EAI Endorsed Transactions
on Scalable Information Systems
Research Article EAI.EU

Knowledge Discovery for Scalable Data Mining

Indu Chhabra[1,*] and Gunmala Suri[2]

[1]Professor, Department of Computer Science and Applications, Panjab University, Chandigarh
[2]Professor, University business School, Panjab University, Chandigarh
Email: [1]indu_c@pu.ac.in,chhabra_i@rediffmail.com, [2]g_suri@pu.ac.in

Abstract

The scalable diverse data and increasing levels of complexity in engineering and management science have given a boost to Data mining technology. The purpose of the proposed research is to evaluate the rule-based technique to develop solutions for analyzing customer Post Purchase behavior through knowledge discovery paradigm of Association rule mining. Over the years, it has proved a good tool to predict because of the incorporation of actual mined patterns. The current work is focused on extracting knowledge about the customer purchasing psychology and behaviour for the most frequent item combinations. For the purchase implementation, association rule framework is assessed for its performance analysis. The inferences of this automated intelligent system are based on of real life data set of 120 item-set combinations of five computer peripherals. This knowledge will help in framing and executing the most appropriate market laws and rules for the overall business growth.

Keywords: Association Mining, Knowledge Discovery, Influential Factors, Post Purchase Behavior and Retail Industry.

Received on 01 February 2019, accepted on 16 April 2019, published on 06 May 2019

doi: 10.4108/eai.19-3-2019.158527

*Corresponding author. Email:chhabra_i@rediffmail.com

1. Prologue

Finding business analytics is an important problem in a variety of disciplines from crucial applications of banking environment to ensure privacy to the day-to-day post-office mail sorting problems. The intelligent usage of data science and analysis has powered the analyst and executives to decide on the critical dynamics of competitors and user target groups to predict about the inter relationships among various internal issues of product setup, pricing and staff capabilities, which are the major contributors to improve on productivity as well as competitiveness.

Developing supportive relationship with customers is more important in the current era of intense competition. Various research issues and its implementation challenges are discussed through the present study. The purpose of the proposed research is to evaluate a rule-based technique to develop solutions for analyzing customer information about the most frequent items to be purchased together to derive most interesting verified market facts. Hence based on this knowledge, the most appropriate market laws and rules can be executed for the overall business growth.

This proposal is an effort to improve data mining analysis by applying association rule mining paradigm. Many traditional classification and prediction techniques like nearest neighbor and decision trees exhibit varied ambiguity and uncertainty and hence restrict the research far from being a general problem solver. Sometimes the situation becomes even graver due to subjectivity as well as few organizational and environmental noise artifacts. Hence the purpose is to investigate the factors which affect the association mining and can be improved.

2. Problem Definition

Association Rule Mining (ARM), the one of the best knowledge discovery method, applies association rules, to predict about most frequently occurred item combinations while preserving their due significance within that item-set. However some items are different in some aspects in real applications like where rare items are less frequent items to be purchased together.

To know consumer behavior as well as one's psychological condition at the time of purchase, the present study analyzes the database to derive pertinent information that can be utilized to enhance business returns at lower cost. *"If the customers find their favorite item-set nearby on the same shelf then they are more inclined towards buying their required itemsets from the same retailer instead of going to different places to shop for different items of same basket"*. This observation has been utilized to derive the intentions and psychology of the customer behavior to generate predictions about the various business avenues.

Developing application knowledge semantics from user knowledge and domain information of what are the patterns users most likely to purchase?" is the theme of the proposed work. In real life, the data, to be interpreted, is from different age groups to affect the business sales. For depicting the psychological state of a user at the time of purchase and to reason why one selects from various available alternatives, the data mining method is applied to improve the corresponding conventional method of knowledge derivation. The objective is to design a system that can retain the interest of all age-groups represented through the basket of item sets. To improve the business trust and for the interesting pattern extraction a generalized real world database is experimented.

3. Data Mining

Data mining is a science of studying real time business data with the prime objective of extracting knowledge from heterogeneous data sources to assist in the business processes. Recent advancements in information technology have facilitated the users to work upon large scale data as generated by the companies like Amazone, Google, eBAY and Facebook. For example to predict and recommend new friend connections, Facebook uses features and attributes of people to whom you may know. This information is really vital for predicting groups of like-minded people.

Data mining is capable of identifying the changing pattern relationships due to its centralized as well as decentralized control. It analyzes data from different perspectives to summarize the relationships into the knowledge rather than making the random guess. Data mining tools predict future trends and behavior allowing business to make proactive, knowledge driven decisions through the visualization of complex relationships existing in wealthy multidimensional databases.

In many businesses faster response means better customer service. Many well established descriptive and predictive data mining techniques are implemented by practitioners to utilize this potential of valuable data for the better management of their activities. Hence with these correct domain facts of historical patterns, the effective metrics can be derived to characterize the future market movements.

For the knowledge discovery, the four mostly applied algorithms by researches are the associations, sequencing, clustering and predictions. To predict about customer behavior, association mining known as "Affinity analysis" seeks the natural linking of the similar object qualities of items to find common groupings. Affinities among items are represented by association rules. These rules find those frequent items which imply the presence of other items. Technically, this quantification is done by the frequency parameter of support. Similarly, another accuracy parameter of confidence validates these found item-combinations through its exhaustive review.

3.1 Association Rule Mining Issues

The association rule mining discovers correlations among the explicit and implicit attributes of the database entities as required by user queries. Precisely this subjective correlation is converted into association rules of $X \Rightarrow Y$ for generalization. The strength of these rules is verified and validated by the interestingness parameters of support(s%) and confidence (c %) respectively. The support for rule $X \Rightarrow Y$ is the s % of database transactions containing X whereas the confidence is the c % fraction of transactions confirming the presence of Y after the occurrence of X only.

Association mining prunes the itemsets based on their frequencies. The frequent items, whose frequencies exceed the minimum support threshold, are retained while other infrequent are pruned. But this pruning technique may be insufficient to help market analyst in making decisions like planning and laying the supermarket shelf space and changing and the store plans. This is due to the modalities involved in the pruning process itself to extract the changing pattern-relationships. Hence the changing customer choices pose a challenge in the identification of most relevant patterns where each customer is treated as an independent entity.

Also the mining procedures require intensive computations for data analysis and comparisons with the complexity ranging from $\log(n)$ to (n^2). Hence the computing platform should be well equipped in terms of data and computing processors for the efficient access of the database items.

Market Basket Association analysis creates a worth from complex subjective relationships through the interesting analytical operators of support and confidence.

4. Research Initiatives and Projects

In 2012, to handle the complex data mining voluminous data, the US National Science Foundation announced the BIGDATA solicitation.

To implement the announcement made by President Obama and Prime Minister Shri Manmohan Singh during the Indo-US Open Government Dialogue in 2010, the Open Government Data (OGD) platform, India was developed jointly by India and US government in 2013 and it was regained in 2018 to add new plans.

The Department of Science and Technology, Govt. of India took a big data initiative in 2016 to promote the research in data mining. In 2017, another data science research initiative was taken to demonstrate their actual impact on data-intensive organizations, business and economy. For quality assessment, on-line data analysis and visualization were emphasized for complexity, efficiency and scalability metrics.

5. Literature Review

Various knowledge discovery techniques are inherited from the emerging artificial life-disciplines for their verification through computer automated conditions for societal and economic growth. Different classes of computing algorithms are provided which are capable enough to handle the routine problems concerning forecasting, firm turnover over the different time spans as well as the identification of credit bonus for banks [1]. How the data mining domain and organizational intelligence can be applied for the firm growth is well illustrated through the complete human and machine cooperated system [2]. These systems provide better prediction accuracy rather than the random guess which bring significant business values to the developers [3]. For real world applications concerning customer buying behavior, there are two main problems. The first is to check the relevance of basket item set combination such as seasonal items and another is exploring the most occurred item-set pattern in that subset of items [4]. These problems are well analyzed for automatic knowledge discovery through association mining algorithm [5]. The case study of online purchasing behavior is analyzed for the high valued customers. A prediction tool is developed to justify why they will prefer a specific air travel agency for their product and services in contrast to their competitors, pricing policies and package customization. This tool will assist the agencies to tailor their product packages according to the mined travel patterns [6].

The important flaws of existing traditional methods are critically analyzed by Ahn and Kim [7]. The importance of data normalization and outlier

detection and the advantage of dimensionality reduction to reduce the feature set are well demonstrated through data mining soft computing paradigms of decision trees, support vector machines and genetic algorithms. To predict the previously unseen trends in diagnosing different lifestyle based diseases the emphasis on in-time analysis of affluent medical data sources is justified through the Google Scholar survey [8]. Association rule mining, a challenging data mining research orientation, sometimes lacks in the candidate generation procedure where multiple passes have to be performed over the source data hence Waiswa and Baryamureeba [9] have illustrated the different ways of deriving interesting and unpredicted rules from large data sets.

The experimentation has successfully mapped the supporting assumptions into the quality measures of support and confidence. The controlling parameters of support and confidence are competent enough to threshold their rating according to the interestingness echelon of the generated rules. Directness and analytical precision are the parameters which can be evaluated in combination for association rule framing [10]. From the user point of view, how customer relationship management (CRM) can be beneficial to develop gainful relationships with specific customers is also well elaborated for the business rule generation [11]. Golchia [12] have tried to understand and improve the customer gaining and customer trustworthiness for overall growth and expansion.

Data mining, a knowledge extraction tool where the business actions are based on learning, the supervised learning paradigm of neural network is experimented to infer the measuring results which are beneficial to the business [13]. Some important supervised and statistical metrics are also applied for the in-depth analysis of fluctuations of Stock dataset. Both descriptive and predictive classifiers are implemented for the accuracy, misclassification rate and sensitivity. It is finally concluded through sufficient validation that logistic regression is capable of mining more promising facts as compared to other classifiers [14].

Few efficient rule mining ways to discover interesting and unpredicted rules from large customer data set are described for their comparative analysis [15]. The support and confidence metrics are applied to limit the level of interestingness in the derived rules. Hence these parameters facilitate in finding those interesting patterns which can improve on comprehensibility and predictive accuracy of the system [16].

6. Design and Working Scenario

For the performance evaluation, a process model is described to delve on various research issues of how to map the qualitative characteristics of consumer facts to some concrete rules, finding the strength of interesting relationships and evaluating the final impact of interestingness parameters on the overall growth. The objective is to frame significant policies especially when employees are from different user group as the real world scenario is. An association mining is implemented to sense the customer purchase preferences. The various addressed questions are “What is preferred with what?”, “Which items are ordered and purchased in combination?". With database D for transaction T, if an association rule X→Y holds then it confirms the transaction set T with confidence c % and s % support to offer X U Y as combined item set. For association rule mining firstly frequent patterns with respect to minimum threshold support are mined and then association rules are generated with respect to minimum confidence as specified by the confidence threshold.

Methodology: Specification of behavioral rules to derive intelligent quotient for retail industries.

Theme: Implementation of Association learning technique to imitate customer behavior to predict about rational and intelligent act for the future business expansion. The mining roadmap is

- Gathering user market database from heterogeneous sources
- Data Staging of preprocessing and filtration through ETL process
- Application of intelligent paradigms for association rule mining

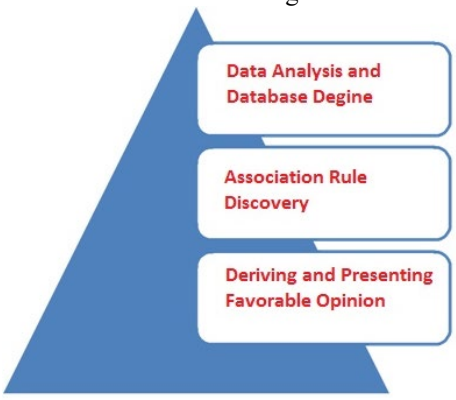

Figure 1: Proposed Knowledge Extraction Phases

Outputs: The knowledge generated from data mining is utilized by the retailing business for their decision making process about future strategies of sales and store layout.

This customized process is implemented through three phases of data analysis and database design, association rule discovery and favorable opinion derivation as shown in Figure 1.

Phase 1: Data Analysis and Design

In this phase various design issues are analyzed such as

- Incorporation of complete or partial historical knowledge: Choosing time-stamped procedure
- Efficiency and scalability of data: To select from parallel, distributed and incremental mining algorithms
- Dealing with the complexity of data as well as noisy and incomplete data: The data is relational but how much.

Phase 2: Association Rule Discovery

The knowledge is mined through

- Candidate Generation
- Frequent set inspection
- Rule Formulation

Phase 3: Favorable Opinion Derivation

The mined knowledge is validated through

- Rule Evaluation
- Pattern Evaluation

Table 1: Data mining Rationale

	Traditional Systems	**Scalable Data Mining Systems**
Purpose	Extraction of Detailed and Summary Data	Knowledge discovery of hidden patterns
Technique	Deduction (Result verification and inferences based on existing data)	Induction (Result prediction through new trial data inducing and training)
Inferences and Results	Data Analysis	Data Science

7. Implementation and Performance Analysis

Technically in implementation, data mining Online Analytical servers (OLAP) are capable enough to scrutinize data to metadata. They provide different performance perspectives to categorize and summarize the acknowledged relationships against their counterparts, the traditional systems, as compared in Table 1. In the present experimentation, for consumer behavior analysis, association rules are employed to lay down the different postulates for trusted customers. The optimized values of controlling parameters are obtained to provide the consistent and generalized solution.

To derive the intelligent quotient for retail industries for improved behavioral rules, a case study of analyzing buying habits of a customer is simulated. Frequent patterns and associated rule discovery is carried out through “market-basket analysis”, to assess the consumer behavior at the time of purchase, hence to facilitate the common man as well as the organization.

7.1 Efficiency Investigation

Efficiency, the one of the vital quality characteristics, has been assessed by the sub-characteristics through the multi-criteria decision making method of association rule mining.

Phase 1: Mining Methodology and Database Design

The sub-characteristics of support and confidence are investigated for the real life dataset of 5 basket transactions for 5 different products through the earlier specified three phases.

Phase 2: Association Rule Discovery

To quantify the intelligent behavior, the corresponding frequencies for the data aggregation are calculated to answer the question, *"Which types of products are purchased most frequently"*, rather than finding the frequency of all product items."

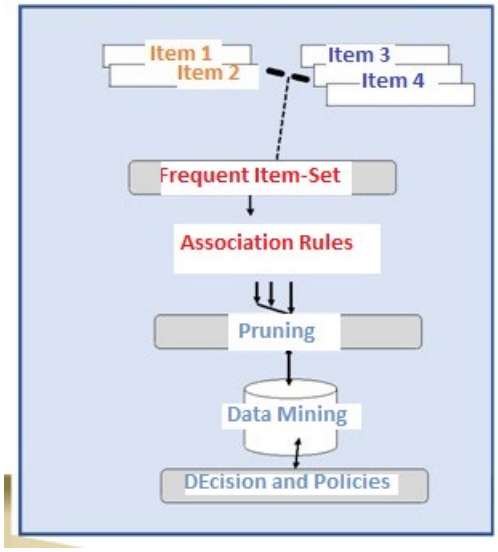

Figure 2: Intelligent Inferences and Predictions from User dataset

Association Rule Formulation:

The data is rolled up into various product categories through

If $((Ai1 = 1) \wedge (Ai2 = 1) \dots \dots (Aik = 1)) \rightarrow Ai(k+1)$

then it holds for true with $1 \leq ij \leq p$ for all j.

Finding Frequent Patterns:

Based on frequency threshold of minimum required support, the frequent set combinations like $\{A\}, \{B\}, \{C\}, \{D\}, \{AC\}, \{BC\}$ are found. Afterwards the rule inferences are articulated.

Phase 3: Deriving and Presenting Favorable Opinion

Knowledge extraction and result visualization are performed as shown in Figure 3.

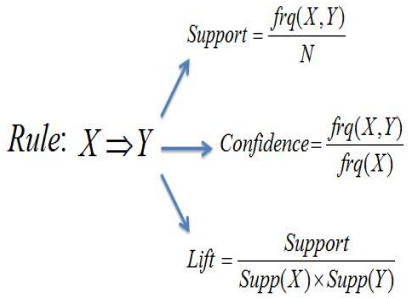

Figure 3: Association Rule Discovery

7.2 General Rule Derivations through Knowledge Discovery

Generalization is achieved through the derived sets of grouped combinations which are recommended and lifted to predict the customer behavior like the chances of buying item X along with item Y or not as shown in Figure

8. Results and Discussion

To achieve repeated success, what should be done to make the right decisions at the right time, a case study is executed.

8.1 Case Study Portraying Real-world Scenario

To check the practical viability of the present study, a database of five basket cart-combinations with five different items is investigated with 1st cart as rare combination, 2nd and 3rd carts for best with maximum sold combinations and 4th and 5th basket carts for regular itemsets purchased in routine.

8.2 Concrete Investigation

To demonstrate the mining process practically, a real life problem of buying computer peripherals in varying combinations according to varying requirements, is posed for the items A: Laptop, B: Podium, C: Projector, D: Antivirus and E: Hard Disk. The 5 item-combinations for 15 customer transactions are analyzed based on their frequency occurrence in the given database of 120 item set combinations. The goal is to identify and retain the most frequent set combinations for the future decision and pruned the non-frequent ones.

To verify and validate the frequency occurrence of five different combinations of { ABC: (Laptop, Podium, Projector)}, {ACD: (Laptop, Projector, Antivirus) }, {BCD: (Podium, Projector, Antivirus) }, {ADE: (Laptop, Antivirus, Hard Disk) } and {BCE: (Podium, Projector, Hard Disk) } with database frequencies (3,3,4,3,2), the vigilance and accuracy parameters of support and confidence are calculated. After applying the three phases of knowledge mining for the extracted support and confidence parameters for the given database, the proposed rules are lifted as shown in Table 2.

Table 2: Association Rule Framing

Rule	Support	Confidence	Lift
$Laptop \rightarrow Antivirus$	0.4 (2/5)	0.66 (2/3)	1.11 (10/9)
$Projector \rightarrow Laptop$	0.4 (2/5)	0.50 (2/4)	0.83 (5/6)
$Laptop \rightarrow Projector$	0.4 (2/5)	0.66 (2/3)	0.83 (5/6)
$Podium \,\&\, Projector \rightarrow Antivirus$	0.2 (1/5)	0.33 (1/3)	0.55 (5/9)

These rules verify that the "best display" and "to be purchased" item set combination is buying laptop with antivirus as indicated by the first row of Table 2 where the confidence is 66% for the 40% support. The row 2nd indicates the person purchasing projector requires laptop for display with confidence of 50% for the same 40% support. But as in this case the calculated confidence is less than the assumed confidence threshold of 65% so it may be overlooked if not beneficial. The benefit should be provided as implied by 3rd row to the buyer who is purchasing laptop and has varying chances of buying projector as indicated by column 2nd with same support but with different lift value in the 4th column of 3rd row for same item set combination of laptop and projector. The 4th rule demonstrates the buying habit that customer purchasing podium and projector may go for antivirus rarely as confirmed by the support of 20% which is the less than the minimum threshold of 30%.

Hence it is concluded that while framing the generalized business promotion polices to improve the sales of all five item-set combinations, the item-sets "Laptop and Antivirus" and "Projector and Laptop" should be given preference over the other frequent combinations and should be displayed besides each other for different models, ranges and prices to enhance the sales.

9. Examining Model Effects

There are two main ways to improve the efficiency of model either by dimensionality reduction, pruning the trivial item-combinations in subsequent iterations or reducing the number of passes. An incremental model is built which substantially refines the pruning process experimentally validated through the statistical parameter of confidence. The given model explains how accurately the variances in the original dataset create accurate estimations of the item-combinations to verify the significance of these effects. The experimentation on real-world scenario has demonstrated the efficacy of designed system.

10. Research Contribution

Technically, the data mining establishes the correlation among dozens of fields in large relational databases of patterns. This fact is established through the real world scenario of rare, best sold and regular combinations of item- sets. Hence to achieve the goal of business growth and market competitiveness, the derived association rules will enable the management to determine the relationships among various internal as well as external factors of price, profit and competition.

The proposed work aims at observing some interesting findings in the field of hybrid intelligence of management and computer discipline to apply it for the practical implementation of hybrid data mining tool for the retail industries to further study the buying habits of a consumer. The psychology of consumers, that how one reasons and selects between different alternatives, is converted into statistical technical frame that can be evaluated for his behavior.

Hence the multidimensional model can directly be analyzed to generate general market rules as shown in Figure 4. As mining is the method of analyzing data from different angles and perspectives, so the

gathered information can be used to improve revenue, costs and both.

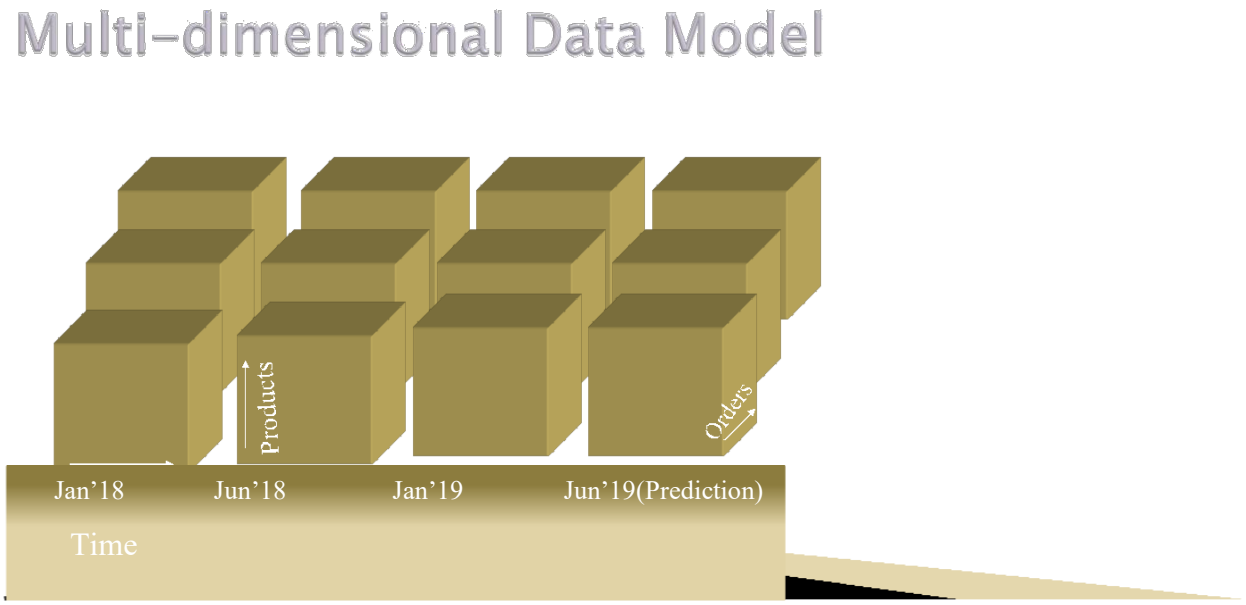

Figure 4: Prediction Deployment

11. Future Scope

For the voluminous database having mountains of metadata about large number of patterns and as well as for long patterns, this algorithm is a bit costly. It prunes the itemsets based on their frequencies which may be misinterpreted in case of huge datasets. Comparative analysis to other various data miming approaches like Apriori approach, capable of identifying suitable item sets with high utility in case of volumes of information, can be performed. For example value-added services like additional privileges, the percentage of benefit in purchasing and selling particular item-combinations can be identified to facilitate the consumers.

Future work will concern on the further investigation of how the system performance can be improved in terms of either considering other interestingness factors of reliability and correctness as well as designing purpose oriented experimental collections. Moreover, a future application of presented model can be "the learning from multi-classifiers".

12. Conclusion

Data mining has been in the news a lot lately, specially related to common man for sustainable development of the country as much discussed by the Prime Minister of India in various public interactions. The present study focuses on how a business can grow and remain in the market. An association rule mining is applied through the case study of customer buying habits to extract the most relevant facts. Hence these factors can be analyzed for the concrete generalized rules to strengthen the organizational decision making for future business growth.

The outcome is, the organizations can set-up various economical and demanding item-set combinations at one stop to retain the customer with the same retailer. Through this study the researchers are capable to draw the earlier unidentified but potentially valuable item-set combinations. Hence the mined factual rules will assist in automating the process of knowledge discovery.

References

[1] Ismail, J. (2002) The design of an e-Learning System. Beyond the hype. *Internet and Higher Education* **volume** (4).

[2] Cao, L. and Zhao, Y. (2009) *Data Mining for Business Applications,* 2nd ed. (Germany: Springer).

[3] Bughin, J. (2010) Clouds, Big Data, and Smart Assets: Ten Tech-Enabled Business Trends to Watch. *McKinSey Quarterly*.

[4] Ahmed, S.R. (2004) Applications of data mining in retail business. *Coding and Computing* **volume (**2).

[5] Woo, J.Y. (2005) Visualization method for customer targeting using customer map. *Expert Systems with Applications* **volume (**28).

[6] Eugene Wong**.** (2018) Customer online shopping experience data analytics: Integrated customer segmentation and customized services prediction model. *International Journal of Retail & Distribution Management* **volume (**46).

[7] Ahn, H. (2006) Hybrid genetic algorithms and case-based reasoning systems for customer classification. *Expert Systems with Applications* **volume (**23).

[8] Sharma M. and Singh G. (2017) Stark Assessment of Lifestyle Based Human Disorders Using Data Mining Based Learning Techniques. *IRBM* **volume (**38).

[9] Wakabi-Waiswa**.** (2008) Extraction of Interesting Association Rules Using Genetic Algorithms. *International Journal of Computing and ICT Research* **volume (**2).

[10] Sandhu Parvinder. (2011) Mining utility-oriented association rules: An efficient approach based on profit and quantity. *International Journal of* Physical *Sciences* **volume (**6).

[11] Turban, E. (2007) *Decision support and business intelligence systems,* 8th ed. (Pearson Education).

[12] Jenabi Golchia. (2013) Using Data Mining Techniques for Improving Customer Relationship Management. *European Online Journal of Natural and Social Science* **volume (2**).

[13] Ansari Azarnoush. (2016) Customer Clustering Using a Combination of Fuzzy C-Means and Genetic Algorithms. *International Journal of Business and Management* **volume (11**).

[14] Sharma M. and Singh G. (2018) Performance Analysis of Statistical and Supervised Learning Techniques in Stock Data Mining. *Data* **volume (3**).

[15] B. Yıldız. (2010) Comparison of Two Association Rule Mining Algorithms without Candidate Generation. *International Journal of Computing and ICT Research*.

[16] Kaur Amandeep (2018) Performance Efficiency Assessment for Software Systems. *Advances in Intelligent Systems and Computing*.