International Journal of Electrical and Computer Engineering (IJECE)
Vol. 7, No. 6, December 2017, pp. 3692~3699
ISSN: 2088-8708, DOI: 10.11591/ijece.v7i6.pp3692-3699

Clustering in Aggregated User Profiles across Multiple Social Networks

Charu Virmani[1], Anuradha Pillai[2], Dimple Juneja[3]
[1]Research Scholar, Ymca University of Science and Technology, Faridabad
[2]Department of Computer Science, Ymca University of Science and Technology, Faridabad
[3]Department of Computer Application, National Institute of Technology, Kurukshetra

Article Info

Article history:

Received Apr 18, 2017
Revised May 30, 2017
Accepted Jun 15, 2017

Keyword:

Clustering,
Ensemble cluster
K-Means
Social network

ABSTRACT

A social network is indeed an abstraction of related groups interacting amongst themselves to develop relationships. However, toanalyze any relationships and psychology behind it, clustering plays a vital role. Clustering enhances the predictability and discoveryof like mindedness amongst users. This article's goal exploits the technique of Ensemble K-means clusters to extract the entities and their corresponding interestsas per the skills and location by aggregating user profiles across the multiple online social networks. The proposed ensemble clustering utilizes known K-means algorithm to improve results for the aggregated user profiles across multiple social networks. The approach produces an ensemble similarity measure and provides 70% better results than taking a fixed value of K or guessing a value of K while not altering the clustering method. This paper states that good ensembles clusters can be spawned to envisage the discoverability of a user for a particular interest.

Corresponding Author:

Charu Virmani,
Associate Professor
Deparment of computer Science and Engineering
Faculty of Engineering and Technology
Manav Rachna International University
Email: charu.fet@mriu.edu.in

1. INTRODUCTION

As the number of social network users increases, a tremendous amount of data is generated by the sharing of information. The intuitive nature of these social networks is the creation of related groups (or clusters) [1]. This has become an area of interest in the discovery of communities in recent times. These patterns are used to mine a variety of information, which is then used in various fields [2]. Cluster analysis, or clustering in a social network context, is the grouping of a set of data objects (for example, friends, connections, communities, or personal information) in such a way that objects in the same group (or clusters) are more similar to each other than to those in other groups (or clusters). The identification of these patterns into clusters has numerous applications in the field of data science.

Many algorithms can be used to cluster data [3]. Populaar clusters include groups with small distances between cluster members, dense areas of the data space, intervals, or particular statistical distribution [4]. Therefore, clustering can be formulated as a multi-objective optimization problem. A suitable clustering algorithm and parameter settings vary from the individual input and expected results. Numerous attempts were made to improve the quality of clusters using ensembling techniques [5] [6] [7] [8] [9] [10] [11] [12]. The main concern of many of these algorithms is to elucidate label correspondence problem. The limitation of many of these algorithms is the assumption of the same number of cluster in each partition and may perform poorly when the information about output cluster is not known in advance.

Literature pertaining to clustering on aggregated publicly-available user profile data of various social networks was thoroughly dwelled and it was discovered that k-means algorithm and ensemble clustering are the most popular algorithms to cluster the data to obtain results The study thus aims to apply k-means clustering on aggregated social network data and ensemble clusters thus formed by grouping different parameters and interpret results. In this work, the partitions are generated with varying number of clusters and thus improving the quality and stability of the consensus partition. Good ensemble cluster are provided by eliminating the dependency of the input parameter like k, the number of clusters. Hungarian algorithm [13] and cumulative voting scheme [14] are used to obtain final clusters. The paper offers two-fold contribution i.e. identifying the skill of a user for particular location across multiple social networks and eliminating the dependency of input parameter like K. The current work uniquely contributes to the limitation of the requirement of equal number of cluster in input partition and the knowledge of the number of clusters to be known in advance.

This paper is structured into five sections: Section 2 throws light on the work of eminent researchers highlighting their substantial contributions. The discussion in section 2 indicates the limitations of k means algorithm. The current work thus finds motivation and resolves the challenge listed above. Section 3 uniquely contributes an ensemble cluster to identify groups of clusters on a measure of similarity. This has been established with a data set in the evaluation section given in section 4. Section 5 finally concludes.

2. RELATED WORK

Traditionally, social network clustering is either hierarchical or partitioning where vertices join into groups of similarity [15]. Community detection in social networks has been an interest for which a successful algorithm is depicted in [16] [17] [18] [19]. As one of the simplest unsupervised clustering techniques, k-means discovers the degree of similarity among k groups assuming k centroids. K-centers are defined and placed spatially as far as possible. Each spatial point is marked to a given data set and associated to the nearest center. New centroids are calculated as barycenter of the clusters and rebounded between same data set points to the nearest new center. Thus, k centers change its location aiming at minimizing an objective function known as squared error function [] by:

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} (||x_{i-} v_j||)^2 \quad (1)$$

where

||xi - vj|| is the Euclidean distance between xi and vj.

ci is the number of data points in ith cluster.

c is the number of cluster centers.

The emerging field of social analysis uses data mining as the key input for analyzing data. Clustering is an important factor in this analysis. It is approached by various clustering algorithms, including: k-means, fuzzy c-mean, and table modeling [20] [21]. While k-means is very fast, its center value depends on the value of k. Different values of k will result in different clusters [22] [23]. Yang et al [24] observed that the K-means learning algorithm requires specification of the number of cluster centers. If two highly-overlapping data exist, then k-means will not be able to resolve the presence of two clusters and also it is not invariant to non-linear transformations.

Zhang et al [4] proposed the mapping of network nodes to identify the overlapping community by Euclidean space and fuzzy c-means clustering. Many researchers have sought community in social networks, as well as proposed metrics for evaluating the structure [25] [26] [27] [28]. Yang et al [24] proposed finding people by using mobile phone usage patterns in a social network. Another researcher proposed a hybrid study to retain customers using clustering [28]. The authors used aggregated data on user profiles from various social networks. With variance clustering, they used k-means and ensemble clustering to group users as per their public information. The study was restricted to cluster the user of a location who has interest in a specific skill. Businesses and marketing strategies can also use this technique for promotional benefits by applying it to other attributes to find user similarities.

Numerous techniques for generating cluster results and combining them have been seen in literature [5] [6] [7] [8] [9] [10] [11] [12]. Generation of input partition followed by integration of all the partitions to obtain final partition is a two way process given by vega-pons et al. [29]. Median partition and object co-occurrence are thw two ways to generate a consensus.In median partition, the final partition maximizes the similarity with all the generated set in the ensemble. This approach is not considered for clustering as defining the Mirkin Distance [30] have been proven NP-hard and computationally expensive. Object-cooccurrence is another approach that obtains the final partition from the generation set depending upon the frequency of occurrence of object together or an object to one cluster followed by similarity based clustering

algorithm. Co-association Matrix followed by clustering mechanism is a way to generate the occurrence of an object. Relabelling and cummulative voting is another choice for attaining the final partition from the generation set depending upon the frequency of occurrence of objects. Relabelling solve label correspondence problem using Hungarian Algorithm [13] following voting process by using cumulative voting [14] to obtain final partition. Other final partitions can be obtained by Genetic algorithm [30], NMF [31] and kernel Method [32] under object co-occurrence that is beyond the consideration of this paper.

It has been observed during the research that no work has been devoted to applying ensemble clustering methods in analyzing a user's publicly available information. However, different strategies have been utilized to recognize community and merge community structures [33]. As data clustering and community detection are very comparative, it ought to be conceivable to merge community in an indistinguishable way from ensembles of clusters with great outcomes. The proposed algorithm performed clustering on aggregated user profiles from various social networks by changing the value of *k* for different parameters. Then, partitions were combined to overcome cluster instability.

3. PROPOSED WORK

A people group or community is a subset of hubs inside a system such that associations between hubs in the subset are denser than associations with rest of the system. Detecting a community is a form of clustering of the information which is similar among neighbors. The aim of this section is to propose method for combining several clusters and generalize this for the user's information. The proposed strategy creates a new feature space utilizing the yields of initial k means algorithm. The phases of the proposed methodology are:

1. Generate Initial clusters using K-means for varying value of k.
2. Generate new components by Hungarian algorithm.
3. Ensemble final clusters on the new generated components.

Unsupervised training is used to partition data on the basis of similarity using k-means. More similar users are grouped into a cluster using Euclidean distance in this technique across all the profiles aggregated by the network. This results in a cluster belonging to a particular location. A particular skill will be found and applied for that location. However, a weighted Euclidean distance is used to cluster the data of more similar belonging to location and skill. A weight was assigned to one parameter and group; the user was assigned based on a different parameter. For mining the skill from the user-generated post, the post extracted is cleaned and converted into a key pair. The pair includes the post ID (or user name) and the post's list of words serving as the skills list that the user applies in the post. The list is converted into a numerical vector; weights are determined using soft TF-IDF.

K-means clustering models are applied on the converted list where k = 3 to 12 for skill and by-variance clusters for skill and location to generate input partitons.These techniques are applied separately on the different variables, thus resulting partitions into different number of clusters. The results of clusters are then combined using Hungarian algorithm and cumulative voting for each cluster. Hungarian algorithm is a multi-objective clustering comprising of multiple clustering partitions with objective functions. It ensembles multiple partitions by combining individual clustering partition and giving a final partition. Final partitions of clusters can be found by applying the voting scheme [16]. Confusion matrix is used to compoute the similarity between clusters. To compute the confusion matrix of two different number of cluster, the remaining cluster of the smaller number of cluster will be kept as empty. Confusion matrix for two clusters (A,B) is of size AxB. The (i,j) th index of the matrix corresponds to the object that are in cluster i of A and in cluster j of B. Maximum element is selected using Hungarian Algorithm. Integration of Element is done by aggregating the aligned partitions by selecting the element that takes the majority cluster label for each observed partition.Majority Voting and plurality voting are the mathods to generate the final clusters that involves selecting an object whose count is greater than a threshold value whereas plurality voting considers the majority cluster label for each observed value.The proposed algorithm is shoum in Algorithm 1.

Algorithm 1

1. Pass the entire dataset and identify the point with the weight assigned to it.
2. Compare the objects and consider it as per *k* (k = 3 to 12).
3. Check the similarity and calculate the mean value from each centroid to the cluster for the object.
4. Each object may reside in the cluster it wins the similarity.
5. Repeat steps 2 to 4 if there is no change.
6. Repeat step for another value of k until K=12
7. Compute confusion matrix based on multiple data partitions from step 5.
8. Find its maximum element, associate the two cluster as per the maximum object. Thus, reduce the matrix upon removal of these clusters.

Error rate, Jacard Index and RAND index are considered to measure the quality of clusters. Error rate depicts the average number of misclassified elements. Partitions are more similar if the error rate is less. Error rate is used to validate the accuracy of the final partition. RAND [34] proposes a measure to validate the quality of the cluster as:

$$r(A, B) = \frac{x+y}{x+y+z+w} \quad (2)$$

Where:

U: set of n clusters
A: partition in U having r subsets
B: partition in U having q subsets
x: number of pair of elements from U which occur in A and B
y: number of pairs of elements from U which are different in A and B
z: number of pair of elements from U which occur in A but not in B
w: number of pair of elements from U which occur in B but not in A

The Jacard index [35] to measure the similarityis computed as:

$$J(A, B) = \frac{y}{y+z+w} \quad (3)$$

4. RESULT AND DISCUSSION

Various social networks are crawled to create raw data on user profile information, including: name, description, location, interests and tweets/news feed. The collected data was aggregated on the vector<User ID, name>. This set of raw data is created in MongoDB. To create enriched data, the data was cleaned for noise removal and stored in the json documents. The proposed clustering algorithm was applied to the data to create desired clusters. Figure 1 shows the architecture for visualizing user information.

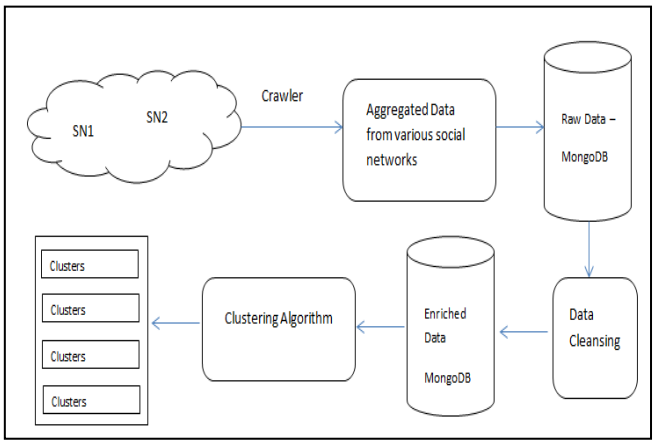

Figure 1. Architecture

Twitter public search and Bing search API acts as the source of data collection. While Twitter search outputs relevant user-generated posts when searched with an input query. The Bing search API allowed creating the mixed inputs of user-variables. For example, user-name + user-location + user-gender + user-description-keywords. This user information is used to extract information from other social networks like Full Connect, Google, and Bing by crawling or using api's of the respective networks. Total 27,956 user profiles extracted; complete data consisted of 45,899 user-generated posts. The data is cleaned i.e. white spaces, stopwords, and common terms (i.e., a, an, and the) are removed and converted into lowercase. User profiles were aggregated by matching user ID and name (public attributes using Jaro-wrinkler). Out of 27,956 user profiles, 18,897 user profiles are aggregated. The complete data statistics is shown in table 1 and the pseudo code to aggregate the profile is depicted in algorithm 2.

Table 1. Statistics of input data

# Input Queries	12
# Raw Documents	27,956
# unique users – Twitter	15,530
# users - Enriched Profile	18,897
# search engine total links	56,896
# search engine user links	21,674

Algorithm 2: Profile Aggregation

1. Initialize Doc1 <- Source1 Raw Document
2. Initialize Doc2 <- Source2 Raw Document
3. Initialize DocN<- SourceN Raw Document
4. Initialize Pairs <- cartesian_pairs of all documents
*a. Pairs <- N*N documents*
5. Iterate in every Pair
a. Rel_var1 <- one of the relevant variable ex – name
b.threshold_score<-Jaro_wrinkler(rel_var1, pair)
c. If score >threshold_score: merge_enrich(rel_var1, pair)
d. else : pass & ignore
6. Update for every pair
a. Pick or replace rel_var values accn to priority.

4.1. Skill Wise Clusters of Keywords by Users

The system has chosen value of k varying from 3 to 12 to generate the partitions, first experiment is carried by passing value of k as 3 resulting in three clusters for each of the 12 queries: Node, NLP, Java, machine learning, database, Python, javascript, big data, deep learning, SQL, Hadoop and Datascience. These models identify repeating patterns in data and organize them into buckets known or "data clusters" and are depicted in table 2. Similar results are obtained from k-mean clustering varying k from 4 to 12. Hence, the similar results are omitted.

Table 2. K-means clusters for k = 3 for the three skills

database 2210
Top terms per cluster: database
Cluster 0: job administrator sql hire database server derby oracle dba disk
Cluster 1: http tungsten dac useful ejnetwork online delete 8i load server
Cluster 2: database sql look dbmnosql 9i sanction opm expect db2

javascript 22446
Top terms per cluster: javascript
Cluster 0: javascriptinspirateebooknjavascriptkomopensourcedisponibleesta
Cluster 1: javascript developer devops job library jquery know use linux design
Cluster 2: ncertificationdmoz webmaster leazysunnyphpjavascriptjavascriptdfranformvalidation

datascience 3636
Top terms per cluster: datascience
Cluster 0: datascience data bigdatamachinelearning analytics iot python business statistic learn
Cluster 1: bigdata cancer beat use artificialintelligencedeeplearningdatascienceiotchatbotfintech
Cluster 2: ronaldvanloon learn machine team mix expert right engineer know

For input queries, user's information is collected and differentiated on the basis of interest and location. Data was collected for three different locations United Kingdom, United States and London. It was analyzed on the basis of java, nlp, Python, javascript, etc. Different parameters are analyzed to the model via k-means clustering on the data set (documents related to user-skills and user-level variables such as location, descriptions, etc.).

In order to identify that the user of a particular location has a particular skill, an approach must be found to identify the skill set of the user of the particular location. The particular location cluster can be created through the k-means algorithm because of its quick convergence to similarity. The skill cluster should define the boundaries of the skill set; this ends in a complex task. To obtain the skill set of the user, one needs to know the interest from the interest attribute (if available from the social network), as well as the user-generated post to mine information for the particular skill. In this study, clusters were obtained for k = 3 to 12 on skill wise user public data collected from various social networks. *K* partitions are generated optimally representing *M* partitions by voting scheme to generate a skilled public group for that particular location. The detailed algorithm is shown in algorithm 1.

Input partitions to the confusion matrix are the clusters obtained from the previously discussed k-means (i.e., k = 3 to 12). In this phase, the clustering results are combined and best clusters by is chosen by computing similarity measure using confusion matrix and voting scheme. Figure 2 represents number of user specialized in skill for different location analyzed from the partitions. Table 3 shows the top five terms of each cluster by combining the results for a particular location London.

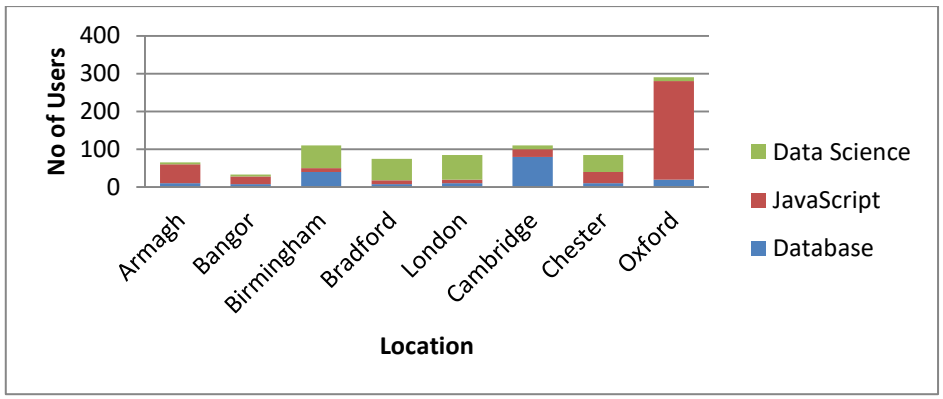

Figure 2. Count of user for different skills for differentlocation

Table 3. Top five clusters

DataScience	Javascript	Database
machinelearning	Jquery	Database
datascience	Formvalidation	Nosql
Bigdata	Nodejs	Sql
deeplearning	Library	Mongodb
analytics	Reactjs	Pymongo

It has been observed that the results produced by ensemble clustering is 70% better than that can be produced by guessing value of k or taking a fixed value of k.Th e complexity of k-means is O(KNId) where k belongs to number of clusters, N belongs to number of samples, I belongs to iterations of k means to converge and d belongto number of components. The complexity of proposed ensemble cluster is $O(k^3)$. The comparison between K-means and ensemble K-means clustering was evaluated using error rate, Jacard index and RAND score which is an extent to evaluate cluster quality as shown in table 4.

Table 4. similarity between k-means and Ensemble K-means

Data set	Method	Error rate	Jacard Index	RAND score
Aggregated user's public information	K-means	45	0.49	0.68
Aggregated user's public information	Ensemble K-means	15	0.97	0.95

5. CONCLUSION AND SCOPE OF FURTHER RESEARCH

This article analyzed two clustering algorithms in the context of clustering social network data when collected from different social networks. It has been reported that it is possible to detect community using ensemble t. This paper proves that the ensemble K-means clustering produces better results in term of error rate, RAND score and Jacard index. This opens up the scope of further research in regards to efficient use for business and marketing strategies

REFERENCES

[1] Otte, E., & Rousseau, R. (2002). Social network analysis: A powerful strategy, also for the information sciences. *Journal of information Science*, 28*(*6), 441-453.

[2] Krebs, V. E. (2002, April). Uncloaking terrorist networks. *First Monday*, 7(4).

[3] Estivill-Castro, V. (2002). Why so many clustering algorithms: A position paper. *SIGKDD Explorations Newsletter*, 4(1), 65-75.

[4] Zhang, S., Wang, R. S., & Zhang, X. S. (2007). Identification of overlapping community structure in complex networks using fuzzy c-means clustering. *Physica A: Statistical Mechanics and its Applications*, 374(1), 483-490.

[5] Strehl, A., & Ghosh, J. (2002). Cluster ensembles---a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research*, 3(Dec), 583-617.

[6] Topchy, A., Minaei-Bidgoli, B., Jain, A. K., & Punch, W. F. (2004, August). Adaptive clustering ensembles. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on* (Vol. 1, pp. 272-275). IEEE.

[7] Topchy, A., Jain, A. K., & Punch, W. (2003, November). Combining multiple weak clusterings. In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on* (pp. 331-338). IEEE.

[8] Topchy, A., Jain, A. K., & Punch, W. (2004, April). A mixture model for clustering ensembles. In *Proceedings of the 2004 SIAM International Conference on Data Mining* (pp. 379-390). Society for Industrial and Applied Mathematics.

[9] Ayad, H. G., & Kamel, M. S. (2008). Cumulative voting consensus method for partitions with variable number of clusters. *IEEE transactions on pattern analysis and machine intelligence*, 30(1), 160-173.

[10] Singh, V., Mukherjee, L., Peng, J., &Xu, J. (2010). Ensemble clustering using semidefinite programming with applications. *Machine learning*, 79(1-2), 177-200.

[11] Bhatnagar, V., &Ahuja, S. (2010, July). Robust clustering using discriminant analysis. In *Industrial Conference on Data Mining* (pp. 143-157). Springer Berlin Heidelberg.

[12] Dudoit, S., &Fridlyand, J. (2003). Bagging to improve the accuracy of a clustering procedure. *Bioinformatics*, 19(9), 1090-1099.

[13] Mahesh, O., &Srinivasan, G. (2002). Incremental cell formation considering alternative machines. *International Journal of Production Research*, 40(14), 3291-3310.

[14] Dimitriadou, E., Weingessel, A., &Hornik, K. (2001, August). Voting-merging: An ensemble method for clustering. In *International Conference on Artificial Neural Networks* (pp. 217-224). Springer Berlin Heidelberg.

[15] Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3), 75-174.

[16] Clauset, A., Newman, M. E., & Moore, C. (2004). Finding community structure in very large networks. *Physical Review E*, 70(6), 066111.

[17] Newman, M. E., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, 69(2), 026113.

[18] Lancichinetti, A., &Fortunato, S. (2009). Community detection algorithms: A comparative analysis. *Physical review E*, 80(5), 056117.

[19] Palla, G., Derényi, I., Farkas, I., & Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043), 814-818.

[20] Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques* (3rded.). The Netherlands: Morgan Kaufmann.

[21] Orme, B., & Johnson, R. (2008). *Improving k-means cluster analysis: Ensemble analysis instead of highest reproducibility replicates* (Sawtooth Software Research Paper Series). Sequim, WA: Sawtooth Software, Inc.

[22] Sahu, M., Parvathi, K., & Krishna, M. V. (2017). Parametric Comparison of K-means and Adaptive K-means Clustering Performance on Different Images. *International Journal of Electrical and Computer Engineering (IJECE)*, 7(2).

[23] Sahu, M., Parvathi, K., & Krishna, M. V. (2017). Parametric Comparison of K-means and Adaptive K-means Clustering Performance on Different Images. *International Journal of Electrical and Computer Engineering (IJECE)*, 7(2).

[24] Yang, X., Wang, Y., Wu, D., & Ma, A. (2010, November). K-means based clustering on mobile usage for social network analysis purpose. In *2010 6th International Conference on Advanced Information Management and Service (IMS)* (pp. 223-228). IEEE.

[25] Oleiwi, W. K. (2016). Using the Fuzzy Logic to Find Optimal Centers of Clusters of K-means. *International Journal of Electrical and Computer Engineering*, 6(6), 3068.

[26] Jain, A. K., & Dubes, R. C. (1988). *Algorithms for clustering data*. Prentice-Hall.

[27] Steinhaeuser, K., &Chawla, N. V. (2010). Identifying and evaluating community structure in complex networks. *Pattern Recognition Letters*, 31(5), 413-421.
[28] Li, X., Huang, Y., Li, S., & Zhang, Y. (2011, May). Hybrid retention strategy formulation in telecom based on k-means clustering analysis. In *2011 International Conference on E-Business and E-Government (ICEE)* (pp. 1-4). IEEE.Vega-Pons, S., Correa-Morris, J., & Ruiz-Shulcloper, J. (2008, September). Weighted cluster ensemble using a kernel consensus function. In *Iberoamerican Congress on Pattern Recognition* (pp. 195-202). Springer Berlin Heidelberg.
[29] Mirkin, B. (1996). Mathematical Classification and Clustering, *Nonconvex Optimization and Its Applications*, Volume 11, Pardalos, P. and Horst, R., editors.
[30] Yoon, H. S., Ahn, S. Y., Lee, S. H., Cho, S. B., & Kim, J. H. (2006, April). Heterogeneous clustering ensemble method for combining different cluster results. *In International Workshop on Data Mining for Biomedical Applications. Springer Berlin Heidelberg.* (pp. 82-92).
[31] Li, T., Ding, C., & Jordan, M. I. (2007, October). Solving consensus and semi-supervised clustering problems using nonnegative matrix factorization. *In Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on* (pp. 577-582). *IEEE.*
[32] Weingessel, A., Dimitriadou, E., &Hornik, K. (2003). An ensemble method for clustering. In Proceedings of the 3rd International Workshop on Distributed Statistical Computing.
[33] Dahlin, J., &Svenson, P. (2013). *Ensemble approaches for improving community detection methods*. arXiv preprint arXiv:1309.0242.
[34] Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336), 846-850.
[35] Ben-Hur, A., Elisseeff, A., &Guyon, I. (2001, December). *A stability based method for discovering structure in clustered data. In Pacific symposium on biocomputing* (Vol. 7, pp. 6-17).