Decker, Gaul (Eds.): Classification and Information Processing at the Turn of the Millenium
Selected Papers from the 23th Annual Conference of the German Classification Society (GfKl)
Bielefeld, March 1999, ISBN 3-540-67589-2, pp. 254-263, ©Springer 2000.

Learning Complex Similarity Measures

B. Stein, O. Niggemann, U. Husemeyer

Dept. of Mathematics and Computer Science—Knowledge-based Systems
University of Paderborn, D–33095 Paderborn, Germany

Abstract: Case-based reasoning is a knowledge processing concept that has shown success in various problem classes. One key challenge in CBR is the construction of a measure that adequately models the similarity between two cases.

Typically, a similarity measure consists of a set of feature-specific distance functions coupled with an underlying feature weighting (importance) scheme. While the definition of the distance functions is often straightforward, the estimation of the weighting scheme requires a deep understanding of the domain and the underlying connections.

The paper in hand addresses this problem. It shows how discrimination knowledge, which is coded within an already solved classification problem, can be transformed towards a similarity measure. Moreover, it demonstrates our approach at the problem of diagnosing heart diseases.

1 Background and Related Theory

Discrimination knowledge that is coded within an already solved classification problem can be transformed towards a similarity measure of a case-based reasoning (CBR) system.

This chapter first points out relationships between classification and similarity assessment in a case-based reasoning system. It then motivates and defines a generic transformation procedure from a case base to a similarity measure; the last two sections of this chapter discuss related realizational aspects.

Chapter 2 presents an application, the diagnosis of heart diseases, to demonstrate the development of a similarity measure at a real-world problem.

1.1 Classification and Case-based Reasoning

Let x denote a problem or some description of a situation. Then a common task is to find another problem y amongst a set $\mathbf{S}$ of problems, such that y is more similar to x than it is to any other $z \in \mathbf{S}$.

Using the terminology of case-based reasoning, we are given a pair $\langle CB, sim \rangle$, where CB, the case base, denotes a set of cases, and sim denotes a similarity measure, $sim : CB \times CB \rightarrow [0, 1]$. With x, y, and $z \in CB$ the semantics of sim is as follows. $sim(x, y) > sim(x, z) \Leftrightarrow$ "*x is more similar to y than it is to z.*"

A case $x \in CB$ usually embodies both a problem description and a related solution. A basic idea of CBR is to exploit previously solved cases when solving a new problem. I. e., the collection of cases, CB, is browsed for the most similar case, whose solution then is adapted to solve the new problem.

However, within a classification task one is not interested in case adaptation: A case's solution simply defines a concept or a class. More precisely, the paper in hand considers cases as tuples $\langle \bar{x}, c_x \rangle$, where $\bar{x}$ denotes a feature vector, say, object description, and c_x, the solution, denotes a particular class or concept from a set of classes C.

Case-based classification classifies new instances based on their similarity to stored cases: When given a new feature vector $\bar{x}$, the set of the k most similar cases $y_1, \ldots, y_k$ retrieved from CB is used to define the class membership of $\bar{x}$.

Observe that case-based classification systems *implicitly* describe the classification concept, say, the discrimination knowledge. This knowledge is distributed over the containers *vocabulary*, *similarity measure*, and *case base*, where each container is able to contain all available knowledge (Richter (1995)).This view on knowledge distribution leads to the central contribution of this paper:

Given a knowledge base with cases of the form $\langle \bar{x}, c_x \rangle$, the feature-vector-class-relation can be used to derive a similarity measure.

Remarks. A classification task can also be tackled by a "symbolic" learning approach (Michalski et al. (1983)). By symbolic learning, Wess and Globig (1994) subsume approaches that code the knowledge of the cases *explicitly*, by means of a symbolic representation of the concept such as formulas or rules. Decision tree induction and version space are two representatives for algorithms that learn an explicit concept description (Quinlan (1986), Mitchell (1982)).

1.2 From a Case Base to a Similarity Measure

Let a case base CB with cases $\langle \bar{x}, c_x \rangle$—but *no* similarity measure be given. The question is, whether CB can be exploited to define such a measure. In the following, a generic procedure for this job is motivated.

Observe that for each two cases $\langle \bar{x}, c_x \rangle, \langle \bar{y}, c_y \rangle \in CB$ a reasonable similarity measure sim would produce a value $sim(\bar{x}, \bar{y})$ close to 1, if $c_x = c_y$ holds. Conversely, $sim(\bar{x}, \bar{y})$ would be closer to zero than to 1, if $c_x \neq c_y$ is true.

A similarity measure for a pair of feature vectors is typically based upon a metric M, which combines the distances between the features' instantiations.

If all features that describe the cases were of equal importance, and if they were homogeneous, continuous-valued, and canonically to normalize, the formation of sim would be straightforward: $sim(\bar{x}, \bar{y}) \equiv 1 - M(\bar{x}, \bar{y})$, where, for example, $M(\bar{x}, \bar{y})$ could denote the Euclidean distance metric $\sqrt{\sum_{i=1}^{|\bar{x}|}(x_i - y_i)^2}$.

However, usually the world is not that simple, and similarity measures are developed, tested, and improved by domain experts (Kolodner (1994)). A commonly used structure of a linear similarity measure is the following:

$$sim_l(\bar{x}, \bar{y}) := 1 - (w_0 + \sum_{i=1}^{|\bar{x}|} w_i \cdot \delta_i(x_i, y_i)), \tag{1}$$

where the w_i define the features' weights, so to speak, their importance, and the δ_i define feature-specific distance computations. Note that it is useful to introduce feature-specific distance functions to handle different types such as nominal, continuous-valued, or linear discrete features.

While distance functions can be operationalized canonically, an estimation of the w_i requires a deep understanding of the domain and the underlying connections. It is possible to estimate the w_i by exploiting the classification knowledge of the case base CB, using the following procedure:

(*i*) Construct a case base CB_Δ of "classified" difference vectors. Each element in CB_Δ is a tuple $\langle \bar{x} \ominus \bar{y}, P_{c_x=c_y} \rangle$, with

$$\begin{array}{ll} \bar{x} \ominus \bar{y} & := (\delta_1(x_1, y_1), \ldots, \delta_m(x_m, y_m)) \\ P_{c_x=c_y} & := 1, \text{if } x \text{ and } y \text{ belong to the same class and 0 otherwise} \end{array}$$

(*ii*) Approximate the relation $\bar{x} \ominus \bar{y} \mapsto P_{c_x=c_y}$, which is implicitly defined by CB_Δ, with the similarity measure $sim_l(\bar{x}, \bar{y})$ by means of regression.

Note that unlike other approaches, which use reinforcement learning to adjust or parameterize a similarity measure (e. g. ISAC Bonzano, Cunningham and Smyth (1997), EACH Salzberg (1991), RELIEF Kira and Rendell (1992), IB4 Aha (1992), GCM-ISW Aha and Goldstone (1992)), our approach transforms the problem onto a standard regression task. As a consequence, the parameterization of a similarity measure is no longer a special enhancement to a CBR system, but can be treated as a separate regression problem. This allows for a direct application of algorithms and results from the fields of statistics and machine learning to the field of automatic parameterization of similarity measures.

The following sections discuss the two steps of our procedure.

1.3 Distance Functions

The learning and classification capability of an instance-based learning system depends decisively on the similarity measure and its underlying distance function(s). Many distance functions are available for such uses, including the Euclidean, the Manhattan, the Canberra, or the Chi-square distance metrics (Wilson and Martinez (1997)).

A feature can be nominal (symbolic), linear discrete, or continuous-valued. Notice that none of the previously mentioned distance functions appropriately handles non-continuous features.

A metric that has been used for nominal features is the *overlap* metric, which is defined as follows.

$$overlap(x_i, y_i) := \begin{cases} 0, & \text{if } x_i = y_i \\ 1, & \text{otherwise} \end{cases}$$

One way to handle feature vectors with both nominal and continuous features is the combination of different distance functions within a heterogeneous distance function. The distance function δ_i^h, which is used in the IB systems (see Aha et al. (1991), Aha (1991)) and which is also used by Giraud-Carrier and Martinez (1995), combines the overlap metric with the range-normalized distance rn_Δ:

$$\delta_i^h(x_i, y_i) := \begin{cases} 1, & \text{if } x_i \text{ or } y_i \text{ is unknown} \\ overlap(x_i, y_i), & \text{if feature } i \text{ is nominal} \\ rn_{\Delta_i}, & \text{otherwise} \end{cases} \tag{2}$$

I. e., if either of the values of feature i is unknown, δ_i^h returns the maximum distance, which is 1. The distance function rn_{Δ_i} is defined as follows.

$$rn_{\Delta_i} := \frac{|x_i - y_i|}{\max_i - \min_i},$$

where $\max_i$ and $\min_i$ are the maximum and minimum values respectively, occurring in the training set for feature i. As a result, δ_i^h typically returns a value in $[0, 1]$.

The $\ominus$-operator can now be defined as

$$\bar{x} \ominus \bar{y} := (\delta_1^h(x_1, y_1), \ldots, \delta_m^h(x_m, y_m)),$$

with $\bar{x}$ and $\bar{y}$ denoting the (possibly heterogeneous) feature vectors of the cases $x, y \in CB$. The overall distance, say, the case distance or similarity between two cases x and y may be defined by means of a combination function $f(\bar{x} \ominus \bar{y})$, or by the Euclidean distance metric, or by some other non-linear function.Applying the Euclidean distance metric leads to the Heterogenous Euclidean-Overlap Metric $HEOM$ as described by Wilson and Martinez: $HEOM(\bar{x} \ominus \bar{y}) := \sqrt{\sum_{i=1}^{|\bar{x}|} \delta_i^h(x_i, y_j)}$

1.4 Computing a Weighting Scheme

Since features may differ in their discrimination quality, it is useful to incorporate a weighting scheme into the distance computation. As argued previously, such a weighting scheme can be estimated by regressing $P_{c_x=c_y}$ on $\bar{x} \ominus \bar{y}$, employing the similarity measure (1) as regression function.

This regression problem can be solved by either a statistical or a neural network (NN) approach. The similarity measure (1), for example, establishes a linear regression model and can be modeled as a perceptron with a linear activation function, as shown in Figure 1 on the left hand side (cf. Sarle (1994), Weisberg (1985), or Myers (1986)).

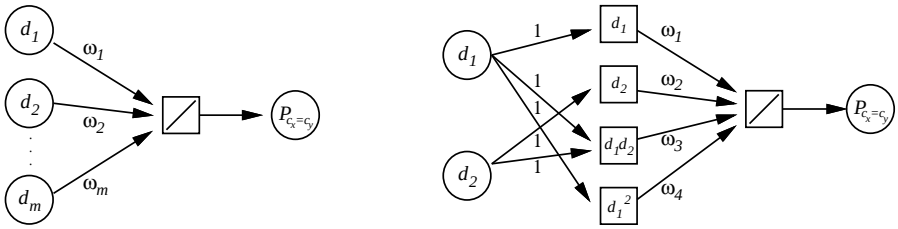

Figure 1: *A simple perceptron for multiple linear regression (left), and a functional link network for polynomial regression (right).*

Our classification problem is dichotomous, i. e., the observed outcome $P_{c_x=c_y}$ is restricted to either the value 0 or 1. Thus it makes sense to apply a logistic regression model, which—from the NN perspective—is modeled by a perceptron with a logistic activation function (cf. Hosmer and Lemeshow (1989)).

As well as that, a similarity measure in the form of a polynomial can be regressed by so-called *functional link networks* (Pao (1989)). Neural networks of this type provide a functional hidden layer that computes the polynomial terms from the input variables (see Figure 1, right hand side).

Remarks. Learning feature importance has also been discussed in Aha (1991). While our approach corresponds to a least squares estimation of the function defined through CB_Δ, Aha pursues in the system CBL4 a k-nearest-neighbor prediction along with some local weight correction. He plans future extensions of CBL4 being able to learn also "context-dependent" feature importance.

Observe that from a statistical perspective, context dependency of features means interaction amongst feature variables. Put another way, context dependency of features can be learned by our approach if the model behind *sim*, a polynomial for instance, reflects feature interaction. In fact, the application presented in this paper pursues this idea.

2 The Heart Diseases Application

Motivation and testbed for our similarity learning approach is a problem from the field of medical diagnosis: Identifying children's heart problems by analyzing examination records with physiological information (= feature vector $\bar{x}$). The respective cases were recorded at a local children's hospital and include a diagnosis as stated by a physician, such as "no sports" or "weight reduction necessary" (= class c_x).

Here, CBR can solve the following problems, which arise when applying a symbolic classification method to the field of medicine: (i) The identification

of explicitly coded expertise, e. g. in the form of rules, is very difficult. (*ii*) The acceptance of medical diagnosis systems, even when combined with strong explanation components, proved to be rather weak.

Also the physicians' requirement, namely the identification of the most similar case given a new case $y = \langle \bar{y}, ? \rangle$ instead of "simply" classifying $\bar{y}$, speaks for the CBR approach. Hence the "only" task to do is the development of an adequate similarity measure.

2.1 Developing a Tailored Similarity Measure

As discussed in Section 1, the development of a similarity measure from the heart diseases case base requires two steps: (*i*) Creating the difference vectors $\bar{x} \ominus \bar{y}$ for cases x, y, and (*ii*) choosing an appropriate structure for a similarity measure and computing its weights.

Since the greater part of the medical features are nominally-valued or cardinally-valued, the distance function (2) will serve most of our purposes. However, some of the features are "tree-valued", i. e., a respective feature value indicates a special taxonomic relationship.

For a feature i of such a type we define the difference between two values x_i and y_i by means of the normalized depth of their first common ancestor in the taxonomy tree. The first common ancestor of two vertices is the first common vertice on the paths from these vertices to the root.

$$tree_{\Delta_i}(x_i, y_i) := \begin{cases} 0, & \text{if } x_i = y_i \\ 1 - \frac{depth(anc(x_i, y_i))}{maxdepth_i}, & \text{otherwise} \end{cases}$$

Function $anc(x_i, y_i)$ denotes the first common ancestor of x_i and y_i, while $maxdepth_i$ denotes the maximum depth (= tree height) related to feature i. Figure 2 shows a fictitious tree-valued feature.

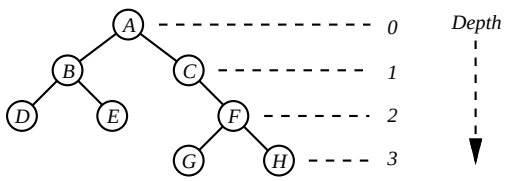

Figure 2: *A tree-valued feature with a maxdepth of* 3. *For instance, the distance between G and H as defined by* $tree_\Delta$ *is* 1/3, *the distance between D and E is* 2/3.

Having completed the heterogenous distance function δ_i^h from Section 1.3 by $tree_{\Delta_i}$, we are ready to construct the case base of difference vectors, CB_Δ. Recall that a case in CB_Δ is of the form $\langle \bar{x} \ominus \bar{y}, P_{c_x = c_y} \rangle$, with

$$\begin{aligned} \bar{x} \ominus \bar{y} &:= (\delta_1^h(x_1, y_1), \ldots, \delta_m^h(x_m, y_m)) \equiv (d_1, \ldots, d_m) \equiv \bar{d} \\ P_{c_x = c_y} &:= 1, \text{if } x \text{ and } y \text{ belong to the same class and 0 otherwise} \end{aligned}$$

As mentioned in Section 1.2, the function most commonly used as a similarity measure is

$$sim_l(\bar{x}, \bar{y}) := 1 - (w_0 + \sum_{i=1}^{|\bar{x}|} w_i \cdot \delta_i^h(x_i, y_i)) \equiv 1 - (w_0 + \sum_{i=1}^{|\bar{d}|} w_i \cdot d_i) \quad (3)$$

The w_i weight the features und thus are used to model the respective feature's importance. Notice that by using this function the independence between features is implicitly presumed. Stated another way, the importance of a feature does not depend on other feature values—an assumption which does not hold in many real world applications.

Also in the field of children heart disease diagnosis various interdependencies between features can be observed: For example the features "age" and "weight" correlate to a high degree—as do the features "weight" and "maximum strain during examination".

To express inter-feature dependency, a stronger and more complex similarity measure of the following structure is needed:

$$sim(\bar{x}, \bar{y}) := 1 - (w_0 + \sum_{i=1}^{|\bar{d}|} f_i(d_1, \ldots, d_{i-1}, d_{i+1}, \ldots, d_m) \cdot d_i) \quad (4)$$

As a consequence, the functional connections f_i (instead of the simple weights w_i) must be estimated in order to define the similarity measure sim. Since the analysis of our data revealed that the majority of existing feature correlations establish pairwise interdependencies, it was reasonable to restrict the f_i to linear type:

$$f_i := \lambda_{i0} + \sum_{\substack{j=1 \\ j \neq i}}^{|\bar{d}|} \lambda_{ij} \cdot d_j \quad (5)$$

Putting (5) into (4) while substituting w_i for λ_{i0} and w_{ij} for $\lambda_{ij} + \lambda_{ji}$ yields:

$$sim(\bar{x}, \bar{y}) := 1 - (w_0 + \sum_{i=1}^{|\bar{d}|} w_i \cdot d_i + \sum_{i=1}^{|\bar{d}|} \sum_{j=i+1}^{|\bar{d}|} w_{ij} \cdot d_i d_j) \quad (6)$$

Actually, (6) is the similarity measure that we have presumed in the heart disease domain. Its corresponding weights can be estimated as described in Section 1.2 by regressing $P_{c_x=c_y}$ onto $sim(\bar{x}, \bar{y})$, using the data base CB_Δ as set of observations.

Technically, the regression is realized by a functional link network with a logistic activation function. The network's functional hidden layer models the $(|\bar{d}|^2 + |\bar{d}|)/2$ polynomial terms of (6); using superwised learning the network was trained by least squares. Chapter 3 contains figures of the underlying technical data as well as the experimental results regarding sim's classification quality.

2.2 Feature Selection as a Means of Preprocessing

Methods for determining a weighting scheme are judged by both the qualitity of the learned weights, say, the learning and classification error, and the effort necessary for computing them.

For the proposed similarity measure (6) the latter criterion may cause problems since $O(|\bar{x}|^2)$ weights must be estimated. However, in many real-world problems the discrimination knowledge is not equally distributed on all features. Feature selection, i. e., leaving out features by setting their weights constantly to zero, provides one way to integrate additional knowledge into the learning process.

The naive approach of testing all $2^{|\bar{x}|}$ feature subsets is usually not practicable, and various algorithms for selecting good subsets have been developed. They consist of a generation function, responsible for feature-subset creation, and an evaluation function for evaluating this subset according to a given criterion (Dash and Liu (1997), Aha and Bankert (1994)). Evaluation functions in turn are either realized as filter methods, which work independently from the classifier, or as wrapper methods, which employ the classifier for testing purposes.

For the domain of heart diseases we pursue a wrapper approach: The absolute values of the learned weights are used for selecting important features. This obviously makes sense because of the direct correlation between weights and feature distances in the structure of the discussed similarity measures.

3 Results and Conclusion

Our database with cases of heart disease diagnosis comprises more than 200 cases; a case consists of 50 features, from which are 10 nominal, 35 cardinal, and 5 hierarchical. These cases are (rather equally) distributed amongst 8 diagnosis classes. We divided this database into a training set and a test set, from which, in a second step, two sets with about 20.000 difference vectors were formed.

Regression was carried out as previously described, and, for comparison purposes, both the simple similarity measure (3) without—as well as the complex similarity measure (6) with interaction terms was learned. Moreover, variants of these measures with only 10 features were constructed according to the outlined feature selection strategy. Applied to the test set, the measures led to classification results as shown in Figure 3.

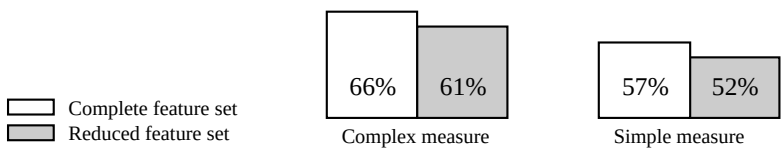

Figure 3: *Classification results of different similarity measures. Note that "misclassified" cases were not assessed completely wrong but assigned to neighbored classes.*

In a nutshell, the paper presents a generic method for learning similarity measures for case-based reasoning by analyzing a given case base.

Starting point was the observation that the weighting scheme behind a similarity measure is essential to adapt CBR to different domains—but hard to be set up manually, because of its need for domain knowledge.

The paper shows how such weighting schemes can be derived automatically. It motivates the method theoretically and shows how it is applied to the difficult domain of medical diagnosis.

References

AHA, D. (1992): Tolerating noisy, irrelevant, and novel attributes in instance-based learning algorithms. *International Journal of Man-Machine Studies*.

AHA, D. W. (1991): Case-Based Learning Algorithms. Proceedings of the 1991 DARPA Case-Based Reasoning Workshop, Morgan Kaufmann.

AHA, D. W. and BANKERT, R. L. (1994): Feature Selection for Case-Based Classification of Cloud Types: An Empirical Comparison. Proceedings of the AAAI-94 Workshop on Case-Based Reasoning, AAAI Press, Seattle, WA, pp. 106–112.

AHA, D. W. and GOLDSTONE, R. (1992): Concept learning and flexible weighting. Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society.

AHA, D. W., KIBLER, D. and ALBERT, M. K. (1991): Instance-Based Learning Algorithms. *Machine Learning*, 6, 37–66.

BONZANO, A., CUNNINGHAM, P. and SMYTH, B. (1997): Using introspective learning to improve retrieval in cbr: A case study in air traffic control. Proceedings of the Second ICCBR Conference.

DASH, M. and LIU, H. (1997): Feature Selection for Classification. *Intelligent Data Analysis*.

GIRAUD-CARRIER, C. and MARTINEZ, T. (1995): An Efficient Metric for Heterogeneous Inductive Learning Applications in the Attribute-Value Language. *Intelligent Systems*, pp. 341–350.

HOSMER, D. W. and LEMESHOW, S. (1989): Applied Logistic Regression. John Wiley & Sons, New York.

KIRA, K. and RENDELL, L. (1992): A practical approach to feature selection. Proceedings of the Ninth International Conference on Machine Learning.

KOLODNER, J. (1994): Case-Based Reasoning. Morgan Kaufmann.

MICHALSKI, R., CARBONELL, J. G. and MITCHELL, T. (Eds.) (1983): Machine Learning: An Artificial Intelligence Approach. Vol. 1, Tioga, Palo Alto, California.

MITCHELL, T. M. (1982): Generalization as search. *Artificial Intelligence*, 18 (2), 203–226.

MYERS, R. H. (1986): Classical and Modern Regression with Applications. Duxbury Press, Boston.

PAO, Y.-H. (1989): Adaptive Pattern Recognition and Neural Networks. Addison-Wesley, Reading, MA.

QUINLAN, J. R. (1986): Induction of Decision Trees. *Machine Learning*, 1 (1), 81–106.

RICHTER, M. M. (1995): The Knowledge Contained in Similarity Measures. Some remarks on the invited talk given at ICCBR'95 in Sesimbra, Portugal.

SALZBERG, S. L. (1991): A nearest hyperrectangle learning method. *Machine Learning*.

SARLE, W. S. (1994): Neural Networks and Statistical Models. Proceedings of the Nineteenth Annual SAS Users Group International Conference, SAS Institute Inc., Cary, NC, USA, pp. 1538–1550.

WEISBERG, S. (1985): Applied Linear Regression. John Wiley & Sons, New York.

WESS, S. and GLOBIG, C. (1994): Case-Based and Symbolic Classification—A Case Study. S. WESS, K.-D. ALTHOFF and M. M. RICHTER (Eds.), Topics in Case-Based Reasoning, LNAI 837, Springer-Verlag.

WILSON, D. R. and MARTINEZ, T. R. (1997): Improved Heterogeneous Distance Functions. *Journal of Artificial Intelligence Research*.