Annals of Operations Research, (in press)

Reformulation of the Support Set Selection Problem in the Logical Analysis of Data

Renato Bruni
Università di Roma "La Sapienza" - D.I.S.
Via M. Buonarroti 12, Roma, Italy, 00185.
E-mail: bruni@dis.uniroma1.it

Abstract

The paper is concerned with the problem of binary classification of data records, given an already classified training set of records. Among the various approaches to the problem, the methodology of the logical analysis of data (LAD) is considered. Such approach is based on discrete mathematics, with special emphasis on Boolean functions. With respect to the standard LAD procedure, enhancements based on probability considerations are presented. In particular, the problem of the selection of the optimal support set is formulated as a weighted set covering problem. Testable statistical hypothesis are used. Accuracy of the modified LAD procedure is compared to that of the standard LAD procedure on datasets of the UCI repository. Encouraging results are obtained and discussed.

Keywords: Classification; Data mining; Logical analysis of data; Massive data sets; Set covering.

1 Introduction

Given a set of *data* which are already grouped into *classes*, the problem of predicting whose class each new data belongs to is often referred to as *classification* problem. The first set of data is generally called *training set*, while the second one *test set* (see e.g. [16]). Classification problems are of fundamental significance in the fields of data analysis, data mining, etc., and are moreover able to represent several other relevant practical problems. As customary for structured information, data are organized into conceptual units called *records*, or *observations*, or even *points* when they are considered within some representation space. Each record has the formal structure of a set of *fields*, or *attributes*.

Giving a *value* to each field, a record instance, or, simply, a record, is obtained [22].

Various approaches to the classification problem have been proposed, based on different models and techniques (see for references [18, 15, 16]). One very effective methodology is constituted by the *logical analysis of data* (LAD), developed since the late 80's by Hammer *et al.* [8, 4, 5, 14, 1]. The mathematical foundation of LAD is in discrete mathematics, with a special emphasis on the theory of Boolean functions. More precisely, LAD methodology uses only binary variables, hence all data should be encoded into binary form by means of a process called *binarization*. This is obtained by using the training set for computing a set of values for each field. Such values are called *cut-points* in the case of numerical fields. Some of such values (constituting a *support set*) are then selected for performing the above binarization and for generating logical *rules*, or *patterns*. This is called *support set selection problem*, and is clearly decisive for the rest of the procedure. Patterns are then generated and used to build a *theory* for the classification of the test set. An advantage of such approach is that theories constitute also a (generally understandable) *compact description* of the data. As a general requirement, instances from the training set should have the same attributes and the same nature than those of the test set. No further assumptions are made on the data-set.

We propose here an original enhancement to the LAD methodology, namely a criterion for evaluating the quality of each cut-point for numerical fields and of each binary attribute for categorical fields. Such quality value is computed on the basis of information directly extractable from the training set, and is taken into account for improving the selection of the support set. Without *a priori* assumptions on the meaning of the data-set, except that it represents some real-world phenomenon (either physical or sociological or economical, etc.), we carry out a general statistical evaluation, and specialize it to the cases of numerical fields having *normal* (Gaussian) distribution or *binomial* (Bernoulli) distribution [12]. The support set selection is therefore modeled as a *weighted set covering* problem [19, 23]. In a related work [6], Boros *et al.* consider the problem of finding essential attributes in binary data, which again reduces to finding a small support set with a good separation power. They give alternative formulations of such problem and propose three types of heuristic algorithm for solving them. An analysis of the smallest support set selection problem within the framework of the probably approximately correct learning theory, together with algorithms for its solution, is also developed in [2].

Notation and the basic LAD procedure, with special attention to the support set selection aspects, is explained in Section 2. We refer here mainly to the "standard" procedure, as described in [5], although several variants of such procedure have been investigated in the literature [14, 1]. Motivations and criteria for evaluating the quality of cut-points are discussed in Section 3. In particular, we derive procedures for dealing with cut-points on continuous fields having normal distribution, on discrete fields having binomial distribution, or on general numerical fields having unknown distribution. This latter approach is used also for qualitative, or categorical, fields. The support set selection problem is then modeled as weighted set covering problem in Section 4. The remaining part of the LAD procedure is afterwards applied. Results are compared to those of the standard LAD procedure on datasets of the UCI repository [3], as shown in Section 5. Advantages of the proposed procedure are discussed in Section 6.

2 The LAD Methodology

A set of records S is given, already partitioned into the set of positive records S^+ and the set of negative records S^-. S is called *training set* and constitutes our source of information for performing the classification of other unseen (or new) records. The structure of records, called *record schema* R, consists in a set of fields f_i, $i = 1 \dots m$. A *record instance* r consists in a set of values v_i, one for each field of the schema. A positive record instance is denoted by r^+, a negative one by r^-.

$$R = \{f_1, \dots, f_m\} \qquad r = \{v_1, \dots, v_m\}$$

Example 2.1. For records representing persons, fields are for instance `age` or `marital status`, and corresponding examples of values can be `18` or `single`.

For each field f_i, $i = 1 \dots m$, its *domain* D_i is the set of every possible value for that field. Fields are essentially of two types: *quantitative*, or *numerical*, and *qualitative*, or *categorical*. A quantitative field is a field whose values are numbers, either continuous or discrete, or at least values having a direct and unambiguous correspondence with numbers, hence mathematical operators can be defined on its domain. A qualitative field simply requires its domain to be a discrete set with finite number of elements.

In order to use the LAD methodology, all fields should be encoded into binary form. Such process is called *binarization*. By doing so, each (non-binary) field f_i corresponds to a set of binary *attributes* a_i^j, with $j = 1 \dots n_i$. Hence, the term "attribute" is not used here as a synonym of "field". A binarized record scheme R_b is therefore a set of binary attributes a_i^j, and a binarized record instance r_b is a set of binary values $b_i^j \in \{0, 1\}$.

$$R_b = \{a_1^1, \dots, a_1^{n_1}, \dots, a_m^1, \dots, a_m^{n_m}\} \qquad r_b = \{b_1^1, \dots, b_1^{n_1}, \dots, b_m^1, \dots, b_m^{n_m}\}$$

For each qualitative fields f_i, all its values are simply encoded by means of a suitable number of binary attributes a_i^j. For each numerical field f_i, on the contrary, a set of cut-points $\alpha_i^j \in \mathbb{R}$ is introduced. In particular, for each couple of values v_i' and v_i'' (supposing w.l.o.g. $v_i' < v_i''$) respectively belonging to a positive and a negative record $v_i' \in r^+ \in S^+$ and $v_i'' \in r^- \in S^-$, and such that not other record $r \in S$ has a value v_i''' between them $v_i' < v_i''' < v_i''$, we introduce a cut-point α_i^j between them.

$$\alpha_i^j = (v_i' + v_i'')/2$$

Note that α_i^j is not required to belong to D_i, but only required to be comparable, by means of $\geq$ and $<$, to all values $v_i \in D_i$.

Example 2.2. Consider the following training set of records representing persons having fields `weight` (in Kg.) and `height` (in cm.), and a positive classifications meaning "to be a professional basketball player".

	`weight`	`height`	`pro.bask.pl.?`
S^+	90	195	yes
	100	205	yes
	75	180	yes
S^-	105	190	no
	70	175	no

Table 1: Training set for Example 2.2.

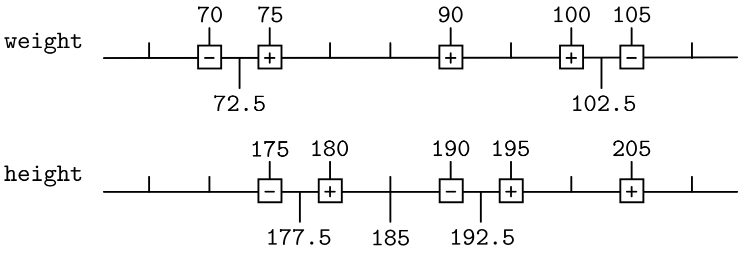

Figure 1: Cut points obtainable from the training set of Table 1.

For each attribute, values belonging to positive (respectively negative) records are represented with a framed + (resp. −). Cut-points obtainable for the above training set are $\alpha^1_{\texttt{weight}}$=`72.5`, $\alpha^2_{\texttt{weight}}$=`102.5`, $\alpha^1_{\texttt{height}}$=`177.5`, $\alpha^2_{\texttt{height}}$=`185`, $\alpha^3_{\texttt{height}}$=`192.5`. Corresponding binary attributes obtainable are $a^1_{\texttt{weight}}$, meaning: `weight` $\geq$ `72.5` Kg., $a^2_{\texttt{weight}}$, meaning: `weight` $\geq$ `102.5` Kg., $a^1_{\texttt{height}}$, meaning: `height` $\geq$ `177.5` cm., $a^2_{\texttt{height}}$, meaning: `height` $\geq$ `185` cm., etc.

Cut-points α_i^j are used for converting each field f_i into its corresponding binary attributes a_i^j, called *level variables.* The values b_i^j of such binary attributes are

$$b_i^j = \begin{cases} 1 & \text{if} \quad v_i \geq \alpha_i^j \\ 0 & \text{if} \quad v_i < \alpha_i^j \end{cases}$$

A set of binary attributes $\{a_i^j\}$ used for encoding the dataset S is a *support set* U. A support set is exactly separating if no pair of positive and negative records have the same binary encoding. Throughout the rest of the paper we are interested in support sets being exactly separating. Clearly, a single dataset admits several possible exactly separating support sets. Since the number of cut-points obtainable in practical problems is often very large, and since many of them may be not needed to explain the phenomenon, we are interested in selecting a small (or even the smallest) exactly separating support set, see also [5, 6]. By using a binary variable x_i^j for each a_i^j, such that $x_i^j = 1$ if a_i^j is retained in the support set, $x_i^j = 0$ otherwise, the following *set covering* problem should be solved. For every pair of positive and negative record r^+, r^- we define $I(r_b^+, r_b^-)$ to be the set of couples of indices (i, j) where the binary representations of r^+ and r^- differ, except, under special conditions [5], for the indices that involve monotone values.

$$\begin{aligned}
&\min \sum_{i=1}^{m} \sum_{j=1}^{n_i} x_i^j \\
&\text{s.t.} \sum_{(i,j)\in I(r_b^+, r_b^-)} x_i^j \geq 1 \qquad \forall I(r_b^+, r_b^-),\ r^+ \in S^+,\ r^- \in S^- \\
&\qquad x_i^j \in \{0, 1\}
\end{aligned}$$

Note that such selection of binary variables does not have the aim of improving the classification power, and actually “the smaller the chosen support set, the less information we keep, and, therefore, the less classification power we may have” [5]. Instead, it is necessary for reducing the computational complexity of the remaining part of the procedure, which may otherwise become impracticable. Indeed, a non-optimal solution to such problem would not necessarily worsen the classification power [5, 6]. Since different support sets correspond to different alternative binarizations, hence to actually different binarized record, the *support set selection* constitutes a key point.

Example 2.3. Continuing example 2.2, by solving to optimality the mentioned set covering problem we have the alternative support sets $U_1 = \{a^2_{\texttt{weight}}, a^1_{\texttt{height}}\}$ and $U_2 = \{a^1_{\texttt{weight}}\, a^2_{\texttt{weight}}\}$. An approximated solution would moreover be $U_3 = \{a^1_{\texttt{weight}}, a^2_{\texttt{weight}}, a^1_{\texttt{height}}, \}$. The corresponding alternative binarizations are:

	U_1		U_2		U_3		
	$b^2_{\texttt{weight}}$	$b^1_{\texttt{height}}$	$b^1_{\texttt{weight}}$	$b^2_{\texttt{weight}}$	$b^1_{\texttt{weight}}$	$b^2_{\texttt{weight}}$	$b^1_{\texttt{height}}$
S^+	0	1	1	0	1	0	1
	0	1	1	0	1	0	1
	0	1	1	0	1	0	1
S^-	1	1	1	1	1	1	1
	0	0	0	0	0	0	0

Table 2: Alternative binarizations obtainable from different support sets.

The selected support set U is then used to create patterns. A *pattern* P is a conjunction ($\wedge$) of literals, which are binary attributes $a_i^j \in U$ or negated binary attributes $\neg a_i^j$. A pattern P *covers* a record r if the set of values $r_b = \{b_i^j\}$

for the binary attributes $\{a_i^j\}$ makes $P = 1$. A *positive* pattern P^+ is a pattern covering at least one positive record r^+ but no negative ones. A negative pattern P^- is defined symmetrically. Patterns admit an interpretation as rules governing the studied phenomenon. Positive (respectively negative) patterns can be generated by means of top-down (i.e. by removing literals from pattern describing single positive (resp. negative) record until the pattern remains positive (resp. negative)), bottom-up (i.e. adding one by one literals until obtaining a positive (resp. negative) pattern), or hybrid procedures (i.e. bottom-up until a certain degree, then top-down using the positive (resp. negative) records not yet covered).

A set of patterns should be selected among the generated ones in order to form a theory. A *positive* theory T^+ is a disjunction ($\vee$) of patterns covering all positive records r^+ and (by construction) no negative record r^-. A negative theory T^- is defined symmetrically. Since the number of patterns that can be generated may be too large, pattern selection can be performed by solving another set covering problem (see [5, 14]), whose solution produces the set of the indices H^+ of selected positive patterns and that of the indices H^- of selected negative patterns. The obtained positive and negative theories are therefore

$$T^+ = \bigvee_{h \in H^+} P_h \qquad T^- = \bigvee_{h \in H^-} P_h$$

Weights $u_h^+ \geq 0$ and $u_h^- \leq 0$ are now assigned to all patterns in H^+ and H^-, by using several criteria [5]. Finally, each new record r is classified according to the positive or negative value of the following weighted sum, called *discriminant*, where $P(r) = 1$ if pattern P covers r, 0 otherwise (see also [5]).

$$\Delta(r) = \sum_{h \in H^+} u_h^+ P_h(r) + \sum_{h \in H^-} u_h^- P_h(r)$$

Example 2.4. By continuing example 2.3, a positive pattern obtained using the support set U_1 is $\neg a^2_{\texttt{weight}} \wedge a^1_{\texttt{height}}$. Another pattern obtained using support set U_3, is $a^1_{\texttt{weight}} \wedge \neg a^2_{\texttt{weight}} \wedge a^1_{\texttt{height}}$. Note that the latter one appears to be even more appropriate, since it means "one is a professional basketball player if has a medium weight (`weight` $\geq$ `72.5` Kg. and `weight` $<$ `102.5` Kg.) and height above a certain value (`height` $\geq$ `177.5` cm.)".

3 Evaluation of Binary Attributes

We noticed that, in the selection of the support set, we may loose some *useful* attribute. We therefore would like to evaluate the *usefulness*, or the *quality* of each attribute, before proceeding to such selection. Let us start with attributes on numerical fields, hence with the corresponding cut-points. We try to evaluate how good cut-point α_i^j *behaves* on field f_i. In the following Figure 2, we give three examples of fields (a,b,c). In each case, we draw "qualitative" distributions densities[1] of a consistent number of positive and negative records' values in the area above the line, and report a smaller sample of positive and negative records having the above distributions on the line. Very intuitively, cut-points obtainable in case a) are the worst ones, while the cut-point of case c) is the best one. Moreover, the various cut-points obtained in case b) do not appear to have all the same utility. We now try to formalize this. Given a single cut-point α_i^j and a record r, denote as $+$ (respectively $-$) the fact that r is actually positive (resp. negative), and denote by $\textit{class} + (\alpha_i^j)$ (resp. $\textit{class} - (\alpha_i^j)$) the fact that r is classified as positive (resp. negative) by α_i^j, i.e. stays on the positive (resp. negative) side of α_i^j.

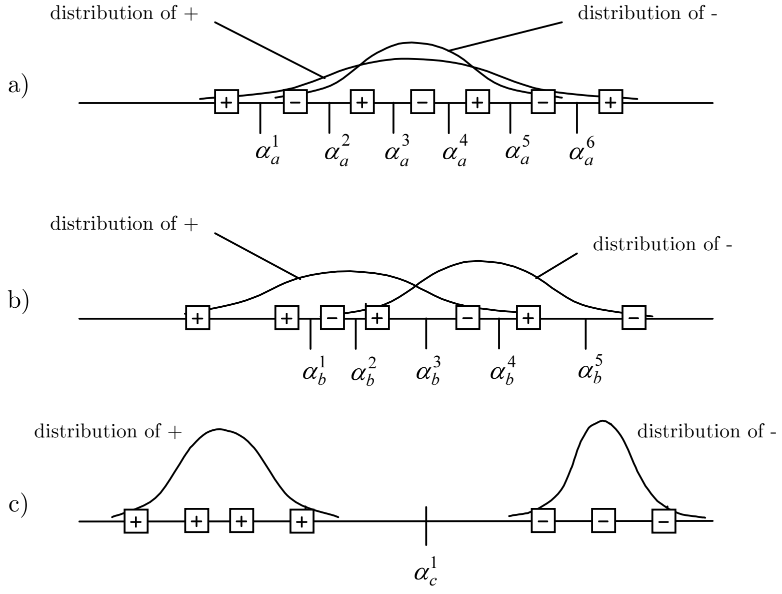

Figure 2: Examples of cut points in different conditions.

Different parameters could be considered for evaluating the quality of each cut-point α_i^j. We evaluate α_i^j on the basis of how it behaves on the training set S, hence how it divides S^+ from S^-, *even if* the real classification step will be conducted by using patterns, as described in previous section.

When classifying a generic set of records N, let A_+ be the set of the records which are $\textit{class}+(\alpha_i^j)$, and A_- be the set of records which are $\textit{class}-(\alpha_i^j)$. Denote instead by N^+ and N^- the (unknown) real positive and negative sets. Errors occur when a negative record is classified as positive, and vice versa. The first kind of errors, called *false positive* errors, are $N_- \cap A_+$. The second kind of errors, called *false negative* errors, are $N_+ \cap A_-$. The representation given in the following Table 3, called *confusion matrix* (see e.g. [16]), helps visualizing

[1] By distribution density we mean the function whose integral over any interval is proportional to the number of points in that interval.

the accuracy of our classification.

		Actual	
		+	−
Classified by α_i^j	+	$N_+ \cap A_+$	$N_- \cap A_+$
	−	$N_+ \cap A_-$	$N_- \cap A_-$

Table 3: Confusion matrix.

The described support set selection problem is a non-trivial decision problem. In order to solve it, it would be convenient to formulate it as a binary linear programming problem. Hence, we would like to obtain for each binary attribute a quality evaluation such that the overall quality value of a set of binary attributes results the sum of the individual quality values. A parameter often used for similar evaluations is the *odds*.

The odds (defined as the number of events divided by the number of non-events) of giving a record a correct positive classification by using only cut point α_i^j is

$$o^+(\alpha_i^j) = \frac{Pr(+ \cap \textit{class} + (\alpha_i^j))}{Pr(- \cap \textit{class} + (\alpha_i^j))}$$

while the odds of giving a correct negative classification using only α_i^j is

$$o^-(\alpha_i^j) = \frac{Pr(- \cap \textit{class} - (\alpha_i^j))}{Pr(+ \cap \textit{class} - (\alpha_i^j))}$$

Clearly, $o^+(\alpha_i^j) \in [0, +\infty)$ and $o^-(\alpha_i^j) \in [0, +\infty)$. The higher the value, the better positive (resp. negative) classification α_i^j provides. In order to have a complete evaluation of α_i^j, we consider the odds product $o^+(\alpha_i^j) \times o^-(\alpha_i^j) \in [0, +\infty)$. Moreover, rather than the numerical value of such evaluation, it is important that the values computed for the different cut-points are fairly comparable. We therefore sum 1 to such odds product, obtaining so a value in $[1, +\infty)$.

$$1 + \frac{Pr(+ \cap \textit{class} + (\alpha_i^j))}{Pr(- \cap \textit{class} + (\alpha_i^j))} \cdot \frac{Pr(- \cap \textit{class} - (\alpha_i^j))}{Pr(+ \cap \textit{class} - (\alpha_i^j))}$$

Denote now by A the set of couples of indices (i, j) of a generic set of cut-points: $\{\alpha_i^j\ : (i, j) \in A\}$. The overall *usefulness* of using such set of cut-points can now be related to the product of the individual terms, hence we have

$$\prod_{(i,j)\in A} \left[1 + \frac{Pr(+ \cap \textit{class} + (\alpha_i^j))}{Pr(- \cap \textit{class} + (\alpha_i^j))} \cdot \frac{Pr(- \cap \textit{class} - (\alpha_i^j))}{Pr(+ \cap \textit{class} - (\alpha_i^j))}\right]$$

As noted above, more than the numerical value, we are interested in fairly comparable values of such evaluation. Therefore, we apply a scale conversion and take the logarithm of the above value. This allows to convert it in a sum.

$$\ln \prod_{(i,j)\in A} \left[1 + \frac{Pr(+ \cap \textit{ class } + (\alpha_i^j))}{Pr(- \cap \textit{ class } + (\alpha_i^j))} \cdot \frac{Pr(- \cap \textit{ class } - (\alpha_i^j))}{Pr(+ \cap \textit{ class } - (\alpha_i^j))} \right] =$$

$$= \sum_{(i,j)\in A} \ln \left[1 + \frac{Pr(+ \cap \textit{ class } + (\alpha_i^j))}{Pr(- \cap \textit{ class } + (\alpha_i^j))} \cdot \frac{Pr(- \cap \textit{ class } - (\alpha_i^j))}{Pr(+ \cap \textit{ class } - (\alpha_i^j))} \right]$$

The quality q_i^j value of a single cut-point α_i^j can now be evaluated as

$$q_i^j = \ln \left[1 + \frac{Pr(+ \cap \textit{ class } + (\alpha_i^j))}{Pr(- \cap \textit{ class } + (\alpha_i^j))} \cdot \frac{Pr(- \cap \textit{ class } - (\alpha_i^j))}{Pr(+ \cap \textit{ class } - (\alpha_i^j))} \right]$$

Clearly, $q_i^j \in [0, +\infty)$. By definition of probability, it can be computed as

$$q_i^j = \ln \left[1 + \frac{\frac{|N_+ \cap A_+|}{|N^+|}}{\frac{|N_- \cap A_+|}{|N^+|}} \cdot \frac{\frac{|N_- \cap A_-|}{|N^-|}}{\frac{|N_+ \cap A_-|}{|N^-|}} \right] = \ln \left[1 + \frac{|N_+ \cap A_+|}{|N_- \cap A_+|} \cdot \frac{|N_- \cap A_-|}{|N_+ \cap A_-|} \right]$$

(Were $| \cdot |$ denotes the cardinality of a set.) However, the above quality evaluation q_i^j for α_i^j could only be computed after knowing the correct classification $\{N^+, N^-\}$ of the dataset N. We would obviously prefer a quality evaluation that is computable *a priori*, that is by knowing the correct classification only for the training set S. We can do this in two different manners, one for fields having a known distribution, the other for fields having unknown distribution, as follows.

In the case of fields for which the hypothesis of a known distribution is satisfactory, their positive and negative density functions can be computed using the training set S. Therefore, the above quantities $|N_+ \cap A_+|$, etc. can be evaluated by using such density functions. There are also *tests* for verifying whether a set of data has a certain distribution (e.g. the χ^2 test) [12]. In particular, for any continuous-valued field f_i, we make the hypothesis of a *normal* (Gaussian) distribution. Such distribution is in fact the most common in nature and somehow describes the majority of continuous real-world values [12]. Denote now by m_{i+} (respectively by m_{i-}) the *mean value* that positive (resp. negative) records have for f_i, by σ_{i+} (resp. by σ_{i-}) the (population) *standard deviation* (defined as $\sqrt{\frac{\sum_{s\in S^+}(v_i^s - m_{i+})^2}{|S^+|}}$ (resp. $\sqrt{\frac{\sum_{s\in S^-}(v_i^s - m_{i-})^2}{|S^-|}}$)) of positive (resp. negative) records for f_i, and suppose w.l.o.g. that cut-point α_i^j represents a transition from $-$ to $+$. By computing the above parameters from the training set S, our evaluation of quality q_i^j becomes

$$q_i^j = \ln \left[1 + \frac{\int\limits_{\alpha_i^j}^{+\infty} \frac{1}{\sqrt{2\pi(\sigma_{i+})^2}} \, e^{-\frac{(t-m_{i+})^2}{2(\sigma_{i+})^2}} dt}{\int\limits_{\alpha_i^j}^{+\infty} \frac{1}{\sqrt{2\pi(\sigma_{i-})^2}} \, e^{-\frac{(t-m_{i-})^2}{2(\sigma_{i-})^2}} dt} \cdot \frac{\int\limits_{-\infty}^{\alpha_i^j} \frac{1}{\sqrt{2\pi(\sigma_{i-})^2}} \, e^{-\frac{(t-m_{i-})^2}{2(\sigma_{i-})^2}} dt}{\int\limits_{-\infty}^{\alpha_i^j} \frac{1}{\sqrt{2\pi(\sigma_{i+})^2}} \, e^{-\frac{(t-m_{i+})^2}{2(\sigma_{i+})^2}} dt} \right]$$

In case of a discrete-valued field f_i, on the contrary, we make the hypothesis of *binomial* (Bernoulli) distribution. This should in fact describe many discrete real-world quantities [12]. Moreover, such distribution is strongly related to the Poisson distribution, and both may be approximated by normal distribution when the number of possible values increases. Denote now by m_{i+} and M_{i+} (respectively by m_{i-} and M_{i-}) the *minimum* and the *maximum* value of positive (resp. negative) values of D_i (such values for the positive records may also coincide with those for the negative ones). Denote also by $n_{i+} = M_{i+} - m_{i+}$ (resp. by $n_{i-} = M_{i-} - m_{i-}$) the *number* of possible positive (resp. negative) values for f_i, and by p_+ (resp. p_-) the characteristic positive (resp. negative) *probability of success* (also called Bernoulli probability parameter, estimated as $|S^+|/n_{i+}$ (resp. $|S^-|/n_{i-}$)). Suppose, again, that cut-point α_i^j represents a transition from $-$ to $+$. By computing the above parameters from the training set S, our evaluation of quality q_i^j becomes in this case

$$q_i^j = \ln\left[1+\frac{\sum\limits_{t=\alpha_i^j-m_{i+}}^{n_{i+}} \binom{n_{i+}}{t}(p_{i+})^t(1- p_{i+})^{n_{i+}-t}}{\sum\limits_{t=\alpha_i^j-m_{i+}}^{n_{i+}} \binom{n_{i-}}{t}(p_{i-})^t(1- p_{i-})^{n_{i-}-t}} \cdot \frac{\sum\limits_{t=0}^{\alpha_i^j-m_{i-}-1} \binom{n_{i-}}{t}(p_{i-})^t(1- p_{i-})^{n_{i-}-t}}{\sum\limits_{t=0}^{\alpha_i^j-m_{i-}-1} \binom{n_{i+}}{t}(p_{i+})^t(1- p_{i+})^{n_{i+}-t}}\right]$$

On the other hand, in the case of fields having unknown distribution (for instance fields where one the above mentioned hypothesis are showed inapplicable by one of the available tests), the expression for q_i^j can be obtained by considering the above cardinalities of the sets. Given a cut point α_i^j, in fact, A_+ and A_- are clearly known (they respectively are the set of the records which are $\mathit{class}+(\alpha_i^j)$ and $\mathit{class}-(\alpha_i^j)$), and the training set S, whose classification is known, is to be used instead of the generic set N.

Finally, the quality of each attribute a_i^j over a numerical field f_i is that of its corresponding cut-point α_i^j, that is the defined q_i^j. The approach used for fields having unknown distribution (considering the training set S instead of N) is also applicable for evaluating attributes a_i^j over qualitative, or categorical, fields f_i.

4 Reformulation of the Support Set Selection Problem

Once the quality values for each attribute are computed, the exactly separating support set selection problem can be modeled as follows. We define the *uselessness* of an attribute as the reciprocal $1/q_i^j$ of the quality q_i^j. We clearly would like to minimize the weighted sum of the uselessness of the selected attributes while selecting at least an attribute for each of the above defined sets $I(r_b^+, r_b^-)$. Moreover, in order to reduce possible *overfitting* problems, we further penalize each attribute a_i^j of a field f_i such that: *i)* a_i^j contains a number ν of positive (resp. negative) records of the training set S smaller then or equal to a certain value $\bar{\nu}$, and *ii*) the adjacent attributes a_i^{j-1} and a_i^{j+1} over the same field f_i respectively contain a number μ_1 and μ_2 of negative (resp. positive) records greater than or equal to a certain value $\bar{\mu}$. Such penalization is obtained by summing to the above uselessness of a_i^j a *penalty* value $t_i^j(\nu, \mu_1, \mu_2)$.

We introduce, as usual, a binary variable x_i^j for each a_i^j, such that

$$x_i^j = \begin{cases} 1 & \text{if } a_i^j \text{ is retained in the support set} \\ 0 & \text{if } a_i^j \text{ is excluded from the support set} \end{cases}$$

Therefore, the following *weighted* set covering problem should be solved, where the weights $w_i^j = \frac{1}{q_i^j} + t_i^j(\nu, \mu_1, \mu_2)$ are non-negative numbers.

$$\begin{aligned}
&\min \sum_{i=1}^{m} \sum_{j=1}^{n_i} w_i^j \; x_i^j \\
&\text{s.t.} \sum_{(i,j)\in I(r_b^+, r_b^-)} x_i^j \geq 1 \qquad \forall I(r_b^+, r_b^-),\ r^+ \in S^+,\ r^- \in S^- \\
&\quad x_i^j \in \{0, 1\}
\end{aligned}$$

Such formulation takes now into account the individual qualities of the attributes. One may observe that this would discard attributes that have a poor isolated effect but may have important effect when combined with other attributes during the pattern generation step. However, such selection is necessary for the computational viability of the entire procedure, and the proposed approach aims at discarding the attributes that appear more suitable to be discarded.

Moreover, such weighted set covering formulation has strong computational advantages on a non-weighted one. Available solution algorithms are in fact considerably faster when the model variables receive different weight coefficients in the objective function. Depending on the size of the model and on available computational time, such weighted set covering problem may be either solved to optimality or by searching for an approximate solution. In the former case, it is guaranteed that the pattern generation step is performed by using a set of attributes U which is a minimal set for which no positive and negative records have the same binary encoding. In the latter case, if the approximate solution is feasible but non-optimal, it is not guaranteed that U is minimal, i.e. it may exist also a proper subset $U' \subset U$ such that no positive and negative records have the same binary encoding. This could have the effect of increasing the computational burden of the pattern generation step, but not of worsening the classification accuracy. If, on the contrary, the approximate solution is (slightly) infeasible, U is such that (few) positive and negative records have the same binary encoding. This could have the effect of accelerating the pattern generation step, but of decreasing the classification accuracy.

5 Implementation and Computational Results

The entire LAD methodology has been implemented in Java. Tests are carried out on a Pentium III 733MHz PC. The data-sets used for the experimentations are "Ionosphere", "Bupa Liver Disorders","Breast Cancer Wisconsin", and "Pima Indians Diabetes", from the UCI Repository of machine learning problems [3].

The first set, Ionosphere, is composed by 351 instances, each having 34 fields (plus the class). In particular, there are 32 real-valued fields and 2 binary ones. All 32 real-valued fields could be considered having normal distribution, one binary fields could be considered having binomial distribution, the other is always 0. They are "data collected by a radar system in Goose Bay, Labrador. The targets were free electrons in the ionosphere. Good radar returns are those showing evidence of some type of structure in the ionosphere. Bad returns are those that do not; their signals pass through the ionosphere.", from [3].

The second set, Bupa Liver Disorders, is composed by 345 instances, each having 6 fields (plus the class), all numerical and discrete. However, 4 of them have a number of possible values high enough, hence 4 field could be considered having normal distribution, while 2 could be considered having binomial distribution. "The first five fields are blood tests which are thought to be sensitive to liver disorders that might arise from excessive alcohol consumption. The last is the number of half-pint equivalents of alcoholic beverages drunk per day.", from [3], the class is presence or absence of liver disorders.

The third set, Breast Cancer Wisconsin, is composed by 699 instances. By eliminating those containing missing values, we obtained 683 instances, each having 9 fields (plus an identifier and the class), all numerical and discrete. All could be considered having binomial distribution. They represent data from the breast cancer databases of the University of Wisconsin Hospitals, Madison, Wisconsin. In particular, fields are the characteristics of the breast cancer, such like "Clump Thickness, Uniformity of Cell Size, etc., and the classification is either benign or malignant" [3].

The fourth set, Pima Indians Diabetes, is composed by 768 instances, each having 8 fields (plus the class). In particular, there are 2 real-valued fields and 6 integer ones. However, since 3 integer fields have a number of possible values high enough, 5 field could be considered having normal distribution, while 3 could be considered having binomial distribution. Fields are medical informations about "females patients of Pima Indian heritage living near Phoenix, Arizona, the class is whether the patient shows signs of diabetes" [3].

The quality values q_i^j are numerically approximated by using C functions [20]. Penalties $t_i^j(\nu, \mu_1, \mu_2)$ have been set to 1/10 of the average uselessness values $1/q_i^j$ of field f_i when $\nu \leq 1$ and $\mu_1, \mu_2 \geq 5$, to 0 otherwise.

Tests are conducted as follows. A certain number of record instances, representing respectively about 15%, 20%, 25%, 30% of the total, are randomly extracted from the data-set, and used as training set. The rest of the data-set constitutes the test-set. Such extraction is performed 10 times, and the results reported in the following tables are averaged on the 10 trials. The weighted set covering problems are solved both by means of ILOG Cplex [13] state-of-the-art implementation of the branch-and-cut procedure [19], and by means of a Lagrangean-based subgradient heuristic for set covering problems (see e.g. [10]). We therefore report percentages of correct classification on test set (Accur.) for:

- the standard LAD procedure solving the non-weighted set covering problems to optimality using branch-and-cut (LAD_I);
- the modified LAD procedure solving the weighted set covering problems to optimality using branch-and-cut (LAD_II);
- the modified LAD procedure solving the weighted set covering problems by finding a feasible sub-optimal solution using Lagrangean subgradient heuristic (LAD_III).

We also report computational times in seconds required by the whole procedure, specifying in parenthesis the percentage of time spent for solving the support set selection problem. A time limit of 3600 seconds (1h) was set for the whole procedure, when exceeded we report '-'.

Training	LAD_I		LAD_II		LAD_III	
Set	Accur.	Time	Accur.	Time	Accur.	Time
53/351	80.8%	480.8 (97%)	82.1%	18.2 (89%)	82.0%	180.2 (53%)
70/351	81.5%	562.3 (98%)	84.3%	20.0 (90%)	84.8%	222.0 (42%)
88/351	83.1%	357.7 (97%)	87.0%	129.0 (87%)	86.8%	3461.0 (20%)
115/351	-	-	90.6%	2163.0 (11%)	-	-

Table 4: Results on Ionosphere (average on 10 trials).

Training	LAD_I		LAD_II		LAD_III	
Set	Accur.	Time	Accur.	Time	Accur.	Time
52/345	58.6%	35.0 (90%)	62.3%	40.8 (95%)	62.5%	80.5 (79%)
69/345	59.5%	50.2 (94%)	63.9%	66.0 (93%)	64.0%	58.2 (94%)
86/345	60.2%	326.0 (90%)	65.3%	145.2 (90%)	65.1%	190.8 (16%)
110/345	61.2%	1886.4 (96%)	65.0%	430.0 (78%)	-	-

Table 5: Results on Bupa Liver Disorders (average on 10 trials).

Training	LAD_I		LAD_II		LAD_III	
Set	Accur.	Time	Accur.	Time	Accur.	Time
102/683	91.1%	7.5 (97%)	92.3%	9.9 (96%)	92.0%	14.2 (97%)
137/683	93.4%	10.0 (97%)	94.0%	15.8 (97%)	94.2%	15.8 (98%)
170/683	93.5%	37.9 (98%)	94.4%	20.0 (97%)	94.4%	480.0 (5%)
205/683	94.1%	409.0 (59%)	95.1%	107.5 (52%)	95.1%	1865.0 (3%)

Table 6: Results on Breast Cancer Wisconsin (average on 10 trials).

Training	LAD_I		LAD_II		LAD_III	
Set	Accur.	Time	Accur.	Time	Accur.	Time
115/768	63.3%	3550.0 (98%)	65.0%	230.1 (90%)	65.0%	2840.0 (10%)
154/768	-	-	68.2%	372.5 (92%)	68.5%	3605.0 (8%)
192/768	-	-	70.1%	1108.0 (28%)	-	-

Table 7: Results on Pima Indians Diabetes (average on 10 trials).

Test results are reported in Tables 4-7 and also plotted for comparison in Figure 3, limiting the representation to LAD_I and LAD_II in order to compare under the same conditions (problems solved to optimality).

As a general result, the effort invested in evaluating the quality of the various binary attributes returns a superior classification accuracy with respect to the standard procedure. Results are anyway of good quality, since, using much larger training set (very often 50% of the data-set or more), the best results presented in the literature on Ionosphere are between 90-95% (e.g. Smooth Support Vector Machine [17], C4.5 [21]), between 65-75% on Bupa Liver Disorders (e.g. Backpropagation Neural Networks [9]), around 90-95% on Breast Cancer Wisconsin (e.g. LMDT [7]), around 70-75% on Pima Indians Diabetes (e.g. ITI [24]). For detailed comparisons see also [11]. Note also that an important additional advantage of a logic-based procedure such as LAD is to provide patterns as understandable rules governing the analyzed phenomenon, whereas other methodologies cannot provide similar interpretations.

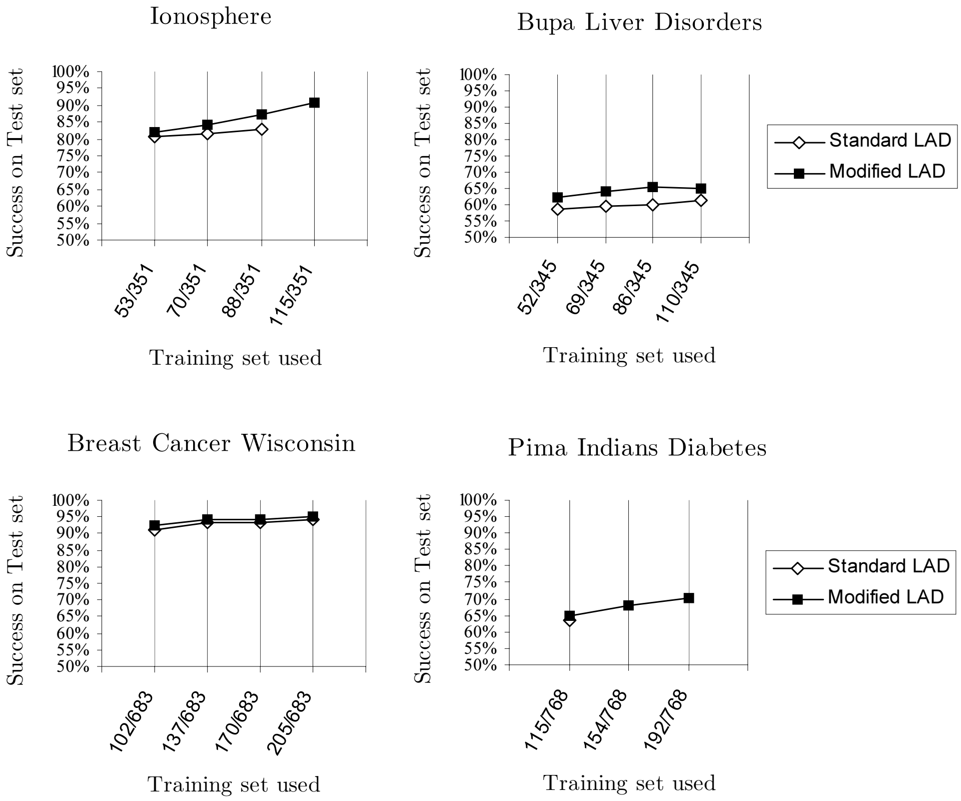

Figure 3: Classification accuracy for standard and modified LAD procedures using 10 cross validation.

From the computational point of view, on the other hand, it can be clearly observed, from Tables 4-7, that weighted set covering problems are solved in times which are much shorter than those needed for the corresponding non-weighted ones. Moreover, when the support set selection problem is not solved to optimality, hence the selected support set retains more binary attributes than it would be strictly needed for an exact separation, the accuracy sometimes slightly increases. However, time needed for the second part of the procedure increases substantially. Therefore, this latter approach appears useful only for very small dataset.

Finally, it can be noticed that the proposed approach is quite effective also when using very reduced training sets. In such case, indeed, a careful selection of the binary attributes to be included in the support set becomes more important. Such characteristic can be of interests in several practical problems were the availability of already classified records is scarce or costly.

6 Conclusions

Among the various approaches to the classification problem, the methodology of the logical analysis of data (LAD) is considered. Such procedure exhibits functional advantages on other techniques, given by the production of understandable and checkable Boolean theories on the analyzed data. Nevertheless, an aspect which is not completely satisfactory consists in the solution of the support set selection problem. Such operation does not increase accuracy but is necessary for computational viability. We propose here a technique for evaluating the quality of each attribute, among which the support set must be selected. Thus, a weighted set covering problem for the selection of the optimal support set is solved. Testable statistical hypothesis on the distributions of numerical fields can be used. Accuracy of the modified LAD procedure is compared to that of the standard LAD procedure on datasets of the UCI repository. The presented techniques are able to increase the classification accuracy. In particular, fairly good results can be achieved by using very reduced training sets. Such advantage can be of interests in several practical problems were the availability of already classified records is scarce. The proposed weighted set covering model has also strong computational advantages on a non-weighted one. This allows a sensible speed-up of the whole classification procedure. As a consequence, larger data sets can be considered.

Acknowledgments The author is grateful to Prof. G. Koch for useful discussions on the probabilistic issues, and to Dr. Z. Falcioni and Dr. A. Robibaro for their important contribution to the implementation work.

References

[1] G. Alexe, S. Alexe, P.L. Hammer, A. Kogan. Comprehensive vs. Comprehensible Classifiers in Logical Analysis of Data. RUTCOR Research Report, RRR 9-2002; DIMACS Technical Report 2002-49; *Annals of Operations Research* (in print).

[2] H. Almuallim and T.G. Dietterich. Learning Boolean Concepts in the Presence of many Irrelevant Features. *Artificial Intelligence* 69:1, 279-306, 1994.

[3] C. Blake and C.J. Merz. UCI Repository of machine learning databases. Irvine, CA: University of California, Department of Information and Computer Science, 1998. URL: `http://www.ics.uci.edu/∼mlearn/MLRepository.html`.

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

[5] E. Boros, P.L. Hammer, T. Ibaraki, A. Kogan, E. Mayoraz, I. Muchnik. An Implementation of Logical Analysis of Data. *IEEE Transactions on Knowledge and Data Engineering*, 12(2):292-306, 2000.

[6] E. Boros, T. Horiyama, T. Ibaraki, K. Makino, M. Yagiura. Finding Essential Attributes from Binary Data. RUTCOR Research Report, RRR 13-2000, *Annals of Mathematics and Artificial Intelligence* (to appear).

[7] C.E. Brodley and P.E. Utgoff. Multivariate decision trees. *Machine Learning*, 19, 45-77, 1995.

[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] Dept. of Internal Medicine, Electrical Engineering, and Computer Science, University of Nevada, Reno. *Nevprop3 User Manual* (Nevada backPropagation, Version 3), 1996.

[10] M.L. Fisher. The Lagrangian relaxation method for solving integer programming problems. *Management Science*, 27:1-18, 1981.

[11] P.W. Eklund. A Performance Survey of Public Domain Supervised Machine Learning Algorithms. KVO Technical Report 2002, The University of Queensland, *submitted*, 2002.

[12] M. Evans, N. Hastings, B. Peacock. *Statistical Distributions* (3rd edition). Wiley series in Probability and Statistics, New York, 2000.

[13] ILOG Cplex 8.0. *Reference Manual.* ILOG, 2002.

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

[15] D.J. Hand, H. Mannila, P. Smyth. *Principles of Data Mining.* MIT Press, London, 2001.

[16] T. Hastie, R. Tibshirani, J. Friedman. *The Elements of Statistical Learning.* Springer-Verlag, New York, Berlin, Heidelberg, 2002.

[17] Y.J. Lee and O.L. Mangasarian. SSVM: A smooth support vector machine. *Computational Optimization and Applications* 20(1), 2001.

[18] T.M. Mitchell. *Machine Learning.* McGraw-Hill, Singapore, 1997.

[19] G.L. Nemhauser and L.A. Wolsey. *Integer and Combinatorial Optimization.* J. Wiley, New York, 1988.

[20] W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery. *Numerical Recipes in C: The Art of Scientific Computing*, Second Edition. Cambridge University Press, 1992.

[21] J.R. Quinlan. *C4.5: Programs for Machine Learning.* Morgan Kaufmann, San Mateo, CA, 1993.

[22] R. Ramakrishnan and J. Gehrke. *Database Management System.* McGraw Hill, 2000.

[23] A. Schrijver. *Theory of Linear and Integer Programming.* Wiley, New York, 1986.

[24] P.E. Utgoff, N.C. Berkman, J.A. Clouse. Decision Tree Induction Based on Efficient Tree Restructuring. *Machine Learning* 29:1, 5-44, 1997.