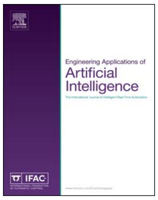
Engineering Applications of Artificial Intelligence 60 (2017) 128–135

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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



# Prototype selection to improve monotonic nearest neighbor

José-Ramón Cano[a,*], Naif R. Aljohani[b], Rabeeh Ayaz Abbasi[b], Jalal S. Alowidbi[c], Salvador García[d]

[a] Department of Computer Science, EPS of Linares, University of Jaén, Campus Científico Tecnológico de Linares, Cinturón Sur S/N, Linares 23700, Jaén, Spain
[b] Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia
[c] Faculty of Computing and Information Technology, University of Jeddah, Jeddah, Saudi Arabia
[d] Department of Computer Science and Artificial Intelligence, CITIC-UGR (Research Center on Information and Communications Technology), University of Granada, ETSII, Calle Periodista Daniel Saucedo Aranda S/N, Granada 18071, Spain

ARTICLE INFO

*Keywords:*
Monotonic classification
Prototype selection
Monotone nearest neighbor
Data reduction
Opinion surveys

ABSTRACT

Student surveys occupy a central place in the evaluation of courses at teaching institutions. At the end of each course, students are requested to evaluate various aspects such as activities, methodology, coordination or resources used. In addition, a final qualification is given to summarize the quality of the course. The prediction of this final qualification can be accomplished by using monotonic classification techniques. The outcome offered by these surveys is particularly significant for faculty and teaching staff associated with the course.

The monotonic nearest neighbor classifier is one of the most relevant algorithms in monotonic classification. However, it does suffer from two drawbacks, (a) inefficient execution time in classification and (b) sensitivity to no monotonic examples. Prototype selection is a data reduction process for classification based on nearest neighbor that can be used to alleviate these problems. This paper proposes a prototype selection algorithm called Monotonic Iterative Prototype Selection (MONIPS) algorithm. Our objective is two-fold. The first one is to introduce MONIPS as a method for obtaining monotonic solutions. MONIPS has proved to be competitive with classical prototype selection solutions adapted to monotonic domain. Besides, to further demonstrate the good performance of MONIPS in the context of a student survey about taught courses.

## 1. Introduction

Classification refers to the problem of predicting the value of a target variable by building a model based on relevant independent input variables Witten et al. (2011). In monotonic classification, the data come from ordered domains Potharst et al. (2009); Gutiérrez et al. (2016); thus, the variable domain is ordered, assuming that the target variable is defined as a monotone function of the describing independent input variables. In addition, it is necessary that the predictions satisfy the monotonicity as it is indicated in Kotlowski and Slowinski (2013), Gutiérrez et al. (2013), Sánchez-Monedero et al. (2014), Gutiérrez and García (2016).

The evaluation of teaching courses based on surveys gathered from students' opinions can be categorized as a monotonic classification problem if it intends to predict a final qualification that summarizes the general quality of the course. The students are asked to evaluate each course according to several aspects related to interest, achieving appropriate class participation, teaching resources, capabilities of the teacher, etc.

The Monotonic Nearest Neighbor classifier (MNN) is one of the most relevant algorithms solving monotonic classification (Duivesteijn and Feelders, 2008). MNN is a nonparametric classifier which uses the entire input data set to establish the monotonic classification rule. Thus, the effectiveness of the classification process performed by MNN depends strongly on the quality of the training data (as in the case of the classical nearest neighbor classification algorithm) (Derrac et al., 2014). The main drawback of MNN is its inefficient execution time making a prediction and low noise tolerance (García et al., 2012). Amongst the most effective techniques for addressing these problems are those that work by preprocessing the data (Cano et al., 2003; García et al., 2015), instead of modifying the computation of the NN rule (MNN rule in this case).

Within data preprocessing, data reduction is widely used. By removing irrelevant data, data reduction can avoid the excessive storage, reducing the execution time of the algorithms, easing and enabling classification techniques to deal with noisy data sets (Cano et al., 2008; García et al., 2008). One of the data reduction techniques

[*] Corresponding author.
*E-mail addresses:* jrcano@ujaen.es (J.-R. Cano), nraljohanig@kau.edu.sa (N.R. Aljohani), frabbasi@kau.edu.sa (R.A. Abbasi), jalowibdi@uj.edu.sa (J.S. Alowidbi), salvagl@decsai.ugr.es (S. García).

http://dx.doi.org/10.1016/j.engappai.2017.02.006
Received 16 May 2016; Received in revised form 30 December 2016; Accepted 9 February 2017
0952-1976/

is the Prototype Selection (PS) family and has shown its valuable capabilities in the past (García et al., 2012).

In this paper, we propose the first monotonic PS algorithm: the Monotonic Iterative Prototype Selection method (MONIPS). We compare it with some of the classical PS solutions which are adapted to be used with monotonicity constraints. This is a novel perspective which is yet to be studied in the monotonic classification discourse. There are two well-representative and known methods for PS: Condensed Nearest Neighbor (CNN, Hart et al., 1968), and Edited Nearest Neighbor (ENN Wilson (1972)). We update their operations to make them useful for monotonic classification, i.e., the subsets they produce should also be monotonic, be smaller than the original training set and achieve similar performances. MONIPS is evaluated and compared with CNN and ENN on a set of 30 benchmarks and in a case study about opinion surveys of teaching courses.

The rest of the paper is organized as follows: Section 2 presents the problem context to which the MONIPS solution has to be applied. This is done through a case study about student surveys. Section 3 discusses the idea of ordinal classification with monotonic constraints that MNN represents, as well as the prototype selection process applied to this domain. Section 4 details the MONIPS algorithm. Section 5 elaborates upon the experimental framework and discusses the results of the corresponding empirical study. Section 6 summarizes the findings.

## 2. Case study: opinion surveys about teaching courses

As stated in the introduction, the case study focuses on the domain of opinion surveys about teaching courses, developed by the Quality Department of an educational institution. The objective is to analyze the learning activities, using the average-scores students gave to their courses.

To address this, the quality department records polls about each course. Students have to answer to 22 questions about different features of the teaching process. Each answer consists of an assessment between 0 (worse) to 10 (best). The questions listed in Table 1, include the corresponding statistical average and mode values of the 22 questions response for all the courses. Considering the average values of the responses for one course, whose values could be similar that those which appear in the Table 1, the quality department assigns a value between 0 (worse) to 5 (best). The data collected, called data set, is composed by rows where each row is the average evaluation of one course by the students who have been enrolled in it, and the assessment decided by the quality department considering the responses (this is the class assigned). The data set contains 276 rows, where each row is composed by 22 values and one class assigned. The presence of missing values would affect to the comparison between courses to analyze the monotonicity constraints. There are not missing values in the data set, all questions are answered by the students.

**Table 1**
Resume of questions about different features of the teaching process.

| | Question | Avg. | Mode |
|---|---|---|---|
| 1 | "It reports on the various aspects of the program of the subject" | 6.318 | 8 |
| 2 | "The classes are given in the attached schedule" | 5.583 | 7 |
| 3 | "Classes are taught regularly" | 6.177 | 8 |
| 4 | "The tutorials are properly met (and virtual)" | 6.141 | 8 |
| 5 | "The planning of the subject is met" | 5.778 | 5 |
| 6 | "There is coordination between the theoretical and practical activities" | 5.590 | 7 |
| 7 | "The evaluation is adjusted as specified at the beginning" | 6.007 | 8 |
| 8 | "The suggested bibliography is useful" | 5.659 | 7 |
| 9 | "The work-out in class are well organized" | 5.949 | 7 |
| 10 | "Various teaching resources that facilitate learning are used" | 5.920 | 7 |
| 11 | "It is explained clearly and highlights the important content" | 6.333 | 8 |
| 12 | "There is interest in the degree of compression of the explanations" | 6.300 | 8 |
| 13 | "Practical examples are presented to accompany the explanations" | 5.956 | 8 |
| 14 | "The contents are explained safely" | 5.952 | 8 |
| 15 | "Doubts that arise are resolved" | 6.126 | 8 |
| 16 | "A work and participation climate is encouraged" | 5.510 | 7 |
| 17 | "A fluid and spontaneous communication is encouraged" | 5.471 | 7 |
| 18 | "Students are encouraged to take an interest in the subject" | 5.942 | 8 |
| 19 | "A climate of respect is favored with students" | 5.938 | 7 |
| 20 | "Students are clear about what they are required to pass the course" | 5.757 | 5 |
| 21 | "The evaluation criteria are appropriate" | 5.550 | 7 |
| 22 | "The objectives are achieved through the proposed activities" | 6.365 | 8 |

The data set demonstrates clear monotonic relations between its features (22 questions) and the class (course assessment). To illustrate this, a course evaluation with higher feature rates must be rated in a higher or equal quality level compared to others course evaluations with lower rates. Or, in other words, a course evaluation with worse feature assessments cannot present higher quality level than other course evaluation with better student responses. If this last situation appears, a monotonic constraint is violated in the qualification of both evaluations, on the first and the second courses.

At present, the data set contains 276 courses but every semester, new courses are evaluated and included in the data set, so in a few years it will increase considerably its size. This large size and the presence of non monotonic information would affect negatively to the efficacy and efficiency of Monotonic Nearest Neighbor.

## 3. Ordinal classification with monotonicity constraints and monotonic prototype selection

In this section we present the monotonic classification problem, the MNN classifier and how to use PS algorithms in this domain.

### 3.1. Monotonic classification

Monotonicity is a property which appears in many areas of our lives. We can see it in domains like natural sciences, natural language, game theory or economics. The classification with monotonicity constraints, also known as monotonic classification (Ben-David et al., 1989), is an ordinal classification problem where a monotonic restriction is present: a higher value of an attribute in an example, fixing other values, should not decrease its class assignment. In literature we can find different algorithms to tackle the monotonic or ordinal classification (Montañés et al., 2014; Stenina et al., 2015). For example, in Ben-David et al. (1989) the Ordinal Learning Model (OLM) was proposed, which is an ordinal classification model working with monotonic and non-monotonic data sets, but ensuring monotonic classification. Other researchers consider the use of monotone decision trees, which present the requirement of produce purely monotone data sets and ensuring monotone classification (Cao-Van and de Baets, 2003; Feelders and Pardoel, 2003; Makino et al., 1999; Potharst and Bioch, 2000). The probabilistic Ordinal Stochastic Dominance Learner (OSDL) proposed in Lievens et al. (2008) deals with non monotonic data sets and provides monotone classification rules. Other algorithms in this domain are the monotonic neural network models which can deal with noisy data but do not guarantee monotone classifications (Daniels et al., 2006). Algorithms that are robust to noise but do not guarantee monotone classification include: Frank and Hall's ordinal class classifier meta-model (Frank and Hall, 2001), classification by function decomposition (Popova and Bioch, 2005), and hybrid algorithm of Ben-David (1995). More recently, Duivesteijn and Feelders (2008) proposed a modification of the classical nearest neighbor algorithm for monotonic contexts. This is the monotone classification algorithm that we have considered in our study. The reasons for this choice are its ability to use the class labels of nearby points for making predictions, not shared by other algorithms (like OLM), and its high robustness and efficacy. The paper (Ben-David et al., 2009) suggests that ordinal