

## CSE601 - Project 2: Microarray Data Analysis: Appendix II

By: Amin Ghadersohi

**Hybrid Cluster** – As opposed to mathematics, in the realm of biology, and its derived sciences, many questions don't exactly have a formula for computing their solution. This maybe due to that fact that such a function has not been discovered yet, or that the problem falls in the category of NP or even NP-hard.

Clustering Microarray data falls in the latter categories. We can prove that clustering is in NP. Simply guess a clustering result and validate its results, by checking internal and external indices of the results. Or if ground truth data is available verify against that. It can further be shown that clustering is NP-hard by a simple reduction from the famous SAT problem. Very briefly, the proof sketch will be along the lines of showing: for all  $x$  in SAT  $\Leftrightarrow f(x)$  is in Clustering, by converting a Boolean formula  $\Phi$  into an instance of the clustering problem via the many-one reduction function  $f$ .

As we explained earlier, after guessing, or more desirably, computing a set of clusters, we must verify the results. Currently, there exist many clustering algorithms such as Diana, DBScan, and KMeans, each of which perform different given different kinds of data. However, it is not good to assume that a property of a dataset is going to be preserved by its sub-datasets. For instance, a very sparse dataset may have a few points that are very densely related. This data may cluster well using a hierarchal based algorithm, but show only one cluster in a density based algorithm.

There already exist many methods for validating results of clustering algorithms (ref 1). However, the problem is still hard because it is not guaranteed that a single algorithm will give best results every time. It is this heuristic nature of the problem that has forced so many people to come up with different algorithms for different kinds of data. But how do we know what kind of data we have?

We can make use of visualization to compare different datasets and decide what algorithm would be most suitable. But is this really efficient, or even accurate? Visualization of Microarray data is usually a transformation of the genetic data from a high number of dimensions to 2 or 3 dimensions. Clearly, there is a loss of information. Moreover, it is not a very easy task for a human to decide properties of a large dataset just by looking at its visualization and calculating its properties.

A more feasible solution is to create a uniform framework for ranking the results of clustering algorithms. As a proposal, the ranking function  $rank(\mathbf{d})$  can be defined as

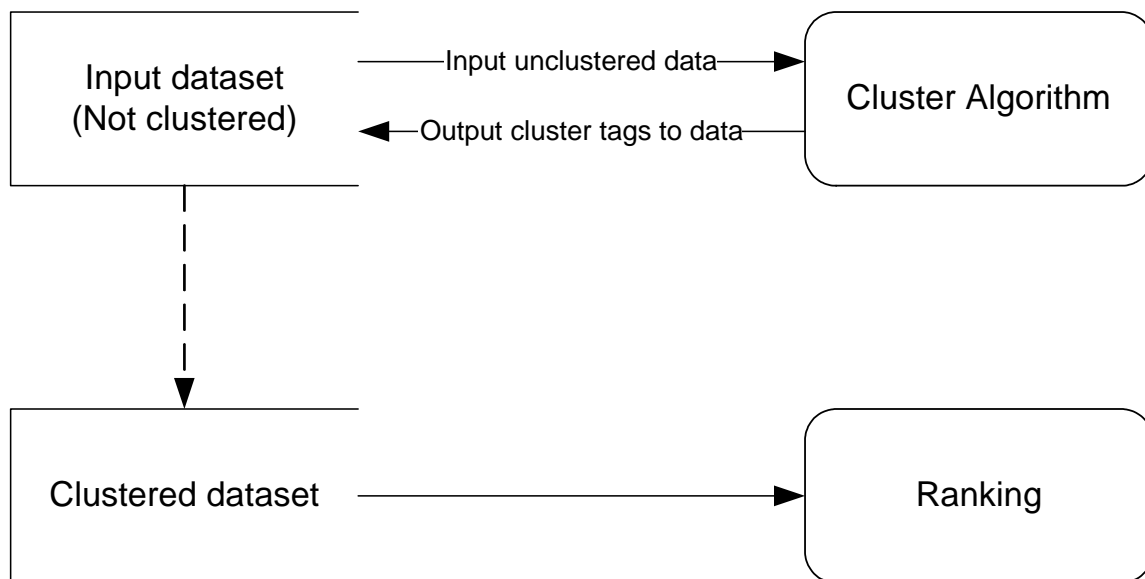
## CSE601 - Project 2: Microarray Data Analysis: Appendix II

By: Amin Ghadersohi

<clustered dataset>  $\rightarrow$  real number  $r$ , where  $r$  is between 0 and 1. To be effective, this ranking function must take into consideration as much information as it can. Also, we can't make use of any ground truth data, because if we knew the results, we wouldn't need to rank anything. We can use ground truth data later to verify the results of the ranking function.

There exist statistical methods such as sum of squared error for calculation relation of data to each other, and the difference of the correlation of the clustering results compared to a random clustering run. So once again, we can see that researchers have already developed methods for evaluating clustering results. If we were to take assign an importance to each of these measurements and take a normalized weighted average of them, we would be left with a real number  $r$  between 0 and 1; the same thing that we wanted  $rank(\mathbf{d})$  to compute. The importance of each component can be calculated from other measures of the dataset.

What we now have is a framework for the evaluating clustering algorithms, ranking their results. It should not be hard to see that a hybrid algorithm can easily make use of this and give us the best results that any algorithm would have given us.



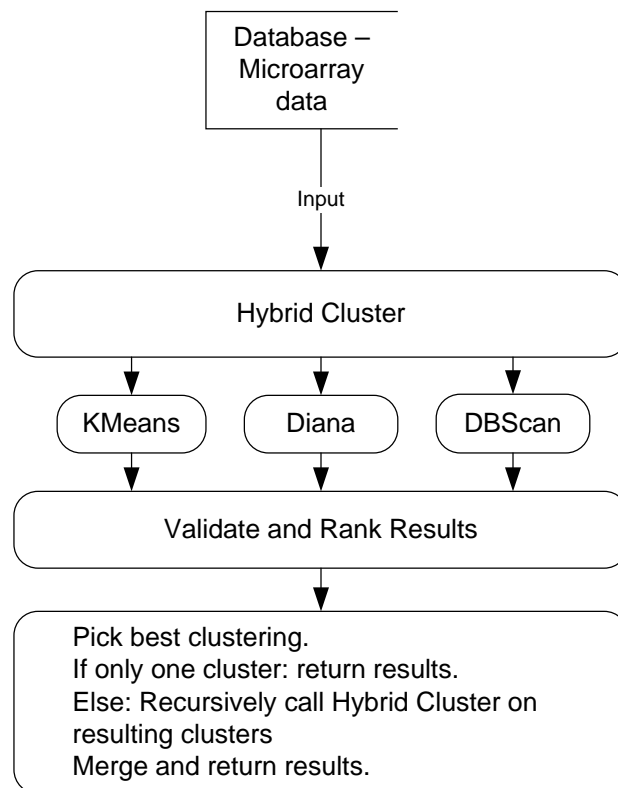
**Figure 1: Generalization of a clustering algorithm.**

## CSE601 - Project 2: Microarray Data Analysis: Appendix II

By: Amin Ghadersohi

We made use of object oriented design to create an abstract notion of a clustering algorithm and its results (figure 1). *Hybrid Cluster* takes as input a dataset and a target cluster count, or minimum rank threshold. The last two parameters tell the program when to stop clustering.

*Hybrid Cluster* can work in one pass, in that it can simply call all the algorithms it knows, rank the results and return the best one, or it can work recursively as follows. Given input dataset  $d$ , call each known algorithm and force it to give 2 to 3 clusters by adjusting its arguments. If after a set number of retries it returns only 1 cluster, move to next algorithm. If it returns more than 3, take that into consideration when ranking the results. In other words if the algorithm does not want to give wanted results it better have a good reason why, namely it better have found really good clusters. Take the results of the algorithm that ranked the best. Call *Hybrid Cluster* recursively on the derived clusters until the desired cluster count is met, or the clusters that are being formed have a rank less than the inputted minimum rank. Then merge the results of the recursion and return the results (figure 2).



## **CSE601 - Project 2: Microarray Data Analysis: Appendix II**

By: Amin Ghadersohi

The first advantage of using *Hybrid Cluster* is that it eliminates the step of manually finding the best algorithm for a given dataset. Moreover, by combining multiple algorithms, we not only have created a function whose range is larger than any other, and whose domain is better defined than other algorithms. Even though, this does not mean that the results are 100% true, it does mean that the results are closer to the truth.

The next logical step would be to define a better ranking function that analyzes the results of clustering algorithms more accurately. A simple approach would be to observe how biologists determine whether they like the results or not. Since the nature of the data is multi-dimensional, any evaluation function will have to take into account different perspectives of the data. The framework that we have created allows for the refinement of the rank function and addition of new algorithms. In the future we hope find better means of finding the right parameters for the sub-algorithms. This will help in the recursive case of the algorithm, in that we don't want it to divide three clusters into two, or two clusters as one.

### **References:**

1. Susmita Datta and Somnath Datta. Comparisons and validation of statistical clustering techniques for microarray geneexpression data. *Bioinformatics* Vol. 19 no. 4 2003, pages 459-466