R U T C O R
R ESEARCH
R E P O R T

PATTERN-BASED FEATURE SELECTION IN GENOMICS AND PROTEOMICS

Gabriela Alexe[a] Sorin Alexe[b] Peter L. Hammer[c]
Bela Vizvari[d]

RRR-7-2003 MARCH 2003

RUTCOR
Rutgers Center for
Operations Research
Rutgers University
640 Bartholomew Road
Piscataway, New Jersey
08854-8003
Telephone: 732-445-3804
Telefax: 732-445-5472
Email: rrr@rutcor.rutgers.edu
http://rutcor.rutgers.edu/~rrr

[a] RUTCOR, Rutgers University, Piscataway, NJ 08854, email: alexe@rutcor.rutgers.edu
[b] RUTCOR, Rutgers University, Piscataway, NJ 08854, email: salexe@rutcor.rutgers.edu
[c] RUTCOR, Rutgers University, Piscataway, NJ 08854, email: hammer@rutcor.rutgers.edu
[d] Eotvos Lorand University, Budapest, Hungary, e-mail: vizvari@cs.elte.hu

RUTCOR RESEARCH REPORT
RRR-7-2003 MARCH 2003

PATTERN-BASED FEATURE SELECTION IN GENOMICS AND PROTEOMICS

Gabriela Alexe Sorin Alexe Peter L. Hammer Bela Vizvari

Abstract. A major difficulty in data analysis is due to the size of the datasets, which contain frequently large numbers of irrelevant or redundant variables. In particular, in some of the most rapidly developing areas of bioinformatics, e.g., genomics and proteomics, the expressions of the intensity levels of tens of thousands of genes or proteins are reported for each observation, in spite of the fact that very small subsets of these features are sufficient for distinguishing positive observations from negative ones. In this study, we describe a two-step procedure for feature selection. In a first "filtering" stage, a relatively small subset of relevant features is identified on the basis of several combinatorial, statistical, and information-theoretical criteria. In the second stage, the importance of variables selected in the first step is evaluated based on the frequency of their participation in the set of all maximal patterns (defined as in the *Logical Analysis of Data*, and generated using an efficient, total-polynomial time algorithm), and low impact variables are eliminated. This step is applied iteratively, until arriving to a Pareto-optimal "support set", which balances the conflicting criteria of simplicity and accuracy.
Keywords: feature selection, genomics, proteomics, logical analysis of data, *LAD*, patterns

Acknowledgements: The partial support provided by ONR grant N00014-92-J-1375 and DIMACS is gratefully acknowledged.

1 Introduction

The presence of large numbers of irrelevant or redundant features in a dataset can create enormous computational difficulties. This is the case, in particular, in genomics and proteomics, two of the most rapidly developing areas of bioinformatics, where, the expressions of the intensity levels of thousands or tens of thousands of genes or proteins are included in the datasets, in spite of the fact that very small subsets of these features are amply sufficient for perfectly distinguishing positive observations from negative ones. For instance, out of more than 15,000 peptides appearing in a recently studied ovarian cancer dataset (http://clinicalproteomics.steem.com), we have identified [1] a subset of only 7 peptides which can distinguish with 100% accuracy the ovarian cancer cases from the controls. Similarly, in a breast cancer dataset (http://www.rii.com/publications/ 2002/vantveer.htm), reporting the expressions of the intensity levels of more than 25,000 genes for each case, we have identified [2] a subset of only 16 genes which can distinguish with an accuracy of 93% the poor prognostic cases from the good ones.

The presence of too many attributes in a dataset, as well as the presence of confounding features, such as experimental noise, may lead – in the absence of adequate ways of selecting small, relevant subsets – to the failure of any attempt to extract knowledge from the data.

The importance of feature selection is well known in statistics, and was recognized for a long time as an important component of research in machine learning, knowledge discovery, data mining, and other related areas. Comprehensive overviews of many existing methods from the 1970's to the present appear in the 1997 survey of M. Dash and H. Liu [3], and the 1998 monographs [4] and [5] of H. Liu and H. Motoda. Some of the classical techniques of statistics (e.g., principal component analysis) are highly pertinent to the area of feature selection. Among the methods specifically designed for this purpose, we mention the exhaustive search heuristics (e.g., sequential forward/backward selection, branch and bound selection [3]), and those using artificial neural networks (e.g., [6], [7]), support vector machines (e.g., [8]), genetic algorithms (e.g., [9]).

One of the complicating factors in the extraction of a relevant subset of features, is the fact that there is a marked difference between the relevance of individual features, and that of subsets of features. As a typical example, we mention the fact that the highly relevant subset of 7 features identified in the ovarian cancer study [1], includes only 4 peptides whose levels are strongly correlated with the presence of ovarian cancer.

The purpose of this paper is to outline a method for identifying subsets of features, which, taken collectively, can distinguish with high accuracy the positive observations from the negative ones. The major concept used in this procedure is that of *collective biomarkers*, or *patterns*, as defined in the *Logical Analysis of Data* (*LAD*).

LAD ([10]) is a combinatorics, optimization, and logic-based methodology for the analysis of data. Specific features of the *LAD* approach include the exhaustive examination of the entire set of features (without the exclusion of those having low statistical correlations with the outcome, or those having low values), focusing on the classification power of the combinations of features (without confining attention only to the individual ones), and on the possibility of extracting novel information on the role of features and of combinations of features through the

analysis of these exhaustive lists. It is important to notice that the potentials offered by *LAD* for feature selection have been fully confirmed by the empirical studies.

Initially, *LAD* was proposed for the analysis of datasets with binary features, i.e., features taking only the value 0 and 1, and included a feature selection component based on the solution of a set-covering problem – a well studied problem in Operations Research. Later, a method was developed ([11]) for "binarizing" data with numerical values. The binarization process replaces a feature with numerical values by several binary ones, and the set-covering approach for feature selection can be applied in this case, too. The distinctive feature of the proposed approach to feature selection is that it is not limited to the analysis of the roles of individual features, but takes also into account the "collective effect" of sets of features in distinguishing the positive and negative observations of a dataset.

While the set-covering based feature selection method turned out to be very useful in the analysis of datasets coming from a variety of areas (economics, seismography, oil exploration, cardiology, design of artificial tissues, etc), its application for genomic and proteomic data ran into two problems.

First, the number of genes or proteins in a typical genomic or proteomic dataset is in the thousands or tens of thousands. Since the number of binary variables to replace a numerical one (corresponding to a gene or a protein) can be hundreds of times larger, it can expected that these problems can lead to set-covering models with millions of binary variables. Further, the constraints in the associated set-covering problems, correspond to the pairs formed by a positive and a negative observation in a dataset. Therefore, the set-covering problem can be expected to have tens of thousands of constraints and millions of variables. Although excellent heuristics are available for the solution of set-covering problems, the size of the problems appearing in the area under investigation is beyond their capacity.

Second, in contrast to the usual set-covering problem, there is a new type of requirements appearing in the binarized forms of data analysis problems coming from genomics or proteomics. In the usual set-covering model, the main problem is to minimize the number of variables (features) which take the value 1 (i.e., are selected) in an optimal solution. In genomics and proteomics, the number of variables to be minimized is not that of the total number of binary variables associated to each gene or protein, but the number of selected genes or proteins. The combinatorial optimization model corresponding to this problem is substantial more complicated than a simple set-covering problem, and no heuristic method for its solution is available in the literature.

The main objective of this paper is to bypass the binarization step, as well as the formulation and the solution of a set-covering problem, in extracting from genomic or proteomic dataset a small, knowledge-preserving set of features, which allows the accurate distinction of positive and negative observations.

2 Selection of Feature Pool

In the first phase of feature selection, we shall apply several criteria for evaluating each one of the features in the dataset, with the aim of creating a pool of selected features to be retained for further analysis. For each of the criteria applied we retain for further consideration the *k* top ranked features; in the examples presented in this paper we have selected *k* to be 50. The pool consists simply in the set of those features which are ranked among the top *k* by at least of the criteria applied. Naturally, many of the selected features are ranked highly by several criteria.

We shall describe below five of the criteria applied in the current implementation of the proposed method.

2.1. Separation Measure

For a feature *x*, let us denote by x^+ , respectively x^-, the average value of *x* taken over the set of all positive, respectively negative, points in the dataset, let x^* be the average of x^+ and x^-, and let us assume that $x^+ \geq x^-$ (the other case being treated in a symmetric way). Let the number of positive (respectively, negative) observations with $x \geq x^*$ (respectively, $x < x^*$) be n_x^+ (respectively, n_x^-). It is conceivable to replace the numerical variable *x* by an associated binary variable ξ, which is equal to 1 for $x \geq x^*$, and 0 otherwise. Therefore, the number of those pairs consisting of a positive and a negative observation which can be distinguished by the corresponding value of ξ, will exceed $n_x^+ n_x^-$. The higher this *separation measure* $\sigma_x = n_x^+ n_x^-$, the more positive/negative pairs of observations can be separated based on the values taken by *x*.

2.2. Envelope Eccentricity

Let us denote by l_x^+ and u_x^+ (respectively by l_x^- and u_x^-) the minimum and the maximum of the values of feature *x* among the positive (respectively, negative) observations in the dataset. The overlap of the intervals $[l_x^+, u_x^+]$ and $[l_x^-, u_x^-]$ in an indication of the separation of positive and classes based on the values of feature *x* alone. We shall define the *overlap index* ω_x of *x* to be the ratio $\omega_x = \dfrac{\min(u_x^-, u_x^+) - \max(l_x^-, l_x^+)}{\max(u_x^-, u_x^+) - \min(l_x^-, l_x^+)}$. Clearly, ω_x takes values between 0 and 1, and the smaller its value the more relevant feature *x* is in the separation of positive and negative observations.

2.3. System Entropy

Information theory provides a valuable measure expressing the nonseparability of positive and negative observations in the dataset based on the values of one of the features. Let *m* be the number of observations in the dataset, and let *t* be an integer approximation of $\sqrt{m}$. Let us order the observations in the dataset according to the values of the variable *x*, and let us partition them into *t*+1 approximately equally sized subsets of consecutive observations $S_0, S_1, ..., S_t$. Assuming that p_j and q_j are the numbers of positive,

respectively negative, observations in S_j, the frequency of positive observations in S_j will be $\frac{p_j}{p_j+q_j}$, and the entropy of $\{S_0, S_1, ..., S_t\}$ will be:

$\varepsilon_x = -\frac{1}{t+1}\sum_{i=0}^{t}\left[\left(\frac{p_i}{p_i+q_i}\right)\ln\left(\frac{p_i}{p_i+q_i}\right)+\left(\frac{q_i}{p_i+q_i}\right)\ln\left(\frac{q_i}{p_i+q_i}\right)\right]$. This number will be called the *entropy* of feature *x*, and it takes values between 0 and ln2. Clearly, variables which "separate" well positive and negative observations will have a low entropy.

2.4. Pearson Correlation

The *Pearson correlation* coefficient π_x between the vector of values of a feature *x* and the outcome vector *o* (whose components are 1's for positive observations, and 0's for negative ones) is indicative of the importance of that feature: the higher the absolute value of this coefficient, the more important the feature is.

2.5. Signal-to-Noise Correlation

The common *signal-to-noise correlation* τ_x provides an additional measure of separation, being defined in the following way: let x^+ and x^- be defined as in 2.1., and let σ^+ and σ^- be the corresponding standard deviations of the values of *x* on the positive and the negative datasets, respectively. The ratio $\tau_x = \frac{x^+ - x^-}{\sigma^+ + \sigma^-}$ is called the *signal-to-noise correlation* ([12]) of feature *x* with the outcome.

Example On the basis of each of the five measures described above, we have selected the best 50 features appearing in the genomic DLBCL dataset long-term follow-up ([13], http://llmpp.nih.gov/lymphoma) and the proteomic dataset Ovarian_Cancer ([14], http://clinicalproteomics.steem.com). Beside these five pools of features, we have introduced a six-th pool consisting of the union of the first five. In the case of the DLDCL dataset, the combined pool consisted of 190 (out of a total of 7,129) features, and in the case of the Ovarian_Cancer dataset it consisted of 157 (out of a total of 15,154) features.

The combined pools distinguish the positive observations from the negative ones more accurately than any of the five pools selected on the basis of a single criterion. In order to validate this statement, we have randomly partitioned each of the two datasets into a "training set" containing about 50% of the positive and 50% of the negative observations, and a "test set" containing the remaining observations. Using the training sets, we have developed on each of the six pools a *LAD* "classifier" (to be defined and discussed in the next section) and applied it to the observations in the test set. The accuracies of the six classifications are presented in Table 1, and show clearly the superior performance of the classifications which used the pool obtained by combining the five criteria.

	Separation Measure (σ_x)	Envelope Eccentricity (ω_x)	System Entropy (ε_x)	Pearson Correlation (π_x)	Simple Correlation (τ_x)	Combined Criteria
DLBCL	64.18%	81.49%	70.67%	70.67%	60.58%	89.18%
Ovarian_Cancer	96.67%	100.00%	91.73%	97.16%	98.77%	99.38%

Table 1. Accuracies based on feature pools selected according to various criteria

3 Elements of *LAD*

The definition of a "support set" of features, i.e., a set which allows the construction of an accurate model capable of distinguishing positive and negative observations, depends on the particular type of model to be used. Since the proposed feature selection method uses concepts and algorithms related to *LAD*, we shall aim our discussion to the creation of support sets which can be used successfully in *LAD*, but shall illustrate on the two datasets DLBCL and Ovarian_Cancer that the support sets obtained in this way are also advantageous for various other data analysis methods. In order to clarify our terminology we shall first present some basic elements of *LAD*.

3.1. Basic Concepts of LAD

One of the first steps of *LAD* is to associate some "cutpoints" and corresponding "binary variables" to each of the original variables. If x is one of the original variables taking numerical values, and c is a cutpoint of it, then the associated binary variable ξ_c is defined to be equal to 1 when $x \geq c$ and 0 otherwise, i.e., it distinguishes "high" and "low" values of x. Several cutpoints (and binary variables) can be associated to the same original variable. The minimization of the number of binary variables to be used in this process was shown in [11] to be NP hard. A *conjunction,* say $C = \xi_c \xi_{c'}^{'} \bar{\xi}_{c''}^{''}$, is a new binary variable, which takes the value 1 if $\xi_c = \xi_{c'}^{'} = 1$ and $\xi_{c''}^{''} = 0$ (i.e., if the corresponding numerical variables x, x', x'' satisfy the conditions $x \geq c, x' \geq c', x'' < c''$), and 0 otherwise.

A distinctive concept of *LAD* is that of *collective biomarkers*, or *patterns*. A conjunction is said to be a *positive pattern* if it takes the value 0 on every negative observation, and if it takes the value 1 on some positive observations. In *LAD*, attention is usually restricted only to patterns which "cover" (i.e., take the value 1 on) a sufficiently large proportion of the positive observations (whose "prevalence" exceeds some prescribed value). *Negative patterns* and their prevalence is defined in a symmetric way. The two most important parameters of patterns are their *prevalences*, and their *degrees*, i.e., the number of variables included in the expression of the pattern (the degree of the conjunction C in the example above is 3).

Let us consider for illustration the Ovarian_Cancer proteomic dataset, in which the features are labeled by their M/Z values. The conjunction requiring the simultaneous fulfillment of the two conditions "*The intensity level of the peptide having the M/Z value 4004.826 is less than 29.899447*" and "*The intensity level of the peptide having the M/Z value 435.46452 is more than 24.0603315* " is fulfilled by 152 of the 162 positive cases in the dataset, and by none of the negative ones. Clearly, the above conjunction describes a positive pattern.

A *reduct* of a conjunction is a conjunction obtained by eliminating one of the variables appearing in the original conjunction; for example $C' = \xi_c \overline{\xi}_{c''}$ is one of the three possible reducts of C. A positive pattern is called *prime* if none of its reducts is a pattern, i.e., if each of its reducts covers some negative observation. *Negative prime* patterns are defined in a symmetric way.

The (*d*,*p*)-*positive pandect* corresponding to a dataset, is the family of all positive prime patterns having degrees of at most *d*, and prevalences of at least *p*. *Negative pandects* are defined symmetrically. An efficient algorithm ([15]) for finding the (*d*,*p*)-positive pandect will be described in Section 3.2.

As an illustration, we present in Table 2 the positive and negative (3, 0.5)-pandects of the Ovarian_Cancer dataset, built on a set of 9 peptides.

Patterns	*M/Z Values of Peptides in Support Set and Pattern Defining Inequalities*									*Prevalence (%)*	
	245.8296	*261.88643*	*336.6502*	*418.8773*	*435.46452*	*437.0239*	*465.97198*	*681.38131*	*4004.826*	*Positive*	*Negative*
P1	≤47.5455				>24.0603315					97	0
P2					>24.0603315				≤29.899447	94	0
P3		≤47.8783305			>24.0603315					94	0
P4	≤51.887381				>25.008547					91	0
P5		≤48.2855705			>24.800402					91	0
P6	≤42.4233285					>16.233283				88	0
P7	≤39.866264								≤29.899447	87	0
P8	≤47.5455	≤46.1940675								86	0
P9	≤51.887381	≤45.871292								83	0
P10	≤40.4012065	≤46.5299145								81	0
P11		≤46.5299145				>16.233283				81	0
P12					>24.0603315		≤14.7199595			78	0
P13	≤47.5455		≤23.3584715							75	0
P14	≤51.887381		≤23.2770235							73	0
P15						>16.233283	≤14.7199595			72	0
P16	≤39.866264						≤14.7199595			72	0
P17			≤23.3584715		>25.920563					59	0
P18	≤39.866264			>30.799397						59	0
P19					>24.0603315			≤30.4082455		52	0
P20	≤40.4012065							≤30.4082455		51	0
N1	>46.306687				≤25.920563					0	91
N2	>46.306687	>45.871292								0	90
N3	>42.953243				≤24.800402					0	90
N4				>30.7853195	≤24.3509305					0	89
N5					≤24.559075			>29.319256		0	89
N6	>47.5455								>22.026144	0	87
N7		>47.8783305			≤25.920563					0	86
N8					≤24.0603315				>23.4580195	0	77
N9		>47.8783305	>22.100553							0	77
N10	>51.65812									0	76
N11		>48.2855705							>23.4580195	0	76
N12					≤25.008547				>27.976873	0	75
N13			>23.2770235		≤25.008547					0	75
N14					≤25.008547		>14.7199595			0	75
N15		>46.1940675					>14.7199595			0	75
N16	>46.306687								>27.976873	0	74
N17	>46.306687		>22.4826545							0	74
N18			>22.100553		≤24.0603315					0	73
N19		>47.8783305							>27.976873	0	68
N20		>46.5299145							>29.899447	0	65
N21	>46.306687					≤16.233283				0	52

Table 2. (3, 0.5) - pandect for the Ovarian_Cancer dataset

3.2. Pandect Generation

We shall briefly describe below the generation method ([15]) of the positive (*d*,*p*)-pandect; the negative pandect is generated in a similar way. For the sake of simplicity, we shall outline the method only for the special case of datasets with two features X and Y, which we shall assume here to take the finite set of values $\{0,1,\ldots,h\}$, and $\{0,1,\ldots,q\}$, respectively.

The algorithm starts by building the matrix $M=(m_{ij})_{i=0,1,...,h;\,j=1,2,...,q}$, having as entries the number m_{ij} of positive points with $X=i$ and $Y=j$. Let us define now for $i \le k$ and $j \le l$, the *interval* $[(i,j),(k,l)]$ as the set of the points (X,Y) having $i \le X \le k$ and $j \le Y \le l$. The acceptable positive patterns correspond to those intervals which include at least p positive, and no negative observations. For every pair of integers $k \in [0,h]$ and $l \in [0,q]$, the procedure constructs recursively matrices $R^{(k,l)}$, the entries (i,j) of which represent the number of positive observations in the interval $[(i,j),(k,l)]$. The recursive computation of $R^{(k,l)}$ is based on the enumeration of the pairs (k,l), $k \in [0,h]$ and $l \in [0,q]$, by using a generalized form of Gray codes ([16]). This procedure allows the enumeration to be performed in such a way that: (a) every pair (k, l) between (0,0) and (h, q) is generated exactly once, and (b) two consecutive pairs differ in exactly one component, and by exactly one unit. As soon as the matrix $R^{(k,l)}$ is computed, the recursion proceeds to the next pair (e.g., $(k, l-1)$) in the Gray sequence. The matrix R produced at the end of this procedure coincides with the matrix $R^{(k,l-1)}$. The algorithm evaluates the (positive and negative) prevalence of all possible intervals in the dataset, generating subsequently all the prime patterns in the (d,p)- pandect.

It was shown in [15] that in the general case of a dataset with n features, the number of operations (additions) necessary to generate all the prime patterns is bounded by $2^n N$, where N is the number of all pairwise distinct intervals in the dataset. Clearly, if the number n of attributes is small, the algorithm is almost linear in the size of the output. For the case of prime patterns of limited degree, the algorithm runs in polynomial time, since the number of intervals N is polynomial in the input size.

As an illustration, we mention that the numbers of positive, respectively negative, patterns of degree at most 2, each covering at least 30% of the positive, respectively negative, observations (i.e., the positive and negative (2, 0.3)-pandects) for the Ovarian_ Cancer dataset build on the pool of 157 features identified in the example in Section 2, are respectively of 3970 and 2046. The computing time for their generation on a 1GHz Pentium III processor is of 52 seconds. Similarly, the numbers of patterns in the positive and negative (2, 0.3)-pandects of the DLBCL dataset, build on the pool of 190 features, are respectively of 2506 and 3963, and the corresponding computing time is of 44 seconds.

3.3. Accuracy of Pandect-Based Classification

The pandect defines ([17]) a *classifier*, i.e., a function which predicts the positive or negative nature of a new (i.e., yet unseen) observation. The classifier will predict an observation to be positive (negative) if it satisfies the defining conditions of some positive (negative) patterns in the pandect, and does not satisfy any of the negative (positive) ones. If the observation satisfies both positive and negative patterns in the pandect then the decision is based ([18], [19]) on the sign of the difference between the proportions of positive and negative patterns satisfied by it. More exactly, if the pandect consists of the positive patterns $P_1, P_2, ..., P_h$ and negative patterns $N_1, N_2, ..., N_q$ then $C(x) = sign(0)\left(\frac{1}{h}\left|\{i \mid x \in P_i\}\right| - \frac{1}{q}\left|\{j \mid x \in N_j\}\right|\right)$, where $x \in P_i$ or $x \in N_j$ mean that

P_i, respectively N_j, cover *x*. Finally, the classifier leaves "unclassified" the (extremely rare) observations *x*, for which $C(x) = sign(0)$.

The *accuracy* of a classifier built on a "training set" of observations is the average of the proportion of correctly classified positive observations and the proportion of correctly classified negative observations in a "test set" of observations. Usually, the test set is disjoint from the training set, although sometimes the two sets are allowed to overlap, or even coincide.

4 Knowledge-Preserving Pool Reduction

4.1. Pandect-Based Feature Evaluation

The number of those patterns in the positive or the negative (*d*,*p*)-pandect which involve a particular feature *x* can give a good indication of its importance. In order to illustrate this point we have ranked on this basis the features appearing in the pools detected in Section 2 for the DCBCL and Ovarian_Cancer datasets. For each of the problems, we have constructed models using the features with highest and with lowest frequencies of participation in the patterns appearing in the positive and negative (2, 0.3)-pandects. For each of the selected feature sets we have developed a model using as "training set" a stratified sample consisting of 50% of the observations, and evaluated the accuracy of its classifications on the "test set" consisting of the remaining 50% of the observations. The accuracy of the resulting classifications is shown in Table 3.

		Number of selected features			
Dataset	Features Ranked	30	40	50	60
DLBCL	Highest	89.18%	85.34%	85.34%	89.18%
	Lowest	63.94%	63.82%	65.87%	64.66%
Ovarian_Cancer	Highest	98.89%	98.89%	98.89%	100.00%
	Lowest	50.00%	50.00%	86.69%	93.58%

Table 3. Accuracies of pools including highest/lowest ranked features

The differences between the accuracies obtained by using the highest and the lowest ranked α (=30, 40, 50, 60) features are significantly different, and strongly support the underlying hypothesis concerning the importance of features with high frequency of participation in patterns. Although it is possible that using different pandect-defining values of *d* and *p*, the accuracies reported in the above table may be slightly different, and in the special case of increased values of *p* perhaps somewhat higher, the size of the gap between the accuracy of classifications based on high and low ranked features outweights the possible influence of using different *d* and *p* values.

The pandect-based evaluation of the role of features is the essential ingredient of the proposed heuristics.

4.2. Iterative Pool Contraction

The pandect-based ranking of features by their importance serves as a guiding principle for the iterative sequence of steps of the proposed algorithm. Starting from the pool defined in Section 2, we define at each step of the procedure a new pool consisting of approximately half of the top ranked features of the current pool, according to the ranking based on the pandect.

The accuracy of the classifier built on the new pool is evaluated on a test set.

- If it is found of acceptable quality, i.e., if the accuracy is approximately equal to that of the parent pool, the new pool replaces the previous pool and the process continues.
- If the accuracy falls substantially below that of the parent pool, the new pool is not rejected automatically, but an attempt is made to find a better performing pandect on the same pool. This can be achieved by varying (usually increasing) the parameters *d* and *p* which define the pandect.
    - If a pandect whose classifying power is of acceptable accuracy is found, then the new pool is accepted and the process continues.
    - If no better performing pandect is identified, then the new pool is "improved" by enlarging it to include approximately 75% of the top ranked features appearing in the parent pool, (and if this pool has still to be augmented, it will include approximately 87.5%, and then approximately 93.75% of the features in the parent pool).
- Finally, if this process does not produce a pool which defines a pandect of acceptable quality, then the process stops with the parent pool as final output.

In some cases it is required that in order to assure robustness, the number of features in the final pool should not fall below a prescribed threshold. In these cases, the process is stopped before the number of feature becomes too small.

We illustrate the method for the DLBCL and Ovarian_Cancer datasets, specifying the number of features selected at each step of the process, along with the corresponding accuracies:

Step	*1*	*2*	*3*	*4*	*5*	*6*
# of Features	190	100	50	25	15	10
Accuracy	89.18%	89.18%	89.18%	89.90%	93.30%	86.78%

Table 4. Pool contraction for the DLBCL dataset

Step	*1*	*2*	*3*	*4*	*5*	*6*	*7*
# of Features	157	80	40	20	10	7	6
Accuracy	99.38%	98.27%	98.89%	99.38%	100.00%	100.00%	99.38%

Table 5. Pool contraction for the Ovarian_Cancer dataset

The pool selected at the end of the process involves 15 features in the case of DLBCL dataset, and 7 features in the Ovarian_Cancer dataset.

5 Feature Selection for Various Data Analysis Methods

It is clear that the selection of a small knowledge-conserving set of features is heavily influenced by the particular data analysis method for which the reduction is carried out. Although the procedure described in Section 4 aims at finding a good (i.e., small + knowledge conserving) set of features for *LAD*, the results of it are advantageous for other methods too.

For illustration, we present below the accuracies (measured on test sets) of classifiers using various data analysis methods, applied to the BLDCL and to the Ovarian_Cancer datasets. In each case, we present the accuracies on the original pool of 190, respectively 157, features as well as on the final pool of 15, respectively 7, features. The 5 methods examined were *LAD*, logistic regression (LR), classification trees (CART), classification neural networks (CNN), and Fisher linear discriminant analysis (LDA). The software used for LR, CART, CNN is *Insightful Miner* 2.0, 2002 (Insightful Corporation), while the one used for *LAD* and LDA was developed by the authors.

Feature Pool	*# of Features*	*LAD*	*Logistic Regression*	*CART*	*CNN*	*LDA*
Original	190	89.18%	65.50%	58.60%	89.70%	45.91%
Final	15	93.30%	65.50%	69.00%	69.00%	62.00%

Table 6. DLBCL: Accuracies of 5 classifiers on original and final pools

Feature Pool	*# of Features*	*LAD*	*Logistic Regression*	*CART*	*CNN*	*LDA*
Original	157	99.38%	99.20%	81.70%	64.30%	N/A
Final	7	100.00%	35.70%	95.20%	100.00%	100.00%

Table 7. Ovarian_Cancer: Accuracies of 5 classifiers on original and final pools

6 Conclusions

The pattern-based feature selection method proposed in this paper was seen to lead to the identification of small feature pools capable of accurately distinguishing positive observations from negative ones in genomic and proteomic datasets. It should be remarked that besides the usual correlation-based feature selection criterion, the proposed method includes several other criteria, emphasizing the role of individually and/or collectively significant features. While the method was developed specifically to serve model-building in the *Logical Analysis of Data*, it is shown that its results provide useful bases on which various other methods can be successfully applied.

7 References

[1] G. Alexe, S. Alexe, P. L. Hammer, L. Liotta, E. Petricoin, M. Reiss. Logical Analysis of the Proteomic Ovarian Cancer Dataset. *RUTCOR Technical Report*, *RTR* 2-2002 (http://rutcor.rutgers.edu/ ~rrr/rtr/2-2002.pdf).

[2] Alexe, G., Alexe, S., Axelrod, D., Boros, E., Hammer P. L., Reiss, M., Combinatorial analysis of breast cancer data from image cytometry and gene expression microarrays, *RUTCOR Technical Report*, *RTR* 3-2002.

[3] Dash, M., Liu, H. Feature selection for classification. *Intelligent Data Analysis*, 1, (3) 1997, 131-156.

[4] Liu, H., Motoda, H. *Feature Extraction, Construction and Selection: A Data Mining Perspective*, Kluwer Academic Publishers, 1998.

[5] Liu, H., Motoda, H. *Feature Selection for Knowledge Discovery and Data Mining*, Kluwer Academic Publishers, 1998.

[6] Leray, P., Gallinari, P. Feature selection with neural networks. *Behaviormetrika*, 26 (1), (1999).

[7] Setiono, R., Liu, H. Neural network feature selector. *IEEE Transactions on Neural Networks*, 8, (3) (1997), 654-662.

[8] Bradley, P.S., Mangasarian, O. L. Feature selection via concave minimization and support vector machines. In J. Shavlik, editor, *Proceedings of the Fifteenth International Conference on Machine Learning*, pp. 82-90. Morgan Kaufmann, San Francisco, CA, (1998).

[9] Chtioui, Y., Bertrand, D., Barba, D. Feature selection by a genetic algorithm. Application to seed discrimination by artificial vision, *Journal of the Science of Food and Agriculture*, 76 (1), (1998), 77-86.

[10] Crama, Y., Hammer, P.L., Ibaraki, T. Cause-Effect Relationships and Partially Defined Boolean Functions. *Annals of Operations Research* 16 (1988), 299-326.

[11] Boros E., Hammer P. L., Ibaraki T., Kogan A. Logical Analysis of Numerical Data. *Mathematical Programming* 79 (1997), 163-190.

[12] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller H., Loh M. L., Downing J. R., Caligiuri M. A., Bloomfield, C. D., Lander E. S. Molecular Classification of Cancer; Class Discovery and Class Prediction by Gene Expression Monitoring. *Science* 286 (5439) (1999), 531-537.

[13] Shipp M. A., Ross, K. N., Tamayo, P., Weng A. P., Kutok, J. L., Aguiar, R. C. T., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G. S., Ray, T. S., Koval, M., Last, A. K. W. , Norton, A., Lister, T.A., Mesirov, J., Neuberg, D.S., Lander, E. S., Aster, J.C., and Golub, T.R. Diffuse Large B-Cell Lymphoma Outcome Prediction by Gene Expression Profiling and Supervised Machine Learning. *Nature Medicine*, Volume 8 1(2002), 68-74.

[14] Petricoin, E. F., Ardekani, A. M., Hitt, B. A., Levine, P. J., Fusaro, V. A., Steinberg, S. M., Mills, G. B., Simone, C., Fishman, D. A., Kohn, E. C., Liotta, L. A. Use of proteomic patterns in serum to identify ovarian cancer. *The Lancet* 359 (9306) (2002), 572-577.

[15] Alexe, S., Hammer, P. L. Accelerated Algorithm for Pattern Detection in Logical Analysis of Data. *RUTCOR Research Report*, RRR 59-2002, *Annals of Operations Research* (2003), (in print).

[16] Koda, Y., Ruskey, F. A Gray code for the ideals of a forest poset. *Journal of Algorithms* 15 (1993) 324-340.
[17] Boros, E., Hammer, P.L., Ibaraki, T., Kogan, A., Mayoraz, E., Muchnik, I. An Implementation of Logical Analysis of Data. *IEEE Transactions on Knowledge and Data Engineering*, 12 (2) (2000), 292-306.
[18] Alexe S., Blackstone E., Hammer, P. L., Ishwaran, H., Lauer, M. S., Pothier Snader, C. E. Coronary Risk Prediction by Logical Analysis of Data. *Annals of Operations Research*, 119 (2003), 15-42.
[19] Alexe, S., Hammer, P. L. Pattern-Based Supervised Learning Classification. *Proceedings of Workshop on Discrete Mathematics and Data Mining*, SIAM-Society for Industrial and Applied Mathematics, San Francisco, May 2003.