RUTCOR
RESEARCH
REPORT

METHODS FOR THE ANALYSIS OF LARGE REAL-VALUED MEDICAL DATABASES BY LOGICAL ANALYSIS OF DATA

Zsolt Csizmadia[a] Béla Vizvári[b]

RRR 42-2004, DECEMBER 2004

RUTCOR • Rutgers Center for Operations Research • Rutgers University • P.O. Box 5062 • New Brunswick New Jersey • 08903-5062
Telephone: 908-932-3804
Telefax: 908-932-5472
Email: rrr@rutcor.rutgers.edu

[a]Mathematical Doctoral School, Eötvös Lóránd University of Budapest, Pázmány Péter sétány 1/C, H-1117 Hungary

[b]RUTCOR, Rutgers University and Department of Operations Research, Eötvös Lóránd University of Budapest, H-1088 Budapest, Pázmány Péter sétány 1/C, H-1117 Hungary

Rutcor Research Report
RRR 42-2004, December 2004

Methods for the analysis of large real-valued medical databases by logical analysis of data

Zsolt Csizmadia Béla Vizvári

Abstract. This report shortly summarizes the works –supported by RUTCOR and DIMACS– of the two authors on the field of LAD and is devoted to the analysis of large real valued databases. The common properties of these databases are the large number of attributes, the non-existence of binary or categorical variables. The large size of the problems requires special algorithmic approach because of the phenomena of combinatorial explosion. The main topics of the report are: (i) the analysis of elementary feature selection methods, (ii) a new general feature selection method based on a polynomial, (iii) generation of LAD objects based on a restricted enumeration tree method used for patterns, perfect patterns, i.e. patterns which are alone a complete LAD theory, and discriminant generation.

Acknowledgements: This research was carried out when the authors visited RUTCOR. They were supported under grant no. NSF-IIS-0312953

1 Introduction

This report is devoted to methods developed in a LAD environment for the analysis of large real valued databases. The common properties of these databases are the large number of attributes, the non-existence of binary or categorical variables. The large size of the problems requires special algorithmic approach because of the phenomena of combinatorial explosion. Both the elaborated methods and the numerical results will be discussed in the paper.

Four publicly available databases were used in this study: ovarian cancer (with 15154 attributes, http://clinicalproteomics.steem.com), large ovarian cancer (approximately 350 thousand attributes, http://clinicalproteomics.steem.com), lymphoma (with 7000 attributes, http://www.genome.wi.mit.edu/MPR/lymphoma) and breast cancer (with 25000 attributes, http://www.rii.com/publications/2002/v_antveer.htm).

The structure of the paper is this: section 2 discusses the analysis and evaluation of elementary feature selection methods including the correlation of the elementary methods and the use of them in composite strategies. Section 3 describes a new general feature selection method based on a polynomial. A general methodology is elaborated in section 4 to generate LAD objects for large databases when not all of the cases can be investigated because of the combinatorial explosion. This method is applied to generate patterns, perfect patterns and linear discriminants.

2 Elementary feature selection methods

1. **Pearson Correlation.** The correlation of the classification vector and each attribute is calculated. The Pearson correlation method orders the attributes to the decreasing order of the absolute values of these correlations.

2. **Average separation.** For each attribute the center of gravity of both the positive and negative observations are determined. A cut point is determined as the average of the two centers. The better the attribute is, if this cut point separates the higher number of positive and negative pairs of observations such that the positive (negative) member of the pair is in the direction of the positive (negative) center.

3. **Best separating cut point.** Assume that observations i and j have the values x_i and x_j at attribute x. Suppose that i is a positive and j is a negative observation. Further on let us suppose that there is no other observation k such that its attribute value x_k is between x_i and x_j. Then any value strictly between x_i and x_j is called a *separation cut point*. Let c be such a cut point. Denote by n_c^u, and n_c^l (p_c^u, and p_c^l) the number of negative (positive) observations above, respectively below c. The value representing the quality of the cut point c is $v(c) = max\{n_c^u + p_c^l, n_c^l + p_c^u\}$. The value of the attribute is $max_c\{v(c)\}$. The better the attribute is, the higher its value is.

4. **Envelope Eccentricity.** This method measures the overlapping of the region of the positive and negative observations for each attributes. Let l_x^+ and u_x^+ (l_x^- and u_x^-) be the minimum and the maximum of the values of feature x among the positive (negative) observations in the dataset. Then the ratio

$$E_x = \frac{min(u_x^-, u_x^+) - max(l_x^-, l_x^+)}{max(u_x^-, u_x^+) - min(l_x^-, l_x^+)}$$

characterizes the overlapping. The smaller its value is, the better the attribute is. If the ratio is negative then the attribute perfectly separates the positive and negative observations. If the two intervals coincide then $E_x = 1$ and it is the possible maximal value.

5. **Signal-to-Noise Correlation.** The signal-to-noise correlation of an attribute is the absolute value of the difference of the two centers mentioned at the average separation divided by the sum of the standard deviations of values of this attribute at the positive and negative observations.

3 Filters

For each dataset, all of the attributes were ranked by five different elementary feature selection methods. These methods are: Pearson correlation, separation cut, best separating cut, envelope, signal-to-noise. The final ranks of the attributes were formed from the ranks of the elementary feature selection methods on the following 3 ways:

1. summing up of the 5 ranks,

2. summing up the best 3 ranks regardless that which one is the best 3,

3. summing up the ranks of the following: correlation, best-cut-separation, envelope.

The smaller the sum is, the better the attribute is. The first 2000 attributes of each dataset were investigated. The 2000 attributes were divided into 40 groups such that the fist 50 formed the first group, the second 50 formed the second group and so on. These 3 different orders are called filters.

4 Correlations among the elementary feature selection methods and the filters

An important observation is that the elementary feature selection methods and the filters are highly correlated to each other in spite of the fact that they are defined with very different mathematical formulae. The only exceptions are the envelope eccentricity and the best 3 filter. The correlations for the individual problems are reported in Tables 1-4, their average is in Table 5. Finally, Table 6 shows the characteristic vectors of those correlations which are higher than 0.4. It is worth to note that all of the correlations of the elementary feature selection methods without the envelope eccentricity is above 0.69.

Ovarian	P. C.	B. Cut	E. E.	Sep.	S2Noise	all 5	best 3	3 sel.
P. C.	1	0.875763	0.398728	0.943862	0.989934	0.276745	0.127487	0.284815
B. Cut		1	0.311459	0.869304	0.88629	0.37176	0.183298	0.381385
E. E.			1	0.355642	0.433322	0.214107	0.085536	0.233169
Sep.				1	0.95192	0.34907	0.174797	0.352903
S2Noise					1	0.367968	0.20069	0.372466
all 5						1	0.513876	0.994164
best 3							1	0.454138
3 sel.								1

Table 1. Correlations of the elementary feature selection methods and filters in the case of the ovarian problem.

Huge O.	P. C.	B. Cut	E. E.	Sep.	S2Noise	all 5	best 3	3 sel.
P. C.	1	0.924035	0.359879	0.909475	0.997173	0.495969	0.061366	0.501827
B. Cut		1	0.338944	0.897143	0.933206	0.553481	0.075304	0.557442
E. E.			1	0.327359	0.376673	0.373137	0.036247	0.415307
Sep.				1	0.916434	0.526442	0.070885	0.520319
S2Noise					1	0.536707	0.073596	0.540944
all 5						1	0.204898	0.981104
best 3							1	0.165641
3 sel.								1

Table 2. Correlations of the elementary feature selection methods and filters in the case of the huge ovarian problem.

Lymph.	P. C.	B. Cut	E. E.	Sep.	S2Noise	all 5	best 3	3 sel.
P. C.	1	0.528975	0.328139	0.670268	0.958732	0.436986	0.151315	0.438212
B. Cut		1	0.02991	0.333059	0.473622	0.403164	0.145559	0.432994
E. E.			1	0.161558	0.389441	0.273914	0.055152	0.325334
Sep.				1	0.743729	0.488305	0.183095	0.415073
S2Noise					1	0.54243	0.206078	0.517888
all 5						1	0.221707	0.969984
best 3							1	0.184443
3 sel.								1

Table 3. Correlations of the elementary feature selection methods and filters in the case of the lymphoma problem.

Breast	P. C.	B. Cut	E. E.	Sep.	S2Noise	all 5	best 3	3 sel.
P. C.	1	0.706163	0.068721	0.638913	0.998021	0.604247	0.118482	0.550356
B. Cut		1	-0.07926	0.679859	0.701628	0.548728	0.116649	0.50916
E. E.			1	-0.05986	0.091444	0.188227	0.005783	0.30705
Sep.				1	0.642102	0.551176	0.114955	0.412629
S2Noise					1	0.61918	0.128329	0.564769
all 5						1	0.118851	0.946913
best 3							1	0.086582
3 sel.								1

Table 4. Correlations of the elementary feature selection methods and filters in the case of the breast cancer problem.

Averages	P. C.	B. Cut	E. E.	Sep.	S2Noise	all 5	best 3	3 sel.
P. C.	1	0.758734	0.288867	0.79063	0.985965	0.453487	0.114662	0.443803
B. Cut		1	0.150262	0.694841	0.748687	0.469283	0.130203	0.470245
E. E.			1	0.196174	0.32272	0.262346	0.04568	0.320215
Sep.				1	0.813546	0.478748	0.135933	0.425231
S2Noise					1	0.516571	0.152173	0.499017
all 5						1	0.264833	0.973041
best 3							1	0.222701
3 sel.								1

Table 5. Average correlations.

Averages	P. C.	B. Cut	E. E.	Sep.	S2Noise	all 5	best 3	3 sel.
P. C.	1	1	0	1	1	1	0	1
B. Cut	1	1	0	1	1	1	0	1
E. E.	0	0	1	0	0	0	0	0
Sep.	1	1	0	1	1	1	0	1
S2Noise	1	1	0	1	1	1	0	1
all 5	1	1	0	1	1	1	0	1
best 3	0	0	0	0	0	0	1	0
3 sel.	1	1	0	1	1	1	0	1

Table 6. Average correlations above 0.4.

The underlying reason of the high correlations is the fact that if the values of two elementary feature selection methods are considered as points of a plane, then the obtained cloud of points is around a line. This phenomenon is illustrated on Figure 1, in the case of elementary feature selection methods, Best Cut and Pearson Correlation. The equation of the regression line is $0.154187\ \underline{BestCut} + 0.74389 = \underline{PearsonCorrelation}$.

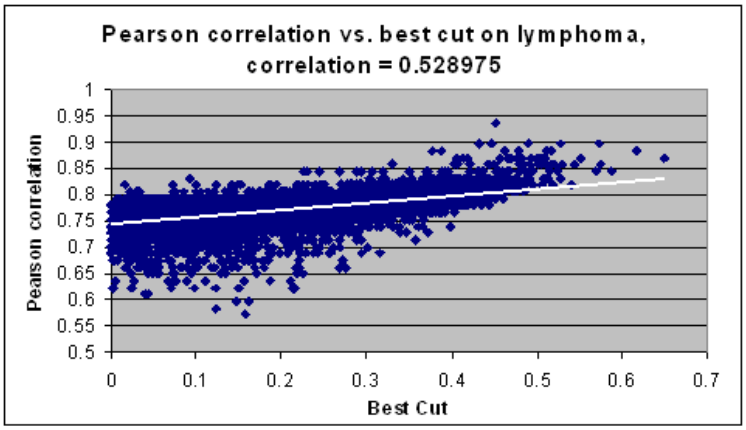

Figure 1. Geometric background of the high correlations.

There is another geometric evidence for the high correlations. Assume that the attributes are indexed to 1 to n where n is the number of attributes. All databases have the property that there are intervals of the indices of the attributes such that the behavior of all feature selection methods is different from the rest of the database. This means that all the feature selection methods are acting to the information encoded into some parts of the database at the same time. In the other parts of the databases the behavior can be described as a noise with a fixed expected value.

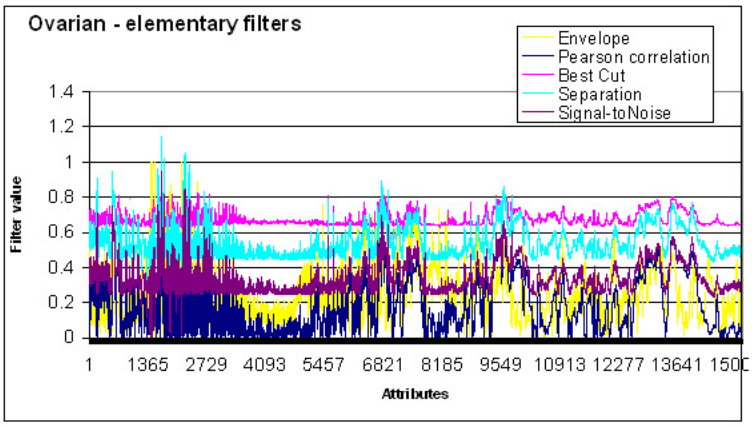

Figure 2. Elementary filters for the ovarian cancer dataset.

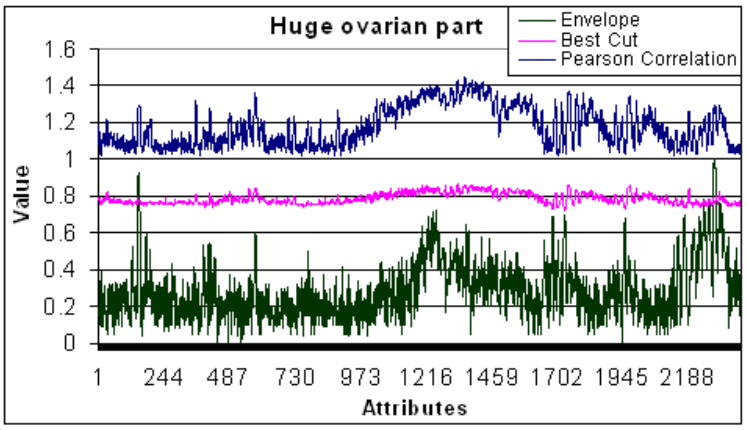

Figure 3. Part of the huge ovarian database. For the sake of better visibility, the values of the Pearson correlations and best cut separation have been increased by 1 and 0.2, respectively.

5 Experiments to measure the effectiveness of elementary feature selection methods

The aim of the computational experiments was to check hypotheses on whether or not the feature selection methods are collecting important information about the classification problem in question. The first hypothesis is directly connected to the logical analysis of data while the second one is independent from the further data processing.

5.1 Measurements based on LAD

Hypothesis 1. If the elementary methods are selecting attributes which contain important information about the classification problem, then the results obtained by LAD, i.e. the errors on the test sets, must have a worsening tendency in position of the above mentioned groups of 50 attributes if LAD uses the attributes of a single group only.

If the hypothesis is true, then the slope of a regression line fitted to the errors obtained on the different groups of the attributes must be positive. Similarly, the group of the first 50 must be one of the best groups. On the other hand, the very best group cannot have a high index.

The results of these "static" groups can be compared with the results of a dynamic group formed on the following greedy way. Initially, a positive integer is determined that how many times must be separated each positive-negative pair of observations. First that attribute is selected, which separates the most pairs. In each further step that attribute is selected which separates the most pairs among those, who are not separated as many times as it is required. This method is called *iterative separation*.

A similar method is the *iterative correlation*. The first attribute is the one with the highest correlation to the classification vector. In any further iteration that attribute is selected which gives the highest correlation to the residuum of the previous linear regression.

The classification method used in the experiments is a simple system where the underlying set covering problems of LAD are solved by the well-known greedy algorithm.

The results are reflecting the average values of 5 folding calculations which were repeated 10 times. Table 1 summarizes the results. The column of the "Position of the 50 obtained by iterative separation/correlation" contains the potential position of the group of 50 attributes selected by the iterative separation/correlation method if it were among the 40 groups.

Database	Filter	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7	Column 8	Column 9	Column 10
Ovarian	all 5	1	100%	1	0.002424247	1	100%	2	110.22%	2	110.31%
	best 3	1	100%	1	0.002618375	2	106.80%	3	115.68%	3	110.34%
	3 selected	1	100%	1	0.002599377	1	100%	3	119.43%	3	117.09%
Huge ovarian	all 5	2	101.69%	5	0.000924639	7	110.25%	5	104.55%	41	156.77%
	best 3	2	102.80%	4	0.00097262	1	100%	28	125.76%	41	159.35%
	3 selected	2	102.34%	4	0.000961649	1	100%	14	117.06%	41	160.01%
Limphoma	all 5	1	100%	1	0.003612179	3	124.34%	2	115.57%	4	126.49%
	best 3	1	100%	1	0.002369578	3	124.19%	2	115.29%	2	122.12%
	3 selected	1	100%	1	0.002369578	3	124.19%	7	139.25%	2	122.12%
Breast	all 5	39	118%	37	-0.00074725	38	116.19%	1	100%	1	100%
	best 3	37	117.06%	23	-0.00096065	30	113.60%	3	101.29%	1	100%
	3 selected	40	117.22%	11	-0.00074338	37	115.10%	1	100%	1	100%

where

Column 1: Position of the first 50
Column 2: The percentage of the result of the first 50 in the result of the best 50
Column 3: Index of the best 50
Column 4: Slope of the regression line
Column 5: Position of the 50 obtained by iterative separation
Column 6: The percentage of the result of iterative separation in the result of the best 50
Column 7: Position of the 50 obtained by iterative correlation on the first best 2000
Column 8: The percentage of the result of iterative correlation in the result of the best 50
Column 9: Iterative correlation on whole dataset
Column 10: The percentage of the result of iterative correlation in the result of the best 50

Table 7. The result of computational experiments checking the hypothesis on the importance and goodness of elementary feature selection methods by LAD.

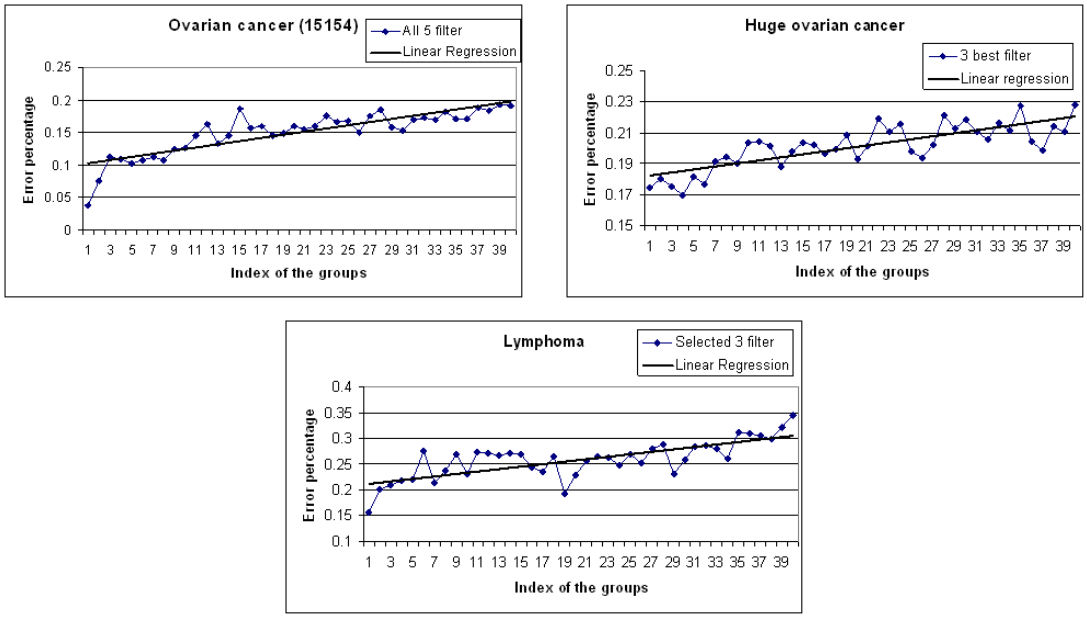

Figure 4. Slope of the regression line for the LAD error percentage.

Conclusions. The hypothesis is justified in the case of the first 3 datasets. The negative slope contradicts the hypothesis in the case of the breast cancer problem. Similarly, here the positions of the first 50 and the iterative separation are against the hypothesis as well. We strongly believe that the explanation is that the breast cancer dataset contains almost no information about the classification problem which can be used by LAD. We do not exclude the possibility that the breast cancer database does not contain usable information even for other classification methods.

5.2 Measurements based on distances

Assume that the observations are the points of the n-dimensional Euclidian space, and all of the points of the space are observations, belonging to either the positive, or the negative class. It is also assumed, that for real life problems the boundary of the two classes are not dense in any n-dimensional subset of the space. Then a classification method my interpreted geometrically as follows. The boundary of the two classes is approximated by a surface depending on the classification method. All the points of the Euclidian space lying on one side of the surface are classified by the method as positive observations and all the points on the other side are classified as negative observations.

Intuitively it is evident that if the distance from the known members of the positive class to the known members of the negative class is higher then it is easier to find the approximate surface. Therefore, the following hypothesis can be formulated.

Hypothesis 2. If the elementary methods are selecting attributes which contain important information about the classification problem, then the average distances must have a shortening tendency on the order of the above mentioned groups of attributes.

If the hypothesis is true, then the slope of a regression line fitted to the average distances obtained on the different groups of the attributes must be negative. Similarly, the group of the first 50 must be one of the best groups. On the other hand, the very best group cannot have a high index.

Database	Filter	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7	Column 8
Ovarian	all 5	1	100.00%	1	-0.02477336	2	98.69%	4	89.39%
	best 3	1	100.00%	1	-0.03028132	2	99.06%	5	90.46%
	3 selected	1	100.00%	1	-0.02357295	2	98.62%	3	89.86%
Huge ovarian	all 5	40	96.10%	3	-0.00264	1	100.00%	4	98.85%
	best 3	1	100.00%	1	-0.00878764	1	100.00%	6	97.46%
	3 selected	40	95.23%	4	-0.00148123	1	100.00%	6	97.33%
Lymphoma	all 5	40	91.33%	15	-0.00333276	1	100.00%	14	98.48%
	best 3	1	100.00%	1	-0.01430385	2	99.37%	7	94.80%
	3 selected	40	97.38%	20	-0.00171189	1	100.00%	10	99.46%
Breast	all 5	7	98.11%	3	-0.0039566	1	100.00%	28	97.38%
	best 3	1	100.00%	1	-0.01110118	6	97.38%	16	95.92%
	3 selected	40	95.51%	3	-0.00219751	1	100.00%	41	96.18%

where

Column 1: Position of the first 50
Column 2: The percentage of the result of the first 50 in the result of the best 50
Column 3: Index of the best 50
Column 4: Slope of the regression line
Column 5: Position of the 50 obtained by iterative separation
Column 6: The percentage of the result of iterative separation in the result of the best 50
Column 7: Position of the 50 obtained by iterative correlation on the first best 2000
Column 8: The percentage of the result of iterative correlation in the result of the best 50

Table 8. The result of computational experiments checking the hypothesis on the importance and goodness of elementary feature selection methods by average distances.

Conclusions. The best 3 filter supports the hypothesis.

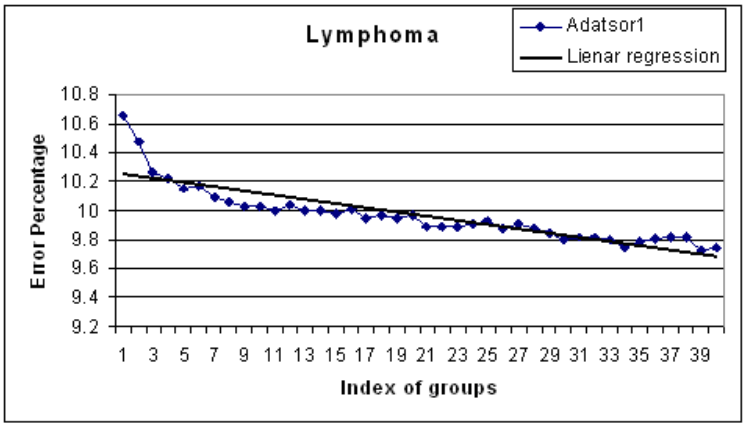

Figure 5. Slope of the regression line for the distances.

6 A general feature selection method

An attribute with a cut point is considered the better if it separates the higher number of positive negative pairs of observations such that the positive observations are on one side of the cut point and negative observations are on the other side. The results of these section are based on this assumption. On the other hand, it will be shown in the next section, that other type of cut points can be useful, too.

E. Boros in [1] raised some general principles of the evaluation of cut points of real valued features. Let c be a cut point. Denote by p_c^u and n_c^u the number of positive and negative observations above c, respectively. Let p and n be the number of positive and negative observations. Denote by x and y the proportion of positive and negative observations above cut point c, i.e. $x_c = \frac{p_c^u}{p}$ and $y_c = \frac{n_c^u}{n}$. A perfect cut point is such that either all of the positive observations are above, and all the negative observations are under it, or vice versa, i.e. either $x_c = 1$ and $y_c = 0$ or $x_c = 0$ and $y_c = 1$. An attribute is considered as good, as good its best cut point.

Let us consider a general cut point evaluation function $f(x, y)$ which evaluates all of the cut points based on the above mentioned proportion. It is supposed that function f is continuously differentiable on the domain $[0, 1] \times [0, 1]$ and its values are between 0 and 1. The higher the value of f, the better the cut point is. It is also supposed that at the points $(1, 0)$ and $(0, 1)$ it has maximum, and similarly it has a minimum at the points $(0, 0)$ and $(1, 1)$. It is claimed that the second order conditions must prove the maximality and minimality, respectively, of the function in the above mentioned points. Then the following equations must be satisfied.

1. $f(1,0) = f(0,1) = 1$
2. $f(0,0) = f(1,1) = 0$
3. $\partial_x f(0,0) = \partial_x f(1,1) = \partial_x f(1,0) = \partial_x f(0,1) = 0$
4. $\partial_y f(0,0) = \partial_y f(1,1) = \partial_y f(1,0) = \partial_y f(0,1) = 0$
5. $\partial_{xx} f(1,0) < 0, \partial_{xx} f(0,1) < 0,$
   $\partial_{xx} f(1,0)\partial_{yy} f(1,0) - (\partial_{xy} f(1,0))^2 > 0,\ \partial_{xx} f(0,1)\partial_{yy} f(0,1) - (\partial_{xy} f(0,1))^2 > 0$

6. $\partial_{xx}f(0,0) > 0, \partial_{xx}f(1,1) > 0,$
$\partial_{xx}f(0,0)\partial_{yy}f(0,0) - (\partial_{xy}f(0,0))^2 > 0$, $\partial_{xx}f(1,1)\partial_{yy}f(1,1) - (\partial_{xy}f(1,1))^2 > 0$

For any function with the above properties, the corresponding attribute evaluation function is $g(a) = max_c f(x_c, y_c)$.

In what follows f is determined in polynomial form. There is no polynomial up to degree 3 with the above mentioned properties. There are infinite many polynomials of degree 4 such that their coefficients satisfy the 12 equations contained in requirements 1-4. This polynomials are determined by 3 quantities, say a, b, c. Their form is this:

$$(3+a)x^2+(2+b)xy+(3+c)y^2-2(1+a)x^3-(6+b)x^2y-(6+b)xy^2-2(1+c)y^3+ax^4+4x^3y+bx^2y^2+4xy^3+cy^4$$

Because of the symmetricity of the polynomial the double constraints contained in requirements 5 and 6, respectively, are the same. These are in the case of requirement 5:

$$a \geq -3, (6+2a)(6+2c) \geq (2+b)^2$$

and in the case of requirements 6:

$$a \leq 3, (-6+2a)(-6+2c) \geq (2-b)^2.$$

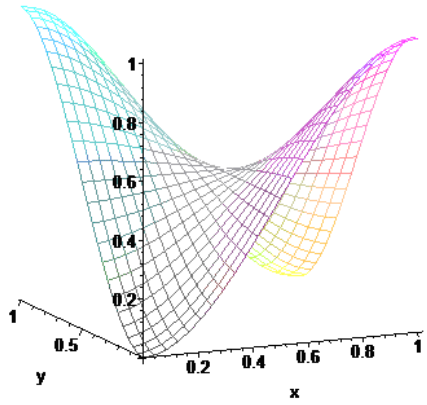

Figure 6. The polynomial in case of a=b=c=0.

f1	Pearson Correlation
f2	Best-Cut Separation
f3	Envelope Eccentricity
f4	Separation
f5	Signal-to-Noise
f6	Poly filter (a=0, b=0, c=0)
f7	Poly filter (a=2.9, b=1, c=0.5)
f8	Poly filter (a=2.9, b=3, c=0.5)
f9	Poly filter (a=1, b=-2, c=1)
f10	Poly filter (a=-2.9, b=-1, c=-0.5)
f11	Poly filter (a=-2.9, b=-3, c=-0.5)
f12	Poly filter (a=0.5, b=1, c=2.9)
f13	Poly filter (a=0.5, b=3, c=2.9)
f14	Poly filter (a=-0.5, b=-1, c=-2.9)
f15	Poly filter (a=-0.5, b=-3, c=-2.9)
f16	Poly filter (a=0, b=3.9, c=0)
f17	Poly filter (a=0, b=-3.9, c=0)

Table 10. Abbreviation used in table 11 and 12

Each elementary feature selection method maps the attributes into the real numbers. Thus each elementary feature selection method determines a real vector having the dimension of the number of attributes. With the exception of the envelope eccentricity this number is the better if its value is the higher. Therefore in the following calculations the value E_x of envelope eccentricity was substituted by $1 - E_x$. The higher the correlation of the vectors of two methods is, the more similar the methods are. Table 11 contains the average correlations on the four databases of the five elementary feature selection methods and 12 different polynomial methods. Table 12 contains 1 where the correlation is above 0.7 and zero otherwise.

Average	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	f11	f12	f13	f14	f15	f16	f17
f1	1.00	0.76	0.29	0.79	0.99	0.78	0.66	0.52	0.78	0.81	0.81	0.66	0.52	0.81	0.82	0.67	0.81
f2	0.76	1.00	0.15	0.69	0.75	0.73	0.61	0.47	0.73	0.75	0.76	0.61	0.47	0.75	0.75	0.62	0.75
f3	0.29	0.15	1.00	0.20	0.32	0.23	0.17	0.11	0.23	0.24	0.25	0.17	0.11	0.24	0.25	0.17	0.25
f4	0.79	0.69	0.20	1.00	0.81	0.89	0.79	0.65	0.89	0.90	0.89	0.79	0.65	0.89	0.88	0.80	0.90
f5	0.99	0.75	0.32	0.81	1.00	0.81	0.69	0.54	0.81	0.83	0.84	0.68	0.54	0.84	0.85	0.70	0.84
f6	0.78	0.73	0.23	0.89	0.81	1.00	0.93	0.80	1.00	0.96	0.94	0.93	0.80	0.96	0.94	0.95	0.98
f7	0.66	0.61	0.17	0.79	0.69	0.93	1.00	0.96	0.93	0.83	0.80	0.93	0.88	0.86	0.82	0.98	0.86
f8	0.52	0.47	0.11	0.65	0.54	0.80	0.96	1.00	0.80	0.67	0.64	0.88	0.89	0.70	0.66	0.93	0.70
f9	0.78	0.73	0.23	0.89	0.81	1.00	0.93	0.80	1.00	0.96	0.94	0.92	0.80	0.96	0.94	0.94	0.98
f10	0.81	0.75	0.24	0.90	0.83	0.96	0.83	0.67	0.96	1.00	0.99	0.86	0.70	0.94	0.93	0.86	0.98
f11	0.81	0.76	0.25	0.89	0.84	0.94	0.80	0.64	0.94	0.99	1.00	0.82	0.66	0.93	0.92	0.82	0.97
f12	0.66	0.61	0.17	0.79	0.68	0.93	0.93	0.88	0.92	0.86	0.82	1.00	0.96	0.82	0.79	0.98	0.86
f13	0.52	0.47	0.11	0.65	0.54	0.80	0.88	0.89	0.80	0.70	0.66	0.96	1.00	0.67	0.64	0.93	0.71
f14	0.81	0.75	0.24	0.89	0.84	0.96	0.86	0.70	0.96	0.94	0.93	0.82	0.67	1.00	0.99	0.85	0.98
f15	0.82	0.75	0.25	0.88	0.85	0.94	0.82	0.66	0.94	0.93	0.92	0.79	0.64	0.99	1.00	0.82	0.97
f16	0.67	0.62	0.17	0.80	0.70	0.95	0.98	0.93	0.94	0.86	0.82	0.98	0.93	0.85	0.82	1.00	0.87
f17	0.81	0.75	0.25	0.90	0.84	0.98	0.86	0.70	0.98	0.98	0.97	0.86	0.71	0.98	0.97	0.87	1.00

Table 11. Average correlations for the polynomial based filters.

Average	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	f11	f12	f13	f14	f15	f16	f17
f1	1	1	0	1	1	1	0	0	1	1	1	0	0	1	1	0	1
f2	1	1	0	0	1	1	0	0	1	1	1	0	0	1	1	0	1
f3	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
f4	1	0	0	1	1	1	1	0	1	1	1	1	0	1	1	1	1
f5	1	1	0	1	1	1	0	0	1	1	1	0	0	1	1	0	1
f6	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
f7	0	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1
f8	0	0	0	0	0	1	1	1	1	0	0	1	1	1	0	1	1
f9	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1
f10	1	1	0	1	1	1	1	0	1	1	1	1	1	1	1	1	1
f11	1	1	0	1	1	1	1	0	1	1	1	1	0	1	1	1	1
f12	0	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1
f13	0	0	0	0	0	1	1	1	1	1	0	1	1	0	0	1	1
f14	1	1	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1
f15	1	1	0	1	1	1	1	0	1	1	1	1	0	1	1	1	1
f16	0	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1
f17	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Table 12. Average correlations for the polynomial based filters above 0.7.

On a 0.7 correlation level four out of the five investigated elementary feature selection methods can be approximated by an appropriate polynomial. The only exception is the envelope eccentricity, where the correlation level of the best approximation is 0.31.

7 Generation of LAD objects by a restricted enumeration tree

Assume that there are some LAD objects, e.g. patterns of different kinds or discriminants, to be generated. These objects are generated by iterative methods such that in each iteration one Boolean variable or an attribute is selected. The temporary effect of the selected Boolean variable or attribute depends on the previously selected Boolean variables and attributes. Generally, the effect can be represented by a numerical value.

Example. In generation of positive patterns, i.e. the pattern must take value 0 on all negative observations and 1 on some positive ones, the effect can be measured with the weighted sum of the negative and positive observations refused by the newly selected Boolean variable but not refused by the previously selected ones. Obviously, the weight of the number of the refused negative (positive) observations must be nonnegative (nonpositive), if it is claimed that higher the weight is, the better the variable is. If two weights are (1,0) then we obtain a greedy method for selecting patterns of low degree, while in the case (1,-1) the method tries to generate a pattern satisfied with many positive observations.

We suggest the following method if the above mentioned circumstances hold. The generation is organized in a restricted enumeration tree. In each node of the tree it is allowed to select not only the best Boolean variable or attribute, but the k best ones (k 2), each of them defining a new branch of the tree. It is allowed that the selected Boolean variables or attributes must satisfied certain constraints. If less than k of them satisfy this constraint then we shall have less then k branches at this node. It is also allowed that if none of them satisfy the constraint then the node becomes a leaf. Computational experiments showed it is also useful to restrict the maximal depth of the tree. E.g. if k=2 and the maximal allowed depth is 10 than the enumerated tree is a binary one having at most 210 leaves.

7.1 Pattern generation

Assume that the database contains either binary variables or the non-binary (the categorical and numerical) variables have been binarized. The traditional way of pattern generation is to check all of the products of length at most k, where k is a small positive integer. If the database has many attributes (more than 1000 binarized variables), then to enumerate even every 3-tuples of attributes cannot be carried out within a reasonable time. It worth to note, that there are many medical databases having much more attributes. In this case the traditional methods are unable to generate patterns of higher degree which might be important for the problem. Then it is very natural to apply the restricted enumeration tree method.

Dataset	Side	Tree width	Max depth	Min hom	Min prev	Degree 1	2	3	4	5	6
breast	0	6	6	0.95	0.75	0	0	19	202	150	0
breast	1	6	6	0.95	0.6	0	3	36	0	0	0
lymphoma	0	10	8	0.95	0.6	0	1	224	149	0	0
lymphoma	1	8	8	0.95	0.85	1	23	291	1853	6274	13702...
ovarian	0	8	8	0.95	0.6	0	60	200	460	711	0
ovarian	1	5	5	0.95	085	0	25	125	287	455	0
l. ovarian	0	3	8	0.95	0.75	0	0	6	81	221	282
l. ovarian	1	3	8	0.95	0.6	0	5	27	67	80	0

Table 13. The number of patterns found by restricted enumeration.

7.2 Perfect patterns

A positive (negative) pattern is called perfect, if it is satisfied by all of the positive (negative) observations and refuses all of the negative (positive) observations. If the attributes in a perfect pattern are numerical ones, then their cut points must have the property that one class is completely on one side of the cut point. It is a natural approach to apply the restricted enumeration tree method for those binary variables which are defined by these cut points.

Perfect patterns have been generated for the investigated datasets. The numbers of the found perfect patterns obtained in a restricted enumeration tree is in Table 13.

Notice, that perfect pattern generation is an important subproblem of LAD where not traditional well-separating cut points are useful.

Dataset	Side	Tree width	Max depth	Degree 2	3	4	5	6	7	8
breast	0	6	6	0	0	0	48	21897	0	0
breast	1	6	6	0	0	1	1745	32180	0	0
lymphoma	0	8	8	1	443	382	0	0	0	0
lymphoma	1	8	8	0	2	513	8180	96578	87147	0
ovarian	0	8	8	0	0	995	15831	30303	0	0
ovarian	1	8	8	3	148	2079	1820	0	0	0
l. ovarian	0	2	8	0	0	0	0	0	7	177
l. ovarian	1	2	8	0	0	0	0	0	0	73

Table 14. The number of perfect patterns found by restricted enumeration.

8 Acknowledgement

This research were basically carried out when the authors visited RUTCOR and DIMACS. Both authors are grateful for the kind hospitality of RUTCOR, DIMACS, NSF grant number NSF-IIS-0312953.

References

[1] G. Alexe, S. Alexe, P. L. Hammer, L. Liotta, E. Petricoin, M. Reiss. Ovarian cancer detection by logical analysis of proteomic data. Proteomics, 3 (2004), 766-783.

[2] Alexe, G., Alexe, S., Axelrod, E. D., Boros, E., Hammer P. L., Reiss, M., Combinatorial analysis of breast cancer data from gene expression microarrays, submitted.

[3] E. Boros, Private communication 2003.

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

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

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

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

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

[9] 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).

[10] 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.

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

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

[13] 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.

[14] 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.

[15] 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.

[16] Alexe, S., Hammer, P. L. Accelerated Algorithm for Pattern Detection in Logical Analysis of Data. RUTCOR Research Report, RRR 59-2002, Discrete Applied Mathematics (2004), (in print).

[17] Koda, Y., Ruskey, F. A Gray code for the ideals of a forest poset. Journal of Algorithms 15 (1993) 324-340.

[18] 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.

[19] 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.

[20] Alexe, G., Alexe, S., Hammer, P. L., Pattern-based clustering and attribute analysis, Proceedings of Workshop on Discrete Mathematics and Data Mining, SIAM-Society for Industrial and Applied Mathematics, San Francisco, May 2003; Soft Computing (in press).

[21] Zeeberg, B.R., et al, GoMiner: A Resource for Biological Interpretation of Genomic and Proteomic Data.Genome Biology, 2003 4(4):R28.