Toward an Explanatory Similarity Measure for Nearest-Neighbor Classification

Mathieu Latourrette

Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier,
UMR 5506 Université Montpellier II CNRS,
161, rue Ada 34392 Montpellier FRANCE
latourre@lirmm.fr

Abstract. In this paper, a new similarity measure for nearest-neighbor classification is introduced. This measure is an approximation of a theoretical similarity that has some interesting properties. In particular, this latter is a step toward a theory of concepts formation. It renders identical some examples that have distinct representations. Moreover, these examples share some properties relevant for the concept undertaken. Hence, a rule-based representation of the concept can be inferred from the theoretical similarity. Moreover, in this paper, the approximation is validated by some preliminary experiments on non-noisy datasets.

1 Introduction

Learning to classify objects is a fundamental problem in artificial intelligence and other fields, one which has been addressed from many sides. This paper deals with the nearest-neighbor methods (Cover and Hart [6]), also known as exemplar-based (Salzberg [8]) or instance-based learning programs (Aha et al. [1]). These algorithms classify each new example according to some past experience (a set of examples provided with their labels) and a measure of similarity between the examples. Actually, they assign to each new example the label of its nearest known example.

At first glance, similarity seems a rather intuitive notion. Examples are denoted by some properties and are similar if they have some properties in common. Thus, the more similar examples are, the more likely they share some relevant properties for the concept to learn. When the size of the dataset increases, new examples and their nearest neighbors become more and more similar. And, in the limit, classification is accurate.

Such a convergence has been studied many times. Despite positive results, such a similarity has been criticized for not being explanatory. It does not identify among the properties shared by some similar examples the ones that are relevant for the concept undertaken.

This paper is focused on the problem of explanation. The concepts to learn are assumed to have some rule-based representations. In this case, the relevant

R. López de Mántaras, E. Plaza (Eds.): ECML 2000, LNAI 1810, pp. 238–245, 2000.

properties are the preconditions of these rules. To explain the classification of each example, the proposed similarity measure enables to infer a rule-based representation of the concept undertaken.

For that matter, we suggest that examples are similar because they satisfy the same rules and, no longer, because the properties they are denoted by are somewhat similar. Such a similarity measure relies on the rules characterizing the concept undertaken. For a classification task, such a similarity is theoretical as the rules are unknown. However, this similarity can be approximated from each dataset and some rules inferred from this latter.

This paper is organized as follows. §2 summarizes some notations and definitions. §3 introduces the theoretical similarity. §4 is devoted to its approximation and to the resulting classifier. §5 deals with some related research.

2 Preliminaries

Let us introduce a few useful definitions. Let F: $\{f_1, f_2, \ldots, f_n\}$ be a set of features, where each feature f_i can take values in its domain Dom_i: a finite unordered set. An example x: $(x_1, x_2, \ldots, x_n)$ is characterized by an instantiation x_i of each feature f_i. The example x satisfies the conjunction x_c: $f_1 = x_1 \wedge f_2 = x_2 \wedge \ldots \wedge f_n = x_n$. Let U denote the universe: the set of all the possible examples. Considering a finite unordered set L of labels, a concept C is a function from U to L. An exemplar e is a couple $(x, C(x))$ of an example and its label. Let E be the set of all the exemplars. A dataset D is a subset of E.

For example, for the monk1 dataset, examples are represented by 6 features. The domain of f_1, f_2 and f_4 is $\{1, 2, 3\}$. The domain of f_3 and f_6 is $\{1, 2\}$ and the domain of f_5 $\{1, 2, 3, 4\}$. The set of labels is $\{0, 1\}$. The universe contains 432 $(=3 \times 3 \times 2 \times 3 \times 4 \times 2)$ examples. The concept undertaken is the boolean function $(f_1 = f_2) \vee (f_5 = 1)$. Two exemplars are:

$$e_1\text{: ((1,1,1,1,1,1), 1)} \quad \text{and} \quad e_2\text{: ((2,2,1,3,2,1), 1)}$$

Definition 1. *A rule r is a partial function from U to a particular label l_r denoted by $c_r \Longrightarrow l_r$. It associates to each example x such that $c_x \Longrightarrow c_r$ the label l_r. c_r is a conjunction of conditions upon the values of each feature. For each feature f_i, its value is required to be in a subset (not empty) of Dom_i.*

On the monk1 problem, a rule r^* is:

$$f_1 \in \{1,2\} \wedge f_2 \in \{1,2\} \wedge f_3 \in \{1\} \wedge f_4 \in \{1,3\} \wedge f_5 \in \{1,2\} \wedge f_6 \in \{1\} \Longrightarrow 1$$

Let us denote such a rule by:

$$\{1,2\}, \{1,2\}, \{1\}, \{1,3\}, \{1,2\}, \{1\} \Longrightarrow 1$$

An example x or an exemplar $e = (x, l)$ is covered by a rule r iff $c_x \Longrightarrow c_r$. Let $U_{/r}$ (resp. $E_{/r}$) be the subset of the examples (resp. exemplars) covered by r. An exemplar refutes r if it is covered by r but has a different label. r is coherent with the dataset D if there is no exemplar in D to refute r. r is coherent with the concept C if all the exemplars of $E_{/r}$ have the label of r. A rule r_1 is more specific than a rule r_2 if $U_{/r_1} \subset U_{/r_2}$. In this case, r_2 is more general than r_1.

Definition 2. *Let the generalization of each subset s of exemplars of the same label be $G(s)$ the most specific rule covering s and coherent with s.*

Notice that the generalization of a subset of exemplars is unique. Actually, the label of a generalization $G(s)$ is the label of the exemplars in s. And, for each feature f_i, the value x_i of an example covered by $G(s)$ is required to be in the union of the values of f_i appearing in s. For example, $G(\{e_1, e_2\})$ is r^*.

The reader shall see that the operator G satisfies the two properties:

1. (**monotonicity**) The generalization of a subset of exemplars covered by a rule coherent with a concept C is coherent with C.

2. (**stability**) Let C be a concept, r a rule and e an exemplar. If $\forall e' \in E_{/r}$, $G(\{e, e'\})$ is coherent with C then $G(\{e\} \cup E_{/r})$ is coherent with C.

In the reminder of this paper, these two properties will be the only ones required for G. As they are rather natural for an operator of generalization, we guess that our approach can be extended to many other representation languages.

3 Similarity with Respect to a Concept

3.1 Definition

This section is devoted to the definition of the theoretical similarity with respect to a concept C. For that matter, we assume that C is *well-defined*.

Definition 3. *A well-defined concept C is a function from a universe U to a set of labels L characterized by a set of rules R. Thus, for each example x of U and each rule r $(c_r \Longrightarrow l_r)$ of R covering x, there is $l_r = C(x)$.*

Many sets of rules characterize a concept. However, as we suggest that examples are similar because they satisfy the same rules, we have to choose these rules.

Definition 4. *Let the definition of a well-defined concept C be the set of all the most general rules coherent with C. For each exemplar e, let $Def_C(e)$ be the subset of the rules covering e and defining C.*

Notice that the rules defining a concept contain only relevant properties. Actually, all the conditions that could have been dropped from the maximal rules have already been. The definition of the monk1 concept is:

$$\begin{array}{cccccccc}
\{1\}, & \{1\}, & \{1,2\}, & \{1,2,3\}, & \{1,2,3,4\}, & \{1,2\} & \Longrightarrow 1 & \text{(I)} \\
\{2\}, & \{2\}, & \{1,2\}, & \{1,2,3\}, & \{1,2,3,4\}, & \{1,2\} & \Longrightarrow 1 & \text{(II)} \\
\{3\}, & \{3\}, & \{1,2\}, & \{1,2,3\}, & \{1,2,3,4\}, & \{1,2\} & \Longrightarrow 1 & \text{(III)} \\
\{1,2,3\}, & \{1,2,3\}, & \{1,2\}, & \{1,2,3\}, & \{1\}, & \{1,2\} & \Longrightarrow 1 & \text{(IV)} \\
\{1\}, & \{2,3\}, & \{1,2\}, & \{1,2,3\}, & \{2,3,4\}, & \{1,2\} & \Longrightarrow 0 & \text{(V)} \\
\{2\}, & \{1,3\}, & \{1,2\}, & \{1,2,3\}, & \{2,3,4\}, & \{1,2\} & \Longrightarrow 0 & \text{(VI)} \\
\{3\}, & \{1,2\}, & \{1,2\}, & \{1,2,3\}, & \{2,3,4\}, & \{1,2\} & \Longrightarrow 0 & \text{(VII)} \\
\{2,3\}, & \{1\}, & \{1,2\}, & \{1,2,3\}, & \{2,3,4\}, & \{1,2\} & \Longrightarrow 0 & \text{(VIII)} \\
\{1,3\}, & \{2\}, & \{1,2\}, & \{1,2,3\}, & \{2,3,4\}, & \{1,2\} & \Longrightarrow 0 & \text{(IX)} \\
\{1,2\}, & \{3\}, & \{1,2\}, & \{1,2,3\}, & \{2,3,4\}, & \{1,2\} & \Longrightarrow 0 & \text{(X)}
\end{array}$$

Definition 5. *The neighborhood of an exemplar e with respect to a well-defined concept C is defined as follows:* $N_C(e) = \{e' \in E \mid Def_C(e) \cap Def_C(e') \neq \emptyset\}$.

Our similarity between two exemplars is measured between their neighborhoods. We choose the ratio between the numbers of exemplars common to the two neighborhoods and the number of examples belonging to one of them:

Definition 6. *Considering two exemplars e and e', their similarity with respect to a well-defined concept C is:* $Sim_C(e, e') = \frac{|N_C(e) \cap N_C(e')|}{|N_C(e) \cup N_C(e')|}$

3.2 An Accurate Similarity for Nearest-Neighbor Classification

Let two exemplars be equivalent if and only if their similarity is 1. First of all, notice that two equivalent exemplars have the same label.

Theorem 1. *Let C be a well-defined concept and e and e' two exemplars. If e and e' are equivalent, they have the same label.*

Proof. By definition of Sim_C, e and e' are equivalent iff $N_C(e) = N_C(e')$. As e belongs to its neighborhood, e belongs to $N_C(e')$. By definition of $N_C(e')$, there is a rule r of $Def_C(e')$ that covers e. Therefore, e and e' have the label of r.

Thus, if the dataset contains an equivalent exemplar for each new example, the nearest-neighbor rule is accurate.

Definition 7. *Let C be a well-defined concept, e an exemplar. Then, the class of equivalence of e considering* Sim_C *is:* $Eq_C(e) = \{e' \in E \mid Sim_C(e, e') = 1\}$

The number of classes of equivalent exemplars does not depend on the dataset (theorem 2). Therefore, when the size of the dataset increases, more and more classes are represented. And, in the limit, the classifier is accurate. For the monk1 concept, there are only 13 such classes and 432 exemplars.

Theorem 2. *Let C be a well-defined concept.*

$$\forall e \in E \qquad Eq_C(e) = \{e' \in E \mid Def_C(e) = Def_C(e')\}$$

Proof. If $Def_C(e) = Def_C(e')$ then $N_C(e) = N_C(e')$ and $Sim_C(e, e')$=1. Now, assume that $Sim_C(e, e') = 1$ (i.e. $N_C(e) = N_C(e')$) and let r be in $Def_C(e)$. Each exemplar e'' covered by r belongs to $N_C(e) = N_C(e')$. Thus, $G(\{e', e''\})$ is coherent (definition of $N_C(e')$ and monotonicity). It follows that $r'' = G(\{e'\} \cup E_{/r})$ is coherent (stability). As r is maximal, it means that r'' is r. Therefore, r belongs to $N_C(e')$. Hence, if $Sim_C(e, e') = 1$, then $Def_C(e)$=$Def_C(e')$.

3.3 An Explanatory Similarity

Considering such a similarity, each exemplar is equivalent to many others. Theorem 3 states that the generalizations of some equivalent exemplars are coherent with the concept. Therefore, among the properties shared by some equivalent exemplars, some of them are relevant. This is the reason why such a similarity is somewhat explanatory.

Theorem 3. *Let C be a well-defined concept.*

$$\forall e \in E, \quad G(Eq_C(e)) \text{ is coherent with } C.$$

Proof. Theorem 2 shows that all the exemplars of $Eq_C(e)$ are covered by the same rules: $Def_C(e)$. Their generalization $G(Eq_C(e))$ is thus more specific than each of the rules of $Def_C(e)$ and, therefore, coherent with the concept C.

In the example, e_2 satisfies the rule II only and is equivalent to all the examples that satisfy only this rule. Therefore, the generalization $G(Eq_C(e_2))$ is

$$\{2\}, \{2\}, \{1,2\}, \{1,2,3\}, \{2,3,4\}, \{1,2\} \Longrightarrow 1$$

It requires each covered example to satisfy $f_1 = 1$, $f_2 = 1$ and $f_5 \neq 1$. The other conditions are trivial as each value is necessary in its domain. The two first properties are relevant for the concept. However, the last one is not. It is present to prevent the exemplars covered from satisfying the rule IV.

4 Application to Nearest-Neighbor Classification

The theoretical similarity depends on the definition of the concept undertaken. In a classification task, such a definition is unknown. However, the previous similarity can be approximated from a dataset.

4.1 An Approximated Similarity Measure

The approximation relies on the ability to approximate each neighborhood by:

Definition 8. *The neighborhood of an example e with respect to a dataset D is:* $N_D(e) = \{e' \in D \mid G(\{e, e'\}) \textit{ is coherent with } D\}$.

Actually, for each exemplar e, the approximated neighborhood $N_D(e)$ converges toward $N_C(e) \cap D$, when the size of the dataset increases. This result follows from the proposition:

Proposition 1. *Let C be a well-defined concept, D a dataset and $e \in D$.*

1. $N_C(e) \cap D \subset N_D(e)$
2. *The probability to be in $N_D(e)$ but not in $N_C(e) \cap D$ decreases when the size of the dataset increases.*
3. *In the limit, D=E and $N_D(e) \subset N_C(e) \cap D$*

Proof. Proposition 1 follows from the monotonicity of G. Proposition 2 states that each generalization is more likely to be refuted when more exemplars are provided. When all the exemplars are provided, generalizations coherent with the dataset are also coherent with the concept, which explains proposition 3.

Therefore, for each exemplar e, the size of $N_C(e)$ is approximated by the average number of exemplars of D that belong to $N_D(e)$. Let us approximate the sizes of the intersection and of the union of two neighborhoods in the same way. The theoretical similarity is, then, approximated by:

Definition 9. *Considering two exemplars e and e', their similarity with respect to the dataset D is:* $Sim_D(e, e') = \frac{|N_D(e) \cap N_D(e')|}{|N_D(e) \cup N_D(e')|}$

4.2 IBLG Classification

On these considerations, we developed a nearest-neighbor classifier based upon the approximated similarity measure and called IBLG (Instance-Based Learning from Generalization). Each new example has several neighborhoods whether it is assumed to have a particular label or another. Hence, IBLG has to compute the nearest-neighbor for each of the possible neighborhoods and choose the nearest one. The pseudo-code of IBLG is shown below. Its complexity is $O(N^3)$ where N is the size of the dataset.

```
For each label $l$,
   initialize $N_l$ as an empty list.
For each exemplar $e = (x, l)$ in $D$,
   compute and add the neighborhood $N_D(e)$ to $N_l$

classify(example $x$)
   for each label $l$
      let $e$ be the exemplar $(x, l)$
      compute the neighborhood $N_D(e)$
      retrieve its nearest neighborhood $N_D(e')$ in $N_l$
      let $Sim_D(e, e')$ be the similarity of $x$ for $l$
   return a label of maximal similarity
```

4.3 Some Experimental Evidences

To validate our approach, some experiments have been carried out to compare IBLG with four other classifiers: CN2 (Clark and Niblett [4]) for rule induction, PEBLS (Cost and Salzberg [5]) and SCOPE (Lachiche and Marquis [7]) for nearest-neighbor. As default classifier, the nearest-neighbor classifier based upon the Hamming distance[1] has been chosen.

As IBLG has no parameter, we have chosen the default parameters of the other algorithms. However, SCOPE has three parameters that are automatically assessed to deal with noisy datasets. Here, datasets are non-noisy and these parameters left to their theoretical values.

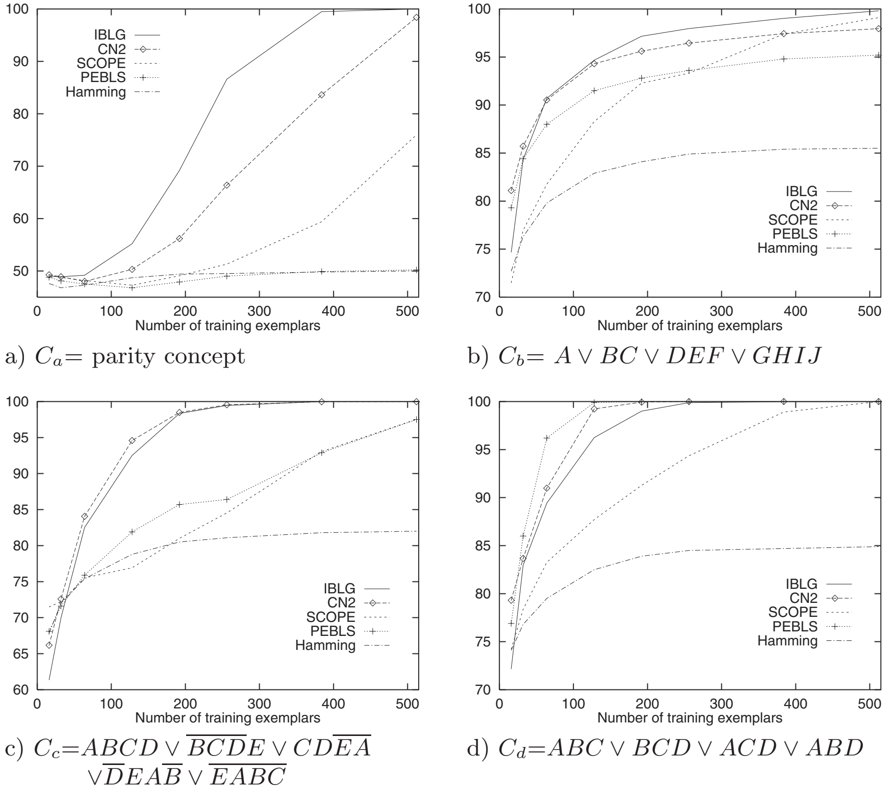
The experiments are summarized figure 1. IBLG appears to be less sensitive to the concept undertaken. Therefore, with respect to the other methods, IBLG performs best for complex concepts.

5 Related Research

5.1 SCOPE Classification

SCOPE (Lachiche and Marquis [7]) is a nearest-neighbor algorithm introduced in 1998. It classifies each new example according to the label of its most numerous

[1] The Hamming distance counts the number of features whose values are different.



a) C_a= parity concept

b) $C_b = A \vee BC \vee DEF \vee GHIJ$

c) $C_c = ABCD \vee \overline{BCD}E \vee CD\overline{EA}$
$\vee \overline{D}EA\overline{B} \vee \overline{EABC}$

d) $C_d = ABC \vee BCD \vee ACD \vee ABD$

Fig. 1. Experimental learning curves for IBLG when target concepts are less and less complex boolean functions of 10 boolean features. Each measure is the average classification accuracy on the unseen examples for 25 trials. The parity concept denotes the parity of the number of features whose value is true among the five first features.

neighborhood. IBLG chooses the label of the most similar known neighborhood. The improvement may appear rather small. However, each neighborhood (a set of exemplars) carries much more information than its size. And, in this paper, this information has been shown to be relevant from both theoretical and experimental points of view.

5.2 Feature Weighting Methods

The usual similarity measure is inversely correlated to the average distance between the values of each feature. However, when too many irrelevant features describe the examples, this similarity is irrelevant as well. The most studied solution is to weight the contribution of each feature to the overall similarity.

The problem is, then, to estimate from the dataset how relevant is a feature or even a value. For example, for the context similarity measure, Biderman ([2])

emphasizes that examples sharing a particular value are perceived more similar if this value is uncommon in the dataset. However, the problem of relevance is still open. The problems raised in this research area are reviewed in (Blum and Langley [3]) and the main contributions to nearest-neighbor methods in (Wettschereck et al. [10]).

For example, PEBLS (Cost and Salzberg [5]) is one of the state-of-the-art nearest-neighbor classifiers for symbolic features. It relies on the Value Difference Metric (Stanfill and Waltz [9]) and outperforms the Hamming classifier on most of the usual datasets but not all. The poor performances of PEBLS on the parity concept (cf fig. 1a) emphasize the difficulties encountered by this approach of similarity.

6 Conclusion

In this paper, we have introduced a new way to measure the similarity between some examples. This similarity measure has some theoretical advantages over the usual ones. Firstly, it becomes more and more accurate when the size of the dataset increases. And, in the limit, similar examples do have the same label. Therefore, convergence does not follow only from the ability to retrieve more and more similar examples. Secondly, this similarity is explanatory: it allows to build a rule-based representation of the concept undertaken. Determining whether these rules make an accurate rule-based classifier will be the scope of another paper. But, preliminary results are promising.

References

1. D. Aha, D. Kibler and M. Albert, Instance-based learning algorithms, *Machine Learning*, 6(1):37-66, 1991. 238
2. Y. Biderman, A Context Similarity measure, *Lecture Note in AI: ECML'94*, Springer Verlag, 1994. 244
3. A. L. Blum and P. Langley, Selection of relevant features and examples in machine learning, *Artificial Intelligence*, 245-271, Elsevier Science BV, 1997. 245
4. P. Clark and T. Niblett, The CN2 induction algorithm, *Machine Learning*, 3(4):261-283, 1989. 243
5. S. Cost and S. Salzberg, A weighted nearest-neighbor algorithm for learning with symbolic features, *Machine Learning*, 10:57-78, 1993. 243, 245
6. T. Cover and P. Hart, Nearest-neighbor pattern classification, *IEEE Transactions on Information Theory*, 13(1):21-27, 1967. 238
7. N. Lachiche and P. Marquis, Scope classification: An instance-based learning algorithm with a rule-based characterization, *In Proceedings of ECML'98*, pages 268-279, Chemnitz, Germany, LNAI 1398, Springer, 1998. 243
8. S. Salzberg, Learning with nested generalized exemplars, *Norwell*, MA: Kluwer Academic Publishers, 1990. 238
9. C. Stanfill, D. Waltz, Toward memory-based reasoning, *Communications of the ACM*, 29:1213-1228, 1986. 245
10. D. Wettschereck, D. W. Aha and T. Mohri, A Review and Empirical Evaluation of Feature Weighting Methods for a Class of Lazy Learning Algorithms, *Artificial Intelligence Review*, 11:273-314, 1996. 245