K-Means vs. Fuzzy C-Means: A Comparative Analysis of Two Popular Clustering Techniques on the Featured Mobile Applications Benchmark

Tuğrul Cabir Hakyemez[1*], Aysun Bozanta[2], Mustafa Coşkun[1]

1 Sakarya University, **2** Boğaziçi University, * Corresponding author, thakyemez@sakarya.edu.tr

Abstract

Over the past few years, mobile applications have become an indispensable part of our daily lives. Noticing this ever-growing market, all those who are engaged in developing attractive applications should make informed decisions along the development process through sophisticated methods in order to survive in the market. As one of these methods, clustering is well suited for identifying the hidden groups existing in huge datasets. In this paper, the Mobile App dataset that contains features of 7196 available applications was clustered using two popular clustering algorithms, namely as k-means and fuzzy-c means. After conducting necessary preprocessing steps (e.g. outlier removal, standardization), these algorithms were run with different parameters in an experimental manner to reach optimal values and their performances were compared based on cluster quality (internal validity), number of iterations and elapsed time. The main findings suggested that fuzzy c- means produced higher quality clusters whereas k-means algorithm converged faster than its counterpart. In the last section, conclusions were made and future studies were discussed.

Keywords: Mobile application, Application mining, Clustering, Fuzzy c-means.

Citation: Hakyemez, T. C., Bozanta, A., Coşkun, M. (2018, October) *K-Means vs. Fuzzy C-Means: A Comparative Analysis of Two Popular Clustering Techniques on the Featured Mobile Applications Benchmark*. Paper presented at the Fifth International Management Information Systems Conference.

Editor: H. Kemal İlter, Ankara Yıldırım Beyazıt University, Turkey

Received: August 19, 2018, **Accepted:** October 18, 2018, **Published:** November 10, 2018

K-Means vs. Fuzzy C-Means: A Comparative Analysis of Two Popular Clustering Techniques on the Featured Mobile Applications Benchmark

[1]Tuğrul Cabir HAKYEMEZ [2]Aysun BOZANTA [3]Mustafa COŞKUN

Abstract

Over the past few years, mobile applications have become an indispensable part of our daily lives. Noticing this ever-growing market, all those who are engaged in developing attractive applications should make informed decisions along the development process through sophisticated methods in order to survive in the market. As one of these methods, clustering is well suited for identifying the hidden groups existing in huge datasets. In this paper, the Mobile App dataset that contains features of 7196 available applications was clustered using two popular clustering algorithms, namely as k-means and fuzzy-c means. After conducting necessary preprocessing steps (e.g. outlier removal, standardization), these algorithms were run with different parameters in an experimental manner to reach optimal values and their performances were compared based on cluster quality (internal validity), number of iterations and elapsed time. The main findings suggested that fuzzy c- means produced higher quality clusters whereas k-means algorithm converged faster than its counterpart. In the last section, conclusions were made and future studies were discussed.

Keywords

Mobile Application, Application Mining, Clustering, Fuzzy c- means

Introduction

[1] Sakarya University, Management Information Systems Department, Sakarya, Turkey (Corresponding Author)
[2] Bogazici University, Management Information Systems Department, Istanbul, Turkey
[3] National Education Department, İzmir, Turkey

Mobile applications have become an inevitable part of our lives such as they are benefitted while keeping an agenda, deciding on which restaurant to go, sharing our experiences via photos and short expressions, listening a music, making bank transactions, communicating, planning a vacation, and even making yoga or tracking a diet list.

The number of mobile application downloads is worldwide in 2017 is 178.1 bn (Mobile App Usage - Statistics & Facts, 2018). This intensive demand prompts the developers and the number of mobile application in the market has been rapidly increasing. There are 3.8 million mobile application in Google Play Store which is the largest application store and there are 2 million mobile application in Apple App Store which is the second-largest store as of the first quarter of 2018 (Number of apps available in leading app stores as of 1st quarter 2018, 2018). (Number of available apps in the Amazon Appstore from 2nd quarter 2015 to 1st quarter 2018, 2018). Table below shows the number of available mobile application in Google Play Store from December 2009 to June 2018 (Number of available applications in the Google Play Store from December 2009 to June 2018, 2018).

INSERT FIGURE 1 HERE

The most popular mobile application categories in Google Play Store are education, entertainment and lifestyle (Most popular Google Play categories, 2018) while gaming, business, and education in Apple Appstore (Most popular Apple App Store categories in May 2018, by share of available apps, 2018) as of the second quarter of 2018. Presented statistics shows that the mobile application market is very huge and it's growing. Therefore, gaining a market share from this market is very profitable for both investors and the developers. Each user chooses an application according to their own criteria. At that point, it is important to understand the preferences of the mobile application users. There can be various parameters such as rating, comments, price, and category of a mobile application that affect the decision of the users. Therefore it is important to analyze the available data of existing mobile applications.

In this paper, the data consisting of the characteristics of the mobile application data is analyzed with the fuzzy c means and k-means clustering techniques. For each clustering technique, experiments are conducted according to the various parameters (e.g. number of clusters) and the results are presented. These two techniques are also compared in terms of eligible performance criteria. By doing so, it is aimed at providing a manual for the developers and entrepreneurs in the field to effectively and efficiently develop viable mobile applications.

This paper is organized as follows: Section 2 provides a detailed analysis of the current literature of mobile application landscape. Then Section 3 explains the methodology. Section 4 presents the results of the analyses. Finally, Section 5 concludes and provides an outlook for the mobile application developers and suggests future research opportunities.

Literature Review

There are a variety of research on mobile applications concept in the literature examining the design, development, test, marketing, usability, and security of mobile applications. The recent research on the factors affecting the preferences of mobile application users are presented in this part.

In study of Naaman & Kaplun (2008), authors examine usage data of Zurfer which is photo sharing application and also conduct extensive user interviews in order to figure out usage trends and patterns which can help for mobile application developers. Some of the necessary characteristics of mobile applications in order to be successful as follows; having wide variety of options and content for "killing time", easily browsing and discovering the content, speed, readiness and responsiveness, social content (but make it easy for users to follow (or prioritize) their closest or favorite contacts), location-based content and finally virtual social interaction.

In study of Wasserman (2010), distinctive features of mobile applications which are not commonly included in traditional software applications are ordered as follows; potential interactions with other applications, sensor handling, native and hybrid (mobile web) Applications, families of hardware and software platforms, security, attractive user interfaces, and more power consumption.

Penttinen, Rossi & Tuunainen (2010) examine the costumers' values, needs, and objectives related to mobile games. In that paper satisfaction of quality expectations, gaming experience, ease of setup, social aspects of games are found as four fundamental objectives of games. In order to achieve these fundamental objectives, games should have these properties; audiovisual effects, ease of shopping and services, customer support, comprehensive product information, trust and triability.

Yang (2013) uses the Theory of Planned Behavior, the Technology Acceptance Model and the Uses and Gratification Theory in order to understand mobile application attitudes, intend and use of young Americans. Perceived enjoyment, usefulness, ease of use, and subjective norm are found as predictors of their mobile application attitudes. Perceived behavioral control, usefulness and mobile internet use have effect on their intent to use mobile applications.

Fuzzy analytical hierarchy process was used to predict the most effective factors for customers to use mobile services (Shieh et al., 2013). These are found as security and privacy, signal quality, comprehensive customer service, handset prices and transmissions fee, advertising, network coverage, transmission speed, service accessibility, real-timeliness, and usefulness with respect to their weights.

It was claimed that the adoption of mobile applications will be more strongly influenced by a consumer's social contacts (friends, compared to family members) (Taylor et al., 2011). This result is supported in another study that human connection and social utility to be more important than entertainment in creating task performance, easiness, and use intention (Kang, 2014).

The impacts of visualizing trust information on mobile application usage were examined in China and Finland (Yan et al., 2010). In both countries participants indicated that displaying an application's reputation value and/or a user's individual trust value could assist customers in the usage of mobile applications. In the study of Ickin et al. (2012), factors that affect the quality of experience of customers about commonly used mobile applications are examined. Application interface design, application performance, application cost, user routines and user lifestyles are found as factors that affect the quality of users' experience of mobile applications.

In the study of Biel et al. (2010), in order to evaluate mobile application usability, "Software ArchitecTure analysis of Usability Requirements realizatioN" (SATURN) method is developed. In this method mobile application usability is evaluated based on three aspects mobile environment, mobile user and mobile task. The components of mobile environment are location orientation, physical properties, social conditions, connectivity and collaboration. The components of mobile user aspect are attention span, motoric capabilities, mental capabilities, preferred location, user type, multimedia usage, and application usage. The components of mobile task are functionality, work-flow, duration, complexity, type, and dependencies.

Lastly in the study of Böhmer et al. (2011), AppSensor, which is a kind of virtual sensor is used for collecting data from 4,100 users of Android-powered mobile devices. The authors determine that users spend almost an hour a day with their phones but the usage time of any app is not more than a minute. Also people use different apps in the different times of the day. For example, they show that news applications are used mostly in mornings, games are used mostly in nights and chat application are used almost all day long.

The studies investigating the factors affecting the use of mobile applications mostly depend on the survey data. However, in this study, Mobile App dataset including features of 7196 available mobile applications was clustered using two popular clustering algorithms; k-means and fuzzy-c means.

There are many studies in the literature using k-means clustering algorithm for customer segmentation (Kim & Ahn, 2008; Hruschka & Natter, 1999), grouping students according to different characteristics (Oyelade et al., 2010; Baradwaj & Pal, 2012), image segmentation (Ng et al., 2006; Dhanachandra et al., 2015). Also, Fuzzy-c means clustering was used for image segmentation (Zhang & Chen , 2004; Cai et al., 2007). Moreover, there are studies comparing the k-means and fuzzy-c means algorithms with various data sets (Mingoti & Lima, 2006; Rong, 2011; Ghosh & Dubey, 2013).

As far as the knowledge of the authors, the analysis of mobile application data set with the clustering algorithms – k-means and fuzzy-c means has not existed in the literature. Therefore, it is believed that the result of this study will fill this gap in the literature and also provide a comprehensive insight to both mobile application developers and marketers.

Methodology

Data

The Mobile App Store dataset (retrieved from URL: *https://www.kaggle.com/ramamet4/app-store-apple-data-set-10k-apps*) contains static (prime genre, Application name etc.) and dynamic (user rating counts, user rating value etc.) features of 7196 mobile applications which are currently available in App Store. The detailed description of the variables included in the analysis is given below.

INSERT TABLE 1 HERE

The data is analyzed through a package program for statistical computing called R (ver. 3.4.4).

Preprocessing

The success of data mining techniques heavily relies on the quality of the data. Considering the huge volume of data, it can be argued that these techniques are susceptible to noise, missing values and inconsistency existing in various data sources. (Han, Pei, & Kamber, 2011). In the current study,therefore, necessary preprocessing procedures were conducted prior to the clustering analysis.

Firstly, k-means clustering and fuzzy c-means clustering algorithms are not robust to noisy data, which means that the performance is highly influenced by the outliers. Thus outlier analysis was conducted to eliminate the negative effects on the performance of the selected clustering algorithms. The data points that remain outside the underlying distribution with a probability of 0.99 were excluded from the analysis. The resulting dataset consists of 6843 mobile applications.

In the next step, the continuous variables that are included in the clustering analysis were standardized to overcome the unit and range differences between the variables. For example, total rating counts and application size are the variables that are measured by completely different units. The former is measured in integer numbers and the latter is measured in bytes.

k-means Clustering

First coined by (MacQueen, 1967), k-means algorithm attempts to iteratively reach the optimal k partition where the squared error between the cluster center and the data points is minimized (Jain, 2010). The main steps of the algorithm are as follows (Wagstaff et al., 2001)

1. Each data point d_i is assigned to its closest cluster center.
2. Each cluster center Cj is updated to be the mean of its constituent instances

The algorithm is quite simple and useful in many situations. However, the algorithm is not without its drawbacks. Nonrobustness to outliers and high level of sensitivity to initial k value are the main problems with this method.

To alleviate this problem, the performances of competing models are measured through an internal validity (quality) index. In current research, Xie and Beni(XB) index were employed to benchmark the algorithms which mainly concerns with the compactness and separation of the clusters (Liu et al., 2010) . Originally developed for evaluating the performances of fuzzy clustering, XB is also applicable for crisp clustering algorithms (Desgraupes, 2011) and is leveraged to compare the performances of both crisp (Dhanalakshmi & Inbarani, 2012). and fuzzy algorithms(Maulik & Bandyopadhyay, 2002; Kim, Lee, & Lee, 2004), hence it was selected as the index that is employed to evaluate the performances of the different models. Developed by Xie & Beni, (1991); XB is explained by a ratio of the total variation of the partition and centroids (U,V) and the separation of the centroids vectors and formulated as follows:

$$u_{XB}(U,V;X) = \frac{\sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik})^{m2}\,(d_{ik})^{2}}{N\left(min_{j\neq l}\{\,|\,|v_j - v_l|\,|^{2}\}\right)} \quad (1)$$

The minimum value of XB is considered to provide a best partition. In the current paper, the k means algorithms were run for different k values, k ϵ {2,3,4,....10}, and the model that produced the smallest XB value is accepted as the prevailing model.

Fuzzy c-means clustering

First developed by (Dunn, 1973) and further modified by (Bezdek, 1981), Fuzzy c-means clustering (FCM) is an unsupervised algorithm that partitions the data points based on a fuzzy rather than crisp membership while minimizing the objective function that is formulated as:

$$J_m(U,V;X) = \sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik})^{m}\,(d_{ik})^{2} \quad (1)$$

where;

J_m (U,V;X)= weighted sum of squared errors within groups

U= fuzzy membership matrix

V= vector of cluster centers

U_{ik} = membership value of x_k to the ith cluster

n= number of data points

c= number of fuzzy clusters

m = fuzziness index that refers to degree to which overlap is allowed between the clusters

d_{ik}= distance between data point x_i to cluster center v_k metric (e.g. Manhattan, Euclidian)

The minimization of the objective function can be achieved iteratively: until the difference between membership values becomes less than the specified threshold, ε (Cannon, Dave, & Bezdek, 1986).

As one can infer from the objective function, c and m values are determined arbitrarily and the initial values assigned to these parameters undoubtedly affects the performance of the FCM algorithm. In current study, an experimental approach was adopted to reach optimal values for c and m parameters. The fuzzy c-means algorithms were applied with different parameter configurations. The set of c value, $c \in \{2,3,4,....10\}$, is and the corresponding m (fuzziness index) values that start with 1.05 and increase by 0.5 to 4.0 - the upper level of the optimal area for m, which is empirically validated by (Wu, 2012). More clearly, the models were created for each pair of (c, m). For example, fcl_{1a} represents the model where c is set to 2 and m to 1.05.

Upon running 63 separate models with different parameters, XB values were calculated for each of them and the one that produced the minimum XB is selected as the best model.

Comparative Analysis

Noting the underlying differences of the fuzzy and crisp clustering, some researchers have investigated the efficiency measures such as number of iterations and time complexity (Ghosh & Dubey, 2013; Cebeci & Yildiz, 2015). Yet, as stated previously, some validity indexes like XB allows for benchmarking these algorithms. In this paper, the best k-means and fuzzy c- means algorithms are compared based on these criteria.

Results

Descriptive statistics

Due to standardized values, an appropriate interpretation of cluster centers requires a clear understanding of the dispersion of the individual variables that are included in the clustering analysis by its own units, thus the means and the standard deviations should be provided.

INSERT TABLE 2 HERE

k-means clustering

k-means clustering algorithm was run for 9 different k values. The XB values for the models are: 218.364 for k=2, 85.092 for k=3, 99.244 for k=4, 163.987 for k=5, 145.450 for k=6, 164.631 for k=7, 152.219 for k=8, 133.498 for k=9 and 363.899 for k=10 respectively. Considering the minimum value supports the best partition, it is obvious that the algorithm with k= 3 outperformed its competitors. So, it can be argued that the optimum value for k is 3 for the current dataset. Resulting cluster centers are displayed below:

INSERT TABLE 3 HERE

In order to properly interpret the figures from the table, one should bear in mind that these are the standardized values. That is to say, the values in each cell represent the difference from the overall variable (column) mean in terms of standard deviation.

Firstly, Cluster 1 contains 4803 applications, which makes it the most crowded group in mobile application market (approximately 70% of the available applications). The most salient characteristic of this cluster is observed at the price and size dimensions of the applications. Another unique point of the cluster is the ratio between rating and rating count, which implies that these algorithms are favored but not frequently downloaded by the users. So, this cluster can be labeled as “paid functional applications”.

The 1703 mobile applications grouped under Cluster 2 are below the overall mean in almost all variables. The most striking difference occurred in the overall (1.35 standard deviation below) and latest version (1.5228 standard deviation below) mean of user rating, which indicates that mobile application users do not favor these applications. The overall and latest version counts that are significantly below the overall mean support this hypothesis. Thus, the cluster can be labeled as “poor applications”.

Cluster 3 has 337 applications. The most important difference from the mean was observed in the number of overall and latest version rating counts (3.123 and 2.904 standard deviation above respectively) and supported languages (0.524 standard deviation above the average). Also, these applications are rated above general mean. As a result, it can be named as “worldwide populars”.

INSERT FIGURE 2 HERE

Fuzzy c-means clustering

As mentioned previously, 63 fuzzy c-means models with different parameters (c and m) were run and the results are obtained. The model selection was made based on clustering performances measured by XB indices. The one that produced the smallest value is considered to be the best model. The values for different parameters are illustrated below:

INSERT TABLE 4 HERE

The model fcl1c outputs the smallest XB index, which indicates that it outperformed the others and, therefore, was flagged as can be seen from the table. The parameters of fc1b are c=4 and m=1.05 respectively. The resulting cluster centers are displayed below:

INSERT TABLE 5 HERE

Cluster 1 has 1597 members and its primarily distinguishing characteristic is its consistently poor performance at almost all dimensions. The most striking difference of Cluster 1 occurs at the mean of overall and latest version user ratings, whose means remain 1.42 standard deviation below in overall user ratings and 1.55 standard deviation below the mean latest version ratings respectively. Due to these facts, the Cluster 1 can be called as “poor applications”.

Cluster 2 grouped 322 applications. The Cluster 2 has the highest values in terms of mean supported languages, user ratings for overall and latest version, overall and latest version rating counts. Taking these characteristics into consideration, it can be claimed that the applications in Cluster 2 are more popular than those in other clusters. As a result, the applications in Cluster 2 can be labeled as “Worldwide Populars”.

Cluster 3 has 689 applications. There exist a huge difference in price and sizes for these applications from other clusters, 1.640 standard deviation above the mean size and 1.4359 standard deviation above the mean price. The other aspect that Cluster 3 applications have higher values is the screenshots provided to the users. Interpreting all these results as a whole, these applications probably belong to a class that is closely related to games or educational genre. Therefore, the cluster 3 can be called “paid games”.

The last cluster, Cluster 4, representing the majority of the mobile applications available on Mobile App Store can be characterized by higher ratings for overall and latest version, whereas their rating counts are less than the mean. The main distinguishing feature of these applications is the number of supported devices, highest among all clusters. In the light of these findings, these applications can be regarded as "less popular high performers". The resulting clusters are visualized below:

INSERT FIGURE 3 HERE

Comparative analysis

In terms of internal validity measure, fuzzy c- means algorithm produced a far smaller XB value, 0.0000925, whereas k-means clustering exhibited an inferior performance with an 85.902. It clearly shows that fuzzy c-means is better at partitioning the mobile applications into compact and separate groups. In addition to the quality of indices, there are efficiency measures that can be used as a basis for comparing the computational performances of the algorithms. For the best models, k-means algorithms converged after only 4 iterations whereas it takes 25 iterations for fuzzy c- means algorithm to reach the optimal cluster centers. To compute the time complexities of k-means and fuzzy c-means algorithm these notations can be used (Ghosh & Dubey, 2013):

$O(ncdi)$ and $O(ndc^2i)$ respectively, where;

n=data points;

c= number of clusters;

d= number of dimension;

i=number of iterations

The elapsed time for k-means algorithm to converge is 0.2228 seconds and it is 1.1882 seconds for fuzzy c-means algorithm, which is not surprising at all as one can simply understand from the time complexities.

As for the structure of the clusters, the best k-means algorithm produced 3 clusters as the best fuzzy c-means partition the applications into 4 different clusters. These clusters exhibited a high level of similarity as one can recall from the findings. For example, the cluster 2 produced by k-means and cluster 1 by fuzzy c-means group the poor applications under the same cluster. On the other hand, the cluster 3 in k-means and cluster 2 in fuzzy c-means mostly consist of the worldwide popular programs. Also, their sizes are very close to each other (337 in k-means and 322 in fuzzy c- means). The cluster that is unique to fuzzy c-means algorithm is cluster 3 that distinguishes from the clusters those produced by k-means in terms of their huge sizes and prices. This extra cluster with 689 members was named as paid games.

INSERT TABLE 6 HERE

Conclusion

The aim of the study is to investigate and compare the efficiency and effectiveness of two separate clustering algorithms, namely k-means and fuzzy c- means in grouping the mobile applications. Drawing upon various features of 7196 applications available on Mobile App store, an experimental approach was adopted to determine optimal parameter values for these algorithms. After removing the outliers and standardization, there remained 6843 applications. In the clustering analysis step, k-means and fuzzy c-means algorithm were run with these cleansed data. For k-means algorithm, 9 different k-values from 2 to 10 were tried and the k value that produced the highest quality clusters was selected as the best model. It was 3 for k-means clustering. On the other hand, 63 different pairs of parameters (cluster number, fuzziness index) were run and the optimal values for these parameters were determined. They are 4 for cluster numbers (c) and 1.05 for fuzziness index.

When the generated clusters are analyzed, both algorithms successfully identified and grouped poor performers and worldwide popular applications. The main difference was observed in the grouping applications that has relatively higher rating, size and prices. Fuzzy c- means algorithm additionally discovered an underlying group that mostly consists of high size-high price applications that provide the highest number of screenshots to their users. In k-means clustering this group is merged with the highly functional applications. As a general conclusion, it can be argued that the most influential factors on the grouping of the applications are mean ratings and rating counts, size and price.

The performance indicators suggest that fuzzy c-means prevailed over the k-means algorithm in terms of cluster quality measured by Xie and Beni index. (0.0000925 for fuzzy c-means and 85.902 for k-means algorithm) But, from the efficiency point of view, k-means algorithm is far better than its counterpart. More clearly, it converges more rapidly with less iteration, which may cause huge efficiency differences in larger datasets, thus being more preferable in some cases.

The current study is not without its limitations. First, the dataset contains a sample of applications available in Mobile App store. The representativeness of this sample may not be adequate. On the other hand, without including the Google Play Store apps, it won't be possible to reach generalizations about the underlying characteristics of the application groups.

As a future study, various methods for both fuzzy (e.g. algorithms proposed by Gath & Geva (1989), Gustafson & Kessel (1979)) and crisp clustering(CLARA, DBSCAN etc.) will be tested to see whether a significant improvement is achieved in terms of cluster quality. In addition, the effectiveness and efficiency of these proposed algorithms will be run for different datasets. In doing so, some external validity measures will also be employed to evaluate the individual performances from a broader perspective. Lastly, a comparative analysis by using Google Play store applications will help acquiring a more comprehensive understanding of the preferences of Android users, thus providing more fruitful insights to the mobile application developers and companies.

References

Baradwaj, B., Pal, S. (2012). Mining educational data to analyze students' performance. *arXiv* , preprint arXiv:1201.3417.

Bezdek, J. C. (1981). Objective function clustering. J. C. Bezdek içinde, *In Pattern recognition with fuzzy objective function algorithms* (s. 43-93). Boston,MA: Springer.

Biel, B., Grill, T., & Gruhn, V. (2010). Exploring the benefits of the combination of a software architecture analysis and a usability evaluation of a mobile application. *Journal of Systems and Software*, *83*(11), 2031-2044.

Böhmer, M., Hecht, B., Schöning, J., Krüger, A., Bauer, G. (2011, August). Falling asleep with Angry Birds, Facebook and Kindle: a large scale study on mobile application usage. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services* (pp. 47-56). ACM.

Cannon, R. L., Dave, J. V., Bezdek, J. C. (1986). Efficient implementation of the fuzzy c-means clustering algorithms. *IEEE transactions on pattern analysis and machine intelligence*, 248-255.

Cebeci, Z., Yildiz, F. (2015). Comparison of K-means and Fuzzy C-means algorithms on different cluster structures. *AGRARINFORMATIKA/JOURNAL OF AGRICULTURAL INFORMATICS, 6*(3), 13-23.

Cai, W., Chen, S., Zhang, D. (2007). Fast and robust fuzzy c-means clustering algorithms incorporating local information for image segmentation. *Pattern recognition, 40*(3), 825-838.

Desgraupes, B. (2011). *Clustering Indices.* Paris: Lab Modal X.

Dhanachandra, N., Manglem, K., Chanu, Y. (2015). Image segmentation using K-means clustering algorithm and subtractive clustering algorithm. *Procedia Computer Science, 54*, 764-771.

Dhanalakshmi, K., Inbarani, H. (2012). Fuzzy soft rough k-means clustering approach for gene expression data. *arxiv preprint*, 1-7.

Dunn, J. (1973). A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. *Journal of Cybernetics, 3*(3), 32-57.

Gath, I., Geva, A. B. (1989). Unsupervised optimal fuzzy clustering. *IEEE Transactions on pattern analysis and machine intelligence, 11*(7), 773-780.

Ghosh, S., Dubey, S. K. (2013). Comparative analysis of k-means and fuzzy c-means algorithms. *International Journal of Advanced Computer Science and Applications, 4*(4), 35-38.

Gustafson, D. E., Kessel, W. C. (1979). Fuzzy clustering with a fuzzy covariance matrix. *17th Symposium on Adaptive Processes, 1978 IEEE Conference* (s. 761-766). IEEE.

Han, J., Pei, J., Kamber, M. (2011). *Data mining: concepts and techniques.* New York: Elsevier.

Hruschka, H., Natter, M. (1999). Comparing performance of feedforward neural nets and K-means for cluster-based market segmentation. *European Journal of Operational Research*, 114(2), 346-353.

Ickin, S., Wac, K., Fiedler, M., Janowski, L., Hong, J. H., Dey, A. K. (2012). Factors influencing quality of experience of commonly used mobile applications. *Communications Magazine, IEEE*, *50*(4), 48-56.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern recognition letters, 31*(8), 651-666.

Kim, D. W., Lee, K. H., Lee, D. (2004). On cluster validity index for estimation of the optimal number of fuzzy clusters. *Pattern Recognition, 37*(10), 2009-2025.

Kim, K., Ahn, H. (2008). A recommender system using GA K-means clustering in an online shopping market. *Expert systems with applications*, 34(2), 1200-1209.

Liu, Y., Li, Z., Xiong, H., Gao, X., Wu, J. (2010). Understanding of Internal Clustering Validation Measures. *IEEE International Conference on Data Mining* (s. 911-916). IEEE.

MacQueen, J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability* (s. 281-29). University of California Press.

Maulik, U., Bandyopadhyay, S. (2002). Performance evaluation of some clustering algorithms and validity indices. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24*(12), 1650-1654.

Mingoti, S., Lima, J. (2006). Comparing SOM neural network with Fuzzy c-means, K-means and traditional hierarchical clustering algorithms. *European Journal of Operational Research, 174*(3), 1742-1759.

Mobile App Usage - Statistics & Facts. (2018). on July 27, 2018 retrieved from Statista: https://www.statista.com/topics/1002/mobile-app-usage/

Most popular Apple App Store categories in May 2018, by share of available apps. (2018). on July 27, 2018 retrieved from Statista: https://www.statista.com/statistics/270291/popular-categories-in-the-app-store/

Most popular Google Play categories. (2018, July 26). on July 27, 2018 retrieved from AppBrain: https://www.appbrain.com/stats/android-market-app-categories

Naaman, M., Nair, R., Kaplun, V. (2008, April). Photos on the go: a mobile application case study. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1739-1748). ACM.

Ng, H., Ong, S., Foong, K., Goh, P., Nowinski, W. (2006). Medical image segmentation using k-means clustering and improved watershed algorithm. *Image Analysis and Interpretation* (s. 61-65). IEEE.

Number of apps available in leading app stores as of 1st quarter 2018. (2018). on July 27, 2018 retrieved from Statista: https://www.statista.com/statistics/276623/number-of-apps-available-in-leading-app-stores/

Number of available applications in the Google Play Store from December 2009 to June 2018. (2018). on July 27, 2018 retrieved from: Statista: https://www.statista.com/statistics/266210/number-of-available-applications-in-the-google-play-store/ adresinden alındı

Number of available apps in the Amazon Appstore from 2nd quarter 2015 to 1st quarter 2018. (2018). on July 27, 2018 retrieved from: Statista: https://www.statista.com/statistics/307330/number-of-available-apps-in-the-amazon-appstore/

Oyelade, O., Oladipupo, O., Obagbuwa, I. (2010). Application of k Means Clustering algorithm for prediction of Students Academic Performance. *arXiv* .

Penttinen, E., Rossi, M., Tuunainen, V. K., 2010. Mobile Games: Analyzing the Needs and Values of the Consumers. *Journal of Information Technology Theory and Application*, vol. 11, no. 1, pp. 5-22.

Rong, C. (2011). Using Mahout for clustering Wikipedia's latest articles: A comparison between k-means and fuzzy c-means in the cloud. *Cloud Computing Technology and Science (CloudCom), 2011 IEEE Third International Conference* (s. 565-569). IEEE.

Shieh, L., Chang, T. H., Fu, H. P., Lin, S. W., Chen, Y. Y. (2013). Analyzing the factors that affect the adoption of mobile services in Taiwan. *Technological Forecasting and Social Change*.

Wagstaff, K., Cardie, C., Rogers, S., Schrödl, S. (2001). Constrained k-means clustering with background knowledge. *International Conference on Machine Learning*, (s. 577-587). Williamstown,MA.

Wasserman, T. (2010). Software engineering issues for mobile application development. *FoSER 2010.*

Wu, K. (2012). Analysis of parameter selections for fuzzy c-means. *Pattern Recognition, 45*(1), 407-415.

Yan, Z., Liu, C., Niemi, V., Yu, G. (2010). Effects of displaying trust information on mobile application usage. In *Autonomic and Trusted Computing* (pp. 107-121). Springer Berlin Heidelberg.

Yang, H. C., 2013. Bon Apetit for Apps: Young American Consumers'Acceptance of Mobile Application. *The Journal of Computer Information Systems*, vol. 53, no. 3, pp. 85-96.

Xie, X. L., Beni, G. (1991). A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 841-847.

Zhang, D., Chen , S. (2004). A novel kernelized fuzzy c-means algorithm with application in medical image segmentation. *Artificial intelligence in medicine, 32*(1), 37-50.

Figures and Tables

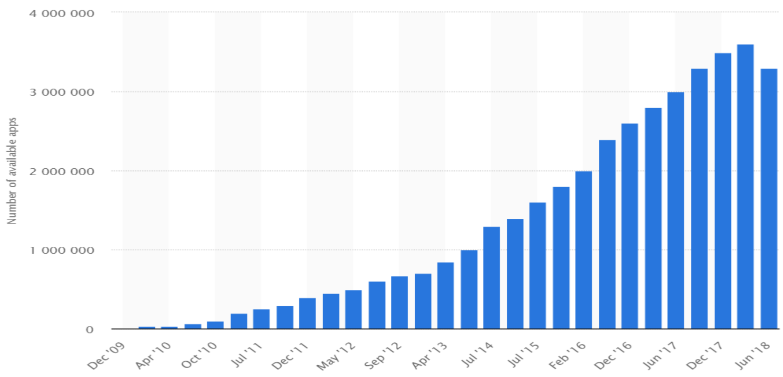

Figure 1: Number of available applications in the Google Play Store from December 2009 to June 2018

Table 1: The description of included variables

Variable	**Description**
id	Mobile App_id
size_bytes	Size (in Bytes)
price	Price amount
rating_count_tot	User Rating counts (for all version)
rating_count_ver	User Rating counts (for current version
user_rating	Average User Rating value (for all version)
user_rating_ver	Average User Rating value (for current version)
sup_devices.num	Number of supporting devices

ipadSc_urls.num	Number of screenshots showed for display
lang.num	Number of supported languages

Table 2: Summary statistics of the selected variables

Variable	***Mean***	**Standard Deviation**
Size_bytes	*199148055*	359230022
Price	*1.7264*	5.8333
Sup_devices.num	*37.3618*	3.7379
Lang.num	*5.4342*	7.9199
User_rating	*3.5270*	1.5180
User_rating_ver	*3.2538*	1.8092
Rating_count_tot	*12886.27*	75742.58
Rating_count_ver	*460.43*	3920.725
IpadSc_urls.num	*3.7069*	1.9860

Table 3: Cluster centers and sizes for k-means clustering

$(Variables/Clusters)$	**Cluster1** **N_1=4803**	**Cluster2** **N_2=1703**	**Cluster3** **N_3=337**
size_bytes	0.1021	-0.2968	0.044
Price	0.1328	-0.2829	-0.4299
Sup_devices.num	-0.0355	0.0847	0.0784
Lang.num	0.0939	-0.3688	0.524
User_rating	0.4419	-1.3520	0.5342
User_rating_ver	0.4980	-1.5228	0.5967
Rating_count_tot	-0.1246	-0.2663	3.123
Rating_count_ver	-0.0846	-0.335	2.9046
IpadSc_urls.num	0.2104	-0.6357	0.2128

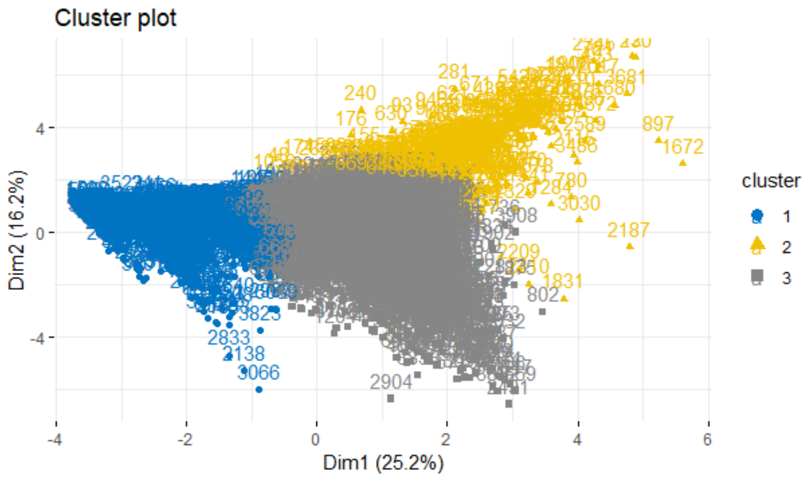

Figure 2: Scatter plot of resulting k-means clustering algorithm

Table 4: XB values for different fuzzy c means models

c/m	m=1.05	m=1.5	m=2	m=2.5	m=3	m=3.5	m=5
c=2	$123*10^{-6}$	$183*10^{-6}$	$362*10^{-6}$	2.93	34.8	47	65.1
c=3	$107*10^{-6}$	$427*10^{-6}$	6.02	$3.17*10^{14}$	20183	22496	10345
c=4	**$92.5*10^{-6**}$**	$249*10^{-6}$	222	$2.62*10^{10}$	5172553	1155	2283
c=5	$138*10^{-6}$	$173*10^{-6}$	186	3405455	$1.2*10^{13}$	$3.67*10^{10}$	7089
c=6	$123*10^{-6}$	$293*10^{-6}$	18463897	1497	∞	12502	$1.56*10^{8}$
c=7	$127*10^{-5}$	$246*10^{-6}$	15947	1059429	$6.56*10^{20}$	$2.82*10^{8}$	3860337
c=8	$111*10^{-6}$	$227*10^{-6}$	1450089	67928	320314	$2.27*10^{12}$	58554
c=9	$132*10^{-6}$	$196*10^{-6}$	443311	$2.3*10^{12}$	239026	$9.87*10^{9}$	988
c=10	$159*10^{-6}$	$159*10^{-6}$	18958	$2.85*10^{21}$	$1.81*10^{11}$	189743	396334

Table 5: Cluster centers and sizes for fuzzy c-means algorithm

$(Variables/Clusters)$	**Cluster1** **N_1=1597**	**Cluster2** **N_2=322**	**Cluster3** **N_3=689**	**Cluster4** **N_4=4235**
size_bytes	-0.3072	-0.0047	1.640	-0.149
Price	-0.3096	-0.4493	1.4359	-0.0822
Sup_devices.num	0.1391	0.0853	-1.8360	0.2376
Lang.num	-0.3813	0.5280	0.0327	0.0978
User_rating	-1.4298	0.5363	0.3049	0.4481
User_rating_ver	-1.550	0.5979	0.3221	0.4897
Rating_count_tot	-0.2700	3.2200	-0.1406	-0.1190
Rating_count_ver	-0.3372	2.9560	-0.1049	-0.0792
IpadSc_urls.num	-0.6611	0.2187	0.4361	0.1620

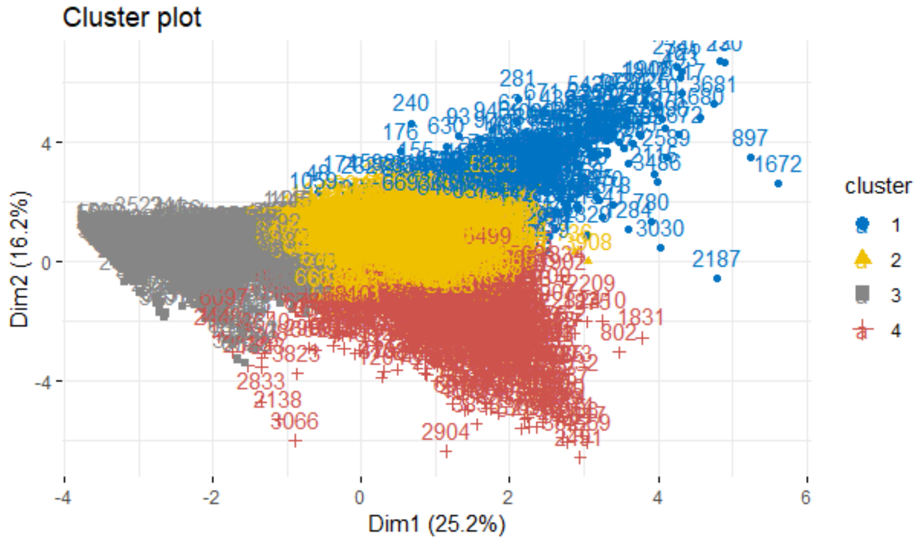

Figure 3: The scatter plot for fuzzy c-means clustering algorithm

Table 6: Comparison of the selected performance criteria for k-means and fuzzy c means algorithm

$\boldsymbol{Algorithm}/_{\boldsymbol{Criteria}}$	**Quality(XB)**	**Time Complexity**	**Elapsed Time**	**Iteration**	**Number of clusters**
k-means	85.902	O(ncdi)	0.2228	4	3
fuzzy c means	0.0000925	O(ndc^{2}i)	1.1882	25	4