Erciyes Üniversitesi Sosyal Bilimler Enstitüsü Dergisi XLIV, 2018/1, 41-57
Geliş Tarihi: 21.12.2017, Yayın Tarihi: 30.06.2018

COMPARING CLUSTERINGS: A STORE SEGMENTATION APPLICATION[*]

Emrah BİLGİÇ[a] Özgür ÇAKIR[b]

Abstract

This study focuses on one of the clustering comparison techniques, pair counting measures such as Rand Index, Adjusted Rand Index and Fowlkes Mallows Index. The aim of the study is discussing the properties of mentioned indexes and indicating a marketing application of the techniques. For the case study, a retail chain company's supermarket stores are segmented using cluster analysis with two different approach. The first clustering approach is segmenting the stores based on socioeconomic factors and the second approach is segmenting based on purchasing behaviors of customers. Since consumer purchases are influenced strongly by socioeconomic factors, this study expects to find an agreement between two clusterings. The results indicates that while Rand Index value found an agreement, Fowlkes-Mallows Index value found a weak agreement and Adjusted Rand Index value could not find any agreement between two clusterings.

Keywords: Comparing Clusterings, Clustering Agreements, Store Segmentation.

Introduction

Clustering, one of the most popular techniques in Data Mining, partitions data points into natural groups called clusters such that points are quite similar within a group (cluster) and dissimilar across clusters. It has a widespread usage area from data summarization at the broadest level to specific topics such as

[*] This paper has been derived from a dissertation entitled "Retail Store Segmentation With Rule Based And Socioeconomic Approaches: An Application On A Chain Company".

[a] Asst. Prof., Mus Alparslan University, e.bilgic@alparslan.edu.tr

[b] Assoc. Prof., Marmara University, ocakir@marmara.edu.tr

customer segmentation, outlier detection, social network analysis and other data mining problems.[1,2]

The most important steps to develop a clustering process can be summarized as follows; 1) Feature selection, 2) Algorithm selection, 3) Cluster Validation and 4) Interpretation of the results.

1) Feature selection: Features (variables) which can clearly distinguish data points from each other will ease the clustering task. Picking the variables that are non-distinguishing may cause not to find any cluster structures.

2) Algorithm selection: In algorithm selection step the similarity measure which will be used to define the relationship between data points is decided first. Generally the relationship between the data points are represented in a proximity (similarity or dissimilarity) measure. After deciding which similarity/dissimilarity measure to use one should select one of the clustering algorithms. Although there are no unique objective "true" or "best" clusters especially in a multi-dimensional data set, researchers should select the most appropriate clustering algorithm according to the aim and what kind of clusters they are looking for.[3] There exist a great number of clustering algorithms available;[4] perhaps k-means clustering and hierarchical clustering are the two best known ones.

3) Cluster Validation: The third step, validity of a clustering can be expressed in terms of the three types of criteria. The first one is *external measures* that compare the clustering to a given true clustering, the second one is *internal measures* which only utilize the information of feature vectors of data points and the last one *hybrid measures* which take both information into account.[5,6] The methods in external criteria are also used in comparing clusterings task, for

[1] Zaki Mohammed J, & Meira Jr Wagner. (2014). *Data mining and analysis: fundamental concepts and algorithms*: Cambridge University Press.

[2] Aggarwal Charu C. (2015). *Data mining: The textbook*. Switzerland: Springer.

[3] Hennig Christian, & Liao Tim F. (2013). How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 62*, 309-369.

[4] Jain A. K., Murty M. N., & Flynn P. J. (1999). Data clustering: a review. *ACM Computing Surveys, 31*, 264-323.

[5] Jain Anil K, & Dubes Richard C. (1988). *Algorithms for clustering data*: Prentice-Hall, Inc.

[6] Xiang Qiaoliang, Mao Qi, Chai Kian Ming, Chieu Hai Leong, Tsang Ivor, & Zhao Zhendong. (2012). A Split-Merge Framework for Comparing Clusterings. *arXiv preprint arXiv:1206.6475*.

assessing the goodness/quality of clustering solutions according to a reference clustering.[7,8]

4) Interpretation of the results: Experts in the relevant fields may interpret the partitioned data and proceed for further analysis.

This study focuses on comparing clusterings issue. One of the most popular techniques, pair counting measures are handled by applying them to a marketing problem, retail store segmentation. We first segment 175 stores of a supermarket company into six segments using socioeconomic variables. Then we segment same stores into five segments using customer purchasing behaviors of each store. Note that two clusterings may have different numbers of clusters when comparing them. [9,10]

While clustering evaluation measures are commonly used to compare the performance of different algorithms, they should be able to compare clusterings of different data sets. [6] In this study although we use two different data sets, both of them belong to the same supermarket stores. The first clustering's data set consists of variables about customers of each store (such as demographics) and about store area (such as competitor number and the region they are placed). The second clustering's data set consist of customer purchasing behaviors of each store which are obtained using association rule analysis from the transaction data set of each store.

Since socioeconomic factors effects purchasing behaviors of customers, one can expect that two clusterings described above should be in agreement.[11]

[7] Vinh Nguyen Xuan, Epps Julien, & Bailey James. (2009). Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *Proceedings of the 26th Annual International Conference on Machine Learning* (pp. 1073-1080): ACM.

[8] Romano Simone, Bailey James, Nguyen Xuan Vinh, & Verspoor Karin. (2014). Standardized Mutual Information for Clustering Comparisons: One Step Further in Adjustment for Chance. In *ICML* (pp. 1143-1151).

[9] Rabbany Reihaneh, & Zaïane Osmar R. (2015). Generalization of clustering agreements and distances for overlapping clusters and network communities. *Data mining and knowledge discovery, 29*, 1458-1485.

[10] Meilă Marina. (2007). Comparing clusterings—an information based distance. *Journal of multivariate analysis, 98*, 873-895.

[11] Kotler Philip, & Gary Armstrong. (2012). *Principles of marketing* (Vol. 14th ed.). New Jersey: Prentice Hall.

A. Literature

There exist a great number of clustering algorithms available in the literature. Since each of the algorithms generate different results, outputs should be investigated by an expert. Apart from that, there are measures which evaluate clusterings results to find the best partitioning.

A very basic approach for comparing clusterings is counting pairs of objects that are clustered in the same way in both clusterings.[12] There exists many pair-counting based measures such as the Rand Index,[13] Jaccard coefficient, [14] Fowlkes-Mallows [15] and Huberts.[16] Apart from pair counting techniques, there are also information theoretic measures such as Variation of Information which is an external measure as counting pairs of objects measures.[10]

As mentioned before, in external criteria approach, the results of a clustering algorithm is evaluated based on a pre-specified structure. An external measure can also measure the degree to which data confirm a priori ideas without a cluster analysis being performed. [4]

Consider a dataset D consisting of n data items $D = \{d_1, d_2, d_3 \dots d_n\}$. A partitioning X partitions D into k mutually disjoint subsets, $X = \{X_1, X_2, X_3 \dots X_k\}$. Let Y denote another partitioning of the same dataset D, $Y = \{Y_1, Y_2, Y_3 \dots Y_r\}$. Each pair (d_i, d_j) of data items is classified into one of four groups based on their co-memberships in X and Y; which results in the following pair-counts.

Table 1: Pair Counts

	Same in Y	**Different in Y**
Same in X	M_{11}=*TP*	M_{10}= *FP*
Different in X	M_{01}=*FN*	M_{00}=*TN*

[12] Wagner Silke, & Wagner Dorothea. (2007). *Comparing clusterings: an overview*: Universität Karlsruhe, Fakultät für Informatik Karlsruhe.
[13] Rand William M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association, 66*, 846-850.
[14] Jaccard Paul. (1908). *Nouvelles recherches sur la distribution florale*.
[15] Fowlkes Edward B, & Mallows Colin L. (1983). A method for comparing two hierarchical clusterings. *Journal of the American Statistical association, 78*, 553-569.
[16] Hubert Lawrence, & Arabie Phipps. (1985). Comparing partitions. *Journal of Classification, 2*, 193-218.

Here, M_{11}/M_{00} counts the number of pairs that are in the same/different partitions in both X and Y. M_{10}/M_{01} sums up those that belong to the same/different partitions in X but are in different same/partitions according to Y. True/false, positive/negative scores, denoted by TP, FP, TN, and FN in the table.

These pair countings are derived using the following contingency table. [18] The table is a k^*r matrix of all the possible overlaps between each pair of clusters in X and Y, where its ijth element shows the intersection of cluster X_i and Y_j, i.e. $n_{ij} = \left|X_i \cap Y_j\right|$.

Considering co-membership of data points in the same or different clusters as a binary variable, Jaccard agreement between clustering X and Y can be defined as;

$$J = \frac{TP}{FP+FN+TP} = \frac{M_{11}}{(M_{01}+M_{10}+M_{11})} \qquad (1)$$

Rand Index (RI) is defined similarly to Jaccard, but it also values pairs that belong to different clusters in both partitionings, i.e. true negatives:

$$RI = \frac{M_{11}+M_{00}}{M_{11}+M_{01}+M_{10}+M_{00}} \qquad (2)$$

which gives;

$$RI = 1 + \frac{1}{(n^2-n)}\left(2\sum_{i=1}^{k}\sum_{j=1}^{r} n_{ij}^2 - \left(\sum_{i=1}^{k} n_{i.}^2 + \sum_{j=1}^{r} n_{.j}^2\right)\right) \qquad (3)$$

Table 2: Contingency Table

	Y_1	Y_2	…	Y_r	marginal sums
X_1	n_{11}	n_{12}	…	n_{1r}	$n_{1.}$
X_2	n_{21}	n_{22}	…	n_{2r}	$n_{2.}$
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
X_k	n_{k1}	n_{k2}	…	n_{kr}	$n_{k.}$
marginal sums	$n_{.1}$	$n_{.2}$	…	$n_{.r}$	n

A problem with the Rand Index is that the expected value of the Rand Index of two random partitions does not take a constant value. Hubert and

Arabie, [18] proposed the Adjusted Rand Index (ARI) with a constant expected value which is bounded above one, and takes value of zero when the index equals its expected value:

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2}-\left[\sum_i\binom{n_{i.}}{2}\sum_j\binom{n_{.j}}{2}\right]/\binom{n}{2}}{\frac{1}{2\left[\sum_i\binom{n_{i.}}{2}+\sum_j\binom{n_{.j}}{2}\right]}-\left[\sum_i\binom{n_{i.}}{2}\sum_j\binom{n_{.j}}{2}\right]/\binom{n}{2}} \tag{4}$$

Yeung and Ruzzo explain in detail why value of the Rand Index takes much higher value than the Adjusted Rand Index as follows: [17] Since the Rand Index lies between zero and one, the expected value of the Rand Index (although not a constant value) must be greater than or equal to zero. On the other hand, the expected value of the Adjusted Rand Index has value zero and the maximum value of the Adjusted Rand Index is also one. Hence, there is a wider range of values that the Adjusted Rand Index can take on; this issue increases the sensitivity of the index.

Another pair counting technique Fowlkes-Mallows Index is introduced for comparing hierarchical clusterings. [17] However it can also be used for flat clusterings. Wagner and Wagner,[14] generalized the formula for comparing the clusterings which have different numbers of clusters as follows:

Let *X* be a finite set with cardinality |*X*| = *n*. A clustering *C* is a set {C_1, … , C_k) of non-empty disjoint subsets of *X* such that their union equals *X*. The set of all clusterings of *X* is denoted by *P(X)*. For a clusterings C= {C_1, … , C_k) we assume $|C_i > 0|$ for all *i* = *1, . . . , k*. A trivial clusterings is either the one-clustering that consist of just one cluster or the singleton clustering in which every element forms its own cluster.

Let C'= $\{C_1', \dots, C_l'\}$ denote a second clustering of *X*. The confusion matrix M= (m_{ij}) (or contingency table) of the pair *C* and C' is a *k* * *l* matrix whose *ij*-th entry equals the number of elements in the intersection of the clusters C_i and C_j'.

$$FM(C,C') = \frac{\sum_{i=1}^{k}\sum_{j=1}^{l} m_{ij}^2-n}{\sqrt{(\sum_i|C_i|^2-n)\left(\sum_j\left|C_j'\right|^2-n\right)}} \tag{5}$$

$$= \frac{M_{11}}{\sqrt{(M_{11}+M_{10})(M_{11}+M_{01})}}$$

[17] Yeung Ka Yee, & Ruzzo Walter L. (2001). Details of the adjusted Rand index and clustering algorithms, supplement to the paper "An empirical study on principal component analysis for clustering gene expression data". *Bioinformatics, 17*, 763-774.

B. Methodology

In this section; data source, data description and data mining tools used for implementation are briefly explained. For the purpose of clustering the stores by two different approach, empirical transaction data is obtained from a well-known supermarket chain company. Although the stores analyzed in this study are in Istanbul Turkey, the case firm has many stores with complex and detailed product categories all around the country. Company's 175 stores in Istanbul are selected for the analysis.

In the first clustering approach (store segmentation based on socioeconomic factors) socioeconomic variables for 175 stores are used. In the second approach (store segmentation based on customer purchasing behaviors) purchasing behaviors for each store are used [18].

The data of socioeconomic variables such as age distributions, marital status, financial status and education level of the people who reside around each store are obtained. Furthermore, competition level and the economic environment (if the store is in a factory, university, finance or touristic area or not) for each store are investigated. The data for socioeconomic segmentation are obtained from Turkish Statistical Institute, Google Maps and real estate companies (see the data for five sample stores in Table 3).

Table 3: Socio-economic Data For Five Stores

Features	Store 1	Store 2	Store 3	Store 4	Store 5
Rental of an apartment (TL)	*1200*	*2200*	*900*	*3000*	*600*
Factory area	*0*	*0*	*0*	*1*	*0*
University area	*1*	*0*	*0*	*0*	*1*
Trade Area	*0*	*0*	*0*	*0*	*0*
Touristic Area	*0*	*0*	*0*	*1*	*0*
Number of Competitor	*5*	*1*	*7*	*1*	*1*
Age 0-4	*0.0854*	*0.0546*	*0.0668*	*0.0376*	*0.0928*
Age 5-14	*0.1714*	*0.1502*	*0.1268*	*0.0749*	*0.1702*

[18] Bilgic Emrah. (2016). Retail store segmentation with rule based and socioeconomic approaches: An application on a chain company. Unprinted PhD dissertation, Istanbul, Marmara University.

Age 15-34	*0.4498*	*0.2675*	*0.3363*	*0.2632*	*0.3647*
Age 35-54	*0.2272*	*0.3631*	*0.3069*	*0.3246*	*0.2921*
Age 55+	*0.0663*	*0.1646*	*0.1631*	*0.2996*	*0.0802*
Single	*0.5062*	*0.4475*	*0.4156*	*0.6010*	*0.3724*
Married	*0.4938*	*0.5525*	*0.5844*	*0.3990*	*0.6276*
No Education	*0.1826*	*0.0932*	*0.0908*	*0.0579*	*0.1356*
Low Educated	*0.3832*	*0.2049*	*0.4821*	*0.2252*	*0.5622*
Middle Educated	*0.3224*	*0.2246*	*0.2629*	*0.2954*	*0.1757*
High Educated	*0.1118*	*0.5021*	*0.1642*	*0.4481*	*0.1283*

After applying Ward's Hierarchical Clustering Algorithm[19] six segments are formed. Istanbul map in Figure 1 shows the distribution of the stores, each symbol represents a different segment. The Ward hierarchical clustering method has been widely used since its first description by Ward. This technique is the only one among the other clustering methods which is based on sum-of-squares criterion. It produces groups that minimize within-group dispersion at each binary fusion.

For the second clustering approach, a one day transactional data of each store are obtained from the company (see Table 4 for a sample of row data). Some association rules are generated from the main transaction data to be searched in each store. Apriori Algorithm[20] is used for association rule mining. These rules are called *strong rules* since they have a threshold support, confidence and lift value. Strong rules are identified and searched one by one in transaction data of each store for finding out purchasing behaviors of customers. The matrix of support values is the input for clustering analysis. After applying Ward's algorithm, five segments are formed. Since consumer purchases are influenced strongly by cultural (Hispanic American, African American…), social (small groups and networks, family, social roles and status), personal (age, occupation, lifestyle) and psychological (perception, beliefs, attitudes…)

[19] Ward Jr Joe H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical association, 58*, 236-244.

[20] Agrawal Rakesh, & Srikant Ramakrishnan. (1994). Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB* (Vol. 1215, pp. 487-499).

characteristics, [13,21] one may consider that those two clusterings are in agreement.

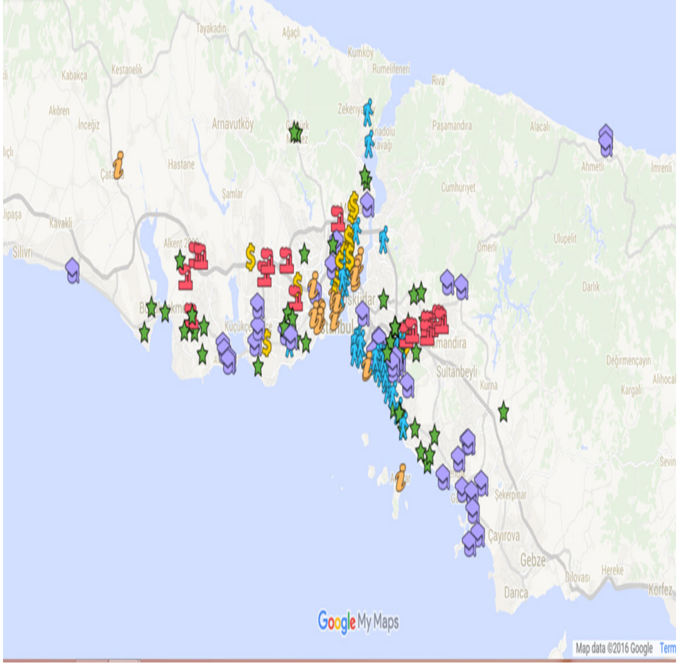
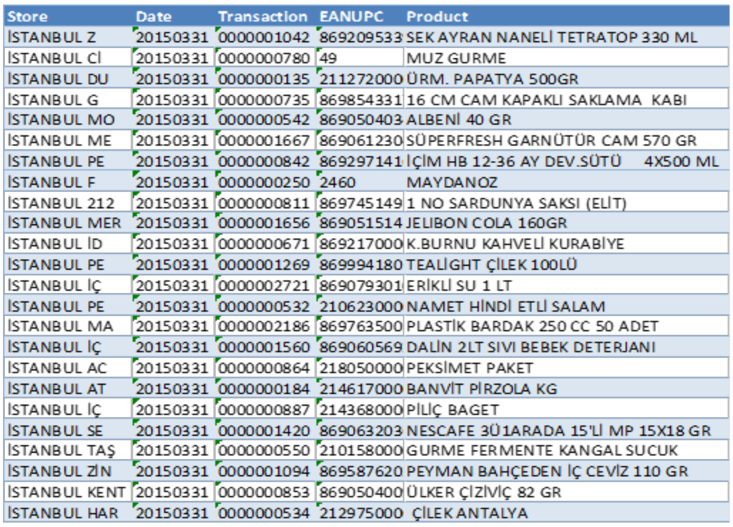


Figure 1: Stores on Istanbul map, each segment represented with a different symbol.

Table 4: A Sample Of Raw Data



Store	Date	Transaction	EANUPC	Product
İSTANBUL Z	20150331	0000001042	869209533	SEK AYRAN NANELİ TETRATOP 330 ML
İSTANBUL Cİ	20150331	0000000780	49	MUZ GURME
İSTANBUL DU	20150331	0000000135	211272000	ÜRM. PAPATYA 500GR
İSTANBUL G	20150331	0000000735	869854331	16 CM CAM KAPAKLI SAKLAMA KABI
İSTANBUL MO	20150331	0000000542	869050403	ALBENİ 40 GR
İSTANBUL ME	20150331	0000001667	869061230	SÜPERFRESH GARNÜTÜR CAM 570 GR
İSTANBUL PE	20150331	0000000842	869297141	İÇİM HB 12-36 AY DEV.SÜTÜ 4X500 ML
İSTANBUL F	20150331	0000000250	2460	MAYDANOZ
İSTANBUL 212	20150331	0000000811	869745149	1 NO SARDUNYA SAKSI (ELİT)
İSTANBUL MER	20150331	0000001656	869051514	JELIBON COLA 160GR
İSTANBUL İD	20150331	0000000671	869217000	K.BURNU KAHVELİ KURABİYE
İSTANBUL PE	20150331	0000001269	869994180	TEALİGHT ÇİLEK 100LÜ
İSTANBUL İÇ	20150331	0000002721	869079301	ERİKLİ SU 1 LT
İSTANBUL PE	20150331	0000000532	210623000	NAMET HİNDİ ETLİ SALAM
İSTANBUL MA	20150331	0000002186	869763500	PLASTİK BARDAK 250 CC 50 ADET
İSTANBUL İÇ	20150331	0000001560	869060569	DALİN 2LT SIVI BEBEK DETERJANI
İSTANBUL AC	20150331	0000000864	218050000	PEKSİMET PAKET
İSTANBUL AT	20150331	0000000184	214617000	BANVİT PİRZOLA KG
İSTANBUL İÇ	20150331	0000000887	214368000	PİLİÇ BAGET
İSTANBUL SE	20150331	0000001420	869063203	NESCAFE 3Ü1ARADA 15'Lİ MP 15X18 GR
İSTANBUL TAŞ	20150331	0000000550	210158000	GURME FERMENTE KANGAL SUCUK
İSTANBUL ZİN	20150331	0000001094	869587620	PEYMAN BAHÇEDEN İÇ CEVİZ 110 GR
İSTANBUL KENT	20150331	0000000853	869050400	ÜLKER ÇİZİVİÇ 82 GR
İSTANBUL HAR	20150331	0000000534	212975000	ÇİLEK ANTALYA

[21] Lamb Charles W, Hair Joe F, & McDaniel Carl. (2011). *Essentials of marketing*: Cengage Learning.

C. Experiments and Results

A contingency table in which the number of stores placed in each cluster is needed for comparison of two clusterings. As shown in the table below, the first clustering have six clusters while the second clustering have five. The number of stores occurred in the first cluster of both two clusterings are 18. The number of stores are two which occurred in the first cluster of first clustering and second cluster of the second clustering. The number of stores are 13 which occurred in the first cluster of the first clustering and third cluster of the second clustering and so on.

Table 5: Contingency Table of Stores

Clusters	**1**	**2**	**3**	**4**	**5**	**6**	*Total*
1	18	6	12	14	3	4	57
2	2	0	3	3	0	4	12
3	13	5	15	11	11	9	64
4	6	5	7	15	3	1	37
5	1	0	1	2	1	0	5
Total	40	16	38	45	18	18	**175**

Value of Rand Index, Adjusted Rand Index, Fowlkes Mallows Index and Jaccard Index are calculated in both MS Excel and R programming language. *Arandi* function in *mclust* package is used for Rand Index and Adjusted Rand Index. While the Rand Index has a value of 0.63, the Adjusted Rand Index has 0.0043. As mentioned before, it is usual that the Adjusted Rand Index can take wider range of values. [19]

AdjustedRand function in *clues* package is used for calculating Rand Index again and also Jaccard Index and Fowlkes Mallows Index. Respectively 0.63, 0.13 and 0.24 are found. Furthermore another function *FM_index* in *dendextend* package calculated the Fowlkes Mallows Index, its expected value and its variance respectively 0.24, 0.23 and 0.0000493.

Conclusions and Discussions

There are many researches that conclude socio-economic factors of consumers affects their purchasing behaviors. In this study we followed this idea and implemented an application for comparing clusterings issue. 175 stores of a retailer have been segmented into six groups with socio-economic factors of customers. 175 stores segmented again into five groups with purchasing behaviors of customers this time. This study expected to find a strong agreement between two clustering approaches.

According to the results of comparing clustering measures, Rand Index value (0.63) indicates that two clusterings (socioeconomic segmentation and purchasing behavior based segmentation) are in agreement. However the Adjusted Rand Index value (0.0043) which is corrected for chance of Rand Index implies that there is not any agreement between two clusterings. Jaccard Index value (0.13) and Fowlkes Mallows Index value (0.23) shows that even it is weak, there is an agreement between two clusterings as this study expects.

There are some problems about counting pairs methods mentioned in the literature and there are no guidelines for their best application scenarios.[12,22] Furthermore the properties of the techniques such as variance of ARI are still being researched.[23]

When the value of all measures are considered one can conclude that there is not a strong evidence for agreement of two clusterings. It seems that finding the relationship between socio-economic factors and purchasing behaviors is not possible with just one day transaction data since consumers may not reflect their socio-economic status that day. Thus, by increasing the quality of the data i.e. increasing the number and quality of input variables of each clustering, there is a clue in hand that results would be more satisfying.

✵ ✵ ✵

[22] [16]Romano Simone, Vinh Nguyen Xuan, Bailey James, & Verspoor Karin. (2015). Adjusting for Chance Clustering Comparison Measures. *arXiv preprint arXiv:1512.01286*.

[23] [17]Steinley Douglas, Brusco Michael J, & Hubert Lawrence. (2016). The variance of the adjusted Rand index. *Psychological methods, 21*, 261.

REFERENCES

Aggarwal Charu C. (2015). *Data mining: The textbook*. Switzerland: Springer.

Agrawal Rakesh, & Srikant Ramakrishnan. (1994). Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB* (Vol. 1215, pp. 487-499).

Bilgic Emrah. (2016). *Retail store segmentation with rule based and socioeconomic approaches: An application on a chain company*. Unprinted PhD dissertation. Istanbul: Marmara University

Fowlkes Edward B, & Mallows Colin L. (1983). A method for comparing two hierarchical clusterings. *Journal of the American Statistical association, 78*, 553-569.

Hennig Christian, & Liao Tim F. (2013). How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 62*, 309-369.

Hubert Lawrence, & Arabie Phipps. (1985). Comparing partitions. *Journal of Classification, 2*, 193-218.

Jaccard Paul. (1908). *Nouvelles recherches sur la distribution florale*.

Jain A. K., Murty M. N., & Flynn P. J. (1999). Data clustering: a review. *ACM Computing Surveys, 31*, 264-323.

Jain Anil K, & Dubes Richard C. (1988). *Algorithms for clustering data*: Prentice-Hall, Inc.

Kotler Philip, & Gary Armstrong. (2012). *Principles of marketing* (Vol. 14th ed.). New Jersey: Prentice Hall.

Lamb Charles W, Hair Joe F, & McDaniel Carl. (2011). *Essentials of marketing*: Cengage Learning.

Meilă Marina. (2007). Comparing clusterings—an information based distance. *Journal of multivariate analysis, 98*, 873-895.

Rabbany Reihaneh, & Zaïane Osmar R. (2015). Generalization of clustering agreements and distances for overlapping clusters and network communities. *Data mining and knowledge discovery, 29*, 1458-1485.

Rand William M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association, 66*, 846-850.

Romano Simone, Bailey James, Nguyen Xuan Vinh, & Verspoor Karin. (2014). Standardized Mutual Information for Clustering Comparisons: One Step Further in Adjustment for Chance. In *ICML* (pp. 1143-1151).

Romano Simone, Vinh Nguyen Xuan, Bailey James, & Verspoor Karin. (2015). Adjusting for Chance Clustering Comparison Measures. *arXiv preprint arXiv:1512.01286*.

Steinley Douglas, Brusco Michael J, & Hubert Lawrence. (2016). The variance of the adjusted Rand index. *Psychological methods, 21*, 261.

Vinh Nguyen Xuan, Epps Julien, & Bailey James. (2009). Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *Proceedings of the 26th Annual International Conference on Machine Learning* (pp. 1073-1080): ACM.

Wagner Silke, & Wagner Dorothea. (2007). *Comparing clusterings: an overview*: Universität Karlsruhe, Fakultät für Informatik Karlsruhe.

Ward Jr Joe H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical association, 58*, 236-244.

Xiang Qiaoliang, Mao Qi, Chai Kian Ming, Chieu Hai Leong, Tsang Ivor, & Zhao Zhendong. (2012). A Split-Merge Framework for Comparing Clusterings. *arXiv preprint arXiv:1206.6475*.

Yeung Ka Yee, & Ruzzo Walter L. (2001). Details of the adjusted Rand index and clustering algorithms, supplement to the paper "An empirical study on principal component analysis for clustering gene expression data". *Bioinformatics, 17*, 763-774.

Zaki Mohammed J, & Meira Jr Wagner. (2014). *Data mining and analysis: fundamental concepts and algorithms*: Cambridge University Press.

✵ ✵ ✵

Erciyes Üniversitesi Sosyal Bilimler Enstitüsü Dergisi XLIV, 2018/1, 54-57
Arrival Date: 21.12.2017, Publishing Date: 30.06.2018

KÜMELEMELERİN KARŞILAŞTIRILMASI: BİR MAĞAZA SEGMENTASYONU UYGULAMASI[*]

Emrah BİLGİÇ[a] Özgür ÇAKIR[b]

Öz

Kümeleme analizi için birbirinden farklı birçok algoritma vardır ve her biri farklı farklı kümelerin, dolayısıyla farklı sonuçların oluşmasını sağlamaktadır. Bu sebeple kümelemenin uygunluğunu, performansını ve ortaya çıkardığı sonuçları iyi değerlendirmek gerekmektedir. Kümeleme sonucunda ortaya çıkan gruplar bir uzman tarafından incelenebilir veya bir kümelemenin geçerliliğini ölçebilecek ölçütler kullanılabilir. Genel olarak kullanılan ölçütler Dışsal Ölçütler, İçsel ölçütler ve Göreceli Ölçütler olarak üçe ayrılırlar.

Dışsal Ölçütlerde bir kümeleme algoritmasının sonuçları daha önceden belirli olan veya insanlar tarafından bir şekilde bilinen veya tahmin edilen sonuçlarla karşılaştırılır. Bu ölçüt için en yaygın kullanılan ölçüler Rand'ın İndeksi, Jaccard'ın Katsayısı, Fowlkes-Mallows'un İndeksi, Hubert'in İstatistiği, Ball İndeksi, Duda İndeksi, Gap İndeksi, Pseudo T^2 İndeksi, Tau İndeksi ve Dunn İndeksi'dir.

İçsel Ölçütlerde ise bir algoritma tarafından üretilen kümelemenin yapısı sadece veri kümesinden intikal eden bazı özellikler ve nicelikler kullanılarak değerlendirilir. Hiyerarşik bir kümelemede oluşan hiyerarşiler bu duruma örnek olarak verilebilir. Bu ölçütte en çok kullanılan ölçüler ise Kophenetik

[*] Yazarın "Kural Bazlı ve Sosyoekonomik Yaklaşımla Mağaza Segmentasyonu: Bir Süpermarket Zincirinde Uygulama" isimli doktora tezinden türetilmiştir.

[a] Dr. Öğr. Üyesi, Muş Alparslan Üniversitesi, e.bilgic@alparslan.edu.tr

[b] Doç. Dr., Marmara Üniversitesi, ocakir@marmara.edu.tr

Korelasyon Katsayısı, İstatistik Önem Testleri, Çapraz Geçerlilik Yöntemleri ve Birleştirici Katsayısıdır (Agglomerative Coefficient).

Göreceli Ölçütler üçüncü ve son olarak kümeleme sonuçlarının değerlendirilmesi konusunda karşımıza çıkmaktadır. Bu ölçütteki temel fikir bir kümeleme yapısının aynı algoritmayla fakat başka parametrelerle üretilmiş olan kümeleme yapısıyla karşılaştırılmasıdır. Bu noktada akla ilk gelen örnek, hiyerarşik kümelemede karşımıza çıkan Tek, Tam ve Ortalama gibi bağlantı yöntemleridir.

Literatürde kümelemenin geçerliliği başlığı altında yer alan dışsal ölçütler aynı zamanda kümelemelerin karşılaştırılması başlığı altında da incelenmektedir.

İki kümeleme sonuçlarının karşılaştırılabilmesi için her iki kümelemede elde edilen küme sayısının eşit olması zorunluluğu yoktur. Kümelemelerin karşılaştırılması konusunda yapılan çalışmalar incelendiğinde genel amaç aynı veri kümesine ait farklı algoritmaların performanslarının karşılaştırılması iken, bu karşılaştırmada kullanılan ölçülerin farklı veri kümelerinin kümelenmelerini de karşılaştırabiliyor olması beklenebilir.

Her ne kadar çalışmamızdaki iki farklı segmentasyon yaklaşımı için kullanılacak kümelemelerin girdi değişkenleri farklı olsa da, her iki yaklaşım için de kümelenecek olan veri noktaları aynı olan 175 mağazadır. "İki kümeleme yönteminde veri noktası çiftleri eğer her iki kümeleme yönteminde de aynı kümede yer almışsa veya her iki yöntemde de farklı kümelerde yer almışsa bu kümelemeler birbirine benzerdir" ana fikrini savunan Çiftleri Sayma Yöntemlerine göre, çalışmamızda kullanılacak olan kümeleme yöntemlerinin benzerlik göstermesi beklenmektedir. Çünkü ilk segmentasyon yaklaşımında kümeleme girdisi olarak kullanılan sosyoekonomik faktörlerin ikinci segmentasyon yaklaşımında kümeleme girdisi olarak kullanılan satın alma davranışlarını etkilediği düşünülmektedir. Dolayısıyla mağazaların bir kısmının her iki yaklaşımda da aynı kümelerde veya her iki yaklaşımda da farklı kümelerde yer alması beklenmektedir. Bunun tespiti için Çiftleri Sayma yöntemi kullanılabilir.

Çiftleri Sayma yönteminde yapılan iş, her iki kümelemede de aynı yolla sınıflanmış veri noktası çiftlerinin sayısının hesaplanmasıdır. Her iki kümelemede de veri çiftlerinin aynı kümede yer almaları veya farklı kümelerde yer almış olmaları hesaplanır. Vurgulandığı gibi, Rand İndeksi bu hesaplamayı yapabilmek için önerilmiş bir indeks olup, Rand'ın bu çalışmasından sonra indeksin değerini şanstan arındırmak amacıyla araştırmacılar tarafından birçok farklı düzeltme faktörü de önerilmiştir.

Kümeleme sonuçlarının karşılaştırılması için kullanılan ölçülerden olan çiftleri sayma yöntemindeki indekslerin hesaplanması Microsoft Excel programında da yapılabilmektedir. İhtiyaç olunan, her iki kümeleme yaklaşımında mağazaların hangi kümelerde yer aldığının sayılması ile hazırlanan kontenjans (eşleştirme) tablosudur.

Çalışmamızda bir perakendecinin İstanbul'daki 175 mağazası, bahsedildiği gibi iki farklı yaklaşımla segmentlere ayrılmıştır. İlk yaklaşımda kullanılan değişkenler sosyoekonomik değişkenlerdir. Her bir mağazanın bulunduğu konumdaki sosyoekonomik göstergeler (bölgede yaşayan halkın yaş dağılımı, eğitim seviyesi, medeni hali, bölgedeki konutların ortalama kirası gibi) çeşitli kaynaklardan elde edilmiştir. Bu yaklaşımla mağazalar 5 segmente ayrılmıştır. İkinci yaklaşımda ise çalışmamızda söz konusu olan perakende şirketinden her mağazaya ait birer günlük satış verileri alınmış ve mağazalar birlikte satılan ürünlere, başka bir deyişle müşterilerinin satın alma davranışlarına göre 6 segmente ayrılmıştır. Her iki kümelemede kümelere düşen mağaza sayıları aşağıdaki kontenjans tablosunda verilmiştir. Örneğin birinci segmentasyon yaklaşımı ile birinci kümede/segmentte ve ikinci segmentasyon yaklaşımı ile de birinci segmentte yer alan mağaza sayısı 18'dir.

	1	**2**	**3**	**4**	**5**	**6**	Toplam
1	18	6	12	14	3	4	57
2	2	0	3	3	0	4	12
3	13	5	15	11	11	9	64
4	6	5	7	15	3	1	37
5	1	0	1	2	1	0	5
Toplam	40	16	38	45	18	18	**175**

Bu tablo yardımıyla, R programlama dilinde yer alan "Clues" paketindeki "adjustedRand" işlevi sayesinde Rand İndeksi değeri 0,63 hesaplanmışken, Düzeltilmiş Rand İndeksi ise 0,0043 bulunmuştur. Ayrıca Jaccard İndeksi ve Fowlkes-Mallows İndekslerinin ise sırasıyla 0,13 ile 0,235 değerlerini hesaplandığı görülmüştür. Bahsedilen indeksler 0 ile 1 arasında değer almaktadır. Ayrıca bu değerler Microsoft Excel yardımıyla da hesaplanabilir.

Fowlkes-Mallows değerini hesaplamak için hazırlanan başka bir paket olan "dendextend" paketindeki "FM_index" işlevi sayesinde indeks değeri 0,2367 olarak tekrar hesaplanmış ayrıca indeksin beklenen değeri ve varyansı da sırasıyla 0,2333 ile 0,0000493 olarak hesaplanmıştır.

Çalışmamızın sonuçlarına göre Rand İndeksi, iki yaklaşımımız arasında orta derecede kuvvetli bir uyum olduğuna işaret etmektedir. Ancak doğrudan Rand indeksi ile yapılacak yorum yanıltıcı olacağından şanstan arındırılmış uyumu veren Düzeltilmiş Rand indeksini ölçü almak daha uygun olacaktır. Bu ölçüye göre iki kümeleme yaklaşımımızın sonuçları arasında bir uyumdan söz etmek oldukça güçtür.

Fowlkes-Mallows indeksi dikkate alınacak olursa hesaplanan beklenen değer ve varyans doğrultusunda iki kümeleme yaklaşımımız sonuçları arasında zayıf da olsa bir uyumdan söz etmek mümkündür.

Tüm ölçütlerden elde edilen sonuçları genelleştirecek olursak kullanmış olduğumuz veri kümesi üzerinde uygulanan iki farklı yaklaşımın uyum içinde olduğunu iddia edebilmek için yeterli delil olmadığını söyleyebiliriz. Ancak çalışmada veri kalitesini daha üst seviyelere taşıyarak ya da her iki yaklaşımda da girdi değişkenlerinin sayısını ve niteliğini arttırarak yeterli delil elde edilebileceğine dair ipuçları söz konusudur.

Anahtar Kelimeler: Kümelemelerin Karşılaştırılması, Kümelemelerde Görüş Birliği, Mağaza Segmentasyonu.

✵ ✵ ✵