Ambiguous Metaclustering Algorithm and Crisp Recursive Profiling for Online Reviewers Using Mahalanobis distance

Kaushalya D.korake [1], Prof. Vilas S. Gaikwad[2]
[1]student , Department of Computer Engineering, JSPM Narhe Technical Campus, Savitribai Phule Pune University,Pune,Maharashtra,India.
[2] Prof, Department of Computer Engineering ,JSPM Narhe Technical Campus, Savitribai Phule Pune University,Pune,Maharashtra,India.

---***---

Abstract - *The crisp and fuzzy met clustering algorithms is used to create more refined profiles of businesses and reviewers on a website for such as e.g. yelp.com. The resulting profiles include the some of the outlying objects into mainstream clusters by allowing them to partially belong to multiple clusters. The objective is to profile the businesses and reviewers by grouping them based on similar characteristics. Business is represented by static information obtained from the database and dynamic information obtained from the clustering of reviewers who reviewed the business. In the same way reviewer representation uses the static representation from the database with profiles of businesses that are reviewed by these reviewers.*

In proposed system work is extended to the implementation of mahalanobis distance for resulting profile. It reduces to the familiar Euclidean distance for uncorrelated variables with unit variance. It accounts for the fact that the variances in each direction are different and covariance between variables in cluster analysis. System uses mahalanobis distance for profiling of online reviewer and business clustering. This approach offers more compact intracluster distance and separate clusters .It increases the inter clustering and decreases the intra clustering of business and reviewers.

Key Words: **Fuzzy-c-mean, K-means, Mahalanobis distance**

1. INTRODUCTION

The ubiquitous nature of web-based systems means that popular websites deal with a large number of entities such as users and resources. Creating meaningful representation and profiles of these users and resources is an important part of web-based system analysis.

Users can also record check-ins at local businesses via their mobile phones. These reviews cover all types of businesses, from automotive to medical services and hotels. However, the majority of reviews center on restaurants and dining. The no. of visitors to yelp.com in the first three months of 2014 was 132 million, with millions of distinct reviewers and businesses., a business can be represented by an information granule consisting of the number of five-star, four-star…, one-star reviews received. A reviewer can be similarly represented by an information granule consisting of votes, and five-star, four-star,…., one-star reviews submitted, where votes are the sum of votes users have given the reviewer's reviews, based on the categories "Cool," "Funny," and "Useful." Traditionally, the data mining Process begins with such a representation of objects based on raw data from the dataset. However, in a typical web-based system, we will have millions of such objects making analytics Unwieldy. A better solution is to use clustering to create a manageable number of profiles of entities such as businesses and reviewers.

Clustering, an unsupervised learning process that groups similar objects, can be used to create profiles of businesses based on type of reviews received, or profiles of reviewers based on types of reviews submitted. However, this only grades the businesses based on the reviews. It does not tell us how easy or hard the reviewers were. This paper addresses that issue by evolving business profiles in parallel with profiles of reviewers. These profiles of both businesses and reviewers recursively enhance the static information obtained from the database. For example, we add the profiles of businesses that were reviewed by a reviewer in the reviewer representation. Similarly, we add the profiles of reviewers who reviewed a business in the representation of the business, which means that the users not only know how the businesses are graded but how easy or hard the reviewers were as well. Similarly, if a user chooses to follow a reviewer, they can find out how easy or hard the reviewer is, a well as the popularity of the businesses graded by the reviewer. The profiles from the clustering are further refined by additional filters that will help users focus on types of businesses, typical hours of operation, and location. The resulting service provides a facility for users to find similar businesses/reviewers based on the category of the business, rating, number of reviews, and number of check-ins. It also provides a succinct profile of a business or reviewer based on these factors so that the users can put the reviews in context.

2. RELATED WORK

2.1 Review of Crisp and Fuzzy Clustering

The goal of both the conventional and fuzzy clustering is to minimize the distances between objects that belong to the same cluster and maximize the distances between clusters. Depending upon the application, we can choose any distance function. Two popular distance functions are Euclidean distance and the inverse of cosine similarity function. This study uses Euclidean distance [1].

2.2 Iterative metaclustering through granular hierarchy of supermarket customers and products

This paper proposes a novel iterative meta-clustering technique that uses clustering results from one set of objects to dynamically change the representation of another set of objects. The proposal evolves two clustering schemes in parallel influencing each other through indirect recursion. The proposal is based on the emerging area of granular computing, where each object is represented as an information granule and an information granule can hierarchically include other information granules. The paper describes the theoretical and algorithmic formulation of the iterative meta-clustering algorithm followed by its implementation. The proposal is demonstrated with the help of a retail store dataset consisting of transactions involving customers and products. A customer granule is represented by static information obtained from the database and dynamic information obtained from clustering of products bought by the customer. Similarly, the product granule augments the static representation from the database with clustering profiles of customers who buy these products. The algorithm is tested for a synthetic dataset to explore various nuances of the proposal, followed by an extensive experimentation with a real-world retail dataset [3].

2.3 Pattern Recognition using the Fuzzy c-means Technique

In the field of pattern recognition due to the fundamental involvement of human perception and inadequacy of standard Mathematics to deal with its complex and ambiguously defined system, different fuzzy techniques have been applied as an appropriate alternative. A pattern recognition system has to undergo basically the steps of preprocessing, feature extraction and selection, classifier design and optimization. In this work the data we have analyzed is in the form of numerical vectors, with a number of clusters predefined. Therefore the fuzzy c-means technique of Bezdek has been considered for this work. Although in the fuzzy c-means technique Euclidean distance has been used to obtain the membership values of the objects in different clusters, in this present work along with Euclidean distance author has used other distances like Canberra distance, Hamming distance to see the differences in outputs[4].

2.4 Temporal recursive meta-clustering

Temporal data have many distinct characteristics, including high dimensionality, complex time dependency, and large volume, all of which make the temporal data clustering more challenging than conventional static datasets. In this paper, proposed a HMM-based partitioning ensemble based on hierarchical clustering refinement to solve the problems of initialization and model selection for temporal data clustering. Their approach results four major benefits, which can be highlighted as: (i) the model initialization problem is solved by associating the ensemble technique;

(ii) the appropriate cluster number can be automatically determined by applying proposed consensus function on the multiple partitions obtained from the target dataset during clustering ensemble phase;(iii)no parameter re estimation is required for then merged pair of cluster, which significantly reduces the computing cost of its final refinement process based on HMM agglomerative clustering and finally (iv)the composite model is better in characterizing the complex structure of clusters. Our approach has been evaluated on synthetic data and time series benchmark, and yields promising results for clustering tasks[5].

3. OBJECTIVE

The objectives of this paper are :

1. The effect of Mahalanobis distance on resulting profiles.
2. To extend existing system to implement the Mahalanobis distance for clustering.
3. Profiling temporal patterns for a given business and reviewer clustering.

4. PROPOSED WORK

4.1 Architectural design

We propose a new approach mahalanobis distance based on the crisp and fuzzy clustering method to evolve the two recursively define clustering of both business and reviewers using a real world dataset. User take input as data, that data take reviewer in the form of static representation and compute the centroid of cluster Ck of reviewer and determine the mahalanobis distance of reviewer and also compute business count and get better cluster reviewer data and vise versa.

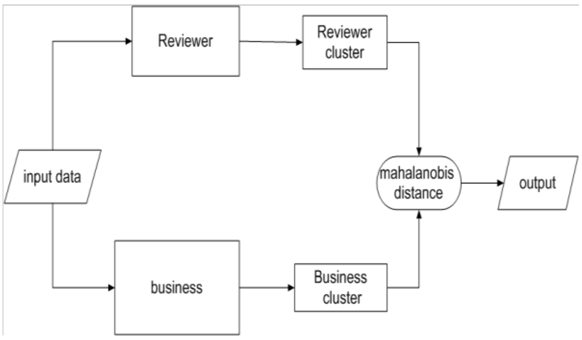


Fig -1: Architecture design

The Mahalanobis distance has the following properties:

- Mahalanobis distance considers the fact that the variances in each direction are different.
- MD considers the covariance between variables in cluster analysis.
- MD minimizes to the familiar Euclidean distance for uncorrelated variables with unit variance [1].

4.2 Methodology:

Reviewer and Business Clustering:

Let R be the set of reviewers, R = r1, r2, . . . ,rnr and B =b1 ,b2, . . . ,bnb be the set of businesses. Let RC be the clustering scheme of reviewers, RC = rc1, rc2, . . . , rckr and BC = bc1 , bc2, . . . , bckb be the clustering scheme of businesses.

The reviewer rj is represented by a static data srj and dynamic data drj, i.e. rj= (srj ,drj). Static part i.e. srj is a data that is taken from the data set such as review types(total,*,**,***,****, *****, votes). Data part drj is derived from the clustering of businesses. Dynamic part is presented by drj= (mj1,mj2, . . . , mjkb). Where; mji is the count of businesses that the reviewer rj reviewed from bci cluster of businesses in crisp clustering. In fuzzy clustering mji is the average count of businesses reviewed by reviewer rj from bci cluster of businesses. The business bi is represented by a static data part sbi and dynamic data part dbi, i.e.,bi= (sbi , dbi). The static part represent the number of reviews (total, *, **, ***, ****, *****) for a business from the dataset. The dynamic part dbi is derived from the clustering of reviewers. For crisp clustering, dbi= (mi1,mi2, . . . , mikr). Mij is the count of reviewers who reviewed the business bi that falls in rcj cluster from the clustering of reviewers.

4.3 Algorithms:

```
Algorithm 1: Reviewer Clustering
Input: R, Srj, Drj
Output: Clusters
1: Compute K number of centroid using Euclidean distance
2: for all reviewer ri ; i =1 to n do
3: Assign each reviewer to the nearest centroid cluster according to the Euclidean
4: distance criterion.
5: end for
```

```
6: for all k = 1 to k do
7: Compute centroid Ck of each cluster
8: if nk > 1,wherenkisnumberofreviewers assigned to that cluster k then
9: Compute the Mahalanobis distance with other cluster.
10: end if
11: end for
```

```
Algorithm 2: Business Clustering
Input: R, Srj,Drj
Output: Clusters
1: Compute K number of centroid using Euclidean distance
2: for all all business bi ; i =1 to n do
3: Assign each business to the nearest centroid cluster according to the Euclidean distance criterion.
4: end for
5: for all k = 1 to k do
6: Compute centroid Ck of each cluster
7: if nk > 1, where nk is number of business assigned to that cluster k then
8: Compute the Mahalanobis distance with other cluster.
9: end if
10: end for
```

5. Result and discussion:

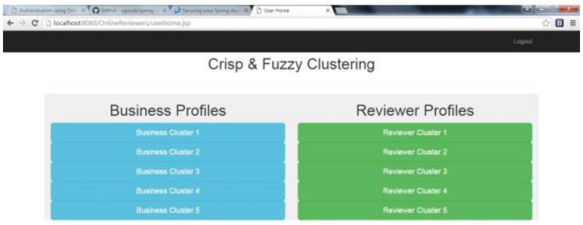
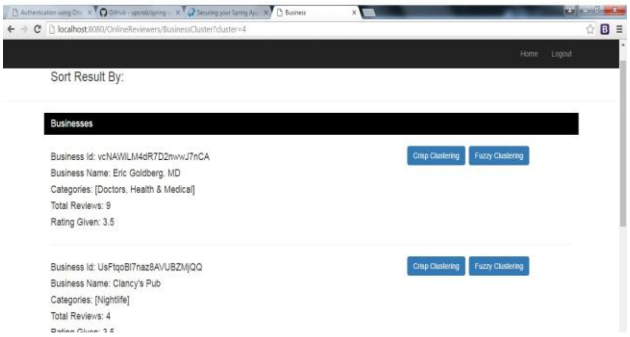


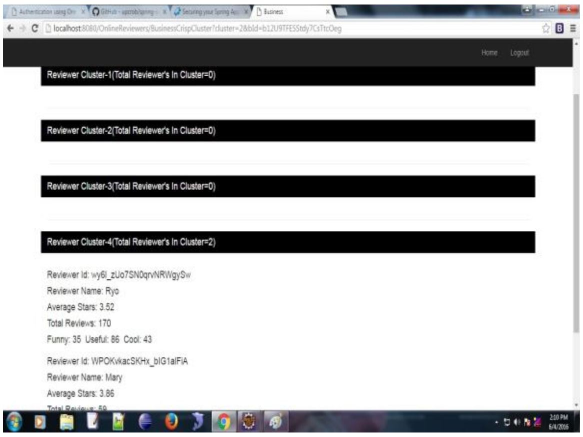
Fig -2: Static clusters of business profiles and reviewer profiles



Business Id: vcNAWiLM4dR7D2nwwJ7nCA
Business Name: Eric Goldberg, MD
Categories: [Doctors, Health & Medical]
Total Reviews: 9
Rating Given: 3.5

Business Id: UsFtqoBl7naz8AVUBZMjQQ
Business Name: Clancy's Pub
Categories: [Nightlife]
Total Reviews: 4

Fig -3: Crisp clustering and fuzzy clustering


localhost:8080/OnlineReviewers/BusinessCrispCluster?cluster=2&bId=b12U9TFESStdy7CsTtcOeg
Home
Logout
2:10 PM
6/4/2016

Reviewer Cluster-1(Total Reviewer's In Cluster=0)

Reviewer Cluster-2(Total Reviewer's In Cluster=0)

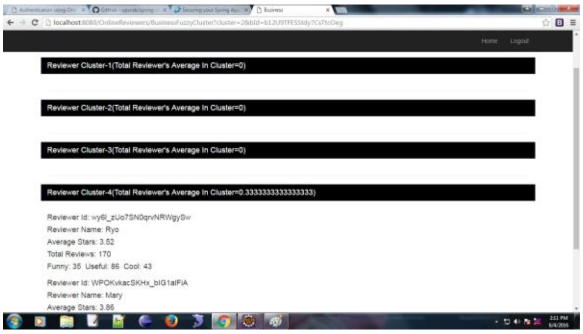
Reviewer Cluster-3(Total Reviewer's In Cluster=0)

Reviewer Cluster-4(Total Reviewer's In Cluster=2)

Reviewer Id: wy6l_zUo7SN0qrvNRWgySw
Reviewer Name: Ryo
Average Stars: 3.52
Total Reviews: 170
Funny: 35 Useful: 86 Cool: 43

Reviewer Id: WPOKvkacSKHx_bIG1alFiA
Reviewer Name: Mary
Average Stars: 3.86

Fig -4: Reviewer cluster list in total reviewer (Crisp clustering)


Business
localhost:8080/OnlineReviewers/BusinessFuzzyCluster?cluster=2&bId=b12U9TFESStdy7CsTtcOeg
Home
Logout

Reviewer Cluster-1(Total Reviewer's Average In Cluster=0)

Reviewer Cluster-2(Total Reviewer's Average In Cluster=0)

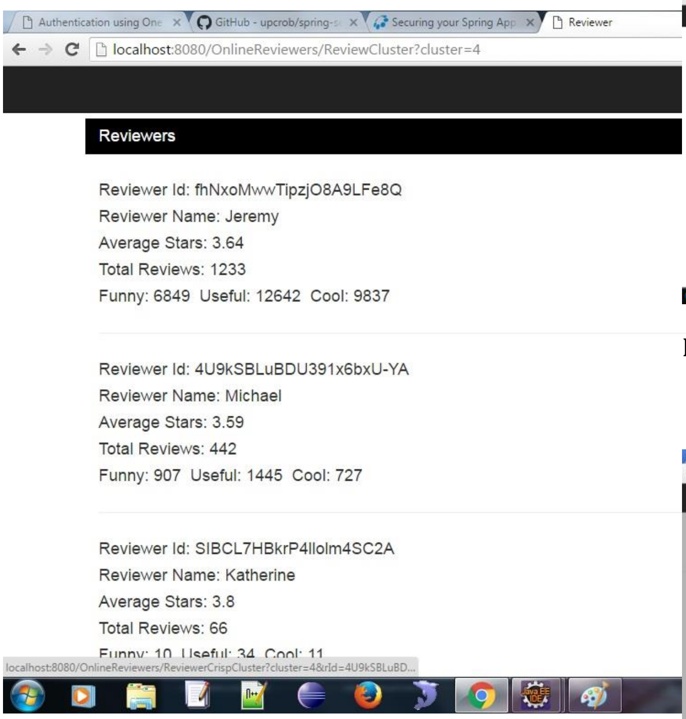
Reviewer Cluster-3(Total Reviewer's Average In Cluster=0)

Reviewer Cluster-4(Total Reviewer's Average In Cluster=0.3333333333333333)

Reviewer Id: wy6l_zUo7SN0qrvNRWgySw
Reviewer Name: Ryo
Average Stars: 3.52
Total Reviews: 170
Funny: 35 Useful: 86 Cool: 43

Reviewer Id: WPOKvkacSKHx_bIG1alFiA
Reviewer Name: Mary
Average Stars: 3.86

Fig -5: Reviewer cluster list in total reviewer average (Fuzzy clustering)



Reviewers

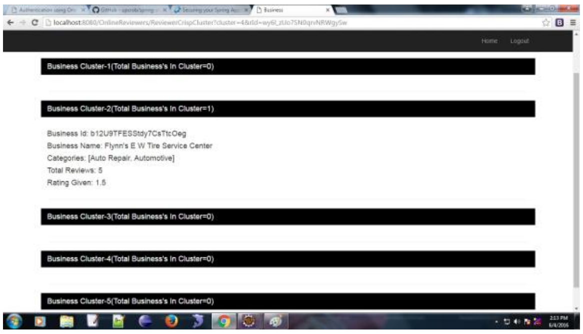
Reviewer Id: fhNxoMwwTipzjO8A9LFe8Q
Reviewer Name: Jeremy
Average Stars: 3.64
Total Reviews: 1233
Funny: 6849 Useful: 12642 Cool: 9837

Reviewer Id: 4U9kSBLuBDU391x6bxU-YA
Reviewer Name: Michael
Average Stars: 3.59
Total Reviews: 442
Funny: 907 Useful: 1445 Cool: 727

Reviewer Id: SIBCL7HBkrP4lloIm4SC2A
Reviewer Name: Katherine
Average Stars: 3.8
Total Reviews: 66
Funny: 10 Useful: 34 Cool: 11

localhost:8080/OnlineReviewers/ReviewerCrispCluster?cluster=4&rId=4U9kSBLuBD...
Crisp Clustering
Fuzzy Clustering

Fig -6: Reviewers information


Business
localhost:8080/OnlineReviewers/ReviewerCrispCluster?cluster=4&rId=wy6l_zUo7SN0qrvNRWgySw
Home
Logout

Business Cluster-1(Total Business's In Cluster=0)

Business Cluster-2(Total Business's In Cluster=1)

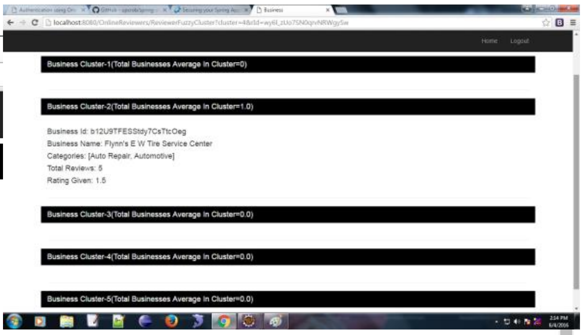
Business Id: b12U9TFESStdy7CsTtcOeg
Business Name: Flynn's E W Tire Service Center
Categories: [Auto Repair, Automotive]
Total Reviews: 5
Rating Given: 1.5

Business Cluster-3(Total Business's In Cluster=0)

Business Cluster-4(Total Business's In Cluster=0)

Business Cluster-5(Total Business's In Cluster=0)

Fig -7: Business cluster user list(Crisp clustering)


Business
localhost:8080/OnlineReviewers/ReviewerFuzzyCluster?cluster=4&rId=wy6l_zUo7SN0qrvNRWgySw
Home
Logout

Business Cluster-1(Total Businesses Average In Cluster=0)

Business Cluster-2(Total Businesses Average In Cluster=1.0)

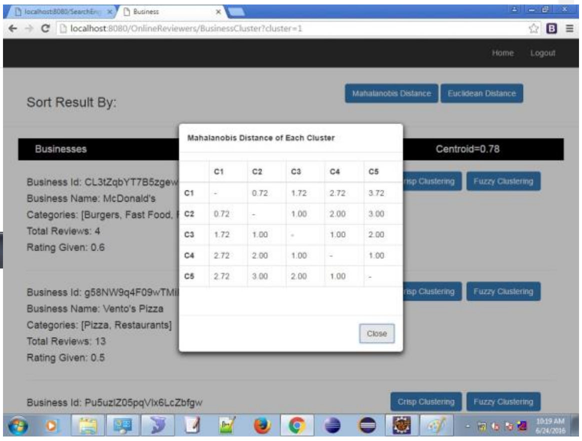
Business Id: b12U9TFESStdy7CsTtcOeg
Business Name: Flynn's E W Tire Service Center
Categories: [Auto Repair, Automotive]
Total Reviews: 5
Rating Given: 1.5

Business Cluster-3(Total Businesses Average In Cluster=0.0)

Business Cluster-4(Total Businesses Average In Cluster=0.0)

Business Cluster-5(Total Businesses Average In Cluster=0.0)

Fig -8: Business cluster total average (Fuzzy clustering)



Sort Result By:

Businesses

Centroid=0.78

Business Id: CL3tZqbYT7B5zgew
Business Name: McDonald's
Categories: [Burgers, Fast Food,
Total Reviews: 4
Rating Given: 0.6

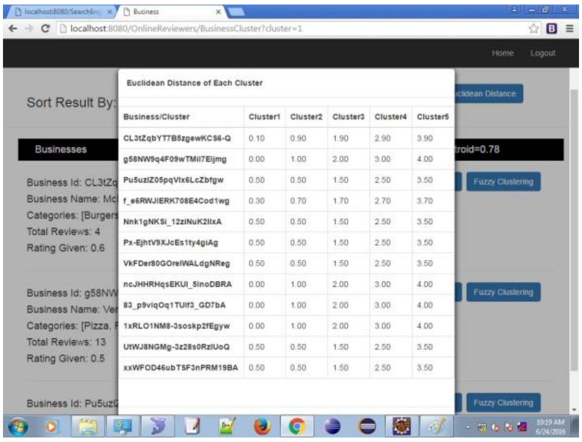
Business Id: g58NW9q4F09wTMi
Business Name: Vento's Pizza
Categories: [Pizza, Restaurants]
Total Reviews: 13
Rating Given: 0.5

Business Id: Pu5uzlZ05pqVlx6LcZbfgw

Mahalanobis Distance of Each Cluster

	C1	C2	C3	C4	C5
C1	-	0.72	1.72	2.72	3.72
C2	0.72	-	1.00	2.00	3.00
C3	1.72	1.00	-	1.00	2.00
C4	2.72	2.00	1.00	-	1.00
C5	2.72	3.00	2.00	1.00	-

Fig -9: Mahalanobis Distance of Each Cluster



Business Id: CL3tZq
Business Name: Mc
Categories: [Burgers
Total Reviews: 4
Rating Given: 0.6

Business Id: g58NW
Business Name: Ve
Categories: [Pizza, F
Total Reviews: 13
Rating Given: 0.5

Business Id: Pu5uzi

Euclidean Distance of Each Cluster

Business/Cluster	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
CL3tZqbYT7B5zgewKCS6-Q	0.10	0.90	1.90	2.90	3.90
g58NW9q4F09wTMil7Eljmg	0.00	1.00	2.00	3.00	4.00
Pu5uziZ05pqVlx6LcZbfgw	0.50	0.50	1.50	2.50	3.50
f_e6RWJIERK708E4Cod1wg	0.30	0.70	1.70	2.70	3.70
Nnk1gNKSi_12ziNuK2lIxA	0.50	0.50	1.50	2.50	3.50
Px-EjhtV9XJcEs1ty4giAg	0.50	0.50	1.50	2.50	3.50
VkFDer80GOrelWALdgNReg	0.50	0.50	1.50	2.50	3.50
ncJHHRHqsEKUI_5lnoDBRA	0.00	1.00	2.00	3.00	4.00
83_p9vlqOq1TUlf3_GD7bA	0.00	1.00	2.00	3.00	4.00
1xRLO1NM8-3soskp2fEgyw	0.00	1.00	2.00	3.00	4.00
UtWJ8NGMg-3z28s0RzlUoQ	0.50	0.50	1.50	2.50	3.50
xxWFOD46ubTSF3nPRM19BA	0.50	0.50	1.50	2.50	3.50

Fig -10: Euclidean Distance of Each Cluster

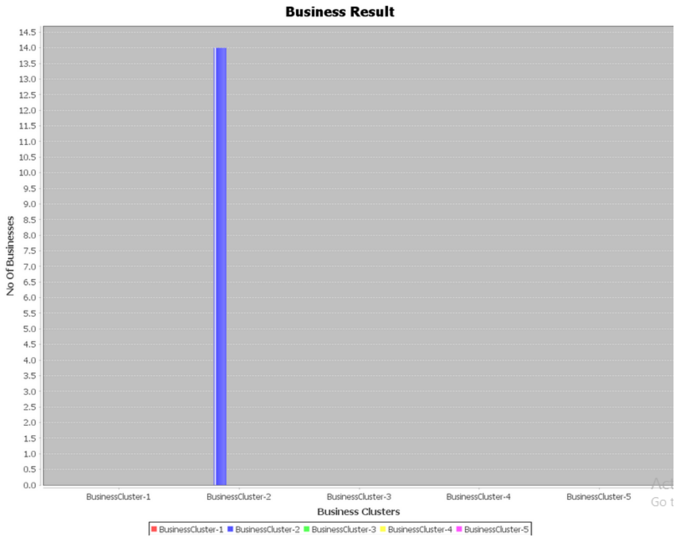


Fig -11: Business Graph

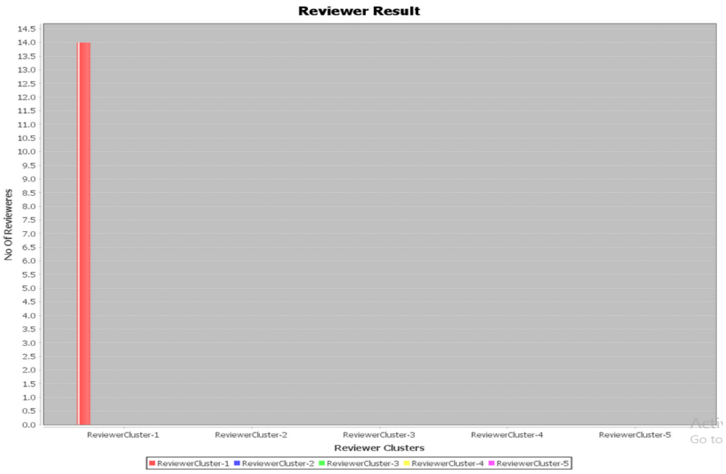


Fig-12: Reviewer Graph

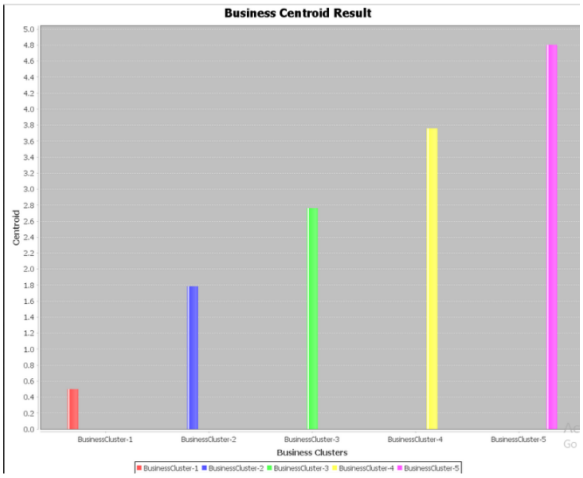


Fig-13: Initial business centroid

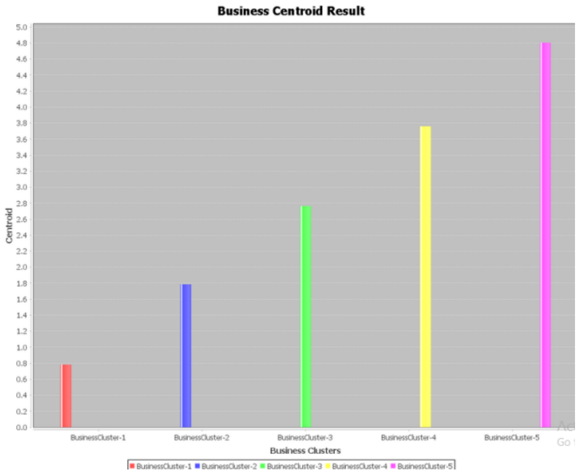


Fig-14: After applying algorithms business centroid

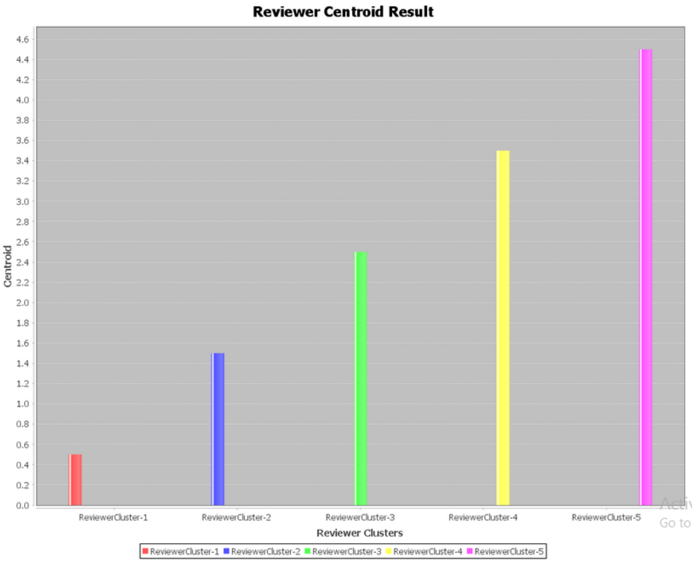


Fig-15: Initial reviewer centroid

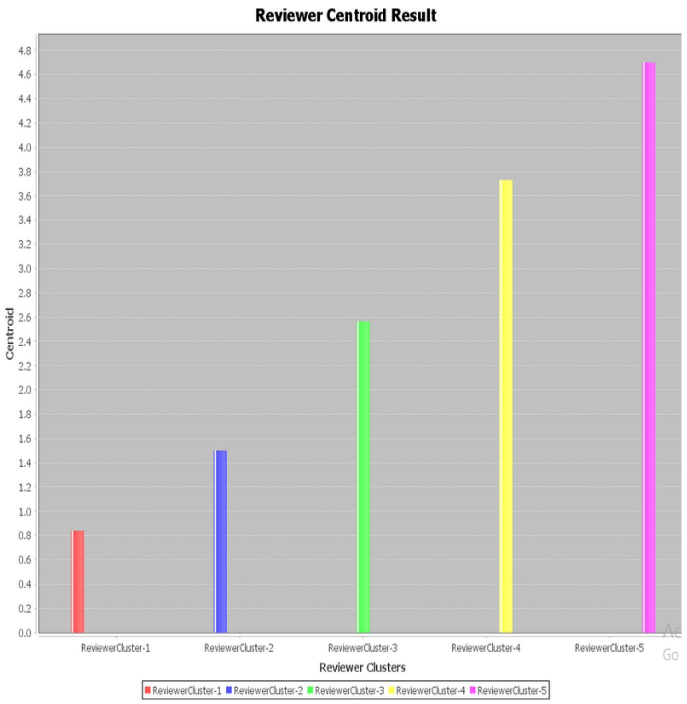


Fig-16: After applying algorithms reviewer centroid

6. CONCLUSION:

Our conclusion is that use of the Mahalanobis distance should become a standard option of the available fuzzy K-means clustering for nonhierarchical cluster analysis. We have use mahalanobis distance for profiling of online reviewer and business clustering. It increases the inter clustering and decreases the intra clustering of business and reviewers.

7. Future Work:

Future studies should work with locations based clustering that check results against a map

REFERENCES

[1] Pawan Lingras and Matt Triff ,“ Fuzzy and Crisp Recursive Profiling of Online Reviewers and Businesses”, VOL. 23, NO. 4, AUGUST 2015.

[2] Haider, Farhana,“Recursive temporal meta-cluster of daily time series”,August 2015, Halifax, Nova Scotia.

[3] P. Lingras, A. Elagamy, A. Ammar, and Z. Elouedi,“metaclustering through granular hierarchy of supermarket customers”, International journal of research and science, Vol. 7, Issue 1, February, 2014.

[4] SamarjitDas,“PatternRecognitionusingtheFuzzyc-means Technique”, International Journal of Energy, Information and Communications Vol. 4, Issue 1, February, 2013 .

[5] D. I. Ignatov, S. O. Kuznetsov, J. Poelmans, and L. E. Zhukov, “Can triconcepts become triclusters?” Int. J. Gen. Syst., vol. 42, no. 6, pp. 572–593, 2013.

[6] D. V. Gnatyshak, D. I. Ignatov, and S. O. Kuznetsov, “From triadic FCA to triclustering: Experimental comparison of some triclustering algorithms,” in Proc. Concept Lattices Appl., 2013, pp. 249–260.

[7] D. I. Ignatov, S. O.Kuznetsov, and J. Poelmans, “Concept-based biclustering for internet advertisement,” in Proc. IEEE 12th Int. Conf. DataMining Workshops, 2012, pp. 123–130.

[8] D. Gnatyshak,D. I. Ignatov, A. Semenov, and J. Poelmans, Gaining Insight in Social Networks With Biclustering and Triclustering, Perspectives in Business Informatics Research. New York, NY, USA: Springer, 2012, pp. 162–171.

[9] Y. Y. Yao,“Artificial intelligence perspectives on granular computing,”in Granular Computing and Intelligent Systems: Design With Information Granules of Higher Order and Higher Type, W. Pedrycz and S.-M. Chen, Eds. Berlin, Germany: Springer, 2011, pp. 1734

[10] D. Ramirez-Cano, S. Colton, and R. Baumgarten, “Player classification using a meta-clustering approach,” in Proc. 3rd Annu. Int. Conf. Comput.

Games, Multimedia Allied Technol., 2010, pp. 297–304.

[11] Y. Y. Yao,“Granular computing: Past, present, and future,” in Rough Set and Knowledege Technology. Berlin, Germany: Springer-Verlag, 2008, pp. 2728.

[12] J. T. Yao, “A ten-year review of granular computing,” in Proc. IEEE Int. Conf. Granular Comput., 2007, pp. 734739