Effective Classification using a small Training Set based on Discretization and Statistical Analysis

Aishwarya B. Jadhav  
PG Student  
Comp Engg, PVPIT  
SPPU

V.S.Nandedkar  
Assistant Professor  
Comp Engg, PVPIT  
SPPU

ABSTRACT

In this paper, we depict the work with the issue of creating a quick and precise information order, gaining from little arrangement of records. The proposed methodology depends on the system of the alleged Logical Analysis of Information (LAD), however advanced with data got from measurable contemplations on the information. Various discrete streamlining issues are illuminated in the diverse strifes of the system, yet their computational interest can be controlled. The precision of the proposed methodology is contrasted with that of the standard LAD calculation, of Support Vector Machines and of Label Propagation calculation on openly accessible datasets of the UCI storehouse.

Keywords
Classification Algorithms, Data Mining, Machine Learning, Discrete Mathematics, Optimization.

1. INTRODUCTION

Set of data are grouped into classes, the problem of predicting which class new data should receive is called Classification problem. There are many approaches to do so. Like Neural Networks, SupportVector Machines, k-Nearest Neighbors, Bayesian approaches, Decision Trees, Logistic regression, Boolean approaches. One approach that is generally considered quite effective for many practical applications is Support Vector Machines (SVM).[1]The bigger is the preparation set, the more data it contains, the more exact the scholarly classifier can be create. Shockingly, in numerous critical applications, marked information are troublesome or costly to get. On inverse, unlabeled information might be moderately simple to gather. In this manner, systems have been created for using so as to enhance a characterization likewise a lot of unlabeled information, that is called approval methodology. Another real system in semi supervised learning is LabelSpread (LP).In any case, no single calculation is at present capable to provide the best execution on all datasets, and this is by accounts unavoidable. In this way, systems taking into account the conglomeration of an arrangement of various (and ideally corresponding) classifiers have been explored.[2] Those procedures create numerous frail learners and join their yields all together to get a grouping that is both exact and hearty. Those frail learners might be founded on a few arrangement approaches.

On the other side Boolean approach classification by using LAD method is useful to make system better learning from examples, as humans learns.

2. LITERATURE SURVEY

Grouping is the information mining undertaking of anticipating the estimation of an all out variable (class or target). Boolean way to deal with grouping is the Logical Analysis of Data(LAD). It is roused by the mental procedures that an individual applies when gaining from illustrations. In this methodology, information ought to be encoded into twofold shape by method for a discretization process called binarization. The preparation set for registering particular qualities for each field, called cut-focuses on account of numerical fields, that split every field into paired properties. They chose parallel properties constitute a support set, and are joined for producing intelligent rules called designs. Examples are utilized to characterize each unclassified record, on the premise of the indication of a weighted total of the examples enacted by that record.

2.1 Classifying With The Lad Methodology

The structure of records, called record scheme R, consists of a set of fields fi, with i= 1 . . . m. A record instance r, also simply called record, consists of a set of values vi, one for each field. A record r is classified if it is assigned to an element of a set of possible classes C. A positive record instance r is denoted by r+, a negative one by r−. A training set S. S+ the set of its positive erecords and by S− the set of its negative ones. Sets S+ and S− constitute our source of information. A set of records used for evaluating the performance of the learned classifier is called test set T. A positive training record is denoted by s+, a negative one by s−. A positive test record is denoted by t+, a negative one by t−. LAD methodology begins with encoding all fields into binary form. This process, called binarization, [11] converts each (non-binary) field fi into a set of binary attributes aij with j = 1 . . . ni. The total number of binary attributes is n = Σj=1..ni. Note that the term “attribute” is not used here as a synonym for “field”.

A binary record scheme Rh is therefore a set of binary attributes aij, and a binary record instance ri is a set of binary values bi ∈ {0, 1} for those attributes.

\[ R_h = \{ a_{i1}, \ldots, a_{i1}, \ldots, a_{in}, \ldots, a_{in} \} \]

\[ r_i = \{ a_{i1}, \ldots, a_{i1}, \ldots, a_{in}, \ldots, a_{in} \} \]

For each qualitative fields fi, all values can simply be encoded by means of a logarithmic number of binary attributes aij, so that ni binary attributes can binarize a quantitative field having up to 2ni different values. For each numerical field fi , on the contrary, we introduce ni thresholds called cut-points

\[ a_{ij}^1, \ldots, a_{ij}^ni \in \mathbb{R}, \]  

and the binarization of a value vi is obtained by considering whether vi lies above or below each aij . Cut-points aij should be set at values representing some kind of watershed for the analyzed phenomenon. Generally, aij are placed in the middle of specific couples of data values v′ and v′′ :

\[ a_{ij} = \frac{(v′ + v''')}{2}. \]
This can be done for each couple vi and v′i belonging to records from opposite classes that are adjacent on fi. Cut-points a′i are then used for binarizing each numerical field fi into the binary attributes ai (also called level variables). The values bi of such ai are

\[ b_i' = \begin{cases} 1 & \text{if } v_i > a_i' \\ 0 & \text{if } v_i = a_i' \end{cases} \]

3. EXISTING SYSTEM

In existing framework there are numerous methodologies for order issue, it incorporates Neural Networks, Support Vector Machines, k-Nearest Neighbors, Bayesian approaches, Decision Trees, Logistic relapse. Every methodology is particular fit for particular order, yet one for the most part consider is Support Vector Machine(SVM). SVM depend on finding an isolating hyperplane that boosts the edge between the great preparing information of inverse classes. Another significant structure in semi-managed learning strategies is Label Spread (LP). This method works by building likeness diagram overall record.

3.1 Disadvantages of Existing System.

- Each approach has several variants and algorithms, specific approach may better fit for specific classification.
- Large data contain large information so the more accurate the learned classifier will be, but labeled data are difficult or expensive to obtain.
- No single algorithm is currently able to provide the best performance on all datasets, and this seems to be inevitable.

4. PROPOSED SYSTEM

In this paper, we propose the improvements to the LAD approach. To begin with, assessing the nature of every cut-point for numerical fields and of every parallel characteristic for straight out fields. In a related work, consider the issue of discovering fundamental characteristics in parallel information, which again lessens to finding a little backing set with a decent division power. The grouping of the test set is most certainly not given here just on the premise of the indication of the weighted aggregate of actuated examples, yet by looking at that weighted aggregate to a suitable characterization limit. Design weights and arrangement edge are truth be told parameters for the order system.

4.1 Advantages of Proposed System

- Small training sets will provide good degree of accuracy on variety of practical applications.
- The proposed system will enhance the classification accuracy and reduces the computational time with respect to the LAD methodology.

4.2 Evaluation Of Binary Attributes

We remarked that selecting a small support set is computationally necessary, but that excluding attributes means losing information. Therefore, we propose to evaluate the quality (the separating power) of each attribute and to perform such a selection taking into account this evaluation. In following example in numeric field (a,b,c), we draw (in the area above the horizontal line) “qualitative” distributions densities of a large number of values from positive and negative records, and report (on the same line) a smaller sample of those values. a) are the worst ones (they do not appear very useful for separating the two classes), while the cut-point of case. c) is the best one (it has a good “separating power”). Moreover, the different cut-points of case b) do not have the same quality.

To estimate this, we analyze how ai j divides the two classes, even if the real classification step will use patterns. Different estimators could of course be designed, however results show that the proposed technique is able to improve accuracy with respect to the standard LAD procedure.

Since the described support set selection problem is a non-trivial decision problem, it seems reasonable to model it as a binary linear programming problem. For doing so, we need to use a criterion for evaluating the quality of each binary attribute such that the overall quality value of a set of binary attributes can be given by the sum of their individual quality values. We obtain this as follows.

\[ o^{+}(\alpha_i^j) = \frac{Pr(\text{+ class } | \alpha_i^j)}{Pr(\text{- class } | \alpha_i^j)} \]

A similar measure can evaluate the accuracy of the negative classification obtained from ai.

\[ o^{-}(\alpha_i^j) = \frac{Pr(\text{- class } | \alpha_i^j)}{Pr(\text{+ class } | \alpha_i^j)} \]

In conclusion, the quality qi j of a single cut-point ai j can be evaluated as follows (so that the quality of a set of cut-points results in the sum of their individual quality values).

\[ q_i^j = \ln \left( \frac{1 + Pr(\text{+ class } | \alpha_i^j)}{1 + Pr(\text{- class } | \alpha_i^j)} \right) \]

Clearly, \( q_i^j \in [0, +\infty) \). Computing the above probabilities by counting instances (and denoting by \( | \cdot | \) the cardinality of a set), we have:

\[ q_i^j = \ln \left( \frac{|N_+ \cap A_i|}{|N_+|} \cdot \frac{|N_- \cap A_i|}{|N_-|} \right) = \ln \left( \frac{|N_+ \cap A_i|}{|N_+|} \cdot \frac{|N_- \cap A_i|}{|N_-|} \right) \]
In particular, for any continuous-valued field \( f_i \), we make the hypothesis of a normal (Gaussian) distribution. Such distribution can indeed model the majority of real-world values, as a consequence of the central limit theorem.[5] Denote now by \( m_i^+ \) the mean value that positive records have for \( f_i \) and by \( a_i^- \) their (population) standard deviation (defined as

\[
\sqrt{\sum_{s \in \mathcal{S}} \frac{(V_i - M_i)^2}{|\mathcal{S}|}}
\]

denote by \( m_i^- \) and \( a_i^- \) the same quantities for the negative records, and suppose w.l.o.g. that cut-point \( a_i^+ \) represents a transition from \(-\) to \(+\). By computing the above parameters from the training set \( \mathcal{S} \), our evaluation of quality \( q_i^+ \) becomes:

\[
q_i^+ = \ln \left( 1 + \frac{1}{a_i^+} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}(a_i^+)^2} e^{-\frac{(t-m_i^+)^2}{2(a_i^+)^2}} \, dt \right) \]

More precisely, each time an attribute from \( f_i \) is selected, we put \( q_i^+ = q_i^- / 2 \) for every still unselected attributes of \( f_i \). Finally, for fields having a considerable overlapping between the two classes, cut-points cannot be generated when inverting the class, because almost every region of the field contains both classes.

### 4.3 Reformulations of the Support Set Selection Problem

We would like to minimize a weighted sum (and not only the number) of selected attributes, where the weights are the reciprocal \( 1/q_i^+ \) of the quality \( q_i^+ \), while selecting at least an attribute for each of the above defined sets \( \mathcal{R}_i^+ \), \( \mathcal{R}_i^- \). When no specific evaluations can be done, those sizes could be set all at \( 1 \). Moreover, we can establish a maximum affordable computational burden \( b \), for instance on the basis of the time available for performing the classification, or of the available computing hardware, etc. Note that such requirement may be independent from the minimum size of an exactly separating support set: the available resources are limited, and, if they allow obtaining an exactly separating support set, the better, but this cannot be imposed. By using the same binary variables \( x_i^+ \), the support set selection problem can now be modeled as binary knapsack problem:

\[
\max \sum_{i=1}^{m} \sum_{j=1}^{n_i} q_i^+ x_i^+ \\
\text{s.t.} \sum_{i=1}^{m} \sum_{j=1}^{n_i} s_i^- x_i^+ \leq b \\
x_i^+ \in \{0, 1\}
\]

In this case, attributes can be selected sequentially, and the weights be modified after each single attribute selection, in order to incorporate penalty techniques such as the one described in the end of previous Section. The above selections are performed independently on positive and negative attribute.

### 5. PATTERN GENERATION AND USE

A pattern \( P \) is a logic function of attributes \( a_i^+ \), typically a conjunction of literals, which are binary attributes \( a_i^+ \in \mathcal{U} \) or negated binary attributes \( a_i^- \). Given a binarized record \( r_0 \), that is a set of binary values \( \{b_i^+\} \), each literal of a generic pattern \( P \) receives a value, and so \( P \) itself receives a value, denoted by \( P(r) \in \{0, 1\} \). We say that a pattern \( P \) covers a record \( r \) if \( P(r) = 1 \), and that pattern \( P \) is activated by \( r \). In the standard LAD procedure,[11] a positive pattern \( P + \) has to cover at least one positive record \( r^+ \) but no negative ones, and a negative pattern \( P^- \) is defined symmetrically. This, however, can lead to improper pattern generation in the case of noisy or otherwise difficult datasets. In our procedure, patterns are built in a bottom-up fashion, as described below. For obtaining a positive pattern, we generate every possible logic conjunction grouping up to \( p \) literals, using one after another all literals obtainable from \( \mathcal{U} \). When a conjunction \( P^- \) verifies the following coverage conditions:

- \( P^- \) covers at least \( n_c \) positive records of \( S \)
- \( P^- \) covers at most \( n_n \) negative records of \( S \)

We need constraints imposing that \( \sum_{s \in \mathcal{S}} \delta \) reproduce in \( T \) the class distribution of \( S \), so \( |T| \cdot \gamma \) should be as similar as possible to \( |\mathcal{S}| \cdot |T| \), and connecting the difference to the introduced \( \gamma \):

\[
\sum_{t \in T} c_t \leq \sum_{s \in \mathcal{S}} c(s) \cdot \frac{|T|}{|S|} + |T| \gamma + \rho
\]

\[
\sum_{t \in T} c_t \geq \sum_{s \in \mathcal{S}} c(s) \cdot \frac{|T|}{|S|} - |T| \gamma - \rho
\]

Note that, when we need to classify just one or a few records, obtaining the same class distribution of \( S \) could be impossible. For example, if we need to classify two records, and the fraction of positive \( \frac{|\mathcal{S}|}{|T|} \) is 0.2, targeting at that class distribution is clearly useless. Hence, above equation should have no effect when \( T \) is very small. This is obtained by using value \( \rho \), that, when set for instance at 3, relaxes constraints of 3 units. For large \( |T| \) this relaxation is negligible, while for small \( |T| \) the problem gradually reduces to minimizing only the classification error on \( S \).
As a general result, our examinations demonstrate that the exertion put resources into assessing the nature of the diverse paired characteristics gives back a better arrangement exactness with deference than the standard LAD methodology. In the totality of the examined cases, undoubtedly, SLAD is more exact than LAD. That extra exertion obviously required an extra computational time, however that was practically insignificant, and in addition, in the arrangement of the backing set determination issue, weighted set covering issues can by and large be illuminated in times which are much shorter than those required for the comparing non-weighted ones, so the parity is supportive of performing the above quality assessment. Moreover, the arrangement of the backing set choice issue as twofold backpack.

6. CONCLUSION
To order in brief times with a decent level of precision on the premise of little preparing sets is required in an assortment of useful applications. Sadly, getting these three alluring elements together can be exceptionally troublesome. We consider here the structure of the Logical Analysis of Data (LAD), and propose a few improvements to this system in view of measurable contemplations on the information.

7. ACKNOWLEDGEMENTS
I take this golden opportunity to owe our deep sense of gratitude to my project guide Prof. V. S. Nandedkar, for her instinct help and valuable guidance with a lot of encouragement throughout this paper work, right from selection of topic work up to its completion. My sincere thanks to Head of the Department of Computer Engineering Prof. N.D. Kale and Prof. Y.B. Gurav who continuously motivated and guided us for completion of this paper. I am also thankful to our PG Coordinator, all teaching and nonteaching staff members, for their valuable suggestions and valuable co-operation for partially completion of this work. I specially thank to those who helped us directly-indirectly in completion of this work successfully.

8. REFERENCES