Information Sciences 179 (2009) 4089–4096

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Loss optimal monotone relabeling of noisy multi-criteria data sets

Michaël Rademaker [a,*], Bernard De Baets [a,**], Hans De Meyer [b]

[a] *Department of Applied Mathematics, Biometrics and Process Control, Ghent University, Coupure links 653, B-9000 Gent, Belgium*
[b] *Department of Applied Mathematics and Computer Science, Ghent University, Krijgslaan 281 S9, B-9000 Gent, Belgium*

ARTICLE INFO

Article history:
Received 22 September 2008
Received in revised form 27 February 2009
Accepted 14 August 2009

Keywords:
Machine learning
Monotonicity constraints
Graph algorithms
Maximum independent set
Monotone relabeling
Data cleaning
Multi-criteria decision making

ABSTRACT

A method to relabel noisy multi-criteria data sets is presented, taking advantage of the transitivity of the non-monotonicity relation to formulate the problem as an efficiently solvable maximum independent set problem. A framework and an algorithm for general loss functions are presented, and the flexibility of the approach is indicated by some examples, showcasing the ease with which the method can handle application-specific loss functions. Both didactical examples and real-life applications are provided, using the zero-one, the L1 and the squared loss functions, as well as combinations thereof.

1. Introduction

We consider the problem of noise in the form of non-monotonicity. Three options present themselves given a data set subject to such noise: keep the data set as it is, identify the noisy objects and remove them from the data set, or identify the noisy objects and relabel them. It is this last option we will discuss here, for the specific case where an integer-valued loss function is defined on the labels. Popular loss functions are the zero-one and L1 loss functions, for which we provide some examples. The technique we describe is more general however: a loss function need not be a metric or distance (the squared loss function, for example, is indeed not a metric). We use the maximum independent set concept to facilitate monotone relabeling. Of key importance is the transitivity of the non-monotonicity relation, permitting the formulation of the problem as one solvable by network flow algorithms and the design of loss optimal monotone relabeling algorithms.

The remainder of this paper is constructed as follows: Section 2 contains the required introduction to the notation and a short mention of the problems non-monotone multi-criteria data pose. Next, we provide a brief introduction to the maximum independent set problem in Section 3, followed by an in-depth description of how we are able to translate the relabeling problem to a maximum independent set problem in Section 4, where we also supply an algorithm which compares favorably to the existing methods, as these are less flexible and/or computationally less efficient. The computational complexity is discussed in Section 5, and some options for fine-tuning are given in Section 6. We provide some example applications to toy problems and a real-life data set in Section 7, followed by our conclusions in Section 8.

* Corresponding author. Tel.: +32 9 2645941; fax: +32 9 264 60 20.
** Corresponding author.
E-mail addresses: michael.rademaker@ugent.be (M. Rademaker), bernard.debaets@ugent.be (B. De Baets).

0020-0255/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.ins.2009.08.018

2. Problem setting

We consider a data set $\langle\mathscr{S},\mathscr{X},f\rangle$, where $\mathscr{S}$ is a set of N objects from an object space Ω, represented in an n-dimensional vector space $\mathscr{X}$ and a labeling function $f:\mathscr{S}\to\mathscr{L}$, assigning labels from a label set $\mathscr{L}$. Objects are described by a set of n criteria (features) $q_i:\Omega\to\mathscr{X}_i$ making up the feature space $\mathscr{X}$, in which an object a is represented by a feature vector $\mathbf{a}=(q_1(a),\ldots,q_n(a))$. As is common for multi-criteria data sets [4], for each feature q_i we have a linear order relation $\leqslant_{\mathscr{X}_i}$, the combination of which yields the natural product ordering $\leqslant_{\mathscr{X}}$. The set of labels $\mathscr{L}=\{\ell_1,\ldots,\ell_m\}$ is equipped with a strict linear order relation $<_{\mathscr{L}}:\ell_1<_{\mathscr{L}}\ell_2<_{\mathscr{L}}\cdots<_{\mathscr{L}}\ell_{m-1}<_{\mathscr{L}}\ell_m$.

Two objects a and b from $\langle\mathscr{S},\mathscr{X},f\rangle$ can now be compared on the basis of their feature vectors $\mathbf{a}$ and $\mathbf{b}$, or their labels $f(a)$ and $f(b)$. We call the labeling function f monotone (w.r.t. $\mathscr{X}$), if for all objects a and b from $\langle\mathscr{S},\mathscr{X},f\rangle$ it holds that $\mathbf{a}\leqslant_{\mathscr{X}}\mathbf{b}\Rightarrow f(a)\leqslant_{\mathscr{L}}f(b)$ (implying in turn the absence of doubt**:** objects with identical feature vectors should be assigned identical labels). This is the monotonicity requirement [6] extended to safeguard against doubt. A couple of objects (a,b) is said to be a non-monotone couple, denoted by $a\prec b$, if $\mathbf{a}\leqslant_{\mathscr{X}}\mathbf{b}\wedge f(a)>_{\mathscr{L}}f(b)$. The non-monotonicityrelation $\prec$ constitutes a strict partial order relation. Observe that non-monotonicity is defined w.r.t. both $\mathscr{X}$ and f. We call an object monotone if it does not make up a non-monotone couple with any other object, and a data set monotone if it contains no non-monotone objects.

Suppose, for example, that the objects have been labeled by a panel of experts. Each of these experts can have taken different characteristics into account, which may or may not be present in $\mathscr{X}$. This can result in f being non-monotone. The object of this paper is to determine a monotone labeling function f' that differs as little as possible from f. Crucial is then the existence of an object-wise loss function D on the labels, more formally $D:\mathscr{L}\times\mathscr{L}\to\mathbb{N}$. As we are dealing with a discrete set of labels, we feel it is natural to only consider integer-valued loss functions. Some loss functions used in this paper are the zero-one loss $=\min(1,|i-j|)$, the L1 loss $=|i-j|$ and the squared loss $=|i-j|^2$, in which cases $\mathscr{L}$ is identified with the set of first m integers. The loss function need not be a metric: it need not be symmetric, nor does it need to satisfy the triangle inequality (as is the case for the squared loss). We consider a loss function any function $D:\mathscr{L}\times\mathscr{L}\to\mathbb{N}$ satisfying the following three properties for all $\ell_i,\ell_j,\ell_k\in\mathscr{L}$:

(p1) $D(\ell_i,\ell_j)\ \geqslant\ 0$

(p2) $D(\ell_i,\ell_j)=0\iff\ell_i=\ell_j$

(p3) $\ell_i<_{\mathscr{L}}\ell_j<_{\mathscr{L}}\ell_k\vee\ell_i>_{\mathscr{L}}\ell_j>_{\mathscr{L}}\ell_k\Rightarrow D(\ell_i,\ell_j)\leqslant D(\ell_i,\ell_k)\wedge D(\ell_j,\ell_k)\leqslant D(\ell_i,\ell_k)$

The first of these natural conditions (p1) is non-negativity, the second (p2) is known as identity of indiscernibles, and the third (p3) we consider a monotonicity-like property. We will sometimes refer to a strict version of (p3), denoting a version where the inequalities on the right-hand side are strict:

(p4) $\ell_i<_{\mathscr{L}}\ell_j<_{\mathscr{L}}\ell_k\vee\ell_i>_{\mathscr{L}}\ell_j>_{\mathscr{L}}\ell_k\Rightarrow D(\ell_i,\ell_j)<D(\ell_i,\ell_k)\wedge D(\ell_j,\ell_k)<D(\ell_i,\ell_k)$

It is useful to also write $D(f,f')$ as shorthand to denote $\sum_{a\in\mathscr{S}}D(f(a),f'(a))$. Finding a monotone f' that minimizes this expression amounts to identifying a D-optimal monotone relabeling scheme for the data set $\langle\mathscr{S},\mathscr{X},f\rangle$.

The interest in monotone data sets arises from the use of machine learning algorithms that are unable to be trained on partially non-monotone data sets, such as the "Tool for Ordinal Multi-Attribute Sorting and Ordering" (TOMASO) algorithm [12], and the monotone decision tree induction algorithm [15]. Nevertheless, noise is often present in real-life data, which can, because of the monotonicity constraint, be all the more readily apparent in multi-criteria data sets (indeed, Brodley and Friedl [5] mention that "for some learning tasks, domain knowledge exists such that noisy objects can be identified because they go against the laws of the domain"). As an alternative to restricting oneself to algorithms that are able to process partially non-monotone data sets [10,11] when faced with noise, monotone relabeled real-life data sets allow use of any algorithm. Another area of application for monotone data sets is the formulation of monotone benchmark data sets, where monotone (relabeled if needed) data sets could be of more interest than truly random monotone data sets [8].

3. The maximum independent set problem

The independent set concept is relevant to the discussion of non-monotonicity in defining an optimal cleanup [17]. This problem from graph theory deals with a graph G, comprised of a set of vertices $V=V(G)$ and a set of edges $E=E(G)$ (if the edges are directed, they are commonly called arcs). We denote $G=(V,E)$, and for our application, the graph G can be considered a simple directed graph: finite, loopless and without duplicate arcs or vertices. Finding a biggest subset of vertices for which no arc has both the end and startpoint in the subset, is called the maximum independent set problem. In general, multiple maximum independent sets exist for a graph G. The intersection of all such maximum independent sets is called the core, while the union is called the corona [3]. In our application, the set of objects corresponds to the set of vertices V, and the arcs correspond to the non-monotonicity relation $\prec$. A data set $\langle\mathscr{S},\mathscr{X},f\rangle$ is thus monotone if there are no such arcs in its graph representation. Consequently, finding a biggest monotone subset (of a data set), is in fact a maximum independent set problem.