DOKUZ EYLÜL UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

A DECISION SUPPORT SYSTEM FOR CUSTOMER SEGMENTATION BASED ON A HYRBID APPROACH

by

Oğuzhan KAYA

May, 2019

İZMİR

A DECISION SUPPORT SYSTEM FOR CUSTOMER SEGMENTATION BASED ON A HYRBID APPROACH

A Thesis Submitted to the
Graduate School of Natural and Applied Sciences of Dokuz Eylül University
In Partial Fulfillment of the Requirements for the Master of Science
Philosophy in Industrial Engineering, Industrial Engineering Program

by

Oğuzhan KAYA

May, 2019

İZMİR

M. Sc. THESIS EXAMINATION RESULT FORM

We have read the thesis entitled **"A DECISION SUPPORT SYSTEM FOR CUSTOMER SEGMENTATION BASED ON A HYBRID APPROACH"** completed by **OĞUZHAN KAYA** under supervision of **ASST. PROF. DR. ŞEBNEM YILMAZ BALAMAN** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Asst. Prof. Dr. Şebnem Yılmaz BALAMAN

Supervisor

Prof. Dr. Hasan SELİM

(Jury Member)

Asst. Prof. Dr. Hülya GÜLDEMİR

(Jury Member)

Prof. Dr. Kadriye ERTEKİN

Director

Graduate School of Natural and Applied Sciences

ACKNOWLEDGEMENTS

I thank my supervisor Asst. Prof. Dr. Şebnem Yılmaz BALAMAN for her guidance and support me throughout my study.

Oğuzhan KAYA

A DECISION SUPPORT SYSTEM FOR CUSTOMER SEGMENTATION BASED ON A HYRBID APPROACH

ABSTRACT

With the rapid growth in information systems and technology which makes big data collection and management easier, all industries especially service industries begin to pay more attention to the analytic techniques for Customer Relationship Management (CRM). The purpose of this study is to develop a two level decision support system to derive meaningful and useful information from data and enhance decision making based on this information in CRM. To this aim, heuristic Algorithms have been integrated with Clustering techniques and Multi Criteria Decision Making Methods in this study. In the first level, to cluster the customers a hybrid approach that integrates Simulated Annealing and K Means Algorithms is used. In the second level, the highest value customers are determined using another hybrid approach that integrates Analytic Hierarchy Process to reach the weights of the criteria and TOPSIS to prioritize the clusters. The decision support system will enhance the process of planning business activities and the companies will be able to make strategic decisions considering the most important customers. Companies can also gain a significant competitive advantage by differentiation of the products and services to meet the specific needs of every single customer. To explore the applicability of the decision support system, the computational experiments are conducted that handle a tourism case, in which data regarding the city of Jakarta / Indonesia is used.

Keywords: Customer segmentation, AHP, TOPSIS, CRM, multi criteria decision making, tourism industry, simulated annealing, k means algorithm, clustering, DSS

MÜŞTERİ SEGMENTASYONU İÇİN HİBRİT BİR YAKLAŞIMA DAYALI BİR KARAR DESTEK SİSTEMİ

ÖZ

Teknolojideki ve bilgi sistemlerindeki hızlı gelişme ile birlikte, veri toplamanın, toplanan veriyi yönetmenin ve işlemenin kolaylaşması, hizmet sektörü başta olmak üzere tüm sektörlerde Müşteri İlişkileri Yönetimi (MİY)'ne olan ilgiyi artmıştır. Bu çalışmanın amacı, MİY nde kullanılmak üzere, veri setinden anlamlı ve kullanılabilir bilgi türetilmesi ve bu bilgiye dayalı olarak karar verme süreçlerini kolaylaştırmak için iki seviyeli bir karar destek sistemi geliştirmektir. Bu amaçla, sezgisel algoritmalar, sınıflandırma ve çok değişkenli karar verme tekniklerine entegre edilmiştir. Birinci seviyede, müşterileri segmentlere ayırmak için Tavlama Benzetimi ve K Means algoritmaları kullanılmıştır. İkinci seviyede, en değerli müşterileri belirlemek için TOPSIS ve Analitik Hiyerarşi Süreci metotlarını entegre eden bir yöntem kullanılmıştır. Analitik Hiyerarşi Süreci ile kriterlerin ağırlıkları belirlenmiş, TOPSIS ile de müşteri segmentleri sıralanmıştır. Bu karar destek sistemi, iş aktivitelerinin planlanması sürecini geliştirecek ve şirketlerin stratejik kararları en değerleri müşteri segmentlerini dikkate alarak verebilir duruma gelmesini sağlayacaktır. Ayrıca şirketler en değerli sınıftaki müşteri segmentlerine özel ihtiyaçları karşılamak için ürünlerini onlara göre özelleştirerek önemli bir rekabet avantajı kazanabileceklerdir. Geliştirilen karar destek sisteminin uygulanabilirliğini test etmek amacıyla, turizm sektöründe Jakarta / Endonezya şehri ile ilgili bir veri setini içeren vaka çalışması ele alınmıştır.

Anahtar Kelimeler: Müşteri segmentasyonu, AHP, TOPSIS, CRM, çok değişkenli karar verme, turizm endsütrisi, simulated annealing, k means algorithm, clustering, DSS

CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER ONE

INTRODUCTION

The increase in the competition between the companies, especially in similar sectors, enforces them to touch the customers' needs more effectively and to keep them closer to themselves. There is a wide range of CRM techniques and strategies to detect the needs of customers accurately so as to provide customized products and services for every single customer to fulfill the specific needs of customers and to gain an edge over competitors. Due to the advances in the information systems and technology, companies in a range of sectors have started to develop and employ analytical techniques to perform a variety of CRM processes in a more cost effective and timely manner to maximize their revenues for the long run profitability. Customer segmentation is one the most widely used CRM techniques to cluster and segment the customers according to their characteristics. In this section, after an overview of the CRM concept and the customer segmentation technique is proposed, the aims, scope, novelties and importance of this thesis will be explained.

1.1 Customer Relationship Management

In the related literature, CRM is described in plenty of ways. According to Chen & Popovich (2003) CRM is an adaptable system and set of people, process and technologies to define the needs of customers. Lefebure & Venturi (2001) view CRM as a process, a strategy or a technology. They have defined CRM as managing the relations with the customers based on the technology and business strategies to meet their exact needs which mean products and services in differentiated and customized ways. In that way, the companies increase their ability to identify, acquire and maintain the most valuable customers with the aim of increased sales and higher profits. CRM has very significant importance for competitive advantage in business, it helps to focus on the expectations of the customers (Ozgener & Iraz, 2005). CRM is also described as a combination of methods and tools to manage the relationships with customer in a systematic way. (Lawson-Body & Limayem, 2004). CRM as a system is business

applications which administrate and control the customer relationships and interactions with the customer-oriented business processes such as sales & marketing and services (Gefen & Riding, 2002). Shih & Liu (2003) states that Customer Relationship Management (CRM) is a customer-oriented enterprise mindset that includes analysis, plan and control of customer interactions and relationships through an advanced information systems and improved communication techniques.

The main contribution of CRM to the companies is that it changes traditional marketing strategies from product orientation approach to customer orientation approach. It provides companies to understand the needs of their customers accurately with the help of the data collected by using the advanced information technology which makes data collection and management easier. Therefore, they can apply customer oriented marketing strategies and gain a crucial competitive advantage against the other firms that produce and sell the alternative products or services.

CRM offers several benefits which is going to aid companies to describe, figure out and support their customers, therefore companies never have to worry about losing their customers and money as a result of incomplete data. The key under the success to meet the needs of the customers lies in truly figuring out their needs. CRM is doing that. The core benefits of CRM applications and software can be described as follows:

1. Improved Organization in terms of Information: The more the companies have information about their customers, the more opportunities they will have to completely provide the needs of customers that they pay off. Every interaction that the customers have with the companies' organization need to be identified, documented and recorded. With the aid of CRM, the methods are better than the sticky notes and messy filled cabinets and to begin enjoying developed enterprise technology that can accurately analyze and cluster the data for easy future reference and can share relevant information about the customers with the related departments.

2. Increased Communication with CRM: Because CRM provides the same customer data to every relevant people in the enterprise, service level of the sales or service people does not differentiate much compared to the condition without CRM.
   Usually valuable customers have one single contact in companies and because the contact knows the customer very well, the customer is always happy to see someone understands them. However, it is possible for that contact to have some annual leave, rotation to some other position or quit job. When it happens, valuable customer might be unhappy to contact with someone new and less experience. CRM is a preventive action for this risk by providing the customers' own unique requirements, solutions which are already prepared for them and experiences from the past which were recorded. Finally with CRM it doesn't matter who works with the customer, they will use the same unique information to assist the customer and everyone will be satisfied even including the employee. CRM can be used cloud base and used by any device such as phones, laptops or tablets and enterprises can enjoy benefits of out of office tasks.
3. CRM Improves Customer Service: Time has significant importance for the customers. Accordingly, if a customer is having a trouble with a problem and needs a quick solution, they will be unhappy if the correct and fast solution can not be provided. However, as an enterprise if you have all the information about the customer such as relevant experience in the past, interested products, most valuable solutions which were provided and nice feedbacks are received for it, problems or needs which were already requested before, then enterprise will take a quick action and will meet the need of the customer in a customized way. In the end, customer will be happy and the loss of time will avoid with the aid of CRM.
4. Improved Analytical Data and Reporting: Customer data can be analyzed in an organized way and valuable information can be derived by CRM. It avoids irrelevant data and miscalculations which can make the company unsuccessful in the long term. Moreover, with enhanced reporting by CRM, companies can make resourceful and effective decisions to reap the rewards in terms of customer loyalty and long run profitability.

In addition to the above mentioned benefits of CRM, there is another advantage of it for especially large scale companies. They are currently able to make many studies to analyze their customers' behavior and to collect data, however without CRM, big data means excess confusion, and eventually it means loss of time and money. CRM helps to avoid irrelevant data in the huge data sets, to cluster the customers by using the most useful data and to provide relevant and valuable information.

To summarize, CRM mainly enhances the following five implications in business:

1. Completion of orders which makes the sales teams and the other departments of the companies to manage the flow of the order across the departments in the organization.
2. It enables the companies to keep records of transactions of the customers and also it will be easy to display and check history of billing and past interactions with the customer.
3. Catalogue which makes easy to access information related to enterprises' prices of products and services, promotions, FAQ and relevant answers.
4. Tracking customers' activities which enables companies to track their customers' all activities related to the interaction with the company and gather profitable information.
5. Segmentation of the customers which enables companies to cluster their customers based on the big data which is gathered and analyzed via CRM applications and to put each new customer into a relevant segment.

Among those core implications of CRM, segmentation of the customers is concentrated on in this thesis, considering its essence for the companies in terms of constituting a base for the further business activities such as marketing and management to gain competitive advantage.

Wahlberg, Strandberg, Sundberg and Sandberg (2009) splitted CRM into 4 sub topics which are as below:

Brief explanations of those four branches of CRM are given as follows:

1. CRM Strategic: It covers strategy of the company's top management level entirely and it focuses on the customer. As main task of CRM Strategic, it supports and applies this focus considering it as CRM strategy, while it pays attention on the systematic analysis of big data related to the customers. It is also known that customer information which is gathered from customer's all activities related to interactions with the company is considered as a platform of marketing and management.

2. CRM Operational: It is mainly related to technology based front office activities which mean operational activities in the field such as sales and service. It needs to be used by the staff in call centers and by the sales team as well. Aim of CRM Operational is to maintain a measurable and optimized systems for the customer services team and sales team With that adi, the system will have opportunities for continuous improvement. Therefore, CRM Operational helps to improve sales and to develop functional activities in sales team.

3. CRM Collaborative: Recent advances in technology and rapid increase on internet usage of the people who might be potentially customers made it easy to collect information more about customers and to communicate with them simply. Those part of CRM is named as CRM Collaborative. With CRM Collaborative, it is possible to communicate customers via e-mail, phone contacts, SMS and social media. The best example for it would be online shopping. It enables customer to create their product configuration. If the customer does not like anything about the company, it is very easy to contact to the seller or write a complaint. There are more examples in mobile computing field as well.

4. CRM Analytical: Recent days, companies want to collect their customers' information and create a big data related to them. It is a real asset for the companies and it is very important to collect that in a systematic way because if the company have enough and valuable big data, then it is not very hard to make an analysis of it and improve marketing activities in customized ways. Base of CRM Analytical is data mining. There are many kinds of data collection techniques and with the help

of them big data is created. After that data can be analyzed in order to make it more meaningful and in a pattern. Main aim of CRM Analytical is to cluster the customers into segments. After clustering, adding more value to the offerings can be done by understanding the customers and meeting their exact needs with customized or differentiated products. Research and implementation area of CRM-Analytical is very open to statistical and mathematical methods. Such methods can be used to analyze a big data as, fuzzy clustering, signaling game, Walrasian general equilibrium approach, text analysis, novelty detection, Bayesian equilibrium, Markov Chain transmission matrix and neural networks.

In this thesis, Analytical CRM is studied due to the fact that it incorporates sophisticated methods and it is used for developing more structured, statistical and mathematical based decision support systems. Analytical CRM can also be integrated with the other approaches in order to create a novel CRM methodology. Customer segmentation, which is mentioned in detailed in Section 1.2, is at the center of Analytical CRM.

1.2 Customer Segmentation

In recent years, most of the companies adopt the customer retention strategy, which refers to the ability of a company or product to retain its existing customers over some specified period by improving their relationships with them, pointing out that the easiest way to grow their customers is not to lose them and repeated orders coming from the existing customers are more profitable than the orders from new customers due to the significant effort to create relationships with the new customers. When the amount of time and money spent for advertisement projects, marketing activities and research and development activities to gain new customers are considered, it can be proved in a simple way that acquisition of new customers is more costly than the retention of the existing customers. Therefore, for the competition with the other firms in the industry, it is essential for the companies to enhance customer loyalty and to decrease the rates of customer churn.

One of the core strategies in CRM is customer segmentation. Customer segmentation should be in the center of the business strategy because of the need for taking into account different expectations of varying customer groups. Segmented customer data is one of the most significant inputs in CRM practices. For better understanding of the expectations of their customers, the companies should apply customer segmentation. Kotler (1980) mentioned that segmentation of the customers is a sort of division of the market into defined sub groups of customers, then any segment can be selected as high value customer group as a main strategy of an enterprise. Customer segmentation can be seen as the practice of dividing a customer base into groups of individuals that have common characteristics so as to enable companies market to each group effectively and appropriately according to their characteristics. In marketing, a company might segment customers according to a wide range of factors relevant to their marketing strategy and properties of their product or service, such as age, gender, interests and spending habits…etc. To effectively segment the customers, customer features should be homogenous in the segment and heterogeneous among the segments. According to Kahreh et al. (2014), segmentation based on demographics and socio-economics uses age, social class, marital status, education, religion and so on. Therefore, those factors are available to be used to make a direct need analysis.

While clustering the customers, data mining techniques and statistical data analysis techniques are widely used to define and group the customers' features, their expectations and their sensitivities related to the features of the products or services as well as the company itself. Methods such as K- Means Algorithm, K-Medoids Algorithm, Hierarchical Clustering Algorithm, Decision Trees, Neural Networks and Clustering Heuristics are applied commonly in the related literature and in practical cases to determine the customer segments accurately. A well-designed segmentation process that is supported by analytical procedures enable effective and easy customization of the products or services to meet every single customer's need. Also, it provides better understanding of the customers' expectations for the long lasting customer retention, strong relationships with customers and profitability. It must be noted that segmentation is a long run process, which means the data related to specific

features of the customers should be gathered repeatedly due to the possible changes related to these features, and accordingly segmentation groups and expectations should be revised and updated regularly. According to the updates, business strategies should be changed and improved.

On the other hand, standardized products or services are more cost effective in terms of production or service costs since they do not require differentiated production or service processes. However, in the recent era, demands significantly change according to the customers and from time to time depending on the system dynamics. The companies which can differentiate their products or services in a timely manner to meet those changing needs possess a significant competitive advantage. It is notable that segmentation of the customers is a necessity in order to have this competitive advantage, because the way which lies on the road to customer satisfaction and customer loyalty is only possible with an accurate understanding of the customer needs, which depends on better segmentation.

However, segmenting the customers is a challenging process. In order to obtain reliable and accurate segments, companies need to consider some input data related to varying features of their customers to cluster them, such as;

- Demographic information
- Personal information
- Products which customer is willing to have
- History of interaction, past purchasing data.
- Recent interests
- Customer's strategy, vision and mission
- Main competitors of the customer
- Their budget
- Financial situation
- Recent investments
- Customer's contact employee's features

As mentioned in the previous section, data collection constitutes the base of customer segmentation. However, data collection is usually a problematic issue because the customers are not always eager to share many information. To encourage the customers to share those information, companies may offer several kinds of promotions and discounts such as shopping cards which are distributed by supermarkets and offer numerous discounts. Most of the customers are not even aware of how they provide very large amount of information to the companies, because through the card, the company can track all shopping attitudes and all activities of the customers and therefore is able to segment its customers according to these attitudes.

For the supermarkets or other businesses that have a high interaction with individuals, it is relatively easy to collect data in comparison with other companies, especially the ones which work Business-to-Business. In this case, collecting data belonging to other business enterprises related to their recent projects, strategy of the company, their budget or financial situations or the products and services would be harder, hence these data should be gathered through some smart ways without annoying the customers. Mostly, it is done by tracking the interactions and try to make inferences for better understanding.

After the collection of the customer data, the second challenge is to select the most appropriate method to analyze the data for segmentation. As mentioned before, a variety of segmentation methods exist in the literature and in practical applications. In this study, K-Means algorithm is chosen due to its advantages such as easy implementation to any kind of data, easy integration and use with other algorithm(s) and it is easily programmable. However, it has some drawbacks as well, one of them is randomness that it starts with randomly selected clusters, which might result in a narrow solution set. To overcome this weakness, Simulated Annealing Algorithm is integrated with K Means Algorithm and a novel hybrid method is introduced in this thesis. Simulated Annealing Algorithm is described in Section 2.1.1.

1.3 Scope and Purpose of This Thesis

1.3.1 Aims and Scope of This Thesis

To gain a competitive advantage against the competitors, the companies still seek to create different marketing strategies based on modified and customer-oriented models of conventional marketing strategies however to be more successful and move the company forward, it is insufficient to create another type of strategy which is not an improved version of the traditional ones. The companies must also create mix of customer centric strategies as different than the traditional ones. Using proper customer-centric marketing and sales strategies companies can retain their existing customers and attract new customers. The first condition of an effective marketing strategy is to determine the most important (or highest value) customer segment(s) based on a variety of unique customer characteristics to detect their needs and expectations, identify trends, to customize marketing plans and product/service development activities according to these expectations and trends, to advertise campaigns and deliver relevant products, and to offer rewards to the customers in these segments for transactions or visits. Customer segmentation personalizes the messages to individuals to improve the communication with the intended groups. The profits of the company can be significantly improved by an efficient customer segmentation model.

Considering these facts, in this thesis, an integrated decision support system for analytical CRM based on customer segmentation is developed with the main aim of facilitating the customer segmentation process by grouping (or clustering) the customers according to their specific features; to specify the highest value customer segments (i.e. clusters), to detect the trends as well as the needs and expectations of the customers in these clusters truly and hence to improve the marketing practices of the companies. Companies in a wide range of sectors can utilize the decision support system developed in this study to make enhanced decisions related to sales and marketing activities.

1.3.2 Brief Description of the Methodology Developed in This Thesis

The decision support system developed in this thesis combines two consecutive levels. In the first level, data clustering is performed by a novel hybrid approach that integrates Simulated Annealing Heuristic and K Means Algorithms. K-Means algorithm has a drawback in terms of randomness because of randomly selected first clusters which are then used to calculate the final clusters. With the aid of the integration of Simulated Annealing Heuristic to the initial part of the K Means Algorithm, randomness is reduced and the hybrid algorithm is created which is more systematic and reliable. The second level handles the problem of selecting the most important cluster combining two well-known Multi Criteria Decision Making methods, namely Analytic Hierarchy Process (AHP) and TOPSIS. In this level, AHP is used to determine the weights of each feature related to customers representing the relative importance pf these features and TOPSIS is used to obtain the final ranking of clusters specified in the first level taking into account the customer features and corresponding weights. Percin (2009) proposes a hybrid method of AHP and TOPSIS to evaluate third party logistics providers in their study. AHP enables the users to give their preferences or opinions as an input by utilizing the scale 1-9. The scale is already very beneficial for experts or decision makers to conclude a decision. That's why, hybrid method of AHP and TOPSIS offers several benefits. It is a systematic and reliable method due to the fact that it is capable of capturing the expert's opinions or considering previous experiences, therefore once AHP weights are used in TOPSIS, it makes the decision process more realistic and rational. On the other hand, during raking process of the clusters, hierarchical structure, pairwise comparisons and consistency checks are taken into account.

To facilitate the entry of data and monitoring the results, a user friendly interface is also designed. Although the decision support system developed in this thesis can be used in a wide range of industries, a case from tourism sector is handled to show the practicability of the method which is proposed. To this aim, the case presented in Srihadi, Hartoyo, Sukandar & Soehadi (2016) is utilized along with the data set related to tourism customers. In the study, the data is collected through questionnaires from

the foreign visitors who come to Jakarta / Indonesia from different regions. Demographic profiles of respondents such as age, gender, race, employment status, purpose trip etc. and AIO (Attitudes, Interests and Opinions) Statements are proposed in their study. The decision support system is developed using VB.Net software.

1.3.3 The Main Novelties, Originalities and Contributions of This Thesis

The main novelties and contributions of this thesis are summarized as follows;

1. This study proposes a novel customer segmentation approach by integrating Simulated Annealing Heuristic and K-Means Clustering Algorithm. In the current literature on the related field, the studies that employ K-Means Clustering Algorithm suffer from the major drawback of this algorithm; excess randomness while it is working, especially to find the initial cluster. On the other hand, Simulated Annealing Heuristic is commonly used to approximate the solutions to the optimal value. In this study, to the best of our knowledge, first time in the literature, by integrating a metaheuristic algorithm, namely Simulated Annealing Algorithm, to the K-Means Clustering algorithm, randomness in finding the initial cluster in K-Means Clustering algorithm is decreased, so that the solutions are further approximated to the optimal values and the algorithm works faster.
2. The use of MCDM methods is commonly experienced in CRM. To the best of our knowledge, the integration of two multi-criteria decision making methods is firstly applied to determine the criteria weights and then rank the important clusters is rare in the literature.
3. Along with the abovementioned contributions to the related literature, this thesis has important contributions in terms of implications for practice. A user friendly software (in VB.net) is developed to be used in real life cases by the companies which need to segment their customers based on specific features and specify their highest ranked customers. By using the software and demonstrating the results based on academically proved methods, companies will be able to improve their marketing strategies and meet their customers' needs more accurately. The software has open source codes and it can be tailored easily to any company.

CHAPTER TWO

METHODS

2.1 Heuristic Approaches

Based on the traditional approaches, in order to find the optimum solution of the decision problem or a solution space involving the feasible solutions to solve a decision problem, many mathematical or statistical models are available such as linear programming and regression analysis. However, real life problems, especially in large scale problems with a large data set, many decision variables and nonlinear relationships between those variables and parameters, impose a complex mathematical or statistical model which includes too much information and many assumptions. However, there is still a possibility that it may not give accurate or reliable solutions. On the other hand, it generally takes too much time to create the model, to collect all data, to program it and to take the solutions. Heuristic approaches are widely utilized in such cases, to get a reliable solution in a timely manner to a complex and large scale problem. In their study, Gigerenzer & Gaissnaier (2011), concluded that a complex mathematical programming model has much more information than a heuristic and it is highly possible to make errors in its extensive assumptions and computations (Gigerenzer & Gaissnaier, 2011).

Heuristic approaches can be described as techniques to be used to determine not the optimum solution but the approximated value to the optimum solution more quickly and even more accurately than many other complex mathematical or statistical methods to solve large scale and/or nonlinear decision problems. Heuristics may also need less effort in terms of constructing the model and programming and also can be used for many types of decision problems. It would not be reasonable to claim that heuristics are always better than the other solution approaches for decision making, however it's very clear that it has so many benefits if the problem is large scale, complex and/or nonlinear.

Silver (2004) presented a study on the overview of the heuristic solution approaches used in the literature and categorized them into:

- Solutions which are generated randomly
- Decomposing or partitioning the problem
- Inductive methods
- Techniques which reduce the field of feasible solutions
- Methods for approximation
- Constructive methods
- Methods which aims to reach local improvements

A variety of heuristics are used to solve the decision problems in the literature, some of widely used approaches include taboo search algorithm, genetic algorithm, neural network, multilevel refinement, variable neighborhood search, ant colony search and simulated annealing algorithm.

Taboo search algorithm and genetic algorithm are commonly used in the literature. The logic and the procedure of those algorithms are described as follows:

1. Taboo search algorithm: Taboo Search algorithm is a memory-based method. It searches for the best solution by comparing the alternatives to each other. During the search, it records each move in the memory. Main principle in this algorithm is to avoid returning to same and already visited solutions alternatives. This is called as cycling, therefore the algorithm prevents cyclic movements. In this way it aims to find the best solution beyond the local optimum points.

The pseudo code representing the procedure of the Taboo search algorithm is given as follows;

```
Input: TabuList_size
Output: S_best
S_best ← ConstructInitialSolution()
TabuList ← ∅
While (¬ StopCondition())
  CandidateList ← ∅
  For (S_candidate ∈ Sbest_neighborhood)
    If (¬ ContainsAnyFeatures(S_candidate, TabuList))
      CandidateList ← S_candidate
    End
  End
  S_candidate ← LocateBestCandidate(CandidateList)
  If (Cost(S_candidate) ≤ Cost(S_best))
    S_best ← S_candidate
    TabuList ← FeatureDifferences(S_candidate, S_best)
    While (TabuList > TabuList_size)
      DeleteFeature(TabuList)
    End
  End
End
Return (S_best)
```

Figure 2.1 Taboo Search Algorithm - Procedure

2. Genetic algorithm: Genetic algorithm is another kind of search heuristic. It is inspired by Charles Darwin's theory which is called as natural evolution. The algorithm considers the problem as a natural selection problem and during the search of solutions, it seeks to evolve the solution. In the natural selection, individuals which are the fittest are selected for regeneration to provide offspring to reach new generation. More fit parents, better new generation and their offspring will be better than the parents. In the algorithm selects the best algorithm among the others based on fitness function.

Five phases are considered in a genetic algorithm:

- Starting solution (population)
- Function of fitness
- Election
- Evolution
- Alteration (Mutation)

The general procedure of the Genetic Algorithm is given as follows:

```
START
Generate the initial population
Compute fitness
REPEAT
    Selection
    Crossover
    Mutation
    Compute fitness
UNTIL population has converged
STOP
```

Figure 2.2 Genetic algorithm – Procedure

In this study, Simulated Annealing Algorithm is used to support K-Means Algorithm in order to decrease the randomness in finding the initial clusters, to get a more narrow feasible solution space and finally to get the clustering results more quickly. Simulate Annealing Algorithm is chosen due to the fact that it is easily programmable, it is easy to integrate to K-Means algorithm, it is easy to modify, it does not involve a complex or complicated procedure. Simulated Annealing Heuristic is applied to many hard combinatorial optimization problems in the literature. Simulated Annealing Heuristic is explained in detailed in the Section 2.1.1.

2.1.1 Simulated Annealing Algorithm

Metropolis et al. introduced Simulated Annealing (SA) Algorithm firstly (1953). It is a heuristic which helps to reach not the optimum solution but the approximated solution to the global optimum solution while searching in the large feasible solution area.

SA procedure includes probability to decide if it needs to search for another candidate solution in the feasible solution area or accept the current solution as the solution.

Annealing process of metals in metallurgy science inspires simulated annealing heuristic. Annealing process is a method for treatment which is used to change physical or chemical specifications of a metal or a materials and the material becomes more durable and workable. This method helps to heat the material till the temperature of its recrystallization and accordingly to decrease the temperature slowly and controlled. Heating process raises up the inner energy of the material and makes the atoms free. After the atoms move to other locations inside of the material and inner temperature is decreased, new and stable structure of the material becomes the current configuration of the material. Different configurations of the material are observed during the heating and cooling process because of the inner energy change and different locations of the atoms. In fact, Simulated Annealing does the same as annealing process. The temperature of the system increases and during the process SA observes all configurations of the alternatives and compares them with each other in terms of a condition which is determined before and once the system reaches to the lowest temperature, it accepts the last solution as the closest solution to the global optimum. Yu & Lin (2015) stated that SA firstly constructs an initial solution and then improves the initial solution by the above mentioned procedure through three neighborhood search mechanisms.

SA is much simple method for implementation, that's why it has a wide implementation field in design and optimization studies.

Another advantage of the SA algorithm is that it is easy to integrate with the other methods to improve the performance of those methods and it makes it possible to implement it on many kinds of decision problems with several different topics.

The standard procedure of the SA algorithm is illustrated in the following figure. In the beginning of the algorithm, initial temperature, final temperature and rate of decrease in temperature are determined, then the local optimum point is being searched by decreasing the temperature, as in the annealing process, until the algorithm reaches the final temperature. At the final temperature, the last solution that the algorithm finds

is selected as the best solution. It is not a usually global optimum point but relatively very close to the global optimum solution. Although it seems as a user dependent algorithm because of the user selection of the temperatures and the decrease rate, in the literature a range of modified simulated annealing algorithms which overcome this shortage are studied.

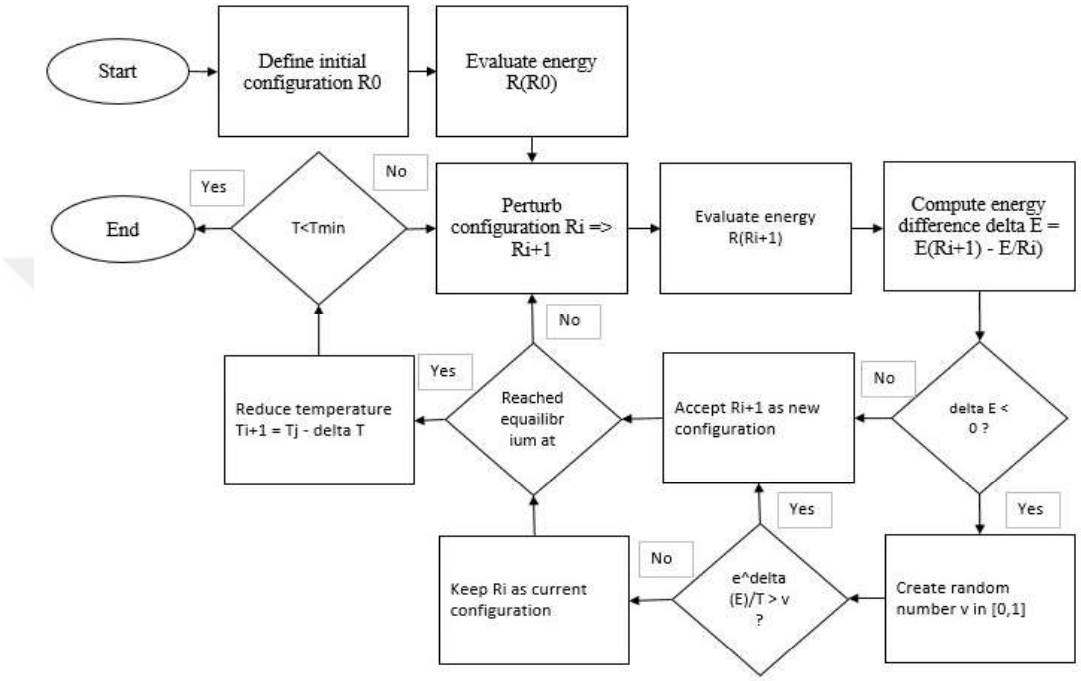


Figure 2.3 Simulated Annealing Algorithm Flow

2.2 Data Mining Methods

Rygielski et al. (2002) defined Data Mining as an advanced data search ability which is provided by statistical algorithms to find out patterns, trends and correlations in data. On the other hand data mining can be described in simple terms as it is a way to determine meaning in data. Data Mining organizes the data and creates a structured source to perform further processes such as predictions, forecasting or clustering.

In recent years Data Mining is considered as a powerful way of gaining competitive advantage for the companies. By mining the related data, companies can reach

significant amount of meaningful information regarding to their customers. It can help them to predict the behavior of customers and to understand their individual needs and expectations. In this way, companies can increase the retention rate of existing customers and accordingly the number of customers which are likely to remain loyal.

According to Injadat, Salo and Nassif (2016), there are 19 main data mining techniques widely used in the literature including;

- Artificial Neural Network (ANN), with high level of abilities that are suitable for the analysis of high dimensional data.
- Bayesian Networks (BN), which is very effective for text clustering.
- Decision Trees (DT), which is especially beneficial for forecasting the importance of the variables.
- Density Based Algorithm, which do not require the pre-specified number of clusters and noise filtering.
- Hierarchical Clustering, which relies on a fully specified similarity matrix.
- K-Nearest Neighbors, which is considered as one of the simplest and most discriminative classifiers.
- K-Means Clustering is doing well especially for figuring out a low amount of clusters. It is efficient in terms of speed and converges to local maxima quickly. In addition, it performs well on multi-dimensional sets of data as well.

2.2.1 Big Data

In the early 2000s, with the rapid growth in information and data systems/technologies, the concept of “big data” has gained momentum. Big data is defined as information assets, both structured and unstructured, which have high-volume, complexity, variety and which needs advanced technologies to store, distribute, manage and analyze (Gandomi & Haider, 2015).

"Big data" is a new term relatively, however the behavior of recording and storing large amounts of information for analysis is very old. The industry analyst Doug Laney articulated the big data as the three Vs, Volume, Velocity and Variety. The volume of the data collected has increased due to a wide range of data sources such as social media, transactions or sensors. The velocity of data flows which refers as the speed of the data and high velocity information streams have to be dealt with in a timely manner especially in real time applications. Data may be collected in a variety of formats such as structured data (numeric data in traditional databases) and unstructured text documents (video, audio, financial transactions, emails). In addition to the increasing varieties and velocities of data, data streams may be highly instable due to periodic/seasonal peaks and trends. This data loads can be quite challenging to manage. All these factors increase the complexity of data streams, which makes it difficult to match, link and transform data to make it more useful. Due to the increased complexity, it is necessary to correlate and connect the data and construct the necessary hierarchies and linkages, or large data can quickly spiral out of control.

Not only the volume of data is important for companies but also what organizations do with the data matters. It is essential to emphasize that the valuable part of it does not come from the raw form of the data. The major value comes after processing the data and its analysis and management. The advances in management of big data has to be together with developments in the acts of supporting the decisions and innovation of products or services.

Big data collected from any source should be analyzed reliably to get different perspectives to make better decisions in a timely manner for strategic business moves such as new product development, optimized offerings to customers and suppliers, cost reductions, time reductions. More specifically, the analysis and management of big data leads to;

- Determining main causes of failures, problems and defects quickly,

- Specifying customer's behaviors and buying habits to generate new coupons at the point of sale
- Recalculating the risk portfolios and redesigning the procedures according to the findings
- Detecting failures in advance before they affect other processes and the organization

Managing big data is important for all organizations across every industry. Some of the industries that the analysis and management of big data is highly important include banking, education, health care, manufacturing and retail.

In the new era, as the people move around the world, they generate megabytes of data in every second. For instance, mobile phones, surfing in search engines, the time spent in social media (pages visited, comments under photos, interests given in timeline), the subjects of our posts, blogs etc. are able to give very large insight to the companies to determine their strategies in an accurate way by analyzing this big data with statistical and heuristic algorithms. According to the last researches, people send 2.9 million emails in each second, they upload 20 hours of video in each minute and share dozens of millions of tweets in a week. When you read those, probably those numbers are already increased.

Social media and website activities are simple but helpful for better understanding of the big data however the other common way to collect big data is to track the customers' activities related to the interactions with the company. When a customer invests for a product or service, and when a customer searches for a different product or service, records of the phone and mail interactions etc. can also compose the big data. A good and structured analysis of the big data will give a meaningful information to the company to be used in CRM activities.

Benefits of using and analysis of big data in terms of CRM can be summarized as follows:

1. Empowers management to make better and structured decisions
2. Helps identify trends to stay competitive
3. Enhances the productivity, efficiency and loyalty of staff in doing valuable tasks and creating solutions for the issues
4. Identifies and acts upon opportunities
5. Very low risk in special secured systems and more plans and decisions which are based on data.
6. Validates decisions
7. Helps in selecting target customers with a structured segmentation and ranking of them

To do those all, big data management needs statistical techniques and heuristics to achieve meaningful information and a base for structured segmentation and improved decisions for the companies. Statistical methods for management of big data in CRM are described in the following section.

2.2.2 Statistical Methods For Data Mining

Data mining is a way of discovering a meaningful pattern through big data using different kinds of statistical methods. For example, in a supermarket, data can be collected with thousands of bills which are printed out to the customers. As a real life example The American retailer Wal-Mart makes more than 20 million transactions daily (Babcock 1994). However, in order to find out a pattern, a relationship or a meaningful information data mining is needed. It will be impossible to see any pattern or to make managerial decisions with big data without data mining. Statistical methods to find a pattern or cluster the data make the big data more meaningful and more useful.

Hand (1998) states that there are two aims of statistical data mining methods, finding global models and finding patterns. Data mining methods for finding global models include analysis of regression and clustering, methods of supervised classification methods, Bayesian networks. Strategies of pattern finding is actually about detection of anomalies. It contains dynamic and interactive graphical methods. In this case, computer search is vital to find a pattern with less effort and less errors.

Ngai et al.(2009) overviewed 900 articles (between 1992 and 2007) related to data mining in CRM and concludes that the data mining techniques which are studied in those articles can be classified into the types given as follows:

1. Association
2. Classification
3. Clustering
4. Forecasting
5. Regression
6. Sequence discovery
7. Visualization

According to Ngai et al.(2009), the most popular data mining algorithms in the literature are given as follows:

1. Rule of Association
2. Decision tree
3. Genetic algorithm
4. Neural Networks
5. K-Nearest neighbor/ K-Means algorithm
6. Linear / Logistic Regression

It's clearly seen that data mining techniques are very popular among researchers and there is a variety of statistical methods to perform data mining. Data mining with statistical methods constitutes a vital part of customer relationship management practices, because the data base will be constructed and organized by data mining and

if the base is well-designed then there will be no problem to build more on it. Therefore, with those statistical methods, firstly a clear base for the customer data should be constructed for an efficient CRM application. Then,in the second step, the aim will be to decide the most valuable customers and then to define and meet their individual needs. Accordingly, customer retention will be maintained to sustain a long term and pleasant relationship with customers.

2.2.3 Statistical Data Mining Methods for Data Clustering

Aim of clustering (or cluster analysis) is to group a set of objects which are relatively similar to each other in the same group but different than the ones in the other groups. In other words, clustering of big data refers to the division of a given data set into a number of groups which have objects as similar as possible. For a set of data, clustering algorithm can be used to assign each point into a cluster. Basde on the theory, the data points in the same cluster have to have similar characteristics, at least very close to each other, and the data regarding to each group must represent and cover all data belonging to each member of the group as well. Whereas points of data which are in different groups have to have highly different characteristics.

Clustering is a phenomenon in the literature related to data mining because it has a wide application area in plenty of decision problems in varying fields, which struggle to discover the subsets of their data set and to seek to find out a meaningful information from it. It can be used in such fields as computer vision, gene expression analysis, Web page clustering, text mining and many companies which occupy in service industry.

In order to cluster the big data, numerous methods exist and are used in different kinds of studies in the literature and in real life applications. According to the type and amount of data as well as the way to determine the segments, each algorithm has its own advantages. Some of those clustering algorithms are given as follows:

Agglomerative algorithms and divisive algorithms can be grouped as hierarchical methods. Probablistic clustering, K-Medoid methods and K Means can be grouped as

Methods of Partitioning Relocation. Density based Connectivity Clustering and Density Functions Clustering can be grouped as Methods of Density based Partitioning. Other clustering techniques can be summarized as constraint-based clustering, graph portioning, clustering algorithms and supervised learning, clustering algorithms in machine learning.

Big data has no meaning without further processing, it would seem as only many different numbers and letters which occupies much space in the software. However, with the use of statistical data mining and clustering techniqueks, the big data will be more organized and meaningful and would constitute a good base to make a decision. The advantage of the clustering methods in CRM is basically that they are able to process the data and reach some remarkable and beneficial thoughts about the customers.

2.2.3.1 K-Means Clustering

K-means clustering (Mac Queen, 1967) is a popular and widely used method to simply divide a set of data into k numbers of groups. It firstly selects k numbers of clusters randomly and then iteratively seeks to refine the clusters based on the algorithm's procedure.

As it is mentioned above, the aim of the algorithm is to divide a given data set into clusters, then the clustering area will be optimized. Clusters can be identified as high density fields in the analytical space separated by low density fields. There are many benefits of K-Means algorithm. Firstly, algorithm procedure is not very complicated and it is simple to apply. That's why it can simply be integrated into any heuristic approach or other statistical method. It does not find the optimum solution always but searches for a feasible solution which is approximated to the optimum solution. Based on the data volume, variety and complexity, computational time for searching the solutions might take much time. To cope with this situation and to decrease the computational time, it can be combined with another algorithm (heuristic or exact). It

is also important to note that, there is no clustering algorithm which dominates the others and which can

be selected as the best. Almost all algorithms, including K-Means, are valid and justifiable clustering algorithms. K-Means procedure is depicted in the following figure:

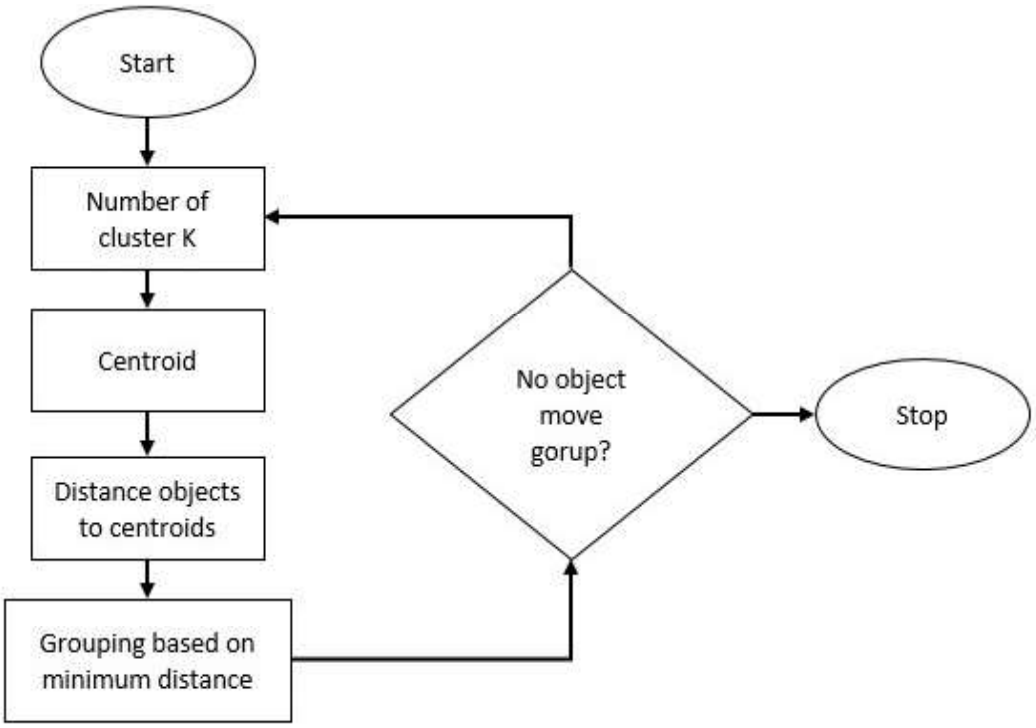


Figure 2.4 K-Means Algorithm Procedure

The steps of K-Means Algorithm are given as follows:

1. Define the number (n) of clusters which is aimed to determine,
2. Assign n customers as draft clusters to use in the next steps,
3. Calculate the distances between each customer to each cluster,
4. Assign each customer to a cluster, depending on minimum distance (Basically, the closest cluster to the customer must be the cluster of that customer),
5. Calculate each clusters' analytical data points by taking average of all data points of customers which belong to that cluster in first iteration,
6. Calculate the distances between each customer to each cluster,
7. Assign each customer to a cluster, depending on minimum distance,

8. If the data points of clusters are the same with the previous iteration, then stop and accept the clusters as finalized clusters. Else, go to step 5,

In order to clarify the logic of the algorithm and to demonstrate the clustering procedure, a simple example is given in the following:

Assumptions & Notations:

- A data set about tourism industry is handled which includes 15 customers and 5 criteria (Age, Gender, Race Group, Employment Status, Purpose of Trip). A code is given to each criterion, for example for Gender criterion, codes are 1 and 2. The customers are represented with codes such as [6,2,1,3,2].
- The aim is to perform segmentation and cluster the customers into 3 clusters
- The notations are given as follows;
  C1: Cluster 1, C2: Cluster 2, C3: Cluster 3, Nx: Customer #x, N1: Customer #1, N2: Customer #2, N3: Customer #3 … etc.
- C1, C2 and C3 are *randomly* selected among the customers and determined as [1,1,2,1,1], [3,1,3,2,1] and [6,2,1,4,3] from the customer data N1, N8 and N14.
- The Euclidean distances are calculated as follows:

The distance from N1 is equal to C1 = [1,1,2,1,1] to N2 = [3,2,3,3,1]:

(1-3) ^ 2 + (1-2) ^ 2+(2-3) ^ 2+(1-3) ^ 2+(1-1) ^ 2= 3.16

After all calculations were completed, summary table with the distances will be as following:

Table 2.1 K Means Example, Summary table of first iteration

Age	Gender	Race Group	Employement Status	Purpose of Trip	Customer Code					Distance to Cluster1	Distance to Cluster2	Distance to Cluster3	Min Distance	Assigned Cluster
A1	B1	C2	D1	E1	1	1	2	1	1	***0,00***	6,32	2,06	0,00	C1
A3	B2	C3	D3	E1	3	2	3	3	1	3,16	4,24	1,41	1,41	C3
A6	B2	C2	D3	E1	6	2	2	3	1	5,48	2,45	3,46	2,45	C2
A5	B1	C1	D2	E2	5	1	1	2	2	4,36	2,65	3,00	2,65	C1
A1	B2	C1	D1	E3	1	2	1	1	3	2,45	5,83	3,74	2,45	C1
A2	B2	C3	D1	E1	2	2	3	1	1	1,73	5,74	1,73	1,73	C3
A4	B1	C2	D1	E1	4	1	2	1	1	3,00	4,36	1,73	1,73	C3
A3	B1	C3	D2	E1	3	1	3	2	1	2,45	4,69	***0,00***	0,00	C3
A6	B1	C2	D4	E2	6	1	2	4	2	5,92	1,73	3,87	1,73	C2
A1	B2	C3	D4	E2	1	2	3	4	2	3,46	5,48	3,16	3,16	C1
A2	B1	C2	D1	E1	2	1	2	1	1	1,00	5,57	1,73	1,00	C1
A4	B2	C3	D1	E1	4	2	3	1	1	3,32	4,58	1,73	1,73	C3
A1	B2	C1	D1	E2	1	2	1	1	2	1,73	5,92	3,32	1,73	C1
A6	B2	C1	D4	E3	6	2	1	4	3	6,32	***0,00***	4,69	0,00	C2
A3	B1	C2	D3	E1	3	1	2	3	1	2,83	4,00	1,41	1,41	C3

- In each row, minimum distances are found and accordingly each customer is assigned to a cluster. For example; for customer [2,1,2,1,1], Euclidean distances are calculated as follows as in the Table 2.1:

  C1: 1.00

  C2: 5.57

  C3: 1.73

  Minimum distance among those is the distance to C1, that's why this customer [2,1,2,1,1] is assigned to cluster 1.

- Current cluster mediums are calculated and updated based on the method given in step 5 in K Means Algorithm steps, as in following:

Cluster1 = [1.83, 1.5, 1.66, 1.83], Cluster2 = [6.00, 1.67, 1.67, 3.67, 2] and

Cluster3 = [3.16, 1.5, 2.66, 1.83, 1].

Calculation of cluster 2 data points is given as follows;

Cluster 2 has 3 customers as shown in Table 2.2 as N3, N9 and N14.

Cluster 2 new data points must be;

C2 = [avg (6; 6; 6), avg (2; 1; 2), avg (2; 2; 1), avg (3; 4; 4), avg (3; 4; 4), avg (1; 2; 3)], therefore

Cluster2 = [6.00, 1.67, 1.67, 3.67, 2]

Below Table 2.2 shows the codes of customers in the first iteration (rows: customer numbers, columns: criteria, values: weights of each criteria per customer).

Table 2.2 K-Means Example –Clusters after first iteration (Cluster1)

Customer	Cluster 1				
N1	1	1	2	1	1
N4	5	1	1	2	2
N5	1	2	1	1	3
N10	1	2	3	4	2
N11	2	1	2	1	1
N13	1	2	1	1	2
avg.	1,833	1,500	1,667	1,667	1,833

Customer	Cluster 2				
N3	6	2	2	3	1
N9	6	1	2	4	2
N14	6	2	1	4	3
avg.	6,000	1,667	1,667	3,667	2,000

Customer	Cluster 3				
N2	3	2	3	3	1
N6	2	2	3	1	1
N7	4	1	2	1	1
N8	3	1	3	2	1
N12	4	2	3	1	1
N15	3	1	2	3	1
avg.	3,167	1,500	2,667	1,833	1,000

- After the second iteration, the new distances and clusters are changed as given in the following table:

Table 2.3 K Means Example, Summary table of second iteration

Age	Gender	Race Group	Employement Status	Purpose of Trip	Customer Code					Distance to Cluster1	Distance to Cluster2	Distance to Cluster3	Min Distance	Assigned Cluster
A1	B1	C2	D1	E1	1	1	2	1	1	1,48	5,80	2,47	1,48	C1
A3	B2	C3	D3	E1	3	2	3	3	1	2,42	3,51	1,32	1,32	C3
A6	B2	C2	D3	E1	6	2	2	3	1	4,49	1,29	3,18	1,29	C2
A5	B1	C1	D2	E2	5	1	1	2	2	3,30	2,16	2,72	2,16	C2
A1	B2	C1	D1	E3	1	2	1	1	3	1,79	5,80	3,52	1,79	C1
A2	B2	C3	D1	E1	2	2	3	1	1	1,79	5,10	1,55	1,55	C3
A4	B1	C2	D1	E1	4	1	2	1	1	2,49	3,56	1,44	1,44	C3
A3	B1	C3	D2	E1	3	1	3	2	1	2,05	3,87	0,65	0,65	C3
A6	B1	C2	D4	E2	6	1	2	4	2	4,82	0,82	3,80	0,82	C2
A1	B2	C3	D4	E2	1	2	3	4	2	2,86	5,20	3,28	2,86	C1
A2	B1	C2	D1	E1	2	1	2	1	1	1,24	4,97	1,66	1,24	C1
A4	B2	C3	D1	E1	4	2	3	1	1	2,80	3,74	1,32	1,32	C3
A1	B2	C1	D1	E2	1	2	1	1	2	1,36	5,72	3,07	1,36	C1
A6	B2	C1	D4	E3	6	2	1	4	3	4,99	1,29	4,44	1,29	C2
A3	B1	C2	D3	E1	3	1	2	3	1	2,05	3,32	1,44	1,44	C3

- Distances and minimum distances are determined and each customer is assigned to the clusters accordingly in iteration 2.
- Current cluster mediums are calculated and updated and finally the clusters seem as given in the following tables:

Table 2.4 K-Means Example –Clusters after second iteration

Customer	Cluster 1				
N1	1	1	2	1	1
N5	1	2	1	1	3
N10	1	2	3	4	2
N11	2	1	2	1	1
N13	1	2	1	1	2
avg.	1,200	1,600	1,800	1,600	1,800

Customer	Cluster 3				
N2	3	2	3	3	1
N6	2	2	3	1	1
N7	4	1	2	1	1
N8	3	1	3	2	1
N12	4	2	3	1	1
N15	3	1	2	3	1
avg.	3,167	1,500	2,667	1,833	1,000

Customer	Cluster 2				
N3	6	2	2	3	1
N9	6	1	2	4	2
N4	5	1	1	2	2
N14	6	2	1	4	3
avg.	5,750	1,500	1,500	3,250	2,000

- Iteration 3 is performed and the results are depicted in the following table:

Table 2.5 K Means Example, Summary table of third iteration

Age	Gender	Race Group	Employement Status	Purpose of Trip	Customer Code					Distance to Cluster1	Distance to Cluster2	Distance to Cluster3	Min Distance	Assigned Cluster
A1	B1	C2	D1	E1	1	1	2	1	1	1,20	5,40	2,47	1,20	C1
A3	B2	C3	D3	E1	3	2	3	3	1	2,73	3,51	1,32	1,32	C3
A6	B2	C2	D3	E1	6	2	2	3	1	5,08	1,29	3,18	1,29	C2
A5	B1	C1	D2	E2	5	1	1	2	2	3,95	2,16	2,72	2,16	C2
A1	B2	C1	D1	E3	1	2	1	1	3	1,62	5,80	3,52	1,62	C1
A2	B2	C3	D1	E1	2	2	3	1	1	1,80	5,10	1,55	1,55	C3
A4	B1	C2	D1	E1	4	1	2	1	1	3,04	3,56	1,44	1,44	C3
A3	B1	C3	D2	E1	3	1	3	2	1	2,42	3,87	0,65	0,65	C3
A6	B1	C2	D4	E2	6	1	2	4	2	5,41	0,82	3,80	0,82	C2
A1	B2	C3	D4	E2	1	2	3	4	2	2,73	5,20	3,28	2,73	C1
A2	B1	C2	D1	E1	2	1	2	1	1	1,43	4,97	1,66	1,43	C1
A4	B2	C3	D1	E1	4	2	3	1	1	3,23	3,74	1,32	1,32	C3
A1	B2	C1	D1	E2	1	2	1	1	2	1,11	5,72	3,07	1,11	C1
A6	B2	C1	D4	E3	6	2	1	4	3	5,57	1,29	4,44	1,29	C2
A3	B1	C2	D3	E1	3	1	2	3	1	2,50	3,32	1,44	1,44	C3

- Clusters after iteration 3 are given as follows:

Table 2.6 K-Means Example –Clusters after third iteration

Customer	Cluster 1				
N1	1	1	2	1	1
N5	1	2	1	1	3
N10	1	2	3	4	2
N11	2	1	2	1	1
N13	1	2	1	1	2
avg.	1,200	1,600	1,800	1,600	1,800

Customer	Cluster 3				
N2	3	2	3	3	1
N6	2	2	3	1	1
N7	4	1	2	1	1
N8	3	1	3	2	1
N12	4	2	3	1	1
N15	3	1	2	3	1
avg.	3,167	1,500	2,667	1,833	1,000

Customer	Cluster 2				
N3	6	2	2	3	1
N9	6	1	2	4	2
N4	5	1	1	2	2
N14	6	2	1	4	3
avg.	5,750	1,500	1,500	3,250	2,000

The iterations should be performed till the iteration gives the results which is the same with the previous iteration. Therefore, it seems that the results in iteration 2 and iteration 3 are the same in this example. It means that K means algorithm is completed and the data is clustered.

Iteration 2 Results:

Cluster1 = [1.2, 1.6, 1.8, 1.6] Cluster2= [4.84, 1.52, 1.56, 2.92, 1.96]

Cluster3= [3.167, 1.5, 2.66, 1.83, 1]

Iteration 3 Results (Solution):

Cluster1= [1.2, 1.6, 1.8, 1.6] Cluster2 = [4.84, 1.52, 1.56, 2.92, 1.96]

Cluster3= [3.167, 1.5, 2.66, 1.83, 1]

2.3 Multi Criteria Decision Making Methods

Decision can be described as a reasonable choice among alternatives. Decision making can be defined as a process of selection among alternatives of an action for a purpose. Large number of complex factors in the problem determines the level of complexity of the problem. Multi Criteria Decision Making Problems are more complex problems than those with single criterion. Multi criteria decision making (MCDM) was found to be a beneficial research area in the early 1970s. Since that time, a variety of methods and approaches, which have different decision making procedures, have been developed, improved, integrated with other methods and published in the literature. MCDM methods are beneficial to solve MCDM problems because those methods take all the complex factors into account and are able to find the most preferred solution in the feasible solution region.

MCDM is commonly used to solve decision making problems. MCDM methods make use of Operation Research models to cope with complex decision problems including different alternatives and a set of decision criteria. Multi criteria decision making methods are classified into two major divisions which are called Decision Making of Multi Objective (MODM) and Multi Attribute (MADM). Numerous methods which are developed under those major divisions exist in the literature. Although the methods have a common way to approach the decision problem based on the rational decision making strategies, which is shown by the following figure, each method has its own characteristics and solution procedure.

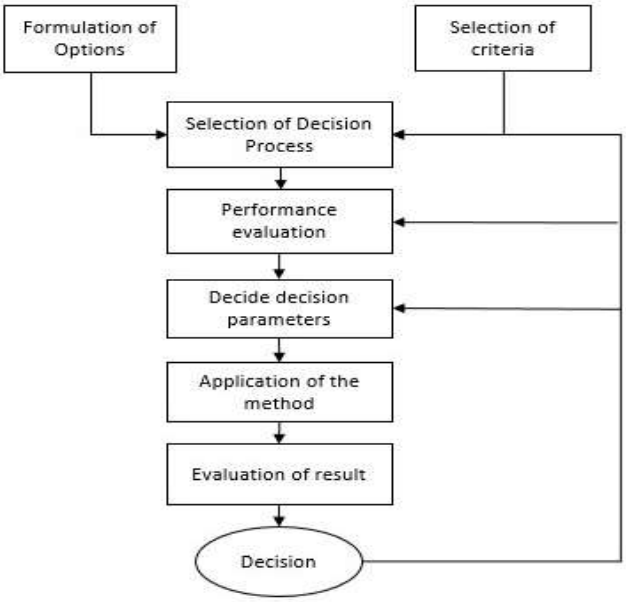


Figure 2.5 MCDM process flow

According to Opricovic & Tzeng (2004), the main steps of MCDM process are as follows:

a. Construction of a system and defining a criterion for evaluation that relates the abilities of system to achieve goals
b. Generation of alternatives
c. Judging the alternatives based on the weighted values of criteria or evaluation functions depending on the MCDM method
d. Implementation of a proper multi criteria analysis algorithm
e. Reaching one of the alternatives as the best solution
f. If the reached solution is not good enough based on the evaluation criteria or function of evaluation then iterate the previous steps

A variety of MCDM methods are studied in the literature. Some of the most widely used MCDM methods are explained in the following table along with their benefits (Pohekar & Ramachandran, 2003):

Table 2.7 K Means Example, Summary table of third iteration

MCDM Method	Abbreviation	Brief Explanation
Weighted sum method	WSM	Single spatial problems; n number of alternative solutions and k criteria.
Weighted Product Method	WPM	N number of alternative solutions and k crtieria, in fact similar as WSM but it is used also for multiple spatial problems depending on the customer.
The Elimination and Choice Translating Reality	ELECTRE	This method is capable of handling discrete criteria of both quantitative and qualitative in nature and provides complete ordering of the alternatives. The problem is to be so formulated that it chooses alternatives that are preferred over most of the criteria and that do not cause an unacceptable level of discontent for any of the criteria. The concordance, discordance indices and threshold values are used in this technique. Based on these indices, graphs for strong and weak relationships are developed.It only produces a core of leading alternatives. This method has a clearer view of alternatives by eliminating less favorable ones, especially convenient while encountering a few criteria with a large number of alternatives in a decision making problem.

Table 2.7 continues

Analytical Hierarchy Process	AHP	Reconstruction of a problem into hierarchical structure. Top of it, the goal stays. The other levels of it, there are criteria and sub criteria. Finally, in the bottom, solution alternatives stay. Pairwise comparisons are performed to determine the weights of the criteria based on the relative preference. During comparison, determination of scores, Saaty's (1980) 1-9 scale. Scale 1 means equal importance, three means relatively more, five means heavily more, seven means definitely more and finally nine means excessively more importance. The other values two, four , six, eight mean sub values if it is hard to decide. The technique calculates eigenvectors until the values are stable. Lastly, high score alternative with the highest weight is taken as the best solution or alternative.
Preference Ranking Organization Method For Enrichment Evaluation	PROMETHEE	Ranking of the solution alternatives is simple to perform. Pairwise comparisons are done as well and maximum ranked solution is chosen as the best alternative. Logic is similar with the others.

Table 2.7 continues

The Technique for Order Preference by Similarity to Ideal Solutions	TOPSIS	In this technique, there are constantly increasing and decreasing utility for each attribute. Based on that, it easy to determine the positive ideal and negative ideal solutions. Therefore, the ranking of alternatives can be found by comparing the Euclidean distances of each other. A matrix is created with n number of alternatives and k criteria. It is normalized and multiplied with weight-criteria matrix. Then the ideal solutions (positive and negative) are calculated. After that relative closeness to the ideal solution is computed. To determine the best solution alternative is the solution which is farthest from the negative ideal solution and closest to the positive ideal solution.
Compromise Programming	CP	In this method, the goal is to determine an alternative which is closest to the ideal solution. It selects the best alternative among in a group of alternatives which are close to the ideal point.

Table 2.7 continues

Multi Attribute Utility Theory	MAUT	This technique allows users to give their preferences as a utility function. The goal is to determine the utility value by finding single utility functions. Then independent and preferred cases are verified. After that multi utility funtions are derived. Highest value alternative solution is chosen as the best one.

In this thesis, after clustering the customer data in the first level of the decision support system, MCDM is used to rank the clusters obtained in the first level and to determine the most important clusters by considering the cluster ranking problem as a MCDM problem. To achieve this goal, TOPSIS and AHP algorithms are combined. Analytical Hierarchy Process is utilized to determine the weights of criteria that demonstrate the importance of each criterion based on pairwise comparisons which are given by the decision maker as an input. And then, TOPSIS is performed to order the alternatives (different clusters in our problem) based on their ranks which are determined during the algorithm and select the most important alternative according to the criteria of which weights are determined by AHP. As stated previously, thanks to AHP decision makers can defines their opinions with the aid of 1-9 scale. The scale is very beneficial if the study needs experts' or an individual's opinions or experiences. That's why, hybrid method of AHP and TOPSIS offers several benefits. It is a systematic and reliable method due to the fact that it is capable of capturing the expert's opinions or considering previous experiences, therefore with an integrated method of AHP weights in TOPSIS algorithm, the process is more realistic and rational. In the next sections, AHP and TOPSIS methods are explained in detail.

2.3.1 Analytical Hierarchy Process

Analytical hierarchy process, which is developed by Saaty (1980), is able to solve complex MCDM problems. It is known to be a beneficial method for handling both qualitative and quantitative assessment criteria to solve MCDM problems. The main characteristic of the method is judgement based on pairwise comparisons. It is more commonly used compared to the other MCDM methods due to its simplicity and adaptability to many types of decision problems. Sipahi and Timor (2010) showed that, among the 232 articles related to MCDM published in reputed international academic journals during 2005–2009, AHP was used in 169 articles and other techniques was used in the rest. AHP is used by both individuals and groups to deal with not only individual decision making problems but also group decision making problems. In addition, a high emphasize on considering the relative priorities of criteria and alternatives in calculations increases the effectiveness of this technique. According to Fong & Choi (2000), there are three steps of AHP:

a. Constructing the hierarchic structure
b. Prioritization procedure
c. Calculation of the results

The MCDM problem is handled in a hierarchic structure by AHP in the first step. The main objective of the problem is on the top of the hierarchy, whereas the criteria and sub-criteria are in sub-levels and the solution alternatives are in the bottom of the hierarchy. Saaty (1994) mentioned that it is beneficial to work with a hierarchic structure for a decision maker since it is the key to have an overall view of the complex relationships between the criteria and also in the judgement process. Therefore, the decision maker is able to compare the issues of the same order of importance. In the second step of AHP, it defines the weights of each criterion. Each criterion is compared in pairwise and the weights of each criterion is determined by the decision maker with respect to the numbers from 1 to 9. Table 2.8 defines the scale of relative importance of attributes along with the related definitions.

Table 2.8 The pairwise importance scale which is introduced by by Saaty (1995)

Scale of Scores Given By User Based on Preferences	**Meaning**
1	Equally important
3	Relatively more
5	Essential or strong importance
7	Strongly more importance
9	Definitely more importance
2,4,6,8	Sub values for relative importance if it is hard to decide

Muralhidar et al. (1990) mentioned that it is an advantage to use the pairwise comparisons method because the decision maker is able to focus on the comparisons between two criteria and it makes the process free from irrelevant impacts. To prepare the results of the pairwise comparisons, it is better to place them in a matrix in order to be ready for the calculation and to make the calculations easier and faster. In the last step of the algorithm, the weights which were determined by the AHP process in the previous step and converted into the matrix form are used to calculate the eigen vector matrix. Then the eigen vector matrix is normalized and the decision can be made according to the sequence of the values in the normalized eigen vector matrix. The flow of the AHP is depicted in the following figure;

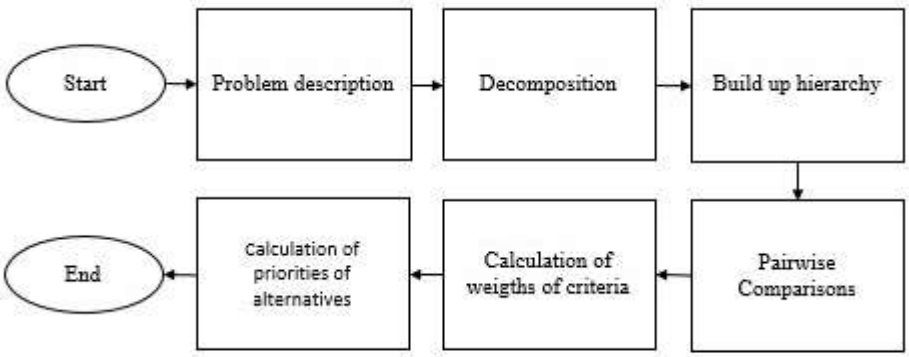


Figure 2.6 Process flow of AHP is as below

In contrast to the other MCDM methods, with the help of pairwise comparisons, it requires decision makers to think more detailed about the weights of the criteria and analyze the problem deeper. It is also an advantage for AHP that it is able to measure the indicators of quality and quantity while using mental preferences, expertise and objective information. Another advantage of it, it is able to incorporate some opinions of experts and the results of some questionnaires or researches, which can be input for pairwise comparisons. This means that it allows subjectivity to some extent. Due to those advantages, AHP has been more commonly used individually and in combination with other decision making techniques to solve multi criteria problems compared to the other methods.

2.3.2 Technique for Order Preference by Similarity to Ideal Solution (TOPSIS)

In the literature TOPSIS has firstly been mentioned by Hwang and Yoon (1981). Based on its procedure, the alternative which is finally chosen must have the farthest distance from the negative ideal solution and the shortest distance from the ideal solution. TOPSIS is very popular and is used in a wide area for solving decision making problems.

In the TOPSIS procedure, the distances of the solutions to PIS (positive-ideal solution) and NIS (negative-ideal solution) are continuously evaluated, and a preference order is ranked based on their relative closeness to PIS and NIS and combination of these two distances. The best alternative will be the one which is the closest to the PIS and farthest from the NIS. PIS should be the one which maximizes the benefit criteria and minimizes the cost criteria. NIS should be the one which maximizes the cost criteria and minimizes the benefit criteria. In other words, the PIS includes all best values and the NIS contains all worst values attainable of criteria.

The procedure of TOPSIS is shown in the following figure;

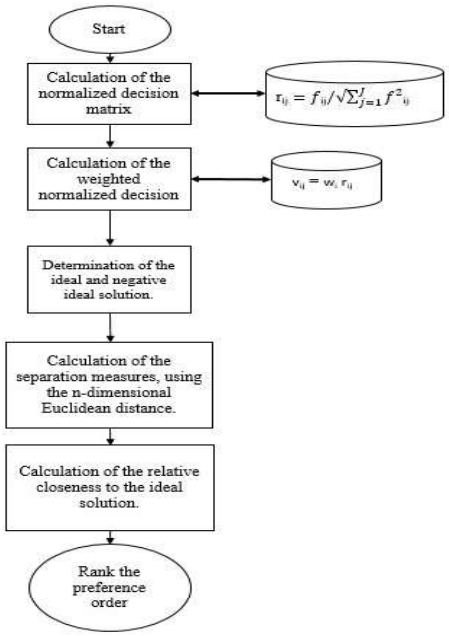


Figure 2.7 The procedure of TOPSIS

To clarify the method of TOPSIS, steps of the method is given as follows (Du et al, 2013):

1. A decision matrix D = (x_{mn}), which includes alternatives and criteria to be created.
2. Normalization:

$$r_{ij} = \frac{x_{ij}}{\sqrt{\sum_{j=1}^{m} x_{ij}^2}}, \quad i=1,\ldots,m \; ; \; j=1,\ldots n \tag{2.1}$$

3. Multiplication of the decision matrix with related weights and find the weighted decision matrix A = v_{mn} :

$$v_{ij} = w_j \text{ x } r_{ij} \text{ , } i=1,\ldots,m \; ; \; j=1,\ldots n \tag{2.2}$$

Note that weight of criterion j is demonstrated as w_j

4. Negative and positive ideal solutions are calculated as A^+ and A^- as follows:

$$A^+ = \{V_1^+ + V_1^+ \ldots\ldots V_1^+\} = \left\{ \left(\underbrace{\max}_{i} v_{ij} \mid j \in K_b\right) \left(\underbrace{\min}_{i} v_{ij} \mid j \in K_c\right) \right\} \tag{2.3}$$

$$A^- = \{V_1^- + V_1^- \ldots\ldots V_1^-\} = \left\{ \left(\underbrace{\min}_{i} v_{ij} \mid j \in K_b\right) \left(\underbrace{\max}_{i} v_{ij} \mid j \in K_c\right) \right\} \tag{2.4}$$

Note that the set of benefit criteria is Kb and the set of cost criteria is Kc.

5. Calculate Euclidean distances of S^+ and S^- from alternatives to the ideal solutions:

$$S_i^+ = \sqrt{\sum_{j=1}^{n}(v_j^+ - v_{ij})^2}\,,\ \text{i} = 1, \ldots, \text{m};\ \text{j} = 1, \ldots, \text{n}. \tag{2.5}$$

$$S_i^- = \sqrt{\sum_{j=1}^{n}(v_j^+ - v_{ij})^2}\,,\ \text{i} = 1, \ldots, \text{m};\ \text{j} = 1, \ldots, \text{n}. \tag{2.6}$$

6. Calculate relative closeness to the ideal solution:

$$C_i = \frac{S_i^-}{S_i^- + S_i^+}\,,\ \text{i} = 1,\ldots,\text{m}. \tag{2.7}$$

7. The best alternative will be determined based on relative closeness to the ideal solution. It means higher value of C_i, higher rank of the alternative i.

TOPSIS is widely used to solve MCDM problems in fields such as, human resources management, transportation, product design, manufacturing, water management, quality control, supplier selection and sustainable and renewable energy. According to Shih et al. (2007), TOPSIS has four main advantages: It is based on sound logic which demonstrates the behavior of choice of a human, an exact value that is calculated for the alternatives with highest value and lowest value in the same time, easy logic to program so it is programmable and measures of performances of each alternative can be visualized simply. Due to those advantages, TOPSIS has been more widely used individually and in combination with other decision making techniques to solve multi criteria problems compared to the other methods.

CHAPTER THREE

LITERATURE REVIEW ON ANALYTIC CUSTOMER SEGMENTATION METHODS: LITERATURE REVIEW

Studies on development and application of analytic CRM techniques based on analysis of big data have significantly increased in the last decades due the improvements in the information systems and technology. CRM has extracted from developments in information technology (Chen & Popovic, 2003). Dimitriadis & Stevens (2008) mentioned that CRM is still a high interest topic the same as in the past decade. Researchers in the academy and also practitioners in a wide range of industries have a high attention on the CRM experiences and applications as well as new studies on different methods.

There are many definitions of CRM in the literature as it is mentioned in Chapter 1 of this thesis. The variety of definitions show that many kinds of perspectives regarding CRM exist. However, the common point of these perspectives is that CRM helps the companies to improve the relationships with customers, provides long run profitability, customer loyalty and enables both the customers and the company a better understanding of each other to establish strong relationships.

Data mining and data segmentation are important sub-fields of the CRM, and they are related to each other as well. Data mining is an essential technique to predict the customer behavior and profile which helps companies to detect and to analyze the business events (Verhoef et al., 2001). Market segmentation concept is firstly mentioned by Smith (1956), it is stated as segmentation of the market is required because of the need to view a heterogeneous market as smaller heterogeneous markets to simply understand and meet differentiated needs of the customers with customized products and services.

Market segmentation involves viewing a heterogeneous market as a number of smaller homogeneous markets, in response to differing preferences, attributable to the desires of consumers for more precise satisfaction on their varying wants''. The following sections present a review on customer segmentation studies consentrating on the approaches that are also handled in this thesis; statistical approaches, heuristic approaches, MCDM methods and hybrid approaches.

3.1 Literature Review on Statistical Approaches For Customer Segmentation

With the help of segmentation, companies can identify competitive marketing strategies and design their products and services to meet the needs of different profile customers (Bull, 2003). An accurate segmentation of customers into groups, as a part of the most important activities in CRM, is necessary to approach each customer individually and to determine the high value segments and customers (Kolarovszki, Tengler and Majercakova, 2016). To segment the customers into customer groups, a variety of methods are developed and applied in the studies in literature and in practical applications in business. Du (2010) reviewed all the methods of neural network clustering in his paper and provided illustrative examples related to them. For the researchers who deal with customer segmentation, it is a very beneficial paper for a detailed understanding of the methods.

The most widely customer segmentation techniques are Hierarchical Clustering, K-Means Clustering and self-organizing maps. Hierarchical Clustering begins by taking into account each individual as a separate segment (Mondal, 2018). Then, it consecutively performs the following:

1. Selection of two clusters which are the closest to each other
2. Combine two closest clusters as one cluster.
3. Iteration

It continues till all of the clusters are combined and the final solution of the algorithm is as in the Figure 3.1 (Cluster 1: 1, 2, 3; Cluster 2: 4, 5, 6) as an example. Hierarchical Clustering can be applied to a variety of data.

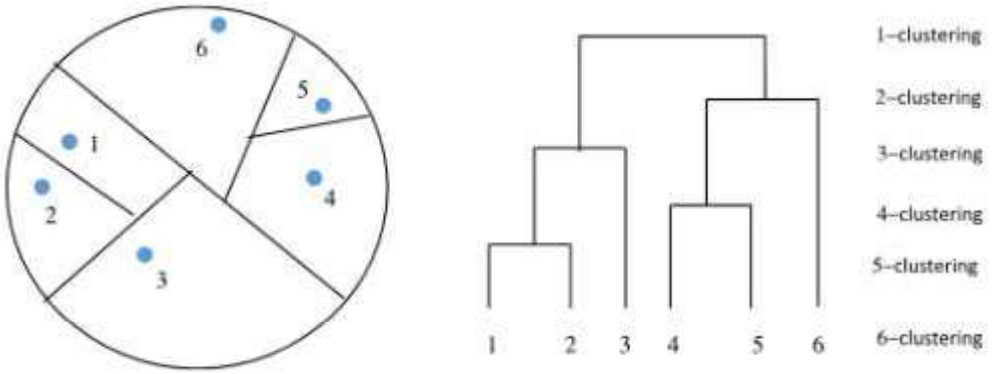


Figure 3.1 Output of Hierarchical Clustering (Mondal, 2018)

Self-organizing map is a type of artificial neural network which is used for clustering a "messy data set". In Logeswari and Karnan (2010)'s paper, Self-organizing Map (SOM) is proposed for medical image segmentation. In their study, components which are called neurons form the self-organizing map. They use the weights of the neurons as input data and regular locations of neurons is a usual arrangement in rectangular and hexagonal grid. With the aid of implementation of Self-organizing map, input data (individual weights of the same dimension) is clustered and brain tumor is detected.

Segmentation by testing hypothesis with respect to the questionnaires is one of the most widely used methods in the literature. For instance; in their paper related to the segmentation of airport customers in Slovakia, Csikosova et al. (2015) applied testing hypothesis method. This method works based on the statement of that closely associated documents tend to be related to the same requests. In their paper which aims to apply segmentation to the mobile phone customers, Hamka et al. (2014) took information from the customers about their mobile phone usage purposes through a mobile application with 129 users. Then, they applied Latent Class Analysis (LCA), which utilizes expectation maximization and Newton-Raphson method through an

iterative process to calculate maximum-likelihood estimates. The results of the analysis specified 6 clusters of mobile phone users. In mobile phone industry, another study Hong, Duan, Li and Wang (2016) apply a novel bi-clustering based market segmentation method by using customer pain points which is obtained from a web forum about a mobile phone brand and they divided the customers into 8 clusters, each of which involves customers who are suffered from similar customer pain points such as internet, game, sound etc.

Velmurugan & Santhanam (2010) studied on the differences on K means and K medoids algorithms and in their numerical examples it is observed that for the smaller sets of data K means algorithm is efficient but for large sets of data K medoids algorithm performs relatively better. Different than widely used K Means algorithm, Brito et al. (2015) used K-Medoids clustering algorithm for segmentation of data regarding to 10,775 customers of an online customized fashion business and obtained 6 clusters. They concluded that the main difference of the K Medoids algorithm from the K means algorithm is that in the second step of K medoids algorithm the customer closest to the existing cluster (median) is selected as the new cluster while K means algorithm selects the mean point. The numerical example in this paper is also beneficial to see that K medoids algorithm gives better results than K means algorithm with large data set. Konu, Loukkanen and Komppula (2011) apply K means algorithm to cluster the ski resort customers in Finland by using the ski destination choice criteria and obtain 6 segments. The remarkable point in the paper is that they used F-Statistics to determine the effects of each criterion to the segmentation. Hong, Duan, Li and Wang (2013) study on a data set related to a large-scale logistics network. To apply the customer segmentation they use Axiomatic Fuzzy-based customer clustering algorithm with a hierarchical analysis structure. They conclude the most significant contribution of this paper is that they determine initial clusters not randomly but analytically based on a threshold “alfa” value which refers to the value of each different location matrix. They also state that the other widely used methods such as K means algorithm, K medoids algorithm or hierarchical clustering algorithm determine the initial clusters randomly by a process dependent to the user, which influences the result significantly.

According to Chen & Shixiong (2009), K Means algorithm has many limitations due to the randomness. Initial randomly selected set of the clusters affect the quality of the final solution therefore there is a possibility that it could be much worse than the global optimum. However, the algorithm is extremely fast and the other alternative which is to run the algorithm several times to overcome the randomness might be considered and find the best clusters by comparing the previous solutions with the last one. Randomness of selection of initial clusters is also mentioned in Yao, Duan, Li and Wang (2013)'s paper as limitation of the algorithm. They applied K Means Algorithm to image segmentation theory which is on the basis of the aggregation of image pixels in the feature space. In their study, Lopez et al. (2010) proposes a clustering algorithm which is called Hopfield-Clustering Algorithm. They mentioned about benefits and limitations of K-Means algorithm and SOM. Based on their research, the most significant advantage of K-Means is the speed of the algorithm, however randomness of the solution which is obtained with K-Means or SOM is a big limitation. Based on Bradley & Fayyad (1998)'s study, selection of initial clusters in K Means is very sensitive on the finalized clusters which are found in the end of algorithm. They used a method to refine the initial solutions however it still includes randomness.

From the literature review, it can be obviously seen that, randomness in the initial part of the K-Means algorithm is a big limitation, that's why in this thesis, Simulated Annealing is used to overcome the randomness of K-Means algorithm.

3.2 Literature Review on Heuristic Approaches for Customer Segmentation

Heuristic approaches can be used for customer segmentation, especially in cases that large data sets have to be captured in complex problems. With heuristic approaches, run time for clustering is usually shorter than mathematical or statistical models. However, they are used in a few studies in customer segmentation field. Among them, Vincentelli, Chen and Chua (1977) applied a greedy heuristic to cluster a large-scale network (tearing the network). Greedy heuristic is an algorithm that

determines the direction for iteration by checking for the cheapest local condition. Based on their computational experiments, greedy algorithm for clustering works with high efficiency and finds near optimal solutions. Hatamlou (2013) comes with an interesting new heuristic optimization algorithm which is called as “Black Hole (BA)” heuristic. Black hole heuristic is inspired by the attitudes of stars through a black hole phenomenon in the space. Hatamlou (2013) applied the heuristic to solve clustering problem on six benchmark datasets with a variety of complexity. According to the study, the new heuristic performs well based on the results of clustering on six different data sets and it has advantages such as simplicity and easy to implement. Zhang, Ouyang and Ning (2010) applied another heuristic, namely Artificial Bee Colony (ABC) Algorithm, to perform clustering of 3 different data sets and tested the heuristic. They also applied other heuristics such as Genetic Algorithm and Taboo Search. Based on the computational results to evaluate the performance of ABC algorithm, it was clearly seen that the experience is very encouraging in terms of the quality of the solution, the average number of function evaluation and the processing time.

3.3 Literature Review on MCDM Methods and Their Usage in Customer Segmentation

In real life business applications and in the scientific studies related to a variety of topics, people generally face with such cases that they must apply decision-making procedures in order to evaluate all the alternatives and to select the best or the most important one. Customer segmentation is one of these topics, in which generally lots of complicated and complex criteria exist with respect to different alternatives. To make a rational decision and to find the best or preferred solution, MCDM methods are applied.

The most widely used MCDM methods are TOPSIS (Hwang & Toon, 1981), AHP (Saaty,1980), Analytical Network Process (ANP) (Saaty, 1996), Elimination et Choice Translating Reality (ELECTRE) (Roy, 1990), Vlse Kriterijumska Optimizacija I Kompromisno Resenje technique (VIKOR) (Opricovic & Tzeng, 2004), and Decision Making Trial and Evaluation Laboratory (DEMATEL) (Fontela &

Gabus, 1976). In the literature, researchers study on these methods frequently, in a wide range of fields because these methods could be applied numerous decision making problems, improved results can be obtained due to the usage of an analytic viewpoint when making a decision rather than to decide irrationally or randomly or intuitional. Baykasoglu et al. (2016) stated that people usually face with problems related to selection among two or more alternatives in their social, personal and professional lives, however, selection of the best solution alternative among the other alternatives might need to consider multiple, incommensurate and contradictory criteria which exceeds cognitive limitations of decision makers and becomes a complex problem as well, therefore MCDM deals with such complicated problems with multiple and conflicting criteria.

Kusumawardani & Agintiara (2015) study a problem on a human resources manager selection among 35 candidates with respect to 9 criteria by using a hybrid approach. They used Fuzzy AHP to determine the weights defining the relative importance of criteria and Fuzzy TOPSIS to calculate the ideal candidate based on the scores that have been assigned according to the output of AHP weighted criteria. Isık et al. (2015) applied Fuzzy AHP to the problem on selection of Learning Management System among 9 alternatives with 7 criteria by using 5 Triangular Fuzzy Numbers, therefore they obtained the best system alternative. Yong (2006) aimed to solve plant location selection problem by applying Fuzzy TOPSIS and AHP and obtained positive ideal and negative ideal solutions. Lee et al. (2014) presented an efficient decision making procedure utilizing fuzzy AHP to select the best open source CRM software among 3 alternatives considering 4 criteria. They utilized the comments from related department managers.

It is obvious that MCDM methods are used frequently to solve decision making problems in almost all fields in business and science and even in personal issues. In customer segmentation, these methods can be applied to rank the clusters and to enable companies to determine the customer groups with high value. Therefore, convenient marketing and business strategies can be applied to every single customer, moreover with the possibility to differentiate the products in order to meet the specific needs of

each customer in high value groups, the companies can have significant competitive advantage over their rivals in the related industry. However, according to the review of the related literature, it seems that very few studies demonstrate the usage of MCDM methods to rank the customer segments and to determine the high value customer groups. Among them, Chiang (2014) developed a FMLC (Frequency, Monetary, Low Price / Discount Products, Cancellation times) model and applied it to the customer segmentation case, afterwards, by applying AHP, he ranked customer segments and decided the high value customers. Mostafa, Batool and Parvaneh (2013) studied on segmentation and apply TOPSIS to rank the factors of the segments. Based on the results, they give a perception to the company related to the most important factor of the customers. In order to demonstrate segmentation of their data, Chiu, Tzeng and Li (2012) utilized a hybrid method which is combination of Decision Making Trial and Evaluation Laboratory (DEMATEL), DEMATEL-based Analytic Network Process (DANP), and VIšekriterijumsko KOmpromisno Rangiranje (VIKOR). Weights of criteria is determined by DEMATEL and the ranking of the segments is determined by VIKOR.

3.4 Hybrid Analytic Approaches for Customer Segmentation

In their paper, Lopez, Aguado and Martin (2011) studied on a data set of 230 electricity customers by using a hybrid algorithm which is called as Hopfield-K-Means algorithm in order to overcome the randomness of the initial solution provided by K-Means algorithm. According to the results of the numerical example, 13 clusters are obtained. They also compared their novel hybrid method with the other widely used algorithms by using the indexes in the process of segmentation of big data, and it is claimed that Hopfield-K-Means algorithm provides better results than Hierarchical Algorithms, Follow the Leader Algorithm and Hopfield's Recurrent Neural Network. Li et al. (2011) provided a two stage algorithm for segmentation of a data set with 477 observed values in a company related to textile industry. The first stage is Ward's algorithm (Ward, 1963) to estimate the number of clusters and the second stage is to perform the clustering operation with K means algorithm. They obtained 5 clusters by

using customer life time value as criteria. This study is one of the most significant contributions to the literature in terms of overcoming the drawback of random selection of initial clusters by K-means algorithm, however, it seems that a shortage still exists because they are selecting the first clusters randomly.

There are other hybrid methods developed for segmentation in the related literature. Among them, Chipman and Tibshirani (2005) applied a hybrid method of tree-structured vector quantization and mean linkage agglomerative. With the aid of the hybrid method, the result of the computational experiments shows that both algorithms work together in hybrid method better than individually. Bhanu, Lee and Das (1994) 's paper proposes a hybrid method as a combination of genetic algorithm and hill climbing to cluster their data. In the computational experiments, they make a comparison between their hybrid method and pure genetic algorithm and the result shows that hybrid method is more robust and quicker. Bose & Chen (2009) clustered their data with another hybrid method of SOM and C5.0 decision tree. C5.0 decision tree is chosen as the second step of the hybrid method due to the fact that it needs less effort in data preprocessing and generated sufficient results.

3.5 Summary of the Literature Review

In this literature review, it can be concluded that through the advances in information technology, the significance of analytic CRM has recently been increased and accordingly the researches on this area increased dramatically. Major sub fields of analytical CRM are data mining and customer segmentation. In the related literature, there are numerous methods to perform data mining and also clustering the customers. K-Means method is one of the most commonly used ones among them.

Based on our literature review, the second most commonly used method in the literature is Hierarchical Clustering method. Zhao and Krypis (2009) stated that

solutions derived from Hierarchical Clustering which are in the form of trees called dendrograms, are overy popular and finds place in many fields for its implementation.

Customer segmentation methods are currently applied mostly in e-commerce industry. The reason is that it is easier to collect data from e-commerce users thanks to recent advances in technology. In the last decades, tourism industry was far more popular than other sectors interm of customer segmentation applications, still it is in the second rank.

Moreover, developing and applying hybrid methods have been of particular interest to cluster the customer data is recent years. The first algorithm of the hybrid method is commonly chosen from conventional data mining methods such as K-Means and Hierarchical trees, whereas the second method is commonly chosen among MCDM methods, which they are also very popular in this field. A variety of MCDM methods such as DEMATEL and VIKOR are used with data mining techniques to create more robust, reliable and fast hybrid methods.

3.6 Gaps in the Current Literature and Possible Ways to Fulfill These Gaps

Compared to the other fields such as transportation, supplier selection, human resources management, manufacturing, water management, product design, quality control, and sustainable and renewable energy, in customer segmentation field, MCDM methods are rarely used both in scientific studies by researchers and in real life segmentation applications by practitioners. Moreover, among the articles related to segmentation and clustering in the literature, it has been observed that there is no article which uses an integration of AHP and TOPSIS methods to determine the ranking of the clusters. In addition, based on the findings of this literature review, there is no study related to a hybrid method integrating K-Means Clustering Algorithm and a metaheuristic algorithm. As it is already mentioned in the previous sections, a metaheuristic algorithm, namely Simulated Annealing Algorithm, is used to overcome

the randomness of K-Means in finding the initial clusters to obtain a more systematic and robust clustering approach.

In order to fulfill this gap in the literature, a new decision support system has been developed to solve customer segmentation problems by using a hybrid two level approach which involves a combination of a heuristic algorithm, a statistical data mining method and MCDM methods. In the first level of the decision support system, a statistical data mining method, namely K-Means Clustering Algorithm and a heuristic, namely Simulated Annealing Algorithm is combined to segment the customer data. Then, in the second level, AHP and TOPSIS methods are integrated to rank the customer segments determined in the first level and specify the most important (i.e. the highest value) segments. Also to the best of our knowledge, there is no decision support system with a user friendly interface for customer segmentation and ranking. Our decision support system, which is based on a VB.Net program, also supported by a user friendly interface to facilitate it usage in any real life problems.

In the next section, the proposed decision support system for customer segmentation is explained.

CHAPTER FOUR

THE PROPOSED DECISION SUPPORT SYSTEM

4.1 Explanation of the Decision Support System (DSS)

There are numerous articles which use K-Means algorithm in the clustering step of their problems. In addition, K-Means algorithm has a wide application area in solving practical clustering problems due to its simplicity and easiness to integrate it to the other algorithms. On the other hand, by the data points of the first draft clusters are chosen by the user randomly K-Means algorithm, which may cause instability in the algorithm. In order to overcome this problem, Simulated Annealing Algorithm is integrated to K-Means algorithm and the data points of the draft clusters are determined by Simulated Annealing Algorithm in our study. Then K-Means algorithm starts by considering those draft clusters. After the clusters are determined, the problem is to decide the most valuable cluster among all clusters. To solve those kinds of MCDM problems, there are several MCDM methods. Among them AHP and TOPSIS are widely used and well known methods. The steps of both methods are simple, easy to integrate to each other and can be programmed easily. In this study, AHP is used to calculate the weights of criterion and TOPSIS is used to determine the ranking of the clusters which are determined by the hybrid method that combines K Means algorithm and Simulated Annealing Algorithm.

In the process flow of K-Means Clustering algorithm, it is obviously seen that (see Figure 4.1) the step of selection and appointment of some customers as data points of initial clusters is a random process and it affects the result and the run time of the algorithm depending on the size of the data.

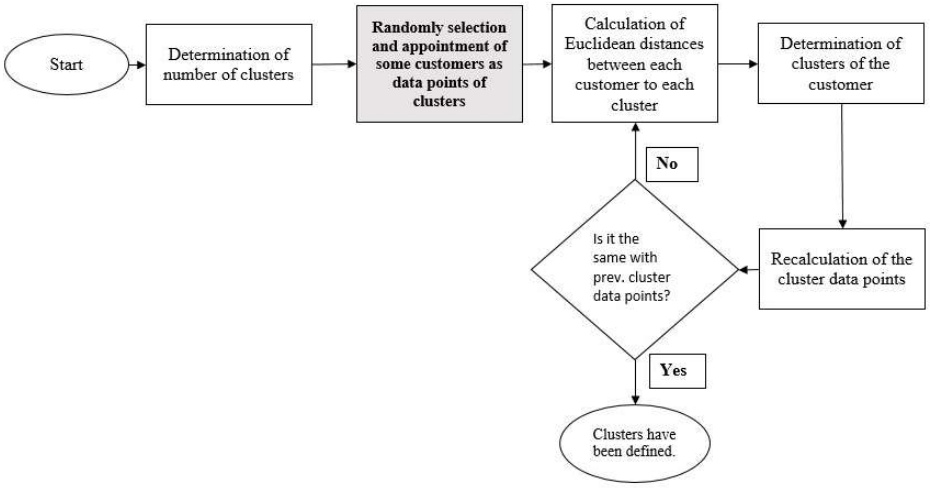


Figure 4.1 Randomness in K Means Algorithm

In order to overcome this randomness, Simulated Annealing Algorithm has been integrated to K-Means Algorithm in our study. Hence, randomness in the step of selection and appointment of some customers as data points of initial clusters has been reduced by using a systematic procedure in this step and run time of the K-Means algorithm has been reduced accordingly.

Simulated Annealing Algorithm is chosen to make this initial step of K Means algorithm more systematic since it is one of the most widely used methods to approximate the result to the optimum solution and to find the result that is closest to the optimum solution in feasible solution space. It is also easy to integrate it to the other methods. In this study, the first clusters which have to be determined randomly by the user in conventional K-Means Algorithm have been determined not manually but by Simulated Annealing Algorithm. After the integration of Simulated Annealing Algorithm to the K-Means algorithm, the process flow of the new hybrid method can be depicted in the following figures (Fig.4.2, Fig.4.3 and Fig.4.4) respectively.

Figure 4.2 demonstrates the flow of Simulated Annealing Algorithm. In this flow, cluster 1 and cluster 2 have been determined as draft clusters based on this algorithm. Fig.4.3 shows the second flow of Simulated Annealing Algorithm, which gives third draft cluster based on the previous Simulated Annealing flow. Through these

procedures found the three draft clusters are obtained to be used in the initial step of the K-Means algorithm. Finally, Fig.4.4 explains the flow of K-Means algorithm. It gives a general overview of K-Means to explain how it works.

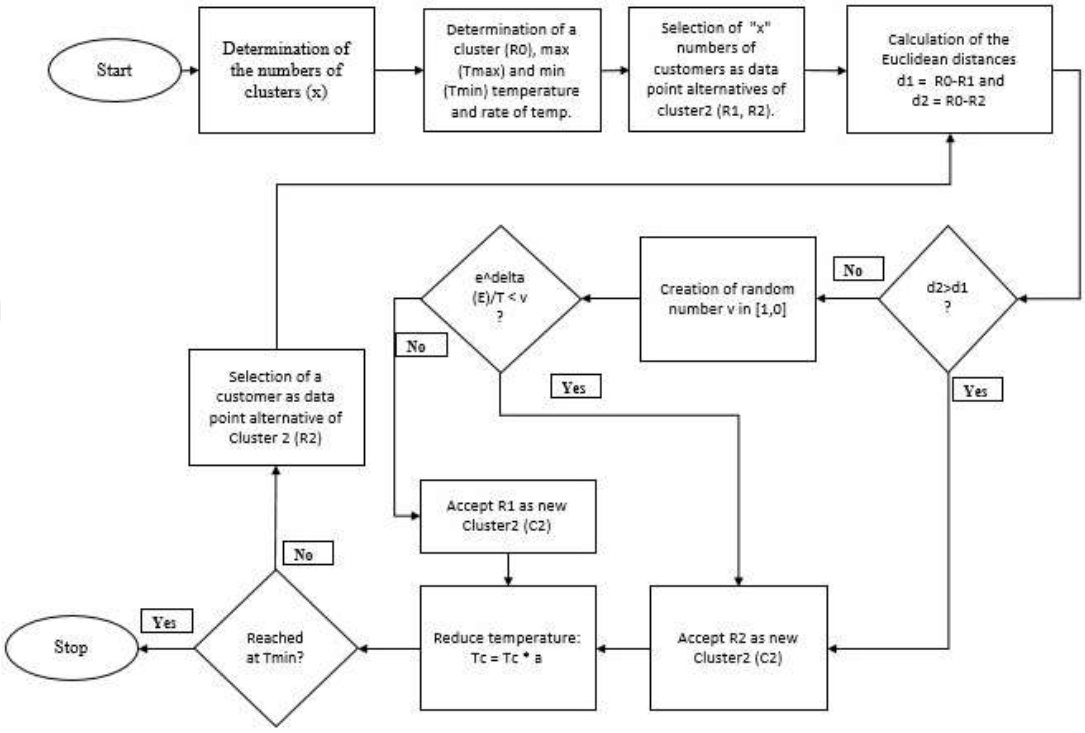


Figure 4.2 Simulated Annealing Algorithm – Determination of draft Cluster 1 and Cluster 2

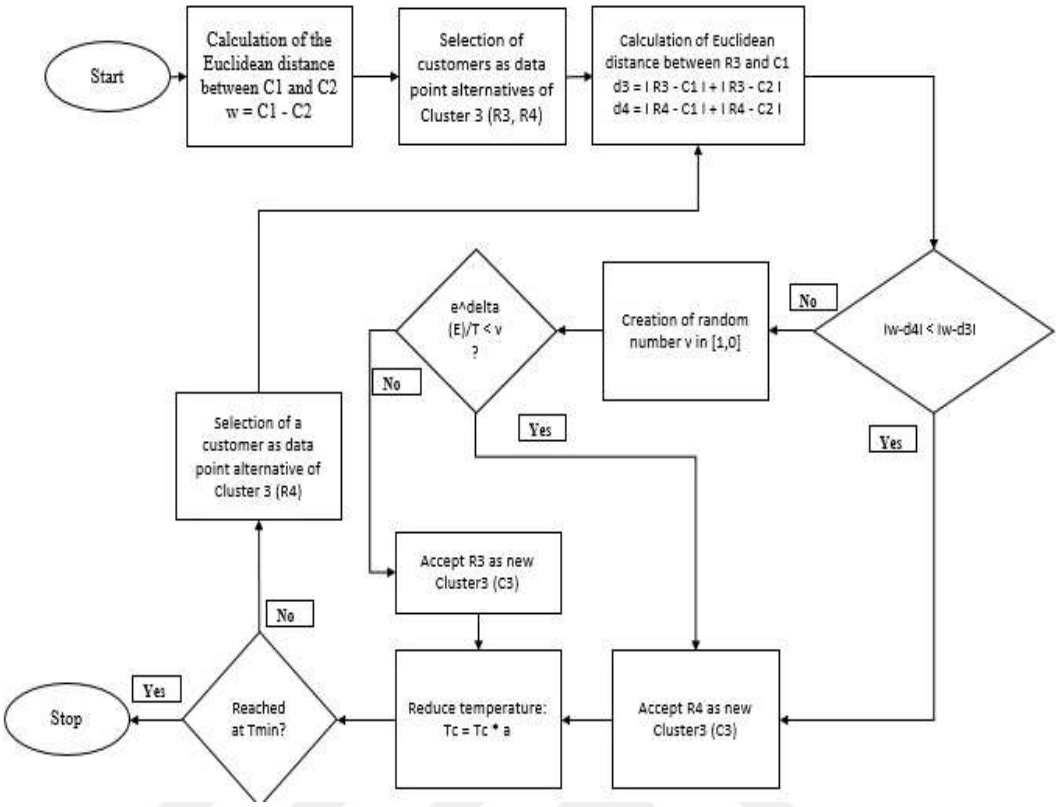


Figure 4.3 Simulated Annealing Algorithm – Determination of draft Cluster 3

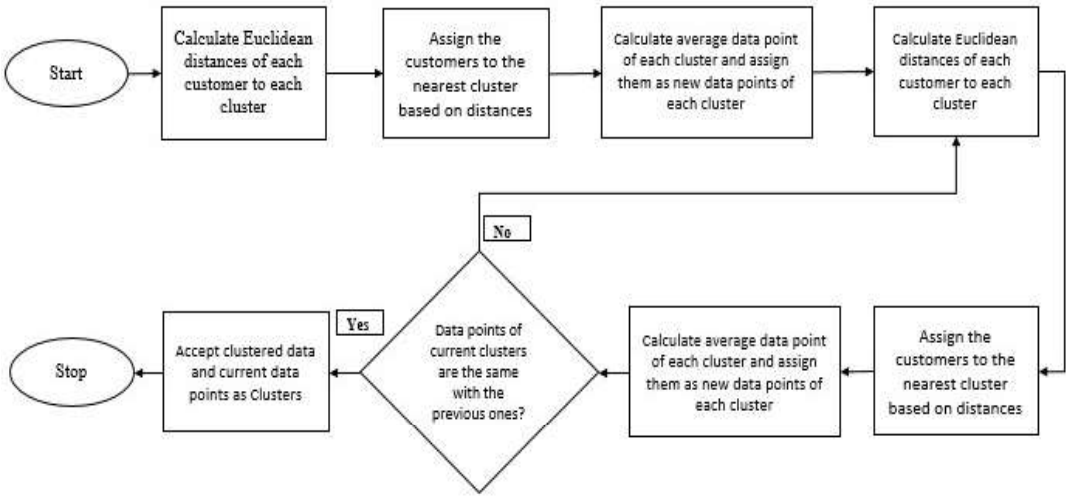


Figure 4.4 K-Means Algorithm – Determination of finalized Cluster 1, Cluster2 and Cluster3

Until here, the proper data points of clusters, which are approximated to the optimum data points by a hybrid method, are determined. In the next step, the decision of the most valuable customer segment will be made.

The customer evaluation criteria are the variables which should be weighted based on their relative importance levels. In this step of the proposed methodology, AHP has been applied to determine the weights. As in conventional AHP, as mentioned in the previous sections, the values from 1-9 (Saaty's scale) are determined through pairwise comparisons. Then, the matrix of pairwise comparison is built. By the matrix computations, the eigenvector matrix is built and finally the weights of the criterion are defined.

The steps of AHP are shown in the following process flow diagram:

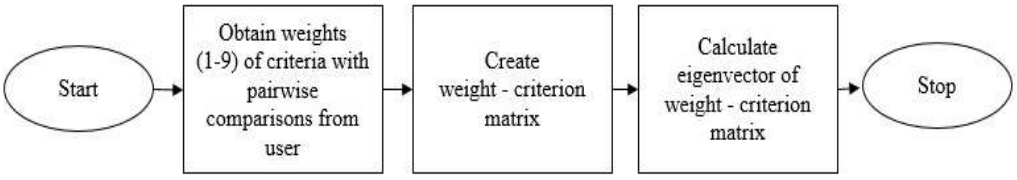


Figure 4.5 AHP Process Flow in the example

Utilizing the determined weights for each criterion, in the next step, the ranks of the clusters are determined by TOPSIS. Each cluster has its criteria and also each criterion has data points in each cluster. Matrix of criteria and clusters are built utilizing the data points of each cluster. Then the matrix is normalized using the formula given as equation 2.1. in section 2.3.2. After that, positive and negative ideal solutions are calculated based on their distances as given in equations in 2.3, 2.4, 2.5 and 2.6 in section 2.3.2. Finally, with the computation of relative closeness to the ideal solution formula 2.7, ranking values are determined and ranking is completed.

The steps of TOPSIS procedure is given in the following process flow diagram:

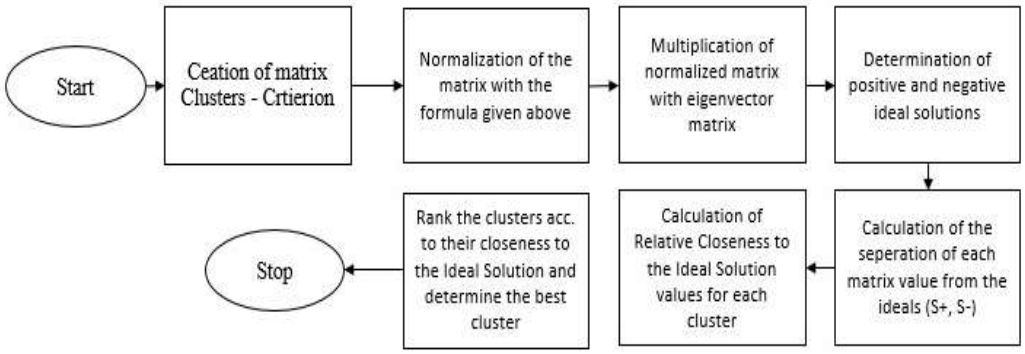


Figure 4.6 TOPSIS Process Flow followed in this example

All of those flows have been programmed in VB.net and the results of the algorithms can be obtained very quickly by the software developed in this study (see appendix for the part of codes).

The main flow which includes all those detailed flows is shown as below:

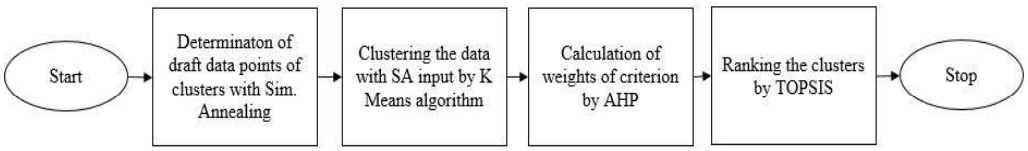


Figure 4.7 Main flow of Decision Support System

4.2 Computational Experiments

4.2.1 Problem Statement

For computational experiments, the study proposed by Srihadi, Hartoyo, Sukandar and Soehadi's (2016) is utilized. To this aim, the data set which is used and presented in this paper is considered to illustrate the applicability of our methodology. The data set is related to the segmentation of the visitors who travel to Jakarta, India. It includes 8 criteria and 393 individual customer's data.

Fig.4.8 depicts the percentages of tourists according to each criterion. There are 8 criteria which are Age, Race Group, Employment Status, Purpose of Trip, Gender, Previous Trips to Jakarta, Travel Companion and Travel Arrangement. Each criterion has their own factors inside of it such as Race Group which includes factors as Asian, Caucasian and Other or Gender whose factors are Male and Female. According to Fig.4.8, there are totally 393 tourists (customers), and 65.9% of them are male, then exact the number of male tourists can be calculated easily. It is the same for each criterion.

	Frequency (%)		Frequency (%)
Age		*Gender*	
18 - 24	107 (27.3)	Male	259 (65.9)
25 - 34	122 (31.0)	Female	134 (34.1)
35 - 44	95 (24.2)		
45 - 54	43 (10.9)	*Previous Trip to Jakarta*	
55 - 64	21 (5.3)	Never	180 (45.8)
65 and above	5 (1.3)	1 time	86 (21.9)
		2 times	40 (10.2)
Race group		3 times	37 (9.4)
Asian	228 (58.0)	4 times and more	50 (12.7)
Caucasian	138 (35.1)		
Other	27 (6.9)	*Travel Companion*	
		Alone	80 (20.4)
Employment Status		With family	44 (11.2)
Employed	204 (51.9)	With friends	193 (49.1)
Self-employed	77 (19.6)	With a partner	49 (12.5)
Housewife	6 (1.5)	With a traveling group	27 (6.8)
Retired/Unemployed	10 (2.5)		
Student/Scholar	96 (24.4)	*Travel Arrangement*	
		All-inclusive package tour	47 (12.0)
Purpose of the trip		Flight and hotel pre-arranged package	48 (12.2)
Vacation/sightseeing	219 (55.7)	Self arrangement for flight and accommodation	298 (75.8)
Business	52 (13.2)		
Convention/exhibition	22 (5.6)		
Vacation and Business	37 (9.4)		
Visiting friends/relatives	21 (5.3)		
Other	42 (10.8)		

Figure 4.8 The data – Percentages of distribution of customers

Age, Gender, Race Group, Employment Status, Purpose of Trip, Previous Trip to Jakarta, Travel Companion and Travel Arrangement are considered as customer evaluation criteria. The associated code of each criterion are determined according to their specific benefits. More clearly, if the benefit of the criterion is high then the code given to that criterion is represented by a high number. For example, a customer for "all-inclusive" hotels brings more money and makes reservation from winter. So if such a customer is the most valuable one for the company, the associated code should

be represented by the highest number in the matrix, which is 3. Values of the other factors which are “flight & hotel pre-arrangent” and “self-arrangement” in the criterion “Travel Arrangement” will be respectively 2 and 1 due to their lower benefits compared to the “all-inclusive” criterion. The codes for each criterion are determined based on this logic. The codes for each criterion can be seen in the following table;

Table 4.1 Criteria and regarding customer codes

Criteria	Travel Arrangement		
	All-inclusive	Flight&hotel pre arrangent	Self arrangement
Cust. Codes	3	2	1

Criteria	Travel Companion				
	Alone	w/ family	w/ friends	w/ a partner	w/ traveling grp
Cust. Codes	1	5	4	3	2

Criteria	Previous Trip to Jakarta				
	Never	1 Time	2 Times	3 Times	4 Times and more
Cust. Codes	5	4	3	2	1

Criteria	Purpose of trip					
	Vacation/Sightseeing	Business	Convention/Exhibiton	Vacation/Business	Visiting Friends/Relatives	Other
Cust. Codes	6	4	3	5	2	1

Criteria	Race Group		
	Asian	Caucasian	Other
Cust. Codes	2	1	3

Criteria	Gender	
	Male	Female
Cust. Codes	2	1

Criteria	Age					
	18-24	25-34	35-44	45-54	55-64	65+
Cust. Codes	4	6	5	3	2	1

Criteria	Employement Status				
	Employed	Self-Employed	Housewife	Retired/Unemployed	Student/Scholar
Cust. Codes	5	3	2	1	4

According to those values, each customer is coded to create a representation of them to be used in the algorithms. Each code of the customer has a meaning based on the specific features of the individual customer data. For example; for the first customer which is shown in Table 4.2, the code of the customer is “42246413”.

It means that Customer#1 has the features as follows:

- Age (4): 35-44
- Gender (2): Male
- Race Group (2): Asian
- Employment Status (4): Student / Scholar
- Purpose of Trip (6): Vacation / Sightseeing
- Prev. trip to Jakarta (4): 1 time
- Travel Companion (1): Alone
- Travel Arrangement (3): All-inclusive

The other customer codes, which are created based on this logic as well, are given in the following Table 4.2:

Table 4.2 The data – Part of data, customer codes

Cust #	Customer Codes	Age	Gender	Race Group	Employement Status	Purpose of trip	Previous Trip to Jakarta	Travel Companion	Travel Arrangement
1	42246413	4	2	2	4	6	4	1	3
2	42236453	4	2	2	3	6	4	5	3
3	42254441	4	2	2	5	4	4	4	1
4	42254443	4	2	2	5	4	4	4	3
5	42254433	4	2	2	5	4	4	3	3
6	42254413	4	2	2	5	4	4	1	3
7	42255441	4	2	2	5	5	4	4	1
8	42253433	4	2	2	5	3	4	3	3
9	42253343	4	2	2	5	3	3	4	3
10	42253241	4	2	2	5	3	2	4	1
11	42253142	4	2	2	5	3	1	4	2
12	42253113	4	2	2	5	3	1	1	3

4.2.2 Results

4.2.2.1 Clustering Results

Firstly, the code for each customer should be entered into the system to using the button “Add New Customer” in the following interface:

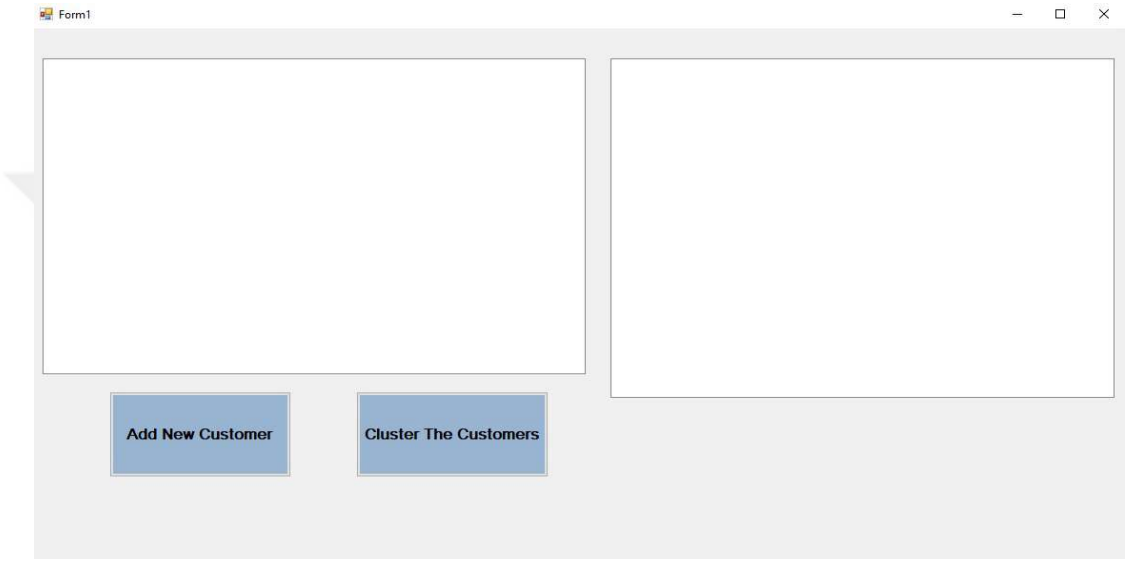


Figure 4.9 Software view – First step: Add New Customer

For example, after adding the codes related to 26 customers, the codes are seen in the program as given in the following figure;

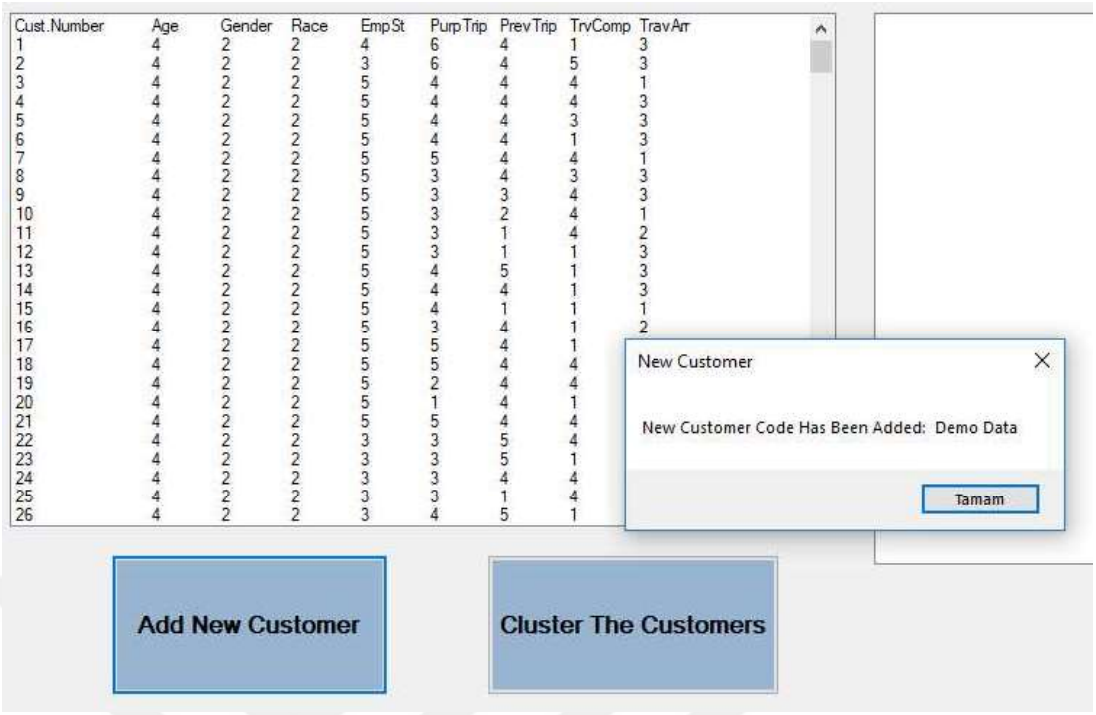


Figure 4.10 Software view – Second Step New Customers Added

After all customers are identified, using the button “Cluster the Customers”, the clustering step is initialized. In the clustering step the process flows which are explained in Section 4.1 are followed. As the first step, Simulated Annealing Algorithm finds the data points of draft clusters. Then, K Means algorithm uses those data points to finalize the clustering and find the latest data points of the clusters. In this computational experiment, the number of clusters which should be found by applying the proposed methodology is determined as 3.

4.2.2.1.1 Simulated Annealing Part. To show the flow of the algorithms in the program through numerical example and to clearly explain the logic of the decision support system, firstly $T_{min,}$ T_{max} and rate of temperature decrease are defined. Then, until the temperature reaches the minimum temperature, the program computes Euclidean distances of each pair of customers and compares them with each other to find the data points which are farthest from each other. There are millions of combinations, that’s why it is nearly impossible to find the optimum solution but it is possible to find the solution which is the best approximation to the optimal.

The Euclidean distances are computed as follows:

Let's say the distance between the customer 1 and customer 2 will be computed, where;

Code for customer 1: 4 2 2 4 6 4 1 3

Code for customer 2: 4 2 2 3 6 4 5 3

Then, the distance between the customer 1 and customer 2 will be;

$d_{1;2} = ((4^2+2^2+2^2+4^2+6^2+4^2+1^2+3^2) + (4^2+2^2+2^2+3^2+6^2+4^2+5^2+3^2))^{1/2}$

$d_{1;2} = (93 + 110)^{1/2}$

$d_{1;2} = 101,5$

After all distances are computed, they will be compared to each other and the ones which are far from each other will be chosen as the draft data points of the clusters. During each computation and comparison, current temperature in the algorithm will be decreased by the rate of temperature decrease which is determined previously. If the temperature reaches to its minimum value which is also determined previously, the algorithm stops and accepts 3 current clusters as the clusters which will be considered in the K Means algorithm.

4.2.2.1.2 K Means Algorithm Part. The number of clusters which should be found at the end of the algorithm was determined as 3. Therefore, the algorithm starts by using the data points of 3 clusters which are found as draft data points of the clusters in Simulated Annealing Algorithm. Then, Euclidean distances of each customer data to each cluster are calculated. If a customer data is closer to a cluster than the other clusters, then that customer is assigned to that cluster which is the closest.

Let's assume that the clusters are determined in Simulated Annealing Algorithm as follows and that we are trying to assign the customer with code 4,1,2,1,1.:

Table 4.3 Example of K-Means

Clusters	Customer Code				
Cluster 1	1	1	2	1	1
Cluster 2	6	2	1	4	3
Cluster 3	3	1	3	2	1

The following table depicts the Euclidean distances of the customer with code 4,1,2,1,1 to each cluster:

Table 4.4: Determination of Cluster of a customer

Customer Number	Customer Code					Distance Cluster1	Distance Cluster2	Distance Cluster3	min	Cluster
N7	4	1	2	1	1	3	4.358	1.732	1.732	C3

After the comparisons, it can be clearly observed that the customer must be clustered in cluster 3.

Through these consecutive calculations, each customer is assigned to 3 clusters and the data points of each cluster are found again with computing the average of the data regarding to the customers assigned to that cluster. Accordingly, the data points of 3 clusters are changed. Then, Euclidean distances of each customer data to the new clusters must be calculated again. After that, each customer will be assigned to new clusters according to their minimum distances and data points of each cluster are calculated again based the average of the data points of the customers. If the data points of the new clusters are exactly the same with the previous data points of clusters, the algorithm stops. If no, the algorithm continues to find the solution which is the best approximation of the optimum solution. Finally, data points of 3 clusters are determined and each customer is assigned to a cluster. Therefore, the clustering process is completed. Next steps apply MCDM methods to find the most valuable cluster.

4.2.2.2 Ranking the Clusters

4.2.2.2.1 AHP Part. After all clusters are determined and the customers are assigned to the most suitable clusters, the decision maker would need to know the most valuable cluster because the user would like to develop strategies to meet the most valuable cluster's needs and increase their retention. In order to do that, firstly it's needed to calculate the weights of the criteria. In this step, AHP, of which process flow is shown in section 4.1, is used to calculate the weights of the criteria. First of all, relative importance values of the criteria are received from the user using Saab (1995)'s scale through an input box.

Fig 4.11 and Fig 4.12 show the input box view of pairwise comparisons during data entrance:

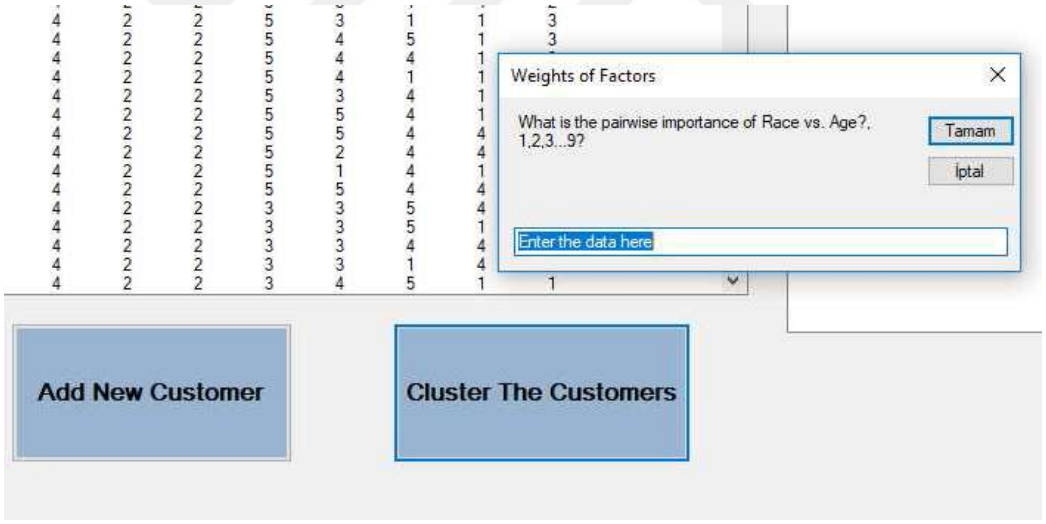


Figure 4.11 Pairwise comparisons input box in program (Race vs. Age)

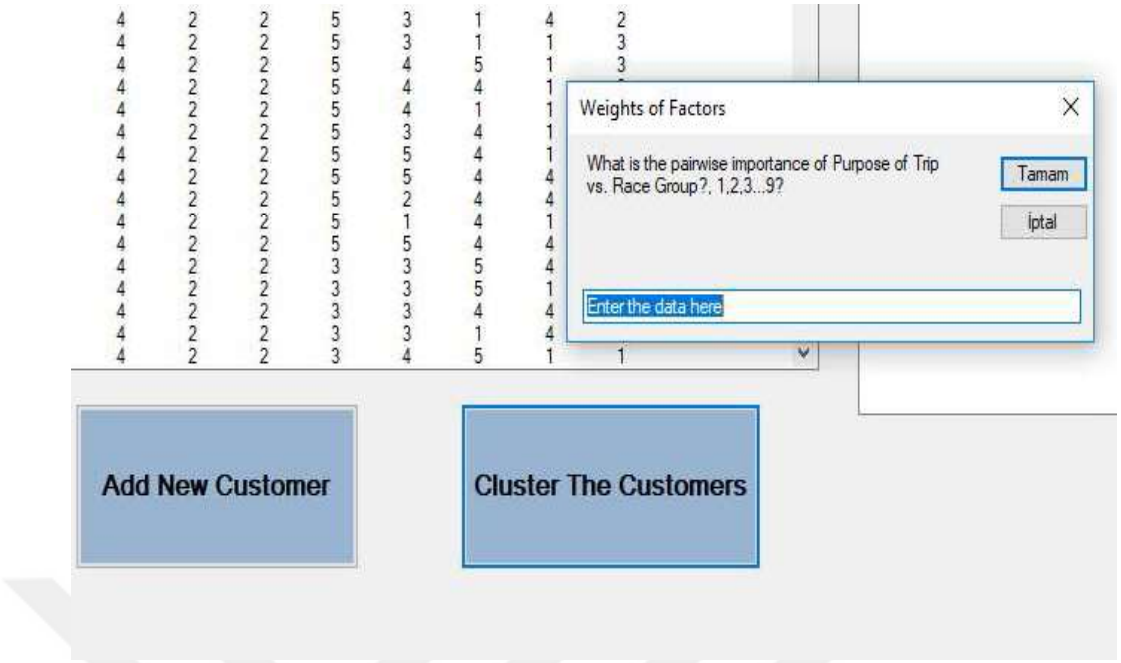


Figure 4.12 Pairwise comparisons input box in program (Purpose of Trip vs. Race)

Finally, the pairwise comparison weights are determined as follows:

Table 4.5 Pairwise comparison weights in AHP

	Age	Gender	Race Group	Employement Status	Purpose of trip	Previous Trip to Jakarta	Travel Companion	Travel Arrangement
Age	1.00	9.00	7.00	0.11	0.20	5.00	2.00	0.11
Gender	0.11	1.00	2.00	0.11	0.20	2.00	0.33	0.11
Race Group	0.14	0.50	1.00	0.14	0.14	2.00	0.33	0.11
Employement Status	9.00	9.00	7.00	1.00	5.00	7.00	7.00	5.00
Purpose of trip	5.00	5.00	7.00	0.20	1.00	7.00	5.00	0.20
Previous Trip to Jakarta	0.20	0.50	0.50	0.14	0.14	1.00	0.20	0.14
Travel Companion	0.50	3.00	3.00	0.14	0.20	5.00	1.00	0.14
Travel Arrangement	9.00	9.00	9.00	0.20	5.00	7.00	7.00	1.00

With the use of the values regarding to pairwise comparison weights, eigenvector matrix is established. According to the results, which are computed by the decision

support system using the AHP flow, which is shown in section 4.1, eigenvector matrix is calculated (See Fig 4.13, events list box) as follows:

Table 4.6 Eigenvector matrix

Age	Gender	Race Group	Employement Status	Purpose of trip	Previous Trip to Jakarta	Travel Companion	Travel Arrangement
0.09	0.07	0.02	0.31	0.14	0.02	0.1	0.26

The values in the matrix represents normalized weights of the criteria. They will be used in TOPSIS during ranking the clusters.

4.2.2.2.2 TOPSIS Part. In this section of the computational experiment, the aim is to determine the ranking of the clusters. The conventional steps of the TOPSIS algorithm includes calculation of the weights of the criteria. However, the weights of the criteria and the eigenvector matrix regarding to them has been already calculated by AHP.

In this step, TOPSIS procedure starts by using AHP weights to calculate positive and negative ideal solutions, which will be gathered from the matrix which is created by multiplying the clusters and criteria matrix with eigenvector (weights). Then, the values in this matrix is normalized with the formula which is given as 2.1 in Section 2.3.2 In each criterion column, the highest and the lowest values are chosen and solution sets are created.

In this numerical experiment, positive and negative ideal solutions are calculated as follows:

Table 4.7 Positive and Negative Ideal Solutions in TOPSIS

Negative Ideal Solutions (V0)							
Age	**Gender**	**Race Group**	**Employement Status**	**Purpose of trip**	**Previous Trip to Jakarta**	**Travel Companion**	**Travel Arrangement**
0.024	0.007	0.002	0.078	0.013	0.003	0.008	0.021

Positive Ideal Solutions (V1)							
Age	**Gender**	**Race Group**	**Employement Status**	**Purpose of trip**	**Previous Trip to Jakarta**	**Travel Companion**	**Travel Arrangement**
0.027	0.007	0.002	0.083	0.047	0.004	0.024	0.022

Afterwards, the separation of each cluster from the ideal points (S_i^+ and S_i^-) are calculated using the formula 2.5 and 2.6 given in the section 2.3.2.

In this computational experiment, they are calculated as follows:

Table 4.8 Calculation of S_i^+ and S_i^-

Cluster1	**Cluster2**	**Cluster3**
0.029	0.001	0.037
0.016	0.034	0.005

Finally, the distances are converted into a single metric which is called as Relative Closeness to the Ideal Solution. It is calculated for each cluster with the formula 2.7 given in the section 2.3.2.

In our computational experiments, calculation results are obtained as follows:

Cluster 1: 0.638

Cluster 2: 0.233

Cluster 3: 0.873

To determine the most valuable cluster, they are ranked according to the values of relative closeness to the Ideal Solution and it is clearly seen that:

Cluster 3 > Cluster 1 > Cluster 2

In the first part of the decision support system, the clusters are determined as below:

- Cluster 1: 4 2 2 4 5 4 1 1
- Cluster 2: 4 2 2 4 1 4 3 1
- Cluster 3: 5 2 2 4 6 4 4 1

Which means,

- Cluster1: Age 18-25; Male; Asian; Student; Vacation/Business; 1 Time Visited Jakarta before; w/ a partner; Self arrangement
- Cluster2: Age 18-25; Male; Asian; Student; Vacation/Sightseeing; Never came before; w/ a partner; Self arrangement
- Cluster3: Age 35-44; Male; Asian; Employed; Vacation/Sightseeing; 1 Time visited Jakarta before; w/friends; Self-arrangement

Finally, the clusters are ranked as:

The most important (or valuable) cluster:

Cluster3: Age 35-44; Male; Asian; Employed; Vacation/Sightseeing; 1 Time visited Jakarta before; w/friends; Self-arrangement

The second most important cluster:

Cluster1: Age 18-25; Male; Asian; Student; Vacation/Business; 1 Time Visited Jakarta before; w/ a partner; Self arrangement

The least important cluster:

Cluster2: Age 18-25; Male; Asian; Student; Vacation/Sightseeing; Never came before; w/ a partner; Self arrangement

Overall results are shown as follows in the “EVENTS” list box:

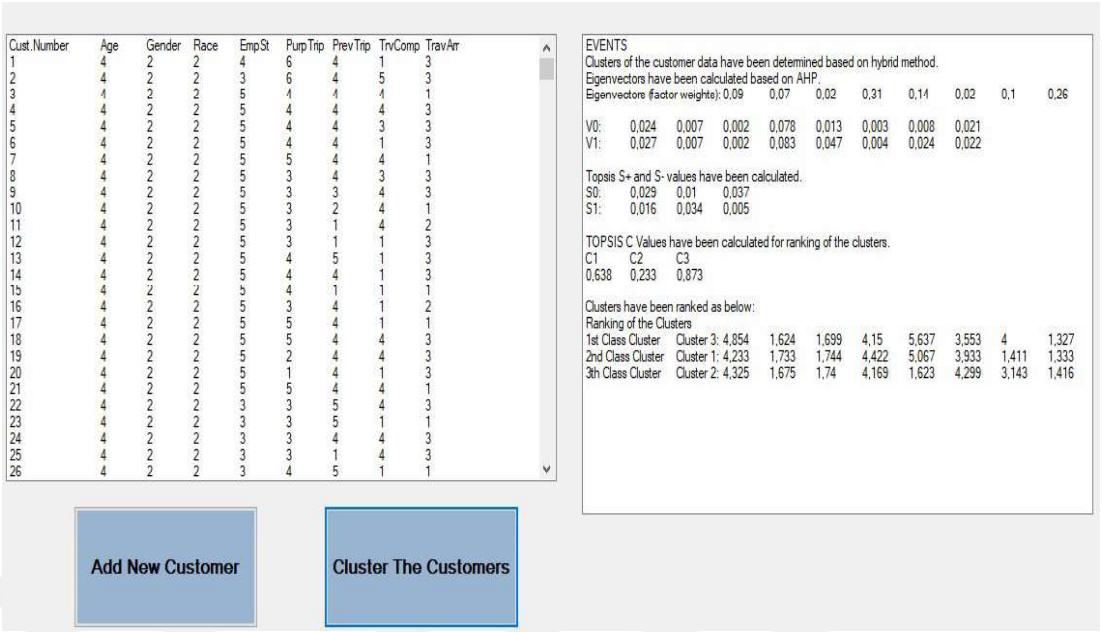


Figure 4.13 The results shown in program

CHAPTER FIVE

CONCLUSIONS

In this thesis, a decision support system is developed to cluster big data by using a hybrid methodology which integrates Simulated Annealing Algorithm and K Means Algorithm and the resulting clusters are ranked by using a Multi Criteria Decision Making methodology integrating Analytical Hierarchy Process and TOPSIS. The algorithms behind the decision support system is programmed and integrated by using Vb.net within the scope of this thesis.

To the best of our knowledge, there is no study that combines K Means and simulated annealing for the purpose of improving the initial clusters determined by K Means. Although it is quite flexible, easy to use and one of the widely used methods in the relevant literature, the conventional K-Means algorithm includes randomness during the selection of the first input data which will be used in the further steps of algorithm to obtain the final clusters. In order to overcome this deficit, Simulated Annealing algorithm is integrated to K-Means algorithm, therefore randomness of K-Means algorithm is reduced and it is modified to result in more robust and reliable solutions.

Two of the most widely used Multi Criteria Decision Making methods, which are TOPSIS and AHP, are used in combination with K-Means and Simulated Annealing Algorithms for the first time in the literature. Moreover, it is also the first time that TOPSIS and AHP are used as an integrated method to rank the clusters and support the decision makers to develop strategies to find out the most valuable customer with their outputs. Another contribution of this thesis is, the process of clustering and ranking the clusters is programmed with a user friendly interface using VB.net. The developed program has many implication areas in practical cases and in the literature as well. The program is rapid, easy to use and flexible for improvements, also its results can simply be analyzed. On the other hand, with small modifications, it is possible to

use it for other purposes instead of clustering and ranking the clusters. It can be widely used in analysis of any data and making a decision with based on the analysis.

For the future studies, our suggestions are; the program is currently developed to find three clusters because 3 is a popular number in practice to determine number of clusters for the companies and also in the literature besides it is simple to manage so that the program could be modified to find more clusters depending on selection of user. It would also be beneficial that the data might be analyzed by using other Multi Criteria Decision Making methods and the results of it might be compared with this one.

REFERENCES

Babcock, C. (1994). Parallel Processing Mines Retail Data. *Computer World, 6,* 95-108.

Bhanu B., Lee, S. and Das, S. (1995). Adaptive image segmentation using genetic and hybrid search methods. *IEEE Transactions on aerospace and electronic systems, 31,* 1268 – 1291.

Bose, I. and Chen, X. (2009). Hybrid models using unsupervised clustering for prediction of customer churn. *Journal of Organizational Computing and Electronic Commerce, 19,* 133-151.

Bradley, F.S. and Fayyad, U.M. (1998). Refining Initial Points for K-Means Clustering. *Proceedings of the 15th International Conference on Machine Learning, ICML98,* 91-99.

Brito, P. Q., Soares, C., Almeida, S., Monte, A. & Byvoet, M. (2015). Customer segmentation in a large database of an online customized fashion business. *Robotics and Computer - Integrated Manufacturing, 36,* 93–100.

Buonamente M., Dindo, H., Chella A., and Johnsson M. (2018). Simulating music with associative self-organizing maps. *Biologically Inspired Cognitive Architectures, 25,* 135-140.

Chen, I.D. & Popovich, K. (2003), Understanding customer relationship management (CRM). People, process and technology. *Business Process Management Journal, 9* (5), 672-688.

Chen, Z. & Shixiong, X. (2009). K Means Clustering Algorithm with improved initial cluster. *Second International Workshop on Knowledge Discovery and Data Mining,* 790-792.

Chiang, W. (2014). Mining high value travelers using a new model designed for online CRM systems: A case study in Taiwan. *Aviation, 18* (3), 157-160.

Chipman, H. and Tibshirani, R. (2005). Hybrid hierarchical clustering with applications to microarray data. *Biostatistics, 7,* 386 – 301.

Chiu, W., Tzeng, G. and Li, Han. (2013). A new hybrid MCDM model combining DANP with VIKOR to improve e-store business. *Knowledge-Based Systems, 37,* 48-61.

Csikosova, A., Antosova, M. & Mihalcova, B. (2015). Segmentation of Airports' Customers in Slovakia. *Procedia Economics and Finance, 23*, 1068-1073.

Dimitriadis, S., Stevens, E. (2008). Integrated customer relationship management for service activities, *Managing Service Quality: An International Journal, 18* (5), 496-511.

Du, K.-L. (2010). Clustering: Neural Network Approach. *Neural Networks, 23,* 89-107.

Fong, P. & Choi, S. (2000). Final contractor selection using the analytical hierarchy process. *Construction Management and Economics, 18* (5), 547-557.

Fontela, E., & Gabus, A. (1976). The DEMATEL observer. Phd Thesis. Battelle Geneva Research Center, Geneva.

Gefen, D., & Ridings, C.M. (2002). Implementation team responsiveness and user evaluation of customer relationship management: A quasi-experimental design study of social exchange theory. *Journal of Management Information Systems, 19* (1), 47-69.

Golcuk, I., Baykasoglu, A. (2016). An analysis of DEMATEL approaches for criteria interaction handling within ANP, *Expert Systems with Applications, 46,* 346-366.

Gandomi, A. & Haider, M. (2015). Data mining techniques in social media: A survey. *International Journal of Information Management, 35,* 137–144.

Hand, D. (1998). Data Mining: Statistics and More? *The American Statistician, 52* (2), 112-118.

Hamka, F., Bouwman, H., Reuver, M. & Kroesen, M. (2014). Mobile customer segmentation based on smartphone measurement. *Telematics and Infromatics, 31*, 220-227.

Hatamlou, A. (2013). Black hole: A new heuristic optimization approach for data clustering. *Information Sciences, 222,* 175-184.

Hong, Y., Duan, Q., Li, D. & Wang, J. (2013). K Means clustering algorithm for fish image segmentation. *Mathematical and Computer Modelling, 58,* 790–798.

Hwang, C. L., & Yoon, K. (1981). *Multiple attribute decision making, methods and applications* (1th ed.). New York: Springer-Verlag.

Isik, A. H., Ince, M. & Yigit, T. (2015). A Fuzzy AHP Approach to Select Learning Management System. *International Journal of Computer Theory and Engineering, 7* (6), 499-502.

Kahreh, M. S., Tive, M., Babania, A. & Hesan, M. (2014). Analyzing the applications of customer lifetime value (CLV) on benefit segmentation for the banking sector. *Social and Behavioral Sciences, 109,* 590 – 594.

Khlif, H. & Jallouli, R. (2014). The Success factors of CRM systems: An Explanatory Analysis, *Journal of Global Business and Technology, 10* (2), 25-42.

Kolarovzski, P., Tengler, J. & Majercakova, M. (2016). The new model of customer segmentation in postal enterprises, *Social and Behavioral Sciences, 230,* 121-127.

Konu, H., Loukkanen, T. & Komppula, R. (2011). Using ski destination choice criteria to segment Finnish ski resort customers. *Tourism Management, 32,* 1096-1105.

Kotler, P. (1980). *Marketing Management: Analysis, Planning, and Control* (4th ed.). New Jersey: Prentice-Hall, Englewood Cliffs.

Kusumawardani, R. P. & Agintiara, M. (2015). Application of fuzzy AHP-TOPSIS mehod for decision making in human resource manager selection process, *Procedia Computer Science, 72,* 638-646.

Lawson-Body, A., & Limayem, M. (2004). The impact of customer relationship management on customer loyalty: The moderating role of web site characteristics. *Journal of Computer Mediated Communication, 9* (4), 20-35.

Lefebure, R. & Ventury, G. (2001). Gestion de la relation client Panorama des produits et conduite de projets. *Eyrolles, 7* (1), 280-302.

Lee, Y., Tang, N. & Sugumaran, V. (2014). Open Source CRM Software Selection using the Analytic Hierarchy Process. *Information Systems Management, 31*, 2-20.

Li, D., Dai, W. & Tseng, W. (2011). A two-stage clustering method to analyze customer, characteristics to build discriminative customer management: A case of textile manufacturing business. *Expert Systems with Applications, 38,* 7186-7191.

Logeswari, T. & Karman, M. (2010). An improved implementation of brain tumor detection using segmentation based on hierarchical self-organizing maps. *International Journal of Computer and Engineering, 2* (4), 1793 – 8201.

Lopez, J. J., Aguado, J. A., Martin, F., Munoz, F., Rodriguez, A. & Ruiz, J. E. (2011). Hopfield K-Means Clustering Algorithm: A proposal for the segmentation of electricity customers. *Electric Power Systems, 81,* 716-724.

Mondal, S. A. (2018). An improved approximation algorithm for hierarchical clustering. *Pattern Recognition Letters, 104*, 23-28.

Mostafa, K., Batool, R. And Parvaneh, P. (2013). Identify and ranking the factors affecting customer satisfaction in carpet industry (Case study: Sahand Carpet Co.). *Research of Journal of Management Sciences, 2 (6),* 1-8.

Muralidhar, K., Santhnam, R. & Wilson, R.L. (1990). Using the analytic hierarchy process for information system project selection. *Information and Management, 2,* 87–95.

Opricovic, S., & Tzeng, G. - H. (2004). Compromise solution by MCDM methods: A comparative analysis of VIKOR and TOPSIS. *European Journal of Operational Research, 156,* 445–455.

Ozgener, S. & Iraz, R. (2006). Customer relationship management in small-medium enterprises: The case of Turkish tourism industry. *Tourism Management, 27,* 1356-1363.

Perçin, S. (2009). Evaluation of third-party logistics (3PL) providers by using a two-phase AHP and TOPSIS methodology. *Benchmarking: An international journal, 16,* 588-604.

Pohekar, S.D. & M. Ramachandran (2004). Application of multi-criteria decision making to sustainable energy planning - A review. *Renewable and Sustainable Energy Reviews, 8,* 365 – 381.

Rowland, T. Moriarty David J. Reibstein. (1986). Benefit Segmentation in Industrial Markets. *Journal of Business Research, 14*, 463-486.

Roy, B. (1990). The outranking approach and the foundations of electre methods. *Readings in Multiple Criteria Decision Aid*, *1*, 155–183.

Rygielski, C., Wang, J. & Yen, D. (2002). Data mining techniques for customer relationship management. *Technology in Society 24,* 483–502.

Saaty, T. L. (1980). *The analytic hierarchy process* (1st ed.). NewYork: Mc Graw-Hill.

Saaty, T.L. (1995) *Decision Making for Leaders: The Analytical Hierarchy Process for Decision in a Complex World* (1st ed.). Pittsburgh: RWA Publications.

Saaty, T. L. (1996). *Decision making with dependence and feedback: The analytic network process* (1st ed.). Pittsburgh: RWS Publications.

Shih, H., Shyur, H. & Lee, E.C. (2007). An extension of TOPSIS for group decision making. *Mathematical and Computer Modelling, 45,* 801–813.

Shih, Y. & Liu, C. (2003). A method for customer lifetime value ranking – Combining the analytical hierarchy process and clustering analysis. *Database Marketing & Customer Strategy Management, 11* (2), 159-172.

Smith, W., (1956). Product differentiation and market segmentation as alternative marketing strategies. *Journal of Marketing 21*, 3–8.

Sipahi, S. & Timor, M. (2010). The analytic hierarchy process and analytic network process: An overview of applications. *Management Decision, 48 (5),* 775–808.

Srihadi, T. F., Hartoyo, Sukandar, D. & Soehadi, A. W. (2016). Segmentation of the tourism market for Jakarta: Classification of foreign visitors' lifestyles typologies. *Tourism Management Perspectives, 19,* 32-39.

Velmurugan, T & Santhanam, T. (2010). Computational Complexity between K-Means and K-Medoids Clustering Algorithms for Normal and Uniform Distributions of Data Points. *Journal of Computer Science, 6 (3),* 363-368.

Verhoef P.C., Franses P.H. & Hoekstra J.C. (2001). The impact of satisfaction and payment equity on cross-buying: A dynamic model for a multi-service provider. *Journal of Retailing 77* (3), 359-378.

Vincentelli, A.S., Chen, L. K. And Chua, L. O. (1977). An efficient heuristic cluster algorithm for tearing large scale networks. *IEEE Transactins on circuits and systems 24, 12,* 709-717.

Wahlberg, O., Strandberg, C., Sundberg, H. & Sandberg, K. W. (2009). Trends, Topics and Under – Researched Areas in CRM Research. *International Journal of Public Information Systems, 09* (3), 191-208.

Wang, Y., Ma, X., Lao, Y. & Wang, Y. (2014). A fuzzy-based customer clustering approach with hierarchical structure for logistics network optimization. *Expert Systems with Applications, 41*, 521-534.

Wang, B., Miao, Y., Zhao, H., Jin, J. & Chen, Y. (2016). A bi-clustering-based method for market segmentation using customer pain points. *Engineering Applications of Artificial Intelligence, 47,* 101–109.

Ward J.H. (1963). Hierarchical grouping to optimize an objective function, *I Am Statist Assoc 58*, 236–244.

Wahlberg O., Strandberg C., Sundberg H., & Sandberg W.K. (2009). Trends, Topics and Under-Researched Areas in CRM Research – A Litreature Review: International Journal of Public Information Systems, *09* (3), 192-208.

Yong, D. (2006). Plant location selection based on fuzzy TOPSIS. *International Journal of Advanced Manufacturing Technologies, 28,* 839-844.

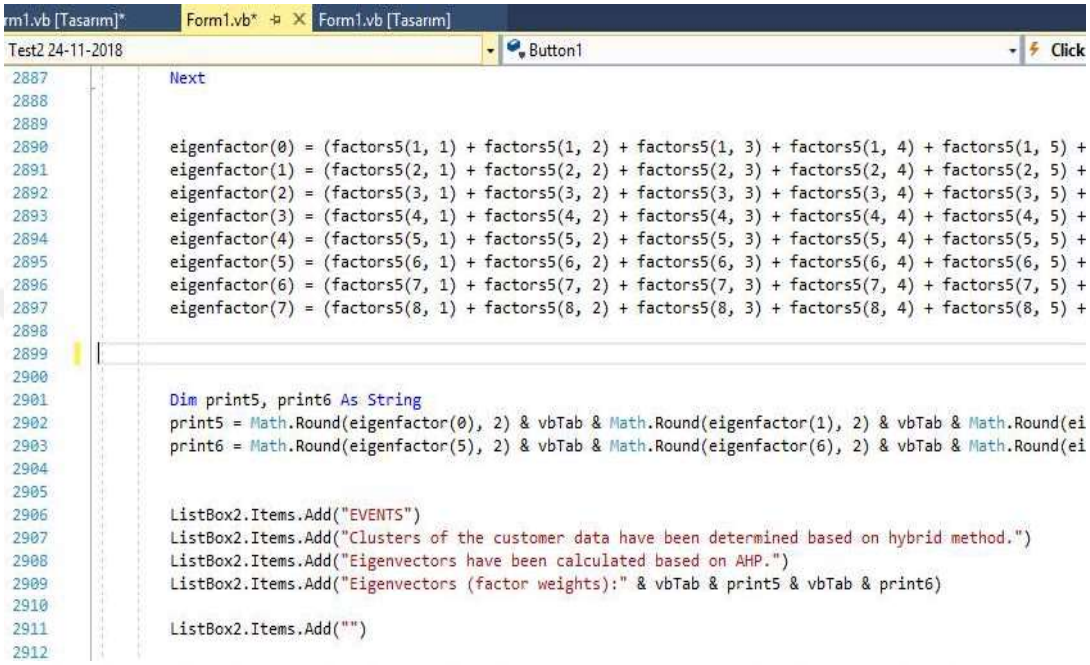
Zhang, C., Ouyang D. And Ning J. (2010). An artificial bee colony for clustering. *Expert systems with Applications, 37,* 4261 – 4267.

Zhao, Y and Karypis, G. (2005). Hierarchical clustering algorithms for document datasets. *Data Mining and Knowledge Discovery 10,* 141-168.

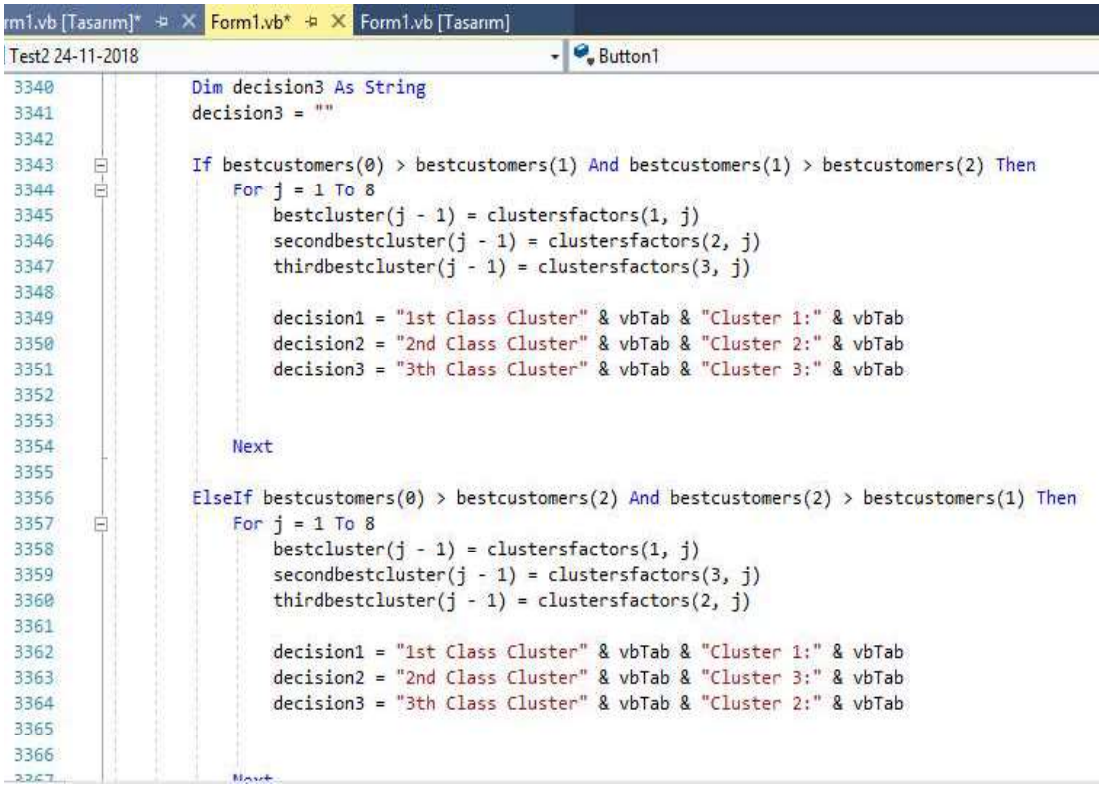
APPENDICES

Appendix A1:

Part of Program Codes:



```
Test2 24-11-2018                                   Button1                                   Click
2887        Next
2888
2889
2890        eigenfactor(0) = (factors5(1, 1) + factors5(1, 2) + factors5(1, 3) + factors5(1, 4) + factors5(1, 5) +
2891        eigenfactor(1) = (factors5(2, 1) + factors5(2, 2) + factors5(2, 3) + factors5(2, 4) + factors5(2, 5) +
2892        eigenfactor(2) = (factors5(3, 1) + factors5(3, 2) + factors5(3, 3) + factors5(3, 4) + factors5(3, 5) +
2893        eigenfactor(3) = (factors5(4, 1) + factors5(4, 2) + factors5(4, 3) + factors5(4, 4) + factors5(4, 5) +
2894        eigenfactor(4) = (factors5(5, 1) + factors5(5, 2) + factors5(5, 3) + factors5(5, 4) + factors5(5, 5) +
2895        eigenfactor(5) = (factors5(6, 1) + factors5(6, 2) + factors5(6, 3) + factors5(6, 4) + factors5(6, 5) +
2896        eigenfactor(6) = (factors5(7, 1) + factors5(7, 2) + factors5(7, 3) + factors5(7, 4) + factors5(7, 5) +
2897        eigenfactor(7) = (factors5(8, 1) + factors5(8, 2) + factors5(8, 3) + factors5(8, 4) + factors5(8, 5) +
2898
2899
2900
2901        Dim print5, print6 As String
2902        print5 = Math.Round(eigenfactor(0), 2) & vbTab & Math.Round(eigenfactor(1), 2) & vbTab & Math.Round(ei
2903        print6 = Math.Round(eigenfactor(5), 2) & vbTab & Math.Round(eigenfactor(6), 2) & vbTab & Math.Round(ei
2904
2905
2906        ListBox2.Items.Add("EVENTS")
2907        ListBox2.Items.Add("Clusters of the customer data have been determined based on hybrid method.")
2908        ListBox2.Items.Add("Eigenvectors have been calculated based on AHP.")
2909        ListBox2.Items.Add("Eigenvectors (factor weights):" & vbTab & print5 & vbTab & print6)
2910
2911        ListBox2.Items.Add("")
2912
```



```
Test2 24-11-2018                                   Button1
3340        Dim decision3 As String
3341        decision3 = ""
3342
3343        If bestcustomers(0) > bestcustomers(1) And bestcustomers(1) > bestcustomers(2) Then
3344            For j = 1 To 8
3345                bestcluster(j - 1) = clustersfactors(1, j)
3346                secondbestcluster(j - 1) = clustersfactors(2, j)
3347                thirdbestcluster(j - 1) = clustersfactors(3, j)
3348
3349                decision1 = "1st Class Cluster" & vbTab & "Cluster 1:" & vbTab
3350                decision2 = "2nd Class Cluster" & vbTab & "Cluster 2:" & vbTab
3351                decision3 = "3th Class Cluster" & vbTab & "Cluster 3:" & vbTab
3352
3353
3354            Next
3355
3356        ElseIf bestcustomers(0) > bestcustomers(2) And bestcustomers(2) > bestcustomers(1) Then
3357            For j = 1 To 8
3358                bestcluster(j - 1) = clustersfactors(1, j)
3359                secondbestcluster(j - 1) = clustersfactors(3, j)
3360                thirdbestcluster(j - 1) = clustersfactors(2, j)
3361
3362                decision1 = "1st Class Cluster" & vbTab & "Cluster 1:" & vbTab
3363                decision2 = "2nd Class Cluster" & vbTab & "Cluster 3:" & vbTab
3364                decision3 = "3th Class Cluster" & vbTab & "Cluster 2:" & vbTab
3365
3366
3367            Next
```