

Application of Affinity Propagation on a large breast cancer data set ¹

D. Soria^a, F. Ambrogi^b, P. Boracchi^b, J.M. Garibaldi^a, E. Biganzoli^{b 2}

^a*School of Computer Science, University of Nottingham, Nottingham, UK*

^b*Istituto di Statistica Medica e Biometria "G.A. Maccacaro",*

Università degli Studi di Milano, Milano, Italy

Abstract: The Affinity Propagation (AP) algorithm is applied to a breast cancer case series subtyping using traditional biological markers. This data set was used to compare the results of AP with the results obtained with standard algorithms. The AP algorithm also provides a procedure to determine the number of profiles to be considered.

Results from Affinity Propagation are consistent with the results already obtained having the advantage of providing an indication about the number of clusters. Moreover, results also provide novel insights, with respect to conventional approaches. As a matter of fact, for the Abd El-Rehim *et al.* data, AP indicates either two, three, four, five, six or eight as an appropriate number of clusters.

Keywords: Clustering for large data-sets, Affinity Propagation, Breast Cancer

1. Introduction

Genomic analysis renewed interest in clustering techniques. After the paper of Eisen *et al.* (1998), proposing hierarchical clustering and the visual inspection of the dendrogram to discover unknown pattern of gene associations, the use of clustering has become more popular especially for discovering profiles in cancer with respect to high-throughput genomic data. An important application of the Eisen method is the work of Perou *et al.* (2000) on breast cancer. One of the main problems connected with cluster analysis is the choice of the number of clusters. In classical cluster analysis it is customary to use indices to compare one cluster solution to other ones and to choose the solution suggested as optimal. In a previous application by Ambrogi *et al.* (2006), different indices were used to select an optimal partition. It is worth noting that the visual inspection of the dendrogram suggested by Eisen *et al.* (1998) is an informal method to determine the number of clusters. Such a procedure was criticized in Goldstein *et al.* (2002) as it can cause difficulty in assessing the reproducibility of the grouping.

In this work, a new clustering algorithm, the Affinity Propagation [AP, Frey and Dueck (2007)], will be adopted to cluster breast cancer patients in order to evaluate its performance with respect to traditional applications. A dataset of breast cancer, already published in literature by Abd El-Rehim *et al.* (2005) was considered in our study. Although AP does not automatically determine the number of clusters, it provides a consistent method to suggest the number of groups to be created.

2. Material and methods

2.1 Case data

Tumours of an independent dataset of 1076 patients from the Nottingham Tenovus Primary Breast Carcinoma Series, which were evaluated by Immunohistochemistry (IHC) for 25 mark-

¹This study was, in part, funded by the BIOPTRAIN FP6 Marie-Curie EST Fellowship (FP6-007597).

²Address of correspondence: School of Computer Science, University of Nottingham, Jubilee Campus, Wollaton Road, Nottingham, NG8 1BB, UK

ers, were analysed. Levels of IHC reactivity were categorized using a modified H-score (values between 0-300). Further description of these data can be found in Abd El-Rehim *et al.* (2005).

2.2 Affinity Propagation

The clustering technique Affinity Propagation [AP, Frey and Dueck (2007)] will be adopted for grouping tumours with similar biological characteristics. As other clustering algorithms, this method uses data to find a set of centers K^* (called exemplars) to maximise the net similarity $\sum_{i \in K^*} \max_{k \in K^*} s(i, k) + \sum_{k \in K^*} p(k)$, where $s(i, k)$ and $p(k)$ are called, respectively, similarities and preferences (they will be defined subsequently). This objective function represents a score of how appropriate the exemplars are for explaining the data. Moreover, the algorithm determines the exemplars from real data points. These exemplars should be representative objects of the data structure among the observations.

AP combines properties of different classes of clustering algorithms. On one hand, algorithms like hierarchical clustering are based on grouping pairs of objects with high affinity. On the other hand model-based clustering uses a probability model based on a mixture of class conditional distributions. AP uses both pairs comparison and a probability model to determine the optimal grouping. According to a more technical point of view, AP can be derived as the sum-product algorithm in a graphical model describing the mixture model [Frey and Dueck (2006)]. This method recursively transmits messages between pairs of data points until a good set of exemplars, and corresponding clusters, emerges.

The first step for the algorithm implementation is to choose a measure of *similarity*, $s(i, k)$, between all pairs of data points. In AP terminology, $s(i, k)$ quantifies how well the data point k is suited to be the exemplar for data point i . Generally, as similarity, it is used the negative Euclidean distance.

The second step is about the choice of the values of *preferences* which represent a measure ($p(i)$) of how much data point i is candidate to be an exemplar. Rather than fixing the number of exemplars, AP assigns to every point a value of preference. In general, data points with larger values of $p(i)$ are more likely to be chosen as exemplars. At the beginning, the AP simultaneously considers all data points as potential exemplars and the initial value of the preferences is usually set equal to the median of all input similarities. However, high values of the preferences $p(i)$ will cause AP to find many exemplars (clusters), while low values will lead to a small number of exemplars (clusters).

The number of identified exemplars is influenced by the values of the input preferences, but also emerges as a result of the message passing structure. There are two kinds of message being passed between each pairs of data points that represent the relationship between them: ‘*responsibility*’ which is sent from data point i to candidate exemplar k and measures how well-suited point k is to be the exemplar for point i , taking into account other potential exemplars for point i ; and ‘*availability*’, sent from candidate exemplar k to point i , which reflects the evidence for point i to choose point k as its exemplar, considered that other points may have k as an exemplar. The messages are recursively updated using specific rules for a fixed number of iterations or until a stable clustering result.

An advanced characteristic of AP is that it determines the number of clusters on the basis of the message passing architecture and of the points that are most representative, given an initial common preference. It is possible to see the effect of the value of the input preference on the number of clusters by a graphic with the value of the common initial preference on the x-axis and the respective number of clusters on the y-axis. In this way, the value to adopt in the analysis can be established in correspondence with plateaus that are observable in this

graphic.

AP differs from K-means in several aspects. Firstly, K-means is based on the minimisation of the Euclidean distance between data points and cluster centers, which are computed as mean points within a cluster, while AP uses general dissimilarities and actual data points as centers. Secondly, K-means begins with an initial set of randomly selected centroids and iteratively refines this set so as to decrease the sum of squared error. AP, instead, simultaneously considers all data points as potential exemplars [Frey and Dueck (2007)]. Finally, the use of squared Euclidean distance as measure of dissimilarity between data points and centroids in K-means not only limits the type of data variables that can be considered, but it can also make the determination of the cluster means non-robust to outliers [Bishop (2006)]. Furthermore, as the K-means algorithm can be thought as an EM algorithm, it is worth noting that the use of a general dissimilarity measure in such a context implies an increased complexity in the M step which is generally faced by considering as possible cluster centers actual data points. By doing so, the algorithm can be implemented for any choice of similarity measure [Bishop (2006)].

3. Results

On the same data described in Abd El-Rehim *et al.* (2005), where a hierarchical algorithm was used to classify patients in six groups, the AP technique was applied and a plot of the effect of the input preference on the number of cluster was produced (Fig. 1(left)). Looking at the plateaus, several solutions were proposed, dividing the dataset in two, three, four, five, six and eight clusters. The solution in two clusters corresponds to the well known grouping of breast cancer in ER positive and negative tumours. The most interesting solution, in terms of comparison with the original work, was the one obtained using a preference value to get six clusters. The original groups determined by Abd El-Rehim *et al.* (2005) can be described as two luminal and ER positive groups, two HER2 positive groups, which show differences in MUC1 and E-cadherin expression, and a strong basal epithelial cluster. The last group, which contains only four patients, appears to be characterised by a basal phenotype with high p53. The distribution of patients assigned to the six clusters by AP and by the original hierarchical clustering is reported in Fig. 1(right). Labels to the AP groups were assigned resorting to the weighted kappa index for the agreement between classifications [Cohen (1960)] and the ordering for which we obtained the highest index was considered (weighted kappa = 0.302). Although there is not a good agreement between the two classifications, a couple of AP groups can also be described in terms of the original work. In particular, AP cluster 1 and 2 seem to express luminal characteristics, while AP group 5 appears to be a basal group.

When using preference values for AP to obtain three, four, five and eight clusters, the groups that emerged were simply subdivisions of previously obtained groups.

4. Discussion

Cluster analysis is a powerful technique to explore complex diseases and improve prognosis. The recent literature on omic data is rich of new methods of cluster analysis able to deal with huge datasets. Moreover techniques of visualization are usually adopted to suggest the number of clusters as highlighted in Eisen *et al.* (1998). At the same time many papers, like Goldstein *et al.* (2002), warn against the possible misuse of clustering techniques. The choice of the number of clusters is one of the main problems to be faced when applying this kind of analysis. The possibility to use algorithms that incorporate a criterion for the choice of the optimal partition is one of the achievement of recent developments in this research field. An example of this kind of techniques is the AP algorithm, which is characterized by a simple software implementation and has the ability to suggest the cluster number.

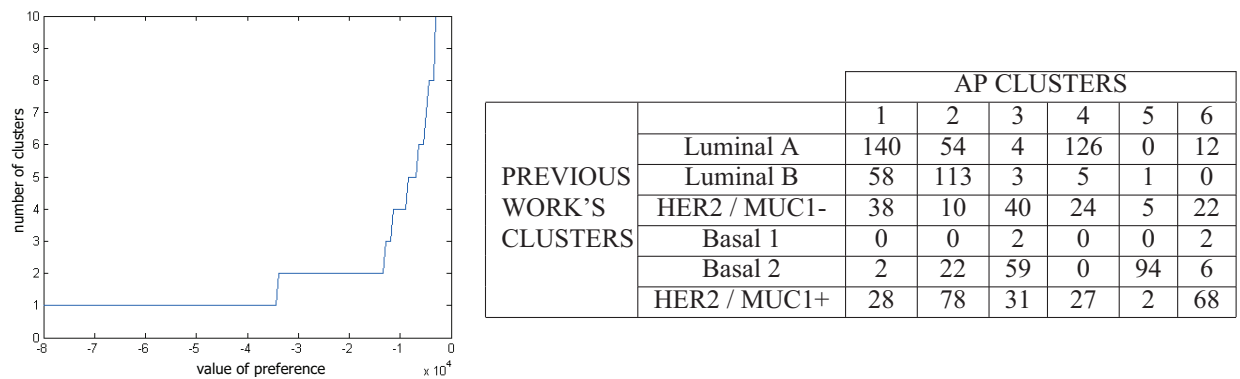


Figure 1: (left): The effect of the input preference value on the number of clusters; (right): The distributions of subjects between new and old classifications

In this work, the AP clustering technique was described and an application over a known breast cancer data set was presented. AP results were compared with those coming from the original hierarchical clustering. Even though there is not a complete agreement between solutions, the AP algorithm suggests novel insights in terms of different number of clusters to consider and of possible split of similar groups.

As a future work, it would be interesting to characterise the AP groups in terms of clinical outcome and response to treatments.

References

- Abd El-Rehim D., Ball G., Pinder S., Rakha E., Paish C., Robertson J., Macmillan D., Blamey R. and Ellis I. (2005) High-throughput protein expression analysis using tissue microarray technology of a large well-characterised series identifies biologically distinct classes of breast cancer confirming recent cDNA expression analyses, *Int. Journal of Cancer*, 116, 340–350.
- Ambroggi F., Biganzoli E., Querzoli P., Ferretti S., Boracchi P., Alberti S., Marubini E. and Nenci I. (2006) Molecular subtyping of breast cancer from traditional tumor marker profiles using parallel clustering methods, *Clinical Cancer Research*, 12, 3, 781–790.
- Bishop C. (2006) *Pattern Recognition and Machine Learning*, New York: Springer edition.
- Cohen J. (1960) A coefficient of agreement for nominal scales, *Educational and Psychological Measurement*, 20, 37–46.
- Eisen M., Spellman P., Brown P. and Botstein D. (1998) Cluster analysis and display of genome-wide expression patterns, *Proc Natl Acad Sci U S A*, 95, 14863–8.
- Frey B. and Dueck D. (2006) Mixture modeling by affinity propagation, *Advances in Neural Information Processing Systems*.
- Frey B. and Dueck D. (2007) Clustering by passing messages between data points, *Science*, 315, 5814, 972–6.
- Goldstein D., Debashis G. and Conlon E. (2002) Statistical issues in the clustering of gene expression data, *Statistica Sinica*, 12, 219–240.
- Perou C., Sorlie T., Eisen M., Van De Rijn M., Jeffrey S., Rees C., Pollack J., Ross D., Johnsen H., Akslen L., Fluge O., Pergamenschikov A., Williams C., Zhu S., Lonning P., Borresen-Dale A., Brown P. and Botstein D. (2000) Molecular portraits of human breast tumours, *Nature*, 406, 747–752.