International Journal of Approximate Reasoning 91 (2017) 36–55

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Analogy-based classifiers for nominal or numerical data

Myriam Bounhas [a,b,*], Henri Prade [c,d], Gilles Richard [c,e]

[a] LARODEC Lab., ISG de Tunis, Tunisia
[b] Emirates College of Technology, Abu Dhabi, United Arab Emirates
[c] IRIT, Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse cedex 09, France
[d] QCIS, University of Technology, Sydney, Australia
[e] British Institute of Technology and E-commerce, London, UK

ARTICLE INFO

Article history:
Received 11 August 2016
Received in revised form 16 August 2017
Accepted 18 August 2017
Available online 31 August 2017

Keywords:
Analogical proportions
Classification
Nominal data
Numerical data

ABSTRACT

Introduced a decade ago, analogy-based classification methods constitute a noticeable addition to the set of instance-based learning techniques. They provide valuable results in terms of accuracy on many classical datasets. They rely on the notion of analogical proportions which are statements of the form "*A* is to *B* as *C* is to *D*". Analogical proportions have been in particular formalized in Boolean and numerical settings. In both cases, one of the four components of the proportion can be computed from the three others, when the proportion holds. Analogical classifiers look for all triples of examples in the sample set that are in analogical proportion with the item to be classified on a maximal number of attributes and for which the corresponding analogical proportion equation on the class has a solution. In this paper when classifying a new item, we specially emphasize an approach where the whole set of triples that can be built from the sample set is not considered. We just focus on a small part of the candidate triples. Namely, in order to restrict the scope of the search, we first look for examples that are as similar as possible to the new item to be classified. We then only consider the pairs of examples presenting the same dissimilarity as between the new item and one of its closest neighbors. In this way, we implicitly build triples that are in analogical proportion on all attributes with the new item. Then the classification is made on the basis of an additive aggregation of the truth values corresponding to the pairs that can be analogically associated with the pairs made of the target item and one of its nearest neighbors. We then only deal with pairs leading to a solvable analogical equation for the class. This new algorithm provides results as good as previous analogical classifiers with a lower average complexity, both in nominal and numerical cases.

* Corresponding author.
E-mail addresses: myriam_bounhas@yahoo.fr (M. Bounhas), prade@irit.fr (H. Prade), richard@irit.fr (G. Richard).

http://dx.doi.org/10.1016/j.ijar.2017.08.010
0888-613X/

1. Introduction

Numerical proportions play an important role in our perception and understanding of reality. Indeed proportions are a matter of comparisons expressed by differences or ratios that are equated to other differences or ratios. Two centuries ago, Gergonne [17,18] was the first to explicitly relate numerical (geometric) proportions to the ideas of interpolation and regression. In fact, geometrical proportions exhibit a simple but effective extrapolation power since, knowing 3 elements a, b, c, we can easily compute a last one d such that $\frac{a}{b} = \frac{c}{d}$ (known as the *rule of three*).

Numerical proportions may be considered as particular instances of the so-called "analogical proportions" which are statements of the form "A is to B as C is to D", often denoted $A : B :: C : D$. This supposes that A, B, C, D refer to the same category of items, which can thus be described in the same terms. Such a proportion expresses that "A differs from B as C differs from D", as well as "B differs from A as D differs from C" [32]. In other words, the pair (A, B) is analogous to the pair (C, D) [19]. More recently, diverse formal views of analogical proportions have been developed in algebraic or logical settings [45,28,29], in such a way that the essential properties of numerical proportions still hold, and especially their extrapolation power. For instance, when analogical proportions are defined in terms of subsets of properties that hold true in a given situation, each variable A, B, C and D refers to a situation described by a vector of feature values [24,31,32,35].

Based on such formal approaches, analogical proportions have proved to be a valuable tool in morphological linguistic analysis [46,23], in solving IQ tests such as Raven progressive matrices [42,12] as well as in classification tasks where results competitive with the ones of classical machine learning methods have been first obtained by [1,30]. In this paper, we focus on this specific type of machine learning application, namely classification.

In this context, we assume that objects or situations A, B, C, D are represented by vectors of attribute values, denoted $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}, \boldsymbol{d}$. The analogical proportion-based approach to classification relies on the idea that the unknown class $x = cl(\boldsymbol{d})$ of a new instance $\boldsymbol{d}$ may be predicted as the solution of an equation expressing that the analogical proportion $cl(\boldsymbol{a}) : cl(\boldsymbol{b}) :: cl(\boldsymbol{c}) : x$ holds between the classes. This is done on the basis of triples $(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$ of examples of the sample set that are such that the analogical proportion $\boldsymbol{a} : \boldsymbol{b} :: \boldsymbol{c} : \boldsymbol{d}$ holds componentwise for all or for a large number of the attributes describing the items.

In this paper, we propose a rather simple and understandable approach to analogy-based classification relying on the use of triples $(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$. For each triple, we compute a global truth value $P(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$ as the average of the truth values obtained in a componentwise manner on each attribute. To aggregate these global truth values and for each class label, we compute the sum of these truth values and we allocate to the new item $\boldsymbol{d}$ to be classified the label having the highest score. Note that we have freedom in choosing $\boldsymbol{c}$ and it is tempting to choose $\boldsymbol{c}$ as a nearest neighbor of $\boldsymbol{d}$. This is what we more particularly experiment in the following. This contrasts with the previous analogical approaches to classification where *all* the triples of examples making an analogical proportion with the new item to be classified were considered in a brute-force, blind and systematic manner.

This paper is structured as follows. In Section 2, we recall the definitions and main properties of analogical proportions, when applied to Boolean-valued and nominal attributes, as well as in the multiple-valued logic case for handling numerical attributes. In Section 3, we investigate the link between analogical proportions and analogical reasoning. We also explain why and how this principle may be used as a tool of interest for classification. In Sections 4, we describe the new analogical proportion-based algorithm (AP-classifier). Section 5 is devoted to an overview of related works. In Section 6, we describe the experimental protocol, the dataset we use and the standard classifiers considered for comparison. In this section, we also provide the experimental results and comment the accuracy of the AP-classifier for Boolean, nominal and numerical data respectively, and compare with state of the art classifiers. In Subsection 6.5, we specifically address the differences between AP-classifier and regular k-NN classifier.

The contents of this paper elaborate on two conference papers respectively devoted to Boolean [2] and to numerical data [4]. In the following, we provide a fully rewritten version of the conference papers, in particular, a deeper investigation of the ideas underlying the procedure, new algorithms and new experiments on a large variety of datasets are reported.

2. Background on analogical proportions

When a, b, c, d are numbers, arithmetical (resp. geometrical) numerical proportions assert equality between two differences: $a - b = c - d$ (resp. ratios: $\frac{a}{b} = \frac{c}{d}$). They are at the root of the idea of analogical proportions. Similarly, a symbolic analogical proportion is a statement of the form "a is to b as c is to d" (usually denoted $a : b :: c : d$ and where the type of a, b, c, d is not specified for now), expressing informally that "a differs from b as c differs from d" and vice versa. As it is the case for numerical proportions, this statement is supposed to still hold when the pairs (a, b) and (c, d) are exchanged, or when the mean terms b and c are permuted (see [36] for a detailed discussion). In the following subsections, we shall first focus on the case where a, b, c, d are Boolean truth variables, i.e. taking their values in $\mathbb{B} = \{0, 1\}$. Then, we will recall how analogical proportions can be extended to graded truth values when variables take their values in $[0, 1]$. This will enable us to deal with Boolean valued and numerical attributes. Lastly, the case of nominal attributes is discussed.