Agris on-line Papers in Economics and Informatics

Volume XII Number 3, 2020

Analysis of Online Consumer Behavior - Design of CRISP-DM Process Model

Emil Exenberger, Jozef Bucko

Department of Applied Mathematics and Business Informatics, Faculty of Economics, Technical University of Košice, Slovakia

Abstract

The basis of the modern marketing of a business entity is to know the behavior of its customers. Advanced artificial intelligence methods, such as data mining and machine learning methods, penetrate data analysis. The application of these methods is most appropriate in the case of online sales of any goods in large quantities and various industries. They are very often used in the sale of electronics, PCs or clothes. However, it is also possible to apply them to the agricultural industry, not only in B2C, but also in B2B in the sale of seeds, agricultural products, or agricultural machinery. Appropriate combinations of offers and knowledge of customers can bring the selling entity higher profits or competitive advantages. The main goal of our study is to design a CRISP-DM process model that will enable small businesses to analyze online customers' behavior. To reach the main goal we perform a data analysis of the online sales data by using machine learning methods as clustering, decision tree and association rules mining. After evaluating the proposed model, we discuss its use of the proposed model in the field of internet sales in the agricultural sector.

JEL Classification: C53, D11, D91, M21, M31

Keywords

Classification, association rules, data analysis, consumer behavior, online shopping.

Exenberger, E. and Bucko, J. (2020) “Analysis of Online Consumer Behavior - Design of CRISP-DM Process Model", *AGRIS on-line Papers in Economics and Informatics*, Vol. 12, No. 3, pp. 13-22. ISSN 1804-1930. DOI 10.7160/aol.2020.120302.

Introduction

In today's rapidly changing world, the information and knowledge gained and used are very valuable. With a high degree of computerization in all areas of our lives, the amount of data generated is also increasing, which is further stored and monitored. Today, the automotive, energy, engineering or agricultural industries are more focused on digital data analysis than ever before. The goal is to extract data from the amount of data that can be effectively transformed into information that is very important for understanding the trend of a business or a company. The gathering of such information is often done to support an entity's decision to move forward in the business process.

We can find many different approaches to knowledge discovery in databases (KDD) by various authors (Brachman and Anand, 1996; Fayyad, Piatetsky-Shapiro and Smyth, 1996; Klösgen and Zytkow, 2002; Mannila, 1997; Simoudis, 1996), but they all follow certain principles. In 1996, as part of an EU project, the basic principles of data mining were formulated as the Enterprise Standard Data Mining Process (CRISP-DM) (Smart Vision Europe, 2015). Business Intelligence (BI) is such data mining (DM) in the KDD process, the results of which are used to support business decision-making in the form of reports.

One area that has tremendous potential and which enables efficient real-time data collection is the sale of products through a website. In the case of online sales, the main task is to ensure customer satisfaction, minimize costs, maximize profits, or streamline processes. Through the acquired data and the application of appropriate machine learning methods, it is possible to analyze customer purchasing behavior.

Our contribution aims to propose a process model of CRISP-DM to analyze business data to support business decision making in the area of business processes. We will apply the proposed process model of CRISP-DM to specific data obtained

from the online sale of electronic components of the international company SOS electronic to increase the effectiveness of the marketing strategy by creating marketing letters, which is expected to increase sales of these components. The main objective is to investigate the relationship between significant association rules and customer size in terms of the volume and financial amount of payments for orders in the online purchasing process.

Clustering is often used to segmentation mainly due to its high success rate, which was evaluated by many authors (Safri, Arifudin, and Muslim 2018; Prashar, Vijay, and Parsad 2018; Suchacka, Skolimowska-Kulig, and Potempa 2015; Keller, Gray and Givens, 1985). The basic idea of clustering is to make elements within one cluster as similar as possible, while the differences between elements from different clusters should be as large as possible. The advantage is regular updating of the training set after each addition of other elements, unlike other forms of classification and prediction based on the original, regularly unchanged training set. As the most commonly used clustering method is considered the k-means method, first used by MacQueen (1967). Prashar, Vijay and Parsad (2018) compare neural network prognostic ability, linear discriminant analysis, and k-means methods to reduce online retailers' vulnerability to market demand. At the end of this study, statistical evidence was provided on the accuracy of the predictions of the methods, with an accuracy of the k-nearest neighbor method of 79.1%. The problem of classifying two user sessions in an online store, a shopping session and a browsing session, was investigated in their work by Suchack, Skolimowska-Kulig and Potempa (2015). By comparing the results of several methods, they evaluated the k-nearest neighbor method (using Euclidean distance) as the most effective in terms of shopping forecasts and overall forecasts.

Decision trees (Quinlan 1986) belong to the basic machine learning classification methods and make it possible to classify data based on decision-making in response to individual tests. This is the most popular form of classifier representation. Komprdová et al. (2012) describe the CART algorithm as one of the best-known algorithms for creating decision trees, which is also a basic representative of binary trees, while also focusing on criterion statistics in particular for regression and classification trees. It explains the basic principles of decision tree creation because other binary trees can be obtained by modifying the rules of the CART tree. The decision tree classification method is most often used to analyze online customer behavior based on past customer behavior data on the seller's website. Sun, Cárdenas and Harrill (2016) present decision trees and software Weka as a new technique and tool that identifies critical attributes that affect the quality level of customer experience when visiting a travel agency website. Raj and Singh (2016) investigate how the demographic situation affects the frequency of online purchases, categorizing customers into three groups - often shoppers, frequent customers, and less frequent customers - using decision tree methods.

Products purchased either in-store or via online sales may have a connection. If the existence of one product affects the existence of another product in the same purchase, this relation expressed in the form of an implication constitutes an association rule (AR). This concept was first introduced by Agrawal, Imielinski and Swami (1993). The process of analyzing customers' shopping carts to find ARs that must meet predetermined conditions determining the significance of ARs is called association rules mining (ARM). ARM is currently used in various fields of research and analysis, such as language studies (Adamov, 2018), electronic transaction security (Askari, Md and Hussain 2020), medicine (Buczak et al., 2015; Soni et al., 2011; Luo et al., 2013), but often also in customer analysis (Kaur and Kang, 2016; Guo, Wang and Li, 2017). One of the most commonly used methods for ARM is the Apriori method, first described by Agrawal et al. (1994). Since then, the Apriori method has been improved to speed up the ARM process, e. g., improving operational efficiency by reducing the number of databases scans needed (Yuan 2017) or reducing the amount of operation needed (Wu et al. 2009). Other modifications were aimed at adapting the Apriori algorithm to real-time online consumer behavior analysis for specific businesses (Kaur and Kang 2016; Guo, Wang, and Li, 2017). For example, Alfian et al. (2019) propose a real-time analysis of consumer behavior for online commerce using ARM. For analysis, they use the proposed system to track customers, product browsing history, and transaction data from digital tagging.

Although ARM has appeared in many publications that analyzed consumer behavior in the purchasing process, it is more often used to track the purchasing process in a brick and mortar store and for specific goods (Avcilar and Yakut 2014; Chen et al. 2015). It is not used so often

in the research of the online purchasing process and we have not found any publication on the research of the sale of agricultural goods by this method. This has become a fundamental motivation for our research.

Despite the widespread use of clustering (k-means), decision-tree (CART) and ARM (Apriori) methods, we did not encounter a combination of these and use them in the analysis of online consumer behavior. However, there have been authors who have proposed or used similar combinations to analyze consumer behavior. Kunjachan, Hareesh and Sreedevi (2018) use a combination of k-means, Apriori and Eclat methods to analyze large amounts of data in the form of online sales data. They recommend this methodology for easier analysis of consumer behavior and the mining of hidden data relationships. The methodology based on the combination of ARM and decision trees in Ma, Haiying and Dong Gang (2011) was to create a model for segmenting customers in an online environment. The benefit was to help managers understand customers, analyze the market and make business management decisions.

Our goal is to show a combination of machine learning methods for understanding the behavior of customers and a targeted offer of a combination of goods. We aim to show that the given method can also be used in agribusiness for the sale of various agricultural goods and products. When studying the current literature, we did not find much application of these methods in the agricultural industry. Their use can be monitored mostly in the field of sales of electronics, PCs and various clothing. Gandhi and Armstrong (2016) provide an overview of the data mining methods used in agriculture, mentioning techniques such as neural networks, bayesian networks and support vector machines. Santosh Kumar and Balakrishnan (2019) use the Apriori algorithm directly to recommend products to customers in the agricultural market. Clustering is also often used in the analysis of agricultural data (Shedthi et al. 2017; Zhao et al., 2009). However, none of these studies focuses on the use of these methods in real-time in the analysis of consumer behavior.

Materials and methods

Based on an analysis of current research in this area, we found that the most common method used to analyze an online purchasing process by customers is clustering. This method is natural because it allows you to create related customer groups and choose the appropriate marketing management for them. It is also possible to assume that there will be specific association rules for different customer clusters. Therefore, we naturally put forward a research hypothesis:

H1: There are specific association rules for each related customer group

In our research, we observed how often a customer who bought the items of interest also bought another in the same purchase. We were interested in the analyzed relationship only on the assumption that the occurrence of such (in our example) pairs was sufficient and would meet the predetermined conditions. When selecting a research site, it was a precondition that the company had the possibility of selling online, with the emphasis being on data being as robust as possible. The dataset represents data from one year of online purchases of the company that distributes electronic components. In total, we obtained data representing **185,706 purchases from 4,111 companies** in one year. Each row of the dataset contains:

- the id of the company that made the purchase;
- id of purchased goods;
- purchase date;
- number of purchased items;
- price per unit of goods.

To keep data anonymous, customer and item names have been replaced by identification numbers. We carried out the whole process of work in the R software environment for its free availability and many such packages that allow performing all the steps of our analysis. From the dataset, we naturally chose the financial volume of payments and the number of orders made as the decision criteria. This selection best describes the customer in terms of creating a marketing strategy and at the same time is possible from the Dataset point of view.

As a first step, we calculated the total number of purchases for each customer and the total amount spent by online purchases in the selected store during the year. Subsequently, we removed the outliers from these values, which could distort the result, while these outliers may be subject to further study. To do this, we use the command *boxplot(Data)$out* (Chambers, 2017; Becker, 2018; Murrell, 2018), which outputs represent the outliers found in the table Data.

In the second step, prepare the data for clustering by **standardizing** the data using the *robustHD package and the standardize()* command. With the NbClust package we find **the optimal number of clusters** (Charrad and Ghazzali, 2014).

It compares the optimal number of clusters detected using twenty-four different methods, evaluates the results and recommends the **optimal number of clusters**. This is then used for **cluster analysis** of k-means using the *kmeans()* (Hartigan and Wong, 1979) command, and the matching results are written to the table including original, non-normalized total amount and purchase data.

The third step is to generate the **decision tree** and the corresponding **classification rules** using *rpart* (Breiman, 2017) package. The columns *number of purchases* and *total amount* will be the input attributes to be tested and the column representing the value of the burst to which the customer belongs – column cluster - will be the target attributes of the decision tree. The result of this step will be a model whereby an enterprise can split new customers into existing clusters.

In the final step, we will be mining the **association rules** for each cluster separately through the *arulez* and *arulezviz* packages. Hahsler and Gruen used the original (Agrawal, Imieliński and Swami, 1993) and more modern (Lepping, 2018; Borgelt and Kruse, 2002; Borgelt, 2003) designs of this method to program the Apriori function into the R environment. First, we divide the Customer Data table into as many tables as there will be the resulting number of clusters, and write down only the rows that belong to them in each table. We will then create a list of transactions (shopping carts) that will include, within each transaction, a set of goods that were purchased in that transaction. Subsequently, we will mine the ARs in the clusters, assuming that one purchase will consist of the goods purchased by one customer in one day. The resulting ARs will be evaluated using the *itemset, frequency, support, confidence* and *lift indicators*.

Itemset make up the goods offered to SOS electronic through online sales. *Frequency (A)* represents the absolute abundance of product A in an *itemset* and *Support (A)* represents the relative abundance of occurrences of product *A* in an *itemset*. *Confidence (A=>B)* expresses the ratio of both *A* and B products in one purchase to those in which there is only product *B* (see equation 1).

$$Confidence(A => B) = \frac{frequency(A,B)}{frequency(A)} \qquad (1)$$

It, therefore, determines how likely product *B* can be expected to be in a purchase that contains product *A*. However, let's have an example where product *B* is in all purchases and product A is only in 5% of purchases. The *confidence (A=>B)* would be equal to 1, but the predictive value of this AR would be low, because the presence of Product *A* in the purchase does not affect the occurrence of Product *B* in the same purchase. For this reason, the strength of AR is analyzed trough the *Lift (A=>B)* indicator by calculating how the presence of product *A* in the purchase affects the occurrence of product *B* in the same (see equation 2).

$$Lift(A => B) = \frac{Support(A,B)}{Support(A) * Support(B)} \qquad (2)$$

In our example, where Product *B* was in all purchases, *Lift (A=>B)* would be equal to 1 indicated that the presence of Product *A* in the purchase does not affect the occurrence of Product *B* in the same purchase. The higher *Lift (A=>B)*, the more product *A* affects product *B* in the same buying and and the lower the *Lift* value is below one, the more product *A* is negatively affected by product *B* in the same purchase. The Apriori method searches for ARs that meet the conditions *min_confidence* and *min_support* (or absolute value *min_frequency*), which the user selects himself. In association rules mining we left the values of *arulez* package at the default settings of *min_frequency* = 7 and *min_confidence* = 1. The result will be a table that expresses the association rules and the associated *support, confidence, lift* and *count (frequency)* values.

Results and discussion

The analysis results consist of these parts:

- **customer clusters** based on the total amount they spent on product purchases and the total number of purchases made;
- **decision tree** as a model for assigning customers to clusters from the previous point;
- **classification rules** as a textual notation of individual branches of decision-making
- tree;
- **association rules** mined for every existing cluster separately.

The result of detecting the optimum number of bursts using the *NbClust()* command was that the best number of clusters is 3.

We then split the customers into three clusters with the *kmeans()* command.

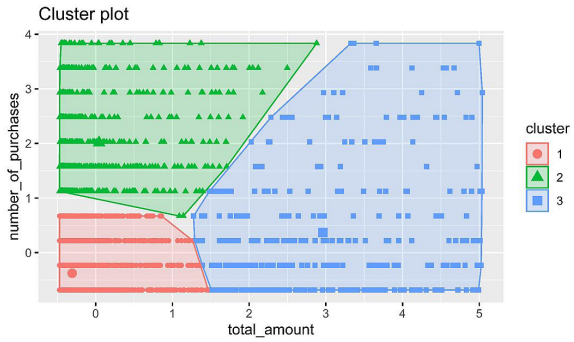
Table 1 shows the absolute and relative (rounded

	Cluster 1	Cluster 2	Cluster 3
Absolute abundance	3237	547	327
Relative abundance	78.74%	13.31%	7.95%

Source: prepared by authors

Table 1: Absolute and relative abundance of customers in clusters.

to 2 decimal places) abundance of customers in clusters.



Source: prepared by authors

Figure 1: Clusters visualization.

A graphical representation of the assignment of customers to the clusters is shown in Figure 1. From the graphical evaluation that Cluster 1 consists of customers whose made a low number of purchases with a low total amount; Cluster 2 is made up of customers whose made more purchases with a low total amount; and there are customers with more purchases for a higher total amount in Cluster 1 in the whole range of the number of purchases. Cluster 1 has the highest absolute abundance because it includes customers who made small purchases at low prices. However, from Table 1 and Figure 1, We cannot determine how many purchases or the total amount constitute the boundaries between individual clusters. To identify these boundaries, we generate a CART decision tree for its simplicity resulting from its binary.

Using package rpart, we have generated a CART decision tree with an absolute abundance of customers in every step showed in Figure 2. The accuracy of classifying customers into individual clusters by using the generated decision tree is 99,37%. For comparison, we also generated C5.0 decision tree with higher accuracy (99.93%). On the other hand, C5.0 decision tree has 10 branches, which makes it more complicated for further use in comparison of 4 branches in CART decision tree and because of that we continued with CART decision tree.

Source: prepared by authors

Figure 2: Decision tree.

Using the same package rpart we get classification rules according to generated decision tree shown in Table 2. Lines end with symbol * represents final assignments customers to clusters expressed by the following listing.

step) test total unfit cluster relatives

*Example: 5) number_of_purchases>=16.5 72 13 3 (0.18 0 0.81) **

where:

- **step (5)** - decision tree branch number (in the example it is step no. 5);
- **test (number_of_purchases>=16.5)** - boolean test performed on the node;
- **total (72)** - total number of customers entering the test;
- **unfit (13)** - number of customers whose are differently assigned to cluster by model. In the - example, 13 customers were assigned to other cluster by k-means as it flows from the decision tree;
- **cluster (3)** - the number of the cluster to which the customer belongs if he passes the test;
- **relatives (0.18 0 0.81)** - relative abundance of unfit customers by clusters. In the example, 18% customers were assigned to cluster 1 by k-means, but to cluster 3 by a decision tree.

Table 3 represented mined association rules for cluster 1 and its values for each mined

1) root 4111 874 1 (0.7873996595 0.1330576502 0.0795426903)
2) number_of_purchases< 5.5 3486 249 1 (0.9285714286 0.0005737235 0.0708548480)
4) total_amount< 238490.7 3236 6 1 (0.9981458591 0.0006180470 0.0012360939) *
5) total_amount>=238490.7 250 7 3 (0.0280000000 0.0000000000 0.9720000000) *
3) number_of_purchases>=5.5 625 80 2 (0.0000000000 0.8720000000 0.1280000000)
6) total_amount< 330888.3 548 8 2 (0.0000000000 0.9854014599 0.0145985401) *
7) total_amount>=330888.3 77 5 3 (0.0000000000 0.0649350649 0.9350649351) *

Source: prepared by authors

Table 2: Classification rules.

lhs		rhs	support	confidence	lift	count
{75125}	=>	{75127}	0.001498041	1	578.53	13
{84025,84027}	=>	{84029}	0.001498041	1	542.38	13
{84029,84033}	=>	{84025}	0.001498041	1	510.47	13
{84031}	=>	{84029}	0.001382807	1	542.38	12

Source: prepared by authors

Table 3: First 4 association rules for Cluster 1 sorted by highest support value.

association rule. In the lhs (left hand side) column is the implication input; rhs (right hand side) is the implication output. In the first association rule of Table 3, we can interpret as: "If the customer bought the product 75125, he also bought the product 75127 with 100% probability (expressed by confidence = 1) and the rule has been applied in 13 cases".

Number of association rules mined for each cluster by k-means with same settings are:

- Cluster 1 = 92 association rules;
- Cluster 2 = 16 association rules;
- Cluster 3 = 12 association rules.

To find out if there are intersections between ARs between we tried to find them and compare the values of the ARs in case of a match. The intersection of ARs between clusters having identical lhs and rhs in both clusters was only C2 ∩ C3 = 1 AR. Despite the intersection, found AR have different values of indicators (support, confidence and lift) between clusters. Based on this, we conclude that there are specific association rules for each related customer group and because of that we do not reject the hypothesis.

The results enable business management to segment customers into clusters for a better understanding of their structure. Clusters can follow association rules that can be used to analyze consumer baskets or predict consumer behavior. First, the customer would be assigned to the cluster according to the decision tree or classification rules, and then it is possible to follow the association rules associated with the cluster. We used a CART decision tree for its simplicity to create a model to quickly assign new customers to the clusters. However, the user can choose different types of decision trees for analysis, depending on his preferences. The work can serve as a CRISP-DM process model that does not require additional application costs and is therefore also suitable for smaller businesses that have not yet analyzed consumer behavior in online sales.

Association rules can serve to **predict consumer behavior** in real-time. After logging into the site, the customer is immediately assigned to one of the clusters with assigned ARs. When a customer puts a product into his shopping cart, all significant ARs that have the product on the lhs side are searched. Those ARs that have as much confidence and support as possible can then be used to create targeted advertising to maximize the likelihood of success. It is confidence and support that talk about the extent to which the AR was valid and how often it occurred in the original data. Targeted advertising in this form can increase sales and business profits. KPIs can be used to monitor the success of the implementation of the proposed CRISP-DM process model, such as:

- KPI 1: Proportion of revenues and costs associated with the implementation of the proposed CRISP-DM process model;
- KPI 2: The number of recommended goods purchased by customers who put in their carts based on targeted advertising;
- KPI 3: Total revenue from recommended goods.

Generated by the ARs can be used for analysis of **frequently occurring customer baskets**. The essence is the analysis of relationships between individual products. An enterprise can analyze why the customer buys product B solely with product A, even though product B has several other alternatives. The results of such an analysis could be used in different ways, e.g. products A and B would be offered in a single package at a discount or the results would serve as an incentive to verify and improve the quality of alternative products B. Process changes based on these analyzes could either bring additional revenues to the business or lead to cost reductions - e.g. goods A and B would be packed together in a single package, thereby saving the company on additional packaging. In such an analysis, when selecting association rules, management should focus on the ARs with the highest Lift value, which evaluates the strength of the relation between the products in each ARs. Verifying the success of the implementation of changes based on shopping cart analysis can be evaluated using KPIs similar to the prediction of consumer behavior described above.

Our results show that the right combination of offers to the right group of customers can increase a company's profits for sales. Our ambition is to apply this combination of methods in an area where it is less common, such as the agro-business sector.

Conclusion

Consumer Behavior Analysis makes it possible to provide business decision-making information that contributes to the achievement of business goals, with the primary benefit being profit. This can be achieved either by increasing revenues or by reducing costs within individual business processes. In our paper, we design and provide a low-cost process model CRISP-DM, that allows both when properly used and implemented. The proposed CRISP-DM process model is based on segmenting existing customers into individual groups using the k-means method; creating a model for segmenting new customers through the decision tree and analyzing the shopping cart using the Apriori method. In the discussion, we present the possible use of results and ways of measuring the success of the alternatives. The article can also serve as a brief overview of the knowledge of machine learning methods today and their use in current scientific analysis.

Although we did our research on the data of a company that sells electronic components, we think that the proposed CRISP-DM model can also be used in the field of agriculture, where it would allow the analysis of consumer behavior, for example in the sale of seeds, agricultural products, or tools usable in the agro-industry. Unlike the other studies mentioned (Gandhi and Armstrong, 2016; Kumar and Balakrishnan, 2019; Shedthi et al., 2017; Zhao et al., 2009), we propose a model that uses a combination of multiple data mining methods and allows real-time research of consumer behavior analysis of sales of agricultural products in the environment of online sales.

In our further research, we will focus on extreme values in the form of customers, from whom we have abstracted the CRISP-DM process model we proposed. We are going to evaluate customers through a modern machine learning method RFM (recency, frequency and monetary value), which assesses the importance of customers based on how often they make purchases when they made their last purchase, and at what value they made during the reporting period. We will test the model in this paper as well as its other planned modifications directly in the environment of agricultural online sales. Segmenting customers based on multiple criteria could ultimately increase the success of the association rules and thus achieve business goals.

Acknowledgements

The research was realized within the national project “Decision Support Systems and Business Intelligence within Network Economy” (Contract No. 1/0201/19) funded by Grant Agency for Science; Ministry of Education, Science, Research and Sport of the Slovak Republic.

Corresponding authors
Emil Exenberger
Department of Applied Mathematics and Business Informatics, Faculty of Economics
Technical University of Košice, Němcovej 32, 040 01 Košice, Slovak Republic
E-mail: emil.exenberger@tuke.sk

References

[1] Adamov, Abzetdin Z. (2018) "Mining Term Association Rules from Unstructured Text in Azerbaijani Language", *In 2018 IEEE 12th International Conference on Application of Information and Communication Technologies (AICT)*, IEEE, pp. 1-4. DOI 10.1109/ICAICT.2018.8747143.

[2] Alfian, G., Ijaz, M. F., Syafrudin, M., Syaekhoni, A., Fitriyani, N. L. and Rhee, J. (2019) "Customer Behavior Analysis Using Real-Time Data Processing: A Case Study of Digital Signage-Based Online Stores", *Asia Pacific Journal of Marketing and Logistics,* Vol. 31, No. 1, pp. 265-290. ISSN 1355-5855. DOI 10.1108/APJML-03-2018-0088.

[3] Askari, S., Md. S. and Hussain, Md. A. (2020) "E-Transactional Fraud Detection Using Fuzzy Association Rule Mining", *Proceedings of the 2nd International Conference on Information Systems & Management Science (ISMS) 2019,* Tripura University, Agartala, Tripura, India, 6 p.

[4] Avcilar, M. Y. and Emre, Y. (2014) "Association Rules in Data Mining: An Application on a Clothing and Accessory Specialty Store", *Canadian Social Science*, Vol. 10, No. 3, pp. 75-83. E-ISSN 1923-6697, ISSN 1712-8056.

[5] Becker, R. A. (2018) "*The New S Language: A Programming Environment for Data Analysis and Graphics*", CRC Press. ISBN 053409192X. DOI 10.1201/9781351074988.

[6] Borgelt, Ch. and Kruse, R. (2002) "Induction of Association Rules: Apriori Implementation", In: Härdle W., Rönz B. (eds) *Compstat*, Physica, Heidelberg. E-ISBN 978-3-642-57489-4. DOI 10.1007/978-3-642-57489-4_59.

[7] Borgelt, Ch. (2003) "Efficient Implementations of Apriori and Eclat", In *Proc. 1st IEEE ICDM Workshop on Frequent Item Set Mining Implementations*, FIMI 2003, Melbourne, FL, CEUR Workshop Proceedings 90.

[8] Brachman, R. J. and Anand, T. (1996) "The Process of Knowledge Discovery in Databases", In *Advances in Knowledge Discovery and Data Mining*, pp. 37-57. ISBN 9780262560979.

[9] Breiman, L. (2017) "Classification and Regression Trees", Routledge. ISBN 1138469521. DOI 10.1201/9781315139470.

[10] Buczak, A. L., Baugher, B., Guven, E., Ramac-Thomas, L. C., Elbert, Y., Babin, S. M. and Lewis, S. H. (2015) "Fuzzy Association Rule Mining and Classification for the Prediction of Malaria in South Korea", *BMC Medical Informatics and Decision Making,* Vol. 15, No. 1, pp. 47. ISSN 1472-6947. DOI 10.1186/s12911-015-0170-6.

[11] Chambers, J. M. (2017) "*Graphical Methods for Data Analysis*", Chapman and Hall/CRC, 410 p. ISBN 9781315893204.

[12] Charrad, M. and Ghazzali, N., Boiteau, V. And Niknafs, A. (2014) "NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set", *Journal of Statistical Software*. ISSN 1548-7660.

[13] Chen, Ch.-Ch., Huang, T.-Ch., Park, J. J. and Yen, N. Y. (2015) "Real-Time Smartphone Sensing and Recommendations towards Context-Awareness Shopping", *Multimedia Systems*, Vol. 21, No. 1, pp. 61-72. E-ISSN 1432-1882, ISSN 0942-4962. DOI 10.1007/s00530-013-0348-7.

[14] Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996) "The KDD Process for Extracting Useful Knowledge from Volumes of Data", *Communications of the ACM,* Vol. 39, No. 11, pp. 27-34. E-ISSN 1557-7317, ISSN 0001-0782. DOI 10.1145/240455.240464.

[15] Gandhi, N. and Armstrong, L. J. (2016) December. A review of the application of data mining techniques for decision making in agriculture. In *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I),* IEEE, pp. 1-6. DOI 10.1109/IC3I.2016.7917925.

[16] Guo, Y., Wang, M. and Li, X. (2017) "Application of an Improved Apriori Algorithm in a Mobile E-Commerce Recommendation System", *Industrial Management & Data Systems,* Vol. 117, No. 2, pp. 287-303. ISSN 0263-5577. DOI 10.1108/IMDS-03-2016-0094.

[17] Hartigan, J. A. and Wong, M. A. (1979) "Algorithm AS 136: A k-Means Clustering Algorithm." *Journal of the Royal Statistical Society. Series C (Applied Statistics),* Vol. 28, No. 1, pp. 100-108. E-ISSN 14679876, ISSN 00359254. DOI 10.2307/2346830.

[18] Kaur, M. and Kang, S. (2016) "Market Basket Analysis: Identify the Changing Trends of Market Data Using Association Rule Mining", *Procedia Computer Science,* Vol. 85, pp. 78-85. ISSN 1877-0509. DOI 10.1016/j.procs.2016.05.180.

[19] Keller, J. M, Gray, M. R. and Givens, J. A. (1985) "A Fuzzy K-Nearest Neighbor Algorithm", *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. SMC-15, No. 4, pp. 580-585. ISSN 21682216. DOI 10.1109/TSMC.1985.6313426.

[20] Klösgen, W. and Zytkow, J. M. (2002) "The Knowledge Discovery Process", In *Handbook of Data Mining and Knowledge Discovery*, 10-21 p. ISBN 978-0-387-09823-4.

[21] Komprdová, K. (2012) "Rozhodovací stromy a lesy", Akademické nakladatelství CERM, 98 p., ISBN 978-80-7204-785-7.

[22] Kumar, M. S. and Balakrishnan, K. (2019) "Development of a Model Recommender System for Agriculture Using Apriori Algorithm", In *Cognitive Informatics and Soft Computing,* pp. 153-163, Springer, Singapore. ISBN 978-981-15-1451-7.

[23] Kunjachan, H., Hareesh, M. J. and Sreedevi, K. M. (2018) "Recommendation Using Frequent Itemset Mining in Big Data", In *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, pp. 561–556. ISBN 9781538628430. DOI 10.1109/ICCONS.2018.8662905.

[24] Lepping, J. (2018) "*Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*". John Wiley and Sons Inc. E-ISSN 1942-4795.

[25] Luo, D., Xiao, Ch., Zheng, G., Sun, S., Wang, M., He, X. and Lu, A. (2013) "Searching Association Rules of Traditional Chinese Medicine on Ligusticum Wallichii by Text Mining", In *2013 IEEE International Conference on Bioinformatics and Biomedicine*, IEEE., pp. 162-167. ISBN 978-1-4799-1309-1. DOI 10.1109/BIBM.2013.6732664.

[26] Ma, Haiying, and Dong Gang. (2011) "Customer Segmentation for B2C E-Commerce Websites Based on the Generalized Association Rules and Decision Tree", In 2011 2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC), pp. 4600-4603, Piscataway, NJ : IEEE. ISBN 9781457705359. DOI 10.1109/AIMSEC.2011.6010255.

[27] MacQueen, J. (1967) "Some Methods for Classification and Analysis of Multivariate Observations", In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Oakland, CA, USA. Vol. 1, pp. 281-297.

[28] Mannila, H. (1997) "Methods and Problems in Data Mining", In *International Conference on Database Theory*, pp. 41-55. ISBN 3540622225. DOI 10.1007/3-540-62222-5_35.

[29] Murrell, P. (2018) "*R Graphics*", CRC Press.

[30] Parsad, Ch., Vijay, T. S. and Prashar, S. (2018) "Predicting Online Buying Behaviour-a Comparative Study Using Three Classifying Methods", *International Journal of Business Innovation and Research,* Vol. 15, No. 1, pp. 62-78. E-ISSN 1751-0260, ISSN 1751-0252. DOI 10.1504/IJBIR.2018.10009022.

[31] Quinlan, J. R. (1986) "Induction of Decision Trees", *Machine Learning,* Vol. 1, No. 1, pp. 81-106. E-ISSN 1573-0565, E-ISSN 0885-6125. DOI 10.1007/BF00116251.

[32] Rakesh, A. and Srikant, R. (1994) "Fast Algorithms for Mining Association Rules", In *VLDB ,94: Proceedings of the 20th International Conference on Very Large Data Bases*, pp. 487-99.

[33] Rakesh, A., Imieliński, T. and Swami, A. (1993) "Mining Association Rules between Sets of Items in Large Databases", In *Acm Sigmod Record*, Vol. 22, pp. 207-16. DOI 10.1145/170036.170072.

[34] Sahil, R. and Singh, D. (2016) "Impact of Demographic Factors on Online Purchase Frequency - A Decision Tree Approach", In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*. ISBN 9789380544205.

[35] Safri, Y. F., Arifudin, R. and Muslim, M. A. (2018) "K-Nearest Neighbor and Naive Bayes Classifier Algorithm in Determining The Classification of Healthy Card Indonesia Giving to The Poor", *Scientific Journal of Informatics,* Vol. 5, No. 1, pp. 18. E-ISSN 2460-0040, ISSN 2407-7658. DOI 10.15294/sji.v5i1.12057.

[36] Shedthi, B. S., Shetty, S. and Siddappa, M. (2017) "Implementation and comparison of K-means and fuzzy C-means algorithms for agricultural data", In *2017 International Conference on Inventive Communication and Computational Technologies (ICICCT),* IEEE, pp. 105-108. ISBN 978-981-15-0146-3. DOI 10.1109/ICICCT.2017.7975168.

[37] Simoudis, E. (1996) "Reality Check for Data Mining", *IEEE Expert*, Vol. 11, No. 5, pp. 26-33. ISSN 1541-1672. DOI 10.1109/64.539014.

[38] Smart Vision Europe (2015) "*About CRISP-DM*", CRISPDM by Smart Vision Europe. [Online]. Available: http://crisp-dm.eu/home/about-crisp-dm/, http://crisp-dm.eu/home/about-crisp-dm/ [Accessed: 20 July, August].

[39] Soni, J., Ansari, U., Sharma, D. and Soni, S. (2011) "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", *International Journal of Computer Applications*, Vol. 17, No. 8, pp. 43-48. ISSN 0975-8887. DOI 10.5120/2237-2860.

[40] Suchacka, G., Skolimowska-Kulig, M. and Potempa, A. (2015) "A K-Nearest Neighbors Method for Classifying User Sessions in e-Commerce Scenario", *Journal of Telecommunications and Information Technology*, Vol. 3, pp. 64-69. E-ISSN 1899-8852, ISSN 1509-4553.

[41] Sun, P., Cárdenas, D. A. and Harrill, R. (2016) "Chinese Customers' Evaluation of Travel Website Quality: A Decision-Tree Analysis", *Journal of Hospitality Marketing & Management,* Vol. 25, No. 4, pp. 476-497. E-ISSN 1936-8631, ISSN 1936-8623. DOI 10.1080/19368623.2015.1037977.

[42] Wu, H., Lu, Z., Pan, L., Xu, R. and Jiang, W. (2009) "An Improved Apriori-Based Algorithm for Association Rules Mining", In *2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery,* Tianjin, China, Vol. 6, pp. 51-55. DOI 10.1109/FSKD.2009.193.

[43] Xiuli, Y. (2017) "An Improved Apriori Algorithm for Mining Association Rules", In *AIP Conference Proceedings*, AIP Publishing LLC. DOI 10.1063/1.4977361.

[44] Zhao, B., Song, Z., Mao, W., Mao, E. and Zhang, X., (2009) "Agriculture extra-green image segmentation based on particle swarm optimization and k-means clustering", *Transactions of the Chinese Society for Agricultural Machinery*, Vol. 40, No. 8, pp.166-169. ISSN 1000-1298.