A Cross Entropy Algorithm for Classification with $\delta-$Patterns

Gabriela Alexe[1], Gyan Bhanot[1,2,3†] and Adriana Climescu-Haulica[4,5]

[1] *Computational Biology Center, IBM T.J. Watson Research Center, Yorktown Hts. NY 10598 USA*
Email:[galexe,gyan]@us.ibm.com
[2] *The Simons Center for Systems Biology, Institute for Advanced Study, Princeton, NJ 08540, USA*
[3] *Department of Biomedical Engineering, Boston University, Boston MA 02215, USA*
[4] *Laboratoire Modélisation et Calcul, IMAG 38041 Grenoble, France*
[5] *Laboratoire Biologie Informtique Mathématiques CEA 17 rue des Martyrs 38054 Grenoble, France*
Email: adriana.climescu@cea.fr

A classification strategy based on δ-patterns is developed via a combinatorial optimization problem related with the maximal clique generation problem on a graph. The proposed solution uses the cross entropy method and has the advantage to be particularly suitable for large datasets. This study is tailored for the particularities of the genomic data.

1 Introduction

A central problem in several areas as artificial intelligence, machine learning, and data mining is the extraction of "patterns" (or rules) from data and their aggregation into a classification model able of distinguishing observations in different classes. In general, in a finite dataset consisting of positive and negative observations represented as real valued n vectors, a positive (negative) pattern is an interval in $\mathbf{R}^n$ with the property that it contains sufficiently many positive (negative) observations, and sufficiently few negative (positive) ones.

There are several classes of patterns which have been studied in the last few years especially in connection with the analysis of genomic or proteomic biomedical data sets (e.g., the class of $\delta-$patterns implemented in the biomedical software packages SPLASH (3) and Genes@Work (4), the classes of spanned and prime patterns defined and studied in the context of the Logical Analysis of Data (LAD) methodology (5).

The aim of this paper is to present a cross-entropy heuristic approach for pattern extraction. Compared to the consensus type algorithms (7) and the SPLASH algorithm (3), the cross-entropy heuristic has the same order of complexity while providing better results for large datasets.

2 δ - Patterns for Classification

Let $D = (A, \Sigma)$ be a two class dataset described by a collection of numerical attributes A and a collection of positive and negative samples $\Sigma = \Sigma^+ \cup \Sigma^-$. According to (3), for any $\delta > 0$ a subset of attributes $a_1, ..., a_n$ and a subset of positive samples $s_1, ..., s_m$ defines a $\delta-$ positive pattern if for each attribute a_i the range of its variation across the samples $s_1, ..., s_m$ is less than δ. Negative $\delta-$patterns are defined in a similar way. The subset of samples defining a $\delta-$pattern d is called the support of d. A $\delta-$pattern d is called maximal if it cannot be extended to a $\delta-$pattern containing more samples nor to a $\delta-$pattern defined by more attributes. Intuitively, if δ is small then all the (numerical) attributes defining a $\delta-$pattern have essentially the same value across the samples in the support set. On the other hand, when δ is large, in particular when $\delta \geq \min_{a \in A}(\max a - \min a)$ where $\max a$ and $\min a$ are the maximum, respectively the minimum value of the attribute a taken across all the samples in the data, a maximal $\delta-$pattern is a maximal spanned pattern as defined in the context of Logical Analysis of Data (LAD) (5). In a more general framework (3), (6) a $\delta-$ positive (negative) pattern might be allowed to contain a small fraction of samples (say up to 20%) from the opposite class.

One of the major applications of $\delta-$patterns is in developing classification models. Current pattern-based classification methodologies include Genes@Work (4) (for δ-patterns) and LAD (5) (for spanned

† Address after Sept. 1, 2006: Department of Biomedical Engineering and BioMaPS, Rutgers University, Piscataway, NJ 08854 USA and Cancer Institute of New Jersey, New Brunswick, NJ 08903 USA

patterns). The classification schemes based on patterns have been shown in numerous studies –especially in the biomedical area – ((3), (4), (5), (6) and (7)) to provide classification models with comparable or superior accuracy than those provided by standard machine learning tools.

To construct highly accurate pattern-based classification models it is necessary to produce a large (ideally exhaustive) collection of maximal patterns and then to select out of this collection pools of reliable positive and negative patterns (i.e. positive and negative patterns with high quality parameters (5), (6)). We have shown in (6) that the collection of all spanned patterns in a dataset can be produced by applying a consensus-type total polynomial algorithm. Clearly a similar consensus scheme could be applied for the generation of all $\delta-$patterns, by simply requiring that the range of variation of the attributes defining a pattern to be bounded by δ. The consensus scheme used for the generation of all patterns can be shown to produce in fact all the maximal bicliques in an associated bipartite graph (8). In (2) we have shown that the problem of enumerating all maximal bicliques in a bipartite graph is equivalent to the problem of enumerating all maximal cliques in a graph. Thus, the problem of enumeration of all δ-patterns can be reduced to the maximal clique problem.

3 A cross-entropy algorithm for maximal clique generation

In this study we propose to adapt a cross-entropy scheme to generate maximal cliques satisfying a threshold selection criteria. This scheme will be applied as a heuristic for the enumeration of all positive and negative reliable patterns. Let (G, V) be an arbitrary undirected graph where $V = \{1, ..., n\}$ is the vertex set of G and $E \subseteq V \times V$ is the edge set. A clique $\{x_1, ..., x_\tau\}$ is represented by a vector $\mathbf{x} = (x_1, ..., x_\tau)$ where $x_i \in \{1, .., n\}$ are pairwise distinct vertices of G.

Let Υ denotes the space of all such vectors. Define $S : \Upsilon \rightarrow \mathbf{N}$ the clique cardinality function, by $S((x_1, ..., x_\tau)) = \tau$. The maximum clique problem is formalized as identifying

$$\tau^* = \max_x S(x) \tag{1}$$

A stochastic learning strategy for this problem is developed in the context of the cross entropy method (1). The cross entropy method starts from associating to the deterministic problem (1) the rare event estimation problem

$$l(\tau) = \mathbf{P}(S(\mathbf{X} \geq \tau) = \mathbf{E}I_{\{S(\mathbf{X})\geq\tau\}} \tag{2}$$

where $\mathbf{X}$ is a random vector and $\mathbf{E}I_{\{S(\mathbf{X})\geq\tau\}}$ represents the expectation of the indicator function associated with the maximum clique problem. Estimating the rare event probability and the associated optimal probability distribution translates back into solving the optimization problem (1). Typically, for the estimation problem the *importance sampling* technique is used. The system is simulated under a different probability distribution, whose parameters are called reference parameters, in order to make the occurrence of the rare event more likely. Usually the reference parameters are difficult to obtain and the variance minimization algorithm are time consuming. The cross entropy method brings a fast and simple procedure for estimating the optimal reference parameters in the importance sampling. This is done by making adaptive changes, according to the Kullback-Leibler cross-entropy, to the probability density function of the random vectors generated at each step. Thus, the cross entropy algorithm has as main part the generation of a sequence of probability density functions which are steered in the direction of the theoretically optimal density. It is shown (1) that, if the corresponding maximizer is unique, say $\mathbf{x}^*$, the cross entropy optimal density is the atomic density at $\mathbf{x}^*$.

In the context of the classification with δ-patterns we are interested on generating all the maximal cliques with their clique value bigger than a fixed threshold T depending on the maximum clique value. In practice, when applied to the enumeration of the pools of reliable positive (negative) patterns, the size of T could be an adjustable parameter used to calibrate the accuracy of the classifier.

In order to introduce the maximal clique trajectory generation and the extraction of the maximum clique value, we define an additional node to the graph G, with label 0, adjacent to all other nodes. Let Π denote the transition probability matrix associated with the extended graph G.

Algorithm 1: Maximal Clique Trajectory Generation Using Node Transitions

1. Define $\Pi^{(1)} = \Pi$ and $X_1 = 0$. Set $k = 1$.

2. Obtain the matrix $\Pi^{(k+1)}$ from $\Pi^{(k)}$ by first setting the X_k-th column of $\Pi^{(k)}$ to 0 and then normalizing the rows to sum up 1. Generate X_{k+1} from the distribution formed by the X_k-th row of $\Pi^{(k+1)}$.

3. If $k = n$ then stop; otherwise set $k = k + 1$ and reiterate from Step 2.

Algorithm 2: Cross Entropy Algorithm for Maximum Clique Value

1. Set $t = 1$. Put $\widehat{\Pi}_t = \Pi$.

2. Generate a sample of vectors $\mathbf{x}_1, ..., \mathbf{x}_N$ of maximal cliques using Algorithm 1 with $\Pi = \widehat{\Pi}_t$.

3. Compute the clique values $S_{(k)} = S(\mathbf{x}_k)$ for each k and order them in increasing order $S_{(1)} \leq ... \leq S_{(N)}$. Let $\widehat{\tau}_t$ be the sample $(1-q)$-quantile of cliques values

$$\widehat{\tau}_t = S_{(\lceil (1-q)N \rceil)}$$

4. Use the same sample to calculate

$$\widehat{p}_{t,ij} = \frac{\sum_{k=1}^{N} I_{\{S(\mathbf{x}_k) \geq \widehat{\tau}_t\}} I_{\{\mathbf{x}_k \in \Upsilon_{ij}\}}}{\sum_{k=1}^{N} I_{\{S(\mathbf{x}_k) \geq \widehat{\tau}_t\}}}$$

where Υ_{ij} is the set of cliques for which the transition from node i to node j is made.

5. If for some $d \leq t$

$$\widehat{\tau}_t = \widehat{\tau}_{t-1} = ... = \widehat{\tau}_{t-d}$$

then output the maximum clique value $\tau = \widehat{\tau}_t$ else set $t = t + 1$ and reiterate from Step 2.

The solution of this algorithm is given by the optimal degenerate transition matrix $\Pi^* = (p^*_{ij})$ defined by $p^*_{ij} = \delta_{x^*}(\Upsilon_{ij})$.

Algorithm 3: Trajectories Generation for Maximal Cliques with Threshold Selection Criteria

1. Define $\Pi^{(1)} = \Pi$ and $X_1 = 0$. Set $k = 1$.

2. Obtain the matrix $\Pi^{(k+1)}$from $\Pi^{(k)}$ by first setting the X_k th column of $\Pi^{(k)}$ to 0 and then normalizing the rows to sum up 1.

3. Generate X_{k+1} from the distribution formed by the X_k- th row of $\Pi^{(k+1)}$.

4. If $k = n$ and the corresponding clique vector has value bigger than the threshold T or equal to T then add it to the list; otherwise if $k < n$ set $k = k + 1$ and reiterate from Step 2. If $k = n$ and the corresponding clique vector has value smaller than T then reiterate from Step 1 with $\Pi = \Pi^{(n)}$.

The three algorithms have polynomial complexity. This is a general feature of the cross entropy algorithms, which are about as fast as the deterministic algorithms and are accurate on finding global optima especially for large size applications.

4 Conclusions

We propose a new strategy for $\delta-$pattern classification, with an increased potential for discrimination on large datasets. Although the $\delta-$pattern classification was oriented in our previous work for solving biomedical problems (e.g., risk evaluation among cardiac patients (8), polymer design for artificial bones (7), genomic-based diagnosis or prognosis of lymphoma (8)), our pattern-based classification approach can be used for any large size data application.

References

[1] R.Rubinstein, D. Kroese. The Cross-Entropy Method, Springer (2004).

[2] G. Alexe, S. Alexe, Y. Crama, S. Foldes, P.L. Hammer, B. Simeone. Consensus algorithms for the generation of all maximal bicliques. Discrete Applied Mathematics 145(1): 11-21 (2004).

[3] A. Califano, G. Stolovitzky, Y. Tu. Analysis of gene expression microarrays for phenotype classification. Proc Int Conf Intell Syst Mol Biol. 8:75-85 (2000).

[4] J. Lepre, J.J. Rice, Y. Tu, G. Stolovitzky. Genes@Work: an efficient algorithm for pattern discovery and multivariate feature selection in gene expression data. Bioinformatics 20(7):1033-44 (2004).

[5] E. Boros, P.L. Hammer, T. Ibaraki, A. Kogan, E. Mayoraz, I. Muchnik. An implementation of Logical Analysis of Data. IEEE Trans.Knowledge and Data Eng. 12:292-306 (2000).

[6] G. Alexe, P.L. Hammer. Spanned patterns for the logical analysis of data. Discrete Applied Mathematics 154(7): 1039-1049 (2006).

[7] S.D. Abramson, G. Alexe, P.L.Hammer, J. Kohn. A computational approach to predicting cell growth on polymeric biomaterials. J Biomed Mater Res A. 73(1):116-24 (2005).

[8] G. Alexe. Combinatorial Knowledge Discovery Efficient Algorithms and Applications. Scientific Seminar, Siemens Corporation (2004).