On the specificity of distance-based generalisation operators

V. Estruch C. Ferri J. Hernández-Orallo M.J. Ramírez-Quintana

DSIC, Univ. Politècnica de València , Camí de Vera s/n, 46020 València, Spain.
{vestruch,cferri,jorallo,mramirez}@dsic.upv.es

Abstract. Learning methods based on distances are widely used to deal with structured information, since several distance functions can be found for the most common sorts of data. In these algorithms the justification of the labelling of a new case is normally guided by a pattern expressing the similarity to a prototype. Other patterns based on the structure of the data (e.g. one saying "all the lists headed by the item i") could be more interesting but some requirements must be taken into account [2]. Among them, it is convenient to know how specific a pattern is. Here, we briefly present a general framework where the concept of specificity is expressed in terms of the distance and, hence, can be defined for every sort of data embedded in a metric space. In this line, it can be shown that Plotkin's lgg is a specific case of minimal generalisation.

1 Introduction

In some learning problems data is not only described by nomimal and numerical features, but also using other sorts of data (sets, lists, etc.). This forces to define new algorithms coping with structured representations [1]. Distance-based methods are really popular for this purpose because they can be directly employed by defining an adequate distance. Despite the fact that these methods are quite intuitive and have successfully been tested in several domains, a model explaining why a new example belongs to one class or another is missing.

The problem of providing descriptions for distance-based algorithms is addressed in [2] for binary generalisation operators. In [3], we extend the idea for n-ary operators, we explore the notion of comprehensible description in more detail and we study the notion of minimality. We briefly introduce the concept of distance-based generalisation operator presented in this second work. The elements to be generalised are embedded in a metric space (X, d) satisfying certain conditions (such as being connected, see [3]). A generalisation operator Δ maps a finite set of elements $E \subset X$ to a pattern p belonging to a pattern language $\mathcal{L}$, i.e. $\Delta(E) = p$ where $E \subseteq Set(p)$ and $Set(p)$ denotes the set of elements in X represented by p. In fact, every pattern represents a set in X. Among all the possible generalisation operators, we are interested in those computing a pattern which is "consistent" with the distance employed. That is, the so-called distance-based generalisation operator. For instance, let (Σ^*, d) be the space of all the finite words defined over the alphabet $\Sigma = \{a, b, c\}$ and d the edit distance permitting

insertions and deletions only. Given the words $w_1 = cabab$ and $w_2 = ababc$ a distance-based generalisation operator Δ_1 could compute $\Delta_1(w_1, w_2) = *ab*$. That is, all the words having the subsequence ab. This pattern somehow shows why $d(w_1, w_2) = 2$ because the subsequence ab has been taking into account by the distance. However, this is different for an operator like $\Delta_2(w_1, w_2) = *c*$ (all the words having the symbol c) since the common sequence c is not considered by the distance. From this example, a basic issue related to distance-based operators arises. The pattern $*ab*$ computed by Δ_1 looks excessively general w.r.t another "consistent" pattern such as $*abab*$. Thus Δ_1 could be redefined to return more restrictive patterns. But the opposite problem (overfitting) must be considered as well.

Although the idea of generality has been deeply studied when data is represented by means of first-order predicates it does not happen the same for the rest of sorts of data and specially when data is in a metric space. Thus, we propose a general way (inspired on the MDL principle)of defining minimal distance-based generalisation operators (mg operators) for data embedded in metric spaces.

2 Minimal distance-based generalisation operators

Determining when a generalisation operator computes minimal generalisations is important if we want a generalisation to "fit" a group of elements as much as possible. Here, we will proceed as follows. First, we will establish a criterion to determine, given two patterns computed by the distance-based generalisation operators $\Delta(E)$ and $\Delta'(E)$ respectively, which one is less general. Finally, we will say that Δ is a mg operator if for every set E and for every operator Δ', $\Delta(E)$ is less general than $\Delta'(E)$.

Intuitively, we could utilise the inclusion operation between sets ($\subset$). That is, a generalisation of E computed by $\Delta(E)$ is less general than a generalisation computed by $\Delta'(E)$, if $Set(\Delta(E)) \subset Set(\Delta'(E))$. However, this leads to several problems. First, most generalisation are incomparable, since neither $Set(\Delta(E)) \subseteq Set(\Delta'(E))$ or vice versa. Secondly, the inclusion between sets ignores the underlying distance. Finally, the minimal generalisation may not exist. For instance, consider $\mathcal{R}^2$ with the Euclidean distance and $\mathcal{L}$ as the set of all the rectangles in $\mathcal{R}^2$ along with their finite unions. If we look at Figure 1 given the pattern p_0 computed by $\Delta_0(E)$, we can always define another operator Δ_1 such that $Set(\Delta_1(E)) \subset Set(\Delta_0(E))$ and so on. Note that, it is enough to draw a connected chain of smaller rectangles which is included in the previous generalisation and links both A and B. Therefore, if we define the generality in terms of the inclusion between sets, then the mg operator does not exist in this case.
We need a more abstract principle as a generality criterion. A possibility could be as follows. The level of "complexity" of a pattern is reasonable only if a sufficient number of examples justify it, as the MDL/MML principle states. For this purpose, we introduce a special function, called the cost function and denoted by $k(E, p)$, which is usually expressed as the sum of two functions $c(p)$ and $c(E|p)$. The first one informs about the complexity of the pattern p and the second one how good p fits E. One novel point in our approach is that the

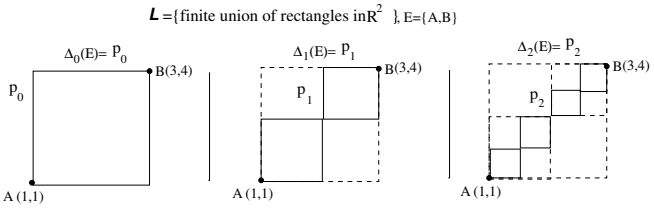

Fig. 1. Generalising $E = \{A(1,1), B(3,4)\}$ by means of Δ_i.

function $c(E|p)$ is expressed in terms of the distance employed. Although there are several possibilities, one choice is to define this function as the sum of all the distances from the elements in E to its nearest point at the border of $Set(p)$. Other more specific definitions for $c(E|p)$ can be found in [3]. The advantage of defining $c(E|p)$ in this way is that it is independent w.r.t. the sort of data employed and it can be applied for any metric space. However, $c(p)$ usually depends on the sort of data [3]. For instance, regarding the pattern language $\mathcal{L}$ formed by rectangles in $\mathcal{R}^2$, $c(p)$ could be the number of rectangles of the pattern p multiplied by a scale factor (e.x. p_0 has 1 rectangle, p_1 has 2, etc.).

Once the function $k(E, p)$ has been specified, we will say that a pattern $\Delta(E)$ is less general than $\Delta'(E)$ if $k(E, \Delta(E)) \leq k(E, \Delta'(E))$. For instance, using both $c(p)$ and $c(E|p)$ definitions shown above and regarding the patterns depicted in Figure 1, it is easy to check that $k(E, \Delta_0(E)) \leq k(E, \Delta_i(E))$, $i = 1, 2$.

To conclude, let us consider the metric space of the first-order atoms induced by the distance [4]. If we set $\mathcal{L}$ the Herbrand's base with variables, $c(p)$ a constant function and $c(E|p)$ as the one defined so far, it can be shown that $\Delta(E) = lgg(E)$, where E is any set of ground atoms, is a mg operator [3].

3 Conclusions

This work introduce the notion of mg operator for every sort of data which is embedded in a metric space. Here we have exposed the basic ideas. In [3] we have a detailed definition of the framework, the notion of generality in terms of the MDL/MML principle, and we define several minimal generalisation operators for different sorts of data. Additionally, we prove that the lgg is a particular case of mg operator for atoms embedded in the metric space defined in [4].

References

1. S. Dzeroski and N. Lavrac, editors. *Relational Data Mining*. Springer-Verlag, Berlin, September 2001.
2. V. Estruch, C. Ferri, J. Hernández-Orallo, and M. J. Ramírez-Quintana. Distance based generalisation. In *Proc. of the 15th International Conference on Inductive Logic Programming, ILP*, volume 3625 of *LNCS*, pages 87–102. Springer, 2005.
3. V. Estruch, C. Ferri, J. Hernández-Orallo, and M. J. Ramírez-Quintana. On the relationship between distance and generalisation. Technical report, DSIC, Universidad Politecnica de Valencia, 2006.
4. J. Ramon, M. Bruynooghe, and W. Van Laer. Distance measures between atoms. In *CompulogNet Meet. on Comp. Logic and ML*, pages 35–41. Uni. Manchester, 1998.