FP- International Journal of Computer Science & Research
Volume 1, Issue 1, Pages 13-16, 2014

# Improved hierarchical clustering using message passing algorithm

Sadhana Bijrothiya
Department of Information Technology
Samrat Ashok Technological Institute , Vidisha, India
sadhanaengg@gmail.com

## ABSTRACT

We present a new improved hierarchical clustering algorithm using message passing algorithm in top down approach. Message passing is an exemplar based clustering algorithm that finds a set of data points that best exemplify the data and associates each data points with one exemplar. We extent message passing in a principled way to overwhelm the hierarchical clustering problem. We derived an algorithm that operates by message exchanging information in between every layer of hierarchy. Based on it, we demonstrate that our method is a new and outperforms agAP HC algorithm (agglomerative Hierarchical Clustering) that cluster iteratively with cluster closeness.

## Keywords

agAP HC algorithm, NC, FM.

## INTRODUCTION

Clustering is a main task of explorative data mining and a common technique for statistical data analysis used in many field's including machine learning, pattern recognition etc. Clustering analysis as such is not an atomic task, but an iterative process of knowledge discovery that involves try and failure. It will often be necessary to modify preprocessing and parameters until the result achieves the desired properties [10]. There are many clustering algorithms such as partitioning and hierarchical clustering algorithms. Hierarchical clustering creates a hierarchical decomposition of database. It iteratively split the database into smaller subset, until some termination condition is satisfied. It does not need *k* as an input parameter, which is an obvious advantage over partitioning clustering algorithms [9]. How many clusters should there be? It is hard to choose a distance metrics, It not finds optimal set of cluster. From the limitations of some hierarchical clustering algorithms which are not capable of finding the predefined structures [4]. A large number of research papers have been published on the improvements of the hierarchical clustering algorithm but no one found to eliminate the problem of hierarchical clustering in top down approach using message passing algorithm. So, from there we have picked agglomerative hierarchical clustering based on affinity propagation algorithm [10] as our base algorithm. To solve the problem above, this paper proposes an improved hierarchical clustering using message passing algorithm is a new simplest algorithm in top down approach. It brings the hierarchical clustering into message passing algorithm and merges the output cluster achieved by the message passing with new clusters whenever a singleton cluster is not formed.

## 2. MESSAGE PASSING

Message passing algorithm simultaneously considers all data points as possible exemplars, exchanging real-valued messages between them until a high quality set of exemplars (and corresponding clusters) emerges. It takes as input a collection of real value similarity between data points s(i,k) where similarity between data points is calculated using negative squared euclidean distance $s(i,k) = -\|x_i - x_k\|^2$. The numbers of clusters need not to be prespecified. It takes as input a real number for each data point is the a priori suitability of point to serve as an exemplar. This value is referred as "preference". The number of identified cluster is influenced by this value. High values of the preferences will cause m to find many exemplars (clusters), while low values will lead to a small number of exemplars (clusters). A good initial choice for the preference is the minimum similarity or the median similarity. After calculating similarities and preferences, two types of messages are exchanged between data points. 1. responsibility r(i,k) – sent from data point i to candidate exemplar point k, reflects the accumulated evidence for how well-suited point k is to serve as the exemplars for point i, taking into account other potential exemplars for point i.

$$r(i,k) \leftarrow s(i,k) - \max_{k' \text{ s.t. } k' \neq k} \{a(i,k') + s(i,k')\}.$$

2. availability a(i,k) – sent from candidate exemplar point k to point i, reflects the accumulated evidence for how appropriate it would be for point i to choose point k as its exemplar, taking into account the support from other points that point k should be an exemplar.

$$a(i,k) \leftarrow \min\{0, r(k,k) + \sum_{i' \text{s.t.} i' \epsilon} \max\{0, r(i',k)\}\}.$$

These messages exchanged between data points, and each takes into account a different kind of competition. Messages can be combined a(i,k)+r(i,k) at any stage to

decide which points are exemplars and, for every other point, which exemplar it belongs to. r(i,k) and a(i,k) can be viewed as log-probability ratios.

## 3. IMPROVED HIERARCHICAL CLUSTERING USING MESSAGE PASSING

Steps of Proposed Algorithm

Input data: Similarity matrix S, Preference value P, Dataset D. Output: Improved hierarchical clustering solution.

Step1: Construct the similarity matrix s based on the negative euclidean distance of dataset. $s(i,k) = -||x_i - x_k||^2$

Step2: Calculate the preference value P (median of the similarities)

Step 3: Calculate the responsibility r(i,k) using this matrix.

$r(i,k) \leftarrow s(i,k) - \max_{k's.t.\ k' \neq k}\{a(i,k')+s(i,k')\}$

Step 4: Calculate the availability a(i,k) using this matrix.

$a(i,k) \leftarrow \min\{ 0, r(k,k) + \sum_{i's.t.\ i' \in} \max\{0,r(i',k)\ \}\}$

The self-availability, a(k,k) is updates as-

$a(k,k) \leftarrow \sum_{i's.t.i' \neq k} \max\{0,r(i',k)\}$

Step5: Set damping factor, convergence iteration and maximum iteration. Damping factor (damp fact): When updating the messages, it is important that they be damped to avoid numerical oscillations that arise in some circumstances.

msg_new = (dampfact) (msg_old) + (1-dampfact) $*$ (msg_new)

Step6: At any point during proposed algorithm availabilities and responsibilities can be combined to identify exemplars. for point i, the value of k that maximizes a(i,k)+r(i,k) either identifies point i as an exemplar if k = i, or identifies the data point that is the exemplar for point i.

Step7: After identify the exemplars. It given the number of N clusters. In proposed, firstly find the first cluster in N clusters. Then we apply again AP algorithm on first cluster whenever first cluster does not given a single cluster and remaining all N-1 clusters are added in new clusters. (N-1)+N_new clusters.

Step8: If AP algorithm generates only one cluster then it returns step8.

Step9: Again this process repeat iteratively from step 3 to 9 whenever all clusters are not generated. This algorithm is halt after fixed number of iterations or the exemplars do not change for a given number of iterations.

Step10: After apply function for display hierarchical clustering dendrogram.

Stop.

Architecture of the proposed method

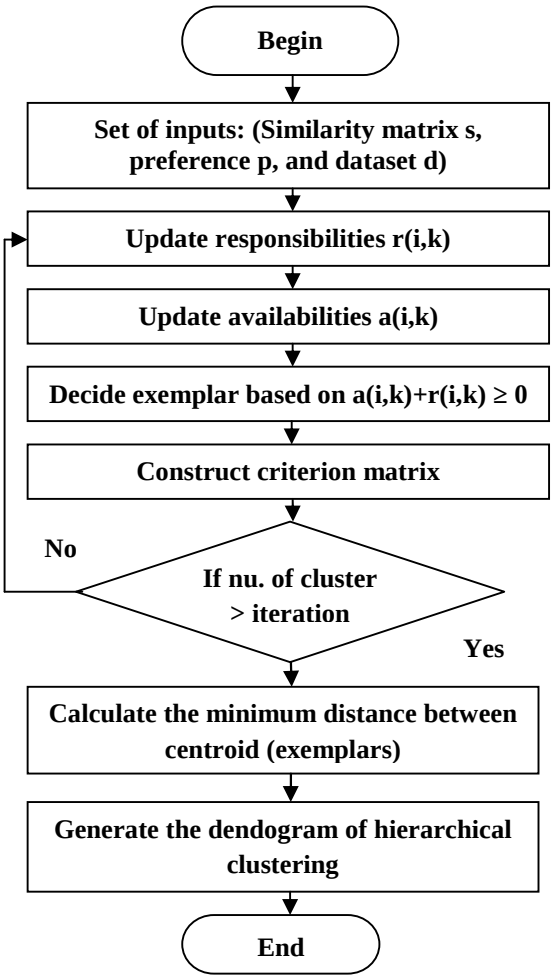


**Figure-1 Flowchart of proposed algorithm**

In this proposed method after finding the clusters, these clusters further pass to message passing algorithm with similarity and preference value and find another clusters. This process iteratively proceeds whenever

defined levels are not proceeding. Finally generate the dendogram of improved hierarchical clustering.

## 4. EXPERIMENT RESULTS AND ANALYSIS

### 4.1 Description of used dataset

In this paper, we have used four datasets of different size that are iris, redwine, ionosphere and whitewine from UCI repository for testing of proposed method. For evaluating the clustering results of hierarchical algorithm the detailed information about the datasets used in this paper is given in table1.1.

### 4.2 Experiment result and analysis

| Datasets | No. of instances | No. of attributes | No. of Actual Classes | References |
|---|---|---|---|---|
| Iris | 150 | 4 | 3 | http://archive.ics.uci.edu/ml/machine-learning-databases /iris |
| Redwine | 1599 | 11 | 4 | http://archive.ics.uci.edu/ml/machine-learning-databases/ redwine |
| Ionosphere | 351 | 34 | 2 | http://archive.ics.uci.edu/ml/machine-learning-databases /ionosphere |
| Whitewine | 4898 | 11 | 7 | http://archive.ics.uci.edu/ml/machine-learning-databases /whitewine |

**Table1.1 Dataset information**

In this paper, measure the performance of proposed algorithm in the following parameters. F-measure is a measure of a test's accuracy. It combines the precision and recall ideas in information retrieval. Recalling $C=\{C_1, C_2, C_3, \ldots, C_k\}$ is a clustering of data set D of N documents, let $C^* = \{C^*_1, C^*_2, \ldots C^*_l\}$ designate the "correct" class set of D. Then, the recall of cluster j with respect to class i, as rec(i,j) is $|C_j \cap C_i^*|/|C_i^*|$. The precision of cluster j with respect to class i, prec(i,j) is defined $|C_j \cap C_i^*|/|C_j|$. F-measure combines both values according to the following formula combines both values according to the following formula.

$$F(i,j) = 2*rec(i,j)*prec(i,j)/prec(i,j)+rec(i,j)$$

Based on this formula, the F-measure for overall quality of cluster set C is defined by the following formula.

$$F = \Sigma[C_i^*/N)*\max(F(i,j))]$$

Another parameter is time(s). Time just wasted divisive clustering, not including the time for message passing partition. It calculates by using tic-toc function in matlab and the numbers of clusters are denoting the actual size of cluster that is generated after the result interpretation. The observation, table1.2 represents the comparative result in between the Ag AP algorithm and proposed algorithm.

| Dataset | Size | Agglomirative hierarchical algorithm | | | Proposed algorithm | | |
|---|---|---|---|---|---|---|---|
| | True NC | NC | Ag AP time (s) | FM | NC | Dv. MP time (s) | FM |
| Iris | 3 | 3 | 0.02 | 0.936 | 3 | 0.033 | 0.937 |
| Redwine | 4 | 11 | 0.03 | 0.823 | 5 | 0.023 | 0.947 |
| Ionosphere | 2 | 14 | 0.26 | 0.589 | 9 | 0.004 | 0.947 |
| Whitewine | 7 | 15 | 0.94 | 0.934 | 4 | 0.006 | 0.947 |

**Table1.2 Comparison of NC, Time (wasted time) and F-measure for agAP algorithm and proposed algorithm**.

In table1.2, the experimental results show that the no. of parameters: time(s), NC denotes the number of classes / clusters and FM (f-measure), are improved in this research paper. We observe that the f-measure of agAP hierarchical clustering is larger than proposed algorithm by using figure1.1. It can be easy seen that the proposed algorithm time is far less than agAP time in figure1.2 and also observe that the size of no. of clusters are far less in compare to agAP. Finally, consequence that the proposed algorithm outperforms agAP hierarchical clustering.

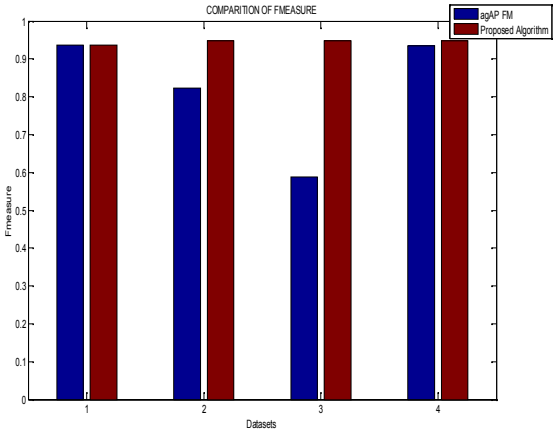


**Figure2 Comparative graph of f-measure**

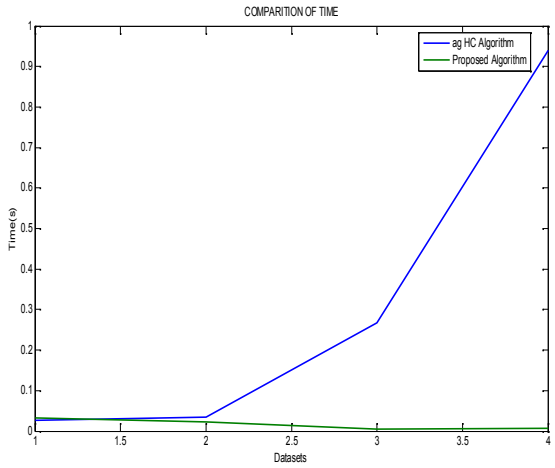


**Figure-3 Comparative graph of time**

## 5. CONCLUSION

We derived an algorithm that operates by message exchanging information in between every layer of hierarchy. The proposed algorithm can merge the close cluster effectively as compare to agAP algorithm. The experimental results show that the parameters: time(s), number of cluster (NC), and f-measure, are improved in this research paper. Based on experimental results, we demonstrate that our method is a new and outperforms agAP HC algorithm that cluster iteratively with cluster closeness. In future, we will try to extend the proposed algorithm for others several types of large and real datasets.

## 6. REFERENCE

[1] Hamidreza Mirzaei, "A Novel Multi-View Agglomerative Clustering Algorithm Based on Ensemble of Partitions on Different Views", School of Computing Science, Simon Fraser University, Burnaby, Canada, DOI 10.1109/ICPR.2010.252, 1051-4651/10, 2010.

[2] Mrs. M. Sreebala, Mr. G. Nageswara Rao, Dr. S. Sai Satyanarayana Reddy, "D-Metric Sparces in Agglomerative Clustering", Journal of Theoretical and Applied Information Technology, Jatit & Lls, All rights reserved, 2005 - 2010.

[3] Inmar E. Givoni and Brendan J. Frey. “A binary variable model for message passing algorithm”. Accepted to Neural Computation, 2009.

[4] Rui Xu and Donald C. Wunsch II, “Clustering” Copyright Institute of Electrical and Electronics Engineers, 2009.

[5] Michele Leone and Martin Weigt. “Unsupervised and semi-supervised clustering by message passing: Soft-constraint message passing algorithm”, Institute for Scientific Interchange, Viale Settimio Severo 65, Villa Gualino, I-10133 Torino, Italy 2008.

[6] Reynaldo J. Gil-García, José M. Badía-Contelles and Aurora Pons-Porrata, "A General Framework for Agglomerative Hierarchical Clustering Algorithms", The 18th International Conference on Pattern Recognition (ICPR'06) 0-7695-2521-0, 2006.

[7] Chris Ding and Xiaofeng He, “Cluster merging and splitting in hierarchical clustering algorithms”, Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM’02) 0-7695 1754-4/02, 2002.

[8] Everitt, B., Landau, S., and Leese, M. “Cluster analysis”, fourth edition, London: Arnold, 2001.

[9] A. K. Jain, M. N. Murty, and P. J. Flynn. “Data clustering: a review”. ACM Computing Surveys, 31(3):264-323, 1999.

[10] Jain & Dubes, Jain, K Anil., & Dubes, Richard C. “Algorithms for clustering data”. Prentice Hall, 1988.