An Analogical Interpolation Method for Enlarging a Training Dataset

Myriam Bounhas[1,2(✉)] and Henri Prade[3]

[1] Emirates College of Technology, Abu Dhabi, UAE
myriam_bounhas@yahoo.fr
[2] LARODEC Lab., ISG de Tunis, Tunis, Tunisia
[3] IRIT, Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse cedex 09, France
prade@irit.fr

Abstract. In classification problems, it happens that the training set remains scarce. Given a data set, described in terms of discrete, ordered attribute values, we propose an interpolation-based approach in order to predict *new* examples useful for enlarging the original data set. The proposed approach relies on the use of continuous analogical proportions that are statements of the form "a is to x as x is to c". The prediction is made on the basis of pairs of examples (a, c) present in the data set, for which one can find a value for x for each attribute value as well as for the corresponding class label of the example thus created. The first option that we consider is to select x as the *midpoint* between a and c, attribute by attribute. To extend the search space, we may also choose x as any randomly selected value *between* the values of a and c. We first propose a basic algorithm implementing these two interpolation definitions, then we extend it to two improved algorithms. In the former, we only consider the nearest neighbor pairs (a, c) to x for prediction, while, in the latter, we further restrict the search to those pairs (a, c) having the same class label. The experimental results, for classical ML classifiers applied to the enlarged data sets built by the proposed algorithms, show the effectiveness of analogical interpolation methods for enlarging data sets.

1 Introduction

Analogical proportions are statements of the form "a is to b as c is to d". In the *Nicomachean Ethics*, Aristotle makes an explicit parallel between such statements and geometric proportions of the form "$\frac{a}{b} = \frac{c}{d}$", where a, b, c, d are numbers. It also parallels arithmetic proportions, or difference proportions, which are of the form "$a - b = c - d$". The logical modeling of an analogical proportion as a quaternary connective between four Boolean items appears to be a logical counterpart of such numerical proportions [15]. It has been extended to items described by vectors of Boolean, nominal or numerical values [2].

A particular case of such statements, named *continuous* analogical proportions, is obtained when the two central components are equal, namely they are

© Springer Nature Switzerland AG 2019
N. Ben Amor et al. (Eds.): SUM 2019, LNAI 11940, pp. 136–152, 2019.
https://doi.org/10.1007/978-3-030-35514-2_11

statements of the form "a is to b as b is to c". In case of numerical proportions, if we assume that b is unknown, it can be expressed in terms of a and c as $b = \sqrt{a \cdot c}$ in the geometric case, and as $\frac{a+c}{2}$ in the arithmetic case. Note that similar inequalities hold in both cases: $\min(a, c) \leq \sqrt{a \cdot c} \leq \max(a, c)$ and $\min(a, c) \leq \frac{a+c}{2} \leq \max(a, c)$. This means that the continuous analogical proportion induces a kind of interpolation between a and c in the numerical case by involving an intermediary value that can be obtained from a and c.

General analogical proportions when d is unknown provides an extrapolation mechanism, which with numbers yields $d = \frac{b \cdot c}{a}$ and $d = b + c - a$ in the geometric and arithmetic cases respectively. We recognize the expression of the well-known Rule of Three in the first expression. Analogical proportions-based inference [2] offers a similar extrapolation device relying on the parallel between (a, b) and (c, d) stated by "a is to b as c is to d".

The analogical proportions-based extrapolation has been successfully applied to classification problems. It may be used either directly as a new classification paradigm [2,12], or as a way of completing a training set on which classical classification methods are applied once this set has been completed [1,4]. This paper investigates the effectiveness of the simpler option of using only continuous analogical proportions that involve pairs instead of triples of items, in order to enlarge a training set.

The paper is organized as follows. Section 2 provides a short background on analogical proportions and more particularly on continuous ones. Then Sect. 3 surveys related work on analogical interpolation or extrapolation. Section 4 presents different variants of algorithms for completing a training set based on the idea of continuous analogical proportions. Section 5 reports the results of the use of different classical classification techniques on the corresponding enlarged training sets for various benchmarks.

2 Background: Continuous Analogical Proportion

The statement "a is to b as c is to d", here denoted $a : b :: c : d$, expresses that "a differs from b as c differs from d, and b differs from a as d differs from c". The logical counterpart of the latter statement, where a, b, c, d are Boolean variables, is given by:

$$a : b :: c : d = (\neg a \wedge b \equiv \neg c \wedge d) \wedge (\neg b \wedge a \equiv \neg d \wedge c)$$

See [13,16] for justifications. This expression is true for only 6 patterns of values for $abcd$, namely $\{0000, 0011, 0101, 1111, 1100, 1010\}$. This extends to nominal values where $a : b :: c : d$ holds true if and only if $abcd$ is one of the following patterns $ssss$, $stst$, or $sstt$, where s and t are two possible distinct values of items a, b, c and d.

Regarding continuous analogical proportions, it can be easily checked that the unique solutions of equations $1 : x :: x : 1$ and $0 : x :: x : 0$ are respectively $x = 1$ and $x = 0$, while $1 : x :: x : 0$ or $0 : x :: x : 1$ have no solution in the

Boolean case. This somewhat trivializes continuous analogical proportions in the Boolean case. The situation for nominal values is the same.

The case of numerical values is richer. a, b, c, d are now supposed to be normalized values in the real interval $[0,1]$. The reader is referred to [6] for a general discussion of multiple-valued logic extensions of analogical proportions. They can be associated with the following expression:

$$a:b::c:d = \begin{cases} 1 - |\,(a-b)-(c-d)\,|, \\ \quad \text{if } a \geq b \text{ and } c \geq d, \text{ or } a \leq b \text{ and } c \leq d \\ 1 - \max(|\,a-b\,|,|\,c-d\,|), \\ \quad \text{if } a \leq b \text{ and } c \geq d, \text{ or } a \geq b \text{ and } c \leq d \end{cases} \tag{1}$$

It coincides with $a : b :: c : d$ on $\{0,1\}$. As can be seen, $a : b :: c : d$ is equal to 1 if and only if $(a-b) = (c-d)$. For instance, $0.2 : 0.5 :: 0.6 : 0.9$, or $0.2 : 0.5 :: 0.2 : 0.5$ holds true. Because $|a-b| = |(1-a)-(1-b)|$, it is easy to check that the code independence property: $a : b :: c : d = (1-a) : (1-b) :: (1-c) : (1-d)$ holds (0 and 1 play symmetric roles, and it is the same to encode an attribute positively or negatively).

Then the corresponding expression for continuous analogical proportions is [16]:

$$a:b::b:c = \begin{cases} 1 - |\,a+c-2b\,|, \\ \quad \text{if } a \geq b \text{ and } b \geq c, \text{ or } a \leq b \text{ and } b \leq c \\ 1 - \max(|\,a-b\,|,|\,b-c\,|), \\ \quad \text{if } a \leq b \text{ and } b \geq c, \text{ or } a \geq b \text{ and } b \leq c \end{cases} \tag{2}$$

As can be seen $a : b :: b : c = 1$ if and only if $b = (a+c)/2$ (which includes the case $a = b = c$). The proportions $0 : \frac{1}{2} :: \frac{1}{2} : 1$ or $0.3 : 0.6 :: 0.6 : 0.9$ are examples of continuous analogical proportions. Moreover, $1 : 3 :: 3 : 5$ is an example of continuous analogical proportion between nominal ordered grades. Thus this extension captures the idea of betweenness implicit in statements of the form "a is to b as b is to c". Note that we have $0 : 1 :: 1 : 0 = 0$ and $1 : 0 :: 0 : 1 = 0$, as expected.

Analogical proportions extend to vectors in a component-wise manner. Let $\boldsymbol{a} = (a_1, \ldots, a_m)$, where each a_i belongs to $\{0,1\}$ (Boolean case), or to a finite set with more than 2 elements (nominal case), or to $[0,1]$ (numerical case). $\boldsymbol{b}, \boldsymbol{c}, \boldsymbol{d}$ are defined similarly. Then $\boldsymbol{a} : \boldsymbol{b} :: \boldsymbol{c} : \boldsymbol{d}$ has a truth value which is just $\min_{i=1}^{m} a_i : b_i :: c_i : d_i$.

In this paper, we deal with classification. So each vector $\boldsymbol{a}$ in a training set is associated with its class $cl(\boldsymbol{a})$. Thus saying that the continuous analogical proportion $\boldsymbol{a} : \boldsymbol{x} :: \boldsymbol{x} : \boldsymbol{c}$ holds true amounts to say:

$$\begin{array}{l} \boldsymbol{a} : \boldsymbol{x} :: \boldsymbol{x} : \boldsymbol{c} = 1 \text{ iff} \\ a_j : x_j :: x_j : c_j = 1 \text{ for each attribute } j \text{ and } cl(\boldsymbol{a}) : cl(\boldsymbol{x}) :: cl(\boldsymbol{x}) : cl(\boldsymbol{c}) = 1 \end{array} \tag{3}$$

Moreover, since continuous analogical proportions are trivial for a Boolean or a nominal variable, we shall also use a more liberal extension of betweenness for

the vectorial case [10] in this paper. Namely, we shall say $\boldsymbol{x}$ is *between* $\boldsymbol{a}$ and $\boldsymbol{c}$ defined as:

$$between(\boldsymbol{a}, \boldsymbol{x}, \boldsymbol{c}) = 1 \text{ iff } a_j \leq x_j \leq c_j \text{ or } c_j \leq x_j \leq a_j \text{ for each attribute } j. \quad (4)$$

Then we can define the set Between$(\boldsymbol{a}, \boldsymbol{c})$ of vectors between two vectors $\boldsymbol{a}$ and $\boldsymbol{c}$. For instance, we have Between$(01000, 11010) = \{01000, 11000, 01010, 11010\}$. Note that in case of Boolean values, the betweenness condition can also be written as $\forall i = 1, \cdots, m$, $(a_i \wedge c_i \rightarrow x_i) \wedge (x_i \rightarrow a_i \vee c_i) = 1$.

3 Related Work

The idea of generating, or completing, a third example from two examples can be encountered in different settings. An option, quite different from interpolation, is the "feature knock out" method [23], where a third example is built by modifying a randomly chosen feature of the first example with that of the second one. A somewhat related idea can be found in a recent proposal [3] which introduces a measure of oddness with respect to a class that is computed on the basis of pairs made of two nearest neighbors in the same class; this amounts to replace the two neighbors by a fictitious representative of the class.

Reasoning with a system of fuzzy if-then rules provides an interpolation mechanism [14], which, from these rules and an input "in-between" their condition parts, yields a new conclusion "in-between" their conclusion parts, by taking advantage of membership functions that can be seen as defining fuzzy "neighborhoods".

Moreover, several approaches based on the use of interpolation and analogical proportions have been developed in the past decade. In [17], the problem considered is to complete a set of parallel if-then rules, represented by a set of condition variables associated to a conclusion variable. The values of the variables are assumed to belong to finite sets of ordered labels. The basic idea is to apply analogical proportion inference in order to induce missing rules from an initial set of rules, when an analogical proportions hold between the variable labels of several parallel rules. Although this approach may seem close to the analogical interpolation-based approach proposed in this paper, our goal is not to predict just the conclusion part of an incomplete rule, but rather a whole example including its attribute-based description and its class. Moreover, we restrict our study to the use of pairs of examples for this prediction, while in [17] the authors use both pairs or triples of rules for completing rules. An extended version of the above-mentioned work has been presented in [22] where the authors also propose a more cautious method that makes explicit the basic assumptions under which rule conclusions are produced from analogical proportions. Along the same line, see also [21] on interpolation between default rules.

Let us also mention the general approach proposed by Schockaert and Prade [20] to interpolative and extrapolative reasoning from incomplete generic knowledge represented by sets of symbolic rules, handled in a purely qualitative manner, where labels are represented in conceptual spaces. This work is an extended

version of [19] in which only interpolative inference is considered. The same authors present an illustrative case study in [18] in the music domain. In the context of natural language modeling, Derrac and Schockaert [5] have proposed a data-driven approach that exploits betweenness and a fortiori inference to derive semantic relations within conceptual spaces.

Besides, some previous works have considered, discussed and experimented the idea of an analogical proportion-based enlargement of a training set, based on triples of examples. In [1], the authors proposed an approach to generate synthetic data to tune a handwritten character classifier. Couceiro et al. [4] presented a way to extend a Boolean sample set for classification using the notion of "analogy preserving" functions that generate examples on the basis of triples of examples in the training set. The authors only tested their approach on Boolean data.

In a more recent work, Lieber et al. [10] have extended the paradigm of classical Case-Based Reasoning to link the current case to either pairs of known cases by performing a restricted form of interpolation, or to triples of known cases by exploiting extrapolation, taking advantage of betweenness and analogical proportion relations.

Lastly, in the context of deep learning, Goodfellow et al. [7] invented the idea of a generative adversarial network (GAN) as a class of machine learning systems. Given a training set, two neural networks, contesting with each other in a game, are learnt in order to generate new data with the same statistics as the training set. More recently, Inoue [9] presented a data augmentation technique for image classification that mix two randomly picked images to train a classifier.

4 Analogical Interpolation-Based Predictor (AIP)

Analogical proportions have been recently applied to classification problems and have shown their efficiency for classifying a variety of datasets [2]. In this paper, we aim to investigate if continuous analogical proportions could be useful for a prediction purpose, namely enlarging a training set with made examples, and if standard classification methods applied to this enlarged set can compete with the direct application of analogical proportions-based inference for classification. As said before, the basic idea of the paper is to apply an interpolation method for predicting *new* examples not present in the original data set which is just enlarged.

In the following, we describe the basic principle of our predicting approach.

4.1 Basic Procedure

Consider a set E of n classified examples i.e., $E = \{(\boldsymbol{x^1}, y^1), ..., (\boldsymbol{x^i}, y^i), ..., (\boldsymbol{x^n}, y^n)\}$ such that the class label $y^i = cl(\boldsymbol{x^i})$ is known for each $i \in 1, ..., n$. The goal is to predict a new set of examples $S = \{(\boldsymbol{x^k}, y^k) \notin E\}$ by interpolating examples from the set E. The new set S will serve for enlarging E.

The basic idea is to find pairs of examples $(\boldsymbol{a}, \boldsymbol{c}) \in E^2$ with known labels such that the analogical proportion (3) is solvable attribute by attribute i.e., there exists $\boldsymbol{x}$ such that $a_j : x_j :: x_j : c_j = 1$ for each attribute $j = 1, ..., m$, and the class equation has $cl(\boldsymbol{x})$ as a solution, i.e., $cl(\boldsymbol{a}) : cl(\boldsymbol{x}) :: cl(\boldsymbol{x}) : cl(\boldsymbol{c}) = 1$.

As mentioned before in Sect. 2, the solution for the previous equation $a_j : x_j :: x_j : c_j = 1$ in the numerical case is just the midpoint $x_j = (a_j + c_j)/2$ for each attribute $j = 1, ..., m$. We are interested in the case of ordered nominal values in this paper. Moreover, we assume that the distances between any two successive values in such an ordered set of values are the same. Let $V = \{v_1, \cdots, v_k\}$ be an ordered set of nominal values, then, v_i will be regarded as the midpoint of v_{i-j} and v_{i+j} with $j \geq 1$, provided that both v_{i-j} and v_{i+j} exist. For instance, if $V = \{1, \cdots, 5\}$, the analogical proportions $1 : 3 :: 3 : 5$ or $2 : 3 :: 3 : 4$ hold, while $2 : x :: x : 5 = 1$ has no solution. So it is clear that some pairs $(\boldsymbol{a}, \boldsymbol{c})$ will not lead to any solution since we restrict the search space to the pairs for which the midpoint (attribute by attribute) exists.

This condition may be too restrictive especially for datasets with high number of attributes which may reduce the set of predicted examples. In case of success, the predicted example $\boldsymbol{x} = \{x_1, ..., x_j, ...x_m\}$ will be assigned to the predicted class label $cl(\boldsymbol{x})$ and saved in a *candidate* set.

Since different voting pairs may predict the same example $\boldsymbol{x}$ more than once ($\boldsymbol{x}$ may be the midpoint of more than one pair $(\boldsymbol{a}, \boldsymbol{c})$), a candidate example may have different class labels. Then has to perform a *vote* on class labels for each candidate example classified differently in the candidate set. This leads to the final predicted set of examples where each example is classified uniquely.

This process can be described by the following procedure:

1. Find solvable pairs $(\boldsymbol{a}, \boldsymbol{c})$ such that Eq. 3 has a unique *non null* solution $\boldsymbol{x}$.
2. In case of ties (an example $\boldsymbol{x}$ is predicted with different class labels), apply voting on all its predicted class labels and assign to $\boldsymbol{x}$ the success label.
3. Add $\boldsymbol{x}$ to the set of predicted examples (together with $cl(\boldsymbol{x})$).

In the next section, we first present a basic algorithm applying the process described above, then we propose two options that may help to improve the search space for the voting pairs.

4.2 Algorithms

The simplest way is to systematically consider *all* pairs $(\boldsymbol{a}, \boldsymbol{c}) \in E^2$, for which Eq. 3 is solvable, as candidate pairs for prediction. Algorithm 1 implements a basic *Analogical Interpolation-based Predictor*, denoted AIP_{std}, without applying any filter on the voting pairs.

Considering all pairs $(\boldsymbol{a}, \boldsymbol{c})$ for prediction may seem unreasonable especially when the domain of attribute values is large since this may blur prediction results. A first improvement of Algorithm 1 is to restrict the search for pairs to those that are among the *nearest neighbors* (in terms of Hamming distance) to the example to be predicted.

Algorithm 1. Analogical Interpolation-based Predictor (AIP_{std})

Input: A set E of classified instances
CandidatesSet = $\emptyset$
S = $\emptyset$
for each pair $(\boldsymbol{a}, \boldsymbol{c})$ in E^2 **do**
  if $cl(\boldsymbol{a}) : cl(\boldsymbol{x}) :: cl(\boldsymbol{x}) : cl(\boldsymbol{c}) = 1$ has solution l **then**
    if $\boldsymbol{a} : \boldsymbol{x} :: \boldsymbol{x} : \boldsymbol{c} = 1$ has solution b **then**
      $cl(b) = l$
      CandidatesSet.add(b)
    end if
  end if
end for
S = VoteOnclasses(CandidatesSet)
Comp(E)= $E + S$
return (Comp(E))

Let us consider two different pairs $(\boldsymbol{a}, \boldsymbol{c})$ and $(\boldsymbol{d}, \boldsymbol{e}) \in E^2$. We assume that $\boldsymbol{a} : \boldsymbol{x} :: \boldsymbol{x} : \boldsymbol{c} = 1$ produces as solution an example $\boldsymbol{b}$ and $\boldsymbol{d} : \boldsymbol{x} :: \boldsymbol{x} : \boldsymbol{e} = 1$ produces an other example $\boldsymbol{b'} \neq \boldsymbol{b}$. If $\boldsymbol{b'}$ is closest to $(\boldsymbol{d}, \boldsymbol{e})$ than $\boldsymbol{b}$ is to $(\boldsymbol{a}, \boldsymbol{c})$ in terms of Hamming distance, it is more reasonable to consider *only* the pair $(\boldsymbol{d}, \boldsymbol{e})$ for prediction. This means that example $\boldsymbol{b'}$ will be predicted while $\boldsymbol{b}$ will be rejected. We denote AIP_{NN} this second improved Algorithm 2 in the following.

Algorithm 3 (that we denote $AIP_{NN,SC}$) is exactly the same as Algorithm 2 in all respects, except that we look for only pairs $(\boldsymbol{a}, \boldsymbol{c})$ belonging to the same class in this case. Note that the two algorithms only differ for non binary classification problems, since $s : x :: x : t = 1$ has no solution in $\{0, 1\}$ for $s \neq t$.

4.3 Another Option

As can be seen in the next section, searching for the best pairs (described in Algorithms 2 and 3) limits the number of accepted voting pairs. Moreover, there is a second constraint to be satisfied, that is limiting the solutions of Eq. 3 to the values of $\boldsymbol{x}$ that are the midpoint of $\boldsymbol{a}$ and $\boldsymbol{c}$ which is hard to be satisfied in the ordered nominal setting. To relax this last constraint, we may think to use the "betweenness" definition given in Eq. 4. In this definition, the equation $between(\boldsymbol{a}, \boldsymbol{x}, \boldsymbol{c}) = 1$ has, as a solution, *any* $\boldsymbol{x}$ such that $\boldsymbol{x}$ is *between* $\boldsymbol{a}$ and $\boldsymbol{c}$ for each attribute $j \in 1, ..., m$. This last option is implemented by the algorithm denoted AIP_{Btw} which is exactly the same as Algorithm 3 except that we use the definition (4) to solve the analogical interpolation.

Algorithm 2. Analogical Interpolation-based Predictor using Nearest Neighbors pairs for prediction (AIP_{NN})

```
Input: A set E of classified instances
CandidatesSet = ∅
PredictedSet = ∅
Min_HD = NbrAttribute
for each pair (a, c) in E^2 do
  if cl(a) : cl(x) :: cl(x) : cl(c) = 1 has solution l then
    if a : x :: x : c = 1 has solution b then
      cl(b) = l
      HD = Max(HammingDistance(b,a), HammingDistance(b,c))
      if HD < Min_HD then
        Min_HD = HD
        CandidateSet.clean()
        CandidatesSet.add(b)
      else if HD =Min_HD then
        CandidatesSet.add(b)
      end if
    end if
  end if
end for
S = VoteOnclasses(CandidatesSet )
Comp(E)= E + S
return (Comp(E))
```

5 Experimentations and Discussion

In this section, we aim to evaluate the efficiency of the proposed algorithms for predicting new examples. For this purpose, we first run different standard ML classifiers on the original dataset, then we apply each AI-Predictor to generate a new set of predicted examples that is used to enlarge the original data set. This leads us to four different enlarged datasets, one for each proposed algorithm. Finally, we re-evaluate again ML classifiers on each of these completed datasets. For both original and enlarged datasets, we apply the testing protocol presented in the next sub-section.

In this experimentation, we tested with the following standard ML classifiers:

- **IBk:** a k-NN classifier, we use the Manhattan distance and we tune the classifier on different values of the parameter $k = 1, 2, ..., 11$.
- **C4.5:** generating a pruned or unpruned C4.5 decision tree. We tune the classifier with different confidence factors used for pruning $C = 0.1, 0.2, ..., 0.5$.
- **JRip:** propositional rule learner, Repeated Incremental Pruning to Produce Error Reduction (RIPPER). We tune the classifier for different values of optimization runs $O = 2, 4, ...10$ and we apply pruning.

Algorithm 3. Analogical Interpolation-based Predictor using Nearest Neighbors pairs in the same class for prediction ($AIP_{NN,SC}$)

```
Input: A set E of classified instances
CandidatesSet = ∅
PredictedSet = ∅
Min_HD = NbrAttribute
for each pair (a, c) in E^2 do
  if cl(a) = cl(c) then
    if a : x :: x : c = 1 has solution b then
      cl(b) = cl(a) //or cl(c)
      HD = Max(HammingDistance(b,a), HammingDistance(b,c))
      if HD < Min_HD then
        Min_HD = HD
        CandidateSet.clean()
        CandidatesSet.add(b)
      else if HD =Min_HD then
        CandidatesSet.add(b)
      end if
    end if
  end if
end for
S = VoteOnclasses(CandidatesSet )
Comp(E)= E + S
return (Comp(E))
```

5.1 Datasets for Experiments

The experimental study is based on several datasets taken from the U.C.I. machine learning repository [11]. A brief description of these data sets is given in Table 1.

To apply the analogical interpolation, we have chosen to deal only with ordered nominal datasets in this study (the extension to the numerical case is the topic of a future work). Table 1 includes 10 datasets with ordered nominal or Boolean attribute values. In terms of classes, we deal with a maximum number of 5 classes.

- `Balance`, `Car`, `Hayes-Roth` and `Nursery` are multiple classes datasets.
- `Monk1`, `Monk2`, `Monk3`, `Breast Cancer`, `Voting` and `W. B. Cancer` datasets are binary class problems. `Monk3` has noise added (in the sample set only). `Voting` data set contains only binary attributes and has missing attribute values. As a missing value, in this dataset, simply means that this value is not "yes" nor "no", we replace each missing value by a third value other than 0 and 1. These data sets are described in Table 1.

5.2 Testing Protocol

To test ML classifiers, we apply a standard 10 fold cross-validation technique. As usual, the final accuracy is obtained by averaging the 10 different accuracies

(computed as the ratio of the number of correct predictions to the total number of test examples) for each fold. However, each ML classifier requires a parameter to be tuned before performing this cross-validation.

Table 1. Description of datasets

Datasets	Instances	Nominal Att.	Binary Att.	Classes
Balance	625	4	0	3
Car	743	6	0	4
Monk1	432	4	2	2
Monk2	432	4	2	2
Monk3	432	4	2	2
Breast Cancer	286	6	3	2
Voting	435	0	16	2
Hayes-Roth	132	5	0	3
W. B. Cancer	699	9	0	2
Nursery	1102	8	0	5

In order to do that, we *randomly* choose a fold (as recommended by [8]), we keep only the corresponding training set (i.e. which represents 90% of the full dataset). On this training set, we again perform a 10-fold cross-validation with diverse values of the parameters. We then select the parameter values providing the best accuracy. These tuned parameters are then used to perform the initial cross-validation. As expected, these tuned parameters change with the target dataset. To be sure that our results are stable enough, we run each algorithm (with the previous procedure) 10 times so we have 10 different parameter optimizations. The displayed parameter p is the average value over the 10 different values (one for each run). The results shown in Table 2 are the average values obtained from 10 rounds of this complete process.

5.3 Experimental Results

In the following, we first provide a comparative study of the overall accuracies for ML classifiers obtained with original and enlarged datasets. This study aims to check if examples predicted by the AIP are of *good quality* (namely labeled with the suitable class). In such case, the efficiency of ML classifiers should be improved when applied to enlarged datasets. Then we also report the main characteristics of these predicted datasets. Finally, we compare ML classification results with enlarged datasets to the ones obtained by directly applying Analogy-based Classification [2] to the original datasets. In this last study, we wonder if using ML classifiers with enlarged datasets may perform similarly as Analogy-based Classification [2] to the original datasets while maintaining a reduced complexity.

Results of ML-Classifiers. Accuracy results for IBk, C4.5 and JRIP are obtained by using the free implementation of Weka software to the enlarged datasets obtained from *AI*-Predictors. To run IBk, C4.5 and JRIP, we first optimize the corresponding parameter for each classifier, using the meta CVParameterSelection class provided by Weka using a cross-validation applied to the training set only. This enables us to select the best value of the parameter for each dataset, then we train and test the classifier using this selected value of this parameter.

Table 2 provides classification results of ML classifiers obtained with a 10-fold cross validation and for the best/optimized value of the tuned parameter (denoted p in this table).

Results in the previous table show that:

Table 2. Results for ML classifiers obtained with the enlarged datasets

Datasets		KNN		C4.5		JRIP	
		Accuracy	p	*Accuracy*	p	*Accuracy*	p
Balance	$AIP_{NN,SC}$	$\mathbf{85.7 \pm 2.13}$	1	$\mathbf{74.15 \pm 2.42}$	0.5	$\mathbf{76.05 \pm 2.85}$	9
	AIP_{NN}	85.31 ± 3.24	1	73.73 ± 4.12	0.5	75.09 ± 3.23	6
	AIP_{Std}	78.16 ± 1.15	3	65.92 ± 2.73	0.5	68.45 ± 3.73	6
	AIP_{Btw}	83.04 ± 3.42	3	75.44 ± 3.89	0.5	75.21 ± 4.64	7
	Orig.	84.05 ± 2.6	11	63.79 ± 4.33	0.3	72.74 ± 3.48	6
Car	$AIP_{NN,SC}$	91.4 ± 1.84	1	92.78 ± 1.28	0.4	88.6 ± 2.82	8
	AIP_{NN}	91.5 ± 1.95	1	$\mathbf{93.14 \pm 1.95}$	0.5	$\mathbf{89.13 \pm 2.55}$	8
	AIP_{Std}	91.51 ± 1.91	3	92.26 ± 1.85	0.3	89.09 ± 1.93	6
	AIP_{Btw}	86.74 ± 2.71	4	88.74 ± 1.99	0.4	85.61 ± 2.38	8
	Orig.	$\mathbf{92.38 \pm 2.51}$	1	90.84 ± 3.61	0.5	86.58 ± 3.67	8
Monk1	$AIP_{NN,SC}$	94.58 ± 2.7	5	94.11 ± 2.88	0.2	$\mathbf{93.75 \pm 2.48}$	2
	AIP_{NN}	94.82 ± 2.37	3	94.53 ± 2.35	0.1	93.62 ± 1.9	2
	AIP_{Std}	87.07 ± 4.48	3	87.35 ± 2.49	0.1	83.21 ± 4.34	6
	AIP_{Btw}	85.34 ± 3.91	3	88.15 ± 4.78	0.3	89.46 ± 3.66	4
	Orig.	$\mathbf{98.37 \pm 2.78}$	2	$\mathbf{99.36 \pm 0.64}$	0.4	90.99 ± 13.15	2
Monk2	$AIP_{NN,SC}$	82.41 ± 4.77	1	72.44 ± 0.19	0.1	71.91 ± 3.32	5
	AIP_{NN}	$\mathbf{82.49 \pm 7.56}$	1	72.44 ± 0.19	0.1	71.87 ± 3.8	3
	AIP_{Std}	76.12 ± 4.28	3	77.03 ± 0.0	0.1	76.6 ± 0.43	4
	AIP_{Btw}	80.86 ± 0.79	3	$\mathbf{80.79 \pm 0.78}$	0.1	$\mathbf{80.56 \pm 0.82}$	3
	Orig.	65.29 ± 1.74	11	67.13 ± 0.61	0.1	64.64 ± 3.69	4
Monk3	$AIP_{NN,SC}$	98.38 ± 1.31	3	98.41 ± 1.31	0.1	98.24 ± 1.49	2
	AIP_{NN}	98.38 ± 1.41	3	98.41 ± 1.41	0.1	98.27 ± 1.42	2
	AIP_{Std}	92.91 ± 2.47	3	93.58 ± 3.09	0.1	92.09 ± 2.63	4
	AIP_{Btw}	97.75 ± 1.76	3	97.71 ± 1.76	0.1	97.87 ± 1.79	2
	Orig.	$\mathbf{99.14 \pm 1.49}$	1	$\mathbf{99.82 \pm 0.18}$	0.2	$\mathbf{98.95 \pm 1.48}$	2

(*continued*)

Table 2. (*continued*)

Datasets		KNN		C4.5		JRIP	
		Accuracy	*p*	*Accuracy*	*p*	*Accuracy*	*p*
Breast Cancer	$AIP_{NN,SC}$	75.57 ± 8.31	4	74.01 ± 7.29	0.2	71.9 ± 8.6	4
	AIP_{NN}	75.59 ± 4.95	5	73.68 ± 6.85	0.2	71.0 ± 7.49	5
	AIP_{Std}	$\mathbf{83.0 \pm 3.19}$	6	$\mathbf{82.47 \pm 3.93}$	0.1	$\mathbf{80.3 \pm 7.01}$	3
	AIP_{Btw}	75.86 ± 5.27	4	75.94 ± 5.99	0.2	72.61 ± 5.84	4
	Orig.	72.81 ± 7.65	9	71.58 ± 6.55	0.2	70.11 ± 8.59	2
Voting	$AIP_{NN,SC}$	93.32 ± 3.58	4	95.65 ± 2.67	0.2	95.62 ± 2.85	3
	AIP_{NN}	93.05 ± 3.17	3	95.79 ± 3.59	0.3	95.67 ± 3.07	3
	AIP_{Std}	$\mathbf{93.89 \pm 2.31}$	2	96.12 ± 2.02	0.3	$\mathbf{96.1 \pm 2.04}$	3
	AIP_{Btw}	93.22 ± 3.84	2	95.45 ± 2.37	0.2	95.73 ± 2.13	2
	Orig.	92.5 ± 3.59	4	$\mathbf{96.38 \pm 2.63}$	0.2	95.84 ± 2.39	4
Hayes-Roth	$AIP_{NN,SC}$	$\mathbf{74.62 \pm 8.84}$	1	74.4 ± 9.63	0.2	84.79 ± 7.65	4
	AIP_{NN}	73.91 ± 8.0	1	74.13 ± 7.65	0.2	85.12 ± 6.58	5
	AIP_{Std}	60.45 ± 11.59	3	70.62 ± 9.3	0.4	78.78 ± 9.67	4
	AIP_{Btw}	69.87 ± 7.77	1	$\mathbf{80.43 \pm 12.53}$	0.1	$\mathbf{88.52 \pm 8.8}$	2
	Orig.	61.41 ± 10.31	3	68.2 ± 6.66	0.2	83.26 ± 9.04	4
W. B. Cancer	$AIP_{NN,SC}$	95.92 ± 1.69	1	94.38 ± 3.38	0.4	94.57 ± 2.15	5
	AIP_{NN}	96.12 ± 2.47	1	94.05 ± 2.82	0.3	94.5 ± 2.31	4
	AIP_{Std}	$\mathbf{96.82 \pm 1.22}$	3	$\mathbf{97.37 \pm 1.23}$	0.5	$\mathbf{96.56 \pm 2.19}$	5
	AIP_{Btw}	95.99 ± 1.17	2	94.43 ± 1.49	0.4	94.44 ± 2.16	5
	Orig.	96.7 ± 1.73	3	94.79 ± 3.19	0.2	95.87 ± 2.9	4
Nursery	$AIP_{NN,SC}$	98.23 ± 0.96	1	98.69 ± 0.56	0.4	97.78 ± 1.12	6
	AIP_{NN}	$\mathbf{98.25 \pm 0.78}$	1	$\mathbf{98.74 \pm 0.64}$	0.5	$\mathbf{97.83 \pm 1.25}$	5
	AIP_{Std}	97.73 ± 0.88	1	98.0 ± 0.96	0.5	97.74 ± 0.99	5
	AIP_{Btw}	95.9 ± 0.97	3	96.51 ± 1.34	0.4	95.78 ± 1.5	6
	Orig.	97.45 ± 1.34	3	97.7 ± 1.36	0.5	95.58 ± 2.04	4
Average	$AIP_{NN,SC}$	**89,01**	–	86,90	–	87,32	–
	AIP_{NN}	88,94	–	86,86	–	87,21	–
	AIP_{Std}	85,76	–	86,07	–	85,89	–
	AIP_{Btw}	86,45	–	**87,35**	-	**87,57**	–
	Orig.	86,01	–	84,96	–	85,46	–

– The accuracy results have been improved when applying ML classifiers on the new predicted data instead of the original data. This is noticed for all datasets except for Monk1 and Monk3 datasets. The highest improvement percentage was noticed with the IBk classifier for the dataset Monk2 (17%), Hayes-Roth (13%) and Breast Cancer (11%).
– Regarding the two artificial datasets Monk1 and Monk3, it is known in the original dataset, that only two attributes among 6 are involved to define the class label for each example. We may think that using the *midpoint* value for each attribute as well as the class label, applied in the proposed analogical

interpolation which treat equally *all* attributes, is not compatible with this kind of classification.
- The good improvement observed for Monk2 dataset confirms our previous intuition since, contrary to Monk1 and Mon3, in Monk2 all attributes are involved in defining the class label in this dataset.
- The standard Algorithm 1 outperforms other algorithms in case of Cancer and Breast Cancer datasets. It is important to note that only these two datasets include attributes with large range of values (with maximum of 10 different values for Cancer and 13 different values for Breast Cancer). Moreover the number of attributes is also high if compared to other datasets. We expect that, in case ordered nominal data is represented by a large scale, using only nearest neighbor pairs for prediction seems too restrictive and leads to a local search for new examples.
- There is no particular algorithm that provides the best results for all datasets.
- We computed the average accuracy for each proposed algorithm and for each ML classifier over all datasets. Results are given at the end of Table 2. We can note that IBk classifier performs the best accuracy when using the enlarged data built from the $AIP_{NN,SC}$ Algorithm. While C4.5 and JRIP perform better when applied to the dataset built from AIP_{Btw} Algorithm.
- Overall, the IBK classifier shows the highest classification accuracy over all datasets.

In this first study, the improved results of ML classifiers when applied to enlarged datasets show the ability of the proposed algorithms (especially, $AIP_{NN,SC}$ and AIP_{Btw}) to predict examples that are labeled with the suitable class.

Characteristics of the Predicted Datasets. To have a better understanding of the previous shown results, in this subsection we aim to investigate more the new predicted datasets. For this end, we compute the number of predicted examples for each dataset and the proportion of these examples that are assigned to the correct/suitable class label. This proportion is computed on the basis of the predicted examples that are compatible with the original set. For this new experimentation, we only consider examples predicted by Algorithm $AIP_{NN,SC}$ (and AIP_{Std} for some datasets). We save these additional results in Table 3. From these results, we can see that:

- In seven among ten datasets, the proportion of predicted examples that are successfully classified is 100%. This means that *all* predicted examples that match the original set are assigned to the correct class label and thus are $fully$ compatible with the original set (see for example Monk2, Breast Cancer, Hayes Roth and Nursery).
- Predicting accurate examples in these datasets may explain why ML classifiers show high classification improvement when applied to the new enlarged dataset.
- Although $AIP_{NN,SC}$ Algorithm succeeds to predict accurate examples, the number of predicted examples is very reduced for some datasets such as for

Breast Cancer, Voting and Cancer. This due to the fact that we restrict the search for only nearest neighbors pairs belonging to the same class in this Algorithm. It is important to note that these datasets contains large number of attributes which make the process of pairs filter more constraining.

- As can be seen in Table 3, the size of the predicted sets is considerably increased, for these three datasets, when applying AIP_{Std} Algorithm which is less constraining than $AIP_{NN,SC}$ (520 examples instead of 46 are predicted for Cancer dataset). In Table 2, we also noticed that, only for these three cited datasets, IBK performs considerably better when applied to the datasets built from the standard algorithm AIP_{Std} (producing larger sets). It is clear that in case the predicted set is very reduced, the enlarged dataset remains similar to the original set that's why the improvement percentage of ML classifiers cannot be clearly noticed in the case of datasets predicted from $AIP_{NN,SC}$ Algorithm.
- Lastly for some datasets such as Monk1 and Monk3, the proportion of predicted examples that are compatible with the original set is low if compared to other datasets. As explained before, in the original sets, the classification function involves only 2 among 6 attributes which seems incompatible with continuous analogical interpolation assuming that all attributes as well as class label are the midpoint of the attributes and the class label of the pair used for prediction.

Table 3. Nbr. of predicted examples, proportion of predicted examples that are compatible with the original set

Datasets	Nbr. predicted	Prop. of success
Balance	529	85.82
Car	630	93.44
Monk1	288	87.5
Monk2	221	100
Monk3	320	96.25
Breast Cancer-$AIP_{NN,SC}$	14	100
Breast Cancer-AIP_{Std}	152	83.78
Voting-$AIP_{NN,SC}$	38	100
Voting-AIP_{Std}	95	100
Hayes-Roth	27	100
Cancer-$AIP_{NN,SC}$	46	100
Cancer-AIP_{Std}	520	100
Nursery	883	99.89

Comparison with AP-Classifier [2]**.** Finally, we provide a comparative study of ML classifiers results, reported in Sect. 5.3, to the results obtained with a direct application of analogical proportions for a classification purpose [2]. Note that in [2], analogical proportions-based extrapolation has been directly applied to define a new classification paradigm while in this paper we exploit analogical proportions-based interpolation to enlarge datasets on which classical ML classifiers are applied. Classification accuracies of analogical proportions-based classifiers [2] are given in Table 4 and compared to the *best* result of each ML classifier applied to the enlarged datasets. Results in Table 4 shows that AP-Classifier outperforms classic ML classifiers on five datasets especially on the three Monks datasets. However enlarged datasets, using analogical interpolation, helped to reduce the gap between AP-Classifier and other ML classifiers once they were applied to these enlarged data. On the other side, ML classifiers provides better accuracies on four other datasets (see for example the Breast cancer (resp. Hayes-Roth) dataset for which the IBK (resp. JRIP) is largely better than AP-Classifier).

Table 4. Results for ML classifiers obtained with the enlarged datasets and comparison with AP-Classifier [2]

Datasets	AP-Classifier [2]		KNN		C4.5		JRIP	
	Accuracy	*p*	*Accuracy*	*p*	*Accuracy*	*p*	*Accuracy*	*p*
Balance	$\mathbf{86.35 \pm 2.27}$	11	85.7 ± 2.13	1	74.15 ± 2.42	0.5	76.05 ± 2.85	9
Car	$\mathbf{94.16 \pm 4.11}$	11	91.5 ± 1.95	1	93.14 ± 1.95	0.5	89.13 ± 2.55	8
Monk1	$\mathbf{99.77 \pm 0.71}$	7	94.82 ± 2.37	3	94.53 ± 2.35	0.1	93.75 ± 2.48	2
Monk2	$\mathbf{99.77 \pm 0.7}$	11	82.49 ± 7.56	1	80.79 ± 0.78	0.1	80.56 ± 0.82	3
Monk3	$\mathbf{99.63 \pm 0.7}$	9	98.38 ± 1.41	3	98.41 ± 1.41	0.1	98.27 ± 1.42	2
Breast Cancer	73.68 ± 6.36	10	$\mathbf{83.0 \pm 3.19}$	6	82.47 ± 3.93	0.1	80.3 ± 7.01	3
Voting	94.73 ± 3.72	7	93.89 ± 2.31	2	$\mathbf{96.12 \pm 2.02}$	0.3	96.1 ± 2.04	3
Hayes-Roth	79.29 ± 9.3	7	74.62 ± 8.84	1	80.43 ± 12.53	0.1	$\mathbf{88.52 \pm 8.8}$	2
W. B. Cancer	97.01 ± 3.35	4	96.82 ± 1.22	3	$\mathbf{97.37 \pm 1.23}$	0.5	96.56 ± 2.19	5

This comparison firstly shows the interest of analogical proportions as a classification tool for some datasets and secondly as way for enlarging datasets for other cases. Identifying on which dataset each of these methods may be better applied should be deeply investigated in future.

In terms of complexity, the proposed Analogical Interpolation approaches (which are quadratic due to the use of pairs of examples) if combined with the IBK classifier for example (which is linear), leads to a improved classifier. This latter shows better classification accuracy and enjoining reduced complexity if compared to the AP-classifier having cubic complexity (that may be computationally costly for large datasets [2]).

6 Conclusion

This paper has studied the idea of enlarging a training set using analogical proportions as in [4], with two main differences: we only consider pairs of examples by using continuous analogical proportions which contribute to reduce the complexity to be quadratic instead of cubic, and we test with ordered nominal datasets instead of Boolean one.

On the one hand the results obtained by classical machine learning methods on the enlarged training set generally improve those obtained by applying these methods to the original training sets. On the other hand, these results, obtained with a smaller level of complexity, are often not so far from those obtained by directly applying the analogical proportion-based classification method on the original training set [2].

References

1. Bayoudh, S., Mouchère, H., Miclet, L., Anquetil, E.: Learning a classifier with very few examples: analogy based and knowledge based generation of new examples for character recognition. In: Kok, J.N., Koronacki, J., Mantaras, R.L., Matwin, S., Mladenič, D., Skowron, A. (eds.) ECML 2007. LNCS (LNAI), vol. 4701, pp. 527–534. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-74958-5_49
2. Bounhas, M., Prade, H., Richard, G.: Analogy-based classifiers for nominal or numerical data. Int. J. Approximate Reasoning **91**, 36–55 (2017)
3. Bounhas, M., Prade, H., Richard, G.: Oddness-based classification: a new way of exploiting neighbors. Int. J. Intell. Syst. **33**(12), 2379–2401 (2018)
4. Couceiro, M., Hug, N., Prade, H., Richard, G.: Analogy-preserving functions: a way to extend Boolean samples. In: Proceedings 26th International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, 19–25 August, pp. 1575–1581 (2017)
5. Derrac, J., Schockaert, S.: Inducing semantic relations from conceptual spaces: a data-driven approach to plausible reasoning. Artif. Intell. **228**, 66–94 (2015)
6. Dubois, D., Prade, H., Richard, G.: Multiple-valued extensions of analogical proportions. Fuzzy Sets Syst. **292**, 193–202 (2016)
7. Goodfellow, I., et al.: Generative adversarial nets. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 27, pp. 2672–2680. Curran Associates, Inc. (2014)
8. Hsu, C., Chang, C., Lin, C.: A practical guide to support vector classification. Technical report, Department of Computer Science, National Taiwan University (2010)
9. Inoue, H.: Data augmentation by pairing samples for images classification. CoRR abs/1801.02929 (2018). http://arxiv.org/abs/1801.02929
10. Lieber, J., Nauer, E., Prade, H., Richard, G.: Making the best of cases by approximation, interpolation and extrapolation. In: Cox, M.T., Funk, P., Begum, S. (eds.) ICCBR 2018. LNCS (LNAI), vol. 11156, pp. 580–596. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01081-2_38
11. Mertz, J., Murphy, P.: UCI repository of machine learning databases (2000). ftp://ftp.ics.uci.edu/pub/machine-learning-databases

12. Miclet, L., Bayoudh, S., Delhay, A.: Analogical dissimilarity: definition, algorithms and two experiments in machine learning. JAIR **32**, 793–824 (2008)
13. Miclet, L., Prade, H.: Handling analogical proportions in classical logic and fuzzy logics settings. In: Sossai, C., Chemello, G. (eds.) ECSQARU 2009. LNCS (LNAI), vol. 5590, pp. 638–650. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-02906-6_55
14. Perfilieva, I., Dubois, D., Prade, H., Esteva, F., Godo, L., Hodáková, P.: Interpolation of fuzzy data: analytical approach and overview. Fuzzy Sets Syst. **192**, 134–158 (2012)
15. Prade, H., Richard, G.: From analogical proportion to logical proportions. Logica Universalis **7**(4), 441–505 (2013)
16. Prade, H., Richard, G.: Analogical proportions: from equality to inequality. Int. J. Approximate Reasoning **101**, 234–254 (2018)
17. Prade, H., Schockaert, S.: Completing rule bases in symbolic domains by analogy making. In: Galichet, S., Montero, J., Mauris, G. (eds.) Proceedings 7th Conference European Society for Fuzzy Logic and Technology (EUSFLAT), Aix-les-Bains, 18–22 July, pp. 928–934. Atlantis Press (2011)
18. Schockaert, S., Prade, H.: Interpolation and extrapolation in conceptual spaces: a case study in the music domain. In: Rudolph, S., Gutierrez, C. (eds.) RR 2011. LNCS, vol. 6902, pp. 217–231. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-23580-1_16
19. Schockaert, S., Prade, H.: Qualitative reasoning about incomplete categorization rules based on interpolation and extrapolation in conceptual spaces. In: Benferhat, S., Grant, J. (eds.) SUM 2011. LNCS (LNAI), vol. 6929, pp. 303–316. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-23963-2_24
20. Schockaert, S., Prade, H.: Interpolative and extrapolative reasoning in propositional theories using qualitative knowledge about conceptual spaces. Artif. Intell. **202**, 86–131 (2013)
21. Schockaert, S., Prade, H.: Interpolative reasoning with default rules. In: Rossi, F. (ed.) IJCAI 2013, Proceedings 23rd International Joint Conference on Artificial Intelligence, Beijing, 3–9 August, pp. 1090–1096 (2013)
22. Schockaert, S., Prade, H.: Completing symbolic rule bases using betweenness and analogical proportion. In: Prade, H., Richard, G. (eds.) Computational Approaches to Analogical Reasoning: Current Trends. SCI, vol. 548, pp. 195–215. Springer, Heidelberg (2014). https://doi.org/10.1007/978-3-642-54516-0_8
23. Wolf, L., Martin, I.: Regularization through feature knock out. MIT Computer Science and Artificial Intelligence Laboratory (CBCL Memo 242) (2004)