Bayes Optimal Instance-Based Learning

Petri Kontkanen, Petri Myllymäki, Tomi Silander, and Henry Tirri

Complex Systems Computation Group (CoSCo)
P.O.Box 26, Department of Computer Science
FIN-00014 University of Helsinki, Finland

Abstract. In this paper we present a probabilistic formalization of the instance-based learning approach. In our Bayesian framework, moving from the construction of an explicit hypothesis to a data-driven instance-based learning approach, is equivalent to averaging over all the (possibly infinitely many) individual models. The general Bayesian instance-based learning framework described in this paper can be applied with any set of assumptions defining a parametric model family, and to any discrete prediction task where the number of simultaneously predicted attributes is small, which includes for example all classification tasks prevalent in the machine learning literature. To illustrate the use of the suggested general framework in practice, we show how the approach can be implemented in the special case with the strong independence assumptions underlying the so called Naive Bayes classifier. The resulting Bayesian instance-based classifier is validated empirically with public domain data sets and the results are compared to the performance of the traditional Naive Bayes classifier. The results suggest that the Bayesian instance-based learning approach yields better results than the traditional Naive Bayes classifier, especially in cases where the amount of the training data is small.

1 Introduction

Machine learning research aims at constructing automated methods for deducing useful information from sample data. In principle the work can be divided into two main subareas of research. In the first research area, the goal is to find useful high-level knowledge representations from the data through exploratory data analysis (this *descriptive* aspect is very related to the research performed in the field of *data mining* [10]). In the second research area, the goal is to predict the outcome of some future event by using the data given. In this paper, we are motivated purely by this latter, *predictive* aspect of machine learning.

The standard approach to machine learning can be viewed as a three phase modeling process. Initially, the models to be considered are restricted to some limited set of models, the *model family.* Examples of common model families include the set of feedforward neural network models, the set of Bayesian networks, or the set of decision trees. In the second phase, some specific *model class*, i.e., a skeleton or a template structure for a model without fixing any parameter values, is selected from the chosen model family. In the third phase,

the parameter values for the selected model class are estimated from the sample data. The resulting full model (model structure + parameter values) can then be used for making predictions. Bayesian probability theory provides a unifying theoretically solid framework for choosing a proper model family, model class, and parameter instantiation during all the three phases of the machine learning process, as discussed for example in [14, 20].

In contrast to the traditional (*eager*) approach described above, in the *instance-based* (also known as the *memory-based* or the *case-based*) approach [29, 24, 1, 4], 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 referred to as *lazy learning*, since the algorithms defer all the essential computation until the prediction phase [2].

For making predictions, instance-based learning algorithms typically use a distance function (e.g., Euclidean distance) for determining the most relevant data items for the prediction task in question. Some simple function, such as majority voting in classification problems, is then used for determining the prediction from the most relevant data items. It has been shown in various studies (see e.g., [23] for references) that this type of an approach in some cases produces quite accurate predictions, when compared to alternative machine learning methods. The method suffers, however, from several drawbacks when applied in practice (see, e.g., the discussion in [32]). Most importantly, the performance of instance-based learning algorithms seems to be highly sensitive to the selection of distance function to be used as demonstrated in recent work reported in [34, 13, 5].

In [32, 31, 26, 25] we proposed a Bayesian framework for instance-based learning based on probability theory and the finite mixture model family [9, 33]. The approach suggested in those studies can be seen as a "partially lazy" approach [2], i.e., a hybrid between the traditional machine learning and the instance-based learning approach, which is based solely on the given data. The studies were based on the probabilistic viewpoint, where the given data vectors are transformed into local distributions, which can be seen as sample points in a distribution space. Thus the predictive distributions required for making predictions could then be computed by using the instance-based learning approach in the distribution space, i.e., by introducing a probabilistic "distance metric". Somewhat similar frameworks have been suggested in [16, 30, 11, 12].

The goal of this paper is to present a novel alternative probabilistic formalization of the *purely lazy learning approach*. The framework suggested here extends 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 this 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 averaging over all the possible models.

In the Bayesian framework, prediction can be viewed as a missing data problem, where the criterion for filling in the missing data (for making the predic-

tions) is the integral over all the possible models. Intuitively, the joint probability distribution produced by a (possibly infinite) mixture of individual models reflects the true (unknown) problem domain distribution better than any single model, thus it should also produce more accurate predictions. In fact, it is known that this type of integrated predictions are optimal from the Bayesian point of view, given a fixed model family. To avoid terminological confusion it should be observed that any type of learning is with respect to some model family. In traditional instance-based learning approaches the model family is implicitly induced by the combination of the distance function and the domain of the data, also in the cases where the distance function is allowed to vary locally.

At first glance the Bayes optimal instance-based learning framework may seem to be a computationally infeasible approach — after all, the integration (summation) may go over an infinite number of models. Nevertheless, it turns out that for some simple model families the integral over all the model instantiations, i.e., the so called *evidence* or the *marginal likelihood* [6], can in fact be solved analytically, and calculated with modest computational effort. An example of such a model family is the family of Bayesian networks (see e.g. [7, 15]), where the model family is determined by defining a set of independence relations between the problem domain variables. For more complex model families including those with latent variables, there exist several computationally feasible methods for approximating the evidence integral — see e.g., the discussion in [19].

It should be emphasized that we make no claims about having invented the idea of making predictions by marginalizing over all the (possibly infinitely many) models, which is a known technique in the Bayesian community. *The main goal of this paper is to point out how the Bayesian formalism can be used for developing a theoretically solid framework for instance-based learning.* A formalization of this Bayesian instance-based learning approach is given in Section 2. As an illustrative example of our general approach, we demonstrate in Section 3 how the Bayesian instance-based learning approach can be applied with a set of strong independence assumptions underlying a simple Bayesian network model structure, the naive Bayes classifier. Section 4 shows the results of an empirical comparison between the Bayesian instance-based learning approach presented here, and the traditional approach, based on a single maximum a posteriori model, in this special case. The tests were performed by using publicly available real-world classification datasets.

2 Bayesian Instance-Based Learning

Let D denote a random sample of N i.i.d. (independent and identically distributed) data vectors $\boldsymbol{d}_1, \ldots, \boldsymbol{d}_N$. For simplicity, we assume here that the data is coded by using only discrete, i.e., finite-valued, attributes $X_1, \ldots, X_m$, although the Bayesian approach described can be extended to continuous attributes as well. More precisely, we regard each attribute X_i as a random variable with possible values from the set $\{x_{i1}, \ldots, x_{in_i}\}$. Consequently, each data vector $\boldsymbol{d}$ is

represented as a value assignment of the form $(X_1 = x_1, \ldots, X_m = x_m)$, where $x_i \in \{x_{i1}, \ldots, x_{in_i}\}$.

In the following, let $\mathcal{M}$ denote a *model family*, a set of models each determining some probability distribution on the problem domain. Examples of model families include the set of feedforward neural network models, the set of Bayesian networks, and the set of decision trees. For notational convenience, it is often useful to partition the models within a model family $\mathcal{M}$ to some number of subsets, *model classes*, where all the models within a model class share the same parametric form (the same number of parameters). Consequently, the model classes usually correspond to some specific model structure. Examples of such structures are the topology of a feedforward neural network, a Bayesian network, or a decision tree. A *model* Θ is here defined as a parameter instantiation within some model class M, fully determining a probability distribution in the data vector space. Consequently, a single model is defined by fixing the parameters attached to a given model structure, e.g., by fixing the weights of a neural network architecture or the decision rules of a decision tree.

In traditional (eager) machine learning, the model family, model class, and the model parameters must all be fixed in order to produce a single model for making predictions. Bayesian probability theory provides a theoretically solid framework for these tasks, as demonstrated, e.g., in [21] in the neural network model family case, and in [32] in the finite mixture model family case. In the Bayesian approach, the model parameters to be used within a model class M are taken to be the *maximum a posteriori (MAP)* values $\hat{\Theta}$ of the parameters,

$$\hat{\Theta} = \arg\max_{\Theta} P(\Theta \mid D, M, \mathcal{M}).$$

Similarly, the model class to be used is the class with the maximal posterior probability,

$$\hat{M} = \arg\max_{M} P(M \mid D, \mathcal{M}).$$

Nevertheless, as

$$P(M \mid D, \mathcal{M}) = \frac{P(D \mid M, \mathcal{M}) P(M \mid \mathcal{M})}{P(D \mid \mathcal{M})},$$

it is sufficient to find the model class M maximizing $P(D \mid M, \mathcal{M})$, the *evidence* or *marginal likelihood* of the data,

$$P(D \mid M, \mathcal{M}) = \int P(D \mid \Theta, M, \mathcal{M}) P(\Theta \mid M, \mathcal{M}) d\Theta, \tag{1}$$

assuming that all the model classes are equally probable a priori.

In principle, the model family can be chosen by maximizing a similar posterior probability $P(\mathcal{M} \mid D)$ over all model families $\mathcal{M}$,

$$P(\mathcal{M} \mid D) = \sum_{k} P(\mathcal{M}, M_k \mid D).$$

Nevertheless, computing the model family posterior probability for all the possible model families is obviously intractable in practice. Instead, some individual model family, determined by a set of assumptions made about the problem domain, is normally fixed in advance. The assumptions are based on some prior knowledge, or just on practitioner's personal preferences.

Having fixed the model family $\mathcal{M}$, model class $\hat{M}$, and the model parameters $\hat{\Theta}$, the Bayes optimal predictive distribution for a new *test vector* $\boldsymbol{d}$ is the conditional distribution

$$P(\boldsymbol{d} \mid D, \hat{\Theta}, \hat{M}, \mathcal{M}) = \frac{P(\boldsymbol{d}, D \mid \hat{\Theta}, \hat{M}, \mathcal{M})}{P(D \mid \hat{\Theta}, \hat{M}, \mathcal{M})}. \tag{2}$$

As the probability $P(D \mid \hat{\Theta}, \hat{M}, \mathcal{M})$ can be regarded as a constant with respect to the test vector $\boldsymbol{d}$, for predictive purposes it is sufficient to be able to compute the joint probability distribution $P(\boldsymbol{d}, D \mid \hat{\Theta}, \hat{M}, \mathcal{M})$. In the sequel, we call distribution (2) the *MAP predictive distribution.*

In instance-based lazy learning approach, on the other hand, we wish to base our predictions directly on the data D, without having to determine individual models Θ. In the Bayesian framework, we can express this by marginalizing out the individual models — in other words, instead of computing the MAP predictive distribution (2), we wish to compute

$$P(\boldsymbol{d} \mid D, \hat{M}, \mathcal{M}) = \int P(\boldsymbol{d} \mid D, \Theta, \hat{M}, \mathcal{M}) P(\Theta \mid D, \hat{M}, \mathcal{M}) d\Theta. \tag{3}$$

Furthermore, if the model class M is not to be fixed, we need to sum over different model classes within the chosen model family, yielding

$$P(\boldsymbol{d} \mid D, \mathcal{M}) = \sum_k P(\boldsymbol{d} \mid D, M_k, \mathcal{M}) P(M_k \mid D, \mathcal{M}). \tag{4}$$

In the sequel, by *Bayesian instance-based learning (BIBL)* we mean the approach based on formulas (4) and (3).

Note that formula (4) offers a formal motivation for the idea of *model averaging* (see, e.g., [22, 3] and the references therein), i.e., for combining multiple predictors for increasing the prediction accuracy: the individual predictions $P(\boldsymbol{d} \mid D, M_k, \mathcal{M})$ produced by different predictors M_k (for example, model classes determined by different decision tree structures), are combined by summing the individual predictions weighted by the model class probabilities $P(M_k \mid D, \mathcal{M})$. As a matter of fact, from the probability theory point of view the Bayesian instance-based learning predictive distribution (4) produces optimally accurate predictions within the chosen model family. In [17, 18], we described how the recently published new coding scheme by Rissanen [28] for representing the stochastic complexity measure [27] offers an alternative definition for an optimal predictive distribution. This definition can be justified by information theoretic arguments, but this approach will not be addressed in this paper.

3 Bayesian IBL with the Naive Bayes Assumptions

For notational simplicity, in the sequel we drop the variable names, and denote a value assignment $(X_1 = x_1, \ldots, X_{m-1} = x_{m-1})$ by writing $(x_1, \ldots, x_{m-1})$. In addition, instead of explicitly stating for each specific model the corresponding model class M and model family $\mathcal{M}$, we use $P(\cdot \mid \Theta)$ for denoting $P(\cdot \mid \Theta, M, \mathcal{M})$.

The MAP predictive distribution distribution (2), and the BIBL predictive distribution (4) can be used for solving various predictive inference problems. As an example, let us consider the standard classification problem, where the goal is to predict the value of the *class variable*, denoted here by X_m, given the values of other variables $X_1, \ldots, X_{m-1}$. In the MAP case (using a given MAP model $\hat{\Theta}$), we need to find a classification x_m (value assignment for variable X_m) maximizing the conditional probability

$$P(x_m \mid x_1, \ldots, x_{m-1}, D, \hat{\Theta}) \propto P(x_m, x_1, \ldots, x_{m-1} \mid D, \hat{\Theta}) = P(\boldsymbol{d} \mid D, \hat{\Theta}), \quad (5)$$

denoting $\boldsymbol{d} = (x_m, x_1, \ldots, x_{m-1})$. Consequently, the resulting conditional probability is a predictive distribution of the form (2). Equivalently, in the BIBL case we wish to compute

$$P(x_m \mid x_1, \ldots, x_{m-1}, D, \mathcal{M}) \propto P(x_m, x_1, \ldots, x_{m-1} \mid D, \mathcal{M}), \quad (6)$$

which corresponds to formula (4).

As an illustrative example of the Bayesian instance-based learning approach, in the following let us consider the model family determined by the set of independence assumptions underlying behind the well-known Naive Bayes classifier. In this case, the variables $X_1, \ldots, X_{m-1}$ are assumed to be independent given the value of the class variable X_m. It follows that the joint probability distribution for a data vector $\boldsymbol{d}$ can be written as

$$P(\boldsymbol{d} \mid \Theta) = P(x_1, \ldots, x_m \mid \Theta) = P(x_m) \prod_{i=1}^{m-1} P(x_i \mid x_m).$$

Consequently, in the Naive Bayes model family, a single predictive distribution can be uniquely determined by fixing the values of the model parameters $\Theta = (\alpha, \Phi)$, where

$$\alpha = (\alpha_1, \ldots, \alpha_K) \text{ and } \Phi = (\Phi_{11}, \ldots, \Phi_{1m}, \ldots, \Phi_{K1}, \ldots, \Phi_{Km}),$$

K $(= n_m)$ is the number of possible values for the class variable X_m, and

$$\alpha_k = P(X_m = x_{mk}), \Phi_{ki} = (\phi_{ki1}, \ldots, \phi_{kin_i}),$$

where $\phi_{kil} = P(X_i = x_{il} \mid X_m = x_{mk})$.

In the following we assume that $\alpha_k > 0$ and $\phi_{kil} > 0$ for all k,i, and l. Furthermore, both the class variable distribution $P(X_m)$ and the intra-class conditional distributions $P(X_i \mid X_m = x_{mk})$ 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., [8]) to the family of multinomials, i.e., the functional form of parameter distribution is invariant in the prior-to-posterior transformation, we assume that the prior distributions of the parameters are from this family. More precisely, let $(\alpha_1, \ldots, \alpha_K) \sim \mathrm{Di}(\mu_1, \ldots, \mu_K)$, and $(\phi_{ki1}, \ldots, \phi_{kin_i}) \sim \mathrm{Di}(\sigma_{ki1}, \ldots, \sigma_{kin_i})$, where $\{\mu_k, \sigma_{kil} \mid k = 1, \ldots, K; i = 1, \ldots, m; l = 1, \ldots, n_i\}$ are the *hyperparameters* of the corresponding distributions. Assuming that the parameter vectors α and Φ_{ki} are independent, the joint prior distribution of all the parameters Θ is

$$\mathrm{Di}(\mu_1, \ldots, \mu_K) \prod_{k=1}^{K} \prod_{i=1}^{m-1} \mathrm{Di}(\sigma_{ki1}, \ldots, \sigma_{kin_i}).$$

Having now defined the prior distribution, the predictive distributions (2) and (4) can be written more explicitly. Let $\boldsymbol{d}[k] = (X_m = x_{mk}, \boldsymbol{q})$ denote a data vector where the values of variables $X_1, \ldots, X_{m-1}$ correspond to the given query $\boldsymbol{q}$, and the value of the class variable X_m is set to x_{mk}. The MAP predictive distribution (2) needed for computing the predictive distribution (5) for $\boldsymbol{d}[k]$ is in the Naive Bayes case

$$P(\boldsymbol{d}[k] \mid D, \hat{\Theta}) \overset{\text{i.i.d.}}{=} P(\boldsymbol{d}[k] \mid \hat{\Theta}) = \hat{\alpha}_k \prod_{i=1}^{m} \hat{\phi}_{kix_i}, \text{ where} \tag{7}$$

$$\hat{\alpha}_k = \frac{h_k + \mu_k - 1}{N + \sum_{k'=1}^{K} \mu_{k'} - K}, \quad \hat{\phi}_{kil} = \frac{f_{kil} + \sigma_{kil} - 1}{h_k + \sum_{l=1}^{n_i} \sigma_{kil} - n_i},$$

and h_k and f_{kil} are the *sufficient statistics* of the training data D: h_k is the number of data vectors in D where $X_m = x_{mk}$, and f_{kil} is the number of data vectors where $X_m = x_{mk}$ and $X_i = x_{il}$. If we assume a uniform prior distribution for the parameters, we get $\hat{\alpha}_k = h_k/N$, $\hat{\phi}_{kil} = f_{kil}/h_k$, which produces the standard maximum likelihood Naive Bayes classifier with the parameters set according to the relative frequencies of different variable values.

The above analysis gives us the Bayesian maximum posterior probability answer to the question of how to determine the parameters of the Naive Bayes classifier. We now turn our focus on the BIBL case. First it should be noted that the formal assumptions listed above can be expressed by, e.g., using a simple two-level tree-structured Bayesian network, where the class variable corresponds to the root of the tree, and the other variables form the leaves. This means that the Naive Bayes assumptions determine a single model class M, so there is no need to sum over the model classes in (4), and hence the predictive distribution (4) reduces to (3). Producing this predictive distribution corresponds to the case where instead of using a single set of parameters for the Naive Bayes classifier, as in the MAP predictive distribution (7), we sum over all the (infinitely many) parameter alternatives for the Naive Bayes classifier. As shown in [7, 15, 19], with the assumptions listed above the marginal likelihood (1) of the data can be computed by

$$P(D \mid \mathcal{M}) = \frac{\Gamma\left(\sum_{k=1}^{K} \mu_k\right)}{\Gamma\left(N + \sum_{k=1}^{K} \mu_k\right)} \prod_{k=1}^{K} \frac{\Gamma(h_k + \mu_k)}{\Gamma(\mu_k)} \cdot \prod_{k=1}^{K} \prod_{i=1}^{m-1} \left(\frac{\Gamma\left(\sum_{l=1}^{n_i} \sigma_{kil}\right)}{\Gamma\left(h_k + \sum_{l=1}^{n_i} \sigma_{kil}\right)} \prod_{l=1}^{n_i} \frac{\Gamma(f_{kil} + \sigma_{kil})}{\Gamma(\sigma_{kil})} \right), \quad (8)$$

where $\Gamma(\cdot)$ denotes the gamma function, a generalization of the common factorial function. By using this result it is relative easy to see that the predictive distribution (4) can in this case be written as

$$P(\boldsymbol{d}[k] \mid D, \mathcal{M}) = \int P(\boldsymbol{d}[k] \mid \Theta, D, \mathcal{M}) P(\Theta \mid D, \mathcal{M}) d\Theta = \bar{\alpha}_k \prod_{i=1}^{m} \bar{\phi}_{kix_i}, \quad (9)$$

where the parameters α_k and ϕ_{kix_i} are set to their expected (*not* maximum probability) values:

$$\bar{\alpha}_k = \frac{h_k + \mu_k}{N + \sum_{k'=1}^{K} \mu_{k'}}, \quad \bar{\phi}_{kil} = \frac{f_{kil} + \sigma_{kil}}{h_k + \sum_{l=1}^{n_i} \sigma_{kil}}.$$

Consequently, in this special case the BIBL approach leads to the somewhat surprising result that if one wishes to sum over all the possible Naive Bayes classifiers (summing over all the infinitely many parameter settings), the resulting predictive distribution is the same as the one obtained by using a single Naive Bayes classifier where the parameters are set to their expected values! Hence the time and space complexity requirements of the BIBL approach are in this case exactly the same as with the standard Naive Bayes classifier. Nevertheless, it is important to realize that this phenomenon is not true in general, and only caused by the specific independence assumptions made above — with other sets of assumptions the BIBL approach would not necessarily lead to a predictive distribution that can be obtained by using a single model.

4 Empirical Results

To validate the Bayesian instance-based learning approach described in the previous sections, we performed a series of experiments with a set of public domain classification datasets from the UCI repository[1]. Each classification query was classified by using both the single MAP Naive Bayes model with uniform priors (MLNB) given by formula (7), and the Bayesian IBL approach with the Naive Bayes assumptions (BIBL) defined in (9), again with uniform priors. Description of the datasets used, and the results obtained can be found in Table 1. The results are averages over 100 independent crossvalidation runs, and the number of

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

folds used was the same as in [23]. By the 0/1-score we mean the relative number of the correct classifications made, while the log-score is obtained by computing minus the logarithm of the probability given to the correct class (thus the smaller the score, the better the result).

Table 1. The datasets used in the experiments, and the averages of the corresponding crossvalidated classification accuracies obtained.

					0/1-SCORE				LOG-SCORE			
				CV	100% data		10% data		100% data		10% data	
Dataset	N	m	K	folds	MLNB	BIBL	MLNB	BIBL	MLNB	BIBL	MLNB	BIBL
Australian	690	15	2	10	**85.0**	84.9	76.2	**83.0**	0.7	**0.5**	9.5	**0.5**
Breast cancer	286	10	2	11	71.7	**72.3**	62.3	**69.4**	2.5	**0.6**	16.1	**0.8**
Diabetes	768	9	2	12	**75.7**	**75.7**	70.4	**72.4**	0.6	**0.5**	8.1	**0.6**
Glass	214	10	6	7	**66.9**	66.4	38.5	**50.5**	7.5	**1.0**	17.2	**1.6**
Heart disease	270	14	2	9	83.4	**84.1**	70.1	**80.0**	1.2	**0.4**	14.1	**0.5**
Hepatitis	150	20	2	5	**84.1**	81.5	63.1	**78.6**	4.2	**0.7**	3.4	**0.7**
Iris	150	5	3	5	93.4	**94.4**	77.5	**94.1**	1.9	**0.1**	2.9	**0.2**
Lymphography	148	19	4	5	78.9	**84.3**	39.3	**72.2**	5.7	**0.4**	6.8	**0.7**
Primary tumor	339	18	21	10	45.7	**48.8**	20.9	**32.2**	18.5	**2.0**	38.8	**3.0**

To see how the methods rely on the size of the data set available for training, we repeated the 100 independent crossvalidation runs, but used at each stage of the crossvalidation cycle only a (randomly selected) subset of the data in the training folds for classifying the data in the test fold. In Table 1 we list the results obtained when only $F = 10\%$ of the training data was used at each stage. Figure 1 illustrates the typical behavior of the averages of the crossvalidated scores as a function of F.

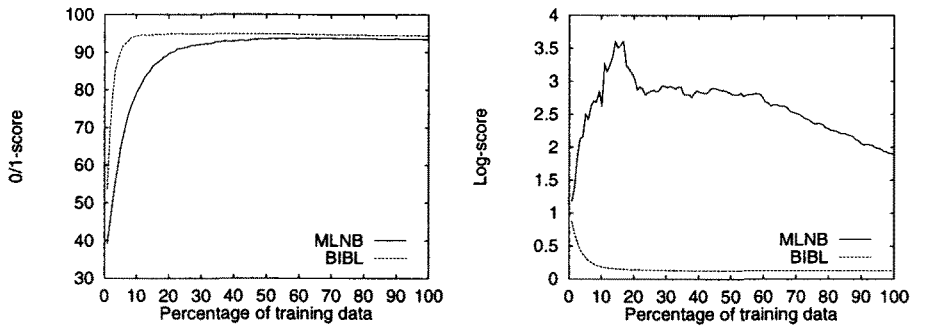

Fig. 1. Average crossvalidated 0/1-scores (left) and log-scores (right) in the Iris dataset case as a function of the percentage of the training data used.

The results show that first of all, although the model family used for the experiments was determined by the strong independence assumptions underlying the structurally simple naive Bayes model, the results are quite competitive when compared to the results obtained by using much more elaborate model families (see, e.g., the results collected in [23]). Secondly, we can see that when full datasets are used, in the 0/1-score sense the difference between the performance of the BIBL classifier and the MLNB classifier is not very large, whereas in the log-score sense the BIBL approach produces consistently better results. The experiments with restricted training data sets show that the BIBL approach is very effective in extracting regularities present in the data, and it clearly outperforms the standard single model approach in cases with very small amounts of data, both in the 0/1-score and in the log-score sense. This is due to the fact that BIBL is much more "conservative" than the single model MLNB. For small samples it is well known that the traditional MLNB classifier is too dependent on the observed data and does not take into account that future data *may* turn out to be different. A more detailed discussion on this topic can be found in [17, 18].

5 Conclusion

In this paper we proposed a Bayesian framework for defining the instance-based learning approach. The framework is based on the observation that moving from a model based learning approach, such as decision tree learning, to an instance-based learning approach that relies solely on data, is in probabilistic terms equivalent to averaging over all the (possibly infinitely many) models. We presented the formalization of the general framework, and illustrated how the approach can be implemented in the special case with a set of strong independence assumptions.

Our experiments with public domain classification data sets indicate that the Bayesian instance-based approach outperforms the (eager) use of a single model from the respective model class, especially in cases where only a small amount of training data is available. It turns out that the Bayesian instance-based learning prediction is very effective in extracting the regularities present in the data sets, and requires sometimes order of magnitude less data than what is actually available in the data sets to predict essentially as well as with the full data set.

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

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, editor. *Lazy Learning.* Kluwer Academic Publishers, Dordrecht, 1997. Reprinted from Artificial Intelligence Review, 11:1–5.
3. K. Ali and M. Pazzani. Error reduction through learning multiple descriptions. *Machine Learning*, 24(3):173–202, September 1997.
4. 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.
5. C. Atkeson, A. Moore, and S. Schaal. Locally weighted learning. In Aha [2], pages 11–73.
6. J.O. Berger. *Statistical Decision Theory and Bayesian Analysis.* Springer-Verlag, New York, 1985.
7. G. Cooper and E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347, 1992.
8. M.H. DeGroot. *Optimal statistical decisions.* McGraw-Hill, 1970.
9. B.S. Everitt and D.J. Hand. *Finite Mixture Distributions.* Chapman and Hall, London, 1981.
10. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors. *Advances in Knowledge Discovery and Data Mining.* MIT Press, Cambridge, MA, 1996.
11. D. Fisher. Noise-tolerant conceptual clustering. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 825–830, Detroit, Michigan, 1989.
12. D. Fisher 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.
13. 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.
14. A. Gelman, J. Carlin, H. Stern, and D. Rubin. *Bayesian Data Analysis.* Chapman & Hall, 1995.
15. 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.
16. S. Kasif, S. Salzberg, D. Waltz, J. Rachlin, and D. Aha. Towards a better understanding of memory-based reasoning systems. In *Proceedings of the Eleventh International Machine Learning Conference*, pages 242–250, New Brunswick, NJ, 1994. Morgan Kaufmann Publishers.
17. 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. Also: NeuroCOLT Technical Report NC-TR-97-004.
18. P. Kontkanen, P. Myllymäki, T. Silander, H. Tirri, and P. Grünwald. On predictive distributions and Bayesian networks. In W. Daelemans, P. Flach, and A. van den Bosch, editors, *Proceedings of the Seventh Belgian-Dutch Conference on Machine Learning (BeNeLearn'97)*, pages 59–68, Tilburg, the Netherlands, October 1997.

19. P. Kontkanen, P. Myllymäki, and H. Tirri. Comparing Bayesian model class selection criteria by discrete finite mixtures. In D. Dowe, K. Korb, and J. Oliver, editors, *Information, Statistics and Induction in Science*, pages 364–374, Proceedings of the ISIS'96 Conference, Melbourne, Australia, August 1996. World Scientific, Singapore.
20. P. Kontkanen, P. Myllymäki, and H. Tirri. Experimenting with the Cheeseman-Stutz evidence approximation for predictive modeling and data mining. In D. Dankel, editor, *Proceedings of the Tenth International FLAIRS Conference*, pages 204–211, Daytona Beach, Florida, May 1997.
21. D. Mackay. *Bayesian Methods for Adaptive Models.* PhD thesis, California Institute of Technology, 1992.
22. D. Madigan, A. Raftery, C. Volinsky, and J. Hoeting. Bayesian model averaging. In *AAAI Workshop on Integrating Multiple Learned Models*, 1996.
23. D. Michie, D.J. Spiegelhalter, and C.C. Taylor, editors. *Machine Learning, Neural and Statistical Classification.* Ellis Horwood, London, 1994.
24. A. Moore. Acquisition of dynamic control knowledge for a robotic manipulator. In *Seventh International Machine Learning Workshop.* Morgan Kaufmann, 1990.
25. 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.
26. 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.
27. J. Rissanen. *Stochastic Complexity in Statistical Inquiry.* World Scientific Publishing Company, New Jersey, 1989.
28. J. Rissanen. Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, 42(1):40–47, January 1996.
29. C. Stanfill and D. Waltz. Toward memory-based reasoning. *Communications of the ACM*, 29(12):1213–1228, 1986.
30. K. Ting and R. Cameron-Jones. Exploring a framework for instance based learning and Naive Bayes classifiers. In *Proceedings of the Seventh Australian Joint Conference on Artificial Intelligence*, pages 100–107, 1994.
31. 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. Springer-Verlag, Berlin Heidelberg, November 1996.
32. 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.
33. D.M. Titterington, A.F.M. Smith, and U.E. Makov. *Statistical Analysis of Finite Mixture Distributions.* John Wiley & Sons, New York, 1985.
34. D. Wettschereck, D. Aha, and T. Mohri. A review and empirical evaluation of feature-weighting methods for a class of lazy learning algorithms. In Aha [2], pages 273–314.