BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 20 2010, pages 2645–2646
doi:10.1093/bioinformatics/btq489

Data and text mining

Advance Access publication August 24, 2010

Obtaining better quality final clustering by merging a collection of clusterings

Selim Mimaroglu* and Ertunc Erdil

Department of Computer Engineering, Bahcesehir University, Ciragan Caddesi 34353 Besiktas, Istanbul, Turkey

Associate Editor: John Quackenbush

ABSTRACT

Motivation: Clustering methods including *k*-means, SOM, UPGMA, DAA, CLICK, GENECLUSTER, CAST, DHC, PMETIS and KMETIS have been widely used in biological studies for gene expression, protein localization, sequence recognition and more. All these clustering methods have some benefits and drawbacks. We propose a novel graph-based clustering software called COMUSA for combining the benefits of a collection of clusterings into a final clustering having better overall quality.

Results: COMUSA implementation is compared with PMETIS, KMETIS and *k*-means. Experimental results on artificial, real and biological datasets demonstrate the effectiveness of our method. COMUSA produces very good quality clusters in a short amount of time.

Availability: http://www.cs.umb.edu/~smimarog/comusa

Contact: selim.mimaroglu@bahcesehir.edu.tr

Received on June 13, 2010; revised on August 17, 2010; accepted on August 18, 2010

1 INTRODUCTION

Clustering is the process of organizing objects into groups which have similar members; a distance metric is used for evaluating the similarity. Clustering is also known as unsupervised classification in the literature.

Clustering has a long and rich history in a variety of scientific fields (Jain, 2010). Taxonomists, social scientists, psychologists, biologists, statisticians, mathematicians, engineers, computer scientists, medical researchers and others who collect and process real data have all contributed to clustering methodology. There are many well-known clustering algorithms, as stated earlier. Each clustering method provides some benefits. Therefore, multiple clusterings can be obtained for the same dataset, and these benefits can be combined into a final clustering. COMUSA takes a collection of clusterings as input, and it creates a final clustering with better overall quality. In other words, COMUSA combines the benefits of multiple clusterings into a final clustering.

COMUSA is a general purpose software; it can be used on biological and non-biological datasets. However, in this article, we will evaluate COMUSA's performance on biological datasets.

*To whom correspondence should be addressed.

Table 1. Π, a collection of clusterings in binary format

Clustering ID	Cluster ID	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8
1	1	1	0	1	0	0	1	0	0
	2	0	0	0	1	1	0	0	0
	3	0	1	0	0	0	0	1	1
2	4	1	1	0	1	0	0	0	0
	5	0	0	0	0	0	0	0	1
	6	0	0	0	0	1	0	1	0
	7	0	0	1	0	0	1	0	0
3	8	0	0	1	0	0	1	0	0
	9	1	1	0	1	0	0	0	1
	10	0	0	0	0	1	0	1	0

2 APPROACH

A collection of clusterings is provided as input to COMUSA. Using this input, COMUSA creates a similarity graph of objects. Similarity graph is an undirected and weighted graph that represents the co-association of objects. Similarity graph is constructed by accumulating the evidence in multiple clusterings. In this graph, an edge incident to vertices v_i and v_j represents the number of times these vertices are assigned to the same cluster and it is shown as $\mathsf{weight}(v_i, v_j)$. Each vertex (object) v_i is represented by a pair: First component of a vertex, $\mathsf{df}(v_i)$, is the number of edges that are incident to the vertex, called *degree of freedom*. Second component, $\mathsf{sw}(v_i)$, is the total sum of weights of these edges, called *sum of weights*. The attachment of a vertex v_i is defined as $\mathsf{attachment}(v_i) = \frac{\mathsf{sw}(v_i)}{\mathsf{df}(v_i)}$.

3 COMUSA

Below, we explain the general characteristics of COMUSA.

An unmarked object v_p having the highest attachment values is selected as pivot (seed) for starting a new cluster. High values of degree of freedom mean that the node is connected to many other vertices. Similarly, large values of sum of weights indicate high similarity with several other vertices. The ratio, sum of weights over degree of freedom, indicates attachment. Therefore, an object having high attachment value is strongly connected to somewhere. A vertex v_i is added to the cluster of v_p if it is more closely attached to v_p than to any other vertex.

For the multiple clusterings of a dataset in Table 1, the similarity graph is constructed as shown Figure 1a. Nodes having the highest attachment value (2.0) are d_3 and d_6. Randomly, d_3 is picked

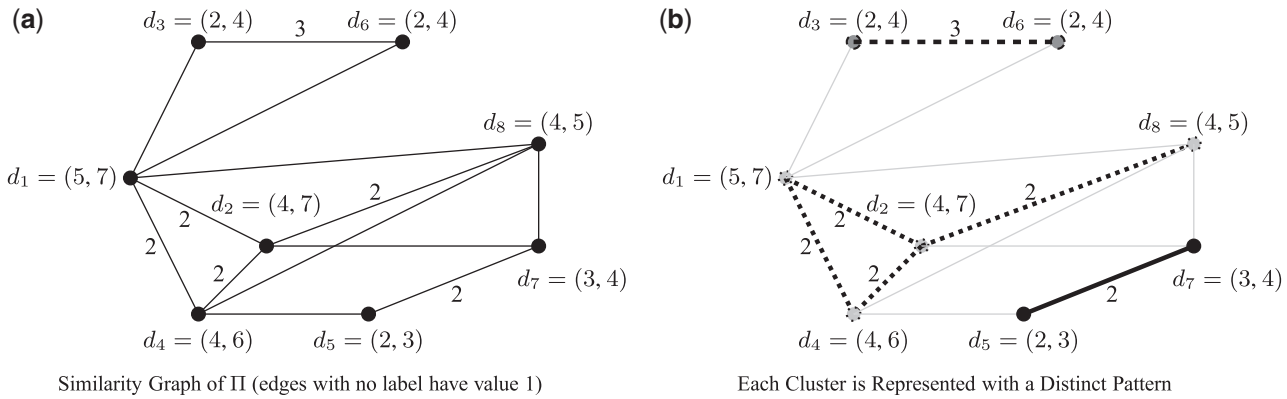

Fig. 1. Constructing clustering using Table 1.

as pivot for starting a new cluster. Next, objects that will be in the same cluster with d_3 are discovered. Starting from the nearest neighbors, we have to evaluate the edge weights and include objects into first cluster if they are strongly connected to d_3. Immediate neighbors of d_3 are d_6 and d_1. d_6 is added to the cluster since $\mathsf{weight}(d_3,\ d_6) \geq \mathsf{weight}(d_6,\ d_1)$. d_1 is not included in the first cluster, because $\mathsf{weight}(d_3,\ d_1) \ngeq \mathsf{weight}(d_1,\ d_4)$ [similarly, $\mathsf{weight}(d_3,\ d_1) \ngeq \mathsf{weight}(d_1,\ d_2)$]. First cluster, which is shown in dashed pattern in Figure 1b, has two objects $C_{\star 1} = \{d_3,\ d_6\}$.

Since there are objects that are not included in any cluster, COMUSA keeps running. Next, d_2 is selected as a pivot, because among all the unlabeled objects it has the highest attachment value. COMUSA extends the cluster in the same way as explained earlier and produces the second cluster: $C_{\star 2} = \{d_2,\ d_1,\ d_4,\ d_8\}$, which is shown in Figure 1b in dotted pattern. Finally, COMUSA merges d_5 and d_7 and halts. Third cluster, $C_{\star 3} = \{d_5,\ d_7\}$, is shown in bold straight line in Figure 1b.

4 RESULTS AND DISCUSSION

Selecting a pivot (seed) object for initiating a new cluster is essential in COMUSA. An unmarked object having the highest attachment value is selected as pivot for starting a new cluster. Initially each cluster is a singleton. Pivot object expands the cluster by considering all the immediate neighbors. A neighbor is included in a pivot's cluster if it is most similar to the pivot. Once a neighbor is included, it is marked and then it acts like a pivot by considering its immediate neighbors for further expansion. Expansion of a cluster comes to a stop if pivots cannot add any other objects into the cluster. If there are unmarked objects left, COMUSA starts a new cluster by choosing a new pivot. COMUSA halts when all the objects belong to a cluster.

Arbitrary shape clusters can be found by our algorithm, and we do not make any assumptions about the input dataset. COMUSA works very well because pivot objects are good starting points, and an object is included into a cluster if the object is most similar to a pivot in that cluster.

In some cases it is desirable to have larger clusters: user input *relaxation* rate is used to achieve this, which is defined as follows. On a set of edge weights $\{w_1,\ w_2, \ldots, w_n\}$, w_k is said to have maximum value with relaxation r, if $\forall_i (w_k + w_k.r \geq w_i)$ holds. Maximum edge weight constraint is relaxed with the relaxation value, r.

COMUSA produces very good results on real and synthetically produced datasets, as well as on biological datasets. Experimental results are shown at the home page in great detail. On *Escherichia coli* dataset, COMUSA produces excellent results with respect to adjusted rand index (ARI) (Hubert and Arabie, 1985).

5 CONCLUSION

Using a collection of clusterings produced by various methods, COMUSA produces a final clustering which has better overall quality. COMUSA creates a similarity graph by using the evidence accumulated from the clusterings. By using the similarity graph, COMUSA computes pivot objects for starting a new cluster, and expands the clusters as much as possible. Number of clusters is automatically found by our algorithm with respect to relaxation rate. COMUSA is partitional, exclusive and complete. Extensive experimental evaluations on many real and artificial datasets demonstrate that COMUSA: (i) finds arbitrary shape clusters, (ii) is not affected by the cluster size, (iii) is not affected by noise and outliers, (iv) is not affected by the sparseness of the dataset, (v) is order independent, (vi) is deterministic and (vii) scales well.

Conflict of Interest: none declared.

REFERENCES

Hubert,L. and Arabie,P. (1985) Comparing partitions. *J. Classif.*, **2**, 193–218.

Jain,A. (2010) Data clustering: 50 years beyond k-means. *Pattern Recogni. Lett.*, **31**, 651–666.