Atlantis Studies in Uncertainty Modelling, volume 1
11th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT 2019)

Outlier Detection in Location Based Systems by Using Fuzzy Clustering

Basar Oztaysi [a], Sezi Cevik Onar[b], and Cengiz Kahraman[c]
[a] Department of Industrial Engineering, Istanbul Technical University, oztaysib@itu.edu.tr
[b] Department of Industrial Engineering, Istanbul Technical University, cevikse@itu.edu.tr
[c] Department of Industrial Engineering, Istanbul Technical University, kahramanc@itu.edu.tr

Abstract

Customer segmentation has been one of most important decision in marketing. In general, demographics of customers, monetary value of customer transactions, types of product/service customers use are the sources of segmentation process. In recent years, new technology enabled new sources of data. On of these new data are the customer location data collected from location based systems (LBS). By using these location data an improved customer insight can be provided to the companies. Segmentation is an important tool for creating customer insight but anomalies in LBS data can prevent a well formed segmentation. In this paper we propose a novel approach to outlier detection in LBS data by using fuzzy c-means algorithm

Keywords: Location Based Systems, fuzzy clustering, segmentation, Fuzzy c-Means, outlier detection

1 Introduction

Segmentation is one of the essential strategies for targeted marketing and customer-centric decision making. As a new research paradigm, location-based services rely on geographical information determination of mobile users via new technologies such as global positioning systems (GPS) and Bluetooth (via Beacons). These services pinpoint users' real-time locations and provide valuable insights on customer visiting preferences. Additionally, location-based services offer an alternative gateway by providing the necessary infrastructure for sending special offers with regard to consumers' preferences or previous visits (Lee at al., 2015). To present customer centric opportunities for personalized marketing, effective segmentation conduces to competitive advantage by extracting new marketing directions with limited advertising budgets. Thus, location segmentation for specifying personalized properties from previously visited locations provides valuable insights of customer tendency to further potential visits (D'Urso et al., 2015). As a result of increasing penetration rate of telecommunication strategies, mobile location-based services can use mobile check-in data, geospatial data and customer comments for a specific place. To conduct location-based systems, Junglas and Watson (2008) mentioned that two major steps are essential: location detection aimed models to investigate "potential" future locations from geospatial data processing and prediction algorithms considering previous search patterns, previous location visits, online profiles and online comments. For targeted marketing management activities, location detection and the estimation of future visits are fundamental processes to capture customers' needs and dynamic changes in their interests. Because of location visiting, preferences can be determined from mobile devices, the context of the promotion or message can be defined consistently in order to keep communication channels alive. Therefore, companies are prone to send instant messages to their customers according to their recent locations.

2 Literature Review

Location-based systems are described as the service or application that combine the utilization of the geographic location of the consumer in order to provide a service or a marketing message (Mobile Marketing Association, 2011). In other words, location-based systems provide real time location data of consumers. For instance, location-based systems capture location data, ensure mobile connection to other mobile devices and send related content including promotions or attractive messages to mobile costumers when they are appeared in certain fields, such as shopping malls (Anagnostopoulos et al., 2015). Another aspect of location-based services (LBSs) is that they can be evaluated as a subset of web services that provide location-aware functions. The

utilization of such services focused on extracting knowledge from where the services are constructed.

Until this time, LBSs have been acquainted with a distributed mobile computing infrastructure where the geographic locations of users are specifically used for application-related optimization. From a different viewpoint, sensor data hides previous interactions between users and locations that can be available using an Internet connection or Bluetooth technology. For this reason, large-scale retail companies, such as Wal-Mart, have commonly transferred their retail activities to LBSs for raising brand awareness (Zou and Huang, 2015). Other examples of LBSs include location-based advertising, coordination of traffic flow, natural disaster search and rescue, tourist route recommender systems, nearest available park and ride applications (Cheverst et al. 2000).

The literature review indicates that LBSs can be defined as an emerging research topic in terms of indoor and outdoor navigation including location-based advertising and location-based mobile advertising, mobile shopping (Yang et al., 2008), travel recommendation systems (Sun et al. 2013,Versichele et al., 2014), recognition of places for future predictions (Vu et al., 2009), group buying (Li et al., 2014) and social media based recommender systems . In addition to these studies, user satisfaction of LBS systems was investigated by Kuo et al. (2009), and a novel concept of a mobile ad hoc network was proposed by Ramya and Prasad Babu (2014) using circular data aggregation technique. Also the area has been investigated with respect to technology selection and customer behaviour (Budak et al. 2016, Oztaysi et al. 2016, Dogan and Oztaysi, 2018a, Dogan and Oztaysi, 2018b, Dogan et al., 2019a, Dogan et al. 2019b). From all these studies, automatic location-based applications have been used widespread that ease traceability of individuals' physical moving in different indoor shopping fields. For this reason, personal positioning and segmentation algorithms acquire critical roles in detecting individuals' positions and locations to make a practical analysis on the determination of group behavior of customers, shopping and visiting tendencies. In other words, location-aware systems provide following up customer's shopping needs with location-dependent offers and promotions to cope with competition.

Although remarkable advantages have been realized, location-based systems have some drawbacks which are emerged in practice. Two major problems are aroused as privacy issues and disruption of messages: customers do not prefer to send their location data to service providers and generally, do not dispose to gather instant messages especially when they are not available in certain times. Additionally, the irrelevance of the message is another problem that causes incorrect predictions related to customers. In our case, retailer clustering necessitates collecting data from various sources to describe the relationship between customers, locations (shopping malls), retailers and their relationships with each other. In other words, clustering procedure provides desired services to satisfy customer needs considering customer tendencies (Gavalas et al., 2014). In this respect, the main research question of this study is stating personalized sales suggestions with respect to previously visited location value and also the determination of alternative retailers according to the similarities assigned by common location characteristics. For this purpose, location clustering can be applied for grouping retailers and alternative retailer suggestions can be conducted using the similarities appeared from the location clusters and product segments. Thus, a literature review is given in Section 3 for describing previous studies on location clustering.

3 Methodology

a. Fuzzy c-Means

Clustering procedure can be explained as the method of splitting data into subgroups which are named as "clusters". A considerable amount of techniques are available for clustering and purpose of these techniques is grouping input data/information to gather corresponding objects in a cluster, and distributing different objects to alternative clusters (Han and Kamber, 2011). Crisp clustering algorithms match input data to one specific cluster. On the other hand, fuzzy clustering algorithms assign an object to diversified clusters simultaneously with a membership degree (Oztaysi and Isik, 2014). From this point of view, one of the most applied clustering algorithm for fuzzy clustering is FcM clustering that clusters should be determined in advance (Chen et al., 2014). Note that the input is a set of data or objects, each of which consists of different attributes or features. After that, cluster analysis can be performed using similarity or dissimilarity measures which are extracted from distance measurement such as Euclidean distance to define the similarity of given observations. A fuzzy partition matrix for extracting clusters is defined from

Ruspini (1970) with the conditions given in the following:

$\mu_{ik} \in [0,1], 1 \leq i \leq c,\ 1 \leq k \leq N,$ (Eq. 1a)

$\sum_{i=1}^{c} \mu_{ik} = 1,\quad 1 \leq k \leq N,$ (Eq. 1b)

$0 < \sum_{k=1}^{N} \mu_{ik} < N, 1 \leq i \leq c$ (Eq. 1c)

Equation (1b) defines the sum of each cluster should be equal to 1, and membership degree should be represented with an interval [0, 1]. The main goal of FcM clustering relies on the minimization of the corresponding objective function which comprised of a nonlinear optimization problem:

$J(Z,U,V) = \sum_{i=1}^{c}\sum_{j=1}^{N}(\mu_{ij})^{m}\|z_j - v_i\|^2$ (Eq.2)

where Z is the data set needed to be partitioned, U represents the fuzzy partition matrix, V is the cluster centers' vector. As seen from the given formula, N represents the number of observations, μ denotes the related membership value, c is the number of appeared clusters and m is the parameter called fuzzifier that identifies the fuzziness degree of the final clusters and fuzzifier parameter can get values greater than 1. Besides that, $z_j - v_i$ denotes the distance from observation j to the center of cluster i. Note that the first step of FcM clustering algorithm contains gathering fuzzy partition matrix as U=[uij] matrix and U(0) denotes the fuzzy partition matrix appeared in the first phase. After that, center vectors V(k)=[vi] are calculated with U(k) by considering the center vector formula.

$$v_i = \frac{\sum_{i=1}^{N}\mu_{ij}^{m}.z_j}{\sum_{i=1}^{N}\mu_{ij}^{m}} \quad \text{(Eq.3)}$$

Again, fuzzy partition matrix in kth step (U(k)) is updated for the further step as

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{\|z_j - v_i\|}{\|z_j - v_k\|}\right)^{\frac{2}{m-1}}} \quad \text{(Eq.4)}$$

with a considerable computational error δ.

As seen from literature, fuzzy clustering is widely adapted to sentence similarity detection as seen in Devi and Gandhi (2015)'s study that "Page Rank" algorithm is combined with Expectation-Maximization (EM) framework. In addition to that, textual document archive clustering from Torra et al. (2005) provides an extension to fuzzy clustering techniques using Gambal system based visualization of documents. Real-time flood forecasting study from Ren et al.(2010) utilized a fuzzy clustering model with a back-propagation (BP) neural network training model. Sowmya and Rani (2011) evaluated image segmentation using FcM algorithm combined with possibilistic fuzzy c means (PFcM) algorithm and competitive neural network (CNN). Most of these studies used crisp data for converting input data to mutually exclusive subsets. On the other hand, in our case, the nature of the given problem contains imprecise data and conflicting prices from diverse retailers

b. Steps of the methodology

Outliers or anomaly data may negatively effect the results of clustering. An outlier is defined as a rare item, event or observation which raise suspicions by differing significantly from the majority of the data. In this paper we aim to use results of fuzzy c-means algorithms as a source for identifying the outliers. As mentioned in the previous subsection, fuzzy c-means method provides membership values for each data point to all identified clusters. Later, these membership values are later used to define the data points main cluster. However, these membership values are highly effected by the cluster formation which is a direct result of the number of clusters. In other terms, the membership values may significantly change when the number of clusters change. In this paper we define outlier as a data point which do not show a high membership value to any clusters for different number of cluster numbers. To this end the steps of the methodology can be given as in the following.

Step 1: Preprocess LBS data to prepare the visitor-locations matrix (VLM) which show the amount of time each visitor spend in each location.

Step 2: Define maximum cluster number (C) to be tested and the fuzzifier parmeter.

Step 3: For c=2 to C execute the fuzzy c-means algorithm using VLM.

Step 4: Check the cluster sizes for their validity. If the size of a cluster is less than a treshold value (such as 0.1%) then eliminate this cluster. In other case, a few outliers may form a cluster and their high membership values prevent them from being detected as an outlier.

Step 5: For each different cluster number, determine the membership values of each visitor, and select the highest membership value for each one.

Step 6: Calculate the dominancy ratio (DR) for visitor k when the number of clusters is selected as c. DR of a visitor for a given c shows the relative dominancy of maximum membership to a cluster with respect to other cluster memberships.

$$DR_k^c = \frac{Max\ Memb._k - MPmM^c}{MPmM^c} \qquad \text{(Eq.5)}$$

where $MPmM^c$ shows the minimum possible maximum membership value when the number of cluster is selected as c. $MPmM^c$ value can be calculated by Eq. 6.

$$MPmM^c = \frac{1}{c} \qquad \text{(Eq.6)}$$

Step 7: For each data point find the maximum dominancy ratio (DR_k) by comparing all DR_k^c value.

Step 8: The visitors with DR_k values lower than a threshold value are selected as the outliers.

4. Real World Case Study

4.a Business Problem

With the spread of the internet concept of objects, Beacon is a generic name given to small Bluetooth radio transmitters that offer personalized experiences. Beacon technology provides location information using Bluetooth low energy (BLE) or the trademarked name, Bluetooth R Smart. Each iBeacon device has a coverage zone with 50 meters, and the interaction distance among devices can be fine-tuned in the zone. For that reason, one smart device can receive signals from more than one beacon devices. In this case, proximity distances between the device and beacon are calculated and classified as immediate, near, far.

For the study, we collected data via iBeacon devices. Approximately 300 iBeacon devices are located in the shopping mall for stores by Blesh Company, one of the earliest developers of data distribution technology in indoor locations in Turkey. Figure 1 depicts the architecture of Beacon technology.

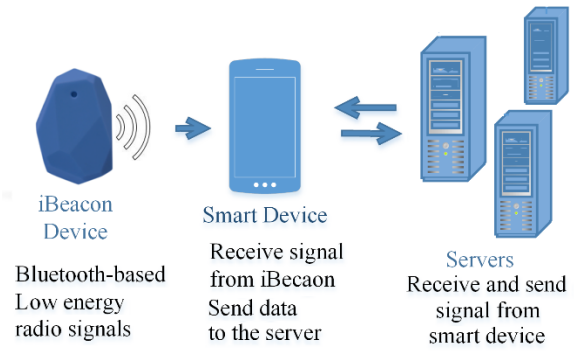

Figure 1: The architecture of Beacon technology

The data used in this paper is collected in one of the major shopping malls in Istanbul. It was opened in 2005 as one of the most prominent and leading shopping and entertainment centers in Europe and the world. The shopping mall has six floors with a total of approximately 300 shops and a parking garage with 2500 cars capacity. It is located at a strategic point in Turkey's metropolis Istanbul. The shopping mall continues to serve domestic and foreign visitors with 293 stores. Before preprocessing steps, collected data had 779877 stores which belong to 12000 unique customers. We grouped the stores concerning their products or services. For example, Store A, Store B and Store C is grouped into Clothing shop group, if they sell mainly clothes. After some other data preparations, we used 9683 customer data that we can determine genders.

4.b Data Structure

iBeacon ILS can produce CSV files that have in each line a positioning value for each event. Table 1 shows a sample part of the localization events provided by Beacons. In each line, the following data are available: ID: For each captured data, a unique number is defined to use different analysis. ID column is not used in our study. Dongle Columns: Because, in some cases, iBeacon devices located near each other, one customer can be seen in different three locations. According to the proximity of the customer to the stores, iBeacon devices gather three different position data. Dongle_1 shows the nearest store that customer data captured, on the other hand, Dongle_3 represents the farthest store location. In the study, we ignore Dongle_2 and Dongle_3 localization data. Timestamp: This is a date and time value that represents the moment at which data is captured by the iBeacon device for the related store. Timestamp format is hh:mm:ss. SubscriberID: This number shows the customer identification number associated with the mobile.

ID	Dongle_1	Dongle_2	Dongle_3	Timestamp	SubscriberID
1028333326	121527			11:14:42	17399446
1028334382	121498			11:16:48	39081930
1028334406	121498	121404		11:16:50	39081930
1028334421	121498			11:16:53	39081930
1028492822	121436	121510	121446	11:23:47	29078632
1028492925	121510	121372		11:23:59	29078632
1028492939	121436	121510	121446	11:24:01	29078632
1028495185	121446	121436	121510	11:28:23	29078632

Table 1: Raw data format used in the paper

5. Application

In order to find the outliers in the data set first the raw data is processed. To this end, all locations are associated with a category and the time duration each customer spend on a category is calculated. As a result a matrix containing 9622 rows (the customers) and 30 columns (the categories) is formed as given in Table 2.

CustomerID	Common area	Clothing	Supermarket	Catering
37600854	0	6	0	9
36997697	0	9	1	3
34612045	5	0	2	5
38585080	3	3	1	9
39329211	2	1	0	2

Table 2: Preprocessed data

The steps of the methodology are implemented by using RStudio 1.1.442. The size of the clusters formed are given in Table 3. For c=7 and c=10 there are clusters with a few members. These clusters are eliminated from the calculations.

c=2	5807, 3815	c=3	2965, 3373, 3284
c=4	2989, 979, 2535, 3119	c=5	2810, 2201, 1310, 2449, 852
c=6	1360, 919, 1556, 2417, 2551, 819	c=7	1821, 804, 1, 1433, 2358, 800, 2405
c=8	2195, 2184, 796, 195, 1050, 78, 1728, 1396	c=9	1488, 347, 2078, 1195, 323, 791, 1406, 49, 1945
c=10	1836, 2007, 33, 475, 790, 1800, 1434, 4, 1243		

Table 3: Cluster sizes for different c values.

The results of the methodology provides DR values for each visitor. The six lowest DR values are given in Table 4.

Customer Id	DR
28476687	0.3390
15886631	0.3391
24329793	0.3391
31903348	0.3391
36789635	0.3391
41598138	0.3391

Table 4: The values with the lowest DR'S

The transactions of the visitors listed in Table 4 is given in Table 5. The values show that the listed customers are displayed on a single category.

Customer Id	…	Computer	Wedding Gown	…
28476687	0	0	12	0
15886631	0	2	0	0
24329793	0	3	0	0
31903348	0	1	0	0
36789635	0	2	0	0
41598138	0	19	0	0

Table 5: the raw data for the outliers

6. Conclusion

In marketing and communication domain customer segmentation is a vital tool for customer management. Traditional methods for customer segmentation use customer buying details, demographics or qualitative data as the source of segmentation. On the other hand, Beacon technology provides data about visitor indoor locations which can provide insight about customer behaviors or buying decisions. In this study we propose a model for automatic detection of outliers. The steps of the methodology is applied to R software and can be used as a part of automatic decision support system.

The data of the case study is obtained from a shopping mall in Istanbul. As we work with a real world dataset, the only way to check the validity of the proposed method is to discuss the results with experts. As a result of our discussion, the experts claim that the outliers defined in the paper are the employees of the shops. Since there are no beacons implemented inside the shops, employees are only displayed in front of the shop they work for.

For further studies, the result of this study can be compared with previously developed outlier detection methods such as one-class SVMs, support vector data description, Principle Component Analysis (PCA) based approached, and Gustafson-Kessel clustering. Another branch of research can focus on optimizing the parameters required for the study. Besides, an analysis can be designed taking into account the seasonality of the visitors.

References

Anagnostopoulos C, Hadjiefthymiades S, Kolomvatsos, K. (2015) Time-optimized user grouping in Location Based Services, Computer Networks 81: 220–244.

Budak, A., Ustundag, A., Oztaysi, B., Cevikcan, E., 2016, A Multi-Criteria Intuitionistic Fuzzy Group

Decision Making Model For Real Time Location System Integration: An Application From Healthcare System, Proceedings of the 12. FLINS Conference, 580-587.

Chen N., Xu ZS, Xia MM. (2014) Hierarchical Hesitant Fuzzy K-Means Clustering Algorithm. Applied Mathematics-A Journal of Chinese Universities 29: 1–17.

Cheverst K, Davies N, Mitchell K, Friday A, Efstratiou C (2000) Developing a context-aware electronic tourist guide: some issues and experiences, CHI '00 Proceedings of the SIGCHI conference on Human Factors in Computing Systems: 17-24 , The Hague, The Netherlands — April 01 - 06, 2000.

D'Urso P, Disegna M, Massari R, Prayag G (2015) Bagged fuzzy clustering for fuzzy data: An application to a tourism market, Knowledge-Based Systems, Volume 73: 335-346

Devi MU, Gandhi GM (2015) An Enhanced Fuzzy Clustering and Expectation Maximization Framework based Matching Semantically Similar Sentences, Procedia Computer Science, Volume 57: 1149-1159

Dogan O., Gurcan O.F., Oztaysi B., Gokdere U. (2019b) Analysis of Frequent Visitor Patterns in a Shopping Mall, Industrial Engineering in the Big Data Era, 217-227.

Dogan O. and Oztaysi B. (2018a). An Application of Process Mining and Association Rule Mining for Indoor Customer Data, 29th European Conference on Operational Research (EURO 2018), 08-11 July, Valencia, Spain.

Dogan O., Bayo-Monton JL , Fernandez-Llatas C., Oztaysi B. (2019a), Analyzing of Gender Behaviors from Paths Using Process Mining: A Shopping Mall Application, Sensors 19 (3), 557.

Dogan O., Oztaysi B. (2018b) In-store behavioral analytics technology selection using fuzzy decision making, Journal of Enterprise Information Management 31 (4), 612-630.

Gavalas D, Konstantopoulos C., Mastakas K., Pantziou G. (2014) Mobile recommender systems in tourism, Journal of Network and Computer Applications 39:319–333

Han J, Kamber M. (2001) Data Mining Concepts and Techniques, Morgan Kauffman Publishers, pp. 5-33.

Junglas, IA. and Watson, RT. (2008) Location-based services, Communications of the ACM, 51 (3): 65–69.

Kuo MH, Chen LC, Liang CW (2009) Building and evaluating a location-based service recommendation system with a preference adjustment mechanism, Expert Systems with Applications 36: 3543–3554.

Lee, S, Kim, KJ., Sundar SS. (2015) Customization in location-based advertising: Effects of tailoring source, locational congruity, and product involvement on ad attitudes, Computers in Human Behavior 51:336–343.

Li YM, Chou CL, Lin LF. (2014) A social recommender mechanism for location-based group commerce, Information Sciences 274: 125–142

Mobile Marketing Association, Mobile Location Based Services Marketing Whitepaper, Tech. Rep, Mobile Marketing Association, 2011.

Oztaysi B, Gokdere U, Simsek EN, Oner SC (2016). A Novel Approach to Segmentation Using Customer Locations Data and Intelligent Techniques Handbook of Research on Intelligent Techniques and Modeling Applications in Marketing Analytics, Eds. Kumar, Anil, Dash, Manoj Kumar, Trivedi, Shrawan Kumar: 21-39.

Oztaysi B. and Isik M. (2014). Supplier Evaluation Using Fuzzy clustering, in Eds. Kahraman C., and Oztaysi B., Supply Chain Management Under Fuzziness: Recent Developments and Techniques: 61-80, Springer.

Ramya AR, Prasad Babu BR (2014) A novel concept of MANET architecture for location based service using circular data aggregation technique. Int J Innov Res Dev;3(1):252–8.

Ren M, Wang B, Liang Q, Fu G (2010) Classified real-time flood forecasting by coupling fuzzy clustering and neural network, International Journal of Sediment Research, Volume 25, Issue 2: 134-148

Ruspini EH (1970) Numerical methods for fuzzy clustering. Information Science, 2: 319–350.

Sowmya B., Rani BS (2011) Colour image segmentation using fuzzy clustering techniques and competitive neural network, Applied Soft Computing, Volume 11, Issue 3: 3170-3178

Sun Y, Fan H, Bakillah, M., Zipf A. (2013) Road-based travel recommendation using geo-tagged

images, Computers, Environment and Urban Systems, http://dx.doi.org/10.1016/j.compenvurbsys.2013.07.006

Torra V, Miyamoto S, Lanau S (2005) Exploration of textual document archives using a fuzzy hierarchical clustering algorithm in the GAMBAL system, Information Processing & Management, Volume 41, Issue 3: 587-598

Versichele M, De Groote L, Bouuaert MC, Neutens T, Moerman I, Van de Weghe N. (2014) Pattern mining in tourist attraction visits through association rule learning on Bluetooth tracking data: A case study of Ghent, Belgium, Tourism Management 44: 67-81.

Vu THN, Ryu KH, Park N (2009) A method for predicting future location of mobile user for location-based services system, Computers & Industrial Engineering 57: 91–105.

Yang WS, Cheng HC, Dia, JB (2008). A location-aware recommender system for mobile shopping environments, Expert Systems with Applications 34: 437–445.

Zou X and Huang KW (2015) Leveraging location-based services for couponing and infomediation, Decision Support Systems 78:93–103.