Affinity Propagation Enhanced by Estimation of Distribution Algorithms

Roberto Santana
Computational Intelligence Group
Universidad Politécnica de Madrid, Spain
roberto.santana@upm.es

Concha Bielza
Computational Intelligence Group
DIA, Universidad Politécnica de Madrid, Spain
mcbielza@fi.upm.es

Pedro Larrañaga
Computational Intelligence Group
DIA, Universidad Politécnica de Madrid, Spain
pedro.larranaga@fi.upm.es

ABSTRACT

Tumor classification based on gene expression data can be applied to set appropriate medical treatment according to the specific tumor characteristics. In this paper we propose the use of estimation of distribution algorithms (EDAs) to enhance the performance of affinity propagation (AP) in classification problems. AP is an efficient clustering algorithm based on message-passing methods and which automatically identifies exemplars of each cluster. We introduce an EDA-based procedure to compute the preferences used by the AP algorithm. Our results show that AP performance can be notably improved by using the introduced approach. Furthermore, we present evidence that classification of new data is improved by employing previously identified exemplars with only minor decrease in classification accuracy.

Categories and Subject Descriptors

G.1 [**Optimization**]: Global optimization; G.3 [**Probabilistic methods**]

General Terms

Algorithms

Keywords

Clustering, classification, estimation of distribution algorithms, cancer, gene expression profiles

1. INTRODUCTION

The analysis of gene expression data can be used to identify the primary anatomical site of origin of human tumors. This identification is important since different types of tumors will require different treatment strategies. However, there are a number of obstacles associated to the automatic classification of new tumor samples. Similar tumors may be produced by different tumor pathways and have different gene expression profiles. Furthermore, in many cases the genes involved in the tumor pathways are only partially known. Classification methods are required not only to distinguish between tumor subtypes, but also to identify exemplary genetic profiles within the same class of tumor subtypes. In this paper we propose a classification algorithm that combines the application of affinity propagation, a very efficient unsupervised classification method, with an estimation of distribution algorithm.

Affinity propagation [5] is a message-passing-based clustering algorithm that, given a set of points and a set of similarity values between the points, finds clusters of similar points, and for each cluster gives a representative example or exemplar. The algorithm has been applied to a variety of problems outperforming other traditional clustering algorithms [5, 9, 18].

Estimation of distribution algorithms (EDAs) [8, 10] are a class of evolutionary algorithms that apply probabilistic modeling of the selected solutions instead of crossover operators. The rationale behind probabilistic modeling is to explicitly model the relationships between the variables of the problem in terms of probabilistic dependencies which are captured by probabilistic graphical models (PGMs). The PGMs are employed to generate new solutions that will likely resemble the selected points.

We apply AP to $12,533$ genes expressed in carcinomas of the prostate, breast, lung, ovary, colorectum, kidney, liver, pancreas, bladder/ureter, and gastroesophagus [20]. We intend to investigate the use of clustering techniques in the context of supervised classification. Our results show that the quality of the affinity propagation clustering can be increased by an automatic selection of its parameters. Concerning the clustering of the gene expression data, even without a previous filtering of genes, the algorithm can achieve a high classification accuracy.

The paper is organized as follows: In the next section, the affinity propagation algorithm is presented and its main components are discussed. In Section 3 exemplar-based classification is introduced. Section 4 presents the estimation of distribution algorithms used in our experiments and how to combine affinity propagation with EDAs to improve classification accuracy. Section 5 introduces the tumor dataset used for our experiments, the experimental framework to evaluate our proposal, and the numerical results. The conclusions of our paper and some lines for future research are given in Section 6.

GECCO'11, July 12–16, 2011, Dublin, Ireland.

2. AFFINITY PROPAGATION

Clustering methods [6] are used to group objects into different sets or clusters, in such a way that each cluster comprises similar objects. Clusters can then be associated to labels that are used to describe the data and identify their general characteristics. Among the the best known clustering algorithms are k-means [7] and k-center clustering [1].

Affinity propagation takes as input a matrix of similarity measures between each pair of points $s(\mathbf{y}^i, \mathbf{y}^k)$. For each data point $\mathbf{y}^k$, a real number $s(\mathbf{y}^k, \mathbf{y}^k)$ is also entered as an initial input. The $s(\mathbf{y}^k, \mathbf{y}^k)$ values are called *preferences* and are a measure of how likely each point is to be chosen as exemplar. The algorithm works by exchanging messages between the points until a stop condition, which reflects an agreement between all the points with respect to the current assignment of the exemplars, is satisfied. These messages can be seen as the way the points share local information in the gradual determination of the exemplars.

There are two types of messages to be exchanged between data points. The *responsibility* $r(i,k)$, sent from data point $\mathbf{y}^i$ to candidate exemplar point $\mathbf{y}^k$, reflects the accumulated evidence for how well-suited point $\mathbf{y}^k$ is to serve as the exemplar for point $\mathbf{y}^i$, taking into account other potential exemplars for point $\mathbf{y}^i$. The *availability* $a(i,k)$, sent from candidate exemplar point $\mathbf{y}^k$ to point $\mathbf{y}^i$, reflects the accumulated evidence for how appropriate it would be for point $\mathbf{y}^i$ to choose point $\mathbf{y}^k$ as its exemplar, taking into account the support from other points that point $\mathbf{y}^k$ should be an exemplar.

The availabilities are initialized to zero: $a(i,k) = 0$. Then, the responsibilities are computed using the rule:

$$r(i,k) \leftarrow s(\mathbf{y}^i, \mathbf{y}^k) - max_{k'|k' \neq k}\{a(i,k') + s(\mathbf{y}^i, \mathbf{y}^{k'})\} \quad (1)$$

Equation (1) allows all the candidate exemplars to compete for ownership of a data point. Evidence about whether each candidate exemplar would be a good exemplar is obtained from the application of the following availability update:

$$a(i,k) \leftarrow min\left\{0, r(k,k) + \sum_{i'|i' \notin \{i,k\}} max\{0, r(i',k)\}\right\} \quad (2)$$

In Equation (2) only the positive portions of incoming responsibilities are added, because it is only necessary for a good exemplar to explain some data points (positive responsibilities), regardless of how poorly it explains points with negative responsibilities. To limit the influence of incoming positive responsibilities, the total sum is thresholded so that it cannot go above zero.

The *self-availability* $a(k,k)$ is updated differently:

$$a(k,k) = \sum_{i'|i' \neq k} max\{0, r(i',k)\} \quad (3)$$

For a point $\mathbf{y}^i$, the value of k that maximizes $a(i,k) + r(i,k)$ either identifies point $\mathbf{y}^i$ as an exemplar if $k = i$ ($c^i = i$, where c^i refers to the exemplar of point i), or identifies the data point that is the exemplar for point $\mathbf{y}^i$.

Update rules described by Equations (1), (2) and (3) require only local computations. Additionally, messages are exchanged only between pairs of points with known similarities. AP also takes advantage of a sparse matrix of similarities when such distribution of similarity values is available.

The message-passing procedure may be terminated after a fixed number of iterations, when changes in the messages fall below a threshold, or after the local decisions stay constant for some number of iterations. The pseudocode of affinity propagation algorithm is shown in Algorithm 1.

Algorithm 1: **Affinity propagation**

```
1 Initialize availabilities a(i,k) to zero ∀i,k
2 do {
3     Update, using Equation (1), all the responsibilities
      given the availabilities
4     Update, using Equation (2), all the availabilities
      given the responsibilities
5     Combine availabilities and responsibilities to ob-
      tain the exemplar decisions
6 } until Termination criterion is met
```

The measure used to compute the similarity between the points and the preference values influence the outcome of affinity propagation. Although a detailed investigation of the effect of the different similarity measures in the convergence results and accuracy of the algorithms has not been conducted, it is clear that some similarity measures could be more suitable to capture the commonalities between the points. In the present work we have investigated the effect of using four ways to measure the similarity between points. These measures are computed from three distance measures between points: Euclidean, cosine, and linear and Spearman correlation.

The Euclidean distance $E(i,j)$ is computed between vector points $\mathbf{y}^i$ and $\mathbf{y}^j$. The cosine distance is equal to one minus the cosine of the angle between vectors $\mathbf{y}^i$ and $\mathbf{y}^j$. The correlation distance between vector points $\mathbf{y}^i$ and $\mathbf{y}^j$ is simply one minus the linear and Spearman correlations $c(i,j)$ are computed taking as observations the components of the vectors. We use as a similarity measure the maximum distance between any pair of points minus the distance between each pair of points. This means that points at a smaller distance are more similar.

A heuristic method to compute initial values for preference measures, is to take the median value of the similarity matrix. This median similarity value is assigned as the preference value for all the points. In this setting, the algorithm is told that all the points are equally likely to be selected as exemplars.

We take the median similarity value as a base preference value s_b and consider three different ways to assign the preferences to each point. Points can either have their preference equal to the base preference value $s(\mathbf{y}^k, \mathbf{y}^k) = s_b$, half of this value ($s(\mathbf{y}^k, \mathbf{y}^k) = \frac{s_b}{2}$) or twice the base preference value ($s(\mathbf{y}^k, \mathbf{y}^k) = 2s_b$). These assignments intend to bias the likelihood that a given point is an exemplar. The way in which the different preferences are assigned to each point is treated in Section 4.

3. EXEMPLAR-BASED CLASSIFICATION

The affinity propagation algorithm will maximize the net similarity $\mathcal{S}$ which is a sum of the similarity between each point to its exemplar plus the preference values of the exem-

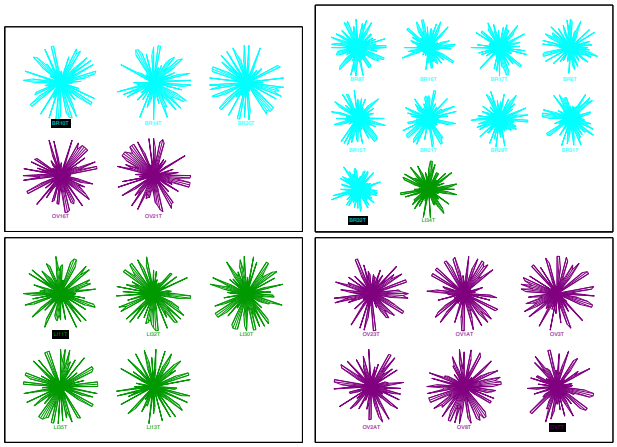

Figure 1: Example of clustering produced by affinity propagation on a set of 26 points with a-priori known classes. Each of the four charts corresponds to a different cluster. The exemplar of each cluster is highlighted with a black label over its name. Points that belong to the same class are represented using the same color.

plars. $\mathcal{S}$ can be used to measure the quality of the clustering [5]. However, there are some situations where we would like to take into consideration other criteria to evaluate the clustering quality. One of these criteria can be the number of clusters. The user may be interested in good clusterings that have a constrained number of clusters. Furthermore, even if AP is generally applied as an unsupervised classification algorithm, it could be the case that some previous classification is known for all or a subset of the points. The a priori classification may not necessarily be related to the similarity measure used for clustering, but we may be interested in evaluating the clustering quality in terms of the known classes.

3.1 Penalized exemplar-based accuracy

We propose a classification accuracy measure for affinity propagation in the cases where a priori information about the classes of the points is known. The measure simultaneously evaluates the capacity of AP to group in the same cluster points that belong to the same (known) class, with its capacity to produce small number of clusters.

The penalized exemplar-based accuracy is defined as:

$$E_{Acc} = \frac{\left(\sum_{i=1}^{m}\sum_{j=1}^{|C_i|} I(c(\mathbf{y}^j), c(\mathbf{y}^e))\right) - m}{n} \quad (4)$$

where m is the number of clusters, $|C_i|$ is the number of points assigned to cluster i, n is the number of points, and I is the indicator function that is one when the (a-priori known) class of point $\mathbf{y}^j$ coincides with the class of the exemplar point $\mathbf{y}^e$ in cluster i, zero otherwise.

The key idea of the penalized exemplar-based accuracy is to assign to each point, as a putative class, the class of its exemplar. Then, to measure the accuracy of the classification, the number of points that are correctly classified is counted. To account for the fact that the classes of the exemplars are assumed to be known for classification, we subtract the number of exemplars from the total number of points correctly classified. This way, clusterings with a few number of clusters are also favored.

Figure 1 shows an example of a possible clustering of 26 points for which the class is a-priori known. In terms of the classification, a good cluster should comprise points that are known to belong to the same class. To compute E_{Acc}, we calculate the number of points correctly classified by the exemplar in each cluster, then $E_{Acc} = \frac{3+9+5+6-4}{26} = 0.73$. Notice that if the penalty on the number of clusters is not considered, then the accuracy is higher (0.88), but this value hides the fact that, since we are using the knowledge of the true class in their classification, the class assignment for the exemplars will be always correct.

4. ESTIMATION OF DISTRIBUTION ALGORITHMS

We evaluate the influence that the preference assigned to each point has on the convergence and accuracy of AP. To this end, we will use EDAs to evolve set of preferences. The optimization problem consists of finding an n-dimensional vector $\mathbf{s}$ of initial preference configurations such that each of its components $s_i = s(\mathbf{y}^i, \mathbf{y}^i)$, $i = 1, \ldots, n$ and $s_i \in \{\frac{s_b}{2}, s_b, 2s_b\}$ and the E_{Acc} measure is maximized. Let $\mathbf{X}$ represent the vector of problem variables and $\mathbf{X}$ one of its possible instantiations. We use a ternary representation $x_i \in \{0, 1, 2\}$ to represent the possible assignments for s_i. By using only three possible values for each variable, we focus on analyzing the effect of the modification of the preference values above and below the median of the similarity measures.

Three different variants of EDAs are used. Each variant captures and uses different relationships between the problem variables, effectively implementing diverse search strategies. The first variant uses a univariate marginal product model in which all variables are independent, i.e. no dependencies are represented in the model. The joint probability distribution of the univariate marginal distribution algorithm (UMDA) [10] over $\mathbf{x} = (x_1, \ldots, x_n)$ can be factorized as follows:

$$p_{UMDA}(\mathbf{x}) = \prod_{i=1}^{n} p(x_i). \quad (5)$$

The second EDA learns a probabilistic model based on a tree. In this model, each variable may depend on no more than one variable that is called the parent. The probability distribution $p_{Tree}(\mathbf{x})$ used by Tree-EDA [19] is defined as

$$p_{Tree}(\mathbf{x}) = \prod_{i=1}^{n} p(x_i \mid pa(x_i)), \quad (6)$$

where $pa(X_i)$ is the parent of variable X_i in the tree, and $p(x_i \mid pa(x_i)) = p(x_i)$ when $pa(X_i) = \emptyset$, i.e. when X_i is the root of the tree. Probabilistic trees can be represented by directed acyclic graphs.

The third EDA is the estimation of distribution Bayesian algorithm (EBNA) [4]. A Bayesian network [12] can be seen as a generalization of a tree where each variable can have multiple parents. The structure St of a Bayesian network is a directed acyclic graph (DAG) that represents a set of conditional independence assertions about the variables on

$\mathbf{X}$. It represents the assertions that X_i and non-descendant variables $\{X_1, \ldots, X_n\} \backslash \mathbf{Pa}_i^{St}$ are conditionally independent given $\mathbf{Pa}_i^{St}$, $i = 1, \ldots, n$. The set of variables $\mathbf{Pa}_i^{St}$ are called the parents of X_i. The parameters of the model are the set of marginal and conditional probability distributions $p(X_i|\mathbf{Pa}_i^{St}$ corresponding to the (in)dependence relationships represented in the structure.

4.1 EDA implementation

All the EDAs were implemented in Matlab® using the MATEDA-2.0 software [16]. Learning of the Bayesian networks within MATEDA-2.0 uses the Matlab Bayes Net (BNT) toolbox [11]. The Matlab® implementation of affinity propagation provided by the authors [5] was used for the clustering experiments. The application of the affinity propagation method is very efficient in terms of time. The computational complexity of the EDA depends on the complexity of the probabilistic models used [17].

4.2 Related work

In the context of evolutionary computation, affinity propagation has been used to learn marginal product models in the learning phase of estimation of distribution algorithms [18].

In [2], the affinity propagation algorithm was applied to a problem of breast cancer subtyping using traditional biologic markers. The authors acknowledged the capacity of the algorithm to automatically determine the number of profiles to be considered. Leone et al [9] use a variant of affinity propagation to classify a test dataset monitoring the expression levels of more than 7000 genes for 42 patients with different subtypes of brain cancer. The data had been previously classified into five diagnosis types [15]. The authors investigate how the preferences influence the number of clusters, but no automatic strategy was proposed to set these values.

5. EXPERIMENTS

5.1 Clustering of gene expression data

The tumor classification problem consists in classifying a set of different carcinomas according to the gene expression data from 12,533 genes. We use the dataset of tumor gene expression data, and the same partitioning strategy of the cases previously used in [20]. In this dataset, the 174 samples are divided into a training data set comprising 100 tumors and a test data set with 74 tumors.

The training set of 100 primary carcinomas was used to directly find the optimal set of preferences using EDAs. This set of tumors comprised 10 prostate adenocarcinomas, 9 bladder/ureter carcinomas, 12 infiltrating ductal breast adenocarcinomas, 10 colorectal adenocarcinomas, 11 gastroesophageal adenocarcinomas, 10 clear cell carcinomas of the kidney, 6 hepatocellular carcinomas, 9 serous papillary ovarian adenocarcinomas, 6 pancreatic adenocarcinomas, and 17 lung carcinomas. The test set of of 74 tumors was used to validate the results found by the EDAs.

5.2 Experimental framework

In the following experiments we investigate whether the optimization of the preference values contributes to enhance the performance of AP. Additionally, we study a number of factors that influence the behavior of the enhanced AP: the type of similarity measure, the class of probabilistic model used by the EDA and the fitness evolution along generations.

We also analyze the composition of the optimal solutions found by the algorithms, in particular the most frequent exemplars identified by AP are computed and compared for the different similarity measures. Finally, we select the subtype of infiltrating ductal breast adenocarcinomasand inspect the clusters in which the samples from this tumor class are grouped. The analysis of the clusters reveals differences in gene expression patterns of tumors assigned to different clusters.

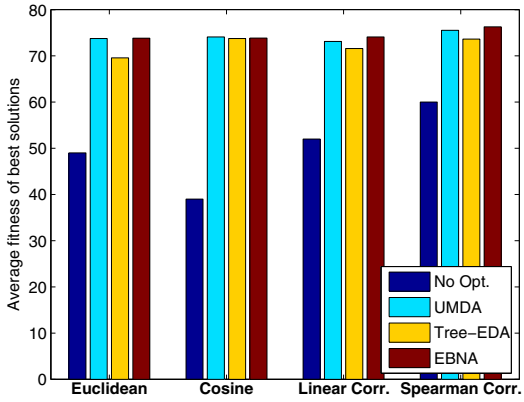

Figure 2: Best penalized exemplar-based classification values achieved by simple AP (without preference optimization) and average of the best clustering values reached by all EDAs for all similarity measures.

5.3 Classification experiments on the training set

Initial experiments were conducted to apply AP to the training set, i.e., 100 tumors were clustered according to their expression levels. Four different distances were used to compute the similarity measures between the 100 tumors. These distances were: Euclidean, cosine, linear correlation and Spearman correlation. From the number of clusters and classification accuracy the clustering quality measure proposed in Section 3.1 was computed.

In Figure 2 the blue bars indicate the E_{Acc} values achieved for each distance. It can be seen the figure that the best results are achieved using the Spearman correlation distance. On the other hand, the lowest E_{Acc} value was achieved using the cosine distance. Below we analyze, how the number of clusters and accuracy values determine the E_{Acc} values for the different measures.

In the next step, we use the different EDAs to find a set of preferences that maximizes quality measure. We apply UMDA, Tree-EDA and EBNA. All EDAs used a population size of 200 individuals and a maximum of 25 generations. Truncation selection with truncation parameter $T = 0.5$ was applied. For each similarity measure and EDA, 30 experiments were conducted. In Figure 2, the average best fitness value found by the EDAs is shown for all combinations of similarity measures and algorithms. All EDAs clearly im-

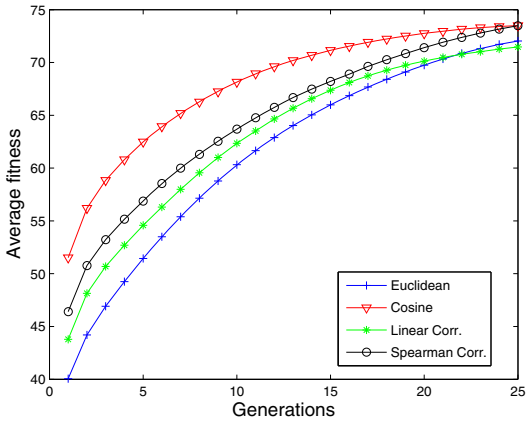

Figure 3: Average fitness of the populations for UMDA using different similarity measures.

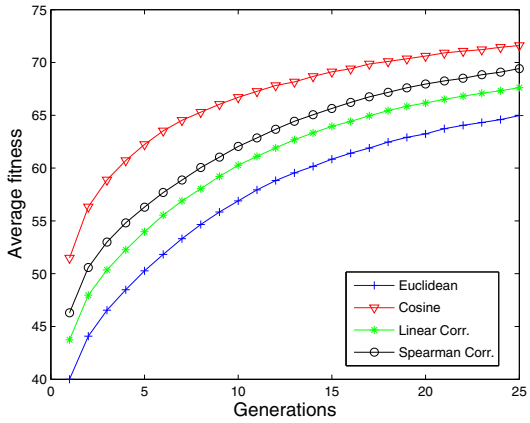

Figure 4: Average fitness of the populations for Tree-EDA using different similarity measures.

prove the results achieved by the simple AP algorithm. In terms of the best fitness, there are no important differences between the performance of the different EDAs, and more importantly, improvements are obtained for all similarity measures.

Figures 3-5 respectively show the average fitness of the population achieved by UMDA, Tree-EDA and EBNA in each generation for all similarity measures. In terms of the average fitness of the population, there are differences between the behavior of the algorithms for the similarity measures. The fitness values which are improved the most at the first generations are those computed from the cosine distance. However, as generations advance the average fitness computed from the cosine distance tends to stagnate and those computed from the Spearman correlation continue to improve. This behavior is particularly evident for EBNA (see Figure 5). Regarding the differences between the EDAs, there are no relevant differences although EBNA seems to slightly outperform the other algorithms. Perhaps the most important conclusion from the analysis of Figures 3-5 is that all EDAs require a relatively small number of evaluations to find preferences values that improve the penalized exemplar-based classification.

5.4 Significant genes in clusters found by AP

The cluster composition may reveal unexpected patterns shared by tumor samples. Particularly relevant is to investigate the way in which tumors of the same class area spread among different clusters. A further identification of the tumor genes that are significantly different within the cluster with respect to the other tumors could serve to discover these characteristic patterns.

To illustrate this type of analysis, we considered the distribution of the 12 infiltrating ductal breast adenocarcinomas among clusters. We chose one of the clusterings with high fitness ($E_{Acc} = 78$). Figure 1(top) shows the two clusters where the breast adenocarcinomas (represented with cyan colors) are located. In the first cluster (Cluster I) there are two ovarian samples mistakenly classified as breast adenocarcinomas. Similarly, in the second cluster (Cluster II) there is one liver tumor mistakenly classified as breast adenocarcinome. However, all breast adenocarcinomas are correctly classified.

To investigate the causes that may determine splitting tumors of the same class in two different clusters, we compute the variables (gene expressions) which are significantly different within the clusters with respect to the rest of tumors (i.e. all the tumors except those comprised in the cluster). The Wilcoxon rank sum test for equal medians was applied. We performed a two-sided rank sum test of the hypothesis that the two independent samples, in the vectors X and Y, come from distributions with equal medians, and returned the p-value from the test. For the sake of space, we report for each cluster only the two most significant genes in terms of the p-value.

For Cluster I, genes CDKN2B and CLCNKB have medians significantly under the values of the median for those tumors that are not included in the cluster, p-values 0.0001777 and 0.0001779, respectively. In the case of Cluster II, genes AI688098:wc92f08.x1 and HERV-K22 have medians significantly over the values of the median for those tumors that are not included in the cluster, p-values 0.000001967 and 0.000002265, respectively.

The CDKN2B gene lies adjacent to the tumor suppressor gene CDKN2A in a region that is frequently mutated and deleted in a wide variety of tumors. It has been previously reported that intragenic mutations of CDKN2B and CDKN2A occur in primary human esophageal cancers [22]. In humans, the CLCNKB gene encodes the chloride channel Kb protein, also known as CLCNKB. In addition, CLCNKA and CLCNKB are closely related (94% sequence identity) and are both expressed in mammalian kidney. Chloride channels regulate the movement of a major cellular anion and maintain intracellular pH and cell volume. Recent work suggest that several chloride channel families may contribute to the cancer phenotype and serve as novel targets for primary cancer therapy [21].

In our analysis, we did not find direct associations between the AI688098:wc92f08.x1 gene and cancer in the literature. However, AI688098:wc92f08.x1 has been included in a metagene approach for breast cancer recurrence prognosis [14]. Metagenes are gene expression signatures derived

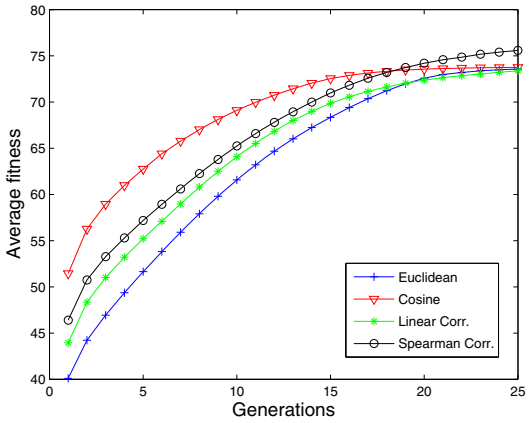

Figure 5: Average fitness of the populations for EBNA using different similarity measures.

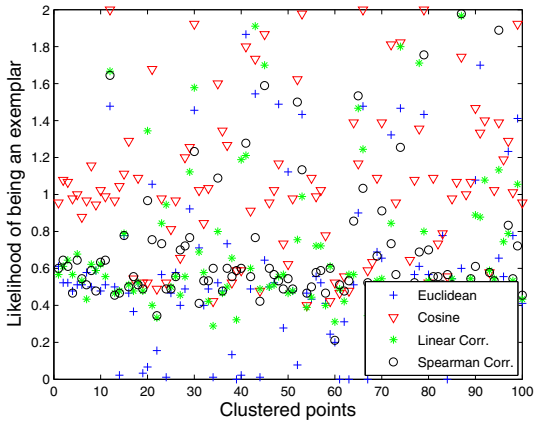

Figure 6: Likelihood that each tumor sample becomes an exemplar. The likelihood is computed from the set of best solutions found by all EDAs, i.e. 120 solutions are used for each similarity measure.

from microarray analyses. The HERV-K22 gene are human endogenous retroviruses (HERV) that are remnants of exogenous retroviruses that entered the germ line millions of years ago. HERVs are believed to be possible pathogenic agents in carcinogenesis [3].

The found genes can provide a possible direction for investigating causative genes and pathways in cancer. It is remarkable the fact that in the two clusters considered, the gene expressions of the significant genes have different signs with respect to the other tumors. It is an open question to investigate the similarity between tumor samples that are mistakenly classified. Also, it will be convenient to use results from the literature to further classify the significant genes found in each cluster in classical oncogenes or tumor suppressor genes [13].

5.5 Influence of the similarity measures

A fundamental question is to know whether there are tumor samples that clearly serve as exemplars accross the different used similarity measures. We consider that a tumor sample is likely to be an exemplar if it is consistently assigned by the EDAs with a high preference value. Since absolute preference values depend on similarity measures, we use relative preference values to compute the exemplar likelihood. Value 0 means that the point has, as preference value, half of the median value, value 1 means that the point has exactly the median value, and value 2 means that the point has twice the median value. We compute the average exemplar likelihood as the average of the relative preferences values of the best solutions found in each of the 30 experiments conducted for each EDA and similarity measure. An average likelihood equal 2 means that in all runs the point was assigned the highest preference value indicating its high probability to be an exemplar.

Figure 6 shows the likelihoods that each tumor sample becomes an exemplar for all tumors and all similarity measures. Candidates for being good exemplar accross different similarity measures can be identified as having exemplar likelihood values over 1.5 for two or more similarity measures. Additionally, from the analysis of Figure 6, different patterns in the behavior of AP for the used similarity measures can be extracted. It can be seen that the Euclidean measures consistently identifies some tumor samples as less likely to be an exemplar. This can be appreciated in the large number of blue crosses whose likelihood is near to zero. On the other hand, the cosine measure consistently identifies the same set of tumor samples as likely to be an exemplar, most of points with likelihood equal two are red triangles. There is also a large concentration of points with likelihood equal 0.5.

5.6 Classification experiments on the test set

The particular characteristics of the AP method make difficult a straightforward application of traditional validation strategies. This can be done in a number of ways. However, a necessary step to validate the AP clustering results is to determine whether exemplars identified for the training set are good at classifying other (unseen) tumor samples. Tumors from the test data may be assigned to already found exemplars based on the similarity to each exemplar. Nevertheless, this procedure does not use information about the similarities among the new tumor samples.

A differente alternative is to apply AP on the whole set of data, but biasing the preferences towards the values previously found on the training set. This way, we can evaluate the benefits of the preference values learned using the training set but the algorithms will also use, for the classification of the new samples, the similarities among them. In this way, more information is infused into the classification process.

Therefore, we applied AP to the complete set of tumor samples formed by 174 samples. Initially, the simple AP with all preference values equal to the median of the similarities, was applied. The results of the algorithm for all similarity measures can be seen in Figure 7 where they are represented using blue bars. It can be seen that clustering quality measures are around 50 for all the similarity measures.

We then used the best solutions found by the application

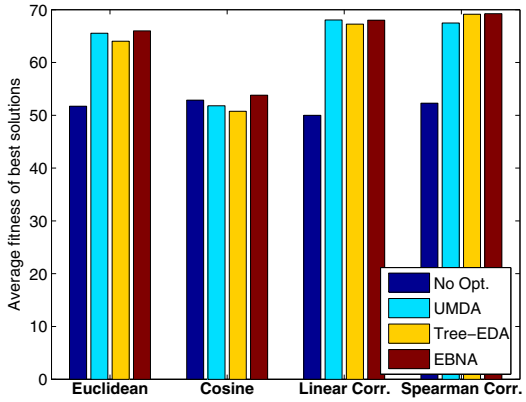

Figure 7: Extended set of tumor samples. Best clustering quality values achieved by simple AP (without preference optimization) and average of the best clustering values reached by all EDAs for all similarity measures.

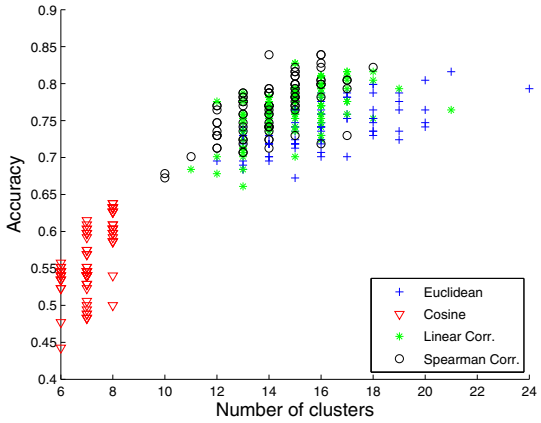

Figure 8: Number of clusters versus classification accuracy for the clusterings found by AP on the complete dataset. Preference values have been biased according to the optimal solutions found by EDA on the test dataset.

of the EDAs on the training set as a way to bias the computation of the clusters. In this case, the preferences values of the tumors in the training data (100 tumor samples) were computed according to the relative preference value of the best solutions (i.e., half, equal or twice the value of the median similarity value computed for the complete dataset). The preference values of the other 74 were assigned negative preferences equal to twice the median of the similarity. Negative preference values guaranteed that none of the new points will be used as an exemplar. Therefore, new points will be assigned to clusters whose exemplar belongs to the training set. The algorithm does not use previously computed clusters, the only bias is passed to AP in the preferences values.

Figure 7 shows the clustering quality results computed from the best solutions found by all the EDAs and all the similarity measures. Notice that in this case, the optimization algorithms have not been applied to compute the preferences. Previous results have been applied to the enlarged dataset. However, it is clear that an improvement in the clustering quality is achieved for three of the four similarity measures used. A different behavior is shown for the cosine measure for which there is no improvement by biasing the preferences in the direction of the best solutions found for the dataset. Notice also that the clustering quality values are not as high as those found for the training set (see Figure 2. We could expect that these values could be improved if optimization of preferences is conducted for the complete dataset but this is against a blind validation approach.

5.7 Analysis of the clusters: test set

We further investigate the differences in the AP clusterings found using the similarity measures. Figure 8 shows the number of clusters versus the classification accuracy of the clusterings found by AP on the complete dataset. This figure helps to understand the reason in the poor clustering quality of the cosine measure. Clusters found by AP using these measures have a low accuracy and the number of clusters is smaller than the optimal one. Notice that, since there are 10 different tumor sample classes, the minimum number of clusters required for an optimal classification must have 10 or more clusters. When AP is applied using the cosine measure, no more than 9 clusters are found. Using the correlation distances, high accuracy values with relatively small number of clusters are found. The Euclidean distance produces some clusters with high accuracy but the number of clusters can also be high. Notice that accuracy values can reach the 85%. This classification rate is not as high as those obtained by the application of supervised classification methods combined with feature selection [20]. However, our evolutionary-enhanced AP approach is faster, it is not so sensitive to redundant features, and more important, it can also be improved by the application of feature selection techniques.

6. CONCLUSIONS

In this paper we have presented an evolutionary approach for improving the clustering quality of the affinity propagation algorithm. We have applied the proposal to the clustering of tumor samples from gene expression data. The main contributions of the paper are the following:

- The penalized exemplar-based classification measure has been introduced. This measure combines the accuracy, computed from previous information about the point classes, with a penalty on the size of the clusters.
- A method for automaticly finding a good set of preference values for the AP algorithm have been proposed. We have shown that evolved preference values improve the clustering quality.
- We have shown that preference values learned on a set of data can be reused to improve the clustering of previously unseen data. This way, information is transferred between problem domains.

- Our results show that high classification accuracy can be achieved from gene expression data in a very short time without the need of a previous filtering step.

Another advantage of using AP clustering is that the identification of (exemplar) tumor samples which are good to represent the characteristic features of a particular tumor tissue may contribute to better characterizations of these tissues. Since tumor samples from the same class may be grouped in different clusters, it could also be possible to identify tumor subtypes and alternative sets of putative causal genes for a common tumor class. Classification rules for the identification of those genes that significantly contribute to tumor classification can be extracted from the analysis of the clusters.

7. ACKNOWLEDGMENTS

This work has been partially supported by the TIN2010-20900-C04-04, Consolider Ingenio 2010 - CSD2007-00018 projects (Spanish Ministry of Science and Innovation) and the CajalBlueBrain project.

8. REFERENCES

[1] P. Agarwal and C. Procopiuc. Exact and approximation algorithms for clustering. *Algorithmica*, 33(2):201–226, 2002.

[2] F. Ambrogi, E. Raimondi, D. Soria, P. Boracchi, and E. Biganzoli. Cancer profiles by affinity propagation. In *Proceedings of the 2008 Seventh International Conference on Machine Learning and Applications*, pages 650–655, Washington, 2008. IEEE Computer Society.

[3] S. Depil, C. Roche, P. Dussart, and L. Prin. Expression of a human endogenous retrovirus, HERV-K, in the blood cells of leukemia patients. *Leukemia*, 16(2):254–259, 2002.

[4] R. Etxeberria and P. Larrañaga. Global optimization using Bayesian networks. In A. Ochoa, M. R. Soto, and R. Santana, editors, *Proceedings of the Second Symposium on Artificial Intelligence (CIMAF-99)*, pages 151–173. Editorial Academia. Havana, Cuba, 1999.

[5] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315:972–976, 2007.

[6] J. Hartigan. *Clustering Algorithms*. John Wiley & Sons, Inc. New York, NY, USA, 1975.

[7] J. Hartigan and M. Wong. Algorithm AS 136: A K-means clustering algorithm. *Applied Statistics*, 28(1):100–108, 1979.

[8] P. Larrañaga and J. A. Lozano, editors. *Estimation of Distribution Algorithms. A New Tool for Evolutionary Computation*. Kluwer Academic Publishers, Boston/Dordrecht/London, 2002.

[9] M. Leone, Sumedha, and M. Weigt. Clustering by soft-constraint affinity propagation: Applications to gene-expression data. *Bioinformatics*, 23(20):2708–2715, 2007.

[10] H. Mühlenbein and G. Paaß. From recombination of genes to the estimation of distributions I. Binary parameters. In H.-M. Voigt, W. Ebeling, I. Rechenberg, and H.-P. Schwefel, editors, *Parallel Problem Solving from Nature - PPSN IV*, volume 1141 of *Lectures Notes in Computer Science*, pages 178–187, Berlin, 1996. Springer.

[11] K. Murphy. The BayesNet toolbox for Matlab. *Computer Science and Statistics: Proceedings of Interface*, 33, 2001.

[12] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, California, 1988.

[13] O. Pickeral, J. Li, I. Barrow, M. Boguski, W. Makałowski, and J. Zhang. Classical oncogenes and tumor suppressor genes: a comparative genomics perspective. *Neoplasia*, 2(3):280, 2000.

[14] J. Pittman, E. Huang, H. Dressman, C. Horng, S. Cheng, M. Tsou, C. Chen, A. Bild, E. Iversen, A. Huang, et al. Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes. *Proceedings of the National Academy of Sciences of the United States of America*, 101(22):8431, 2004.

[15] S. Pomeroy, P. Tamayo, M. Gaasenbeek, L. Sturla, M. Angelo, M. McLaughlin, J. Kim, L. Goumnerova, P. Black, C. Lau, et al. Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, 415(6870):436–442, 2002.

[16] R. Santana, C. Bielza, P. Larrañaga, J. A. Lozano, C. Echegoyen, A. Mendiburu, R. Armañanzas, and S. Shakya. MATEDA: Estimation of distribution algorithms in MATLAB. *Journal of Statistical Software*, 35(7):1–30, 2010.

[17] R. Santana, P. Larrañaga, and J. A. Lozano. Protein folding in simplified models with estimation of distribution algorithms. *IEEE Transactions on Evolutionary Computation*, 12(4):418–438, 2008.

[18] R. Santana, P. Larrañaga, and J. A. Lozano. Learning factorizations in estimation of distribution algorithms using affinity propagation. *Evolutionary Computation*, 18(4):515–546, 2010.

[19] R. Santana, A. Ochoa, and M. R. Soto. The mixture of trees factorized distribution algorithm. In *Proceedings of the Genetic and Evolutionary Computation Conference GECCO-2001*, pages 543–550, San Francisco, CA, 2001. Morgan Kaufmann Publishers.

[20] A. Su, J. Welsh, L. Sapinoso, S. Kern, P. Dimitrov, H. Lapp, P. Schultz, S. Powell, C. Moskaluk, H. Frierson, et al. Molecular classification of human carcinomas by use of gene expression signatures. *Cancer Research*, 61(20):7388, 2001.

[21] K. Suh and S. Yuspa. Intracellular chloride channels: critical mediators of cell viability and potential targets for cancer therapy. *Current pharmaceutical design*, 11(21):2753–2764, 2005.

[22] H. Suzuki, X. Zhou, J. Yin, J. Lei, H. Jiang, Y. Suzuki, T. Chan, G. Hannon, W. Mergner, J. Abraham, et al. Intragenic mutations of CDKN2B and CDKN2A in primary human esophageal cancers. *Human molecular genetics*, 4(10):1883, 1995.