Jurnal Teknologi

Full paper

Using Soft Consensus Clustering for Combining Multiple Clusterings of Chemical Structures

Faisal Saeed[a,b*], Naomie Salim[a]

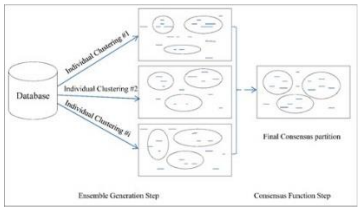
[a]*Faculty of Computing, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia*
[b]*Information Technology Department, Sanhan Community College, Sana'a, Yemen*

*Corresponding author: alsamet.faisal@gmail.com

Article history

Received :31 August 2012
Received in revised form :
28 February 2013
Accepted :15 May 2013

Graphical abstract



Abstract

The consensus clustering has shown capability to improve the robustness, novelty and stability of individual clusterings in many areas including chemoinformatics. In this paper, graph-based consensus method (cluster-based similarity partitioning algorithm CSPA) and soft consensus clustering were examined for combining multiple clusterings of chemical structures. The clustering is evaluated based on the ability to separate active from inactive molecules in each cluster. Experiments suggest that the effectiveness of soft consensus method can obtain better results than the hard consensus method (CSPA).

Keywords: Consensus clustering; graph partitioning; molecular datasets; soft clustering

1.0 INTRODUCTION

Consensus clustering is a framework for combining multiple clusterings of a set of objects to obtain a final consensus partition without accessing the original features of the objects [1]. The consensus clustering has two main steps, partitions generation and consensus function. In the first step, as many as possible individual clusterings will be generated; the collection of these partitions is known as an ensemble. There are different generation mechanisms can be applied including the using of: different object representations, different individual clustering methods, different parameters initialization for clustering methods, and data resampling. In the consensus function step, there are two main approaches: the objects co-occurrence-based and median partition-based approaches [2]. Graph-based and soft consensus clustering are examples of the first approach.

The main advantages of using consensus clustering were summarized by Topchy *et al*. [3] and Fred and Jain [4]. They reported that the consensus clustering can improve the robustness of individual clusterings by obtaining better average performance than the individual clustering algorithms. In addition, the consensus clustering can find solutions unattainable by individual clusterings. Moreover, it obtains results with lower sensitivity to noise and outliers.

In chemoinformatics, it is most unlikely that any single method will yield the best classification under all circumstances [5]. Chu, *et al*. [5] used some consensus methods on sets of chemical structures and concluded that a consensus clustering can outperform Ward's method, which is the current standard clustering method for chemoinformatics applications. However, based on the implemented methods, it was not the case if the clustering is restricted to a single consensus method. Also, Saeed *et al.* [6] examined the using of graph-based methods and concluded that it can improve the robustness of chemical structures clusterings. Moreover, Saeed *et al*. [7] used voting-based consensus method to improve the novelty of chemical clusterings. They concluded that it can outperform Ward's method and other consensus methods.

2.0 MATERIALS AND METHODS

2.1 Dataset

The MDL Drug Data Report (MDDR) database [8], which is the most popular chemoinformatics dataset, was used. This database consists of 102516 molecules. The subset (DS1) was chosen from the MDDR database so that it contains eleven activity classes (8294 molecules), which involves homogeneous and

heterogeneous active molecules. This dataset has been used for many virtual screening experiments [9-11]. The details of this dataset are listed in Table 1. Each row in the table contains an activity index, activity class and the number of molecules belonging to the class. Also, two descriptors were used, which were developed by Scitegic's Pipeline Pilot [12]. These were 120-bit ALOGP and 1024-bit ECFP_4 fingerprints.

Table 1 MDDR activity classes for DS1 dataset

Activity Index	Activity class	Active molecules
31420	Renin Inhibitors	1130
71523	HIV Protease Inhibitors	750
37110	Thrombin Inhibitors	803
31432	Angiotensin II AT1 Antagonists	943
42731	Substance P Antagonists	1246
06233	Substance P Antagonists	752
06245	5HT Reuptake Inhibitors	359
07701	D2 Antagonists	395
06235	5HT1A Agonists	827
78374	Protein Kinase C Inhibitors	453
78331	Cyclooxygenase Inhibitors	636

2.2 Partitions Generation

The generation of partitions was performed on two steps. The first one is to generate the hard partitions by using six individual clustering algorithms on each 2D fingerprint. These algorithms were single-linkage, complete linkage, average linkage, weighted average distance, Ward's and K-means clustering methods. The thresholds of 500, 600, 700, 800, 900 and 1000 were used to generate partitions with different number of clusters. The Jaccard distance measure was used with each clustering method because it was reported that it is the method of choice for partitions generation [7]. The second step is to generate the soft partitions by combining the hard partitions using multiple runs of voting-based consensus method CCVA [7] (b =5 in this experiment), each with random arrangement of partitions. In the voting-based consensus method, the final consensus partition is obtained through a voting process among the objects; so that, each object in the final partition is assigned to each cluster with specific probability or membership value; and the sum of probabilities of assigning each object to all clusters equals to 1.

2.3 Consensus Methods

The graph-based consensus method, Cluster-based Similarity Partitioning Algorithm (CSPA), was proposed by Strehl and Ghosh [13]. It is developed based on transforming the set of clusterings into a graph representation and .establishes a measure of pairwise similarity matrix between the objects. The similarity matrix S is generated so that each two objects have a similarity of 1 if they are in the same cluster and 0 otherwise. The process is repeated for each individual clustering method. Here, we view the similarity matrix as a graph (vertex = object, edge weight = similarity) and cluster it using graph partitioning algorithm METIS [14].

The soft version of CSPA (sCSPA) was proposed by [1], so that it extends CSPA by using values in S to compute pairwise similarities. If we visualize each object as a point in $\sum_{q=1}^{r} k^{(q)}$ dimensional space, with each dimension corresponding to probability of its belonging to a cluster, then SS^T is the same as finding the dot product in this new space. So, the objects are transformed into a label-space, and then the dot product between the vectors representing the objects is considered as their similarity. After creating the similarity matrix, we cluster it using METIS algorithm.

2.4 Performance Evaluation

The results were evaluated based on the effectiveness of the methods to separate active from inactive molecules using the F-measure [15] and Quality Partition Index (QPI) measure [16]. As defined in [5], if the cluster contains *n* compounds, that *a* of these are active and that there is a total of *A* compounds with the chosen Activity. The precision, *P*, and the recall, *R*, for that cluster are:

$$P = \frac{a}{n} \tag{1}$$

$$R = \frac{a}{A} \tag{2}$$

$$F = \frac{2PR}{P+R} \tag{3}$$

This calculation is carried out on each cluster and the F-measure is the maximum value across all clusters. In addition, an active cluster can be defined as a non-singleton cluster for which the percentage of active molecules in the cluster is greater than the percentage of active molecules in the dataset as a whole. Let *p* be the number of actives in active clusters, *q* the number of inactives in active clusters, *r* the number of actives in inactive clusters (i.e., clusters that are not active clusters) and *s* the number of singleton actives. The high value occurs when the actives are separated from the inactive molecules. Then the quality partition index, QPI, is defined to be [16]:

$$QPI = \frac{p}{p+q+r+s}$$

3.0 RESULTS AND DISCUSSION

The generation process was performed on two steps. In the first one, the hard partitions were generated by six individual clusterings. Then, the soft partitions were generated by combining the hard partitions using multiple runs of voting-based consensus method (b=5).

The mean of the F-measure and the QPI values were averaged over the eleven activity classes of the dataset. Tables 2-5 show the effectiveness of clustering of the MDDR dataset using the ALOGP and ECFP_4 fingerprints. The best F-measure and QPI values of consensus methods for each column were bold-faced for ease of reference.

Visual inspection of the F-measure and QPI values in Tables 2-5 enables comparisons to be made between the effectiveness of soft and hard consensus clusterings.

Table 2 Effectivenss of clustering of the MDDR dataset using the F-measure: ALOGP

Clustering Method		No. of clusters					
		500	**600**	**700**	**800**	**900**	**1000**
Consensus	sCSPA	**5.56**	**4.98**	**4.21**	**3.86**	**3.58**	3.08
	CSPA	5.31	4.82	4.15	3.77	3.48	**3.13**
Individual (Ward)		9.93	9.19	8.19	7.17	6.67	6.44

Table 3 Effectivenss of clustering of the MDDR dataset using the F-measure: ECFP_4

Clustering Method		No. of clusters					
		500	**600**	**700**	**800**	**900**	**1000**
Consensus	sCSPA	**5.71**	**5.03**	4.21	**4.00**	3.62	**3.40**
	CSPA	5.51	4.99	**4.25**	3.99	3.62	3.20
Individual (Ward)		11.61	10.71	9.04	8.29	7.64	7.02

In Tables 2-5, for clustering of the MDDR dataset which represented by ALOGP and ECFP4 fingerprints, the performance of soft consensus clustering (sCSPA) outperformed the hard consensus clustering (CSPA) using the F-measure. While, both consensus methods give results that are inferior to Ward's method.

Using the QPI measure, the performance of the soft consensus method (sCSPA) outperformed the CSPA and Ward's methods for ALOGP fingerprint. However, when ECFP_4 is used, sCSPA gives better results than CSPA while its performance is inferior to Ward's method.

Table 4 Effectivenss of clustering of the MDDR dataset using the QPI measure: ALOGP

Clustering Method		No. of clusters					
		500	**600**	**700**	**800**	**900**	**1000**
Consensus	sCSPA	**57.21**	**59.44**	**61.54**	**62.99**	**65.35**	**69.47**
	CSPA	55.03	59.13	60.84	61.03	63.73	67.44
Individual (Ward)		52.33	54.86	56.9	59	61.33	63.17

Table 5 Effectivenss of clustering of the MDDR dataset using the QPI measure: ECFP_4

Clustering Method		No. of clusters					
		500	**600**	**700**	**800**	**900**	**1000**
Consensus	sCSPA	**70.29**	**72.71**	**74.98**	**77.00**	**78.82**	**80.19**
	CSPA	69.91	71.73	74.20	76.01	77.72	79.26
Individual (Ward)		75.83	79.88	83.34	84.25	86.49	88.25

4.0 CONCLUSION

The experimental results show that the performance of soft consensus clustering (sCSPA) can provide better results than hard consensus clustering (CSPA) when combining multiple clusterings of chemical structures. However, the Ward's standard individual clustering method shows superior results than soft consensus method. Therefore, more soft consensus methods are needed to be examined in future work.

Acknowledgement

This work is supported by Ministry of Higher Education (MOHE) and Research Management Centre (RMC) at the Universiti Teknologi Malaysia(UTM) under Research University Grant Category (VOT Q.J130000.7128.00H72).

References

[1] Punera, K., Ghosh, J. 2008. Consensus-based Ensembles of Soft Clusterings. *Applied Artificial Intelligence*. 22(7–8): 780–810.

[2] Vega-Pons, S., and Ruiz-Schulcloper, J. A. 2011. Survey of Clustering Ensemble Algorithms. *International Journal of Pattern Recognition and Artificial Intelligence*. 25(3): 337–372.

[3] Topchy, A., Jain, A. K., Punch, W. 2004. A Mixture Model of Clustering Ensembles. *SIAM Int. Conf. Data Mining*. 379–390.

[4] Fred, A. L. N., Jain, A. K. 2005. Combining Multiple Clustering Using Evidence Accumulation. *IEEE Trans. Patt. Anal. Mach. Intell*. 27: 835–850.

[5] Chu, C-W., Holliday, J., Willett, P. 2012. Combining Multiple Classifications of Chemical Structures Using Consensus Clustering. *Bioorgan Med Chem*. 20(18): 5366–5371.

[6] Saeed, F., Salim, N., Abdo, A., Hentabli, H. 2013. Graph-based Consensus Clustering for Combining Multiple Clusterings of Chemical Structures. *Molecular Informatics*. 32(2): 165–178.

[7] Saeed, F., Salim, N., Abdo, A. 2012.Voting-based Consensus Clustering for Combining Multiple Clusterings of Chemical Structures. *Journal of Cheminformatics*. 4(1): 37.

[8] Sci Tegic Accelrys Inc., the MDL Drug Data Report (MDDR) database is available from at http :// www.accelrys.com / (accessed 1st of November 2012).

[9] Abdo, A., Chen, B., Mueller, C., Salim, N., Willett, P. 2010. Ligand-based Virtual Screening Using Bayesian Networks. *J Chem Inf Model.* 50: 1012–1020.

[10] Abdo, A., Salim, N. 2011. New Fragment Weighting Scheme for the Bayesian Inference Network in Ligand-based Virtual Screening. *J Chem Inf Model.* 51: 25–32.

[11] Abdo, A., Saeed. F., Hentabli, H., Ahmed, A., Salim, N. 2012. Ligand Expansion in Ligand-based Virtual Screening Using Relevance Feedback. *J Comput-Aided Mol Des.* 26: 279–287.

[12] Pipeline Pilot software: SciTegic Accelrys Inc. San Diego: Accelrys Inc website; 2008. http://www.accelrys.com/.

[13] Strehl, A., Ghosh, J. 2002. Cluster Ensembles—A Knowledge Reuse Framework for Combining Multiple Partitions. J. *Machine Learning Research.* 3: 583–617.

[14] Karypis, G., Aggarwal, R., Kumar, V., Shekhar, S. 1997. Multilevel Hypergraph Partitioning: Application in VLSI Domain, DAC'97: Proc. *34th Ann. Conf. Design Automation (ACM, New York, NY, USA*). 5261997529.

[15] Van, Rijsbergen, C. J. 1979. Information Retrieval. 2nd Edition. London: Butterworths;

[16] Varin, T., Saettel, N., Villain, J., Lesnard, A., Dauphin, F., Bureau, R., Rault, S. J. 2008. 3D Pharmacophore, Hierarchical Methods, and 5-HT4 Receptor Binding Data. *Enzyme Inhib Med Chem*. 23: 593–603.