DIMACS Technical Report 2002-49
December 2002

Comprehensive vs. Comprehensible Classifiers in Logical Analysis of Data

by

Gabriela Alexe[1], Sorin Alexe[1], Peter L. Hammer[1,2], and Alexander Kogan[1,2]

RUTCOR-Rutgers Center for Operations Research
Rutgers, The State University of New Jersey
640 Bartholomew Rd.
Piscataway, NJ 08854-8003, U.S.A.
e-mail: {alexe@rutcor.rutgers.edu, salexe@rutcor.rutgers.edu, hammer@rutcor.rutgers.edu, kogan@rutcor.rutgers.edu}

[1] We gratefully acknowledge the partial support from the Office of Naval Research (Grant N00014-92-J-1375) and from DIMACS.
[2] Permanent member

ABSTRACT

The main objective of this paper is to compare the classification accuracy provided large, comprehensive collections of patterns (rules) derived from archives of past observations, with that provided by small, comprehensible collections of patterns. This comparison is carried out here on the basis of an empirical study, using several publicly available datasets. The results of this study show that the use of comprehensive collections allows a slight increase of classification accuracy, and that the "cost of comprehensibility" is small.

1 Introduction

The extraction of knowledge hidden in records of past observations is a common problem in practically every area of science, engineering and business, and is the central object of study of classical disciplines, like statistics, and newer ones, like machine learning and data mining. Numerous methods have been developed to address this type of problems, and substantial success with their use is reported in the literature.

The *Logical Analysis of Data* (*LAD*), introduced in [8], [9], is a combinatorial approach to the analysis of datasets consisting of "positive" and "negative" observations, each of which is represented as a vector of n (usually real) attribute values. It has been established in previous studies ([7]) that *LAD* provides a competitive classification tool comparable in efficiency with the best classification techniques available. Additionally, through the systematic use of the concept of pattern, *LAD* has the advantage of offering clear explanations for which a particular new observation is classified as positive or as negative.

The basic approach of *LAD* is to derive from the set of past observations a large collection of patterns, some being characteristic for the observations having a positive classification, and the other ones being characteristic for the observations with a negative classification. Such a large collection of patterns is then "filtered" by the *LAD* procedure, in order to arrive at a much smaller, irredundant collection of patterns, which provides the same as the initial, large collection. The reasons for *LAD* to reduce (by filtering) the large, comprehensive collection of patterns to a small one, include its objective of providing comprehensible explanations for each classification. While the explanatory power of *LAD* is due to the comprehensibility of each individual pattern, this capacity can be obscured by the consideration of an excessive number of patterns.

It is interesting to note that none of the previous *LAD* studies has examined the effects of pattern filtering on the classification accuracy of *LAD* procedures, when applied to "new" observations, i.e., observations not contained in the original dataset. A systematic evaluation of the relative advantages and disadvantages of the large, comprehensive pattern sets, in contrast with small, comprehensible ones, is the main object of this paper. The comparison of the role of comprehensiveness versus comprehensibility is focused on the two types of patterns ("strong prime" and "strong spanned"), which had been shown in [10] to be of central importance in knowledge extraction.

After introducing some basic concepts of *LAD*, we describe two algorithms for generating comprehensive collections of strong prime and strong spanned patterns. We present after that the results of a computational study, carried out on four publicly available datasets, and consisting in the evaluation of the accuracy of various comprehensive and comprehensible *LAD* classifiers. Our results show on the one hand that comprehensive collections of patterns do provide somewhat higher classification accuracy, and on the other hand that the loss of classification accuracy due to the use of a comprehensible collection of patterns, seems to be relatively small.

2 Basic Concepts

We shall assume in this paper that the dataset Ω appears in the form of two disjoint sets Ω^+ and Ω^- of *positive* and *negative* observations, represented as points in R^m. A transformation, called *binarization*, mapping every point of Ω^+ and Ω^- into a point of $\{0,1\}^n$, for an appropriately chosen *n*, will be described below. The set of binary images of Ω^+ in this transformation, will be called the set *T* of *true points*, while the set of binary images of Ω^- in this transformation, will be called the set *F* of *false points*.

Following [6], let us introduce for every real variable X_j, a set of *cutpoints* $c_{j_1},\dots,c_{j_{k_j}}$ and a set of binary variables $x_{j_1},\dots,x_{j_{k_j}}$, defined by $x_{j_h}=\begin{cases}0, & X_j \le c_h\\ 1, & X_j > c_h\end{cases}$. The binarization is called *proper* if the binary image of any positive observation is different from the binary image of any negative observation. Obviously, if we disregard the total number of binary variables introduced, it is extremely easy to find a proper binarization, e.g., by simply subdividing for every *j* the interval $[\min_{Z\in\Omega} z_j, \max_{Z\in\Omega} z_j]$, into a sufficiently large number of equal size subintervals; here z_j represents the j^{th} coordinate of every vector *Z* in Ω. It was shown in [6] that finding a proper binarization using a minimum number of binary variables is NP-hard. In order to avoid the difficulties of finding an minimum size system of cutpoints, in the computational experiments of this paper, the systems of cutpoints, are consistently based on the same simple heuristic procedure.

We shall briefly describe below several basic concepts in Boolean logic. A *partially-defined Boolean function (T,F)* consists of two disjoints sets of Boolean vectors *T* and *F*, called the *true* and the *false* vectors, respectively; if $T \cup F = \{0,1\}^n$, the pair (*T*,*F*) defines a *Boolean function*. To every Boolean variable *u* we shall associate its *negation* $\bar{u} = 1-u$, and shall refer to both *u* and $\bar{u}$ as the *literal*s. A *term* is a conjunction (product) of distinct literals, which does not contain both a variable and its negation. The *degree* of a term is the number of literals in it. A term of degree *n* will be called a *minterm*. Note that minterms are in one-to-one correspondence with Boolean vectors. Terms can be interpreted geometrically as *subcubes* of the *n*-dimensional cube $\{0,1\}^n$. Whenever it does not cause a confusion, we may refer interchangeably to terms or to the corresponding subcubes.

We shall say that a term *C covers* a point *Z* if and only if $C(Z) = 1$. The Boolean subcube of $\{0,1\}^n$, not necessarily included in $T \cup F$, corresponding to the points covered by a term *C* will be denoted by $S(C)$, while $S(C) \cap (T \cup F)$ will be called the *coverage* of *C,* and denoted by $COV(C)$. Let us further introduce the concepts of positive, respectively negative, coverage of a term *C* defined by $COV^+(C) = COV(C) \cap T$, and $COV^-(C) = COV(C) \cap F$, respectively. The *relative positive* (respectively, *negative*) coverage of a term *C* is defined as $|COV^+(C)|/|T|$ (respectively $|COV^-(C)|/|F|$).

LAD was built around two central concepts: (positive or negative) patterns and (positive or negative) theories. Following the terminology of [8], for any number $\chi \in (0,1]$, a term *C* will be called a *positive χ-pattern* of (*T,F*) if

1. $COV(C) \neq \emptyset$, and
2. $|COV^{+}(C)| \geq \chi |COV(C)|$,

and χ will be called the *homogeneity* of C. Note that condition 1 is equivalent to the condition that $C(Z) = 1$ for at least one vector $Z \in T$. It should be remarked that in most practical situations, the value of χ is "close" to 1. A *negative* χ*-pattern* is defined in a similar way, by replacing condition 2 with the condition that $|COV^{-}(C)| \geq \chi |COV(C)|$. It is frequently of interest to use different χ values for the definition of positive and of the negative χ-patterns, e.g., positive 0.9 - patterns and negative 0.75 - patterns can be produced.

Since the properties of positive and negative patterns are completely symmetric, without loss of generality we shall limit some of the discussions in this paper to the case of positive patterns, and shall frequently refer to positive patterns simply as patterns, whenever this cannot lead to any ambiguity.

Given a set of properties Π, we define a Π*-pattern* as a term which satisfies all the properties Π. Clearly, a χ-pattern is a special case of a Π-pattern, corresponding to the case when Π consists of the two conditions 1 and 2 above. Usually, in addition to conditions 1 and 2 above, Π may also include constraints on the term C, such as a lower bound on $|COV(C)|$, and /or an upper bound on $|Lit(C)|$, where $Lit(C)$ is the set of literals of C.

Notice that in the special case of Boolean functions (in contrast with the case of partially defined Boolean functions), condition 1 is superfluous. In that case, a term which satisfies condition 2 for $\chi = 1$ is called a *positive (negative) implicant* of that Boolean function. Clearly, in the case of Boolean functions the concept of a positive 1-pattern reduces to that of an implicant.

We model the suitability of various types of patterns in *LAD*, by introducing partial preorders defined on the set of patterns. A pattern P is called *Pareto-optimal* with respect to a partial preorder ρ defined on the set of patterns, if there is no distinct pattern P' such that $P' >_{\rho} P$. It has been shown that the most relevant preorder on the set of patterns is defined by the so-called *evidential preference* ([10]), which states that a pattern P is evidentially-preferred to a pattern P' if and only if $COV(P) \supseteq COV(P')$. Patterns which are evidentially Pareto-optimal are called *strong patterns*.

It can happen that the same subset of positive observations is covered by several distinct strong positive patterns. It is therefore important to consider two opposite refinements of the evidential preference. In a loose sense, in the *conservative* case, preference is given to the "longest" pattern having the same positive coverage, while in the *aggressive* case, preference is given to an irreducible pattern having the same positive coverage. More formally, the secondary preference is defined on the basis of the set of literals of a pattern. The *simplicity preference* ([10]) states that a pattern P is preferred to a pattern P' if $Lit(P) \subseteq Lit(P')$. On the other hand, the *specificity preference* ([10]) states that a pattern P is preferred to a pattern P' if $Lit(P) \supseteq Lit(P')$. Those patterns which are Pareto-optimal with respect to evidence refined by specificity are called *strong spanned*, while those patterns which are Pareto-optimal with respect to evidence refined by simplicity are called *strong prime* patterns.

For a given type of Pareto-optimality, the set of all Pareto-optimal positive (negative) Π-patterns is called the *positive (negative) Π-compendium*, and their union is called the *Π-pandect*. In this paper, we shall focus our attention on the strong spanned and the strong prime Π-compendia and Π-pandect.

3 Theories, Models, Compendia, Pandects and Discriminants

Similarly to the concept of theories defined in *LAD*, given a positive Π-compendium of a partially-defined Boolean function (T,F), we shall define a *positive Π-theory* $\mathcal{T}$ as a minimal subset of the compendium $\{P_1, \ldots, P_k\}$ with the property that for every $Z \in \mathcal{T}$ which is covered by at least one of the patterns in the compendium, there exists a $P_i \in \mathcal{T}$ such that $P_i(Z) = 1$. A theory $\mathcal{T}$ is associated with the Boolean function represented by the disjunction of terms (DNF) $\bigvee_{i=1}^{k} P_i$. The value of this Boolean function at a Boolean vector Z, where $Z \notin T \cup F$, can be used to predict whether Z is true or false. Similarly, a positive Π-compendium can also be used to predict whether Z is true or false. Note that if a positive theory predicts that Z is true, then the corresponding positive compendium will also predict that Z is true, but the reverse inequality does not necessarily hold.

A *negative Π-theory* can be defined similarly. Obviously, one can use a negative theory and a negative compendium to predict whether a Boolean vector $Z \notin T \cup F$ is true or false. In this case, if a negative theory predicts that Z is false, then the corresponding negative compendium will also predict that Z is false, but not necessarily vice versa.

A pair consisting of a positive and a negative Π-theory will be called a *Π-model*. It has been remarked above that a positive Π-theory, as well as a negative Π-theory taken separately, can be considered as a "classifier". This use of theories as classifiers is easy to justify; indeed, it is true that every Boolean vector $Z \in T \cup F$ is true (false) if it satisfies some positive (negative) pattern of the positive (negative) χ-theory for $\chi = 1$, and it is false (true) if it does not satisfy any of them. The basic assumption of *LAD* is that the above implications hold not only for a Boolean vector $Z \in T \cup F$, but for many Boolean vectors.

When considering a model, a Boolean vector Z can be in four situations:

Case 1. Z satisfies a pattern in the positive theory, and none in the negative theory.
Case 2. Z satisfies a pattern in the negative theory, and none in the positive theory.
Case 3. Z satisfies a positive and a negative pattern in the model.
Case 4. Z does not satisfy any pattern in the model.

A model can be also used as a classifier, by considering a point as being true if it is in case 1 described above, false in case 2, and "unclassified" in cases 3 and 4. It is interesting to note that if models use 1-theories, because of the definition of the concept of 1-patterns, then every Boolean vector in T or F belongs to one of the first two cases described above.

The role of the pandect in classification is entirely analogous of that of a model; the corresponding classification is defined as above by replacing in each of the four classes listed above the words “theory” and “model” by the words “compendium” and “pandect”, respectively.

Notice that cases 3 and 4 listed above are quite different. Indeed, in case 4 none of the patterns which are taken into consideration provides any information, while in case 3 conflicting information is available. The contradictory information in case 3 may not necessarily be in “balance”, since one of the two pieces of information may be “dominating” the other. For example, if a Boolean vector $Z\notin T \cup F$ satisfies an “important” group of positive patterns, and a “negligible” group of negative patterns, it is reasonable to assume that it is a true vector.

In order to give a precise meaning to the words ‘important” and “negligible” we shall introduce the concept of discriminant. Given a model or pandect consisting of the positive patterns P_i, ($i\in I$), and negative patterns N_j ($j\in J$), a pseudo-Boolean function

$$\Delta(Z)=\sum_{i\in I}\alpha_i P_i(Z)-\sum_{j\in J}\beta_j N_j(Z)$$

will be called a *discriminant*, having given positive “weights” α_i and β_j . The sign of the discriminant is used to classify observations in case 3 above: if $\Delta(Z)>0$ the observation is classified as true, if $\Delta(Z)<0$ the observation is classified as false, and if $\Delta(Z)=0$ the observation is unclassified.

In previous *LAD* studies, various choices of the weights α_i and β_j have been considered. In [11] and [3], such a discriminant -- having all positive (respectively, negative) weights -- was used to distinguish high- and low-risk among cardiac patients. This particular discriminant, called *prognostic index*, or *LAD score*, was proved to have a theoretical connection with the correlation coefficient between the current observation and an "ideal" one, and to be very reliable for classification. Consequently, in the computational experiments reported in this paper, all the α_i ’s (as well as the β_j’s) were considered to be equal, and to add up to 1.

4 Pattern Generation Algorithms

4. 1 Strong Prime *Π* - Patterns

Prime patterns, and in particular those of small degree, play an important role in *LAD*-based classification, both because of their explanatory power, and because of the availability of efficient algorithms for their generation.

In this section, we propose an efficient algorithm (“almost” linear in the number of all possible conjunctions) for the generation of all strong prime *Π*-patterns. For the sake of simplicity, the presentation below will be confined to the case of datasets with two numerical attributes; the details of the general case appear in [4].

Let us consider a proper binarization of two numerical variables X and Y, resulting in two sets of binary variables which are associated to the two sets of real-valued “cutpoints” $c_1 < c_2 < \dots < c_p$, and $d_1 < d_2 < \dots < d_q$, by the relations:

$$x_i=\begin{cases}0, & X\le c_i\\ 1, & X>c_i\end{cases} \qquad (\ i = 1, 2, \dots , p),$$

and

$$y_j = \begin{cases} 0, & Y \le d_j \\ 1, & Y > d_j \end{cases} \qquad (j = 1, 2, \dots, q).$$

It is obvious that, for any $i' < i''$, the corresponding binary variables must satisfy the relations $x_{i'} \ge x_{i''}$ and $\bar{x}_{i'} \le \bar{x}_{i''}$; a similar remark holds for the y's. In view of this fact, it is clear that the longest conjunctions in the binarized space will be of degree four, having the form $C = x_i \bar{x}_j y_k \bar{y}_l$, $i < j$, $k < l$. Clearly, the set of all those points in R^2, which have the property that the conjunction C takes the value 1 in all the binarized points associated to it, form an interval $(c_i, c_j] \times (d_k, d_l]$ in R^2, while if C is a lower degree conjunction, i.e., one in which some of the four literals $x_i, \bar{x}_j, y_k, \bar{y}_l$ are missing, then the corresponding interval will be unbounded.

The algorithm for the generation of all strong Π-prime patterns is based on the evaluation of $|COV^+(C)|$ and $|COV^{--}(C)|$, for all possible conjunctions C, and the subsequent selection of those which verify the properties Π. We shall present the algorithm for the evaluation of the values of

$|COV^+(C)|$ only, since the evaluation of the values of $|COV^{--}(C)|$ is done in a similar way.

First, the algorithm builds the matrix $M = (m_{ij})_{i=0,1,\dots,p;\, j=1,2,\dots,q}$, whose entries m_{ij} are defined by $m_{ij} = |COV^+(x_i \bar{x}_{i+1} y_j \bar{y}_{j+1})|$ (where, if a variable index is out of range, the corresponding literal is missing).

For every pair of integers $k \in [0, p]$ and $l \in [0, q]$, with k between 0 and p and l between 0 and q, the algorithm computes a matrix $R^{(k,l)}$, having as entries (i,j) the values of $|COV^+(x_i \bar{x}_k y_j \bar{y}_l)|$; the pair (k,l) is called the *basis* of the matrix.

The matrices $R^{(k,l)}$, are calculated sequentially, on the basis of a Gray-code enumeration of their bases, in such a way that:

(a) all the bases between (0,0) and (p,q) are generated,

(b) no basis is generated twice, and

(c) two consecutive bases differ in exactly one component, by exactly one unit.

The matrix $R^{(p,q)}$ is calculated using the following algorithm:

```
R := M;
For i := p downto 0 do
    For j := q-1 donwto 0 do
        r_{i,j} := r_{i,j+1} + r_{i,j};
For j := q donwto 0 do
```

For $i := p$-1 downto 0 do

$r_{i,j} := r_{i+1,j} + r_{i,j}$;

As soon as the matrix $R^{(p,q)}$ is computed, the algorithm proceeds by recursion to the next basis in the Gray sequence. Denoting the current basis by (k, l) the next basis is chosen from the set: {(k+1, l), (k-1, l), (k, l+1), (k, l-1)}. Since the four cases are treated in essentially similar ways, we present here only the case when the new basis is (k, l-1). In this case:

```
For  i := 0 to p do
{        For j := 0 to l - 1 do
             r_{i,j} :=  r_{i,j} - r_{i,l};
         For j := l to q do
             r_{i,j} := r_{i,j} + r_{i,l-1}.
}
```

Example. Let us consider $p = 3$, $q = 3$, and the initial matrix

$$M = \begin{pmatrix} m_{00} & m_{01} & m_{02} & m_{03} \\ m_{10} & m_{11} & m_{12} & m_{13} \\ m_{20} & m_{21} & m_{22} & m_{23} \\ m_{30} & m_{31} & m_{32} & m_{33} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 1 & 2 \\ 2 & 1 & 0 & 0 \\ 2 & 2 & 2 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}.$$

Here the first basis is (3,3) -- marked as a bold entry--, and the corresponding matrix is $R^{(3,3)} = \begin{pmatrix} 17 & 11 & 7 & 3 \\ 13 & 8 & 4 & 1 \\ 10 & 7 & 4 & 1 \\ 3 & 2 & 1 & \mathbf{0} \end{pmatrix}$. In this matrix each entry (i, j) represents the sum of the elements of bottom-right submatrix of M having (i, j) as the top-left. More precisely, the entry having the value 8 in $R^{(i,j)}$, representing $|COV^{+}(x_1 y_1)|$, equals the sum of the entries in the [1,3]×[1,3] submatrix $\begin{pmatrix} 1 & 0 & 0 \\ 2 & 2 & 1 \\ 1 & 1 & 0 \end{pmatrix}$ of M. The following step of the algorithm produces the matrix defined by the basis (3,2): $R^{(3,2)} = \begin{pmatrix} 14 & 8 & 4 & 7 \\ 12 & 7 & 3 & 4 \\ 9 & 6 & 3 & 4 \\ 3 & 2 & \mathbf{1} & 1 \end{pmatrix}$. The entry having the value 12, representing $|COV^+(x_1\bar{y}_3)|$, equals the sum of the entries in the [1,3]×[0,2] submatrix of M defined by the position of this entry (0,1) and the basis (4,3): $\begin{pmatrix} 2 & 1 & 0 \\ 2 & 2 & 2 \\ 1 & 1 & 1 \end{pmatrix}$.

For the computational experiments reported in this paper, we have generated the complete collections of strong prime patterns of limited degree d, for several datasets. In order to obtain these collections, we have generated all the subsets of cardinality at most d in the sets of numerical attributes of the given datasets. After binarization we have applied the algorithm described above to evaluate $|COV^+(C)|$ and $|COV^-(C)|$ for every possible conjunction C, involving binary variables associates to the numerical variables in each selected subset. Each pattern produced in the list was checked, and only those satisfying all conditions in properties Π were retained.

4. 2. Strong Spanned Π - Patterns

Spanned patterns are more "conservative" than the prime ones, having usually higher degrees. They "explain" a new observation only if it is contained in the "interval hull" of the observations covered by the pattern. It was shown in [2] that classification based on models using spanned patterns, are more "robust", having usually fewer errors, at the cost of leaving somewhat larger numbers of observations unclassified.

We present below a consensus-type algorithm for the generation of all spanned patterns, along with an implementation of it, which runs in incremental polynomial time. The method is similar to the well-known consensus method of Blake ([5]) and Quine ([12]) for finding prime implicants of a Boolean function. Malgrange ([13]) used a consensus-type approach to find all the maximal submatrices consisting of ones of a 0-1 matrix. Also, a consensus-type algorithm for finding all maximal bicliques of a graph is presented in [1]. An important feature of this consensus-type algorithm for generation of all spanned-patterns is the fact that its complexity does not depend on the number of cutpoints used for binarization.

Consensus-type methods enumerate the family of all maximal objects of a certain collection, by starting from a sufficiently large set of objects, and systematically completing it by the application of two simple operations. The operation of *consensus adjunction* associates to a pair of objects in the given collection one or more new objects, and adds them to the collection. The operation of *absorption* removes from the collection those objects which are "dominated" by other objects in the collection. The two operations are repeated as many times as possible, leading eventually to a collection consisting exactly of all the maximal objects.

We shall briefly present here a particular variant of the consensus method, which produces the complete collection of all its spanned positive patterns of a dataset Ω. For simplicity, we shall restrict only to χ-patterns, for χ = 1, which are usually called *pure* patterns.

Let $P=[a_1,b_1]\times\ldots\times[a_m,b_m]$ and $P'=[a'_1,b'_1]\times\ldots\times[a'_m,b'_m]$ be a pair of positive spanned 1-patterns, and let P'' be the interval $[a''_1,b''_1]\times\ldots\times[a''_m,b''_m]$, where $a''_i=\min\{a_i,a'_i\}$

and $b''_i = \max\{b_i, b'_i\}$, i = 1,..., m. If P'' is a positive 1-pattern, then it is called the *consensus* of the 1-patterns P and P'. Clearly, a pair of positive spanned 1-patterns can have at most one consensus, which is the 1-pattern spanned by the observations in $COV(P) \cup COV(P')$. We say that the positive spanned 1-pattern P *absorbs* the positive spanned 1-pattern P' if $P = P'$.

The proposed algorithm *SPIC*, which generates all positive spanned 1-patterns in incremental polynomial time, runs as follows:

Algorithm *SPIC*

Let C_0 be the collection of patterns spanned by each individual observation in Ω^+.

1. **Initiate** $C := C_0$.
2. **Repeat** the following operation **until** the collection C cannot be furthermore enlarged:

For every pair of patterns P_0 in C_0 and P in C, if their consensus P' exists and it is not already contained in C, then add it to C.

Example. Let us illustrate algorithm *SPIC* for the dataset $\Omega = \{v_1 = [1,0,2], v_2 = [0,2,0], v_3 = [3,1,1], v_4 = [2,0,2]\}$, having all the observations but v_2 positive.

- The input collection C_0 is $\{P_1 = [1,1] \times [0,0] \times [2,2]$, $P_3 = [3,3] \times [1,1] \times [1,1]$, $P_4 = [2,2] \times [0,0] \times [2,2]$. Initialize $C := C_0$.
- Perform consensus adjunction for the pair of patterns P_1 in C_0 and P_3 in C: the candidate for consensus is $P_{1,3} = [1,3] \times [0,1] \times [1,2]$, having $COV(P_{1,3}) = \{v_1, v_3, v_4\} \subseteq \Omega^+$; since $P_{1,3}$ is not contained in C, it is added to C.
- Perform consensus adjunction for the pair of patterns P_1 in C_0 and P_4 in C: the candidate for consensus is $P_{1,4} = [1,2] \times [0,0] \times [2,2]$, having $COV(P_{1,4}) = \{v_1, v_4\} \subseteq \Omega^+$; since $P_{1,4}$ is not contained in C, it is added to C.
- Perform consensus adjunction for the pair of patterns P_3 in C_0 and P_4 in C: the candidate for consensus is $P_{3,4} = [2,3] \times [0,1] \times [1,2]$, having $COV(P_{3,4}) = \{v_3, v_4\} \subseteq \Omega^+$; since $P_{3,4}$ is not contained in C, it is added to C.

The consensus of any other pair of patterns from C and C_0 is contained in C. The algorithm stops and outputs the family of all positive spanned patterns $C = \{P_1, P_3, P_4, P_{1,3}, P_{1,4}, P_{3,4}\}$.

The proof of correctness of algorithm *SPIC* (i.e. of the fact that it stops after a finite number of steps, producing at termination the list containing all the pure positive 1-patterns spanned by subsets of observations in the dataset), as well as its worst-case complexity, are presented in [2].

The algorithm *SPIC* runs in incremental polynomial time, the total running time of it being $O(\beta m^{+}(m+m^{+}n))$, where β is the number of spanned patterns, n is the number of attributes, m is the number of observations, and m^{+} is the number of positive observations in the dataset.

In all real-life applications encountered, the number of spanned patterns was extremely high. In view of this situation, it was important to apply various filtering mechanisms to restrict the number of 1-patterns produced, and to keep in this way both time and memory requirements at an acceptable level. The final list of spanned 1-patterns is obtained from the list *C*, based on several selection criteria, which include restrictions on the number of 1-patterns produced, total time allocation, and the characteristic parameters of the retained 1-patterns.

The implemented version of the algorithm includes several accelerating heuristics. One of the important procedures used for this purpose, partitions the original dataset into several subsets, applies the input-consensus algorithm separately to the subsets, and after eliminating redundancies in the union of these subsets, creates a final list of spanned patterns. Another heuristic included in the current implementation of the algorithm applies a pre-selection mechanism along the process, eliminating from consideration those patterns whose parameters (prevalence, homogeneity, etc) are not sufficiently high.

The list *L* of all strong positive spanned patterns can be easily obtained from the output collection *C*, by selecting from *C* the maximal elements with respect to set inclusion. The list *L* can be also be produced and updated gradually, during the consensus-type procedure: *L* is initialized with the empty set, and whenever a consensus candidate, say *P*, is added to *C*, it is checked whether *P* is already contained in a pattern in *L*. If the test fails, then *P* is added to *L*, and all patterns in *L* which are contained in *P* are deleted from *L*. The selection of all strong pure spanned patterns can be performed in an additional time of order $O(\beta^2)$; however, we are not able to guarantee yet a total polynomial-time for producing all strong spanned patterns. In fact, the dualization problem of a monotone non-decreasing Boolean function can be reduced in quadratic time to the problem of generating all strong spanned patterns of a certain dataset (see [12]). Thus, the existence of a total polynomial-time algorithm for generating all strong spanned patterns would imply the existence of a total polynomial-time algorithm for the dualization problem mentioned above; until now, the best known algorithms are pseudo-polynomial.

5 Evaluation Methodology

A comparative study of the suitability of different types of patterns for classification, carried out in [10], has established the fact that the strong prime and the strong spanned patterns are the two most suitable ones to be used for this purpose. In view of this fact, in the present study the comparison of compendia, pandects, theories and models is done for both strong prime and strong spanned patterns.

The accuracy of *LAD* classifications was evaluated using cross-validation. In each experiment, *LAD* was run on the training set and the classification accuracy was evaluated on the test set using the following measure of accuracy:

$$ACC = \frac{1}{2}\left[a + e + \frac{1}{2}(c + f)\right],$$

here *a*, *b*, *d*, *e* are the percentages corresponding to the entries in the following table, calculated on the test set:

	Predictions		
	True	False	?
% of True Observations	*a*	*b*	*c*
% of False Observations	*d*	*e*	*f*

Note that $a+b+c = d+e+f = 100$. When the *LAD* classification is obtained using a theory or a compendium, then $c = f = 0$, and therefore, the expression for accuracy will be reduced to $ACC = \frac{a+e}{2}$. However, if the *LAD* classification is obtained using a model or a pandect, then *c* and *f* are not neccesarily 0 anymore, showing that that the expression of *ACC* is based on the idea that one unclassified observation can be expected to account for 0.5 errors.

Since different datasets may exhibit different properties in classification problems, this comparative study was carried out on four datasets, publicly available from the UC Irvine repository (http://www.ics.uci.edu/~mlearn/MLRepository.html). These four datasets are the following:

- **Wisconsin breast cancer (bcw)**. In this dataset 683 observations (obtained after the removal of 16 instances which contain missing attributes) represent malignant or benign breast tissues, each observation being represented by nine numerical attributes.

- **BUPA liver disorders (bld).** In this dataset 345 observations represent male patients some of whom had a liver disorder, each patient being represented by 6 numerical attributes corresponding to blood tests and alcohol consumption.

- **StatLog heart disease (hea).** This dataset contains the records of 270 patients, indicating for each of them the presence or absence of heart disease, together with the numerical results of seven medical tests, and six binary ones.

- **Congressional voting records (vot).** This dataset contains the voting records of the 435 members of the U. S. House of Representatives of the 98th Congress, each being classified as a Democrat or a Republican. The 16 attributes represent the votes of the representatives on 16 issues, encoded as 1, 0 and 0.5, the latter corresponding to the absence of vote.

The choice of parameter values in the set of properties *Π* which define the *Π*-patterns, affects the performance of *LAD* classification. It is known from previous *LAD* studies ([7]) that the two most critical parameters influencing the effectiveness of *LAD* are the relative coverage and the homogeneity of patterns, and that the optimal choice of parameter values is dataset dependent.

The first stage of this experimental study consisted in identifying for every dataset a "good" choice of parameter values. The homogeneity parameter χ was varied from 0.7 to 1.0, in steps of 0.05, and the relative coverage of the positive (respectively, negative) patterns, was varied from 0.1 to 0.5, in steps of 0.1. In the generation of strong prime patterns, their degree was limited to 3. The aim of a first set of experiments was to determine the best parameter values for relative coverage and homogeneity for every dataset, and for each of the two types of patterns (strong prime and strong spanned). These values were used then consistently in the second (main) set of experiments aimed at estimating the relative accuracy of *LAD* classifications based on theories, models, compendia and pandects.

The first set of experiments consisted of three independent runs of twofold cross-validations for every combination of the parameter values. The preliminary estimate of *LAD* classification accuracy was determined as the average accuracy of the (3×2=) 6 experiments. For each of these experiments six *LAD* classifications were evaluated, based on the positive and negative theories and compendia, as well as on the corresponding model and pandect. These estimates of the six *LAD* classifications were again averaged, and used for selecting that combination of the two parameter values for which the highest average accuracy was achieved.

The second set of experiments consisted of ten independent runs of 2-fold cross validations for the chosen combination of parameter values**.** The average accuracy was determined separately for each of the six *LAD* classifications mentioned above. Then the six sequences of 10 numbers (corresponding to the 2-foldings) were used to compare the relative accuracies of *LAD* classifications based on positive and negative theories and compendia, as well as on the models and the pandects. This comparison was done on the basis of the paired two-tail *t*-test. All experiments were carried out separately for strong prime and strong spanned patterns.

6 Empirical Results

The specific combinations of the parameter values for each of the datasets and for both pattern-types, which were obtained in the first set of experiments, are reported in Table 1. For example, the best parameter value combination when using strong prime positive patterns for the **hea** dataset was that consisting of the requirement that (*i*) the patterns should cover at least 50% of the true observations in the training set, and (*ii*) to have homogeneity of at least 0.9 (i.e., that the number of false observations covered by the pattern should be at most 10% of its coverage).

Dataset	Strong Prime Patterns		Strong Spanned Patterns	
	Relative Coverage	Homogeneity	Relative Coverage	Homogeneity
vot	0.3	1	0.3	1
bcw	0.2	1	0.2	1
hea	0.5	0.9	0.3	1
bld	0.1	0.95	0.1	0.95

Table 1

The 12 *LAD*-based classifiers used in this study (compendia, pandects, theories and models, based either on strong prime or on strong spanned patterns), are listed in Table 2, and denoted as C1, C2, ... , C12. In all these definitions we have called "true" the smaller of the two given sets of observations *T* and *F*, and called "false" the other one.

Strong Patterns	Compendium		Pandect	Theory		Model
	Positive	Negative		Positive	Negative	
Prime	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**
Spanned	**C7**	**C8**	**C9**	**C10**	**C11**	**C12**

Table 2.

We report in Table 3 the average accuracies of *LAD*-based classifiers, as obtained in the second series of experiments. The rows of the table correspond to the four datasets considered, while the columns correspond to the 12 classifiers, and to the ratio and the difference of the maximum and minimum accuracies.

Dataset	LAD Classifiers												MAX/MIN	MAX-MIN
	Based on Strong Prime Patterns						Based on Strong Spanned Patterns							
	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12		
vot	89.97%	95.53%	94.57%	92.59%	95.15%	94.38%	96.33%	95.95%	96.30%	95.23%	96.04%	95.60%	1.07	6.36%
bcw	95.87%	90.17%	96.18%	95.60%	94.54%	94.95%	85.54%	93.98%	91.64%	91.64%	94.51%	90.35%	1.12	10.64%
hea	75.72%	76.47%	80.50%	77.02%	77.99%	79.44%	72.83%	78.01%	79.62%	75.14%	79.96%	78.71%	1.11	7.67%
bld	62.03%	60.70%	67.23%	64.11%	62.04%	65.29%	67.80%	67.94%	71.47%	68.34%	68.50%	70.27%	1.18	10.77%

Table 3.

We report in Table 4 the critical confidence levels corresponding to the values of the *t*-statistic in the paired two-tail *t*-tests comparing the accuracy of the 12 *LAD*-based classifications for every dataset. A positive (negative) sign indicates that the accuracy corresponding to the row was higher (lower) than that corresponding to the column.

	t-test	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12
vot	C1	-100.00%	-100.00%	-100.00%	-100.00%	-100.00%	-100.00%	-100.00%	-100.00%	-99.99%	-100.00%	-100.00%
	C2		-99.01%	-99.94%	-80.22%	-99.13%	-89.71%	-69.54%	-90.56%	+62.96%	+82.21%	+18.95%
	C3			-99.99%	-84.31%	-64.06%	-99.11%	-98.65%	-99.40%	+79.73%	+98.98%	+96.53%
	C4				-99.75%	-99.94%	-99.94%	-99.94%	-99.96%	+99.52%	+99.93%	+99.89%
	C5					-96.17%	-95.23%	-94.67%	-97.65%	+13.96%	+96.82%	+75.20%
	C6						-98.92%	-98.82%	-99.34%	+85.22%	+99.16%	+97.13%
	C7							-79.40%	-11.76%	+99.91%	+49.21%	+95.69%
	C8								-94.38%	+95.73%	+30.49%	+85.92%
	C9									+99.54%	+56.19%	+96.62%
	C10										-96.31%	-91.45%
	C11											-97.03%

Table 4 a.

	t-test	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12
bcw	C1	+100.00%	-78.91%	+78.90%	+100.00%	+99.84%	+99.95%	+82.04%	+99.24%	+99.88%	+80.09%	+99.81%
	C2		-100.00%	-100.00%	-100.00%	-100.00%	+96.05%	-97.65%	-75.81%	-96.22%	-99.71%	-11.22%
	C3			+84.56%	+99.96%	+99.82%	+99.95%	+89.82%	+99.35%	+99.88%	+90.15%	+99.82%
	C4				+99.25%	+97.23%	+99.91%	+77.09%	+98.41%	+99.82%	+72.62%	+99.58%
	C5					-90.36%	+99.89%	+31.36%	+96.30%	+99.78%	+2.00%	+99.25%
	C6						+99.90%	+55.07%	+97.42%	+99.79%	+34.97%	+99.38%
	C7							-97.84%	-100.00%	-93.03%	-98.90%	-99.99%
	C8								+65.32%	+97.50%	-83.10%	+83.18%
	C9									-99.97%	-79.52%	+99.99%
	C10										-98.59%	+99.92%
	C11											+91.41%

Table 4 b.

	t-test	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12
hea	C1	-71.99%	-100.00%	-98.91%	-99.87%	-100.00%	+58.37%	-93.74%	-97.64%	+16.94%	-99.97%	-94.46%
	C2		-100.00%	-65.61%	-97.99%	-99.98%	+69.53%	-88.49%	-96.38%	+39.17%	-99.97%	-91.31%
	C3			+100.00%	+99.98%	+99.38%	+94.37%	+97.42%	+44.15%	+92.14%	+63.43%	+77.60%
	C4				-94.20%	-99.99%	+74.43%	-69.67%	-88.87%	+49.90%	-99.86%	-74.10%
	C5					-99.95%	+85.12%	-1.81%	+76.29%	-71.04%	-98.65%	-41.39%
	C6						+91.83%	+87.24%	-10.74%	+86.91%	-71.87%	+42.13%
	C7							-87.48%	-98.68%	-91.89%	-94.13%	-96.57%
	C8								-78.68%	+73.97%	-98.86%	-41.01%
	C9									+98.89%	-21.75%	+94.65%
	C10										-91.42%	-96.76%
	C11											+69.13%

Table 4 c.

	t-test	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12
bld	C1	+83.26%	-100.00%	-97.30%	-1.02%	-99.18%	-99.79%	-99.99%	-100.00%	-99.82%	-99.98%	-99.98%
	C2		-100.00%	-98.91%	-96.67%	-99.93%	-99.72%	-99.99%	-99.99%	-99.70%	-99.98%	-99.97%
	C3			+99.78%	+99.99%	+93.14%	-31.33%	-52.11%	-99.08%	-53.88%	-73.10%	-94.89%
	C4				+91.06%	-89.53%	-97.85%	-99.74%	-99.98%	-98.20%	-99.92%	-99.91%
	C5					-99.56%	-99.73%	-99.98%	-100.00%	-99.73%	-99.97%	-99.98%
	C6						-92.59%	-99.40%	-99.96%	-93.99%	-99.64%	-99.73%
	C7							-13.18%	-100.00%	-72.20%	-54.83%	-100.00%
	C8								-99.82%	-28.71%	-61.83%	-98.13%
	C9									+99.85%	+99.47%	+99.23%
	C10										-12.17%	-99.70%
	C11											-95.72%

Table 4 d.

In Table 5, we present the relative rankings of the average accuracies of the 12 *LAD*-based classifiers for each of the four datasets. The columns correspond to the datasets, and the rows correspond to the ranks. The entry in a cell (i, j) represents the classifier having rank i for dataset j.

Rank	vot	bcw	hea	bld
1	C7	C3	C3	C9
2	C9	C1	C11	C12
3	C11	C4	C9	C11
4	C8	C6	C6	C10
5	C12	C5	C12	C8
6	C2	C11	C8	C7
7	C10	C8	C5	C3
8	C5	C9	C4	C6
9	C3	C12	C2	C4
10	C6	C2	C1	C5
11	C4	C7	C10	C1
12	C1	C10	C7	C2

Table 5.

7 Conclusions

The main objective of this paper was to evaluate the comparative advantages and disadvantages of basing *LAD* classification on theories and models, or on compendia and pandects. In the initial *LAD* studies, classification was based on theories. In later developments, instead of just using a positive or a negative theory for classification, the use of models and discriminants associated to them became the norm. Although *LAD* algorithms did generate compendia and pandects, these were only viewed as intermediate collections from which theories and models had to be extracted.

The obvious appeal of theories and models consists in the fact that they contain much fewer patterns than compendia and pandects, and are therefore much easier to comprehend, manipulate, and use. However, it was natural to question the possible loss of classification robustness resulting from the reduction of a large collection of patterns to a minimal subset. It was important therefore to examine empirically whether such a reduction of compendia and pandects to theories and models is indeed beneficial. Such a comparative study was carried out in this paper.

The results of extensive computational experiments which were carried out in this investigation, including the 1,840 runs of *LAD*-based classification tests, and those of a series of comparative studies for the evaluation of the relative performances of the 12 classification systems examined, lead to a number of conclusions.

First, the optimal parameter values, the type of patterns (strong prime or strong spanned), the best types of the LAD-based classifiers are all heavily dataset-dependent, and many of them have high accuracies on some of the datasets. Therefore, when deciding on an optimal classifier for a new dataset, it is essential to evaluate all these choices.

Second, it is clear from the last two columns of Table 3 that the choice of a classifier which performs well on a specific dataset is essential, since a non-optimal choice of the classifier may result in significantly lower accuracy. Substantial research is needed to determine the essential features of datasets, on which the optimal choice of classifiers should be based.

Third, it can be seen from Table 5 that in all of the four datasets examined, classifiers using pandects seemed to dominate the others. More precisely, in each of these datasets the accuracy of one of the two pandect-based classifiers (strong prime or strong spanned) was always statistically one of the best. Actually, in three of the datasets, one of the two pandect-based classifiers was better than all the others, on the average. Even in the remaining dataset (**vot**), the accuracy of a pandect-based classifier, although ranked second on the average, was statistically indistinguishable from the accuracy of the classifier ranked first.

In spite of the fact that the pandect-based classifiers seem to offer a somewhat increased accuracy, theory and model-based classifiers should not be ruled out. The advantage of the theories and models, which usually consist of few patterns, resides in their "explanatory power" due to the ability of the human mind to comprehend them.

In order to evaluate the loss of accuracy when going from compendium- or pandect-based classifiers to theory- or model-based ones, we present in Table 6 for each of the datasets (*i*) the accuracy of the best classifier based on a pandect or a compendium, (*ii*) the accuracy of the best classifier based on a theory or a model, as well as (*iii*) their ratio and (*iv*) their difference; the figures in this table are based on the results given in Table 3. The table shows that the trade-off of performance for comprehensibility seems to be small, and hence acceptable.

Datasets	Best Accuracy		Ratio	Difference
	Compendia & Pandects	Theories & Models		
vot	96.33%	96.04%	0.997	0.29%
bcw	96.18%	95.60%	0.994	0.58%
hea	80.50%	79.96%	0.993	0.54%
bld	71.47%	70.27%	0.983	1.20%

Table 6.

In conclusion, it is reasonable to suggest that in the case when accuracy is the overriding consideration, a classifier based on one of the two pandects is perhaps advisable. However, when the overriding consideration is the comprehensibility of classification, then theory- and model-based classifiers should be viewed as a viable choice.

References

[1] G. Alexe, S. Alexe, S. Foldes, P. L. Hammer, and B. Simeone. Consensus Algorithms for the Generation of All Maximal Bicliques. *Discrete Applied Mathematics* (in print).

[2] G. Alexe and P.L. Hammer. Spanned Patterns for the Logical Analysis of Data. Rutgers University, *RUTCOR Research Report*, RRR 15-2002.

[3] S. Alexe, E. Blackstone, P. L. Hammer, H. Ishwaran, M. S. Lauer, and C. E. Pothier Snader. Coronary Risk Prediction by Logical Analysis of Data. *Annals of Operations Research* -- 2002 (in print).

[4] S. Alexe and P.L. Hammer. Accelerated Algorithm for Pattern Detection in Logical Analysis of Data. Rutgers University, *RUTCOR Research Report*, RRR 59-2002.

[5] A. Blake. *Canonical Expressions in Boolean Algebra*. Ph.D. Thesis, University of Chicago, 1937.

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

[7] E. Boros, P.L. Hammer, T. Ibaraki, A. Kogan, E. Mayoraz, and I. Muchnik. An Implementation of Logical Analysis of Data. *IEEE Transactions on Knowledge and Data Engineering*, **12**, No. 2 (2000), 292-306.

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

[9] P.L. Hammer. Partially Defined Boolean Functions and Cause-Effect Relationships. *International Conference on Multi-Attribute Decision Making Via OR-Based Expert Systems*, University of Passau, Passau, Germany (1986).

[10] P.L. Hammer, A. Kogan, B. Simeone, and S. Szedmak. Pareto-Optimal Patterns in Logical Analysis of Data. Rutgers University, *RUTCOR Research Report*, RRR 7-2001, and *Discrete Applied Mathematics* (in print).

[11] M.S. Lauer, S. Alexe, C.E. Pothier Snader, E.H. Blackstone, H. Ishwaran, and P.L. Hammer. Use of the "Logical Analysis of Data" Method for Assessing Long-Term Mortality Risk After Exercise Electrocardiography. *Circulation* **106** (2002), 685-690.

[12] W. Quine. A way to simplify truth functions. *American Mathematical Monthly*, **62** (1955), 627-631.

[13] Malgrange, Y. *Recherche des sous-matrices premières d'une matrice à coefficients binaires. Applications à certains problèmes de graphe*. In Deuxième Congrès de l'AFCALTI, October 1961, Gauthier-Villars, 1962, 231-242.