ISSN No. 0976-5697

Volume 8, No. 5, May-June 2017

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

Data Mining on Customer Segmentation: A Review

Er. Rupampreet Kaur
M.Tech. Computer Science & Engineering,
Department of Computer Engineering and Technology
Guru Nanak DevUniversity, Amritsar, India

Er.Kiranbir Kaur
Assistant Professor,
Department of Computer Engineering and Technology,
Guru Nanak DevUniversity, Amritsar, India

Abstract: Data mining is used to extract important information from the bulk of data to save it and summarize it in effective manner. The hidden information can be extracted from the large set of data. The goal of this paper is to investigate the methods that are used for efficient grouping of data. The grouping must be done in its manner that the group can recognize the group members and group can also recognize still, not grouped member so far. Different approaches for customer segmentation in data mining are: clustering and subgroup discovery. Because of some limitations and scope of the clustering techniques, it leads to further refinements in methodology in data mining.

keywords: Data mining, Big data, Data segmentation, Clustering, Subgroup discovery.

1. INTRODUCTION

Processing large amount of dataneedsdifferent methods, tools and architectures to store it efficiently, so Data mining is the technique to handle large amount of data having different tools to process it. The key reason for the development of "Large Data" is actually because of increase in processing power and availability of information. The current status of international market segmentation is provided either by big data approaches or data mining techniques [1]. The segments that predict international market segmentation involves studying the geographic configuration, validation, results efforts . Big data techniques may be new and difficult to handle so data mining techniques are preferred by most companies. Companies find data mining like "business as usual" or use data mining techniques to continue evolution toward more data [2]. This may be used in banks to enhance the increasing volume and variety of data about their customers. The frequent use of bank web technologies like online websites from which the customer may proceed its transaction from one account to another or online payments and also mobile channels or applications has led to a steady increase in volume of customer data. So the data mining approach have the great influence in banks and other financial institutions[3]. Most of the financial institutions in North America believe that analytic techniques offered by big data and data mining have a significant competitive advantage for big companies and data mining initiatives will define the significant change in storage of data[4]. The most important goal of Big Data initiatives is to analyze diverse data sources and new data types, not only to manage the large data sets. So, we sought to better understand after knowing deeply about data mining, data segmentation, clustering and subgroup discovery.

1.1 DATA MINING

Data mining considered to be exploratory rather than confirmatory. Data mining is used to extract important information from the bulk of data and save it and summarize it in effective manner. Data mining used to extract hidden information from large set of the data [1]. Table 1 explains different types of data mining techniques.

Table 1- data mining techniques

Direction	Use	Examples
Supervised (Directed)	Used for hypothetical testing	Classification, estimation and prediction
Unsupervised (Undirected)	Used to get new information	Association rules, clustering

Algorithms that classify data mining techniques are
1) Association Rule or Recommendation system
2) Clustering.
3) Classification.
4) Regression
5) Anomaly detection

Recommendation system is used to search relationship between customer and the management from the production side. Most company's focus only the CRM i.e. Customer Relation Management that deals only with customers. It is applied to search most frequently visited items by the user. In short it establishes relationship among objects. The anomaly detection is used to find the text that is completely irrelevant from the entire data[1]. Different data mining algorithm examples are as follows:

Table 2 Data mining algorithms

DATA MINING ALGORITHMS	DEFINATION	EXAMPLES
Association	Association rules are used to specify users the recent trends in market	Apriori, FP growth, ECLAT Partition
Clustering	The process of arrangement of objects in groups whose members characters relate to each other in some way.	K-Means, fuzzy C Means, DB SCAN, Expectation Maximization,
Classification	Classification predicts the certain outcome based on given input	KNN, Naïve Bayes, SVM,C4.5, Decision Trees.
Regression	Regression is used to determine the relation between the dependent variable and independent variables the independent variables may be more than one..	Multivariate Linear regression

1.2 DATA SEGMENTATION

Data Segmentation is division of data that is similar in specific way. This technique is used in market that targets the groups of customers having same demands or likings. In this way the production of product is analyzed i.e. liking and disliking of the product by the customers can be analyzed easily[5]. It is a conventional online strategy, validated plus discussed in each and every manual focused on business. Equally conceptually plus essentially, operators are aware that they can satisfy any shopper entirely. Thus, the rationale will be to separate clients into teams, then concentrate on the marketing and advertising work to the almost all attractive segment. In this case charm indicates profits plus sustainability[6]. The main target of segmentation was to separate the objects that are homogeneous and heterogeneous with the external market (the consumers).On the other hand irrespective of which procedure is often used, the decision is hardly ever computerized or perhaps totally data driven. The outcome of segmentation depends mainly on the knowledge variables, which can be gathered from market, psychographic, regional, life-style, etc.[7]. The customer segmentation is of two types i.e. clustering and subgroup discovery. Target was to divide the external market. The final choice is very rare to be automatic or fully data driven. Many important decisions have to be identified like the segment selection or which segment to choose then to identify the segment and then decide their relative size. Segmentation depends on input variable. The input variables are further subdivided in demographic, psychographic, geographic and life style [5].

1.3 CLUSTERING

From the whole study we can understand clustering is known to be unsupervised data mining technique[6].The clustering can be defined in two different ways i.e. soft clustering and hard clustering. Soft clustering refers to the data or cluster that can also belong to other cluster whereas hard clustering is defined as data or cluster that belongs to its own data or cluster. So in soft clustering the same data can define 10 more clusters [10]. Different types of clusters that are extracted from data mining techniques are k-means clustering, hierarchical clustering, agglomerative clustering, divisive clustering[8].Clustering is used to discover groups and structures in the data. And it classifies that which data belongs to which group. As clustering is unsupervised data mining technique this means that it doesn't require any target variable. The data mining technique with supervised channel need target variable. In supervised learning target variable is further categorized in continuous or classical method [6].It helps consumers to understand the natural group or structure in a very data fixed. Used either as a stand-alone tool for getting insight straight into data submission or as a pre-processing phase for other algorithms [17].

1.4 SUBGROUP DISCOVERY

Relation between a target variable that is dependent on another variable and independent variables that is not dependent on any other variable is identified by the subgroup discovery.This particular technique is some time almost between predictive along with illustrative induction, and it is goal will be to obtain in an along with interpretable manner subgroups to clarify relationships between separate specifics including a number of worth on the goal variable. This algorithm with this process has to produce subgroups for each and every worth on the goal variable. Hence, a strong setup for each and every worth on the varied must be performed[19]. Any rule(R), which consists of activated subgroup information, might be officially described as

R: Cond $\rightarrow Target_{value}$

Where $Target_{value}$ is a value of the variable to which the condition target for the subgroup discovery task and Cond is commonly a collection of features and attributes which are able to describe an unusual statistical distribution with respect to the $Target_{value}$.

2. RELATED WORK

Jong Tak ,et al.[9] discussed as the population increase, health services also increase that can be accessed by personal computers and smart phones to provide health related problems. P2P is based on PBR (Personal Bio

Record) platform for increasing silver population. The health cluster model used to provide services in multi - platform environment and enhance health status of old aged patients having chronic diseases. Noble, et al. [12] analyzed use three clusters to identify the latent class analysis. The methodology of clustering was performed on 377 participants that attend aboriginal community controlled health service (ACCHS) in Australia. In this cluster one survey the low fruit/ vegetable intake , cluster two includes younger unemployed males who have smoking, alcohol addiction problem whereas cluster three include depressed personalities. Cluster three only include younger to mid-aged women.Alzahrani , et al.[15] reviewed two clusters in which one cluster has low fruit consumption and second one include high sweet consumption. The author use hierarchical agglomerative cluster analysis (HASA). Main study was on school students whose age is between 13-14 and 17-19 years. The students were divided in different grades i.e. intermediate and secondary schools. Health related behaviors, demographics characteristics parents occupation was included in questionnaire. The result was further studied on 1150 students.Mehar ,et al.[18]The standard k-means clustering was used by the author. The k-means clustering was divided in three groups i.e. poor, intermediate and good outcomes. The surrogate value is included in this survey. The cluster includesthe data points from the surrogate variable value. The main limitation in this paper is to find the surrogate value. The surrogate value unstably correlated with the health outcomes. So the result may be inappropriate. The main survey was based on socio demographic background and is considered for health service utilization.Brito.et.al.[5]The investigation of the data mining approaches for customer segmentation was the main aim of this paper. k-medoids and CN2-SD are the data managerial approaches. These are used to segment problem and complement each other. The segmentation is further divided externally and internally. The external segment helps the company to redefine communication strategy for sales promotion and internally matching the product of customer preference that will help to redefine product design. Azizpour,et al.[11]Implemented the sources of corporate default clustering in the United States. We reject the hypothesis that firms' default times are correlated only because their conditional default rates depend on observable and latent systematic factors. By contrast, we find strong evidence that contagion, through which the default by one firm has a direct impact on the health of other firms, is a significant clustering source. The amount of clustering that cannot be explained by contagion and firms' exposure to observable and latent systematic factors is insignificant. Guha ,et al.[8] Clustering can be a useful and everywhere tool with data analysis. Commonly conversing, clustering is definitely the difficulty of group any data collection directly into several groups so that, underneath many specification of "likeness," very similar products are with the exact same group and dissimilar products are in different groups.Tsuboi.et al.[14]Within this phase we center on clustering in a very internet streaming predicament in which limited data products are introduced during a period and now we won't be able to retail store most the data points. Therefore, our own algorithms will be limited by a single pass. Space stops is generally sub linear, $i(in)$, in which the quantity of feedback factors can be in.

3. COMPARISION TABLE

Name	Year	Methodology used	Validation	Usage
1.Big Data Alchemy: How can Banks Maximize the Value of their Customer Data?	2013	Big data	analytics offers a significant competitive advantage	Increase volume of customer data
2.P2P-based u-health cluster service model for silver generation in PBR platform	2015	based on PBR (Personal Bio Record) platform	Valid for silver generation	used to provide services in multi - platform environment
3.The clustering of health behaviors in Ireland and their relationship with mental health, self rated health and quality of life	2011	identify 5 clusters which relates health problems	Valid for adults and adolescents	Used to get complete information about Irish people
4. Blended Clustering for Health Data Mining	2010	k-means clustering	based on socio demographic background	Used for health service utilization
5. Patterns of clustering of six-health –	2014	Identify clusters which relates low fruit and high sweet consumption	Valid by using hierarchical agglomerative cluster analysis (HASA).	Used for school students analysis

compromising behavior in Saudi adolescents				
6. A cross-sectional survey and latent class analysis of the prevalence and clustering of health risk factors among people attending an aboriginal community controlled health service	2015	Identify three clusters to identify the latent class analysis	Validation was done by aboriginal community controlled health service (ACCHS) in Australia	Includes younger to mid-aged women.
7. Customer segmentation in a large database of an online customized fashion business	2014	k-medoids and CN2-SD	used to segment problem and methodologies complement each other.	strategy for sales promotion and internally matching the product of customer preference that will help to redefine product design

CONCLUSION

Data mining used to extract hidden information from large set of data. So this paper shows about the comparison of various techniques based on the image segmentation and big data. These approaches are important for highly customized industries that use large amount of data. But still there are some issues that have not considered the use of big data with customer segmentation as well as the use of differential evolution is ignored in existing literature to classify the data. So in near future we will use differential evolution and also will protect our software for the premature results.

REFERENCES

[1] Shobana and Maheshwari and Savithri“Study on Big data with Data Mining” Vol. 4, Issue 4, April 2015,International Journal of Advanced Research in Computer and Communication Engineering(IJARCS),page 1.

[2] Thomas H. Davenport Jill Dyché“Big Data in Big Companies” May 2013, International institute for analytics, Thomas H. Davenport and SAS Institute Inc, page no1,2.

[3] Coumaros, J., J. Buvat, O. Auliard, S. Roys, S. KVJ, L. Chretien, and V. Clerk. "Big Data Alchemy: How can banks maximize the value of their customer data."Capgemini and EFMA, Retail Banking Voice of the Customer Survey, 2013.

[4] Robert Stackowiak, VenuMantha, Art Licht, AmbreeshKhana“Big Data in Financial Services and Banking :-Architect’s Guide and Reference Architecture Introduction” February 2015 ,Oracle Enterprise Architecture White Paper – Improving Banking and Financial Services Business Performance with Big Data, page number 3.

[5] Brito, Pedro Quelhas, Carlos Soares, Sérgio Almeida, Ana Monte, and Michel Byvoet. "Customer segmentation in a large database of an online customized fashion business." Robotics and Computer-Integrated Manufacturing , Elsevier ,2015.

[6] Zyxo“https://zyxo.wordpress.com/2010/07/17/the-difference-between-segmentation-and-clustering/” july 17,2010.

[7] Steenkamp, Jan-Benedict EM, and FrenkelTerHofstede. "International market segmentation: issues and perspectives." International Journal of Research in Marketing 19, no. 3 (2002).

[8] Guha, Sudipto, and Nina Mishra. "Clustering data streams:-In Data Stream Management”. Springer Berlin Heidelberg, 2016.

[9] Kim, Jong Tak, Hee-Jun Pan, and Jonghun Kim. "P2P-based u-health cluster service model for silver generation in PBR platform." Peer-to-Peer Networking and Applications . 14 July 2015 ,SpringerScience,Business Media New York 2015.

[10] Arshad Muhammad, Anthony Maeder, KenanMatawie, and AthulaGinige. "Blended clustering for health data mining." In E-Health, pp. 130-137. Springer Berlin Heidelberg, 2010

[11] Azizpour, Shahriar, Kay Giesecke, and Gustavo Schwenkler. "Exploring the sources of default clustering."June 15, 2011; this draft February 24, 2014.

[12] Noble, Natasha E., Christine L. Paul, Nicole Turner, Stephen V. Blunden, Christopher Oldmeadow, and Heidi E. Turon. "A cross-sectional survey and latent class analysis of the prevalence and clustering of health risk factors among people attending an Aboriginal Community Controlled Health Service."Noble et al. BMC Public Health (2015).

[13] Atzmueller, Martin. "Subgroup discovery." Wiley Interdisciplinary Reviews: Data Mining and Knowledge “Volume 4 Issue 2, March 2014

[14] Tsuboi, Satoshi, Takehito Hayakawa, Hideyuki Kanda, and Tetsuhito Fukushima. "The relationship between clustering health-promoting components of lifestyle and bone status among middle-aged women in a general population." Environmental health and preventive medicine, Springer(2009).

[15] Alzahrani, Saeed G., Richard G. Watt, Aubrey Sheiham, Maria Aresu, and GeorgiosTsakos. "Patterns of clustering of six health-compromising behaviours in Saudi adolescents” volume14(5), Alzahrani et al. BMC Public Health 2014

[16] Conry, Mary C., Karen Morgan, Philip Curry, Hannah McGee, Janas Harrington, Mark Ward, and Emer Shelley. "The clustering of health behaviours in Ireland and their relationship with mental health, self-rated health and quality of life."Conry et al. BMC Public Health (2011).

[17] sauravkaushik,"An Introduction to Clustering and different methods of clustering" Analytics vidhya –learn everything about analytics ,3 November 2016

[18] Grosskreutz, Henrik, Mario Boley, and Maike Krause-Traudes. "Subgroup discovery for election analysis: a case study in descriptive data mining."volume 6332,In International Conference on Discovery Science,Discovery Science pp 57-71.