Pp. 67–72 in *Artificial Intelligence Applications* (Proceedings of the EXPERSYS-97 Conference), edited by P.Smith. IITT International, Gournay sur Marne, 1997.

A Bayesian Approach for Retrieving Relevant Cases

Petri Kontkanen, Petri Myllymäki, Tomi Silander, Henry Tirri
Complex Systems Computation Group (CoSCo)
P.O.Box 26, Department of Computer Science
FIN-00014 University of Helsinki, Finland
Email: cosco@cs.Helsinki.FI
URL: http://www.cs.Helsinki.FI/research/cosco/

Abstract

The problem of finding the set of most relevant cases from a given database, with respect to the decision making situation at hand, is frequently encountered in many real-world domains. In the case-based reasoning framework this task is commonly known as the case matching problem. Case matching is an important problem in several commercially significant application areas, such as industrial configuration and manufacturing problems. Earlier approaches to the case matching problem typically rely on some distance measure, e.g., the Euclidean distance, although there is no a priori guarantee that such measures really reflect the useful similarities and dissimilarities between the cases. In this paper we introduce a novel approach to the case matching problem based on Bayesian probability theory, and propose a Bayesian case matching measure for scoring the cases with respect to a given decision making situation. The Bayesian case matching score discussed is currently being applied in a real-world manufacturing problem. The empirical results show that when encountered with cases where many of the feature values have been removed, even a relatively small number of remaining values is sufficient for retrieving the original case from the case base by using the proposed measure. The experiments also show that the approach is computationally very efficient.

Keywords: Case-based reasoning, Probability theory, Bayesian similarity metrics

1 Introduction

Any given database of feature vectors can be seen as a sample from a problem domain, where the sampled cases are drawn from some unknown problem domain probability distribution. In traditional machine learning, this distribution is typically approximated by using a single model, e.g., a decision-tree or a neural network, which is constructed by using the given sample of problem domain data. In contrast to this model-based approach, in the *case-based reasoning (CBR)* approach [1, 3, 14, 18, 21], the learning algorithms base their predictions directly on the sample data, without producing any specific models of the problem domain. This type of machine learning is often also referred to as *lazy learning*, since the algorithms defer all the essential computation until the prediction phase [2].

For making predictions, the CBR algorithms typically use a distance function (e.g., Euclidean distance) for scoring the cases, i.e., determining the most relevant data items for the prediction task in question (case matching). Some simple function, such as majority voting in classification problems, can then used for determining the prediction from the most relevant data items (case adaptation). It has been shown in various studies (see e.g., [17] for references)

that this type of an approach can in some cases produce quite accurate predictions, when compared to alternative machine learning methods. However, the performance of the algorithms seems to be highly sensitive to the selection of the distance function [4, 11].

In [23, 22, 20, 19] we proposed a Bayesian framework for CBR based on probability theory and the finite mixture model family [8, 24]. The approach suggested can be seen as a *partially lazy approach* [2], i.e., a hybrid between the traditional machine learning and the lazy learning approach, which is based solely on the given data. The studies were based on the probabilistic viewpoint, where groups of data vectors are transformed into distributions, which can be seen as sample points in a distribution space. The predictive distributions required for making predictions could then be computed by using the CBR approach in this distribution space, i.e., by introducing a probabilistic "distance metric". Somewhat similar frameworks has been suggested in [9, 10, 13].

In [15] we presented a new, improved probabilistic formalization of CBR of a purely lazy nature. The framework extended our earlier results by presenting a Bayesian approach for making (discrete) predictions directly from data, *without the transformation step between the original sample space and the distribution space*. Intuitively speaking, the new approach is based on the following notion: if we wish to make predictions by using only the sample data given, avoiding the notion of individual models, from the Bayesian point of view we can take this as a requirement for marginalizing, i.e., integrating, out the individual models. Quite interestingly, in [15] we showed that with the Bayesian CBR approach, the case adaptation problem can be solved directly in one pass, without the standard two-phase CBR methodology where the cases are first ordered according to some scoring function, and the best cases are then used in the adaptation phase for making predictions. Nevertheless, there are many situations where it would also be important to retrieve the most relevant cases from a given database. For example, the list of the relevant cases could be used by an explanation mechanism for increasing the credibility of the predictions. In Section 2 we describe a probabilistic solution to the case matching problem. In the empirical part of the paper, the proposed Bayesian case matching technique is evaluated empirically by using publicly available real-world databases.

2 Bayesian case matching

Similarly to most of the prediction oriented CBR research, we restrict ourselves in the following to the situation where the cases are represented as feature vectors. From the Bayesian perspective, the case base D denotes a random sample of N i.i.d. (independent and identically distributed) data vectors $\vec{d}_1, \ldots, \vec{d}_N$ from some unknown probability distribution $\mathcal{P}(\vec{d})$. For simplicity, we assume that the data is coded by using discrete, finite-valued, attributes from a set $\mathrm{X} = \{X_1, \ldots, X_m\}$, although the Bayesian approach described here can be extended to continuous attributes as well. In probabilistic modeling we regard each attribute X_i as a random variable with possible values from the set $\{x_{i1}, \ldots, x_{in_i}\}$.

In the following, let us restrict the set of possible probability distributions for modeling $\mathcal{P}(\vec{d})$ by a set of assumptions Ψ. It should be observed that in practice we are always forced to make some restricting assumptions about the problem domain, although this fact is not always explicitly recognized. For example, in many traditional CBR systems the algorithms typically use a distance function (e.g., Euclidean distance) for determining the most relevant data items for the prediction task in question. The use of a specific distance function implicitly assumes that the distance function is relevant with respect to the problem domain probability distribution, and hence restricts the set of distributions considered.

In traditional machine learning, the assumptions Ψ are typically used for defining a model class $\mathcal{M}$, i.e., a set of parametric models Θ, where each model (parameter instantiation) determines one probability distribution on the problem domain. Given some data D, the

goal is then to find such a single model $\Theta(D)$ that the corresponding probability distribution $P(\vec{d}|\Theta(D))$ approximates $\mathcal{P}(\vec{d})$ as closely as possible. In contrast to this model-based approach, in the case-based reasoning approach the learning algorithms base their predictions directly on the sample data, without producing any specific models of the problem domain. From the Bayesian point of view, we can regard the CBR approach as a requirement for marginalizing, i.e., integrating, out all the individual models. This means that the problem domain probability distribution is approximated by

$$\mathcal{P}(\vec{d}) \approx P(\vec{d} \mid D, \Psi) = \int_{\Theta \in \mathcal{M}} P(\vec{d} \mid \Theta) P(\Theta \mid D) d\Theta. \quad (1)$$

In [15] we showed how formula (1) can be used for solving case adaptation tasks, i.e., for predicting the values of the unknown variables, denoted here by V, given the values of the known values U = X \ V, and the sample data D. This approach does not follow the standard two-phase CBR methodology, since the case adaptation problem is solved directly without a preceding case matching phase where the cases are ordered according to some scoring function. Nevertheless, there are many situations where it would be important to retrieve the most relevant cases from a given database D, given a query $\vec{U} = \vec{u}$. For example, the list of the relevant cases could be used by an explanation mechanism for increasing the credibility of the predictions. For retrieving relevant cases, we propose here the following Bayesian case matching scoring function $\mathcal{S}(\vec{d_j}|\vec{u})$:

$$\mathcal{S}(\vec{d_j} \mid \vec{u}) = P(\vec{u} \mid \vec{d_j}, D, \Psi) = \int_{\Theta \in \mathcal{M}} P(\vec{u} \mid \Theta) P(\Theta \mid \vec{d_j}, D) d\Theta. \quad (2)$$

Intuitively speaking, the suggested case matching score ranks the cases according to the following question: "Which of the cases in D should be duplicated if one wishes to maximize the probability of the given value assignment $\vec{U} = \vec{u}$?". On the other hand, noting that $\mathcal{S}(\vec{d_j}|\vec{u}) \propto P(\vec{d_j}|\vec{u}, D, \Psi)/P(\vec{d_j}|D, \Psi)$, we see that the cases are ranked according to their "posterior" probability (probability given the query and the sample data), normalized by their "prior" probability (probability given the sample data without a query). This means that also cases with very low initial probability can get a high score if they match the given query well, which is also intuitively reasonable.

It should be noted that the scoring metric suggested is "soft" in the sense that the winning case may violate one or more of the variable-value assignments in the query $\vec{U} = \vec{u}$. It goes without saying that if the query $\vec{u}$ can not be regarded as a "wish-list", but as a set of absolute constraints which should not be violated, then this type of hard constraint situation can be handled easily by restricting the search for the matching case to those cases consistent with the given query.

3 Empirical results

To validate the suggested Bayesian case matching score (2), we used several public domain classification data sets of varying size from the UCI data repository[1]. In order to be able to use the suggested case matching score, the assumptions Ψ have to be such that the integral in (2) can be solved analytically, or at least approximated well. In the classification framework used for our experiments, we used the following assumptions. First of all, the variables were assumed to be independent, given the value of the class variable, denoted here by X_m (this same assumption is used in constructing a model called the *Naive Bayes classifier*). In this

[1] "http://www.ics.uci.edu/~mlearn/".

case, the joint probability distribution for a data vector $\vec{d}$ can be written as

$$P(\vec{d}|\Theta) = P(X_m = x_m) \prod_{i=1}^{m-1} P(X_i = x_i \mid X_m = x_m).$$

Consequently, a single distribution P can be uniquely determined by fixing the values of the parameters $\Theta = (\alpha, \Phi)$, where $\alpha = (\alpha_1, \ldots, \alpha_K)$ and $\Phi = (\Phi_{11}, \ldots, \Phi_{1,m-1}, \ldots, \Phi_{K1}, \ldots, \Phi_{K,m-1})$. Here K denotes the number of values of the class variable X_m, α_k the probability $P(X_m = k)$, and $\Phi_{ki} = (\phi_{ki1}, \ldots, \phi_{kin_i})$, where $\phi_{kil} = P(X_i = x_{il} \mid X_m = x_m)$. In the following we further assume that $\alpha_k > 0$ and $\phi_{kil} > 0$ for all k,i, and l, and that both the class variable distribution $P(X_m)$ and the intra-class conditional distributions $P(X_{i|k}) = P(X_i \mid X_m = k)$ are multinomial, i.e., $X_m \sim \text{Multi}(1; \alpha_1, \ldots, \alpha_K)$, and $X_{i|k} \sim \text{Multi}(1; \phi_{ki1}, \ldots, \phi_{kin_i})$. Since the family of Dirichlet densities is *conjugate* (see e.g. [7]) to the family of multinomials, we assume that $(\alpha_1, \ldots, \alpha_K) \sim \text{Di}(\mu_1, \ldots, \mu_K)$, and $(\phi_{ki1}, \ldots, \phi_{kin_i}) \sim \text{Di}(\sigma_{ki1}, \ldots, \sigma_{kin_i})$, where $\{\mu_k, \sigma_{kil} \mid k = 1, \ldots, K; i = 1, \ldots, m-1; l = 1, \ldots, n_i\}$ are the *hyperparameters* of the corresponding distributions. Further assuming that the parameter vectors α and Φ_{ki} are independent, and applying the results in [6, 12], the Bayesian case matching score (2) for case $\vec{d_j}$ can be computed by

$$\mathcal{S}(\vec{d_j}|\vec{u}) = \sum_{k=1}^{K} P(\vec{u}, X_m = k \mid \vec{d_j}, D, \Psi) = \sum_{k=1}^{K} \frac{h_k + \mu_k}{N + \sum_{c=1}^{K} \mu_c} \prod_i \frac{f_{kiu_i} + \sigma_{kiu_i}}{h_k + \sum_{l=1}^{n_i} \sigma_{kil}}, \tag{3}$$

where the product goes over all the variables instantiated in the query $\vec{u}$, and h_k and f_{kil} are the sufficient statistics of the data $D \cup \vec{d_j}$, i.e., h_k is the number of cases where $X_m = k$, and f_{kil} is the number of cases where $X_m = k$ and $X_i = x_{il}$. In our experiments we have used the noninformative uniform prior: all hyperparameters μ_k and σ_{kij} were set to 1.

To validate the Bayesian approach to the case matching problem, we used the following experimental setup. The goal was to recover cases in a situation where one or more elements of the case vector was destroyed. More precisely, we first selected a case $\vec{d_t}$ from the case base D to be the *target case*, the vector we were trying to recover. We then removed randomly r elements from $\vec{d_t}$, and used the remaining variable settings as a query $\vec{u}$. Using the resulting query, we computed for each datavector $\vec{d_j} \in D$ the Bayesian matching score $\mathcal{S}(\vec{d_j}|\vec{u})$ according to (3), and the cases were ordered in descending order with respect to the score obtained. By the *rank* we mean in the following the order number of the target case in this ordering. Consequently, obtaining rank value 1 means that the original target case has got the highest score, hence we have been able to retrieve it despite of the r missing feature values introduced. The *relative matching score* is defined as the ratio between the score of the target case, and the score of the first case in the ordering, normalized so that the average over all the cases is 0.0, and the maximum is 1.0. The average rank and average relative score were computed by performing the above procedure 100 times with different r random missing elements, and the same was repeated by using in turn each of N cases as the target case.

In Figure 1, the average ranks are plotted as a function of the number of the fixed variables. As can be seen in this picture, with all eight case bases, a relatively small number of fixed variables is sufficient to obtain a rank near 1, which corresponds to retrieving the original target case. Furthermore, the average relative matching scores show that even with a low number of fixed variables (when the rank is far from 1), the target case gets a very high probability compared to the average case probability.

It should be noted that by *not* restricting the search to those cases consistent with the given query, we allowed the assignments $\vec{u}$ to be violated, although it is easy to see that in the artificial case retrieval framework described above, the hard constraint approach would always have given better results. At first sight it may seem that a more appropriate setup for testing

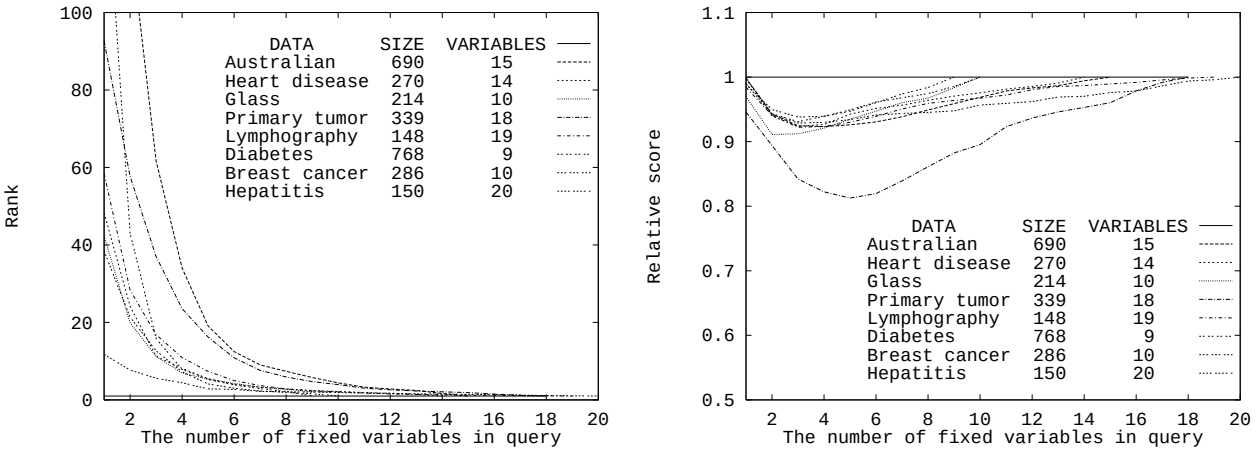

Figure 1: The average ranks (left) and the relative matching scores (right) as a function of the number of the fixed variables $(m - r)$ in the query $\vec{u}$.

the soft constraint case could have been obtained by producing the queries by adding noise to the target cases, but in this case there is no reason why a noisy version of the target case should be closest to the original target in any similarity metric. A more natural experimental testing of the soft constraint approach is currently in progress with one of our industrial partners by using real-world data and problem domain expert evaluations.

To get an idea of the implementation performance for Bayesian case matching, we performed an experiment where we generated a case base of 2 million data vectors, each consisting of 14 discrete attributes having on the average 3-4 values, by duplicating the Heart disease case base several times. Using a 200MHz Pentium machine running Linux, matching a given query against the whole case base, corresponding to performing 2 million case matches, took 27.23 seconds, i.e., about 13.6 microseconds per matching operation (approximately 74000 matches per second). The size of this case base serves also as an example of the scalability of the approach.

4 Conclusion

We have argued that Bayesian probability theory can be used as a formalization for the intuitively appealing case-based reasoning paradigm. In this paper we concentrated on the case matching problem, and proposed a probabilistic scoring metric for this task. In the experimental part of the paper, the Bayesian case matching score was evaluated empirically by using publicly available real-world case bases. The results show that when encountered with cases where some of the feature values have been removed, a relatively small number of remaining values is sufficient for retrieving the original case from the case base by using the proposed measure. It should also be emphasized that although the particular set of assumptions used in the experiments may at first sight appear overly simplifying, the empirical results [16] show that with such assumptions the approach may perform surprisingly well in terms of prediction accuracy, when compared to the results obtained by alternative methods. The experiments also showed that the BCBR approach is computationally very efficient.

Acknowledgments. This research has been supported by the Technology Development Center (TEKES), and the Academy of Finland.

References

[1] D. Aha. *A Study of Instance-Based Algorithms for Supervised Learning Tasks: Mathematical, Empirical, an Psychological Observations.* PhD thesis, University of California, Irvine, 1990.

[2] D. Aha. Editorial for a special issue on lazy learning. *Artificial Intelligence Review (to appear)*, 1997.

[3] C. Atkeson. Memory based approaches to approximating continuous functions. In M. Casdagli and S. Eubank, editors, *Nonlinear Modeling and Forecasting. Proceedings Volume XII in the Santa Fe Institute Studies in the Sciences of Complexity.* Addison Wesley, New York, NY, 1992.

[4] C. Atkeson, A. Moore, and S. Schaal. Locally weighted learning. *AI Review to appear*, 1995.

[5] J.O. Berger. *Statistical Decision Theory and Bayesian Analysis.* Springer-Verlag, New York, 1985.

[6] G. Cooper and E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347, 1992.

[7] M.H. DeGroot. *Optimal statistical decisions.* McGraw-Hill, 1970.

[8] B.S. Everitt and D.J. Hand. *Finite Mixture Distributions.* Chapman and Hall, London, 1981.

[9] D. Fischer. Noise-tolerant conceptual clustering. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 825–830, Detroit, Michigan, 1989.

[10] D. Fischer and D. Talbert. Inference using probabilistic concept trees. In *Proceedings of the Sixth International Workshop on Artificial Intelligence and Statistics*, pages 191–202, Ft. Lauderdale, Florida, January 1997.

[11] J.H. Friedman. Flexible metric nearest neighbor classification. Unpublished manuscript. Available by anonymous ftp from Stanford Research Institute (Menlo Park, CA) at playfair.stanford.edu., 1994.

[12] D. Heckerman, D. Geiger, and D.M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20(3):197–243, September 1995.

[13] S. Kasif, S. Salzberg, D. Waltz, J. Rachlin, and D. Aha. Towards a framework for memory-based reasoning. Manuscript, in review., 1995.

[14] J. Kolodner. *Case-Based Reasoning.* Morgan Kaufmann Publishers, San Mateo, 1993.

[15] P. Kontkanen, P. Myllymäki, T. Silander, and H. Tirri. Bayes optimal lazy learning. Technical Report NC-TR-97-031, ESPRIT Working Group 8556: Neural and Computational Learning (NeuroCOLT), 1997.

[16] P. Kontkanen, P. Myllymäki, T. Silander, H. Tirri, and P. Grünwald. Comparing predictive inference methods for discrete domains. In *Proceedings of the Sixth International Workshop on Artificial Intelligence and Statistics*, pages 311–318, Ft. Lauderdale, Florida, January 1997.

[17] D. Michie, D.J. Spiegelhalter, and C.C. Taylor, editors. *Machine Learning, Neural and Statistical Classification.* Ellis Horwood, London, 1994.

[18] A. Moore. Acquisition of dynamic control knowledge for a robotic manipulator. In *Seventh International Machine Learning Workshop.* Morgan Kaufmann, 1990.

[19] P. Myllymäki and H. Tirri. Bayesian case-based reasoning with neural networks. In *Proceedings of the IEEE International Conference on Neural Networks*, volume 1, pages 422–427, San Francisco, March 1993. IEEE, Piscataway, NJ.

[20] P. Myllymäki and H. Tirri. Massively parallel case-based reasoning with probabilistic similarity metrics. In S. Wess, K.-D. Althoff, and M Richter, editors, *Topics in Case-Based Reasoning*, volume 837 of *Lecture Notes in Artificial Intelligence*, pages 144–154. Springer-Verlag, 1994.

[21] C. Stanfill and D. Waltz. Toward memory-based reasoning. *Communications of the ACM*, 29(12):1213–1228, 1986.

[22] H. Tirri, P. Kontkanen, and P. Myllymäki. A Bayesian framework for case-based reasoning. In I. Smith and B. Faltings, editors, *Advances in Case-Based Reasoning*, volume 1168 of *Lecture Notes in Artificial Intelligence*, pages 413–427, Proceedings of the 3rd European Workshop, Lausanne, Switzerland, November 1996. Springer-Verlag, Berlin Heidelberg.

[23] H. Tirri, P. Kontkanen, and P. Myllymäki. Probabilistic instance-based learning. In L. Saitta, editor, *Machine Learning: Proceedings of the Thirteenth International Conference*, pages 507–515. Morgan Kaufmann Publishers, 1996.

[24] D.M. Titterington, A.F.M. Smith, and U.E. Makov. *Statistical Analysis of Finite Mixture Distributions.* John Wiley & Sons, New York, 1985.