Bayesian Classifiers with Consensus Gene Selection: A Case Study in the Systemic Lupus Erythematosus

Rubén Armañanzas[1], Borja Calvo[1], Iñaki Inza[1], Pedro Larrañaga[1], Irantzu Bernales[2], Asier Fullaondo[2], and Ana M. Zubiaga[2]

[1] ISG - Department of Computer Science and Artificial Intelligence, University of the Basque Country, P.O. Box 649 - 20080 San Sebastián, Spain {ruben, borxa, inza, ccplamup}@si.ehu.es

[2] Department of Genetics, Physical Anthropology and Animal Physiology, University of the Basque Country, P.O. Box 644 - 48080 Bilbao, Spain {ggpbepui, ggpfuela, ggpzuela}@lg.ehu.es

1 Introduction

Within the wide field of classification on the Machine Learning discipline, Bayesian classifiers are very well established paradigms. They allow the user to work with probabilistic processes, as well as, with graphical representations of the relationships among the variables of a problem.

Bayesian classifiers assign the corresponding predicted class of a certain pattern as the one that has the highest a posteriori probability. This a posteriori probability is computed by means of the Bayes theorem in conjunction with assumptions about the density of the patterns conditioned to the class.

In this work three of these classification paradigms are applied to a DNA microarray database of control, systemic lupus erythematosus and antiphospholipid syndrome samples. The number of genes from which the models are induced is considerably reduced by means of a novel consensus filter gene selection technique.

Combining a nonparametric bootstrap resampling technique and the k dependence Bayesian classifier paradigm, we propose a new method to obtain gene interaction networks of high reliability. These gene networks can be seen as a tool to study the relationships among the genes of the domain. In fact, some of the previous knowledge about both pathologies is confirmed by the new approach.

2 Bayesian Classifiers

A supervised classifier is a function that assigns labels to observations,

$$\gamma : (x_1, \ldots, x_n) \rightarrow \{1, 2, \ldots, m\},$$

where $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathcal{R}^n$ conforms the observation and $\{1, 2, \ldots, m\}$ are the range of possible values for the class variable. The main assumption is the existence of an unknown underlying probability joint distribution $p(x_1, \ldots, x_n, c)$ where the observations come from:

$$p(x_1, \ldots, x_n, c) = p(c|x_1, \ldots, x_n)p(x_1, \ldots, x_n) = p(x_1, \ldots, x_n|c)p(c).$$

In practice, this joint probability distribution $p(x_1, \ldots, x_n, c)$ is estimated from a random sample, $\left\{(\boldsymbol{x}^{(1)}, c^{(1)}), \ldots, (\boldsymbol{x}^{(N)}, c^{(N)})\right\}$.

The *naïve Bayes* (NB) classifier [6] is based on two assumptions over the predictive variables and the class to predict: the class variable C can only take one of its m possible values $c_1, \ldots, c_m$; and, if this class value is known, the knowledge of some predictive variables is independent from the knowledge of the rest ones. Therefore, the search for the most probable class value, c^*, once all the variables' values are known, can be reduced to look for

$$c^* = \arg\max_c p(c) \prod_{i=1}^{n} p(x_i|c).$$

The conditional independence assumption of the naïve Bayes paradigm can be a very restrictive condition. So as to overcome this limitation, there are classification paradigms that allow conditional dependencies among the variables. One of them is the *tree augmented network* (TAN) [3], in which a tree-like classification modelization and the Bayesian classification paradigm comes together; first, a tree structure among the predictive variables is built, and then, the class node is related to all the variables.

The metric to configure edges between variables is based on the mutual information conditioned to the class variable,

$$I(X, Y|C) = \sum_{i=1}^{t} \sum_{j=1}^{w} \sum_{r=1}^{m} p(x_i, y_j, c_r) \log \frac{p(x_i, y_j|c_r)}{p(x_i|c_r)p(y_j|c_r)},$$

where X and Y are two discrete predictive variables and C is the class label. The complete learning algorithm is discussed in [3] and makes use of the Kruskall algorithm to build the maximum weight spanning tree.

In order to go trough the wide spectrum from the naïve Bayes to a complete Bayesian network, Sahami (1996) [7] presents an algorithm called *k dependence Bayesian classifier* (kDB). The algorithm has its basis on a naïve Bayes structure that allows each predictive variable to have a maximum number of parent variables. The algorithm extends the TAN algorithm allowing a variable to have a number of parents, excluding the class variable C, bounded by k. In Fig. 1 graphical examples of the three paradigms are gathered.

3 Consensus Gene Selection

Bayesian classifiers deal only with discrete data. This restraint makes it necessary to translate the microarray data from continuous to discrete value-domains. This translation can make the original data lose precision, even

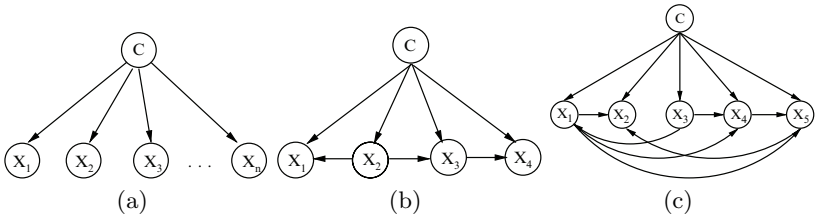

Fig. 1. Graphical structures of a naïve Bayes (**a**), tree augmented naïve Bayes (**b**) and k dependence Bayesian (**c**) classification models

degrading its original quality. Thus, if a discretization process has biased the original data, this bias affects all posterior knowledge discovery processes. Therefore, the search for a robust solution makes us rely on several rather than on a single discretization method.

Let O be the original microarray dataset with continuous features and $S_1, \ldots, S_D$ the results of D different discretizations of the O set. Using a filter subset selection method, N different feature selections are performed on the basis of the $S_1, \ldots, S_D$ discrete datasets, producing the following subsets of genes: $G_1, \ldots, G_D$. The final consensus gene subset Γ is the intersection of all of them, that is $\Gamma = \bigcap_{i=1}^{D} G_i$, with $|\Gamma| = m \leqslant \min_{i=1,\ldots,D} |G_i|$. The complete formulation of this consensus approach can be reviewed in [1].

4 Knowledge Discovery by Means of Bayesian Classifiers

For the final process of knowledge inference, the application of a suited technique that contributes certain level of reliability is crucial. For this purpose, we propose the application of a technique known as *bootstrap*, firstly presented by Efron (1979).

The bootstrap procedure allows us to compute a confidence level for each feature under study on a probabilistic graphical model. These confidence levels are calculated after repetitive runs of the induction algorithm, but, instead of inducing the models in basis of the original dataset, for each run, the original dataset is substituted by N randomly sampled instances with replacement from the original ones. The knowledge discovery process is centered in the detection of the same edges along the different induced graphical models, because this high confidence edges are expected to have a direct biological interpretation. The nonparametric bootstrap algorithm approach implemented for the present study can be found in [4].

5 Results

Both systemic lupus erythematosus (SLE) and antiphospholipid syndrome (APS) are autoimmune diseases with unknown origin. SLE is mainly an

inflammatory disease with clear autoimmune features, and it can affect multiple organs and body systems. Related to the genetic basis of the disease, more than 100 genes are now thought to be involved in SLE genetic susceptibility. APS, also known as "sticky blood" syndrome, is another immunological disease characterized by the repeated appearance of thrombosis, a high number of miscarriages in the second and third gestation quarters, and thrombopenia or hemolytic anemia.

There is no clear diagnosis methodology for SLE and APS: different criteria have to be evaluated in order to assess its presence. Therefore, the study of genes that present different expression profiles among SLE, APS and control subjects is medically and biologically of great interest.

5.1 Data Preprocess

The biochip model used is the Affymetrix® *HGU133A*. The marking and hybridization processes are performed using peripheral blood obtained from 12 different Caucasian women: two with primary APS, four with SLE, and six healthy people, used as controls. Four different criteria are measured to evaluate the biochips reliability: the presence of spike control BioB, the $3'/5'$ relation of the GAPDH housekeeping control, the percentage of present probes in the array and the *dChip*[1] array outlier percentage. From the original 12 biochips, one of them does not reach a sufficient quality level in three out of the four criteria, consequently, it is removed from the dataset.

Filtering the data by the Affymetrix® *detection* algorithm, the amount of valid probes decreases from 22,067 to 8,808; these probes form our starting set of predictive variables. There are a total of 40 comparisons between the samples and the reference microarray ratios, divided in three phenotypes or classes: ten correspond to the control arrays among themselves, another ten between the five control and the two APS patient arrays, and the last 20 correspond to the control and SLE arrays.

5.2 Gene Selection Step

Three different discretization policies are used in the consensus selection: equal width, equal frequency and entropy discretization [2]. Correlation-based feature selection [5] is used as the feature selection method, and the overall process returns eight variables. These genes are considered *statistical prototypes* of gene families showing different behavior profiles over the original data. The members of each family are computed by the classical mutual information metric, obtaining a total set of 150 relevant variables.

[1] DNA-chip analyzer from Harvard University available in `http://biosun1.harvard.edu/complab/dchip/`

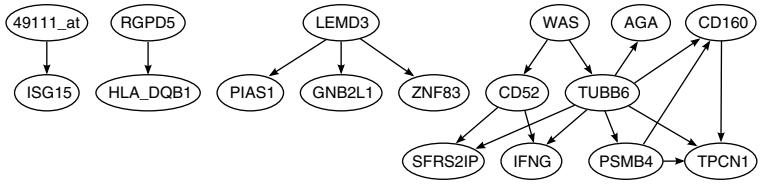

Fig. 2. Dependent structures for a 90% confidence level found by the bootstrap approach over the kDB (k=4) classification models

5.3 Classification Step

By means of the Elvira[2] platform for Bayesian networks, we induce five different models: naïve Bayes, TAN, and kDB with k values of 2, 3 and 4. Each of these models is validated using a leaving one out cross validation. The induced models by naïve Bayes, TAN, or kDB for its three k values, achieve a 100.0% classification accuracy. Due to the low number of instances in the problem, these good results in classification may come from an overfitting effect of the classifiers to the data.

5.4 Knowledge Discovery Step

Starting from the 150 relevant variables identified in the previous process and in basis of the entropy discretization dataset, we perform 1,000 loops in the bootstrap procedure. Thus, 1,000 of random samplings are performed, and 2,000 kDB models are induced (three and four are taken as values for the k parameter).

A total of three different edges are configured always (100% confidence), for both k values. When decreasing the confidence to 90%, the k=4 models configure 18 edges while the k=3 models configure only 13 of them. Notice that the edge between the class and the nodes in the graph is not taken into account. These edges allow us to construct networks of high reliability with respect to their graphical dependencies. Fig. 2 shows the structures found for a k value of four and a 90% of confidence. A deep discussion about the identified genes is collected in [1].

References

1. R. Armañanzas. Solving bioinformatics problems by means of Bayesian classifiers and feature selection. Technical Report EHU-KZAA-IK-2/06, University of the Basque Country, 2006.
2. U. M. Fayyad and K. B. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*, pages 1022–1027. Morgan Kaufmann, 1993.

[2]Elvira system available in `http://leo.ugr.es/elvira/`

3. N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29(2):131–164, 1997.
4. N. Friedman, M. Goldsmidt, and A. Wyner. Data analysis with Bayesian networks: A bootstrap approach. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, pages 196–205, 1999.
5. M. A. Hall and L. A. Smith. Feature subset selection: A correlation based filter approach. In N. Kasabov et al., editor, *Proceedings of the Fourth International Conference on Neural Information Processing and Intelligent Information Systems*, pages 855–858, Dunedin, 1997.
6. M. Minsky. Steps toward artificial intelligence. *Transactions on Institute of Radio Engineers*, 49:8–30, 1961.
7. M. Sahami. Learning limited dependence Bayesian classifiers. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pages 335–338, 1996.