Pp. 507–515 in *Machine Learning: Proceedings of the Thirteenth International Conference*, edited by L. Saitta. Morgan Kaufmann Publishers, San Francisco, CA, 1996.

Probabilistic Instance-Based Learning

Henry Tirri **Petri Kontkanen** **Petri Myllymäki**
Complex Systems Computation Group (CoSCo)
P.O.Box 26, Department of Computer Science
FIN-00014 University of Helsinki, Finland
{Henry.Tirri,Petri.Kontkanen,Petri.Myllymaki}@cs.Helsinki.FI

Abstract

Traditional instance-based learning methods base their predictions directly on (training) data that has been stored in the memory. The predictions are based on weighting the contributions of the individual stored instances by a distance function implementing a domain-dependent similarity metrics. This basic approach suffers from three drawbacks: computationally expensive prediction when the database grows large, overfitting in the presence of noisy data, and sensitivity to the selection of a proper distance function. We address all these issues by giving a probabilistic interpretation to instance-based learning, where the goal is to approximate predictive distributions of the attributes of interest. In this probabilistic view the instances are not individual data items but probability distributions, and we perform Bayesian inference with a mixture of such prototype distributions. We demonstrate the feasibility of the method empirically for a wide variety of public domain classification data sets.

1 Introduction

Traditional instance-based learning methods (Stanfill and Waltz, 1986; Moore, 1990; Aha, 1990; Atkeson, 1992) are a family of learning algorithms which base their predictions directly on (training) data that has been stored in the memory[1]. In their basic form instance-based methods store all the training data in the memory during the learning phase, and at prediction phase use a distance function to determine which data items are relevant for prediction. The predictions of the individual items are then combined, for example by using averaging. This type of algorithms are often called "lazy" as they defer all the essential computation until the prediction phase. Examples of instance-based learning methods are k-nearest neighbor (Cover and Hart, 1967), kernel regression (Franke, 1982) and locally weighted regression (for a recent survey see (Atkeson et al., 1995)). Some neural network approaches can also be viewed as instance-based methods, e.g., the family of Radial Basis Function networks (Moody and Darken, 1989) and Probabilistic networks (Specht, 1990).

This basic approach of storing all the data items at the learning phase suffers from several drawbacks. First, the run-time computational costs for such algorithms are high when the size of the data set grows large. Second, saving all the data items leads easily in overfitting in the prediction phase. Finally, the performance of the instance-based method is sensitive to the selection of a proper, possibly varying, distance function (Friedman, 1994; Atkeson et al., 1995). The high run-time costs due to storing of all the data has lead to methods that attempt to find a smaller set of "prototypes" (Aha, 1990) that represent the data set without sacrificing prediction accuracy. This "reference selection problem" has been addressed by pruning (i.e., storing typical instances) (Zhang, 1992), by exploiting domain knowledge (Kurtzberg, 1987) and by stochastic techniques that perform search in the space of sets of prototypes (Skalak, 1994). A somewhat different approach is presented in (Deng and Moore, 1995) where the data items are grouped by using *kd*-trees thus allowing predictions with costs proportional to the number of groups instead of the number of individual ele-

[1] For this reason they are also known as "memory-based" methods and in more structured domains as "case-based" methods.

ments. Attempts to avoid overfitting have been based on pruning reference items that cause misclassifications and by storing abstractions of instances (Aha, 1990). The problem of proper selection of a distance function is extensively studied in (Friedman, 1994).

The approach presented in this paper is based on the probabilistic viewpoint, where the attributes A_i are interpreted as random variables, and the given set of training instances is used to approximate the underlying joint probability distribution of the attributes. Adopting this point of view allows the following probabilistic interpretation to the prediction phase procedures used in instance-based learning: Given an attribute vector with some unknown attributes, the predictive distributions for these attributes are constructed by summing the predictions of individual instances weighted by the distance function. In this view each instance can be seen as a component distribution, which contributes to the joint distribution, i.e., the joint distribution is represented as a mixture of "instance distributions". Therefore the basic instance-based approach in probabilistic terms can be understood as a form of kernel density estimators (see e.g., (Scott, 1992)), and is thus similar to Radial Basis Function network based estimation.

In practice a case where all the instance distributions are needed for a good approximation is very extreme as the instance set usually exhibits some cluster structure. Therefore usually the joint distribution can be approximated by a simpler mixture of distributions by giving a weighted sum of "cluster distributions", each of which gives the marginal attribute probability distributions conditioned by the cluster index. In fact, construction of prototypes by averaging (Aha, 1990) can be understood as coarse grain approximations for such simpler mixture structures. Thus in the general case the reference selection problem can be understood in probabilistic terms as the statistical problem of finding a finite mixture model (Everitt and Hand, 1981; Titterington et al., 1985) for the data (in our case in a discrete domain). This probabilistic viewpoint has been the starting point of our work.

There are several advantages of using the above probabilistic interpretation of instance-based learning. If a good approximative representation of the problem domain distribution can be found, a minimum risk Bayes decision rule can be used in the prediction phase (Gelman et al., 1995). It should also be observed that although we present here experimental results only for classification tasks, the predictive distributions can also be used directly for regression tasks, since in the learning phase all attributes (including the class attribute) are treated equally. Thus with a probabilistic approach, both the classification and regression tasks can be treated uniformly. In addition, such a prediction computation can be performed efficiently (Myllymäki and Tirri, 1995; Myllymäki and Tirri, 1994; Myllymäki and Tirri, 1993).

For learning the component distributions from the instance set, we have adopted the Bayesian approach (see e.g., (Gelman et al., 1995)) which allows us to make a tradeoff between the complexity of our distribution structure and fit to the data thus resolving the overfitting problem of the traditional instance-based approaches. This combination of finite mixtures with Bayesian model selection is akin to the approach adopted in the Autoclass system (Cheeseman et al., 1988) with the notable difference in our focus to prediction rather than latent class analysis.

The use of component distributions as prototypes implies that a given instance matches to several prototypes simultaneously with different probabilities, and thus the predictive distributions are computed as a weighted estimate from the marginal probability distributions for the attribute in question given by the prototypes. In the case of classification tasks this easily leads to confusion between the class structure and the mixture structure. Sometimes these structures coincide, but in the general case the number of components does not need to match the number of values of the class attribute, a difference that distinguishes our approach from the Naive Bayes classifier (see e.g., (Kononenko, 1993)) in classification tasks.

In this paper we describe a methodology for probabilistic instance-based learning for discrete domains. The methodology addresses all the three drawbacks discussed above: it provides for a computationally efficient prediction algorithm, avoids overfitting by using Bayesian model selection and uses directly probabilities as measures of similarity. The methodology described has been implemented in the D-SIDE software package. We present empirical results of the method's classification prediction performance for a set of public domain data sets (including data sets from the StatLog project (Michie et al., 1994)), and compare the results to the performance of various other machine learning and neural network methods for the same data sets. Our results clearly demonstrate that the probabilistic approach is highly competitive for a wide spectrum of natural data sets.

2 The finite mixture model for the instance space

In this work we confine ourselves to discrete data and discretize continuous attributes by quantization. The problem domain is modeled by m discrete attributes $A_1, \ldots, A_m$, which are regarded as discrete random variables. An *instance* $\vec{d}$ is a vector of attribute-value combinations, $\vec{d} = (A_1 = a_1, \ldots, A_m = a_m)$, where $a_i \in \{a_{i1}, \ldots, a_{in_i}\}$. As we will use a finite mixture to model the instance set, we assume that the instance space is partitioned into a set of K data clusters $c_1, \ldots, c_K$, and the attributes are assumed to be independent within each cluster. The probability distribution on the instantiation space is approximated as a weighted sum of mixture distributions:

$$P(\vec{d}) = \sum_{k=1}^{K} \left(P(C = c_k) \prod_{i=1}^{m} P(A_i = a_i | C = c_k) \right),$$

where the value of the discrete *clustering random variable* C denotes the cluster of the given instance. Consequently, a finite mixture model can be defined by first fixing K, the *model class* (the number of the component distributions), and then by determining the values of the model parameters $\Theta = (\alpha_1, \ldots, \alpha_K, Q_1, \ldots, Q_K)$, where $\alpha_k = P(C = c_k)$ and

$$Q_k = (q_{k11}, \ldots, q_{k1n_1}, \ldots, q_{km1}, \ldots, q_{kmn_m}), \text{where}$$

$$q_{kil} = P(A_i = a_{il} | C = c_k), k = 1, \ldots, K, i = 1, \ldots, m, \\ l = 1, \ldots, n_i.$$

3 Constructing models from instances

Let $\mathcal{D} = \{\vec{d}_1, \ldots, \vec{d}_N\}$ denote a database of N instances used as training data. In our probabilistic interpretation, instance set $\mathcal{D}$ is viewed as a random sample from the instance space probability distribution $\mathcal{P}$. By model construction we mean here the problem of constructing a single finite mixture model $M(\Theta)$ which represents the probability distribution $\mathcal{P}$ as accurately as possible. The model construction task can be divided into two separate phases. In the first phase, we determine the optimal number of component clusters by evaluating the posterior probability for each model class $\mathcal{M}_k$ (i.e., all the k cluster models), given the data:

$$P(\mathcal{M}_k | \mathcal{D}) \propto P(\mathcal{D} | \mathcal{M}_k) P(\mathcal{M}_k), k = 1, \ldots, N,$$

where the normalizing constant $P(\mathcal{D})$ can be omitted since we only need to compare different model classes. The number of clusters can safely be assumed to be bounded by the size of the instance set N, otherwise the sample size is too small for model construction.

Assuming equal priors for the model classes, they can be ranked by evaluating the *evidence* $P(\mathcal{D}|\mathcal{M}_k)$ for each model class,

$$P(\mathcal{D} | \mathcal{M}_k) = \int P(\mathcal{D} | \Theta, \mathcal{M}_k) P(\Theta | \mathcal{M}_k) \, d\Theta,$$

where the integration goes over the whole parameter space. As discussed in (Rissanen, 1989), the evidence can also be understood as an information theoretic measure called *stochastic complexity*. This evidence integral is hard to evaluate due to the very large dimensionality of the parameter space, but the evidence can be approximated by using e.g., Laplace's method (Kass and Raftery, 1994). In the experimental results presented in Section 5 this automatic model class selection has not yet been used, instead in the search process the model classes were selected by manual search in the model class space. For more discussion on estimating the evidence in the finite mixture context, see (Kontkanen et al., 1996a).

In the second phase of the model construction process, we wish to find the optimal set of parameters for the selected model class $\mathcal{M}_k$ by maximizing the posterior probability $P(\Theta|\mathcal{D})$. We assume that both the prior distribution for the cluster random variable $P(C)$ and the intra-class conditional distributions $P(A_i|C = c_k)$ are multinomial, and hence use the *Dirichlet distribution* as the prior for the parameters Θ[2]. The priors are *noninformative*, i.e., before seeing any instances no single model is assumed to be more probable than others. Furthermore, assuming parameter independence, we get

$$P(\Theta | \mathcal{D}) = \text{Dirichlet}\left(\frac{1}{K} + h_1, \ldots, \frac{1}{K} + h_K\right) \\ \cdot \prod_{k=1}^{K} \prod_{i=1}^{m} \text{Dirichlet}\left(\frac{1}{n_i} + f_{ki1}, \ldots, \frac{1}{n_i} + f_{kin_i}\right),$$

where h_k is the size of the cluster c_k and f_{kil} is the number of instantiations in cluster c_k with attribute A_i having value a_{il}.

From the properties of Dirichlet density and the independence assumptions it follows that the maximal probability values for the parameters Θ can be ob-

[2]For the justification of Dirichlet distributions as priors see e.g., (Heckerman et al., 1995).

tained by setting

$$P(C = c_k) = \frac{h_k + \frac{1}{K} - 1}{N + 1 - K}$$
$$P(A_i = a_{il}|C = c_k) = \frac{f_{kil} + \frac{1}{n_i} - 1}{h_k + 1 - n_i}.$$

Naturally, the parameters h_k and f_{kil} are not known, but they can be regarded as missing data and we estimate them by using the EM algorithm (Dempster et al., 1977). The EM algorithm is an iterative algorithm, which monotonically increases the expected value of the posterior corresponding to incomplete data. The derivation of the update formulas in our mixture case can be found in (Kontkanen et al., 1996b).

4 Bayesian inference with the mixture model

Let us assume that the mixture model $\mathcal{M}(\Theta)$ for the instance space has been constructed by the method described above. Furthermore let $\mathcal{I} = \{i_1, \ldots, i_t\}$ be the set of instantiated attribute indices, $\mathcal{A} = \{A_{i_1}, \ldots, A_{i_t}\}$ the set of instantiated attributes and $A_{i_1} = a_{i_1 l_1}, \ldots, A_{i_t} = a_{i_t l_t}$ the query (attribute value assignment) presented to the inference algorithm. Since we assume that attributes are conditionally independent given the cluster variable C, for all $i \notin \mathcal{I}$, A_i's predictive distribution $P(A_i = a_{il}|\mathcal{A})$ can be computed as follows:

$$\begin{aligned} &P(A_i = a_{il}|\mathcal{A}) \\ &= \sum_{k=1}^{K} \left(P(C = c_k|\mathcal{A})\, P(A_i = a_{il}|C = c_k, \mathcal{A})\right) \\ &= \sum_{k=1}^{K} \frac{\alpha_k q_{kil} P(\mathcal{A}|C = c_k)}{P(\mathcal{A})} = \sum_{k=1}^{K} \frac{\alpha_k q_{kil} \prod_{s=1}^{t} q_{ki_s l_s}}{\sum_{r=1}^{K} \alpha_r \prod_{s=1}^{t} q_{ri_s l_s}}. \end{aligned}$$

The predictive distribution can be computed efficiently as the summation is over the number of clusters K. Only in the very extreme case where the number of component distributions is N the computing cost approaches that of traditional instance-based learning. However, such a degenerate case would indicate that the instance space has no nontrivial cluster structure, a situation which seems to be very rare with natural data sets. Typically the number of clusters K is one or several orders of magnitude smaller than the size of the instance set N (see Table 2).

5 Empirical results

The above probabilistic instance-based approach has been implemented in the D-SIDE software package consisting of a model construction module and a prediction module. We have used this software to evaluate the feasibility of our approach, and compared it to alternative methods. Instead of using artificial data we were especially interested in the prediction performance for natural data sets. We readily admit the problems of using real data, especially for comparison purposes (underlying causes for performance differences are hard to identify, high variance in the observed performance differences etc.). However, the main advantage of using natural data sets is that it "keeps one honest", i.e., it is produced without any knowledge of the particular procedures that it will be used to test. In addition with artificial data there is always the danger that they do not correspond to situations that are likely to occur in practice.

We have done extensive experimentation with our probabilistic instance-based method using publicly available data sets for classification problems. We have also collected performance results for alternative methods for these same data sets from the literature. The list of the results of alternative algorithms tries by no means to be exhaustive, however for each data set we have included the best results we have found in the literature. The data sets were partly selected on the basis of their reported use, i.e., we have preferred data sets that have been used for testing many different methods over data with only isolated results. Many of the results are from the StatLog project (Michie et al., 1994), but we have also included more recent results. The descriptions of the data sets, our testing procedures, and the best model classes (the number of clusters) found for each data set are given in Table 1. The default value denotes the success rate of a simple classifier, which classifies all the instances to the most common class.

It should be observed that with the exception of the DNA data set, all our results are crossvalidated, and that for the StatLog data sets we have used the same crossvalidation scheme as described in (Michie et al., 1994). The same does not hold for many of the results for the other methods, as in many cases the testing procedure either was not reported, or the best result with a single test set was given. The actual performance results (measured as classification success percentage) for individual data sets are presented as barcharts in Figures 1 and 2.

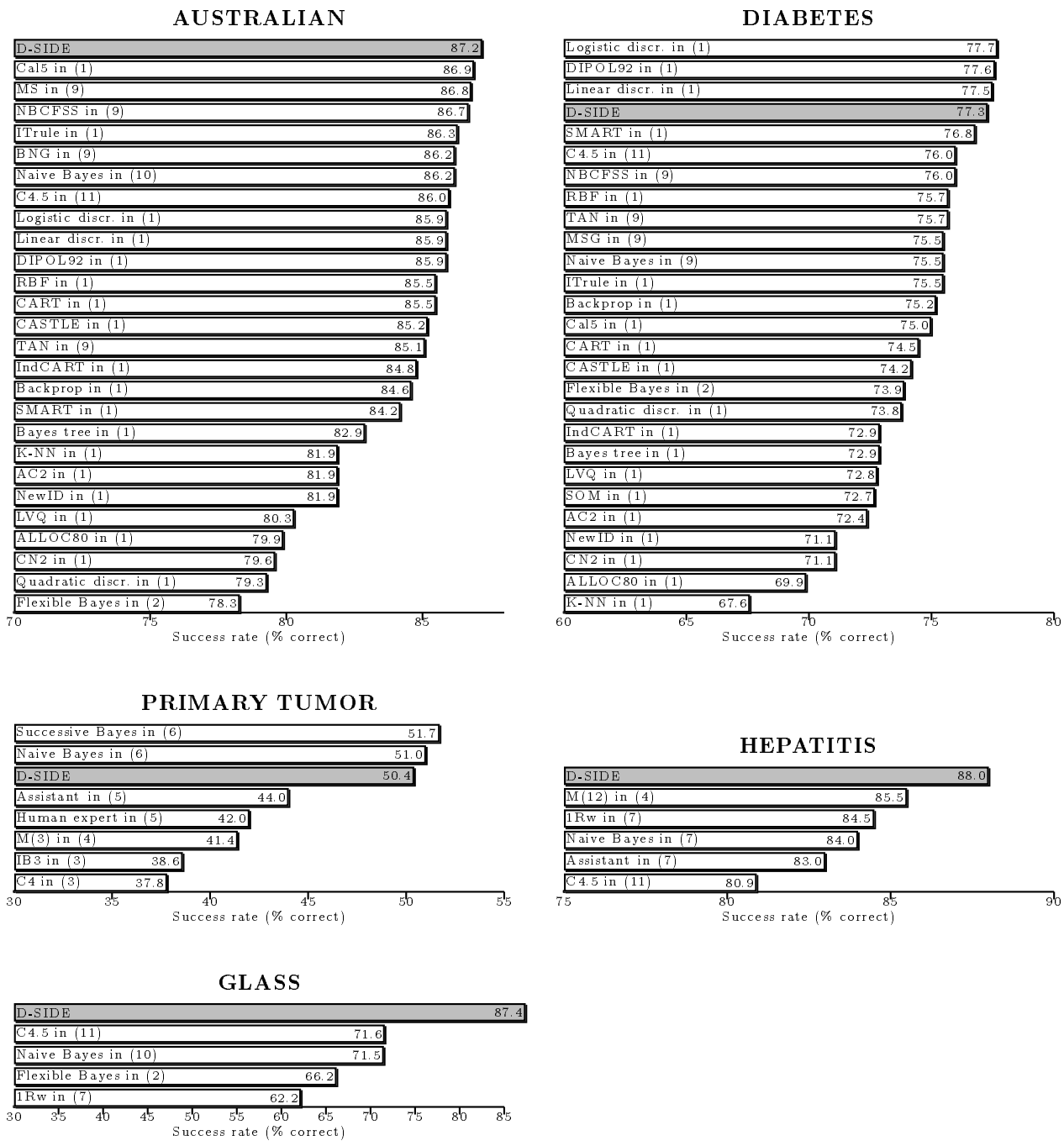

Figure 1: Experimental results on the Australian, Diabetes, Primary tumor, Hepatitis and Glass datasets. The references in the barcharts are as follows: (1) = (Michie et al., 1994), (2) = (John and Langley, 1995), (3) = (Aha et al., 1991), (4) = (Cestnik and Bratko, 1991), (5) = (Kononenko and Bratko, 1991), (6) = (Kononenko, 1993), (7) = (Holte, 1993), (8) = (Skalak, 1994), (9) = (Friedman and Goldszmidt, 1996a), (10) = (Friedman and Goldszmidt, 1996b) and (11) = (Quinlan, 1996).

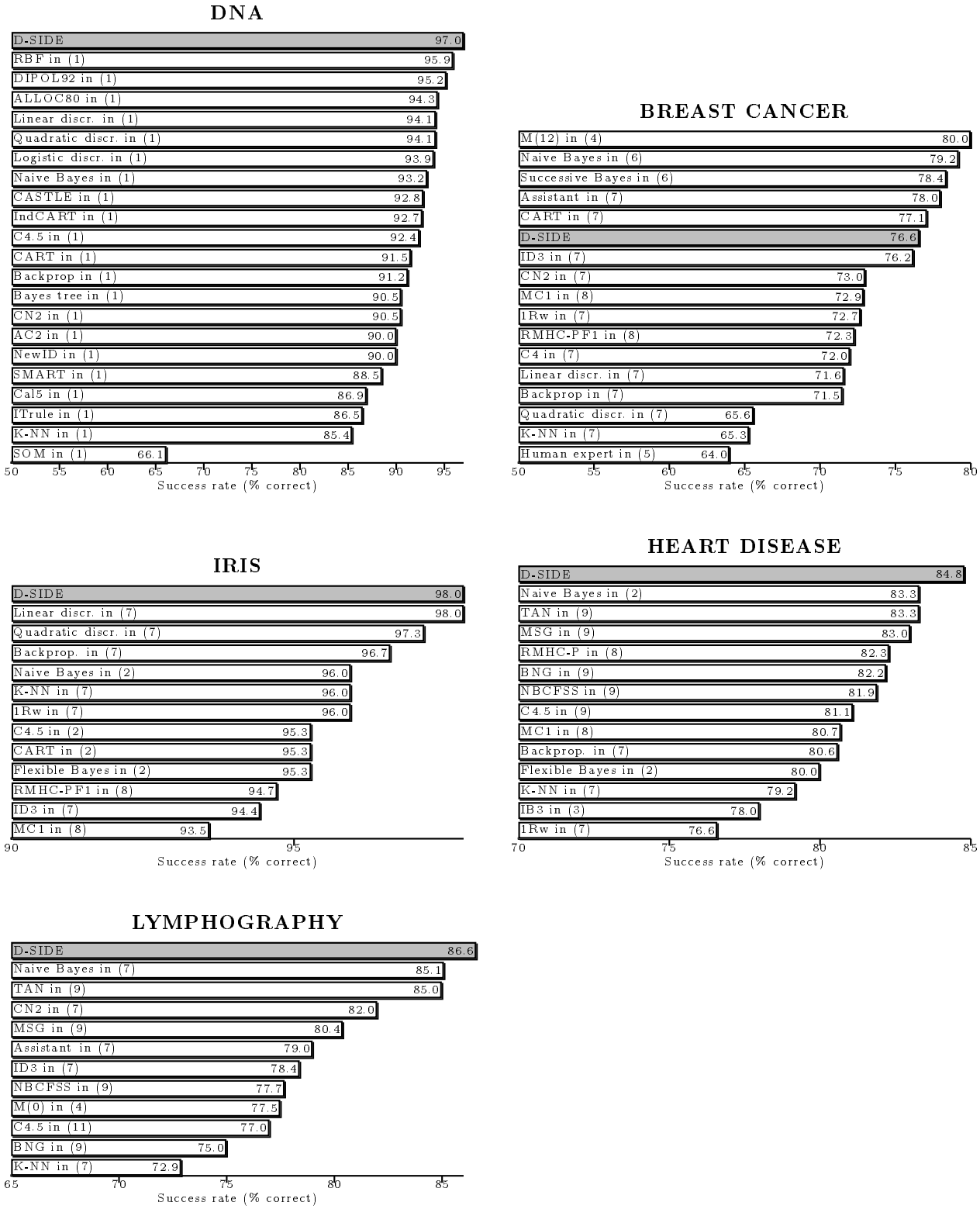

Figure 2: Experimental results on the DNA, Breast cancer, Iris, Heart disease and Lymphography databases. The references in the barcharts are as follows: (1) = (Michie et al., 1994), (2) = (John and Langley, 1995), (3) = (Aha et al., 1991), (4) = (Cestnik and Bratko, 1991), (5) = (Kononenko and Bratko, 1991), (6) = (Kononenko, 1993), (7) = (Holte, 1993), (8) = (Skalak, 1994), (9) = (Friedman and Goldszmidt, 1996a), (10) = (Friedman and Goldszmidt, 1996b) and (11) = (Quinlan, 1996).

Table 1: The datasets and testing methods used in our experiments.

name	size	#attrs	#classes	#clusters	test method	default
Australian	690	15	2	17	10-fold CV	56.0
Breast cancer	286	10	2	21	11-fold CV	70.3
Diabetes	768	9	2	20	12-fold CV	65.0
DNA	3186	181	3	13	train&test	50.8
Glass	214	10	6	30	7-fold CV	40.7
Heart disease	270	14	2	8	9-fold CV	79.4
Hepatitis	150	20	2	9	5-fold CV	55.6
Iris	150	5	3	4	5-fold CV	33.3
Lymphography	148	19	4	19	5-fold CV	54.7
Primary tumor	339	18	21	21	10-fold CV	24.8

Although this experimentation is still ongoing, the empirical results clearly show that the probabilistic instance-based approach performs favorably not only when compared to traditional instance-based methods (K-NN, IB3, ALLOC80), but also with respect to decision tree methods and common neural network approaches such as backpropagation. An interesting observation is that the probabilistic instance-based method outperforms also all other Bayesian approaches present in the StatLog comparison as well as more recent Naive Bayes related algorithms TAN (Tree Augmented Naive Bayes) and NBC^{FSS} introduced in (Friedman and Goldszmidt, 1996a). This supports the common hypothesis that many real data distributions can be naturally modeled as a sum of several component distributions. Table 2 summarizes the performance of those methods for which we could find results in 4 or more data sets. By *performance index* we mean here the relative success percentage of a given method, when compared to the best method in the current classification task (e.g., the performance index of 90.0 of means that the method has achieved a success rate which is 90% of the success rate of the best method in the task in question). It should be noted that this comparison favors methods which are tested on fewer (easier) datasets. Nevertheless, the results confirm the observation that the probabilistic instance-based approach offers the most consistent performance over the various data sets.

6 Conclusion

We have presented a methodology for probabilistic instance-based learning which addresses all the three common drawbacks of traditional instance-based learning: computational efficiency of prediction, overfitting and sensitivity to distance functions. We also presented empirical results of the method's classification prediction performance for a set of public domain data sets, and compared the results to the performance of various other machine learning and neural network methods. Our results clearly demonstrated that our probabilistic approach is highly competitive and offers a very consistent performance over different types of data.

Although the discussion on this paper is confined to discrete data, the approach extends also to the case where attributes are real-valued. However, some limited experimentation indicates that moving from discrete to continuous values does not necessarily improve the prediction performance of the model due to the additional assumptions of the distribution form. This extension is a natural topic for future research.

Acknowledgements

This research has been supported by the Technology Development Center (TEKES). The primary tumor, the breast cancer and the lymphography domains were obtained from the University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia. Thanks go to M. Zwitter and M. Soklič for providing the data.

References

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

Aha, D., Kibler, D., and Albert, M. (1991). Instance-based learning algorithms. *Machine Learning*, 6:37–66.

Table 2: Performance indexes of some of the most commonly used methods for the ten datasets used in our experiments.

method	mean	variance	min	max	#datasets
D-SIDE	99.27	1.94	95.75	100.00	10
TAN	97.85	0.12	97.43	98.23	4
Linear discriminant	96.95	15.01	89.50	100.00	5
MSG	96.86	6.12	92.84	99.54	4
CART	96.38	1.60	94.33	98.05	5
Naive Bayes	96.15	24.10	81.81	99.00	10
NBCFSS	95.89	13.68	89.72	99.43	4
Backpropagation	95.15	8.87	89.38	98.67	6
BNG	94.86	23.29	86.61	98.85	4
C4.5	93.42	28.00	81.92	98.62	8
Quadratic discriminant	92.84	36.93	82.00	99.29	5
CN2	92.18	1.83	91.02	94.69	6
Assistant	92.04	20.94	85.11	97.50	4
Flexible Bayes	90.45	59.94	75.74	97.24	5
K-NN	89.45	29.28	81.62	97.96	7
1Rw	88.96	75.90	71.17	97.96	6

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

Atkeson, C., Moore, A., and Schaal, S. (1995). Locally weighted learning. *AI Review to appear.*

Cestnik, B. and Bratko, I. (1991). On estimating probabilities in tree pruning. In Kodratoff, Y., editor, *Machine Learning EWSL-91*, pages 138–150. Springer-Verlag.

Cheeseman, P., Kelly, J., Self, M., Stutz, J., Taylor, W., and Freeman, D. (1988). Autoclass: A Bayesian classification system. In *Proceedings of the Fifth International Conference on Machine Learning*, pages 54–64, Ann Arbor.

Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13:21–27.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38.

Deng, K. and Moore, A. (1995). Multiresolution instance-based learning. In *International Joint Conference on Artificial Intelligence*, pages 1233–1239.

Everitt, B. and Hand, D. (1981). *Finite Mixture Distributions.* Chapman and Hall, London.

Franke, R. (1982). Scattered data interpolation: Test of some methods. *Mathematics of Computation*, 38(157).

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

Friedman, N. and Goldszmidt, M. (1996a). Building classifiers using Bayesian networks. In *Proceedings of AAAI-96 (to appear).*

Friedman, N. and Goldszmidt, M. (1996b). Discretizing continuous attributes while learning Bayesian networks. In Saitta, L., editor, *Machine Learning: Proceedings of the Thirteenth International Conference (to appear).* Morgan Kaufmann Publishers.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (1995). *Bayesian Data Analysis.* Chapman & Hall.

Heckerman, D., Geiger, D., and Chickering, D. (1995). Learning Bayesian networks: The combination of

knowledge and statistical data. *Machine Learning*, 20(3):197–243.

Holte, R. (1993). Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, 11:63–91.

John, G. and Langley, P. (1995). Estimating continuous distributions in Bayesian classifiers. In Besnard, P. and Hanks, S., editors, *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, pages 338–345. Morgan Kaufmann Publishers.

Kass, R. and Raftery, A. (1994). Bayes factors. Technical Report 254, Department of Statistics, University of Washington.

Kononenko, I. (1993). Successive naive Bayesian classifier. *Informatica*, 17:167–174.

Kononenko, I. and Bratko, I. (1991). Information-based evaluation criterion for classifier's performance. *Machine Learning*, 6:67–80.

Kontkanen, P., Myllymäki, P., and Tirri, H. (1996a). Comparing Bayesian model class selection criteria by discrete finite mixtures. In *Proceedings of the ISIS (Information, Statistics and Induction in Science) Conference*, Melbourne, Australia. (To appear.).

Kontkanen, P., Myllymäki, P., and Tirri, H. (1996b). Constructing Bayesian finite mixture models by the EM algorithm. Technical Report C-1996-9, University of Helsinki, Department of Computer Science.

Kurtzberg, J. (1987). Feature analysis for symbol recognition by elastic matching. *IBM Journal of Research and Development*, 31:91–95.

Michie, D., Spiegelhalter, D., and Taylor, C., editors (1994). *Machine Learning, Neural and Statistical Classification*. Ellis Horwood, London.

Moody, J. and Darken, C. (1989). Fast learning in networks of locally-tuned processing units. *Neural Computation*, 1:281–294.

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

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

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

Myllymäki, P. and Tirri, H. (1995). Constructing computationally efficient Bayesian models via unsupervised clustering. In A.Gammerman, editor, *Probabilistic Reasoning and Bayesian Belief Networks*, pages 237–248. Alfred Waller Publishers, Suffolk.

Quinlan, J. (1996). Improved use of continuous attributes in C4.5. *Journal of Artificial Intelligence Research*, 4:77–90.

Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquiry*. World Scientific Publishing Company, New Jersey.

Scott, D. (1992). *Multivariate Density Estimation. Theory, Practice, and Visualization*. John Wiley & Sons, New York.

Skalak, D. (1994). Prototype and feature selection by sampling and random mutation hill climbing algorithms. In *Machine Learning: Proceedings of the Eleventh International Conference*, pages 293–301.

Specht, D. (1990). Probabilistic neural networks. *Neural Networks*, 3:109–118.

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

Titterington, D., Smith, A., and Makov, U. (1985). *Statistical Analysis of Finite Mixture Distributions*. John Wiley & Sons, New York.

Zhang, J. (1992). Selecting typical instances in instance-based learning. In *Proceedings of the Ninth International Machine Learning Workshop*, pages 470–479, San Mateo, CA. Morgan Kaufmann.