International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163
Issue 4, Volume 2 (April 2015) www.ijirae.com

Application of Data Mining Techniques for Customer Segmentation in Real Time Business Intelligence

Mrs. Pradnya Muley*
MCA Department,
P.E.S Modern College of Engineering, Pune-05

Dr. Anniruddha Joshi
Department of Management Sciences,
Savitribai Phule Pune University

Abstract— Data Mining is an analytic process designed to explore data in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data. The ultimate goal of data mining is prediction - and predictive data mining is the most common type of data mining and one that has the most direct business applications. Customer segmentation is a core process for assisting an marketing strategy. Huge amount of customer data are continuously generated .The paper aimed to study different DM techniques used so far for customer segmentation in online retail Industry. To identify the gaps from previous studies which can be helpful for further research in the field.

Keywords— Data Mining, Predictive data mining, Customer segmentation, Online retail Industry

I. INTRODUCTION

With fast-paced changes and technological advancements, organizations are forced to be prepared and informed with the latest possible trends and application to achieve the competitive advantage. This trend is particularly witnessed among online retailers with the advent of WWW, where major parts of sales information process and transactions are handled electronically. The recent e-commerce sales estimates in India reflect that, approximately, $3.59 billion has been spent between 2012 and 2014, and is projected to grow to $17.52 billion in 2015 (Statistical, 2014). Customers of this segment leave the large amount of information like searching, ordering, and feedback provision resulting in the collection of a huge amount of raw data and storing the same in the database to generate accurate demand forecast trends to drive planning and inform resource allocations (Leventhal, 2010; Acito & Khatri, 2014). However, analysing such large amount of data is a challenging and time consuming process as the analyst often needs to divide data into multiple segments based on certain criteria and also to combine objectives into groups based on needs. It is not possible to determine these data volumes in the conventional way; thus Data Mining (DM) has been introduced, which is commonly used to segment effectively as it enables to differentiate customers' markets to meet their various needs; therefore, marketing efforts are manageable (Bulysheva & Bulyshev, 2012; Tan, Steinbach & Kumar, 2006). It is also apparent that the benefits of analytics will continue to expand and span a variety of dimensions, including overall improvement in the quality and speed of decision, better alignment of resources to strategies, increased revenue, and improvement in cost efficiencies.

II. BACKGROUND OF THE STUDY

Although DM techniques have been used widely to segment the customers, with the advent of big data, the popularity of application of this technique has increased profoundly perhaps due to advancement in technology, the computer processing power, and data storage capacity (Conen, 2011) and its capability to handle and manipulate large data set quickly and easily (Liu & Chen, 2007). Since big data has been recently been emerging, and various authors have documented their challenges (Bedeley & Iyer, 2014) in handling the massive amount of data efficiently and generating insights with real business value (Sun et al., 2014). "Big data is high-volume, high velocity, and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight, and decision making" (Gartner's IT Glossary). According to the 2012 IDC digital universe study ("The digital Universe in 2020"), it showed that between the periods of 2005 to 2020; the digital universe will grow by a factor of 300 from 130 exabytes to 40,000 exabytes or 40 trillion gigabytes (IDC, 2013). Secondary data management systems cannot manage such Big Data; therefore, there is a need for better data mining techniques that follow standard CRISP methodology.

Several machine algorithms are available and generally divided into supervised and unsupervised, where latter is used when a specific output variable is not available to explain or predict (e.g., clustering problems or link analysis). However, supervised algorithms are useful when an output attribute needs an explanation (Jones, 2014). Data mining comprises statistical and machine learning techniques for identifying trends and patterns in huge data sources like (a) classification (for example, example artificial neural network (ANN), decision tree analysis, rune induction, K-nearest neighbouring techniques, (b) estimation, (c) prediction (such as multivariate statistics, ANN), (d) clustering (e.g. k-means, hierarchical, Kohonen Networks), and (e) association rules (particularly for market basket analyses) (Hastie, Tibshirani, & Firedman, 2009; Larose, 2005). Although most competitive corporations are already using data mining to discover new and useful knowledge, the lack of domain specific practicable research has significantly hampered the utility of DM for many knowledge based industries (Wang & Wang, 2009; Pechenizkiy, Puuronen, & Tsymbal, 2008).

III. PREVIOUS WORK

The term segmentation was attributed to Smith (1956) and subsequently several authors (Sun, 2009; Yankelovich & Meer, 2006) have provided the broad overview; but, overall idea was to segment or cluster similar customers, who have similar characteristics of values, behaviours, and demographic pattern (Bailey et al., 2009). In order to segment the customer, various techniques have been proposed such as the cluster (Xia et al., 2010).

Previous studies have used different techniques for customer and market segmentation. For instance, the study by Hung et al. (2007) proposed Support Vector Clustering (SVC), while Kim and Ahn (2008) applied Genetic Algorithms (GAs) to segment the online shopping market. Li, Wang and Xu (2008) used Chameleon Based on Clustering Feature Tree (CBCFT), which is hybrid of the clustering tree of algorithm BIRCH with algorithm Chameleon. Similarly, the study by Farajian and Mohammadi (2010) identified rules in the data, i.e. they used K-means clustering algorithm and developed a new two-stage framework that analyzed the customer behaviour and an association rule inducer for analyzing bank databases. The algorithm identified groups of customers based on monetary, frequency, recency, and behavioural scoring predicators, which was segmented into three different profitable groups. The study used the Apriori association rule inducer to characterize the groups of customers by creating customer profiles and to enhance customer relationship. In line with Huang and Tsai (2008), the proposed thing was the hierarchical self-organizing segmentation model (HSOS) for marketing segmentation of real world multimedia.

The study by Varun et al. (2012) aimed to establish the relation between marketing campaign and customer segmentation along with the enhancement using the Recency Frequency Monetary (RFM) data mining technique. The study also aimed to propose appropriate Customer-Relationship- Model (CRM), by which the targeted customers were identified with the new segmentation method. The study was conducted in two phases. In first phase, K-Means clustering was included, where the customers were clustered according to their RFM. In the second phase, with demographic data, each cluster was again partitioned into new clusters. Finally, LTV was used to generate customer's profile.

Raju et al. (2014) proposed CART algorithm for learning method. Decision Tree was used for customer retention. EM Algorithm Support Vector Machine (SVM) and Logistic Regression were used and proposed in the clustering model in banking sector to detect fraud. Koudehi, Rajeh, Farazmand, Seyedhosseini (2014) developed two phases clustering models based on SOM and K-means technique to observe the traits of demographic and transaction data in each cluster.

Cluster analysis was the effective tool for customer segmentation as it minimized the difference between two different clusters. Clustering is classified into hierarchical method and partitioning method (Witten & Frank, 2000), while later is divided into exclusive (k-means, SOM) and overlapping. Overlapping (e.g. fuzzy c-means) is based on the theory of fuzzy, where an element can be a member of one or more sets (Hwang & Thill, 2007). Although K-means has been widely applied, still its accuracy depends on the choice of initial seeds (Milligan & Cooper, 1980); therefore, accurate designation of the market cluster is challenging. However, fuzzy c means is more flexible than k-means as it shows those objectives that have some interface with more than one cluster in the partition. Fuzzy c-means developed by Bezdek (1981) includes a parameter, i.e. fuzzifer denoted as m, and chose from (o, ∞) in advance. Previous studies have emphasized the benefits of combining two clustering method (Vesanto & Alhoniemi, 2000) and not many studies by combining fuzzy C mean with SOM. In this study, which is based on the limitations of previous K-means, the study would propose and apply SOM and Fuzzy c-means.

IV. CONCLUSIONS

Although most of the studies had used various techniques, but were in weak visualization and not many on visualizing market segments as identified by several authors (e.g., Coene, 2011). Such methodology would be effective and easily transferable to market managers and also helps to visualize knowledge underlying the data set. Previous studies had emphasized the need for combining two clustering method (Vesanto & Alhoniemi, 2000).

REFERENCES

[1] Acito, F. & Khatri, V., 2014. Business analytics: Why now and what next? Business Horizons, 57(5), pp.565–570. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0007681314000871 [Accessed November 27, 2014]J. Breckling, Ed., *The Analysis of Directional Time Series: Applications to Wind Speed and Direction*, ser. Lecture Notes in Statistics. Berlin, Germany: Springer, 1989, vol. 61.

[2] Bailey, C., Baines, P.R., Wilson, H. & Clark, M., 2009. Segmentation and customer insight in contemporary services marketing practice: why grouping customers is no longer enough, J. Marketing Management, 25(3-4), pp. 227-252.

[3] Bedeley, R.T. & Iyer, L.S., 2014. Big data opportunities and challenges: the case of banking industry. In Proceedings of the Southern Association for Information Systems Conference. Macon, GA, USA, pp. 1–6.

[4] Bezdek, J. C., 1981. Pattern recognition with fuzzy objective function algorithms. New York: Plenum Press.
[5] Bulysheva, L. & Bulyshev, A., 2012. Segmentation Modeling Algorithm: A Novel Algorithm in Data Mining. Information Technology & Management, 13(4), pp.263–271.
[6] Cao, L., 2009. Introduction to domain driven data mining. L. Cao, P. S. Yu, C. Zhang, & H. Zhang (Eds.), Data mining for business applications, New York, NY: Springer, pp. 3-10.
[7] Chye, K. H., Chin, T. W. & Peng, G. C., 2004. Credit scoring using data mining techniques. Singapore Management Review, 26(2), pp. 25-47. Available at: ABI/INFORM Global database.
[8] Coenen, F., 2011. Data mining: past, present and future. The Knowledge Engineering Review, 26(01), pp.25–29. Available at: http://www.journals.cambridge.org/abstract_S0269888910000378 [Accessed January 14, 2015].
[9] Datta, R. P., 2008. Data mining applications and infrastructural issues: An Indian perspective. ICFAI Journal of Infrastructure, 6(3), pp. 42-50. Available at: Business Source Complete database.
[10] Farajian, M.A. & Mohammadi, S., 2010. Mining the Banking Customer Behavior Using Clustering and Association Rules Methods. International Journal of Industrial Engineering & Production Research, 21(4), pp.239–245.
[11] Gartner, 2014. IT Glossary. Available at: http://www.gartner.com/it-glossary/big-data/.
[12] Hastie, T., Tibshirani, R. & Friedman, J., 2009. The Elements of Statistical Learning, New York: Springer.
[13] Huang, J.J., Tzeng, G.H. & Ong, C.S., 2007. Marketing segmentation using support vector clustering. Expert Systems with Applications, 32, pp. 313–317.
[14] Huang, Y., K. & Kechadi, T., 2013. An effective hybrid learning system for telecommunication churn prediction. Expert Systems with Applications, 40, pp. 5635–5647.
[15] Hung, C. & Tsai, C.-F., 2008. Market segmentation based on hierarchical self-organizing map for markets of multimedia on demand. Expert Systems with Applications, 34(1), pp.780–787. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0957417406003290 [Accessed February 18, 2015].
[16] Hwang, S. & Thill, J. C., 2007. Using Fuzzy Clustering Methods for Delineating Urban Housing Submarkets. Proceedings of the 15th international symposium on advances in geographic information systems.
[17] IDC, 2013. IDC Big Data and Business Analytics Forum 2013. In Leveraging Data for Agile Business. Elsevier Inc. Available at: http://idc-cema.com/eng/events/50534-idc-big-data-and-business-analytics-forum-2013.
[18] Jones, L.E., 2014. Renewable Energy Integration. In Practical Management of Variability, Uncertainty, and Flexibility in Power Grids. Academic Press, p. 474. Available at: http://store.elsevier.com/Renewable-Energy-Integration/Lawrence-Jones/isbn-9780124079106/.
[19] Kim, K. & Ahn, H., 2008. A recommender system using GA K-means clustering in an online shopping market. Expert Systems with Applications, 34(2), pp. 1200–1209.
[20] Kohonen, T., Mäkisara, K. & Saramäki, T., 1984. Phonotopic maps – Insightful representation of phonological features for speech recognition. Proceedings of 7ICPR, international conference on pattern recognition, Los Alamitos, CA , IEEE Computer Society Press, , pp. 182–185.
[21] Koudehi, F.A. et al., 2014. A Hybrid Segmentation Approach for Customer Value. INTERDISCIPLINARY JOURNAL OF CONTEMPORARY RESEARCH IN BUSINESS, 6(6), pp.142–152. Available at: http://euabr.com/ijcrboct14/142-152oct14.pdf.
[22] Larose, D.T., 2005. Discovering Knowledge in Data, Wiley, New York.
[23] Leventhal, B., 2010. An introduction to data mining and other techniques for advanced analytics. Journal of Direct, Data and Digital Marketing Practice, 12(2), pp.137–153. Available at: http://www.palgrave-journals.com/doifinder/10.1057/dddmp.2010.35 [Accessed February 18, 2015].
[24] Li, J., Wang, K. & Xu, L., 2008. Chameleon based on clustering feature tree and its application in customer segmentation. Annals of Operations Research, 168(1), pp.225–245. Available at: http://link.springer.com/10.1007/s10479-008-0368-4 [Accessed February 18, 2015].
[25] Liu, S.S. & Chen, J., 2009. Using data mining to segment healthcare markets from patients' preference perspectives. International journal of health care quality assurance, 22(2), pp.117–34. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19536963 [Accessed February 18, 2015].
[26] Milligan, G. W. & Cooper, M. C., 1980. An examination of the effect of six types of error perturbation on fifteen clustering algorithms. Psychometrika, 45(3), pp. 159–179.
[27] Pechenizkiy, M., Puuronen, S. & Tsymbal, A., 2008. Towards more relevanceoriented data mining research. Intelligent Data Analysis, 12, pp.237–249.
[28] Raju, P.S., Bai, V.R. & Chaitanya, G.K., 2014. Data mining: Techniques for Enhancing Customer Relationship Management in Banking and Retail Industries. International Journal of Innovative Research in Computer and Communication Engineering, 2(1), pp.2650–2657.
[29] Smith, W., 1956. Product differentiation and market segmentation as alternative marketing strategies, Journal of Marketing, 21, pp. 3–8.
[30] Statista, 2014. The Statistics Portal. Available at: http://www.statista.com/statistics/289770/india-retail-e-commerce-sales/

[31] Sun, N. et al., 2014. iCARE: A framework for big data-based banking customer analytics. IBM Journal of Research and Development, 58(5/6), pp.4:1–4:9. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6964895 [Accessed February 18, 2015].
[32] Sun, S., 2009. An Analysis on the Conditions and Methods of Market Segmentation. International Journal of Business and Management, 4(2), pp. 63-70.
[33] Tan, P., Steinbach, M. & Kumar, V., 2006. Introduction to Data Mining, Pearson/Addison Wesley, Boston.
[34] Varun, K.M., Vishnu, C.M. & Madhavan, M., 2012. Segmenting the Banking Market Strategy by Clustering. International Journal of Computer Applications, 45, p.10.
[35] Vesanto J. & Alhoniemi , E., 2000. Clustering of the Self-Organizing Map, IEEE Transactions on Neural Networks, special issue on data mining, 11(3), pp. 586-600.
[36] Wang, G. & Wang, Y., 2009. 3DM: Domain-oriented data-driven data mining. Fundamenta Informaticae, 90, pp.395–426
[37] Witten, I. H. & Frank, E., 2000. Data mining: Practical machine learning tools and techniques. Morgan Kaufmann series in data management systems.
[38] Yankelovich, D. & Meer, D., 2006. Rediscovering market segmentation. Harvard Business Review, pp. 1-10.

