WEIGHTED CONSENSUS CLUSTERING FOR MICROARRAY DATA ANALYSIS

MEGHANA DEODHAR JOYDEEP GHOSH

Department of Electrical and Computer Engineering
The University of Texas at Austin,
Austin, TX 78712-1084, USA

ABSTRACT

Clustering is powerful tool used for analyzing microarray data in order to discover groups of genes sharing patterns of co-expression across a set of conditions. The results of different clustering algorithms on the same dataset may vary significantly. Different algorithms may find varied, but good quality gene clusters. Therefore, a consensus across several clustering results may be able to capture the merits of the different clustering algorithms and produce more reliable clusters. This paper proposes a novel weighted consensus clustering algorithm that is aimed at providing a good quality clustering of microarray data and evaluates it experimentally.

INTRODUCTION

Common clustering algorithms like k-means and hierarchical clustering do not always work very well on microarray data since is often very noisy and high dimensional. Moreover, the results of different clustering algorithms with different distance measures on the same microarray dataset could vary significantly based on how suitable they are to the dataset. This suggests that a combination of the good quality gene clusters from each of these clustering results could capture the merits of the different algorithms and produce more reliable clusters. A consensus across the base results could be taken in such a way that it weighs better quality base clustering results higher. We propose a weighted consensus clustering algorithm for microarray data and evaluate its effectiveness experimentally.

RELATED WORK

The problem of combining multiple partitionings of a set of objects into a single consolidated clustering has been addressed by Strehl et al. [1]. They pose the consensus clustering problem as finding a consensus function that combines clustering results from a variety of sources without using the original features or algorithms. Three computationally efficient heuristics for combining multiple clusterings are proposed, each of which find an optimal clustering that maximizes the average normalized mutual information (ANMI) with the individual clusterings.

Consensus clustering has recently been applied to microarray data to improve the quality and robustness of the resulting clusters. A resampling based approach is used by Monti et al. [4] for cluster discovery and visualization of microarray data. Swift et al. use a variety of clustering

algorithms on the same dataset to generate different base clustering results [2] and try to find clusters that are consistent across all the base results. Another formulation of the consensus clustering problem is as a median partition problem [3], where the aim is to find a partitioning of the data points that minimizes the distance to all the other partitionings.

MOTIVATION FOR WEIGHTED CONSENSUS

In the resampling based consensus clustering approach, several datasets are obtained from the original dataset by subsampling [4]. A base clustering algorithm is run on each subsampled dataset. Each base clustering run generates a connectivity matrix $C_{N\times N}$, where N is the total number of data points, such that $C_{ij} = 1$ if points i and j are assigned to the same cluster, $C_{ij} = 0$ otherwise. The consensus matrix $M_{N\times N}$ is obtained by simply averaging the connectivity matrices across all the subsampled datasets. The consensus matrix is viewed as a similarity matrix and is used to obtain the final clustering using hierarchical clustering.

One of the main problems with this approach is that it weights all the base clustering solutions equally. However, it is natural that some base clustering solutions are of a better quality than the others. If there were a reliable way of determining the quality of a clustering solution, then a consensus clustering algorithm that weights the base clustering results proportional to their quality would be more appropriate and give better performance. In addition, the resampling based approach uses subsampling to create perturbed datasets, which causes a loss of potentially useful data. This approach is also computationally very intensive since it involves several hundred runs of the base algorithm. Finally, the resampling approach uses just one base algorithm which could give poor performance on datasets for which it is not suitable. Using a set of different base clustering algorithms would give a more stable solution. Other consensus clustering approaches [3] and [2] use different base clustering algorithms but even these techniques weight the base solutions equally while taking a consensus across them. The weighted consensus clustering algorithm is devised to improve on some of the shortcomings of the resampling based approach. It uses different base clustering algorithms like k-means and hierarchical clustering with different distance measures and assigns weights to the base solutions, proportional to their quality.

WEIGHTED CONSENSUS CLUSTERING ALGORITHM

The first step is to determine a suitable measure for evaluating a clustering result. Mutual information is a symmetric measure that quantifies the statistical information shared between distributions [5]. Mutual information can hence be used to provide an indication of the information shared between a pair of clusterings. Let X and Y be two random variables that represent two clustering results. Let $I(X, Y)$ denote the mu-

tual information between X and Y. $I(X,Y)$ can be normalized so that it lies between 0 and 1 and allows easier interpretation and comparison. If $H(X)$ denotes the entropy of X, the normalized mutual information (NMI) between X and Y is computed as $NMI(X,Y) = \frac{I(X,Y)}{\sqrt{H(X)H(Y)}}$. Average normalized mutual information (ANMI) is defined as a measure between a set of N clusterings denoted by S and a single clustering C [1], computed as $ANMI(S,C) = \frac{1}{N}\sum_{i=1}^{N} NMI(S_i,C)$. The experiments carried out by Strehl et al. indicate that the ANMI of a clustering result with the other clustering results on the same dataset is highly correlated with the NMI of the result with the true labels [1]. ANMI can hence be used as an indicator of the quality of a clustering result. The weighted consensus clustering algorithm is as follows:

1. Obtain the clustering results of the individual base clustering algorithms. Express each result as a connectivity matrix $C_{n\times n}$, where n is the number of data points. $C_{i,j} = 1$ if points i and j are assigned to the same cluster, $C_{i,j} = 0$ otherwise.

2. Let N be the number of base clustering algorithms. Let C_i be the connectivity matrix of the i^{th} base clustering result and A_i be the ANMI of the i^{th} base clustering result with the other results. Assign a weight w_i to each connectivity matrix C_i proportional to its ANMI with the other base clustering results. The weights are normalized to ensure that they sum to 1, $w_i = \frac{A_i}{\sum_{j=1,j\neq i}^{j=N} A_j}$ for $i = 1$ to N.

3. Take a weighted average across all the connectivity matrices to produce the consensus matrix $M = \sum_{i=1}^{i=N} w_i C_i$

4. Treat the consensus matrix M as a similarity matrix. Apply average linkage hierarchical clustering to matrix M to obtain the final clustering result.

EXPERIMENTAL RESULTS

The datasets used for evaluating the weighted consensus clustering algorithm are the ones from [4], described in Table 1. For these datasets, the experiments are clustered based on the similarities in the expression of the associated genes. Evaluation is based on comparing the final clustering result with the true class labels, using NMI and the adjusted Rand index (RAND) as measures of agreement.

Table 2 compares the weighted consensus clustering result to the average and best base clustering results. The results have been generated using 8 base solutions - k-means with squared Euclidean, squared cosine and correlation distance, average, single and complete link HAC with cosine distance, average link HAC with squared Euclidean distance and spherical kmeans and are averaged over 20 runs. AvgNmi is the NMI averaged across the base

Dataset	# Classes	# Data Points	# Features
Leukemia	3	38	999
Novartis multi-tissue	4	103	1000
St. Jude Leukemia	6	248	985
Lung cancer	4	197	1000
CNS tumor	5	48	1000
Normal tissue	13	99	1277

Table 1: Description of Datasets

clustering results with AvgNmiStd as the standard deviation. NmiMax is the NMI of the best base solution. NmiCC and NmiCCStd are the mean and standard deviation of the NMI of the weighted consensus clustering result. On all the datasets, the consensus results are significantly better than the average base results. Additionally, the performance of the consensus algorithm is very close to that of the best base clustering solution on most datasets.

Dataset	AvgNmi	AvgNmiStd	NmiMax	NmiCC	NmiCCStd
Leukemia	0.918	0.017	1.000	1.000	0.000
Novartis	0.877	0.005	0.969	0.940	0.000
Jude	0.852	0.007	0.938	0.933	0.001
Cancer	0.456	0.006	0.645	0.571	0.000
CNS	0.654	0.009	0.813	0.811	0.000
Normal	0.762	0.005	0.826	0.817	0.017

Table 2: Consensus Clustering Results vs Base Results

Tables 3 and 4 evaluate the weighted consensus clustering algorithm in terms of the improvement weighting by AMNI provides over just averaging the base results. NmiCC and RandCC are the mean NMI and the mean RAND index of the weighted consensus clustering result respectively. NmiAvg and RandAvg are the mean NMI and the mean RAND index of the clustering result obtained by averaging the base solutions to obtain the consensus matrix in step 3 of the consensus clustering algorithm, rather than computing a weighted average.

Dataset	NmiCC	NmiCCStd	NmiAvg	NmiAvgStd
Leukemia	1.000	0.000	1.000	0.000
Novartis	0.940	0.000	0.939	0.000
Jude	0.933	0.001	0.929	0.003
Lung Cancer	0.571	0.000	0.570	0.000
CNS	0.811	0.000	0.762	0.038
Normal	0.817	0.017	0.817	0.017

Table 3: Evaluation of the Weighted Approach (NMI)

Dataset	RandCC	RandCCStd	RandAvg	RandAvgStd
Leukemia	1.000	0.000	1.000	0.000
Novartis	0.946	0.000	0.946	0.000
Jude	0.948	0.000	0.943	0.003
Lung Cancer	0.404	0.000	0.404	0.000
CNS	0.729	0.000	0.664	0.052
Normal	0.630	0.032	0.630	0.032

Table 4: Evaluation of the Weighted Approach (RAND index)

From Tables 3 and 4 it can be observed that just averaging the base results does equally well as compared to the weighted averaging technique on datasets Leukemia, Novartis, Jude, Lung Cancer and Normal. This is because the quality of the different base solutions does not vary much on these datasets. The ANMI values and consequently the weights assigned to the base solutions are roughly equal. However, in case of the CNS dataset, the k-means solutions have a significantly higher NMI compared to the other solutions as seen in table 5. These good quality solutions are assigned higher weights which brings about an improvement over simple averaging.

Algorithm	k-means sqCosine	k-means correlation	k-means sqEuclidean	avg-HAC cosine
NMI	0.8133	0.8133	0.8115	0.5614
weight	0.1297	0.1297	0.1302	0.1284
Algorithm	avg-HAC sqEuclidean	single-HAC cosine	complete-HAC cosine	sp-kmeans
NMI	0.6139	0.4366	0.6139	0.6183
weight	0.1227	0.1178	0.1227	0.1188

Table 5: Weights Assigned to Base Solutions

Figure 1 shows a comparison of the weighted algorithm with the resampling approach using the adjusted RAND index. The weighted technique uses 8 base solutions. It has been compared to the resample approach with an equal number of base solutions (resample-8) and resampling with 500 solutions (resample-500) [1]. The weighted algorithm does significantly better on the Jude, Lung Cancer, CNS and Normal data sets as compared to the resampling based approach. Both the algorithms perform well on Leukemia and Novartis which are easy, well separated data sets.

CONCLUSIONS

The weighted consensus clustering algorithm provides improvement over the average performance of the base clustering algorithms and comes very

[1] The results we obtained by implementation of the resampling algorithm are slightly different from the ones in the paper by Monti et al. [4], but are still in the ballpark.

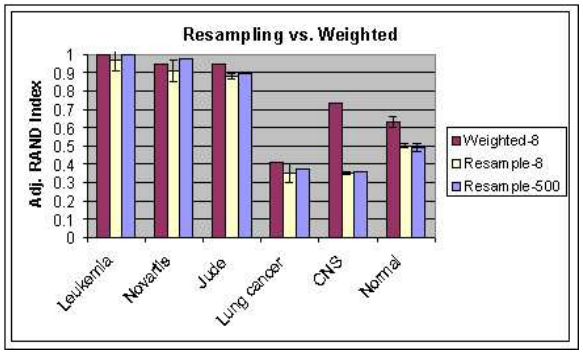

Figure 1: Comparison with the Resampling Approach

close to the best base solution on the microarray datasets in Table 1. Additionally, the weighted algorithm gives significantly better clustering results as compared to the resampling based approach. Another observation is that weighting base solutions by their ANMI causes an improvement only when the base solutions differ significantly in quality. If the quality of the base solutions does not vary much, a simple averaging approach provides results similar to the weighted approach. It would be interesting to apply this algorithm to datasets from different domains and investigate whether similar behavior is observed.

ACKNOWLEDGEMENTS

This work was supported by NSF grant IIS-0325116.

REFERENCES

[1] A. Strehl and J. Ghosh. Cluster Ensembles – A Knowledge Reuse Framework for Combining Multiple Partitions. In *Jl. Machine Learning Research (JMLR)*, pages 583–617, December 2002.

[2] S. Swift, A. Tucker, V. Vinciotti, N. Martin. Consensus Clustering and Functional Interpretation of Gene-expression Data. In *Genome Biology 5:R94*, 2004.

[3] V. Filkov and S. Skiena. Integrating Microarray Data by Consensus Clustering. In *International Journal on Artificial Intelligence Tools (IJAIT).*, 4:863–880, 2004.

[4] S. Monti, P. Tamayo, J. Mesirov, T. Golub. Consensus Clustering-A resampling-based method for class discovery and visualization of gene expression microarray data. In *Journal of Machine Learning* , 52:91–118, 2003.

[5] Thomas M. Cover and Joy A. Thomas. Elements of Information Theory. John Wiley and Sons, Inc., New York, N.Y, pages 193–218, 1991.