Annals of Operations Research (2020) 287:11–35
https://doi.org/10.1007/s10479-019-03389-7

ORIGINAL RESEARCH

Logical analysis of multiclass data with relaxed patterns

Travaughn C. Bain[1] · Juan F. Avila-Herrera[2] · Ersoy Subasi[3] · Munevver Mine Subasi[1]

Published online: 25 September 2019

Abstract
An efficient and robust algorithm based on mixed integer linear programming is proposed to extend the Logical Analysis of Data (LAD) methodology to solve multiclass classification problems, where One-vs-Rest learning models are constructed to classify observations in predefined classes. The proposed algorithm uses two control parameters, homogeneity and prevalence, for identifying relaxed (fuzzy) patterns in multiclass datasets. The utility of the proposed method is demonstrated through experiments on multiclass benchmark datasets. Numerical experiments show that the efficiency and performance of the proposed multiclass LAD method with relaxed patterns is comparable to, if not better than, those of the previously developed LAD based multiclass classification as well as other well-known supervised learning methods.

Keywords Supervised learning · Multiclass classification · Logical analysis of data · Mixed integer linear programming

1 Introduction

With the advent of new technologies, we are facing an exponentially growing volume of complex structured data in diverse fields of science and engineering, where data mining has become ubiquitous in many real-world applications. Wide interest for data analysis com-

✉ Munevver Mine Subasi
msubasi@fit.edu

Travaughn C. Bain
tbain2013@my.fit.edu

Juan F. Avila-Herrera
delagarita@gmail.com

Ersoy Subasi
esubasi@fit.edu

1 Department of Mathematical Sciences, Florida Institute of Technology, 150 W. University Blvd., Melbourne, FL 32901, USA

2 Escuela de Informática, Universidad Nacional Escuela de Matemática, San Jose, Costa Rica

3 Department of Computer Engineering and Sciences, Florida Institute of Technology, 150 W. University Blvd., Melbourne, FL 32901, USA

ing from numerous disciplines motivated an interdisciplinary research approach, which cuts across the disciplines of applied mathematics and computer science and fosters the integration of ideas. These interdisciplinary efforts set the stage for innovation by uniting together to create new tools, develop new disciplines, and ultimately open new avenues of research. A fundamental challenge in data mining is to extract, analyze, and interpret knowledge from large-scale datasets effectively and efficiently. In order to address this challenge, the traditional statistical methods are complemented by sophisticated supervised learning techniques, including, for example, support vector machines (Burges 1998; Schölkopf and Smola 2001), neural networks (Bishop 2007; Fausett 1994), decision trees (Bishop 2007; Duda et al. 2001), and a pattern based method, called Logical Analysis of Data (Alexe et al. 2007; Boros et al. 1997, 2000) that are designed to find a decision boundary from the given samples with known classes to predict the class of a new or unseen observation.

Supervised learning algorithms solve binary classification problems, where a learning model is constructed to separate observations into two predefined classes. However, many real-world problems require the identification of more than two subgroups of observations and the features and patterns associated with each subgroup. Typical examples include (i) identification of different subtypes of human cancers (Hanash and Creighton 2003), (ii) protein fold recognition (Ding and Dubchak 2001), (iii) microscopy images (Boland et al. 1998; Misselwitz et al. 2010), (iv) histogram based image classification (Chapelle et al. 1999), (v) handwritten character recognition (LeCun et al. 1989; Lee and Seung 1997), (vi) part-of-speech tagging (Nakagawa et al. 2002), (vii) speech recognition (Jelinek 1998), (viii) text categorization (Apté et al. 1994), etc.

Since the problem is of practical importance, there have been several attempts to extend well-known binary classification algorithms to multiclass problems. The most common approaches to multiclass classification are the natural extension of binary classification problem known as One-vs-One (OvO) and One-vs-Rest (OvR) (Hastie and Tibshirani 1998). Given a K-class dataset $\Omega \subset \mathbb{R}^{m\times n}$ with m observations and n features, OvO scheme, shown in Fig. 1, assumes that there exists a separator between any two classes and builds $K(K-1)/2$ classifiers, denoted by f_{ij}, to distinguish each pair of classes $\mathcal{C}_i, \mathcal{C}_j \in \mathcal{C}, i \neq j$, where $\mathcal{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$ is the family of classes. The class of a new or unseen observation, $\mathbf{o} \in \mathbb{R}^n, \mathbf{o} \notin \Omega$, is then assigned by the use of the discriminant function:

$$f(\mathbf{o}) = \arg\max_i \sum_j f_{ij}(\mathbf{o}). \tag{1}$$

A less expensive approach OvR, shown in Fig. 2, assumes the existence of a single separator between a class $\mathcal{C}_i$ (for some i) and all other classes in $\mathcal{C}$ and builds K different binary classifiers. Let f_i be the ith classifier separating observations in class $\mathcal{C}_i$ (considered to be positive) and observations in $\mathcal{C} \setminus \mathcal{C}_i$ (forming the set of negative observations). In this case a new or unseen observation $\mathbf{o} \in \mathbb{R}^n, \mathbf{o} \notin \Omega$ is classified by

$$f(\mathbf{o}) = \arg\max_i f_i(\mathbf{o}). \tag{2}$$

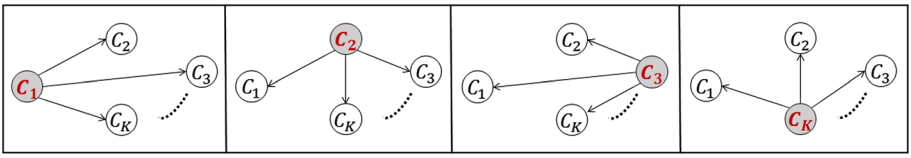

Fig. 1 One-vs-One (OvO) multiclass scheme

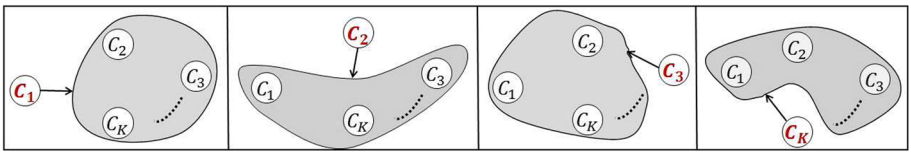

Fig. 2 One-vs-One (OvR) multiclass scheme

Since both approaches are easy to adopt, diverse groups of researchers invented them independently: see, for example, multiclass classification (Beygelzimer et al. 2007; Gehler and Nowozin 2009; Har-Peled et al. 2002; Yang and Tsang 2012), discriminant analysis for multiclass classification (Li et al. 2006; Liu et al. 2011), multiclass learning (Daniely et al. 2012; Even-Zohar and Roth 2001), combining many two-class classifiers into a multiclass classifier (Galar et al. 2011; Platt et al. 2000; Tax and Duin 2002; Tewari and Bartlett 2007; Wu et al. 2004), multiclass classification with applications (Singh-Miller and Collins 2009), mixed integer programming approach to multiclass data classification (Avila-Herrera and Subasi 2013, 2015; Kim and Choi 2015; Üney and Türkay 2006), multiclass classification by using support vector machine (Aiolli and Sperduti 2005) and general multiclass classification methods review (Aly 2005). The choice between the use of OvO and OvR in multiclass problems is largely computational.

Despite the undoubted advancements in the area of multiclass classification, there is still room for developing new approaches to improve the effectiveness and efficiency of the methods and tools to analyze archives of historical records for the purpose of discovering hidden structural relationships in large-scale datasets. In this paper, we integrate the mixed integer linear programming based Logical Analysis of Data (LAD) approach of Ryoo and Jang (2009) with the multiclass LAD method of Avila-Herrera and Subasi (2013, 2015) to develop a new multiclass LAD algorithm, where two control parameters, homogeneity and prevalence, are incorporated to generate relaxed (fuzzy) patterns.

LAD is a pattern-based two-class learning method which integrates principles of combinatorics, optimization, and the theory of Boolean functions. The research area of LAD was introduced and developed by Hammer (1986) whose vision expanded the LAD methodology from theory to successful data applications in numerous biomedical, industrial, and economics case studies, see, for example, Alexe et al. (2003, 2004, 2005, 2006), Hammer et al. (1999, 2011), Hammer and Bonates (2006), Lauer et al. (2002), Reddy et al. (2008, 2009) and the references therein. The implementation of LAD method was described in Boros et al. (1997, 2000), Crama et al. (1988) and several further developments of the original technique were presented in Alexe et al. (2007), Alexe and Hammer (2006), Bonates et al. (2008), Guo and Ryoo (2012), Hammer et al. (2004), Ryoo and Jang (2009). An overview of standard LAD method can be found in Alexe et al. (2007) and Bonates et al. (2008). Various recent applications of LAD are presented in Dupuis et al. (2012), Ghasemi et al. (2013), Lejeune et al. (2018), Lejeune and Margot (2011), Mortada et al. (2011) and Subasi et al. (2017). LAD method has been extended to survival analysis (Kronek and Reddy 2008) and regression analysis (Bonates and Hammer 2007; Lemaire 2011) as well.

Extensions of LAD method to multiclass problems are previously studied by Moreira (2000), Mortada (2010), Mortada et al. (2014). Moreira (2000) proposed two methods to break down a multiclass classification problem into two-class problems using an OvO approach. The first method uses the typical OvO scheme which does not require the alteration of the structure of the standard LAD method presented by Boros et al. (2000). The second OvO-type method modifies the architecture of the pattern generation and theory formation steps

in standard LAD method, where a LAD pattern P_{ij} is generated for each pair of classes $\mathcal{C}_i, \mathcal{C}_j \in \mathcal{C}, i \neq j$.

Mortada (2010) proposed a multiclass LAD method, integrating ideas from the second approach presented by Moreira (2000), which is based on OvO scheme and an implementation of LAD based on mixed integer linear programming (MILP) presented by Ryoo and Jang (2009). The methodology of Mortada (2010) was applied to five multiclass benchmark datasets. Mortada (2010) observed that the MILP based LAD approach of Ryoo and Jang (2009) combined with the second approach of Moreira (2000) provides classification models with higher accuracy than those models obtained by multiclass approach applied to standard LAD algorithm of Boros et al. (2000).

Recent papers by Avila-Herrera and Subasi (2013, 2015) and Kim and Choi (2015) have also considered the multiclass extension of LAD. Avila-Herrera and Subasi (2013, 2015) explored and rectified the limitations of the two-class MILP LAD approach by Ryoo and Jang (2009), relating to its poor differentiating power in two-class classification. Kim and Choi (2015) developed an efficient iterative genetic algorithm with flexible chromosomes and multiple populations to extend LAD to multiclass classification. The performance of the method was evaluated on six benchmark multiclass datasets.

In this paper, we propose a parametrized/relaxed algorithmic approach that builds on the MILP pattern generation approach of Ryoo and Jang (2009) and multiclass LAD approach of Avila-Herrera and Subasi (2013, 2015) that constructs an OvR-type LAD classifier to identify patterns in multiclass datasets. This modification introduces two control parameters, homogeneity and prevalence, to generate fuzzy patterns which we call "relaxed patterns". The organization of the paper is as follows. Section 2 describes the basic principles of the standard LAD method of Boros et al. (2000). Section 3 presents the proposed MILP based relaxed multiclass LAD approach to obtain OvR-type multiclass LAD classifiers. In Sect. 4 we present experiments on multiclass benchmark datasets to demonstrate the utility of our proposed methodology and compare the efficiency and performance of the proposed multiclass LAD method with relaxed patterns with that of the previously developed LAD based multiclass classification as well as other well-known supervised learning methods.

2 Preliminaries: logical analysis of data

Logical Analysis of Data (LAD) is a two-class learning method based on combinatorics, optimization, and the theory of Boolean functions Boros et al. (2000). Given an input dataset, Ω, consisting of two disjoint classes Ω^+ (set of positive observations) and Ω^- (set of negative observations), where $\Omega = \Omega^+ \cup \Omega^-$ and $\Omega^+ \cap \Omega^- = \emptyset$, there exists a hidden function of nature separating the observations in Ω^+ and Ω^-. The goal of LAD is to identify positive and negative patterns to approximate the hidden function of nature (as illustrated in Fig. 3).

The standard LAD methodology presented in Boros et al. (1997, 2000) is a multistep procedure, consisting of five main components outlined below. A more detailed overview of the standard LAD method together with the recent developments in its theory and applications can be found in Lejeune et al. (2018).

2.1 Discretization/binarization and support set selection

This step is the transformation of numeric features (attributes/variables) into binary features without losing predictive power. The procedure consists of finding cut-points for each numeric

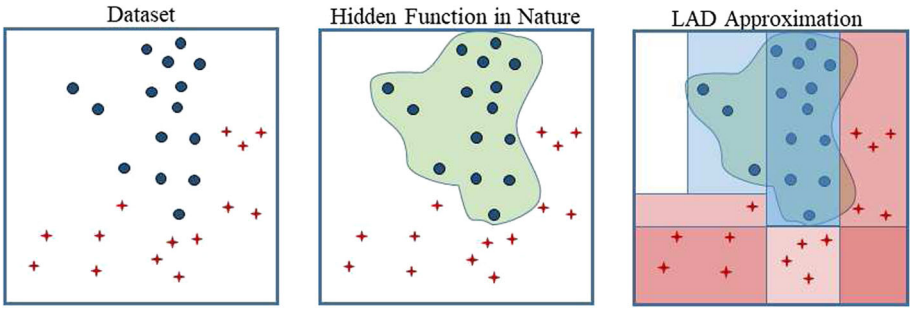

Fig. 3 LAD approximation to hidden function in nature

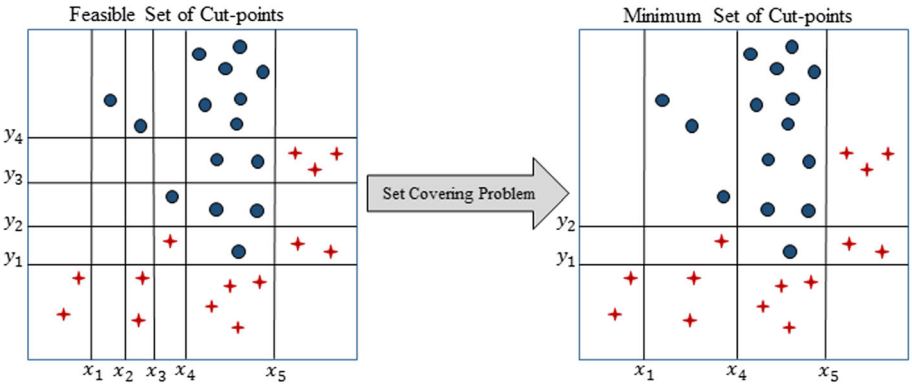

Fig. 4 Illustration of cutpoints

feature. The set of cut-points can be interpreted as a sequence of threshold values collectively used to build a global classification model over all features (Boros et al. 1997, 2000). Given a numeric dataset, there is a large number of "feasible" cut-points, identified by the standard methods such as equal width, equal intervals, based on entropy, chi-square tests, etc. In most cases many of these cut-points may be redundant. In this step a set covering problem is solved to identify an optimal (minimum size) set of cut-points to transform the numerical features into binary ones, illustrated in Fig. 4.

Discretization is a very useful step in data mining, especially for the analysis of medical data (which is very noisy and includes measurement errors)—it reduces noise and produces robust results. The problem of discretization is well studied and many powerful methods are presented in literature, see, e.g., the survey papers (Kotsiantis and Kanellopoulus 2006; Liu et al. 2002). Discretization step may produce several binary features some of which may be redundant. In this step of LAD procedure, a minimum set covering problem is solved to obtain an irredundant smallest subset of binary variables, which can distinguish every pair of positive and negative observations in the dataset. The resulting set is called a *support set* (Boros et al. 1997, 2000). If the input data is categorical (nominal) or text, then there are well-known techniques to binarize the data (Aggarwal 2015).

2.2 Pattern generation

Patterns are the key ingredients of LAD algorithm. This step uses the features in combination to produce rules (combinatorial patterns) that can define homogenous subgroups of interest within the data. The simultaneous use of two or more features allows the identification of more complex rules that can be used for the precise classification of an observation as illustrated in Fig. 5.

Given a binary (or binarized) dataset $\Omega = \Omega^+ \cup \Omega^- \subset \mathbb{B}^{m \times n}$ with m observations and n features, where $\Omega^+ \cap \Omega^- = \emptyset$, a *pattern* P is simply defined as a subcube of $\mathbb{B}^n = \{0, 1\}^n$. A LAD pattern can be described as a Boolean term, that is, a conjunction of literals (binary variables or its negation) which does not contain both a variable and its negation:

$$P = \bigwedge_{j \in M_P} x_j \bigwedge_{j \in N_P} \bar{x}_j$$

where $M_P, N_P \subseteq \{1, \cdots, n\}$, $M_P \cap N_P = \emptyset$, and x_j is the Boolean literal associated with the jth feature in the dataset.

Patterns define homogeneous subgroups of observations and have the following distinctive characteristics.

- Degree: The number of literals involved in the definition of a pattern is called the *degree* of the pattern.
- Homogeneity: The proportion of positive (negative) observations among all those observations covered by a pattern is called the *positive (negative) homogeneity* of the pattern. A *pure positive pattern* has 100% positive homogeneity and 0% negative homogeneity and is defined as a combination of features which covers a proportion of positive observations, but none of the negative ones: $P(\omega^+) = 1$ for at least one $\omega^+ \in \Omega^+$ and $P(\omega^-) = 0$ for every $\omega^- \in \Omega^-$. A pure negative pattern can be defined similarly: $P(\omega^-) = 1$ for at least one $\omega^- \in \Omega^-$ and $P(\omega^+) = 0$ for every $\omega^+ \in \Omega^+$.
- Prevalence: The proportion of positive (negative) observations covered by a pattern is called the *positive (negative) prevalence* of the pattern.
- Coverage: An observation $\omega \in \Omega$ satisfying the conditions of a pattern P, i.e., $P(\omega) = 1$, is said to be *covered* by that pattern. *Coverage* of a pattern P, denoted by $Cov(P)$, is the set of observations covered by the pattern.
- Hazard ratio: The ratio between the proportion of positive observations among all those observations covered by a pattern and the proportion of positive observations among those observations not covered by the pattern is called the *hazard ratio* of the pattern.

Note that a positive homogeneity plus negative homogeneity of a pattern is 100%. A pattern associated with positive (negative) class must exhibit a positive (negative) homogeneity more than 50%. A positive (negative) pattern with positive (negative) homogeneity less than 100% is called a relaxed (fuzzy) pattern. We also remark that a high quality positive (negative) pattern must have high positive (negative) homogeneity and high positive (negative) prevalence. Smaller degree reduces the complexity, allowing easier interpretation of the pattern.

The most straightforward approach to pattern generation is based on the use of combinatorial enumeration techniques, for example, a *bottom-up/top-down* approach (Boros et al. 1997, 2000). The bottom-up approach follows a lexicographic order in generating the patterns in order to reduce the amount of computations necessary. The approach starts with terms of degree one that cover some positive observations. If such a term does not cover any negative observation, it is a positive pattern. Otherwise, literals are added to the term one by

one until generating a pattern of prefixed degree. The top-down pattern generation approach starts by considering all uncovered observations as patterns of degree n and for each of those patterns, literals are removed one by one, until a pattern with smallest degree is reached. The enumeration type pattern generation approach is a costly process. Given a two-class binary dataset with n features, the total number of candidate patterns to be searched is $\sum_{i=1}^{n} 2^i \binom{n}{i}$ and the number of degree d patterns can be $2^d \binom{n}{d}$.

A pattern P is called a *strong pattern* if there is no pattern P' such that $Cov(P) \subset Cov(P')$. Pattern P is called a *prime pattern* if the deletion of any literal from P results in a term that is no longer a pattern. Since patterns play a central role in LAD methodology, various types of patterns have been studied and several pattern generation algorithms have been developed for their enumeration (see Lejeune et al. 2018 and the references therein). Our OvR-type multiclass LAD algorithm is motivated by the MILP approach of Ryoo and Jang (2009) that generates strong LAD patterns in a two-class dataset.

Consider a two-class dataset Ω consisting of m binary observations and n features. Let $I^+ = \{i : \omega_i^+ \in \Omega^+\}$ and $I^- = \{i : \omega_i^- \in \Omega^-\}$, where $\Omega = \Omega^+ \cup \Omega^-$ and $\Omega^+ \cap \Omega^- = \emptyset$. For each observation $\omega_i \in \Omega$, let ω_{ij} denote the binary value of the jth feature in the ith observation ω_i. Let x_j, $j = 1, \cdots, n$, denote the Boolean literal corresponding to the jth feature b_j in Ω and introduce n new features $x_{n+j} = 1 - x_j$, $j = 1, \cdots, n$ (representing $\bar{b}_j$). If $x_j = 1$ for some specific $j = 1, \ldots, n$, then $b_j = 1$ in observation ω_i and if $x_{n+j} = 1$ for some specific $j = 1, \ldots, n$, then $b_j = 0$ in observation ω_i. Let ρ_i, $i = 1, \ldots, m$, be penalties associated with the coverage of the pattern P, defined as

$$\rho_i = \begin{cases} 1 & \text{if} \quad P(\omega_i) = 0, \ i \in I^+ \\ 0 & \textit{otherwise}. \end{cases}$$

Following the pattern generation approach of Ryoo and Jang (2009), we formulate the below MILP to generate a positive LAD pattern with $\alpha\%$ positive homogeneity and $\beta\%$ positive prevalence:

$$\begin{aligned}
&\text{minimize } cd + \sum_{i \in I^+} \rho_i \\
&\text{subject to} \\
&\sum_{j=1}^{2n} \omega_{ij} x_j + n\rho_i \geq d, \quad i \in I^+ \\
&\sum_{j=1}^{2n} \omega_{ij} x_j - y_i \leq d - 1, \quad i \in I^- \\
&\sum_{i \in I^-} y_i \leq \alpha |\Omega^+| \\
&\sum_{i \in I^+} \rho_i \leq (1 - \beta) |\Omega^+| \\
&x_j + x_{n+j} \leq 1, \quad j = 1, \ldots, n \\
&\sum_{j=1}^{2n} x_j = d \\
&1 \leq d \leq n, \quad d \in \mathbb{Z}^+ \\
&0 \leq y_i \leq 1, \quad i = 1, \ldots, m \\
&\rho_i \in \{0, 1\}, \quad i = 1, \ldots, m \\
&x_j \in \{0, 1\}, \quad j = 1, \ldots, 2n
\end{aligned} \tag{3}$$

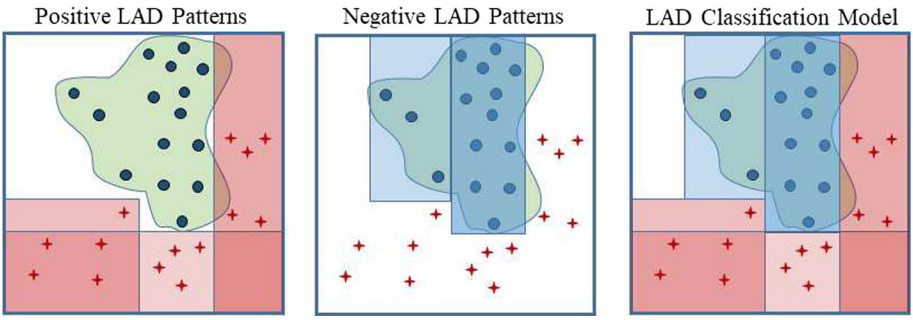

Fig. 5 LAD classification model

where $0 \leq \alpha, \beta \leq 1$ and $c \in \mathbb{R}$ are constants, d is the unknown degree of a positive pattern P, and variables $y_i, i = 1, \ldots, m$ are the relaxation variables, ensuring the generation of pattern with $\alpha\%$ homogeneity and $\beta\%$ prevalence.

Variables x_j and $x_{n+j}, j = 1, \ldots, n$, associated with the jth feature in dataset Ω determine whether the jth feature takes value 1 or 0 in the ith observation $w_i \in \Omega$. A feasible solution of problem (3) is a positive pattern with degree $d, \alpha\%$ positive homogeneity, and $\beta\%$ positive prevalence. When the scaling parameter c is positive, an optimal solution of problem (3) is a positive pattern with the smallest degree and maximum coverage, and hence is a strong prime pattern (Ryoo and Jang 2009):

$$P = \bigwedge_{\{j\,:\,x_j=1\}} b_j \bigwedge_{\{j\,:\,x_{n+j}=1\}} \bar{b}_j \,.$$

Note that if we change the roles of index sets I^+ and I^- in problem (3), an optimal solution of the problem provides us with a negative strong prime pattern with $\alpha\%$ negative homogeneity and $\beta\%$ negative prevalence.

2.3 LAD model

The entire collection of patterns generated in *Pattern Generation Step* is called *pandect* and is denoted by $\mathcal{P} = \mathcal{P}^+ \cup \mathcal{P}^-$, where $\mathcal{P}^+$ and $\mathcal{P}^-$ are disjoint sets of all positive and negative patterns, respectively. A LAD model (illustrated in Fig. 5), denoted by $\mathcal{M}$ is a subset of pandect, that is, $\mathcal{M} = \mathcal{M}^+ \cup \mathcal{M}^-$, where $\mathcal{M}^+ \subseteq \mathcal{P}^+$ and $\mathcal{M}^- \subseteq \mathcal{P}^-$.
Patterns are selected into the LAD model $\mathcal{M}$ so that the model provides the same separation of the positive and negative observations as the pandect $\mathcal{P}$.

In many cases, when constructing a LAD model, every observation in the training dataset is required to be covered at least k times ($k \in \mathbb{Z}^+$) by the patterns in the model. The standard LAD approach of Boros et al. (1997, 2000) produces patterns using enumerative techniques. Greedy-type heuristic approaches are then adopted to select patterns into a final LAD model. As for the MILP approach of Ryoo and Jang (2009), producing patterns as optimal solutions of the MILP's involved, such as problem (3), the authors proposed Algorithm 1 to find a LAD model.

Note that after a pattern is generated as an optimal solution of problem (3), Algorithm 1, proposed by Ryoo and Jang (2009), removes the observations covered by that pattern from the training data to prevent the algorithm from finding the same pattern found in the previous

Algorithm 1: Pattern Generation (Ryoo and Jang 2009)

Data: Training data, Support Features, MILP model (3) for pattern generation
Result: Set of + and − patterns ($\mathcal{M}^+$ and $\mathcal{M}^-$, respectively)

```
for * ∈ {+, −} do
    set M* = ∅ ;
    while I* ≠ ∅ do
        formulate and solve an instance of the MILP problem (3);
        form a pattern P from the solution obtained;
        M* ← M* ∪ {P};
        I* ← I* \ {i ∈ I* : ω_i is covered by P};
return M*;
```

solutions of problem (3). The algorithm terminates when every observation is covered at least once. The resulting set of positive and negative patterns form a LAD model $\mathcal{M}$.

2.4 Prediction and accuracy

In the final step of the LAD framework, LAD classification model $\mathcal{M}$ is used for the classification of a new or unseen observation $\mathbf{o} \in \mathbb{B}^n$, $\mathbf{o} \notin \Omega$ by the use of a discriminant function $\Delta : \{0,1\}^n \rightarrow I\!R$ associated with the model $\mathcal{M}$, where $\Delta(\mathbf{o})$ is defined as the difference between the weighted proportion of positive patterns and negative patterns covering $\mathbf{o}$, that is,

$$\Delta(\mathbf{o}) = \sum_{P_k^+ \in \mathcal{M}^+} \delta_k^+ P_k^+(\mathbf{o}) \; - \sum_{P_k^- \in \mathcal{M}^-} \delta_k^- P_k^-(\mathbf{o}),$$

where $\delta_k^+ \geq 0$ and $\delta_k^- \geq 0$ are the weights assigned to positive patterns $P_k^+ \in \mathcal{M}^+$ and negative patterns $P_k^- \in \mathcal{M}^-$, respectively. The weights δ_k^+ and δ_k^- can be calculated in several ways. One possibility is to use the proportion of positive (negative) observations covered by a positive pattern $P_k^+ \in \mathcal{M}^+$ (a negative pattern $P_k^- \in \mathcal{M}^-$) to the total number of positive (negative) observations (i.e., the prevalence of the pattern):

$$\delta_k^+ = \frac{1}{|\Omega^+|} \sum_{i \in I^+} P_k^+(\omega_i^+) \quad \text{and} \quad \delta_k^- = \frac{1}{|\Omega^-|} \sum_{i \in I^-} P_k^-(\omega_i^-)$$

where $I^+ = \{i : \omega_i^+ \in \Omega^+\}$, and $I^- = \{i : \omega_i^- \in \Omega^-\}$.

The accuracy of the model is estimated by a classical cross-validation procedure, where the dataset Ω is randomly divided into two disjoint subsets called *training* and *test* sets (Aggarwal 2015; Dietterich 1998; Efron and Tibshirani 1986; Hastie et al. 2005). A LAD model is generated on the training data and evaluated on the test data. The experiment is repeated several times (determined by the data analyst) and the accuracy of the LAD model is reported as the average accuracies on the test datasets.

If an external dataset (validation set) is available, a LAD model $\mathcal{M}$ is obtained for the original dataset Ω and the performance of the model is evaluated on the validation set.

Figures 6 and 7 shows the overall framework of LAD method, including the validation step in case of the absence or presence of an external data, respectively.

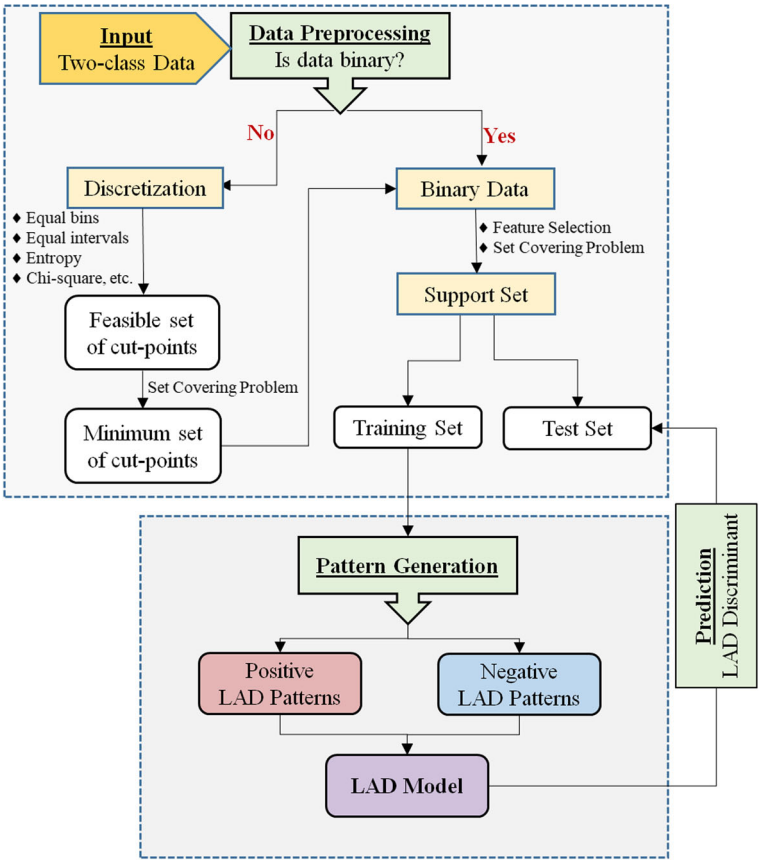

Fig. 6 Cross-validation of LAD model

3 Multiclass LAD method with relaxed patterns

In this section we present an OvR-type extension of LAD algorithm to multiclass classification problems. As in conventional LAD method, our multiclass LAD approach has four steps: (i) binarization and support set selection, (ii) pattern generation, (iii) theory formation, and (iv) prediction. These steps are outlined below.

3.1 Binarization and support set selection

Binarization of a multiclass numeric data is similar to that of two-class data discussed in Sect. 2.1. Binarization step associates several cut-points, α_{v_k}, and the following indicator variables to a numeric feature v to transform it into a set of binary features:

$$x_{v_k} = \begin{cases} 1 & \text{if} \quad v \geq \alpha_{v_k} \\ 0 & \text{if} \quad v < \alpha_{v_k} \end{cases}$$

Transforming the data from discrete levels to indicator variables results in a multiclass binary dataset. For each variable, virtually any numerical value can be considered as a cut-point.

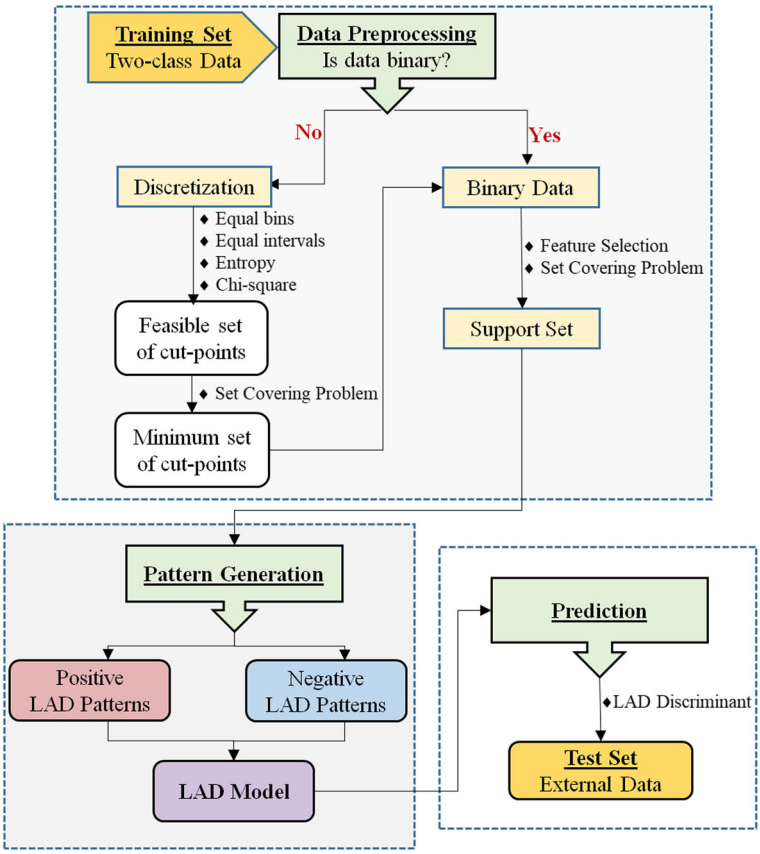

Fig. 7 Validation of LAD model on an external dataset

However, the cut-points are chosen in a way which allows to distinguish between observations in different classes (Kotsiantis and Kanellopoulus 2006). An optimal (minimum size) set of cut-points can be obtained by the use of a set-covering problem. The multiclass discretization problem is extensively studied and there are several different approaches to accomplish this task (Friedman et al. 2000). As in two-class binarization procedure, an irredundant subset of binarized features (support set) that can distinguish every pair of observations in different classes of the K-class dataset Ω can be identified by solving a minimum set covering problem similar to the one presented in Boros et al. (1997, 2000).

In what follows we assume that the input dataset is a binary (or binarized) multiclass dataset.

3.2 Relaxed multiclass LAD pattern generation

Let $\Omega = \Omega_1 \cup \cdots \cup \Omega_K \subset \mathbb{B}^{m \times n}$ be a K-class binary dataset with n features and m observations, where $\Omega_i \cap \Omega_j = \emptyset$ for all $i \neq j$. We introduce the following notations:

- $\mathcal{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$: family of classes in Ω, that is, any observation in Ω_k has class $\mathcal{C}_k$, $k = 1, \ldots, K$.

- $P_{\mathcal{C}_p}$: pattern associated with class $\mathcal{C}_p$, $p = 1, \ldots, K$.
- ω_{ij}: binary value of the jth feature in the ith observation $\omega_i \in \Omega$.
- y_j, $j = 1, \ldots, n$: binary variable representing a Boolean literal corresponding to the jth binary feature b_j in Ω. If $y_j = 1$ for some specific $j = 1, \ldots, n$, then $b_j = 1$ in observation ω_i and the jth literal, x_j, is used in the construction of the pattern.
- $y_{n+j} = 1 - y_j$, $j = 1, \ldots, n$: binary variable representing the negation of the jth binary feature $\bar{b}_j$ in Ω. If $y_{n+j} = 1$ for some specific $j = 1, \ldots, n$, then $b_j = 0$ in observation ω_i and the negation of the jth literal, $\bar{x}_j$, is used in the construction of the pattern.
- $d \in \mathbb{Z}^+$: unknown degree of a pattern $P_{\mathcal{C}_p}$, $p = 1, \ldots, K$.
- $\rho = (\rho_1, \rho_2, \ldots, \rho_m)$: penalty vector associated with the coverage of the pattern $P_{\mathcal{C}_p}$. For all $i = 1, \ldots, m$,

$$\rho_i = \begin{cases} 1 & \text{if } \ \omega_i \in \mathcal{C}_p \text{ is not covered by pattern } P_{\mathcal{C}_p} \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

- z_i, $i = 1, \ldots, m$: relaxation variables, ensuring the generation of a pattern with homogeneity less than 100%.

We shall formulate an MILP to generate a smallest degree and maximum coverage relaxed pattern $P_{\mathcal{C}_p}$ associated with some class $\mathcal{C}_p$ in K-class dataset Ω. In order to achieve this goal we proceed as follows:

- Since a pattern cannot include both literals x_j and $\bar{x}_j$, we impose the condition

$$y_j + y_{n+j} \leq 1, \quad j = 1, \ldots, n. \tag{5}$$

- Degree of pattern $P_{\mathcal{C}_p}$ is $d \in \mathbb{Z}^+$ (unknown):

$$\sum_{j=1}^{2n} y_j = d. \tag{6}$$

- Consider the augmented matrix $B = [\Omega | \overline{\Omega}]$, where $\overline{\Omega}$ is the binary data obtained from Ω by replacing 0 entries by 1 and 1 entries by 0. Define the vector $\mathbf{v} = B\mathbf{y}$. In order to produce a relaxed pattern $P_{\mathcal{C}_p}$ with $\alpha\%$ homogeneity and $\beta\%$ prevalence, we prescribe the following constraints:

$$v_i + n\rho_i \geq d, \quad i \in I_p, \tag{7}$$

$$v_i - z_i \leq d - 1, \quad i \in I_k, \quad k = 1, \ldots, K, \quad k \neq p \tag{8}$$

$$\sum_{i \in I_k} z_i \leq \alpha |\Omega_p|, \tag{9}$$

$$\sum_{i \in I_p} \rho_i \leq (1 - \beta)|\Omega_p|, \tag{10}$$

$$0 \leq z_i \leq 1, \quad i = 1, \ldots, m \tag{11}$$

$$1 \leq d \leq n, \quad d \in \mathbb{Z}^+ \tag{12}$$

$$\rho_i \in \{0, 1\}, \quad i = 1, \ldots, m \tag{13}$$

$$y_j \in \{0, 1\}, \quad j = 1, \ldots, 2n \tag{14}$$

where $I_p = \{i : \omega_i \text{ is in class } \mathcal{C}_p\}$ and $I_k = \{i : \omega_i \text{ is in class } \mathcal{C}_k\}$ for all $k \neq p$.

- Let $c \in I\!R^+$ be a constant. Our goal is to generate a pattern $P_{\mathcal{C}_p}$ with minimum degree and maximum coverage. Therefore, we consider the objective function

$$cd + \sum_{i \in I_p} \rho_i \tag{15}$$

We use conditions in (5)–(14) to formulate the following MILP:

$$\begin{aligned}
&\text{minimize } \ cd + \sum_{i \in I_p} \rho_i \\
&\text{subject to} \\
&\qquad v_i + n\rho_i \geq d, \quad i \in I_p \\
&v_i - z_i \leq d - 1, \quad i \in I_k \ , \ k = 1, \ldots, K, \ \ k \neq p \\
&\qquad \sum_{i \in I_k} z_i \leq \alpha |\Omega_p| \\
&\qquad \sum_{i \in I_p} \rho_i \leq (1 - \beta)|\Omega_p| \\
&y_j + y_{n+j} \leq 1, \quad j = 1, \ldots, n \\
&\qquad \sum_{j=1}^{2n} y_j = d \\
&1 \leq d \leq n, \quad d \in \mathbb{Z}^+ \\
&0 \leq z_i \leq 1, \quad i = 1, \ldots, m \\
&\rho_i \in \{0, 1\}, \quad i = 1, \ldots, m \\
&y_j \in \{0, 1\}, \quad j = 1, \ldots, 2n
\end{aligned} \tag{16}$$

where $c \in I\!R^+$ and $0 \leq \alpha, \beta \leq 1$ are input parameters.

Theorem 1 *Let* $(\boldsymbol{v}^*, \boldsymbol{y}^*, \boldsymbol{z}^*, \rho^*, d^*)$ *be a feasible solution of problem* (16)*. Let*

$$S = \left\{ j : \ y_j^* = 1, j = 1, \ldots, n \right\} \quad \textit{and} \quad \bar{S} = \left\{ j : \ y_{n+j}^* = 1, j = 1, \ldots, n \right\}.$$

Then

$$P_{\mathcal{C}_p} = \bigwedge_{S} x_j \bigwedge_{\bar{S}} \bar{x}_j \tag{17}$$

forms a relaxed (fuzzy) pattern of degree d^**, associated with class* $\mathcal{C}_p$*.*

Proof Let $(\mathbf{v}^*, \mathbf{y}^*, \mathbf{z}^*, \rho^*, d^*)$, where $\mathbf{v}^* = B\mathbf{y}^*$, be a feasible solution of problem (16). First note that the constraint

$$y_j + y_{n+j} \leq 1, \quad j = 1, \ldots, n$$

ensures that the Boolean term $P_{\mathcal{C}_p}$ shown in (17) does not contain both literals x_j and $\bar{x}_j$ associated with the jth feature in dataset Ω and the condition

$$\sum_{j=1}^{2n} y_j = d$$

guarantees that the term $P_{\mathcal{C}_p}$ is of degree d. The constraint

$$v_i + n\rho_i \geq d, \quad i \in I_p$$

ensures that $P_{\mathcal{C}_p}$ covers at least one observation ω_i in class $\mathcal{C}_p$, that is, $P_{\mathcal{C}_p}(\omega_i) = 1, i \in I_p$. If an observation $\omega_i, i \in I_p$, is covered by $P_{\mathcal{C}_p}$, then d number of y_j's are set to 1 and hence, we have $v_i = d, i \in I_p$, where v_i is the ith component of vector $\mathbf{v} = B\mathbf{y}$. However, if an observation is not covered by $P_{\mathcal{C}_p}$, then $v_i < d, i \in I_p$, and the term "$n\rho_i$" is added to the left hand side to compensate it. Similarly, the constraints

$$v_i - z_i \leq d - 1, \quad i \in I_k, \quad k = 1, \ldots, K, \quad k \neq p,$$
$$\sum_{i \in I_k} z_i \leq \alpha |\Omega_p|$$

with the relaxation variables $0 \leq z_i \leq 1,\ i = 1, \ldots, m$, guarantee that up to $(1 - \alpha)\%$ of the observations covered by the term $P_{\mathcal{C}_p}$ may be in $\Omega \setminus \Omega_p$, i.e., the homogeneity of pattern $P_{\mathcal{C}_p}$ is $\alpha\%$.

Moreover, the constraint

$$\sum_{i \in I_p} \rho_i \leq (1 - \beta)|\Omega_p|$$

ensures that at least $\beta\%$ of the observations in Ω_p are covered by pattern $P_{\mathcal{C}_p}$, i.e., the prevalence of pattern $P_{\mathcal{C}_p}$ is $\beta\%$.

Thus, the solution $(\mathbf{v}^*, \mathbf{y}^*, \mathbf{z}^*, \rho^*, d^*)$ can be used to form a relaxed (fuzzy) pattern $P_{\mathcal{C}_p}$, shown in (17), with degree d^*, homogeneity $\alpha\%$, and prevalence $\beta\%$. □

Theorem 2 *Let* $(\mathbf{v}^{opt}, \mathbf{y}^{opt}, \mathbf{z}^{opt}, \rho^{opt}, d^{opt})$ *be an optimal solution of problem* (16)*. Let*

$$S = \left\{ j \ : \ y_j^{opt} = 1, j = 1, \ldots, n \right\} \quad \textit{and} \quad \bar{S} = \left\{ j \ : \ y_{n+j}^{opt} = 1, j = 1, \ldots, n \right\}.$$

Then

$$P_{\mathcal{C}_p}^{opt} = \bigwedge_{S} x_j \bigwedge_{\bar{S}} \bar{x}_j \tag{18}$$

is a relaxed (fuzzy) strong prime pattern of degree d^{opt}*, associated with class* $\mathcal{C}_p$*.*

Proof Let $(\mathbf{v}^{opt}, \mathbf{y}^{opt}, \mathbf{z}^{opt}, \rho^{opt}, d^{opt})$ be an optimal solution of problem (16). Hence, as discussed in the proof of Theorem 1, it can be used to construct a relaxed LAD pattern $P_{\mathcal{C}_p}$ of degree d^{opt} that is associated with class $\mathcal{C}_p$. Note that the objective function of problem (16) minimizes the degree of $P_{\mathcal{C}_p}$, resulting in a prime pattern. Since the objective function simultaneously minimizes the penalties ρ_i associated with observations $\omega_i, i \in I_p$, the resulting pattern has the maximum coverage in class $\mathcal{C}_p$ and is a strong pattern. Thus, an optimal solution to problem (16) used to form a relaxed pattern $P_{\mathcal{C}_p}^{opt}$, shown in (18), is a strong prime pattern with degree d^{opt}, homogeneity $\alpha\%$, and prevalence $\beta\%$. □

3.3 Relaxed multiclass LAD model

As stated in Theorem 2, an optimal solution of problem (16) is a fuzzy strong prime pattern. In order to obtain a multiclass LAD model, patterns must be generated until every observation in class $\mathcal{C}_k$ for all $k = 1, \ldots, K$ is covered at least once. Before we develop an algorithm that produces a multiclass LAD model consisting of relaxed LAD patterns, we recall that in case of two-class MILP approach, Algorithm 1 of Ryoo and Jang (2009) (shown in Sect. 2.3) produces a set of patterns associated with a positive (negative) class that loops as many times

as necessary until all observations in positive (negative) class are covered by at least one pattern. Note that once a positive (negative) pattern P is found as an optimal solution of problem (3), Algorithm 1 of Ryoo and Jang (2009) removes the observations covered by the pattern P, whilst looping through execution. However, this is counterproductive because every time the algorithm loops through again, it uses less information (smaller training set) to generate new patterns. Mortada (2010) has adopted a similar approach to develop an OvO-type multiclass LAD algorithm, where observations covered by a pattern are removed from the training dataset while executing the proposed algorithm. As compared to Algorithm 1 of Ryoo and Jang (2009), the algorithm proposed by Mortada (2010) stops looping when each observation is covered by t patterns, where t is a user-input, i.e., determined by the data analyst.

In order to avoid the removal of observations from the training dataset when generating new patterns that form a multiclass LAD model containing relaxed patterns obtained as the optimal solutions of problem (16), we define κ as an m-vector that keeps track of the number of patterns covering an observation $\omega_i \in \Omega$ for all $i = 1, \ldots, m$. Initially, for each class $\mathcal{C}_k, k = 1, \ldots, K$, we set $\kappa = \mathbf{0}$. This vector shall be updated as new solutions of the MILP problem (16) are found. With the help of new vector κ, constraint (7), i.e., the first constraint in problem (16), can be replaced by

$$v_i + n\,(\rho_i + \kappa_i) \geq d, \quad i \in I_p. \tag{19}$$

where $\kappa_i \geq 0, i = 1, \ldots, m$.

Theorem 3 *Let* $(\boldsymbol{v}', \boldsymbol{y}', \boldsymbol{z}', \rho', d')$ *be an optimal solution of problem* (16)*, where the constraint*

$$v_i + n\rho_i \geq d, \quad i \in I_p$$

is replaced by constraint (19)*. Let*

$$S = \left\{ j : \; y'_j = 1, j = 1, \ldots, n \right\} \quad \textit{and} \quad \bar{S} = \left\{ j : \; y'_{n+j} = 1, j = 1, \ldots, n \right\}.$$

Then

$$P'_{\mathcal{C}_p} = \bigwedge_{S} x_j \bigwedge_{\bar{S}} \bar{x}_j$$

is a degree d' *relaxed strong prime pattern associated with class* $\mathcal{C}_p$*.*

Proof The proof of the assertion follows immediately from the proof of Theorem 1 and Theorem 2 and hence, is left to the reader. □

Theorem 3 enables us to propose an algorithmic approach to generate a multiclass LAD model consisting of relaxed LAD patterns obtained as the optimal solutions of our multiclass MILP problem (16). This approach is presented in Algorithm 2. We remark that, unlike the Algorithm 1 of Ryoo and Jang (2009), our proposed Algorithm 2 does not require the removal of observations from the training dataset at any iteration by adding "NewConstraint" to the relaxed MILP problem (16) each time a new pattern is generated to prevent the algorithm from finding the same pattern found at the previous iterations. This is achieved by introducing κ_i that keeps track of the number of patterns covering observations $\omega_i \in \Omega$ for all $i = 1, \ldots, m$ and "TotCov" that counts the number of observations covered so far.

Let $\mathcal{M}_k$ denote the set of relaxed LAD patterns associated with class $\mathcal{C}_k, k = 1, \ldots, K$ and $\mathcal{M} = \mathcal{M}_1 \cup \cdots \cup \mathcal{M}_K$, where $\mathcal{M}_i \cap \mathcal{M}_j = \emptyset, i \neq j$, be the multiclass LAD model obtained by Algorithm 2.

Algorithm 2: Relaxed Multiclass LAD Algorithm

Input: p: index of current class
Global data: $\Omega \subset \mathbb{B}^{m\times n}$: *binary dataset,* $\mathcal{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$: *set of classes*
Result: MyPats[p] : patterns for class $\mathcal{C}_p$

$B = [\Omega|\overline{\Omega}]$;
$\mathbf{v} = B\,\mathbf{y}$; (* $\mathbf{y}$ unknown variable *)
MyPats[p] = {};
$\kappa = 0$;
$NewConstraint = \{\}$;
$TotCov = 0$;
while $TotCov < \left|I_p\right|$ **do**
  $\mathcal{R} = \{constraintsfrom : \text{Problem}(16)\} \cup NewConstraint$;
  $pat = Minimize\left[cd + \sum_{i\in I_p} \rho_i : \mathcal{R} \text{ and } \mathbf{v}, \mathbf{y}, \mathbf{w}, d \in \mathbb{Z}\right]$
  $\mathbf{y}^*$ part of pat corresponding to variables $\mathbf{y}$;
  for $i = 1$ **to** m **do**
    if $v_i = d$ **then**
      $\kappa_i = \kappa_i + 1$;
  $TotCov = 0$;
  for $i = 1$ **to** m **do**
    if $(i \in I_p) \wedge (\kappa_i \neq 0)$ **then**
      $TotCov = TotCov + 1$;
  $NotFound = True$;
  for $i = 1$ **to** m **do**
    if $(i \in I_p) \wedge (\kappa_i = 0) \wedge (v_i < d) \wedge (NotFound)$ **then**
      $NewConstraint = \{v_i = d\}$;
      (* d and Y as unknown variables *)
      $NotFound = False$;
  MyPats[p] = MyPats[p] $\cup \{\mathbf{y}^*\}$;
return MyPats[p];

The final step of our proposed multiclass LAD method with relaxed patterns is to validate and use the model $\mathcal{M}$ for prediction of new observations. Similar to the two-class classification problem, the accuracy of a multiclass model $\mathcal{M}$ is estimated by the classical cross-validation procedures (Aggarwal 2015; Dietterich 1998; Efron and Tibshirani 1986; Hastie et al. 2005). If an external dataset (test set) is available, the performance of the model can be evaluated on that set and the accuracy of the model is reported as the accuracy on the test set.

3.4 OvR multiclass LAD discriminant and prediction

Given a $K-$class dataset $\Omega = \Omega_1 \cup \cdots \cup \Omega_K$ and a corresponding multiclass LAD model $\mathcal{M} = \mathcal{M}_1 \cup \cdots \cup \mathcal{M}_K$, $(\mathcal{M}_i \cap \mathcal{M}_j = \emptyset, i \neq j)$, the classification of a new (or unseen) observation $\mathbf{o} \in \mathbb{B}^n$, $\mathbf{o} \notin \Omega$ is determined by the value of the discriminant function

$$\Delta(\mathbf{o}) = \arg\max_k \Delta_k(\mathbf{o}) \tag{20}$$

where

$$\Delta_k(\mathbf{o}) = \sum_{P_{\mathcal{C}_k} \in \mathcal{M}_k} \delta_k P_{\mathcal{C}_k}(\mathbf{o}), \quad k = 1, \cdots, K$$

Table 1 Five multiclass datasets from UCI repository

Dataset	Number of observations in class C_i	Number of features
Iris	$\lvert C_1\rvert = \lvert C_2\rvert = \lvert C_3\rvert = 50$	4
Glass ID	$\lvert C_1\rvert = 69, \lvert C_2\rvert = 76, \lvert C_3\rvert = 17$	10
	$\lvert C_4\rvert = 13, \lvert C_5\rvert = 9, \lvert C_6\rvert = 29$	
Wine	$\lvert C_1\rvert = 9, \lvert C_2\rvert = 71, \lvert C_3\rvert = 48$	12
E. Coli	$\lvert C_1\rvert = 143, \lvert C_2\rvert = 77, \lvert C_3\rvert = 52$	34
	$\lvert C_4\rvert = 35, \lvert C_5\rvert = 20, \lvert C_6\rvert = 5$	
Dermatology	$\lvert C_1\rvert = 112, \lvert C_2\rvert = 61, \lvert C_3\rvert = 72$	19
	$\lvert C_4\rvert = 49, \lvert C_5\rvert = 52, \lvert C_6\rvert = 20$	

and $\delta_k \geq 0$ are the weights assigned to patterns $P_{\mathcal{C}_k} \in \mathcal{M}_k, k = 1, \ldots, K$ and can be calculated in various ways. One possibility is to use the prevalence of patterns that is defined by

$$\delta_k = \frac{1}{|\Omega_k|} \sum_{i \in I_{\mathcal{C}_k}} P_{\mathcal{C}_k}(\omega_i)$$

where $\Omega_k \subset \Omega$ is the set of observations in class $\mathcal{C}_k$ and $I_{\mathcal{C}_k} = \{i : \omega_i \in \Omega_k\}, k = 1, \ldots, K$. If $\Delta(\mathbf{o}) = \Delta_p(\mathbf{o}) = \Delta_q(\mathbf{o})$ for some $p \neq q$, then the observation $\mathbf{o}$ is *unclassified*.

4 Experiments

In this section we present experimental results on publicly available datasets to show how Algorithm 2 described in Sect. 3.3 can be used for multiclass classification. Regarding the stopping criterion, Algorithm 2 ends once all patterns for each class $\mathcal{C}_k, k = 1, \ldots, K$, have been computed, i.e., all observations are covered. In the worst case, an adhoc pattern can be built by the algorithm to cover a single observation. In regards to our relaxed MILP LAD algorithm, the experiments are implemented through Python 3.5.7 on a GPU machine, containing the virtual environment and all necessary packages. The specifications for the machine include an Intel CORE i7-6700 CPU with 64 GB Memory running Linux. The goal in these experiments is not only to compare our new LAD based multiclass method with the prior multiclass LAD techniques in LAD literature, but also with the well-known and commonly used supervised learning methods in machine learning literature.

4.1 Experimental results

In order to test our proposed multiclass LAD methodology, we conduct experiments on five multiclass datasets from UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/). Table 1 summarizes the characteristics of these datasets.

We apply our proposed "Relaxed Multiclass LAD Method" (in what follows referred as Relaxed MC-LAD) to these five datasets and report the average sensitivities of 10×10-folding cross-validation experiments in Table 2.

Table 2 Relaxed MC-LAD method: average sensitivity of 10 × 10 cross-validation experiments

Dataset	C_1 (%)	C_2 (%)	C_3 (%)	C_4 (%)	C_5 (%)	C_6 (%)
Iris	90	100	100			
Glass ID	89	80	30	86	89	86
Wine	98	94	88			
E. Coli	95	84	60	70	60	85
Dermatology	95	98	93	100	92	95

Table 3 Relaxed multiclass LAD pattern characteristics for Iris dataset

Iris dataset	# of Patterns	Patterns by relaxed MC-LAD					
		Degree of patterns			Coverage of patterns		
Class		Min	Average	Max	Min (%)	Average (%)	Max (%)
Setosa	1	1	1.00	1	100	100	100
Verisicolor	3	3	3.00	3	5.26	40.35	92.11
Virginica	3	2	2.00	2	35.14	53.15	89.19

4.2 Pattern characteristics

The Relaxed MC-LAD pattern characteristics for all five datasets are shown in Tables 3, 4, 5, 6 and 7. As can be seen from these results, the Relaxed MC LAD method produces accurate classification models with high quality patterns.

4.2.1 Relaxed multiclass LAD pattern characteristics for Iris dataset

Table 3 shows the pattern characteristics for Iris dataset, where the degree of the patterns ranges from one to three and patterns exhibit coverage varying from 5.26 to 100%. Note that for class Setosa, our relaxed multiclass LAD approach produces a single pattern of degree 1 that covers all observations in class Setosa. The relaxed multiclass LAD model contains seven patterns, one for class Setosa, three for class Verisicolor, and three for class Virginica.

4.2.2 Relaxed multiclass LAD pattern characteristics for Wine dataset

Our relaxed multiclass LAD approach produces a LAD model consisting of nine patterns, two for class A, five for class B, and two for class C. Table 4 shows the pattern characteristics for Wine dataset, where the degree of the patterns ranges from 2 to 5 and patterns exhibit coverage varying from 2.82 to 94.91%.

4.2.3 Relaxed multiclass LAD pattern characteristics for Glass dataset

Glass dataset is a noisy dataset. Our relaxed multiclass LAD approach produces a LAD model consisting of 51 patterns, fifteen for class A, sixteen for class B, nine for C, two for D, four for E, and five for class E. Table 5 shows the pattern characteristics for Glass dataset, where the degree of the patterns ranges from 1 to 5 and patterns exhibit coverage varying from 1.96 to 90.00%.

Table 4 Relaxed multiclass LAD pattern characteristics for Wine dataset

Wine dataset	# of Patterns	Patterns by relaxed MC-LAD					
		Degree of patterns			Coverage of patterns		
Class		Min	Average	Max	Min (%)	Average (%)	Max (%)
A	2	4	4.00	4	28.81	61.86	94.92
B	5	5	5.00	5	2.82	31.83	80.28
C	2	2	2.00	2	35.42	63.54	91.67

Table 5 Relaxed multiclass LAD pattern characteristics for Glass dataset

Glass dataset	# of Patterns	Patterns by relaxed MC-LAD					
		Degree of patterns			Coverage of patterns		
Class		Min	Average	Max	Min (%)	Average (%)	Max (%)
A	15	1	2.86	4	1.96	9.28	25.49
B	16	2	3.14	5	3.51	7.46	21.05
C	9	2	3.00	4	7.96	11.97	23.08
D	2	2	3.00	4	10.00	50.00	90.00
E	4	2	3.00	4	14.29	28.57	42.86
F	5	2	2.80	4	4.55	21.82	68.18

Table 6 Relaxed multiclass LAD pattern characteristics for *E. Coli* dataset

E. Coli dataset	# of Patterns	Patterns by relaxed MC-LAD					
		Degree of patterns			Coverage of patterns		
Class		Min	Average	Max	Min (%)	Average (%)	Max (%)
cp	11	3	5.27	8	1.40	16.34	39.16
im	10	4	7.50	11	1.30	11.56	42.86
pp	7	4	6.86	9	1.92	22.25	46.15
imU	7	3	5.86	9	2.86	13.88	42.86
om	4	4	4.00	4	5.00	31.25	80.00
omL	1	3	3.00	3	100	100	100

4.2.4 Relaxed multiclass LAD pattern characteristics for *E. Coli* dataset

Table 6 shows the pattern characteristics for *E. Coli* dataset, where the patterns are generated by our multiclass LAD approach. Note that this model contains more complicated patterns with degrees, ranging from 3 to 11 and coverage, ranging from 1.4 to 100%. The multiclass LAD model contains a total of 40 patterns. Note that only one pattern of degree three is sufficient to cover all observations in class omL.

4.2.5 Relaxed multiclass LAD pattern characteristics for Dermatology dataset

Table 7 shows the pattern characteristics for Dermatology dataset, where the patterns are generated by our multiclass LAD approach. Similar to *E. Coli* LAD model, the multiclass

Table 7 Relaxed multiclass LAD pattern characteristics for Dermatology dataset

Dermatology dataset	# of Patterns	Patterns by relaxed MC-LAD					
		Degree of patterns			Coverage of patterns		
Class		Min	Average	Max	Min (%)	Average (%)	Max (%)
A	4	2	2.75	3	13.25	53.61	79.52
B	6	2	3.33	5	2.22	25.93	57.78
C	2	1	1.00	1	92.45	92.45	92.45
D	7	3	3.86	5	2.78	18.65	72.22
E	2	2	4.00	6	41.67	59.72	77.78
F	2	2	2.50	3	33.33	63.33	93.33

Table 8 Classification accuracy (%) of LAD based multiclass methods on Iris, Wine, and Glass datasets

Methods	Iris	Wine	Glass
Relaxed MC-LAD	97.03 ± 1.90	94.67 ± 2.14	80.37 ± 4.87
Kim and Choi (2015)-OvR	94.80 ± 0.40	96.18 ± 1.76	96.26 ± 1.06
Kim and Choi (2015)-OvO	95.73 ± 0.53	96.86 ± 1.48	93.46 ± 1.48
Moreira (2000)-OvO	n.a.	92.70 ± 2.54	62.41 ± 5.88
Mortada (2010)-OvO	n.a.	93.10 ± 3.20	65.00 ± 5.40
Avila-Herrera and Subasi (2013)-OvR	94.00 ± 2.20	91.33 ± 3.54	79.54 ± 5.35

Table 9 Classification accuracy (%) of LAD based multiclass methods on *E. Coli* and Dermatology datasets

Methods	*E. Coli*	Dermatology
Relaxed MC-LAD	82.50 ± 5.79	96.06 ± 2.85
Kim and Choi (2015)-OvR	75.07 ± 0.60	92.18 ± 0.43
Kim and Choi (2015)-OvO	81.88 ± 1.00	96.26 ± 1.39
Moreira (2000)-OvO	78.34 ± 3.40	89.07 ± 2.84
Mortada (2010)-OvO	79.20 ± 5.35	n.a.
Avila-Herrera and Subasi (2013)-OvR	n.a.	92.00 ± 2.14

LAD model for Dermatology dataset contains more complicated patterns with degrees, ranging from 1 to 6 and coverage, ranging from 2.22 to 92.45%. Note that the method produced two degree one patterns associated with class C and these two patterns cover 92.45% of the observations in class C.

4.3 Experiments comparing LAD based multiclass classification methods

In this section we report the average accuracy of 10×10-folding experiments, where we compare Relaxed MC-LAD method against other LAD based multiclass methods of Kim and Choi (2015)-OvR, Kim and Choi (2015)-OvO, Moreira (2000)-OvO, Mortada (2010)-OvO, and Avila-Herrera and Subasi (2013)-OvR. These results are summarized in Table 8 for Iris, Wine, and Glass datasets and in Table 9 for *E. Coli* and Dermatology datasets.

Table 10 Comparison of average CPU time for LAD based multiclass methods

Methods	Average CPU time (in s)				
	Iris	Wine	Glass	*E. Coli*	Dermatology
Relaxed MC-LAD	0.24	4.4	37.8	18.3	29.1
Kim and Choi (2015)-OvR	15.1	16.9	24.9	31.4	21.4
Kim and Choi (2015)-OvO	21.1	22.8	73.9	129.4	74.2
Moreira (2000)-OvO	n.a	1.2	12.7	17.5	10.1
Mortada (2010)-OvO	n.a	1.0	39.0	49.0	n.a
Avila-Herrera and Subasi (2013)-OvR	16.70	236.5	1578	n.a	41.0

The average CPU times for LAD based multiclass methods, including our proposed method Relaxed MC-LAD, are given in Table 10. Note that the efficiency of our proposed Relaxed MC-LAD Method is comparable if not better than those other existing multiclass LAD methods.

As compared to the previously proposed LAD multiclass approaches, our Relaxed MC-LAD method is more efficient and performs better (higher accuracy) than those of Moreira (2000)-OvO, Mortada (2010)-OvO and Avila-Herrera and Subasi (2013)-OvR for all five datasets. The Relaxed MC-LAD method has smaller CPU times than those reported by the methods of Kim and Choi (2015) for almost all of the datasets, except it is slightly slower than Kim and Choi (2015)-OvR for the Glass dataset. We also observe that our proposed Relaxed MC-LAD method's performance in terms of classification accuracy is very comparable to, if not better than, the performance of the methods of Kim and Choi (2015) for all datasets, except for Glass dataset. However, for Glass dataset, the Relaxed MC-LAD method outperforms the previously developed multiclass LAD methods, excluding the methods of Kim and Choi (2015), as can be seen from Tables 8 and 9.

4.4 Experiments comparing relaxed multiclass LAD method with well-known classification techniques

Similar to the previous section, here we report the average accuracy of 10×10-folding experiments of our Relaxed MC-LAD method as compared to the well-known classification techniques, including Nearest Neighbor, Naïve Bayes, Logistic Regression, Support Vector Machines, Neural Networks, and Decision Trees-C4.5. These experiments were run in WEKA 3.8 Data Mining Software (Frank et al. 2016).

Table 11 shows the average accuracy of the cross-validation experiments for Iris, Wine, and Glass datasets and Table 12 shows the average accuracy of the cross-validation experiments for *E. Coli* and Dermatology datasets.

We observe from Tables 11 and 12 that our proposed Relaxed MC-LAD method's classification accuracy is very comparable to, if not better than, that of the well-known supervised learning methods including Nearest Neighbor, Naïve Bayes, Logistic Regression, Neural Networks, and Decision Trees for all datasets, including Glass data where the performance of all methods is much worse ($\geq$ 10% smaller accuracy) than the Relaxed MC-LAD method. While all of the multiclass methods, including both LAD based and others, but excluding the Relaxed MC-LAD method, poorly classify the observations in Glass dataset, the methods of Kim and Choi (2015) do outperform all of the methods by about more than 10% improvement in accuracy based on the results reported in their paper.

Table 11 Comparison of average accuracy of well-known classification methods for Iris, Wine, and Glass datasets

Methods	Average accuracy (%) of cross-validation experiments		
	Iris	Wine	Glass
Relaxed MC-LAD	97.03 ± 1.90	94.67 ± 2.14	80.37 ± 4.87
Nearest neighbor	95.40 ± 4.80	95.12 ± 4.34	70.30 ± 8.96
Naïve Bayes	95.53 ± 5.02	97.46 ± 3.70	47.75 ± 9.36
Logistic regression	97.07 ± 4.77	97.23 ± 3.83	63.92 ± 8.81
Support vector machines	96.27 ± 4.58	98.76 ± 2.73	57.72 ± 9.06
Neural networks	96.93 ± 4.07	98.02 ± 3.26	65.96 ± 9.11
Decision trees-C4.5	94.64 ± 5.78	89.90 ± 3.11	62.80 ± 4.43

Table 12 Comparison of average accuracy of well-known classification methods for *E. Coli* and Dermatology datasets

Methods	Average accuracy (%) of cross-validation experiments	
	E. Coli	Dermatology
Relaxed MC-LAD	82.50 ± 5.79	96.06 ± 2.85
Nearest neighbor	81.99 ± 6.35	95.60 ± 3.16
Naïve Bayes	87.01 ± 6.23	97.52 ± 2.46
Logistic regression	86.64 ± 5.54	97.24 ± 2.65
Suport vector machines	84.43 ± 4.84	97.59 ± 2.37
Neural networks	87.14 ± 6.54	97.32 ± 2.61
Decision trees-C4.5	80.59 ± 4.14	94.48 ± 2.69

Our numerical experiments suggest that the proposed Relaxed MC-LAD method is an exciting alternative to the multiclass classification literature. The method not only provides efficient and robust results, but also easily interpretable explicit classification models—an essential feature of LAD methodology.

5 Conclusions

In this paper we extend Logical Analysis of Data (LAD) to multiclass classification where relaxed (fuzzy) patterns are generated as optimal solutions of a mixed integer linear programming problem. Our proposed relaxed multiclass LAD approach is motivated by MILP formulation of Ryoo and Jang (2009) that generates LAD patterns for two-class datasets and LAD based multiclass algorithm of Avila-Herrera and Subasi (2013, 2015) that generated pure patterns with minimum degree and maximum coverage. Our multiclass method uses homogeneity and prevalence as two parameters to generate relaxed LAD patterns, aimed at improving the generalization capability. While all of the multiclass methods, including both LAD based and others, but excluding the Relaxed MC-LAD method, poorly classify the observations in Glass dataset, the methods of Kim and Choi (2015) do outperform all of the methods by about more than 10% improvement in accuracy based on the results reported in their paper. All these results suggest that our proposed Relaxed Multiclass LAD method is an exciting alternative to the multiclass classification literature. We demonstrate the advantage

of having the flexibility proposed in our method through experiments on five benchmark multiclass datasets. Experimental results show that the proposed relaxed multiclass LAD algorithm produces highly accurate classification models on the benchmark datasets, where the cross-validation accuracy of the relaxed multiclass LAD algorithm is comparable to, if not better than, those obtained by previously developed multiclass LAD classification methods well as those by the well-known classification techniques, including Nearest Neighbor, Naïve Bayes, Logistic Regression, Support Vector Machines, Neural Networks, and Decision Trees. In addition, our proposed relaxed multiclass LAD method is very efficient as can be seen from the reported CPU time of the experiments.

References

Aggarwal, C. C. (2015). *Data mining*. Berlin: Springer.

Aiolli, F., & Sperduti, A. (2005). Multiclass classification with multi-prototype support vector machines. *Journal of Machine Learning Research*, *6*(1), 817–850.

Alexe, G., Alexe, S., Axelrod, D. E., Bonates, T., Lozina, I., Reiss, M., et al. (2006). Breast cancer prognosis by combinatorial analysis of gene expression data. *Breast Cancer Research*, *8*(4), R41.

Alexe, G., Alexe, S., Axelrod, D. E., Hammer, P. L., & Weissmann, D. (2005). Logical analysis of diffuse large B-cell lymphomas. *Artificial Intelligence in Medicine*, *34*(3), 235–267.

Alexe, G., Alexe, S., Bonates, T. O., & Kogan, A. (2007). Logical analysis of data—the vision of Peter L. Hammer. *Annals of Mathematics and Artificial Intelligence*, *49*(1–4), 265–312.

Alexe, G., Alexe, S., Liotta, L. A., Petricoin, E., Reiss, M., & Hammer, P. L. (2004). Ovarian cancer detection by logical analysis of proteomic data. *Proteomics*, *4*(3), 766–783.

Alexe, G., & Hammer, P. (2006). Spanned patterns for the logical analysis of data. *Discrete Applied Mathematics*, *154*(7), 1039–1049.

Alexe, S., Blackstone, E., Hammer, P. L., Ishwaran, H., Lauer, M. S., & Snader, C. E. P. (2003). Coronary risk prediction by logical analysis of data. *Annals of Operations Research*, *119*(1–4), 15–42.

Aly, M. (2005). Survey on multiclass classification methods. In *Neural Networks* (pp. 1–9).

Apté, C., Damerau, F., & Weiss, S. M. (1994). Automated learning of decision rules for text categorization. *ACM Transactions on Information Systems (TOIS)*, *12*(3), 233–251.

Avila-Herrera, J. F., & Subasi, M. M. (2013). Logical analysis of multiclass data. In *RUTCOR research reports, RRR 5-2013*.

Avila-Herrera, J. F., & Subasi, M. M. (2015). Logical analysis of multiclass data. In *Proceedings of the 2015 Latin American computing conference* (pp. 1–10). IEEE.

Beygelzimer, A., Langford, J., & Ravikumar, P. (2007). Multiclass classification with filter trees.

Bishop, C. M. (2007). *Pattern recognition and machine learning*. Berlin: Springer.

Boland, C. R., Thibodeau, S. N., Hamilton, S. R., Sidransky, D., Eshleman, J. R., Burt, R. W., et al. (1998). A National Cancer Institute Workshop on Microsatellite Instability for cancer detection and familial predisposition: Development of international criteria for the determination of microsatellite instability in colorectal cancer. *Cancer Research*, *58*(22), 5248–5257.

Bonates, T. O., & Hammer, P. L. (2007). Pseudo-Boolean regression. In *RUTCOR research report*, Vol. **3-2007**.

Bonates, T. O., Hammer, P. L., & Kogan, A. (2008). Maximum patterns in datasets. *Discrete Applied Mathematics*, *156*(6), 846–861.

Boros, E., Hammer, P. L., Ibaraki, T., & Kogan, A. (1997). Logical analysis of numerical data. *Mathematical Programming*, *79*(1), 163–190.

Boros, E., Hammer, P. L., Ibaraki, T., Kogan, A., Mayoraz, E., & Muchnik, I. (2000). An implementation of logical analysis of data. *IEEE Transactions on Knowledge and Data Engineering*, *12*(2), 292–306.

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, *2*(2), 121–167.

Chapelle, O., Haffner, P., & Vapnik, V. (1999). Support vector machines for histogram-based image classification. *IEEE Transactions on Neural Networks and Learning Systems*, *10*(5), 1055–1064.

Crama, Y., Ibaraki, T., & Hammer, P. L. (1988). Cause-effect relationships and partially defined boolean functions. *Annals of Operations Research*, *16*(1–4), 299–325.

Daniely, A., Sabato, S., & Shalev-Shwartz, S. (2012). Multiclass learning approaches: A theoretical comparison with implications. In *Neural information processing systems*.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, *10*(7), 1895–1923.
Ding, C. H. Q., & Dubchak, I. (2001). Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics*, *17*(4), 349–358.
Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification*. Hoboken: Wiley.
Dupuis, C., Gamache, M., & Pagé, J. F. (2012). Logical analysis of data for estimating passenger show rates at air canada. *Journal of Air Transport Management*, *18*(1), 78–81.
Efron, B., & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, *1*(1), 54–75.
Even-Zohar, Y., & Roth, D. (2001). A sequential model for multi-class classification. In *EMNLP-2001, the SIGDAT conference on empirical methods in natural language processing* (pp. 10–19).
Fausett, L. V. (1994). *Fundamentals of neural networks: Architectures, algorithms, and applications*. Englewood Cliffs, NJ: Prentice-Hall.
Frank, E., Hall, M. A., & Witten, I. H. (2016). The WEKA Workbench. In *Online Appendix for "Data mining: Practical machine learning tools and techniques"*, 4th edn.
Friedman, N., Linial, M., Nachman, I., & Peer, D. (2000). Using Bayesian networks to analyze expression data. *Journal of Computaional Biology*, *7*(3–4), 601–620.
Galar, M., Fernández, A., Barrenechea, E., Bustince, H., & Herrera, F. (2011). An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on One-vs-One and One-vs-All schemes. *Pattern Recognition*, *44*(8), 1761–1776.
Gehler, P., & Nowozin, S. (2009). On feature combination for multiclass object classification. In *2009 IEEE 12th international conference on computer vision* (pp. 221–228). IEEE.
Ghasemi, A., Esmaeili, S., & Yacout, S. (2013). Development of equipment failure prognostic model based on logical analysis of data (LAD). *Engineering Letters*, *21*(4), 256–263.
Guo, C., & Ryoo, H. S. (2012). Compact MILP models for optimal and pareto-optimal LAD patterns. *Discrete Applied Mathematics*, *160*(16–17), 2339–2348.
Hammer, P. L. (1986). Partially defined Boolean functions and cause-effect relationships. In *International conference on multi-attribute decision making via OR-based expert systems*.
Hammer, P. L., & Bonates, T. O. (2006). Logical analysis of data—An overview: From combinatorial optimization to medical applications. *Annals of Operations Research*, *148*(1), 203–335.
Hammer, A. B., Hammer, P. L., & Muchnik, I. (1999). Logical analysis of chinese labor productivity patterns. *Annals of Operations Research*, *87*, 165–176.
Hammer, P. L., Kogan, A., & Lejeune, M. A. (2011). Reverse engineering country risk ratings: Statistical and combinatorial non-recursive models. *Annals of Operations Research*, *188*(1), 185–213.
Hammer, P., Kogan, A., Simeone, B., & Szedmák, S. (2004). Pareto-optimal patterns in logical analysis of data. *Discrete Applied Mathematics*, *144*(1), 79–102.
Hanash, S., & Creighton, C. (2003). Making sense of microarray data to classify cancer. *The Pharmacogenomics Journal*, *3*, 308–311.
Har-Peled, S., Roth, D., & Zimak, D. (2002). Constraint classification: A new approach to multiclass classification. In *International conference on algorithmic learning theory* (pp. 365–379). Springer. https://doi.org/10.1109/ICCV.2009.5459169.
Hastie, T., & Tibshirani, R. (1998). Classification by pairwise coupling. *The Annals of Statistics*, *26*(2), 451–471.
Hastie, T., Tibshirani, R., Friedman, J., & Franklin, J. (2005). The elements of statistical learning: Data mining, inference and prediction. *The Mathematical Intelligencer*, *27*(2), 83–85.
Jelinek, F. (1998). *Statistical methods for speech recognition*. Cambridge: The MIT Press.
Kim, H. H., & Choi, J. Y. (2015). Pattern generation for multi-class LAD using iterative genetic algorithm with flexible chromosomes and multiple populations. *Expert Systems with Applications*, *42*(2), 833–843.
Kotsiantis, S., & Kanellopoulus, D. (2006). Discretization techniques: A recent survey. *GESTS International Transactions on Computer Science and Engineering*, *32*(1), 47–58.
Kronek, L. P., & Reddy, A. (2008). Logical analysis of survival data: Prognostic survival models by detecting high degree interactions in right-censored data. *Bioinformatics*, *24*(16), i248–253.
Lauer, M. S., Alexe, S., Pothier-Snader, C. E., Blackstone, E. H., Ishwaran, H., & Hammer, P. L. (2002). Use of the logical analysis of data method for assessing long-term mortality risk after exercise electrocardiography. *Circulation*, *106*(6), 685–690.
LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., et al. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, *1*(4), 541–551.
Lee, D. D., & Seung, H. S. (1997). Unsupervised learning by convex and conic coding. In *Advances in neural information processing systems* (pp. 515–521).

Lejeune, M., Lozin, V., Lozina, I., Ragab, A., & Yacout, S. (2018). Recent advances in the theory and practice of logical analysis of data. *European Journal of Operational Research*, *275*, 1–15. https://doi.org/10.1016/j.ejor.2018.06.011.

Lejeune, M. A., & Margot, F. (2011). Optimization for simulation: LAD accelerator. *Annals of Operations Research*, *188*(1), 285–305.

Lemaire, P. (2011). Extensions of logical analysis of data for growth hormone deficiency diagnoses. *Annals of Operations Research*, *186*(1), 199–211.

Liu, D., Yan, S., Mu, Y., Hua, X., Chang, S., & Zhang, H. (2011). Towards optimal discriminating order for multiclass classification. In *2011 IEEE 11th international conference on data mining (ICDM)* (pp. 388–397). IEEE.

Liu, H., Hussain, F., Tan, C. L., & Dash, M. (2002). Discretization: An enabling technique. *Data Mining and Knowledge Discovery*, *6*, 393–423.

Li, T., Zhu, S., & Ogihara, M. (2006). Using discriminant analysis for multi-class classification: An experimental investigation. *Knowledge and Information Systems*, *10*(4), 453–472.

Misselwitz, B., Strittmatter, G., Periaswamy, B., Schlumberger, M. C., Rout, S., Horvath, P., et al. (2010). Enhanced cell classifier: A multi-class classification tool for microscopy images. *BMC Bioinformatics*, *11*, 30.

Moreira, L. (2000). *The use of Boolean concepts in general classification contexts*. Ph.D. Thesis, Universidade do Minho, Portugal.

Mortada, M. (2010). *Applicability and interpretability of logical analysis of data in condition based maintenance*. Ph.D. Thesis, École Polytechnique de Montréal, Canada.

Mortada, M. A., Yacout, S., & Lakis, A. (2011). Diagnosis of rotor bearings using logical analysis of data. *Journal of Quality in Maintenance Engineering*, *17*(4), 371–397.

Mortada, M. A., Yacout, S., & Lakis, A. (2014). Fault diagnosis in power transformers using multi-class logical analysis of data. *Journal of Intelligent Manufacturing*, *25*(6), 1429–1439.

Nakagawa, T., Kudo, T., & Matsumoto, Y. (2002). Revision learning and its application to part-of-speech tagging. In *Proceedings of the 40th annual meeting of the association for computational linguistics* (pp. 497–504).

Platt, J. C., Cristianini, N., & Shawe-Taylor, J. (2000). Large margin DAGs for multiclass classification. *Advances in Neural Information Processing Systems*, *12*(3), 547–553.

Reddy, A., Brannon, A. R., Seiler, M., Irgon, J., Ljungberg, B., Zhao, H., Brooks, J. D., Ganesan, S., Rathmell, W. K., & Bhanot, G. (2009). A predictor for survival in intermediate grade clear cell renal cell carcinoma. In *BIOCOMP*.

Reddy, A., Wang, H., Yu, H., Bonates, T. O., Gulabani, V., Azok, J., et al. (2008). Logical analysis of data (LAD) model for the early diagnosis of acute ischemic stroke. *BMC Medical Informatics and Decision Making*, *8*, 30.

Ryoo, H. S., & Jang, I. Y. (2009). MILP approach to pattern generation in logical analysis of data. *Discrete Applied Mathematics*, *157*(4), 749–761.

Schölkopf, B., & Smola, A. J. (2001). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. Cambrige, MA: The MIT Press.

Singh-Miller, N., & Collins, M. (2009). Learning label embeddings for nearest-neighbor multi-class classification with an application to speech recognition. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, A. Culotta (Eds.), *Advances in neural information processing systems* (Vol. 22, pp. 1678–1686).

Subasi, E., Subasi, M. M., Hammer, P. L., Roboz, J., Anbalagan, V., & Lipkowitz, M. S. (2017). A classification model to predict the rate of decline of kidney function. *Frontiers in Medicine*, *4*, 97.

Tax, D. M. J., & Duin, R. P. W. (2002). Using two-class classifiers for multiclass classification. In *Proceedings of 16th international conference on pattern recognition* (Vol. 2, pp. 124–127). IEEE.

Tewari, A., & Bartlett, P. L. (2007). On the consistency of multiclass classification methods. *Journal of Machine Learning Research*, *8*, 1007–1025.

Üney, F., & Türkay, M. (2006). A mixed-integer programming approach to multi-class data classification problem. *European Journal of Operational Research*, *173*(3), 910–920.

Wu, T. F., Lin, C. J., & Weng, R. C. (2004). Probability estimates for multi-class classification by pairwise coupling. *The Journal of Machine Learning Research*, *5*, 975–1005.

Yang, J. B., & Tsang, I. W. (2012). Hierarchical maximum margin learning for multi-class classification. Preprint arXiv:1202.3770.