Flexible Matching of Boolean Symbolic Objects

Floriana Esposito, Donato Malerba and Francesca A. Lisi

Dipartimento di Informatica, Università degli Studi di Bari, Via Orabona 34, 70126 Bari, Italy

Summary

Matching is the process of comparing two or more structures to discover their likenesses or differences. It is a common operation performed in symbolic classification, pattern recognition, data mining and expert systems. The definition of a matching operator for Boolean symbolic objects is important for the development of symbolic data analysis techniques. In this paper we give the definition of canonical matching of Boolean symbolic objects, and then we extend it in order to take into account only partial matching caused by the presence of noise. The new definition of flexible matching is based on the probability theory. Some experimental results are reported.

Keywords: Symbolic data analysis, Canonical and Flexible Matching.

1 Introduction

Matching is the process of comparing two or more structures to discover their similarities or differences. Similarity judgements in the matching process are directional: They have a *referent*, *a*, and a *subject*, *b*. The former is either a prototype or the description of a *class* of objects, while the latter is either a variant of the prototype or an *instance* of a class of objects. Matching two structures is a common problem to many domains, like symbolic classification, pattern recognition, data mining and expert systems.

The definition of a matching operator for Boolean symbolic objects (BSOs) is deemed important for the development of several symbolic data analysis techniques, such as factor analysis. In general, a BSO represents a class description (Diday 1990) and plays the role of the referent in the matching process. For instance, the following BSO:

a: $[color = \{black, white\}] \wedge [height = [170, 200]]$

describes a group of individuals either black or white, whose height is in the interval [170,200]. Such a set of individuals is called *extension* of the BSO. The extension is a subset of the universe Ω of *individuals*. Given another BSO *b*, corresponding to the intensional description of an individual and playing the role of subject, the problem is that of establishing whether the individual described by *b* can be considered an instance of the class described by *a*. For instance, the following BSO:

b': $[color = black] \wedge [height = 180]$

describes an individual in the extension of *a*, while the following BSO

b": $[color = red] \wedge [height = 160]$

does not. Then we can say that *a* matches *b'* but not *b"*.

The result of a *canonical* matching operator is either *0* (false) or *1* (true). If S denotes the space of BSOs described by a set of *n* variables x_i taking values in the corresponding domains O_i, then the matching operator is a function:

$$Match: S \times S \rightarrow \{0, 1\}$$

such that for any two BSOs $a, b \in S$:

$$a = [x_1 = A_1] \wedge [x_2 = A_2] \wedge \ldots \wedge [x_n = A_n] = \wedge_{i=1}^{n} [x_i = A_i]$$

$$b = [x_1 = B_1] \wedge [x_2 = B_2] \wedge \ldots \wedge [x_n = B_n] = \wedge_{i=1}^{n} [x_i = B_i]$$

it happens that:

$$Match(a,b) = 1 \quad \text{if } B_i \subseteq A_i \text{ for each } i=1, 2, \ldots, n,$$
$$Match(a,b) = 0 \quad \text{otherwise.}$$

It is worthwhile to note that the *Match* function satisfies two out of three properties of a similarity measure:

1. $\forall a, b \in E$: $Match(a, b) \geq 0$
2. $\forall a, b \in E$: $Match(a, a) \geq Match(a, b)$ *(*backward *property)*

while it does not satisfy the commutativity *or* simmetry *property:*

$\forall a, b \in E$: $Match(a, b) = Match(b, a)$

because of the different role played by *a* and *b*.

2 Definition of flexible matching

The requirement $B_i \subseteq A_i$ for each $i=1, 2, \ldots, n$, might be too strict for real-world problems, because of the presence of noise in the description of the individuals of the universe. For instance, in the example above the function *Match* returns zero when *height=169*. Therefore, it becomes necessary to rely on a more flexible definition of matching that aims at comparing two descriptions in order to identify their similarities rather than their equality (Esposito, Malerba & Semeraro 1991a). The result of a flexible matching should produce a number in the unit interval [0,1] that indicates a *degree of match* between two BSOs, that is

$$\textit{flexible-matching: } \mathsf{S} \times \mathsf{S} \rightarrow [0,1]$$

such that, for any two BSOs *a* and *b*,

i) $\textit{flexible-matching}(a,b)=1$ if $\textit{Match}(a,b)=\textit{true}$,
ii) $\textit{flexible-matching}(a,b) \in [0,1)$ otherwise.

The result of the flexible matching can be interpreted as the probability of *a* matching *b* provided that a change is made in *b*. Let *b'* be a BSO obtained from *b* by means of some changes, such that *a* matches *b'*, and let *P(b | b')* be the conditional probability of observing *b* given that the original observation was *b'*. Then it is possible to set

$$\mathsf{S}_a = \{b' \in \mathsf{S} \mid \textit{Match}(a,b')=1\}$$

and

$$\text{flexible-matching}(a,b) \underset{def}{=} \max_{b' \in S_a} P(b \mid b')$$

that is *flexible-matching(a,b)* equals the maximum conditional probability over the space of BSOs matched by *a*.

Now the problem is that of estimating *P(a| b')* for all clauses *b'* matched by *a*. Let *b* be the conjunction of *simple* BSOs, $b_1, b_2, \ldots, b_m$. Then, under the assumption of conditional independence of the variables used to describe individuals, the probability *P(b | b')* can be factored as follows:

$$P(b|b') = \prod_{i=1}^{n} P(b_i|b') = \prod_{i=1}^{n} P(b_i|b'_1 \wedge b'_2 \wedge \ldots\ b'_n)$$

where $P(b_i \mid b')$ denotes the probability of observing the fact b_i given *b'*. Suppose that b_i is $[x_i=Value_i]$. If *b'* contains the conjunct $[x_i=Value'_i]$, $P(b_i \mid b')$ is the probability that the real value was $Value'_i$, but we observed $Value_i$. Unfortunately, the last equation causes pragmatism to rear its ugly head. Indeed it requires knowledge on the conditional probabilities $P(b_i \mid b'_1 \wedge b'_2 \wedge \ldots \wedge b'_n)$. To lighten the burden, we assume that b_i depends exclusively on $[x_i=Value'_i]$. Thus we have:

$$P(b_i \mid b') = P([x_i=Value_i] \mid [x_i=Value'_i])$$

This probability can be interpreted as the *similarity* between [x_i=*Value*$_i$] and [x_i=*Value'*$_i$], in the sense that the more similar they are, the higher

$$P(b_i \mid b'_i) = P([x_i=Value_i] \mid [x_i=Value'_i])$$

Given:

1. a probability distribution of the values in the domain of x_i, and

2. a distance function δ_i defined on the domain itself, than $P(b_i \mid b')$ can be defined as follows:

$$P(b_i \mid b') = P([x_i=Value_i] \mid [x_i=Value'_i]) = P(\delta_i(Value'_i, X) \geq \delta(Value'_i, Value_i))$$

that is, $P(b_i \mid b')$ is the probability of observing a greater distortion than that existing between $Value_i$ and $Value'_i$.

Note that when $Value_i = Value'_i$ it happens that $\delta_i(Value'_i, X) \geq \delta_i(Value'_i, Value_i) = \delta_i(Value'_i, Value'_i) = 0$ for any value taken by X, thus:

$$P(b_i \mid b') = 1$$

For instance, assuming that δ_i is the city block distance for nominal variables:

$$\delta_i(x,y) = \begin{cases} 0 & \text{if x = y} \\ 1 & otherwise \end{cases}$$

and that the values in the domain O_i are uniformly distributed, then:

$$P([x_i = Value_i] \mid [x_i = Value'_i]) = \frac{|\mathrm{O_i}| - 1}{|\mathrm{O_i}|}.$$

This result is coherent with our intuition that the larger the size of O_i the higher the probability of observing a value different from $Value'_i$.

It is interesting to note that all distance functions equivalent to δ_i lead to the same result. Thus our formulation of flexible matching is *scale-invariant*.

To sum up,

$$P(b|b') = \prod_{i=1}^{n} P(b_i|b') = \prod_{i=1}^{n} P([x_i = Value_i]|[x_i = Value'_i]) =$$

$$= \prod_{i=1}^{n} P(\delta_i(Value'_i, X) \geq \delta_i(Value'_i, Value_i))$$

When b contains a missing value [x_i=?] we can say that the information on x_i is *missing* or *unknown*. In this case $P(b_i \mid b')$ is computed as the expected value of $P([x_i=Value] \mid b')$, where *Value* is one of the values in the domain $\mathrm{O_i}$, that is

$$P(b_i|b') = \sum_{Value \in O_i} P([x_i = Value]) \cdot P([x_i = Value] \mid b') =$$

$$= \sum_{Value \in O_i} P([x_i = Value]) \cdot P([x_i = Value] \mid [x_i = Value'_i])$$

where the summation should be intended as an integral for continuos-valued variables. For instance, for nominal variables with a city block distance δ_i we have:

$$P([x_i = ?] \mid [x_i = Value'_i]) = \frac{|O_i|^2 - |O_i| + 1}{|O_i|^2}.$$

A thorough presentation of the problem of missing data is given in (Esposito, Malerba & Semeraro 1991b).

The definition of flexible matching can be easily extended in order to deal with logical dependencies between variables. Indeed, the main change in the definition concerns the space S_a, since only those BSOs *b* such that *Match(a,b)=1* and that satisfy the logical dependencies should be considered.

3 Experimental results

In order to show how the definition of flexible matching can be used, we considered a data set concerning credit card applications. The data set is distributed with the C4.5 decision tree learning system (Quinlan 1993) and contains fifteen variables whose names and values have been changed to meaningless symbols to protect confidentiality of the data. The sixteenth variable concerns the class of the credit card application: positive in case of approval of credit facilities, negative otherwise. By using the systems C4.5 and C4.5rules it is possible to learn some classification rules for each class. In particular rules learned from a training set of 490 cases are:

Rule	*Class*	*Th.*	*Conditions*
41	-	0.89	[A3 > 1.54] ∧ [A9 = f] ∧ [A4 ∈ {u, y}] ∧ ∧ [A6∈ {c,d, cc, i, j, k, m, r, q, w, e, aa, ff}]
43	-	0.85	[A4 ∈ {u, y}] ∧ [A8 <= 1.71] ∧[A9 = f]
6	-	0.95	[A3 <= 0.835] ∧ [A6 ∈ {c,d,i,k,m,q,w,e,aa }] ∧ ∧ [A7 ∈ {v,bb}] ∧ [A14 > 102] ∧ [A15 <= 500]
30	+	1.0	[A9 = t]
34	+	1.0	[A3 <= 0.125] ∧ [A14 > 221]
46	+	1.0	[A4 ∈ {l}]

Such rules can be easily represented by means of BSOs. For instance, the second rule can be represented as follows:

[A1 = *] ∧ ... ∧ [A3 = *] ∧ [A4 = {u,y}] ∧ [A5 =*] ∧ … ∧ [A7=*] ∧ [A8 = [0.0 .. 1.71]] ∧ [A9 = {f}] ∧ [A10 = *] ∧ ... ∧ [A15 = *]

where “*” stands for any value of the domain. The thresholds to be used in the flexible matching are estimated on the training set itself and are reported in the third column (*Th.*) of the table above. It is worthwhile to observe that all thresholds for class “+” equals 1.0, since induced rules are too specific and no condition can be relaxed without covering several observations of class “-”. In order to test the validity of the approach, both the canonical and the flexible matching with estimated thresholds are applied to an independent set of 200 new cases distributed with C4.5, as in the previous case. The experimental results are reported below:

rule	*41*	*43*	*6*	*30*	*34*	*46*
canonical	46/1	65/4	23/13	83/27	7/3	0/0
flexible	86/5	80/5	32/16	83/27	7/3	0/0

where each pair *m/n* represents the number of correct/wrong classifications.

As can be noticed, the application of a flexible matching leads to a significant increase of correct classifications, still keeping the misclassification rate low. Furthermore, from the table we can draw the conclusion that flexible matching cannot improve the performance of “bad” classification rules, such as rule 46. On the contrary, it can be profitably exploited to relax the conditions expressed in rules “good enough”, without restarting a new learning process in order to build rules that classify uncovered cases.

As future work we intend to validate the proposed approach on other datasets.

References

E. Diday (1990), Knowledge representation and symbolic data analysis. *Knowledge, Data and Computer.Assisted Decisions*, Schader, M. & Gaul, W. (Eds.), Springer-Verlag, 17-34.

F. Esposito, D. Malerba, & Semeraro, G. (1991a), Flexible matching of noisy structural descriptions. *Proceedings of the 12th Joint Conference on Artificial Intelligence*, 658-664, Sidney.

F. Esposito, D. Malerba, & Semeraro, G. (1991b), Classification of incomplete structural descriptions using a probabilistic distance measure. *Symbolic-Numeric Data Analysis and Learning*, Diday, E. & Lechevallier, Y. (Eds.), Nova Science Publishers, 469-482.

J.R. Quinlan (1993), *C4.5: Programs for machine learning*, Morgan Kaufmann, San Mateo, CA.