges of [47.7,48.6) and [52.6,53.0) respectively, then egg width (V2) will be in a range [40.0,40.2). It is seen that V1, V2, V3, V6, V7 and V14 traits are frequently included together as antecedent or consequent in the top or important rules obtained according to both support and confidence criteria. This means that more supported associations do exist between the certain value ranges of the traits egg weight, egg width, egg length, shell thickness, and shell breaking strength when compared to the others.

In this study, association rule analysis, which is a data mining field of study was applied on egg traits dataset. The necessary data preprocessing and discretization processes to perform rules mining on a multivariate dataset containing a set of continuous variables are introduced. In this regard, this study not only provides a guide for rule mining in livestock continuous data, but also provides some interesting rules for chicken egg traits.

4. Conclusions

As a general conclusion, some important associations were obtained between the certain value ranges of the traits egg weight, egg width, egg length, shell thickness, and shell breaking strength when compared to the others. In fact, there are numerous research reporting strong correlations between these traits. However, previous studies conducted on the continuous values of the egg traits, reported only the correlations that indicate general measures of relations calculated on the all observations in the datasets. However, association rules analysis can provide more interesting and useful results in terms of showing which associations exist between which value ranges of the interested traits. The appropriate biological and economic interpretations of the rules may contribute to the poultry industry in practice. In this context, it is obvious that more comprehensive studies covering both categorical and continuous traits are also needed.

Ethical Statement

The authors declare that ethical approval is not required for this research.

Funding

This research received no external funding.

Declaration of Author Contributions

Conceptualization, Material, Methodology, Investigation, Data Curation, Formal Analysis, Visualization, Writing-Original Draft Preparation, *F. CERİTOĞLU*; Conceptualization, Material, Methodology, Formal Analysis, Supervision, Writing-Review & Editing, *Z. CEBECİ*. All authors declare that they have seen/read and approved the final version of the article ready for publication.

Declaration of Conflicts of Interest

All authors declare that there is no conflict of interest related to this article.

Acknowledgments

The authors would like to express their kind gratitude to Prof. Dr. Mikail BAYLAN from the Department of Animal Science, Faculty of Agriculture Çukurova University, for providing the dataset used in this study.

References

Agrawal, R., Imielinski, T., Swami, A., 1993. Mining association rules between sets of items in large databases. *SIGMOD '93: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, May 26-28, Washington, USA, pp. 207-216.

Agrawal, R., Srikant, R., 1994. Fast algorithms for mining association rules. *In: Proceedings of the 20th International Conference Very Large Data Bases*, September 12-15, Santiago de Chile, Chile, pp. 487-499.

Align, B.N., Malheiros, R.D., Anderson, K.E., 2023. Evaluation of physical egg quality parameters of commercial brown laying hens housed in five production systems. *Animals*, 13(4): 716.

Anonymous, 2023. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, (https://www.R-project.org/>), (Accessed Date: 25/05/2024).

Balhara, S., Singh, R.P., Ruhil, A.P., 2021. Data mining and decision support systems for efficient dairy production. *Veterinary World*, 14(5): 1258-1262.

Bhatia, J., Gupta, A., 2014. Mining of quantitative association rules in agricultural data warehouse: A road map. *International Journal of Information Science and Intelligent System*, 3(1): 187-198.

Brooks, C.E.P., Carruthers, N., 1953. Handbook of Statistical Methods in Meteorology. HM Stationery Office, London.

Cebeci, Z., Yildiz, F., 2017a. Unsupervised discretization of continuous variables in a chicken egg quality traits dataset. *Turkish Journal of Agricultural-Food Science and Technology*, 5(4): 315-320.

Cebeci, Z., Yildiz, F., 2017b. Comparison of Chi-square based algorithms for discretization of continuous chicken egg quality traits. *Journal of Agricultural Informatics*, 8(1): 13-22.

Cencov, N.N., 1962. Estimation of an unknown distribution density from observations. *Soviet Mathematics*, 3: 1559-1562.

Davies, O.L, Goldsmith, P.L., 1980. Statistical Methods in Research and Production. Longman, London.

Doane, D.P., 1976. Aesthetic frequency classification. *American Statistician*, 30(4): 181-183.

Doran, J.E., Hodson, F.R., 1975. Mathematics and Computers in Archaeology. Massachusetts: Harvard University Press, Cambridge.

Dougherty, J., Kohavi, R., Sahami, M., 1995. Supervised and unsupervised discretization of continuous features. In: Machine Learning: Proceedings of the Twelfth International Conference on Machine Learning, City, California, July 9-12, p. 194-202.

Durmuş, İ., 2014. Effect of egg quality traits on hatching results. *Akademik Ziraat Dergisi*, 3(2): 95-99. (In Turkish).

Elibol, O., 2009. Embryo development and hatching. In: M. Türkoğlu and M. Sarıca (Eds.), *Poultry Science, Breeding, Nutrition, Diseases*, Bey Ofset Matbaacılık, Ankara, Türkiye, pp. 151-188. (In Turkish).

Freedman, D., Diaconis, P., 1981. On this histogram as a density estimator: L2 theory. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 57(4): 453-476.

García, S., Luengo, J., Sáez, J.A., López, V., Herrera, F., 2013. Survey of discretization techniques, taxonomy and empirical analysis in supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, 25(4): 734-750.

Gül, E.N., Altuntaş, E., Demir, R., 2021. Determining the internal and external quality traits of eggs with different weights. *Journal of Agricultural Machinery Science*, 17(2): 55-63. (In Turkish).

Hacibeyoglu, M., Ibrahim, M.H., 2018. EF unique: An improved version of unsupervised equal frequency discretization method. *Arabian Journal for Science and Engineering*, 43(12): 7695-7704.

Han, J., Kamber, M., 2001. Data Mining Concept and Technology. China Machine Press: Beijing, China.

Hahsler, M., Chelluboina, S., 2011. Visualizing Association Rules: Introduction to the R- Extension Package arulesViz. (https://cran.csiro.au/web/packages/arulesViz/vignettes/arulesViz.pdf), (Accessed Date: 25/05/2024).

Hahsler, M., Buchta, C., Gruen, B., Hornik, K., 2016. Arules: Mining Association Rules and Frequent Itemsets. (https://CRAN.R-project.org/package=arules), (Accessed Date: 20.06.2024).

Hahsler, M., Karpienko, R., 2017. Visualizing association rules in hierarchical groups. *Journal of Business Economics*, 87(3): 317-335.

Houtsma, M., Swami, A., 1995. Set-oriented mining for association rules in relational databases. In: *Proceedings of the 11th IEEE International Conference on Data Engineering*, March 6-10, Taipei, Taiwan, pp. 25-34.

Kotsiantis, S., Kanellopoulos, D., 2006. Discretization techniques: A recent survey. *International Transactions on Computer Science and Engineering*, 32(1): 47-58.

Kuhn, M., Quinlan, R., 2023. C50: C5.0 Decision Trees and Rule-Based Models. R Package Version 0.1.8. (https://CRAN.R-project.org/package=C50), (Accessed Date: 20/06/2024).

Kumbhare, T.A., Chobe, S.V., 2014. An overview of association rule mining algorithms. *International Journal of Computer Science and Information Technologies*, 5(1): 927-930.

Lane, D.M., Scott, D., Hebl, M., Guerra, R., Osherson, D., Zimmer, H., 2016. Online Statistics Education: A Multimedia Course of Study. (http://onlinestatbook.com/Online_Statistics_Education.pdf), (Accessed Date: 20/06/2024).

Little, R., 1988. Missing-data Adjustments in large surveys. *Journal of Business and Economic Statistics*, 6(3): 287-296.

Liu, H., Hussain, F., Tan, C.L., Dash, M., 2002. Discretization: An enabling technique. Data *Mining and Knowledge Discovery*, 6(4): 393-423.

Mehta, A., Bura, D., 2020. Mining of association rules in R using Apriori algorithm. *Advances in Communication and Computational Technology*, 668: 181-188.

Molnar, S., Szöllösi, L., 2020. Sustainability and quality aspects of different table egg production systems: A literature review. *Sustainability*, 12(19): 7884.

Narushin, V.G., Romanov, M.N., 2002. Egg physical characteristics and hatchability. *World's Poultry Science Journal*, 58(3): 297-303.

Niu, L., Yang, C., Du, Y., Qin, L., Li, B., 2020. Cattle disease auxiliary diagnosis and treatment system based on data analysis and mining. In: *5th International Conference on Computer and Communication Systems*, May 15-18, Shanghai, China, pp. 24-27.

Nyambo, D.G., Luhanga, E.T., Yonah, Z.O., 2019. Characteristics of smallholder dairy farms by association rules mining based on Apriori algorithm. *International Journal of Society Systems Science*, 11(2): 99-118.

Okon, B., Ibom, L.A., Dauda, A., Ebegbulem, V.N., 2020. Egg quality traits, phenotypic correlations, egg and yolk weights prediction using external and internal egg quality traits of Japanese quails reared in Calabar, Nigeria. *International Journal of Molecular Biology*, 5(1): 21-26.

Ramírez-Gallego, S., García, S., Mouriño-Talín, H., Martínez-Rego, D., Bolón-Canedo, V., Alonso-Betanzos, A, Benítez, J.M., Herrera, F., 2015. Data discretization: taxonomy and big data challenge. *WIREs Data Mining Knowledge Discovery*, 6(1): 5-21.

Pandya, R., Pandya, J., 2015. C5.0 Algorithm to improved decision tree with feature selection and reduced error pruning. *International Journal of Computer Applications*, 117(16): 18-21.

Patel, H., Patel, D., 2014. A brief survey of data mining techniques applied to agricultural data. *International Journal of Computer Applications*, 95(9): 6-8.

Patil, A.B., 2021. A Role of data mining technique in healthcare system of lactating animals. *International Research of Humanities and Interdisciplinary Studies*, August 27-29, Maharashtra, India, pp. 25-29.

Pham, D.T., Dimov, S.S., Nguyen, C.D., 2005. Selection of K in K-means clustering. *Journal of Mechanical Engineering Science*, 219(1): 103-119.

Putri, P.A.R., Prasetiyowati, S.S., Sibaroni, Y., 2023. The performance of Equal-Width and Equal-Frequency discretization methods on data features in classification process. *Sinkron: Jurnal dan Penelitian Teknik Informatika*, 8(4): 2082-2098.

Qiao, L., Peng, C., Guo, X., Wang, Y., 2017. Price association analysis of agricultural products based on Apriori algorithm. *Proceedings of Science, Information Science and Cloud Computing (ISCC 2017)*, December 16-17, Guangzhou, China, pp. 1-7.

Raj, S., Ramesh, D., Sethi, K.K., 2021. A spark-based Apriori algorithm with reduced shuffle overhead. *The Journal of Supercomputing*, 77(1): 133-151.

Savesere, A., Omiecinski, E., Navathe, S., 1995. An efficient algorithm for mining association rules in large databases. In: *Proceedings of 20th International Conference on VLDB*, September 10, San Francisco, United States, pp. 432-444.

Scott, D.W., 1979. On optimal and data-based histograms. *Biometrika*, 66(3): 605-610.

Shin, S., Yoo, S., Kim, H., Lee, T., 2015. Association analysis of technology convergence based on information system utilization. *Journal of Computer Virology and Hacking Techniques*, 11(3): 173-179.

Sturges, H., 1926. The choice of a class-interval. *Journal of the American Statistical Association*, 21(153): 65-66.

Terrell, G.R., Scott, D.W., 1985. Oversmoothed nonparametric density estimates. *Journal of the American Statistical Association*, 80(389): 209-214.

Van Buuren, S., Groothuis-Oudshoorn, K., 2011. Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3): 1-67.

Wang, H., Bah, M.J., Hammad, M., 2019. Progress in outlier detection techniques: A survey. *IEEE Access*, 7: 107964-108000.

Wilson, H.R., 1991. Interrelationships of size, chick size, post hatching growth and hatchability. *World's Poultry Science Journal*, 47: 5-20.

Zaki, M.J., Hsiao, C.J., 2012. CHARM: An efficient algorithm for closed itemset mining. In: *Proceedings of the 12th SIAM International Conference on Data Mining*, 26-28 April, Anaheim, USA, pp. 457-473.

CITATION: Ceritoğlu, F., Cebeci, Z., 2024. Association Rules Analysis for Continuous Chicken Egg Traits Dataset. *Turkish Journal of Agricultural Research*, 11(3): 296-304.