Mach Learn (2012) 86:233–272
DOI 10.1007/s10994-011-5263-6

Learning by extrapolation from marginal to full-multivariate probability distributions: decreasingly naive Bayesian classification

Geoffrey I. Webb · Janice R. Boughton · Fei Zheng · Kai Ming Ting · Houssam Salem

Received: 8 December 2009 / Accepted: 15 September 2011 / Published online: 13 October 2011

Abstract Averaged n-Dependence Estimators (AnDE) is an approach to probabilistic classification learning that learns by extrapolation from marginal to full-multivariate probability distributions. It utilizes a single parameter that transforms the approach between a low-variance high-bias learner (Naive Bayes) and a high-variance low-bias learner with Bayes optimal asymptotic error. It extends the underlying strategy of Averaged One-Dependence Estimators (AODE), which relaxes the Naive Bayes independence assumption while retaining many of Naive Bayes' desirable computational and theoretical properties. AnDE further relaxes the independence assumption by generalizing AODE to higher-levels of dependence. Extensive experimental evaluation shows that the bias-variance trade-off for Averaged 2-Dependence Estimators results in strong predictive accuracy over a wide range of data sets. It has training time linear with respect to the number of examples, learns in a single pass through the training data, supports incremental learning, handles directly missing values, and is robust in the face of noise. Beyond the practical utility of its lower-dimensional variants, AnDE is of interest in that it demonstrates that it is possible to create low-bias high-variance generative learners and suggests strategies for developing even more powerful classifiers.

Keywords Bayesian learning · Classification learning · Probabilistic learning · Averaged one-dependence estimators · Naive Bayes · Semi-naive Bayesian learning · Learning without model selection · Ensemble learning · Feating

Editor: Peter Flach.

G.I. Webb (✉) · J.R. Boughton · F. Zheng · K.M. Ting · H. Salem
Faculty of Information Technology, Monash University, Clayton, VIC, 3800, Australia
e-mail: Geoff.Webb@monash.edu

J.R. Boughton
e-mail: Janice.Boughton@monash.edu

K.M. Ting
e-mail: Kaiming.Ting@monash.edu

1 Introduction

This paper presents a family of learning algorithms that utilize a predefined function to extrapolate from observed marginal distributions to the full multivariate distribution of interest. This stands in contrast to the majority of learning algorithms that instead seek to fit a model directly to the observed multivariate probability distribution. Whereas learning is sometimes cast as a problem of searching through a space of hypotheses to find one that best fits the training data (Mitchell 1982), this new approach does not employ search and does not perform model selection.

The members of this new family of algorithms have a unique combination of features that is well suited to many applications. We discuss these features in more detail below. Notable amongst them are training complexity linear with respect to the number of training examples; single pass learning; direct capacity for incremental learning; and accuracy that is competitive with the state-of-the-art. They are of further theoretical interest because they demonstrate that it is possible to create low bias generative learners.

The family contains algorithms that range from low variance coupled with high bias through to high variance coupled with low bias. Successive members of the family will be best suited to differing quantities of data, starting with low variance for small data, with successively lower bias but higher variance suiting ever increasing data quantities (Brain and Webb 2002). The asymptotic error of the lowest bias variant is Bayes optimal.

One member of this family of algorithms, naive Bayes (NB), is already well known. A second member, Averaged One-Dependence Estimators (AODE) (Webb et al. 2005), has enjoyed considerable popularity since its introduction in 2005 (Nikora 2005; Camporelli 2006; Flikka et al. 2006; Orhan and Altan 2006; Lasko et al. 2006; Hunt 2006; Ferrari and Aitken 2006; Birzele and Kramer 2006; Kunchevaa et al. 2007; Lau et al. 2007; Masegosa et al. 2007; Wang et al. 2007; Garcia et al. 2008; Tian et al. 2008; Kurz et al. 2009; Leon et al. 2009; Shahri and Jamil 2009; Simpson et al. 2009; Affendey et al. 2010; García-Jiménez et al. 2010; Hopfgartner et al. 2010; Liew et al. 2010). The work presented in this paper arises from the realization that NB and AODE are but two instances of a family of algorithms, which we call AnDE.

In Sect. 2 we explain the underlying learning strategy, and define the AnDE family of algorithms. The AnDE family of algorithms build upon the method pioneered by AODE (Webb et al. 2005). In Sect. 3 we discuss how the AnDE algorithms relate to Feating (Ting et al. 2011), a generic approach to ensembling that also builds upon techniques pioneered by AODE. In Sect. 4 we present an extensive evaluation of the AnDE family of algorithms, comparing their performance to relevant Bayesian techniques, to Feating and to the state-of-the-art Random Forests classifier. Section 5 presents conclusions and directions for future research.

2 The AnDE family of algorithms

We wish to estimate from a training sample $\mathcal{T}$ of t classified objects the probability $\mathrm{P}(y \mid \mathbf{x})$ that an example $\mathbf{x} = \langle x_1, \ldots, x_a \rangle$ belongs to class y, where x_i is the value of the ith attribute and $y \in \{c_1, \ldots, c_k\}$. We use $\bar{v}$ to denote the average number of values per attribute. These and other elements of notation are listed in Table 1.

From the definition of conditional probability we have

$$\mathrm{P}(y \mid \mathbf{x}) = \mathrm{P}(y, \mathbf{x})/\mathrm{P}(\mathbf{x}) \tag{1}$$

Table 1 Notation

$\mathrm{P}(e)$	The unconditioned probability of event e
$\mathrm{P}(e \mid w)$	The conditional probability of event e given event w
$\hat{\mathrm{P}}(e)$	An estimate of $\mathrm{P}(e)$
$\hat{\mathrm{P}}_{\mathrm{NB}}(e)$	A naive Bayes estimate of $\mathrm{P}(e)$
$\hat{\mathrm{P}}_{\mathrm{AODE}}(e)$	An AODE estimate of $\mathrm{P}(e)$
$\hat{\mathrm{P}}_{\mathrm{A}n\mathrm{DE}}(e)$	An AnDE estimate of $\mathrm{P}(e)$
a	The number of attributes
c_i	The ith class
k	The number of classes
t	The number of training examples in $\mathcal{T}$
$\bar{v}$	The average number of values per attribute
y	A value from the set of all classes $\{c_1, \ldots, c_k\}$
$\mathcal{T}$	A training sample of t classified objects
$\mathbf{x} = \langle x_1, \ldots, x_a \rangle$	An object
x_i	The value of the ith attribute of $\mathbf{x} = \langle x_1, \ldots, x_a \rangle$
$x_{\{i,j,\ldots,q\}}$	The subset of attributes values from $\mathbf{x}$ with the specified indices. For example, $x_{\{2,3,5\}} = \langle x_2, x_3, x_5 \rangle$
$\binom{\mathcal{A}}{n}$	The set of all size-n subsets of $\{1, \ldots, a\}$
$\delta(x_\alpha)$	A function that is 1 if $\mathcal{T}$ contains an object with the value x_α, otherwise 0

As $\mathrm{P}(\mathbf{x}) = \sum_{i=1}^{k} \mathrm{P}(c_i, \mathbf{x})$, we can always estimate (1) from estimates of $\mathrm{P}(y, \mathbf{x})$ for each class using

$$\mathrm{P}(y, \mathbf{x})/\mathrm{P}(\mathbf{x}) = \mathrm{P}(y, \mathbf{x}) \Big/ \sum_{i=1}^{k} \mathrm{P}(c_i, \mathbf{x}). \tag{2}$$

In consequence, in the remainder of this paper we consider only the problem of estimating $\mathrm{P}(y, \mathbf{x})$, thereby setting the work in a generative framework.

We define the *dimensionality* of a probability or probability estimate as the number of attributes in the distribution to which the probability or estimate relates. Hence, the dimensionality of $\mathrm{P}(y, \mathbf{x})$ is $a + 1$.

If the training data do not include sufficient examples of $\mathbf{x}$ to directly derive accurate estimates of each $\mathrm{P}(c_i, \mathbf{x})$, we must extrapolate these estimates from observations of lower-dimensional probabilities in the data. All other things being equal, an estimate of a lower-dimensional probability from a given finite training set will be more accurate than an estimate of a higher-dimensional probability, and estimates of higher-dimensional probabilities will vary more from training sample to training sample. Hence, models derived from lower-dimensional probability estimates are likely to have lower variance than models derived from higher-dimensional probability estimates. On the other hand, models derived from higher-dimensional probabilities are likely to have lower bias, as less restrictive assumptions are made about the form of the probability distribution.

This is illustrated in Fig. 1, that shows a simple attribute-space with three ternary attributes and a binary class. To classify a new object with attribute-values Age = Old, Pulse = Slow and Temperature = High, one wishes to infer the class distribution in the cell highlighted in Fig. 1(a), which is a four-dimensional probability distribution. If there are insufficient examples to directly estimate that distribution, it might be extrapolated

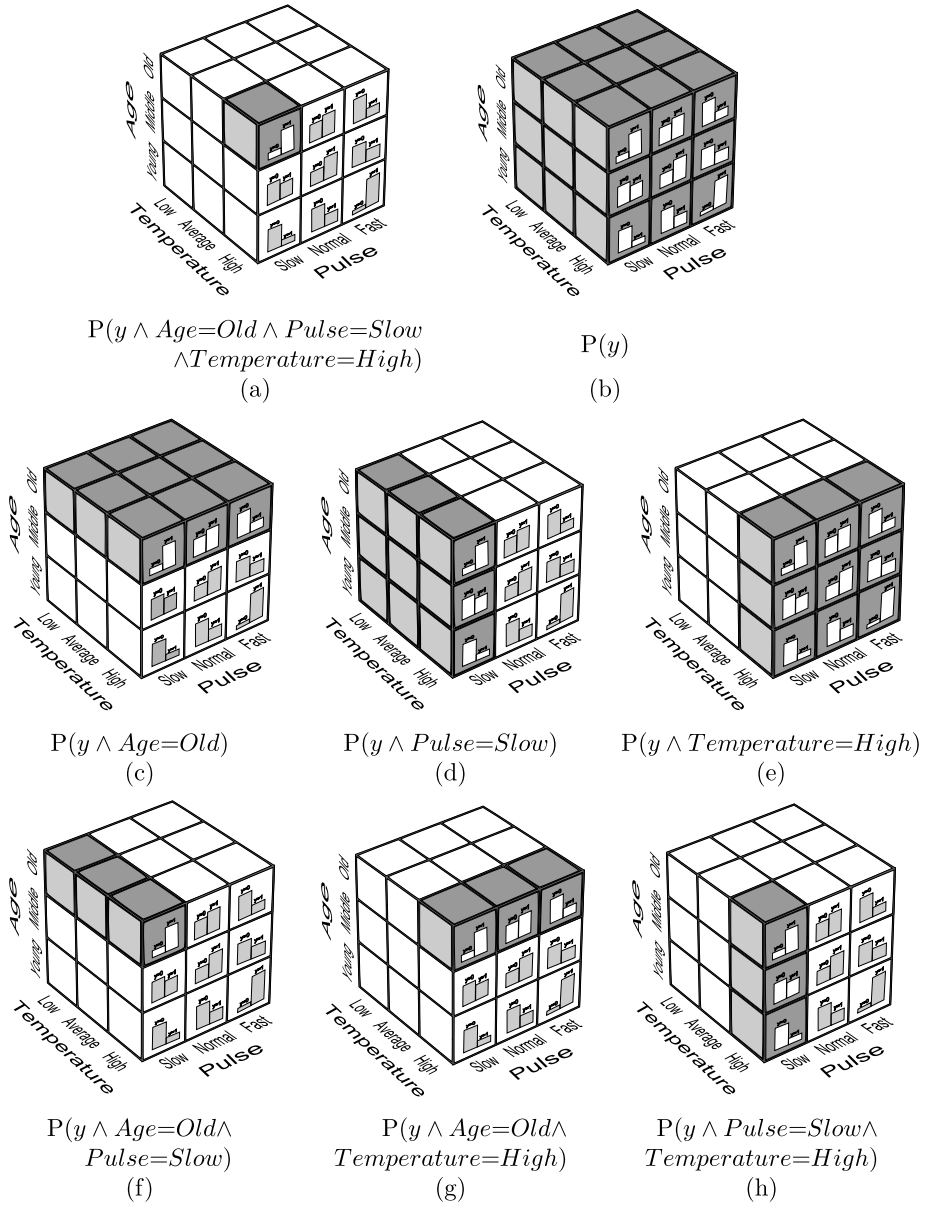

Fig. 1 Probabilities of varying dimensionality for an attribute-space with three ternary attributes and a binary class

from any of a number of lower dimensional probability distributions. The prior class distribution P(y) is a one-dimensional probability distribution that can be estimated from the entire attribute-space (Fig. 1(b)). The two-dimensional probabilities P($y \wedge$ *Age* = *Old*), P($y \wedge$ *Pulse* = *Slow*), P($y \wedge$ *Temperature* = *High*) can be estimated from the regions depicted in Fig. 1(c–e). The regions associated with the three-dimensional probabili-

ties $\mathrm{P}(y \wedge Age = Old \wedge Pulse = Slow)$, $\mathrm{P}(y \wedge Age = Old \wedge Temperature = High)$ and $\mathrm{P}(y \wedge Pulse = Slow \wedge Temperature = High)$ are illustrated in Fig. 1(f–h).

From the definition of conditional probability we have

$$\mathrm{P}(y, \mathbf{x}) = \mathrm{P}(y)\mathrm{P}(\mathbf{x} \mid y). \tag{3}$$

If the number of classes, k, is small, it should be possible to obtain a sufficiently accurate estimate of $\mathrm{P}(y)$ from the sample frequencies. However, we still have the problem that $\mathbf{x}$ may not occur sufficiently frequently in the training data and hence accurate estimates of $\mathrm{P}(\mathbf{x} \mid y)$ cannot be obtained directly from the sample frequencies.

The solution used by NB is to extrapolate to $\hat{\mathrm{P}}(\mathbf{x} \mid y)$ from each two-dimensional probability estimate $\hat{\mathrm{P}}(x_i \mid y)$ by assuming the attributes are independent given the class:

$$\mathrm{P}(\mathbf{x} \mid y) = \prod_{i=1}^{a} \mathrm{P}(x_i \mid y). \tag{4}$$

Hence we classify using

$$\hat{\mathrm{P}}_{\mathrm{NB}}(y, \mathbf{x}) = \hat{\mathrm{P}}(y) \prod_{i=1}^{a} \hat{\mathrm{P}}(x_i \mid y). \tag{5}$$

With reference to Fig. 1, NB estimates the distribution in (a) by extrapolation from the distributions in (b) (that gives $\hat{\mathrm{P}}(y)$), (c) (that gives $\hat{\mathrm{P}}(Age = old \mid y)$), (d) (that gives $\hat{\mathrm{P}}(Pulse = slow \mid y)$) and (e) (that gives $\hat{\mathrm{P}}(Temperature = high \mid y)$).

The independence assumption is a very strong assumption about the underlying probability distribution. As a result, NB has very high bias. However, due to the low dimensionality of the base probabilities from which the model is estimated, it has low variance.

2.1 AODE

Averaged One-Dependence Estimators (AODE) (Webb et al. 2005) extends to three-dimensional probabilities NB's search-free strategy of extrapolation from lower-dimensional probabilities. It does so by averaging the estimates of all of a class of three-dimensional estimators.

A Super-Parent One-Dependence Estimator (SPODE) is a three-dimensional probability estimator that relaxes the assumption of conditional independence by making all other attributes independent given the class and one privileged attribute, the super-parent, x_α. This is a weaker conditional independence assumption than NB's, as it is necessarily true if NB's is true and may also be true when NB's is not.

It uses

$$\mathrm{P}(y, \mathbf{x}) = \mathrm{P}(y, x_\alpha)\mathrm{P}(\mathbf{x} \mid y, x_\alpha) \tag{6}$$

together with a conditional independence assumption

$$\mathrm{P}(\mathbf{x} \mid y, x_\alpha) = \prod_{i=1}^{a} \mathrm{P}(x_i \mid y, x_\alpha). \tag{7}$$

As this is a weaker assumption than (4), the bias of the model should be lower than that of NB. However, it is derived from higher-dimensional probability estimates and hence its variance should be higher.

AODE exploits the lower bias of SPODEs while addressing their higher variance by averaging over all estimates of $\mathrm{P}(y, \mathbf{x})$ produced by using different super-parents. AODE seeks to use

$$\hat{\mathrm{P}}(y, \mathbf{x}) = \sum_{\alpha=1}^{a} \hat{\mathrm{P}}(y, x_\alpha)\hat{\mathrm{P}}(\mathbf{x} \mid y, x_\alpha)/a. \tag{8}$$

However, in practice it is desirable to only use estimates of probabilities for which relevant examples occur in the data. Hence, AODE actually uses

$$\hat{\mathrm{P}}_{\mathrm{AODE}}(y, \mathbf{x}) = \begin{cases} \sum_{\alpha=1}^{a} \delta(x_\alpha)\hat{\mathrm{P}}(y, x_\alpha)\hat{\mathrm{P}}(\mathbf{x} \mid y, x_\alpha) \Big/ \sum_{\alpha=1}^{a} \delta(x_\alpha) & : \quad \sum_{\alpha=1}^{a} \delta(x_\alpha) > 0 \\ \hat{\mathrm{P}}_{\mathrm{NB}}(y, \mathbf{x}) & : \quad \text{otherwise} \end{cases} \tag{9}$$

where $\delta(x_\alpha)$ is 1 if attribute-value x_α is present in the data, otherwise 0. That is, it averages over all superparents whose value occurs in the data, and defaults to NB if there are no such superparents.

As AODE uses all of a predefined family of estimators, each of which extrapolates the desired high-dimensional probability from lower-dimensional probabilities, it does not perform search.

In terms of the example attribute space, AODE extrapolates to Fig. 1(a) from the lower-dimensional probabilities illustrated in Fig. 1(c–h) with (f) conditioned on (c) and (d), (g) conditioned on (c) and (e), and (g) conditioned on (d) and (e).

AODE has demonstrated strong prediction accuracy (both zero-one loss and mean-squared error) with relatively modest computational requirements for low dimensional data (Webb et al. 2005). In consequence, it has enjoyed substantial uptake (Nikora 2005; Camporelli 2006; Flikka et al. 2006; Orhan and Altan 2006; Lasko et al. 2006; Hunt 2006; Ferrari and Aitken 2006; Birzele and Kramer 2006; Kunchevaa et al. 2007; Lau et al. 2007; Masegosa et al. 2007; Wang et al. 2007; Garcia et al. 2008; Tian et al. 2008; Kurz et al. 2009; Leon et al. 2009; Shahri and Jamil 2009; Simpson et al. 2009; Affendey et al. 2010; García-Jiménez et al. 2010; Hopfgartner et al. 2010; Liew et al. 2010).

2.2 A*n*DE

In this paper we generalize to higher-dimensional probabilities the strategy of search-free extrapolation from lower-dimensional probabilities.

For notational convenience we define

$$x_{\{i,j,\ldots,q\}} = \langle x_i, x_j, \ldots, x_q \rangle. \tag{10}$$

For example, $x_{\{2,3,5\}} = \langle x_2, x_3, x_5 \rangle$.

A*n*DE aims to use

$$\hat{\mathrm{P}}(y, \mathbf{x}) = \sum_{s \in \binom{\mathcal{A}}{n}} \hat{\mathrm{P}}(y, x_s)\hat{\mathrm{P}}(\mathbf{x} \mid y, x_s) \Big/ \binom{a}{n}, \tag{11}$$

where $\binom{\mathcal{A}}{n}$ indicates the set of all size-n subsets of $\{1, \ldots, a\}$.

However, in practice we also need to avoid using pairs of superparents whose values do not occur in the data, and hence use

$$\hat{\mathrm{P}}_{\mathrm{A}n\mathrm{DE}}(y,\mathbf{x}) = \begin{cases} \sum_{s\in\binom{\mathcal{A}}{n}} \delta(x_s)\hat{\mathrm{P}}(y,x_s)\hat{\mathrm{P}}(\mathbf{x}\mid y,x_s) \Big/ \sum_{s\in\binom{\mathcal{A}}{n}} \delta(x_s) & : \quad \sum_{s\in\binom{\mathcal{A}}{n}} \delta(x_s) > 0 \\ \hat{\mathrm{P}}_{\mathrm{A}(n-1)\mathrm{DE}}(y,\mathbf{x}) & : \quad \text{otherwise.} \end{cases} \tag{12}$$

Attributes are assumed independent given the superparents and the class. Hence, $\mathrm{P}(\mathbf{x}\mid y,x_s)$ is estimated by

$$\hat{\mathrm{P}}(\mathbf{x}\mid y,x_s) = \prod_{i=1}^{a} \hat{\mathrm{P}}(x_i\mid y,x_s). \tag{13}$$

Note that $\mathrm{P}(x_i\mid y,x_s) = 1$ when $i\in s$. Whereas other probability estimates should be smoothed or regularized, smoothed estimates should not be used in this case, and in practice these values are not included in the calculation.

It should be recalled that A0DE is NB and A1DE is AODE.

In terms of the simple attribute-space depicted in Fig. 1, A2DE extrapolates to (a) from (a) conditioned on each of (f), (g) and (h), and A3DE makes inferences directly from the class distribution in (a).

When $n=a$, $\binom{\mathcal{A}}{n} = \{\{1,\ldots a\}\}$, so $x_s = \mathbf{x}$. Therefore, the ultimate expression of AnDE, AaDE seeks to classify using

$$\hat{\mathrm{P}}_{\mathrm{A}a\mathrm{DE}}(y,\mathbf{x}) = \hat{\mathrm{P}}(y,\mathbf{x})\hat{\mathrm{P}}(\mathbf{x}\mid y,\mathbf{x}) \Big/ \binom{a}{a} \tag{14}$$

where $\hat{\mathrm{P}}(y,\mathbf{x})$ is estimated directly from $\mathcal{T}$, cascading to ever lower dependence estimators should the combination of attribute-values not be present in $\mathcal{T}$. As $\mathrm{P}(\mathbf{x}\mid y,\mathbf{x})$ and $\binom{a}{a}$ both equal 1, it classifies using only its direct estimate of $\mathrm{P}(y,\mathbf{x})$.

Observation 1 The asymptotic classification performance of AaDE equals that of the Bayes optimal classifier.

Proof AaDE classifies using

$$\operatorname*{argmax}_y \left(\hat{\mathrm{P}}(y,\mathbf{x}) \Big/ \sum_{z\in\{c_1,\ldots,c_k\}} \hat{\mathrm{P}}(z,\mathbf{x}) \right)$$

where each $\hat{\mathrm{P}}(\cdot)$ is directly estimated from the observed data and hence approaches $\mathrm{P}(\cdot)$ as the quantity of data approaches infinity. Hence, in the limit, AaDE approaches

$$\operatorname*{argmax}_y \left(\mathrm{P}(y,\mathbf{x}) \Big/ \sum_{z\in\{c_1,\ldots,c_k\}} \mathrm{P}(z,\mathbf{x}) \right)$$

which is the Bayes optimal classifier. □

However, assuming there is sufficient data to compute the necessary probabilities, and we wish to store the necessary probabilities rather than computing them as required for

classification, the space complexity of AaDE is $O(k\bar{v}^a)$. This is because joint frequencies must be stored for every combination of attribute value and class value. Except in cases of low dimensional data, even the computational requirements of A3DE defeat our Weka implementation, and hence in this paper we present primarily results for A2DE with some illustrative examples of A3DE.

AnDE forms an $(n+2)$-dimensional probability table containing the observed frequency for each combination of $n+1$ attribute values and the class labels. The space complexity of the table is $O(k\binom{a}{n+1}\bar{v}^{n+1})$ and the time complexity of compiling it is $O(t\binom{a}{n+1})$, as we need to update each entry for every combination of the $n+1$ attribute-values for every instance. The time complexity of classifying a single example is $O(ka\binom{a}{n})$ as we need to consider each attribute for every qualified combination of n parent attributes within each class.

We demonstrate that as n increases, averaged n-dependence estimators achieve lower bias at the cost of higher variance. In consequence, the ideal dimensionality of dependence will depend on the degree to which the underlying probability distribution fits the assumptions of the n-dependence estimator, the quantity of data available to estimate the base probabilities, and the computational demands of averaging over higher-dimensional estimators.

2.3 Weighted averaging

AODE and its generalization AnDE perform an unweighted average of the component n-dependence estimators. It has been demonstrated that weighted averaging can improve upon the accuracy of AODE's estimates (Cerquides and Mántaras 2005; Jiang and Zhang 2006; Yang et al. 2007). The empirical evidence suggests that the Bayesian model averaging of Maximum a Posteriori Linear Mixture of Generative Distributions (MAPLMG) is the most effective of current approaches (Cerquides and Mántaras 2005; Yang et al. 2007). It seems likely that similar approaches will be equally effective with n-dependence estimators.

It is notable that the introduction of Bayesian model averaging to the AnDE framework introduces both search and discriminative learning, as a search is performed for the set of weights that optimize the posterior probabilities relative to the training data. Doing so can be expected to reduce bias at the cost of introduction of variance.

One of the interesting questions that this paper investigates is the relative payoff for the investment of additional computation in either performing Bayesian model averaging on AnDE or increasing n and using A$(n+1)$DE. Both approaches can be expected to reduce bias at the cost of an increase in both variance and computation. Which provides the more effective trade-off?

2.4 Tree Augmented Naive Bayes

An *n-dependence Bayesian classifier* (n-DBC) (Sahami 1996) is a Bayesian network in which each attribute depends upon the class and at most n other non-class attributes. An n-DBC uses $(n+2)$-dimensional probabilities. Within this framework, NB is a 0-DBC, AODE is a 1-DBC and the full Bayesian classifier is an $(a+1)$-DBC.

An alternative to the AnDE approach to relaxing NB's independence assumption is to use search to select a single model that adds selected interdependencies between attributes. Tree Augmented Naive Bayes (TAN) (Friedman et al. 1997) is a popular approach of this type. It uses conditional mutual information to select a best single parent for each attributes, in addition to the class. Thus, it is a 1-DBC.

It is interesting to consider how search for a single Bayesian classifier model compares with averaging over a class of Bayesian classifier models of the same level of dependence or of a higher level of dependence. This paper also investigates this issue.

3 Relationship to Feating

Feating (Ting et al. 2011) is a generic ensemble learning technique that also builds upon the ensembling strategy of AODE. Like AnDE, Feating operates by building a local model for each combination of n attribute values. To classify a new instance, Feating applies all applicable local models and aggregates the results by performing a majority vote of the resulting classifications. AnDE is similar to Feating NB. However, Feating aggregates the predictions of its base learners by taking the mode of the class predictions. For probabilistic classifiers, these class predictions correspond to the maximum posterior probability. In contrast, AnDE uses the ensemble to estimate the joint probability, $\mathrm{P}(\mathbf{x}, y)$ for each class, and then calculates its estimate of the posterior probability from this ensembled estimate of the joint probability. A generic ensembling technique, such as Feating, cannot work by calculating an ensemble estimate of the joint probabilities because many classifiers do not produce appropriate probability estimates.

Despite the close relationship to Feating, AnDE is worthy of study in its own right for three reasons.

First, irrespective of which aggregation method is used, coupling the search-less ensembling strategy embodied by Feating with search-less base learner NB creates a learner that can deliver low bias using a predefined mapping from low dimensional probabilities to the desired high dimensional probabilities without search or model selection. Hence, AnDE provides an example of an alternative to the traditional search-based learning paradigm which is able to deliver low bias classifiers.

Second, as already noted, AnDE utilizes a different aggregation method to Feating. It is interesting to examine the consequences of these differences. Cerquides and de Màntaras (2005) found that weighted ensembles of joint probability estimates achieved lower error than weighted ensembles of posterior probability estimates, so there is some evidence that the outcomes may be substantially different.

Third, as there is overlap in the information required by each of its local models, AnDE can use a single compiled matrix of joint frequencies, saving considerable space relative to storing all of the local models independently. The space complexity of an AnDE model is $O(k\binom{a}{n+1})\bar{v}^{n+1}$ whereas the space complexity of Feating NB to level n is $O(k(a-n)\binom{a}{n}\bar{v}^{n+1})$, which is $(n+1)$ times the space complexity of AnDE. Most base models formed by Feating will not have this property, and hence AnDE is a notable special case.

4 Evaluation

In this section, we evaluate the efficacy of AnDE. Due to relatively high time complexity of higher-dimensional estimators, the highest level of AnDE with which we perform detailed assessment is A2DE. The primary metrics we use are bias, variance, zero-one loss and RMSE. To assess computational overheads we use total training and classification times divided by the number of examples.

We first study the performance of NB, AODE and A2DE to reveal how performance varies as n increases within the AnDE framework. TAN and MAPLMG are studied to show how the search-free generative AnDE strategy compares with, respectively, discriminative search for a single Bayesian network classifier of the same dimensionality of dependence, and discriminative search for a weighted classifier of the next lower dimensionality of dependence. We also compare A2DE, that estimates the mean joint probability of the submodels, with variants that calculate the mean posterior probability (PA2DE) and perform Feating of NB by taking the mode of the class predictions of the submodels (FA2DE). Finally, to explore how the classification performance of A2DE compares to state-of-the-art classifiers, we also study Random Forests (Breiman 2001) with ten trees (RF10) and Random Forests with 100 trees (RF100).

We compare these algorithms implemented in the Weka workbench (Witten and Frank 2005) on the 62 data sets described in Table 2 that have been used previously in related

Table 2 Data sets used for experiments

No.	Domain	Case	Att	Class
1	Abalone	4177	9	3
2	Adult	48842	15	2
3	Annealing	898	39	6
4	Audiology	226	70	24
5	Auto Imports	205	26	7
6	Balance Scale	625	5	3
7	Breast Cancer (Wisconsin)	699	10	2
8	Car Evaluation	1728	8	4
9	Census-Income (KDD)	299285	40	2
10	Connect-4 Opening	67557	43	3
11	Contact-lenses	24	5	3
12	Contraceptive Method Choice	1473	10	3
13	Covertype	581012	55	7
14	Credit Screening	690	16	2
15	Cylinder Bands	540	40	2
16	Dermatology	366	35	6
17	Echocardiogram	131	7	2
18	German	1000	21	2
19	Glass Identification	214	10	3
20	Haberman's Survival	306	4	2
21	Heart Disease (Cleveland)	303	14	2
22	Hepatitis	155	20	2
23	Horse Colic	368	22	2
24	House Votes 84	435	17	2
25	Hungarian	294	14	2
26	Hypothyroid(Garavan)	3772	30	4
27	Ionosphere	351	35	2
28	Iris Classification	150	5	3
29	King-rook-vs-king-pawn	3196	37	2

Table 2 (*Continued*)

No.	Domain	Case	Att	Class
30	Labor negotiations	57	17	2
31	LED	1000	7	10
32	Letter Recognition	20000	17	26
33	Liver Disorders (Bupa)	345	7	2
34	Lung Cancer	32	57	3
35	Lymphography	148	19	4
36	MAGIC Gamma Telescope	19020	11	2
37	Mushrooms	8124	23	2
38	Nettalk (Phoneme)	5438	8	52
39	New-Thyroid	215	6	3
40	Nursery	12960	9	5
41	Optical Digits	5620	49	10
42	Page Blocks	5473	11	5
43	Pen Digits	10992	17	10
44	Pima Indians Diabetes	768	9	2
45	Postoperative Patient	90	9	3
46	Primary Tumor	339	18	22
47	Promoter Gene Sequences	106	58	2
48	Segment	2310	20	7
49	Sick-euthyroid	3772	30	2
50	Sign	12546	9	3
51	Sonar Classification	208	61	2
52	SPAM E-mail	4601	58	2
53	Splice-junction Gene Sequences	3190	62	3
54	Syncon	600	61	6
55	Teaching Assistant Evaluation	151	6	3
56	Tic-Tac-Toe Endgame	958	10	2
57	Vehicle	846	19	4
58	Volcanoes	1520	4	4
59	Vowel	990	14	11
60	Waveform-5000	5000	41	3
61	Wine Recognition	178	14	3
62	Zoo	101	17	7

research (Webb et al. 2005; Langley and Sage 1994; Pazzani 1996; Domingos and Pazzani 1996; Zheng and Webb 2000; Yang et al. 2006). Each algorithm is tested on each data set using the repeated cross-validation bias-variance estimation method (Webb 2000). In order to maximize the variation in the training data from trial to trial, we use two-fold cross validation. To minimize the variance in our measurements we report average values over 50 cross-validation trials.

We also form learning curves for NB, AODE, A2DE and A3DE on the Adult data set to further investigate how increasing n within the AnDE framework affects performance as the quantity of data increases.

The current implementations of AODE and A2DE are limited to categorical data. A number of approaches have been developed for extending AODE to numeric data (Flores et al. 2009). These could be generalized to the AnDE framework, but how best to do so is a matter for future research. Hence, we assess only the relative capacities of these algorithms with respect to categorical data. To this end, all numeric attributes are discretized. When MDL discretization (Fayyad and Irani 1993), a common discretization method for NB, is used to discretize quantitative attributes within each cross-validation fold, many attributes have only one value. In these experiments, we discretize quantitative attributes using three-bin equal-frequency discretization prior to classification.

The base probabilities are estimated using m-estimation (Cestnik 1990) ($m = 1$), as it often appears to lead to more accurate probabilities than Laplace estimation for NB and AODE. An exception is that we always use 1.0 for $\hat{\mathrm{P}}(x_i \mid y, x_s)$ when $i \in s$.

The above experiments were conducted on a single CPU single core virtual Linux machine running on a Dell PowerEdge 1950 with dual quad core Intel Xeon E5410 processors running at 2333 MHz with 32GB of RAM.[1]

Average values for each combination of metric, algorithm and dataset are provided in the Appendix. Summary results are provided in the text.

4.1 Varying n within AnDE

We first consider the relative performance of the three variants of AnDE. For each performance measure, the number of data sets for which A2DE has lower, equal or higher outcomes relative to AODE and NB are summarized into win/draw/loss records, and likewise for AODE relative to NB. For each win/draw/loss record a binomial sign test is performed to assess the probability of observing the given number of wins and losses if each were equally likely. These results are presented in Table 3. As expected, we see that increasing n from 0 (NB) to 1 (AODE) to 2 (A2DE) consistently decreases bias at the cost of an increase in variance. As we believe that different bias and variance profiles suit different data quantities (Brain and Webb 2002), we believe that the zero-one loss and RMSE results tell us as much about the composition of the data collection as they do about the algorithms. Specifically, we contend that whether one algorithm or another will win on a given dataset is determined

Table 3 Win/draw/loss: AnDE, $n = 0$, 1 and 2, on all 62 data sets

	A2DE vs AODE		A2DE vs NB		AODE vs NB	
	W/D/L	p	W/D/L	p	W/D/L	p
Bias	47/0/15	**<0.001**	49/2/11	**<0.001**	48/0/14	**<0.001**
Variance	19/1/42	**<0.001**	15/0/47	**<0.001**	20/1/41	**0.005**
Zero-one loss	33/2/27	0.259	42/1/19	**0.002**	44/1/17	**<0.001**
RMSE	35/1/26	0.153	15/0/47	**<0.001**	49/1/12	**<0.001**

[1] Due to technical issues including memory leaks in the Weka implementation of Random Forests, it was not possible to complete all 50 runs of 2-fold cross validation for RF10 on Covertype and RF100 on Covertype and Census-Income (KDD). These experiments were instead completed on a Linux Cluster of Xeon 2.8 GHz CPUs, an environment that does not allow reliable time measurements to be taken. For RF10 and RF100 on Covertype, compute times were estimated by averaging over those runs that could be completed on the virtual machine. No runs could be completed on the virtual machine for RF100 on Census-Income (KDD) and so no time results are reported.

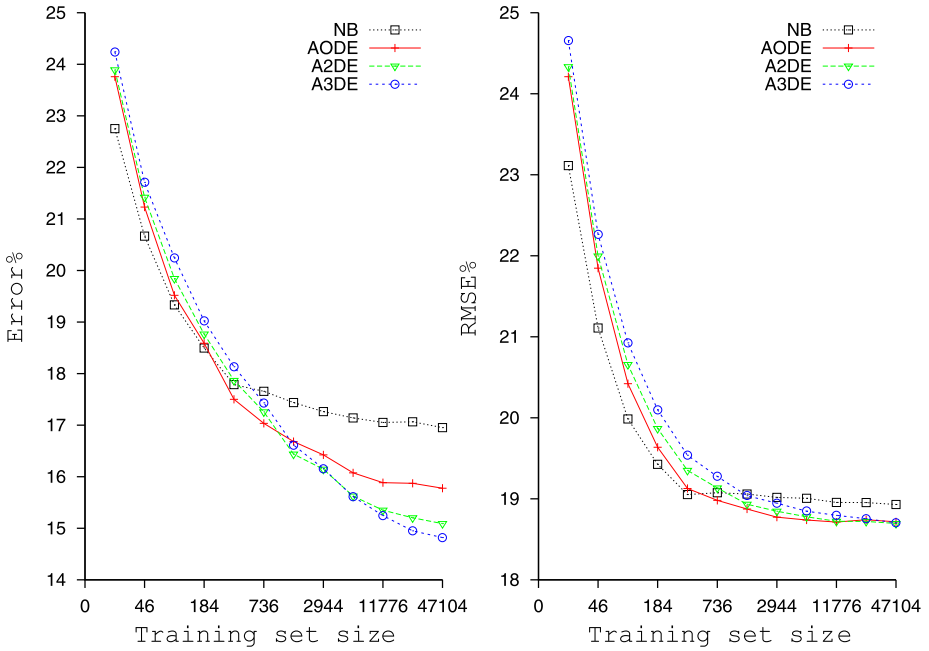

Fig. 2 Zero-one loss and RMSE of NB and AnDE on Adult dataset, as function of training set size

by how well the two algorithms' learning biases match the underlying distribution, by their variance, and by the quantity of data. A low variance algorithm will usually have an advantage for small data while a low bias algorithm will usually be advantaged by large data. For our datasets, both AODE and A2DE reduce both zero-one loss and RMSE significantly often relative to NB. While A2DE obtains lower zero-one loss and RMSE than AODE more often than the reverse, this difference is not found to be significant.

To investigate in greater detail our expectation that algorithms with lower variance will be advantaged for small data and those with lower bias for larger data, we form learning curves for Adult, replicating the method of Webb et al. (2005). 1000 objects are selected at random as a test set and training sets were sampled from the remaining objects. The training set size starts from 23 and then doubles up to 47104, this being a progression that ends with as close to all the available data as possible once the 1000 test cases are removed. This process is repeated 50 times and each algorithm is evaluated on the resulting training-test set pairs. The learning curves of zero-one loss and RMSE for NB, AODE, A2DE and A3DE are presented in Fig. 2.

The plots for zero-one loss clearly show the predicted trade-off for increasing n. At the smallest data size, where low variance is more important than low bias, zero-one loss is minimized by $n = 0$ (NB) and increases as n increases. At the largest data size, where low bias is most important, this dimensionality is reversed. A similar trend is shown with respect to RMSE although the algorithms have not yet achieved their asymptotic rates at the largest data sizes available.

It is interesting to see how the relative bias/variance trade-offs of increasing n play off when NB's attribute independence assumption holds. The LED dataset has a specific configuration of attribute-values for each class, making the attributes conditionally independent given the class. Each attribute has 10% noise added. AODE and A2DE overfit this noise,

Table 4 Win/draw/loss: AnDE, $n = 0$, 1 and 2, on the ten largest data sets

	A2DE vs AODE		A2DE vs NB		AODE vs NB	
	W/D/L	p	W/D/L	p	W/D/L	p
Zero-one loss	10/0/0	**0.001**	10/0/0	**0.001**	10/0/0	**0.001**
RMSE	10/0/0	**0.001**	10/0/0	**0.001**	10/0/0	**0.001**

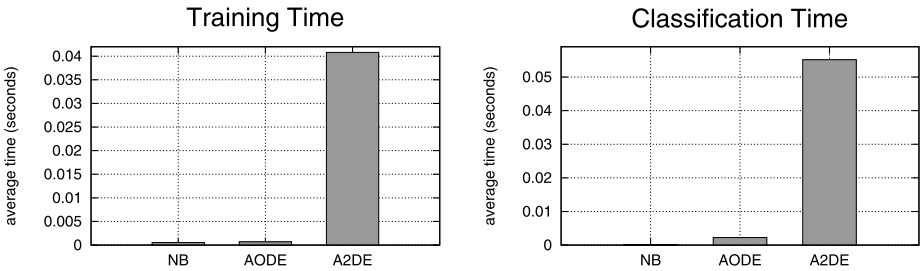

Fig. 3 Average per-example training and classification times for NB, AODE and A2DE

leading to increased error. NB's zero-one loss is 0.2627, AODE's is 0.2639 and A2DE's is 0.2667. These outcomes are with training set sizes of 500. Using the UCI data generator, we generated 10 LED datasets comprising 2,000 and 10 comprising 4,000 instances and repeated the cross-validation experiments thereon. For the datasets of 2,000, the training set size is 1,000 and the mean and standard deviation of the respective zero-one loss is NB: 0.2603±0.0099, AODE: 0.2601±0.0101 and A2DE 0.2603±0.0102. For the datasets of 4,000, the training set size is 2,000 and the mean and standard deviation of the respective zero-one loss is NB: 0.2597±0.0049, AODE: 0.2598±0.0051 and A2DE: 0.2603±0.0053. It seems clear that increasing training set sizes rapidly reduces the error advantage that NB enjoys in this context where its conditional attribute assumption is satisfied.

As final confirmation that higher n is best suited to larger data, on the ten largest datasets, those with more than 8,000 examples, A2DE always achieves lower zero-one loss and RMSE than AODE ($p = 0.001$), see Table 4.

As expected, both training and classification compute times increase as n increases. Figure 3 shows the grand averages for the per-example training and classification times for each algorithm.

4.2 Comparison with TAN

We here explore the relative benefits of discriminative search for a single best Bayesian classifier model against AnDE's search-free approach of averaging over a class of Bayesian classifier models. To this end we compare AODE and A2DE with TAN. Table 5 presents win/draw/loss results comparing AODE and A2DE to TAN.

Overall, TAN has an advantage in bias but a disadvantage in variance relative to AODE. When using search to select a single 1-DBC model is compared to averaging over a class of 2-DBCs, the bias advantage is lost but the variance disadvantage remains. The relative bias-variance tradeoffs of AODE and TAN result in a general error advantage to AODE. Comparing TAN to A2DE, TAN no longer has a bias advantage, and at this higher value of n, the error advantage of the AnDE classifier becomes even more consistent.

Figure 4 shows the relative training and classification times for AODE, A2DE and TAN. It is clear that A2DE has a considerably greater computational requirements both for training

Table 5 Win/draw/loss: AnDE, $n = 1$ and 2 vs TAN on all 62 data sets

	A2DE vs TAN		AODE vs TAN	
	W/D/L	p	W/D/L	p
Bias	34/0/28	0.263	20/1/41	**0.005**
Variance	48/0/14	**<0.001**	52/1/9	**<0.001**
Zero-one loss	48/0/14	**<0.001**	43/1/18	**0.001**
RMSE	43/1/18	**0.001**	40/1/21	**0.010**

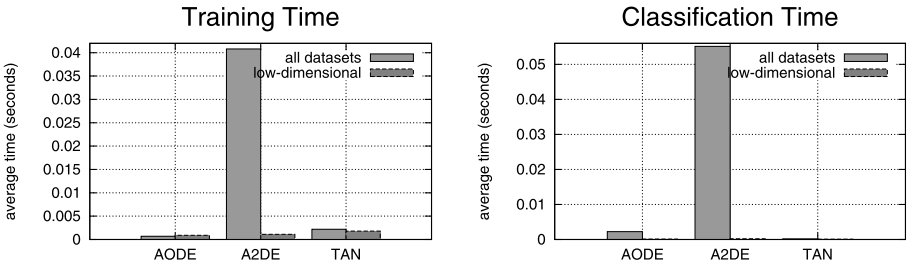

Fig. 4 Average per-example training and classification times for AODE, A2DE and TAN. (Two values are shown for each algorithm, the average across all datasets and the average across the ten lowest-dimensional (4–7 attributes) datasets)

Table 6 Win/draw/loss: AnDE, $n = 1$ and 2 vs MAPLMG on all 62 data sets

	A2DE vs MAPLMG		AODE vs MAPLMG	
	W/D/L	p	W/D/L	p
Bias	40/0/22	**0.015**	17/4/41	**0.001**
Variance	19/1/42	**0.002**	36/5/21	**0.031**
Zero-one loss	30/1/31	0.500	22/4/36	**0.043**
RMSE	34/1/28	0.263	19/0/39	**0.006**

and classification. However, this disadvantage disappears when we consider only the ten lowest dimensional datasets, also illustrated in this figure.

4.3 Comparison with MAPLMG

As discussed above, we wish to investigate the relative payoffs obtained by investing additional computation to that required by AODE by respectively using discriminative learning of weights or, alternatively, increasing the dimensionality of the probabilities from which the posterior probability is extrapolated. To this end, Table 6 presents win/draw/loss results comparing A2DE and AODE to MAPLMG.

As established by previous research (Cerquides and Mántaras 2005), MAPLMG's approach of using discriminative learning of weights for the AODE linear combination significantly reduces bias relative to AODE at the cost of an increase in variance. However, relative to this discriminative approach to extrapolating from three-dimensional probabilities, A2DE's search-free approach to extrapolating from four-dimensional probabilities further reduces bias at the cost of an increase in variance. While the resulting difference in error is not found to be significant across the full suite of 62 datasets, when the ten largest datasets

Table 7 Win/draw/loss: AnDE, $n = 1$ and 2 vs MAPLMG on the ten largest data sets

	A2DE vs MAPLMG		AODE vs MAPLMG	
	W/D/L	p	W/D/L	p
Zero-one loss	10/0/0	**0.001**	1/1/8	**0.039**
RMSE	10/0/0	**0.001**	0/0/10	**0.001**

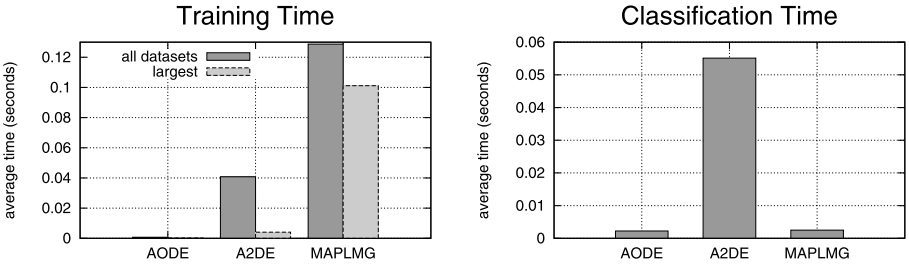

Fig. 5 Average per-example training and classification times for AODE, A2DE and MAPLMG. (In addition to the times for all datasets, training times are shown for the ten largest datasets)

are considered, the lower bias algorithm, A2DE, consistently achieves lower zero-one loss and RMSE than MAPLMG ($p = 0.001$) (see Table 7).

MAPLMG's Bayesian model averaging comes at considerable cost in training time. Figure 5 shows the average per-example training and test times for AODE, A2DE and MAPLMG. Note that MAPLMG is implemented as an external function to Weka, and hence is likely to be inherently more efficient. The training and test times include a substantial fixed overhead, and hence the per-instance training times should decrease if the complexity is linear with respect to the training set size. However, MAPLMG's super-linear training complexity minimizes this effect, demonstrating that it will not be feasible to apply it to very large data.

4.4 Comparison with Feating

To understand how the AnDE approach performs relative to Feating NB, we compare A2DE, that calculates the mean of the joint probabilities, with a variant PA2DE, that calculates the mean of the posterior probabilities, and another variant FA2DE, that calculates the mode of the classes predicted by the submodels. As described in Sect. 3 and the start of Sect. 4, these embody the two main differences between AnDE and Feating NB.

Table 8 shows the win/draw/loss results comparing A2DE to these variants. It is clear that both variants have higher bias but lower variance than A2DE. It is straightforward to understand why Feating would have lower variance. The mode is a much more stable estimator of central tendency than the mean, which can be greatly influenced by a single outlier. It is less obvious why lower variance should result from averaging over the estimates of the posterior rather than of the joint probability. Nonetheless, the result is consistent with Cerquides and de Màntaras' (2005) finding that a linear combination of joint probability estimates resulted in higher accuracy than a linear combination of posterior probability estimates. This remains an interesting unexplained phenomena worthy of further investigation.

Over the full range of datasets these differences in bias and variance profiles do not result in statistically significant differences on either measure of error, except with respect

Table 8 Win/draw/loss: A2DE vs PA2DE and FA2DE

	A2DE vs PA2DE		A2DE vs FA2DE	
	W/D/L	p	W/D/L	p
Bias	41/2/19	**<0.001**	46/1/15	**<0.001**
Variance	21/3/38	**0.018**	22/1/39	**0.020**
Zero-one loss	33/1/28	0.304	36/2/24	**0.078**
RMSE	30/0/32	0.449	49/0/13	**<0.001**

Table 9 Win/draw/loss: A2DE vs PA2DE and FA2DE on the ten largest datasets

	A2DE vs PA2DE		A2DE vs FA2DE	
	W/D/L	p	W/D/L	p
Zero-one loss	8/0/2	0.109	9/0/1	**0.021**
RMSE	8/0/2	0.109	8/1/1	**0.039**

to RMSE for Feating. This reflects the manner in which Feating selects a single class rather than producing a distribution of class probabilities.

Due to its lower bias, A2DE achieves lower zero-one loss than PA2DE on eight and FA2DE on nine of the ten largest datasets. This outcome is statistically significant at the 0.05 level with respect to FA2DE, but misses out on being statistically significant with respect to PA2DE. A2DE achieves lower RMSE than both the alternatives on eight of the ten largest datasets, and draws on one of the remaining datasets with respect to FA2DE, again attaining statistical significance at the 0.05 level relative to FA2DE but failing to do so relative to PA2DE. Hence, the evidence is suggestive that for large data the AnDE approach is preferable to the selection of the mode, and calculating the mean of the joint probabilities is preferable to calculation the mean of the posteriors.

4.5 Comparison with the state-of-the-art

In addition to the relative performance of these related algorithms, it is useful to understand how the performance compares to well known examples of the state-of-the-art. We choose Random Forests (Breiman 2001) as the comparator algorithm because it is relatively unparameterized and hence readily produces clearly understood performance outcomes. We use Random Forests with both the default setting of 10 trees (RF10) and with 100 trees (RF100), allowing us to explore the relative computational/accuracy trade-offs. Table 10 shows the win/draw/loss results for each of A2DE, AODE and NB against RF10 and RF100 for each of zero-one loss, Bias, Variance and RMSE.

All three levels of AnDE have higher bias but lower variance than both levels of Random Forests. This trade-off delivers lower error significantly more often than not for both A2DE and AODE relative to RF10. Both deliver lower error more often than RF100, but not significantly so. Notably, relative to both RF10 and RF100, NB achieves higher error almost as often as lower. This illustrates the weaknesses of such 'bake-offs' with respect to error. As we have argued above, low variance algorithms such as NB will be advantaged by the relatively small data sets used in this study. To assess this effect, we repeated the error comparisons using only the ten largest datasets, those containing more than 8000 examples. The results are shown in Table 11. For these larger datasets, both R10 and RF100 achieve lower error more often than all three of A2DE, AODE and NB, significantly so with respect to AODE and NB and when comparing RF100 to A2DE on zero-one loss. This suggests that

Table 10 Win/draw/loss: AnDE, $n = 0$, 1 and 2, vs RF10 and RF100 on all 62 data sets

		AnDE vs RF10		AnDE vs RF100	
		W/D/L	p	W/D/L	p
A2DE	Bias	18/1/43	**0.001**	22/2/38	**0.026**
	Variance	57/0/5	**<0.001**	45/1/16	**<0.001**
	Zero-one loss	42/0/20	**0.004**	36/3/23	0.059
	RMSE	40/0/22	**0.015**	35/0/27	0.187
AODE	Bias	16/0/46	**<0.001**	20/0/42	**0.004**
	Variance	57/1/4	**<0.001**	47/0/15	**<0.001**
	Zero-one loss	41/0/21	**0.008**	33/1/28	0.304
	RMSE	39/0/23	**0.028**	34/0/28	0.263
NB	Bias	14/1/47	**<0.001**	16/1/45	**<0.001**
	Variance	56/0/6	**<0.001**	51/0/11	**<0.001**
	Zero-one loss	33/0/29	0.352	30/1/31	0.500
	RMSE	30/0/32	0.450	28/0/34	0.263

Table 11 Win/draw/loss: AnDE, $n = 0$, 1 and 2, vs RF10 and RF100 on the ten largest data sets

		AnDE vs RF10		AnDE vs RF100	
		W/D/L	p	W/D/L	p
A2DE	Zero-one loss	2/0/8	0.055	1/1/8	**0.020**
	RMSE	3/0/7	0.172	2/0/8	0.055
AODE	Zero-one loss	0/0/10	**0.001**	0/0/10	**0.001**
	RMSE	1/0/9	**0.011**	0/0/10	**0.001**
NB	Zero-one loss	0/0/10	**0.001**	0/0/10	**0.001**
	RMSE	0/0/10	**0.001**	0/0/10	**0.001**

for very large training data, in the absence of any prior knowledge of the nature of the multivariate probability distribution that the data embodies, Random Forests are likely to achieve lower error than an AnDE classifier, although the data quantity at which this is achieved will be ever greater as the dimensionality of AnDE is increased.

However, Random Forests' error advantage for large data comes at a cost in training time. Figure 6 shows the training and classification times for AODE, A2DE, RF10 and RF100. It is apparent that, overall, RF100 has very high training times. While A2DE's training time does approach RF100's for high dimensional data, for small data and low dimensional data its training times are competitive with RF10. On the other hand, A2DE requires substantially more classification time on average than Random Forests. This requirement grows greatly with high-dimensional data. A2DE will not be feasible for classification of large numbers of high-dimensional objects. In contrast, its classification time is very competitive on low-dimensional data.

5 Conclusions and directions for future research

AnDE provides an attractive framework for developing machine learning techniques. A single parameter n controls a bias-variance trade-off such that $n = a$ provides a classifier whose

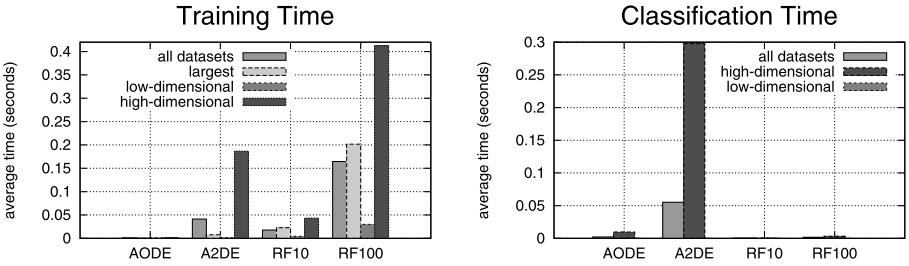

Fig. 6 Average per-example training and classification times for AODE, A2DE, RF10 and RF100. (Training times are presented for all, the ten largest (excluding Census Income, for which RF100 could not be executed on a machine for which reliable times could be obtained, thus 5,620–581,012 examples), the ten lowest dimensional (5–7 attributes) and the ten highest dimensional (43–70 attributes) datasets. Classification times are presented for all, the ten lowest dimensional and the ten highest dimensional datasets)

asymptotic error is the Bayes optimal error rate. However, for high-dimensional data only very low-dimensional forms of AnDE are feasible. Nonetheless, we have established that higher-dimensional variants are likely to deliver greater accuracy than lower-dimensional alternatives when the number of training examples is high. In consequence, a promising direction for future research is to develop computationally efficient techniques for approximating AnDE for high values of n.

A further unresolved issue is how to select an appropriate value of n for any specific dataset $\mathcal{T}$. Are there more computationally efficient approaches than a simple wrapper-based comparison of each possible value?

A number of techniques have been developed for extending AODE to handle numeric data (Flores et al. 2009). There is a need to extend this work to the more general AnDE framework.

We have presented a strategy for learning without fitting the full multivariate probability distribution. We do not argue, however, that fitting the full multivariate probability distribution should necessarily be avoided. Indeed, it has been demonstrated that it is possible to reduce the error of AODE both by appropriate feature selection (Zheng and Webb 2006, 2007; Yang et al. 2007) and weighting of the submodels (Cerquides and Mántaras 2005; Jiang and Zhang 2006; Yang et al. 2007) in order to better fit the full multivariate probability distribution. Therefore, it is likely to be worthwhile to explore efficient methods for each of these strategies for higher values of n. If fast classification is required, and time for training is less constrained, approaches that use search to select a small number of submodels from an AnDE model are likely to be desirable. Where there is sufficient training time available, search for appropriate submodel weights is also likely to be useful.

We have developed a generative learning algorithm that generalizes the principles that underlie AODE to ever higher levels of dimensionality. It has the following desirable features:

- both time and space complexity are linear with respect to the number of training examples;
- it learns in a single pass through the training data;
- it performs direct prediction of class probabilities;
- it has integrated handling of missing values;
- it is robust in the face of noise;
- other than the choice of which instantiation (choice of n) and choice of smoothing technique, the approach uses no tunable parameters;
- it does not perform model selection;

- a simple mechanism controls the bias/variance trade-off;
- it supports incremental learning;
- learning and classification can readily utilize parallel computation; and
- there is a direct theoretical basis that provides optimal prediction except insofar as clearly specified assumptions are violated.

A single parameter n provides control over a bias-variance trade-off, such that higher values of n are appropriate for greater numbers of training cases. AnDE demonstrates that it is possible to develop competitive learners without using search. Of further interest, this family of algorithms show that it is possible to develop low bias algorithms in a generative framework. Finally, A2DE proves to be a computationally tractable version of AnDE that delivers strong classification accuracy for large data without any parameter tuning.

Acknowledgements This research has been supported by Australian Research Council grant DP110101427. We are grateful to Mark Carman, Joao Gama, Kevin Korb and Nayyar Zaidi for insightful discussions on this research and feedback on drafts of this paper.

Appendix: Detailed results

Detailed results for Bias, Variance, zero-one Loss, RMSE, Training Time and Classification Times are presented in Tables 8 to 13. The datasets are listed in ascending order on number of instances.

Table 12 Bias

Dataset	NB	AODE	A2DE	PA2DE	FA2DE	TAN	MAPLMG	RF10	RF100
Contact-lenses	0.1897	0.1970	0.1967	0.2018	0.2002	0.3245	0.1913	0.1710	0.1728
Lung Cancer	0.3413	0.3423	0.3443	0.3702	0.3726	0.3174	0.3428	0.3454	0.3739
Labor negotiations	0.0555	0.0608	0.0650	0.0662	0.0654	0.0807	0.0627	0.1009	0.1011
Postoperative Patient	0.2730	0.2893	0.2949	0.2941	0.2923	0.2717	0.2878	0.2809	0.2907
Zoo	0.0382	0.0308	0.0311	0.0317	0.0315	0.0518	0.0308	0.0302	0.0329
Promoter Gene Sequences	0.0692	0.0902	0.2299	0.1624	0.1737	0.0808	0.0934	0.0901	0.0576
Echocardiogram	0.2653	0.2678	0.2628	0.2613	0.2637	0.2416	0.2616	0.2500	0.2543
Lymphography	0.1286	0.1202	0.1164	0.1153	0.1181	0.1256	0.1185	0.1307	0.1373
Iris Classification	0.0308	0.0286	0.0296	0.0307	0.0293	0.0270	0.0288	0.0306	0.0308
Teaching Assistant Evaluation	0.3567	0.3075	0.2997	0.2999	0.3066	0.3280	0.3109	0.2969	0.3001
Hepatitis	0.1368	0.1262	0.1230	0.1276	0.1267	0.1008	0.1244	0.1307	0.1327
Wine Recognition	0.0430	0.0387	0.0405	0.0391	0.0413	0.0406	0.0388	0.0434	0.0435
Auto Imports	0.2615	0.1775	0.1680	0.1724	0.1740	0.1996	0.1768	0.1380	0.1362
Sonar Classification	0.2227	0.1610	0.1422	0.1501	0.1508	0.1439	0.1593	0.1221	0.1344
Glass Identification	0.2133	0.2124	0.2110	0.2075	0.2105	0.2142	0.2126	0.2041	0.2074
New-Thyroid	0.0997	0.0944	0.0960	0.0933	0.0927	0.1026	0.0947	0.0981	0.0998
Audiology	0.2193	0.2196	0.2193	0.2179	0.2179	0.1817	0.2196	0.1568	0.1712
Hungarian	0.1232	0.1225	0.1232	0.1240	0.1237	0.1262	0.1219	0.1458	0.1496
Heart Disease (Cleveland)	0.1490	0.1470	0.1454	0.1483	0.1483	0.1397	0.1468	0.1514	0.1530
Haberman's Survival	0.2416	0.2408	0.2413	0.2425	0.2420	0.2406	0.2418	0.2342	0.2369
Primary Tumor	0.3811	0.3819	0.3810	0.3827	0.3815	0.3764	0.3833	0.3719	0.3824
Liver Disorders (Bupa)	0.3269	0.3209	0.3204	0.3213	0.3201	0.3137	0.3202	0.3203	0.3242
Ionosphere	0.1758	0.0688	0.0615	0.0633	0.0636	0.0476	0.0672	0.0547	0.0585
Dermatology	0.0107	0.0120	0.0123	0.0113	0.0116	0.0130	0.0119	0.0219	0.0195
Horse Colic	0.1735	0.1527	0.1457	0.1446	0.1448	0.1412	0.1516	0.1283	0.1343

Table 12 (*Continued*)

Dataset	NB	AODE	A2DE	PA2DE	FA2DE	TAN	MAPLMG	RF10	RF100
House Votes 84	0.0917	0.0471	0.0367	0.0436	0.0408	0.0450	0.0415	0.0299	0.0339
Cylinder Bands	0.1548	0.1317	0.1315	0.1378	0.1374	0.3326	0.1293	0.1862	0.2007
Syncon	0.1018	0.0536	0.0436	0.0728	0.0740	0.0300	0.0486	0.0384	0.0382
Balance Scale	0.1381	0.1453	0.1505	0.1503	0.1508	0.1432	0.1454	0.1429	0.1436
Credit Screening	0.1266	0.1197	0.1191	0.1204	0.1209	0.1181	0.1188	0.1137	0.1165
Breast Cancer (Wisconsin)	0.0255	0.0276	0.0263	0.0261	0.0255	0.0269	0.0272	0.0307	0.0314
Pima Indians Diabetes	0.2499	0.2406	0.2276	0.2259	0.2295	0.2189	0.2393	0.2107	0.2122
Vehicle	0.3849	0.2948	0.2821	0.2876	0.2854	0.2678	0.2901	0.2404	0.2458
Annealing	0.0908	0.0676	0.0671	0.0647	0.0654	0.0625	0.0673	0.0969	0.0986
Tic-Tac-Toe Endgame	0.2509	0.2010	0.0719	0.1003	0.1114	0.1693	0.1961	0.0400	0.0275
Vowel	0.3308	0.1370	0.1194	0.1200	0.1244	0.1674	0.1189	0.1054	0.1056
LED	0.2261	0.2268	0.2270	0.2272	0.2266	0.2241	0.2269	0.2261	0.2277
German	0.2100	0.2010	0.1970	0.2023	0.2035	0.1840	0.2012	0.1826	0.1977
Contraceptive Method Choice	0.4683	0.4339	0.4074	0.4168	0.4105	0.3830	0.4247	0.3682	0.3754
Volcanoes	0.5034	0.4630	0.4577	0.4577	0.4583	0.4557	0.4656	0.4539	0.4577
Car Evaluation	0.1061	0.0591	0.0411	0.0379	0.0417	0.0482	0.0520	0.0395	0.0388
Segment	0.1605	0.1211	0.1162	0.1177	0.1178	0.1104	0.1156	0.0925	0.0933
Splice-junction Gene Sequences	0.0383	0.0304	0.0266	0.0581	0.0569	0.0376	0.0302	0.0355	0.0259
King-rook-vs-king-pawn	0.1074	0.0716	0.0541	0.0722	0.0858	0.0626	0.0458	0.0079	0.0062
Hypothyroid (Garavan)	0.0743	0.0738	0.0737	0.0740	0.0742	0.0746	0.0739	0.0739	0.0741
Sick-euthyroid	0.0663	0.0624	0.0613	0.0618	0.0623	0.0589	0.0603	0.0590	0.0590
Abalone	0.4737	0.3925	0.3902	0.3860	0.3903	0.3907	0.3921	0.3745	0.3769
SPAM E-mail	0.3322	0.3324	0.3323	0.3322	0.3322	0.3310	0.3324	0.3308	0.3312
Waveform-5000	0.2070	0.1692	0.1625	0.1854	0.1879	0.1569	0.1629	0.1312	0.1531

Table 12 (*Continued*)

Dataset	NB	AODE	A2DE	PA2DE	FA2DE	TAN	MAPLMG	RF10	RF100
Nettalk (Phoneme)	0.1957	0.1476	0.1512	0.1322	0.1373	0.1878	0.1361	0.1063	0.1098
Page Blocks	0.0721	0.0771	0.0774	0.0767	0.0781	0.0795	0.0776	0.0775	0.0776
Optical Digits	0.0804	0.0286	0.0229	0.0410	0.0421	0.0382	0.0279	0.0243	0.0221
Mushrooms	0.0205	0.0002	0.0000	0.0000	0.0000	0.0002	0.0002	0.0000	0.0000
Pen Digits	0.1562	0.0601	0.0474	0.0516	0.0508	0.0770	0.0531	0.0330	0.0341
Sign	0.4530	0.4001	0.3661	0.3685	0.3678	0.3859	0.3833	0.3378	0.3391
Nursery	0.0901	0.0655	0.0433	0.0454	0.0427	0.0546	0.0668	0.0105	0.0084
MAGIC Gamma Telescope	0.2387	0.2171	0.1992	0.2073	0.2057	0.2135	0.2166	0.2045	0.2049
Letter Recognition	0.4450	0.3740	0.3473	0.3568	0.3635	0.3668	0.3697	0.2564	0.2591
Adult	0.1777	0.1668	0.1579	0.1652	0.1661	0.1552	0.1600	0.1397	0.1416
Connect-4 Opening	0.2646	0.2250	0.2107	0.2421	0.2584	0.2224	0.2203	0.1304	0.1427
Census-Income (KDD)	0.2358	0.0853	0.0604	0.1318	0.1224	0.0628	0.0806	0.0429	0.0447
Covertype	0.3444	0.3180	0.3075	0.3335	0.3444	0.3233	0.3114	0.2363	0.2369
Mean	0.1971	0.1690	0.1632	0.1679	0.1693	0.1693	0.1663	0.1493	0.1521

Table 13 Variance

Dataset	NB	AODE	A2DE	PA2DE	FA2DE	TAN	MAPLMG	RF10	RF100
Contact-lenses	0.1419	0.1472	0.1675	0.1415	0.2098	0.0972	0.1379	0.1873	0.1647
Lung Cancer	0.1794	0.1921	0.1963	0.1967	0.1967	0.2157	0.1916	0.2596	0.2218
Labor negotiations	0.0393	0.0483	0.0470	0.0415	0.0433	0.0902	0.0468	0.1001	0.0827
Postoperative Patient	0.0859	0.0782	0.0876	0.0832	0.0822	0.1027	0.0810	0.1091	0.0961
Zoo	0.0343	0.0308	0.0302	0.0302	0.0324	0.0472	0.0308	0.0478	0.0386
Promoter Gene Sequences	0.0693	0.1119	0.1180	0.0905	0.0897	0.1562	0.1140	0.1661	0.0909
Echocardiogram	0.0636	0.0697	0.0823	0.0858	0.0857	0.0868	0.0742	0.1109	0.1043
Lymphography	0.0410	0.0444	0.0479	0.0421	0.0443	0.1041	0.0457	0.0898	0.0622
Iris Classification	0.0112	0.0110	0.0119	0.0131	0.0152	0.0171	0.0116	0.0126	0.0127
Teaching Assistant Evaluation	0.1649	0.1958	0.1945	0.1980	0.1911	0.1998	0.1965	0.2151	0.2108
Hepatitis	0.0237	0.0327	0.0374	0.0333	0.0338	0.0622	0.0334	0.0605	0.0464
Wine Recognition	0.0174	0.0183	0.0207	0.0199	0.0194	0.0364	0.0184	0.0389	0.0203
Auto Imports	0.1371	0.1375	0.1378	0.1325	0.1311	0.1273	0.1375	0.1550	0.1289
Sonar Classification	0.0756	0.0883	0.0781	0.0843	0.0847	0.1197	0.0865	0.1380	0.0904
Glass Identification	0.0484	0.0400	0.0449	0.0466	0.0455	0.0565	0.0428	0.0618	0.0616
New-Thyroid	0.0192	0.0258	0.0269	0.0247	0.0242	0.0338	0.0256	0.0298	0.0287
Audiology	0.1014	0.0995	0.0986	0.0990	0.1008	0.1588	0.0997	0.1619	0.1104
Hungarian	0.0249	0.0291	0.0381	0.0370	0.0363	0.0596	0.0313	0.0710	0.0568
Heart Disease (Cleveland)	0.0209	0.0299	0.0395	0.0348	0.0343	0.0635	0.0316	0.0867	0.0708
Haberman's Survival	0.0411	0.0569	0.0639	0.0646	0.0595	0.0645	0.0505	0.0853	0.0808
Primary Tumor	0.1594	0.1582	0.1607	0.1576	0.1589	0.2001	0.1580	0.2361	0.2029
Liver Disorders (Bupa)	0.0968	0.1075	0.1100	0.1111	0.1133	0.1186	0.1086	0.1183	0.1155
Ionosphere	0.0326	0.0117	0.0168	0.0096	0.0097	0.0198	0.0131	0.0378	0.0177
Dermatology	0.0095	0.0111	0.0112	0.0106	0.0102	0.0288	0.0111	0.0364	0.0179
Horse Colic	0.0245	0.0369	0.0436	0.0335	0.0332	0.0587	0.0382	0.0499	0.0313

Table 13 (*Continued*)

Dataset	NB	AODE	A2DE	PA2DE	FA2DE	TAN	MAPLMG	RF10	RF100
House Votes 84	0.0061	0.0096	0.0137	0.0098	0.0113	0.0204	0.0112	0.0173	0.0084
Cylinder Bands	0.1065	0.1112	0.1073	0.0954	0.0960	0.0703	0.1100	0.1272	0.0910
Syncon	0.0092	0.0199	0.0196	0.0166	0.0163	0.0259	0.0197	0.0582	0.0254
Balance Scale	0.0555	0.0606	0.0712	0.0724	0.0718	0.0672	0.0622	0.0828	0.0808
Credit Screening	0.0256	0.0270	0.0281	0.0278	0.0272	0.0451	0.0275	0.0578	0.0441
Breast Cancer (Wisconsin)	0.0016	0.0095	0.0127	0.0076	0.0073	0.0241	0.0105	0.0138	0.0076
Pima Indians Diabetes	0.0254	0.0349	0.0508	0.0513	0.0478	0.0576	0.0369	0.0772	0.0729
Vehicle	0.0857	0.0986	0.1049	0.0997	0.1023	0.1233	0.1024	0.1461	0.1305
Annealing	0.0117	0.0174	0.0169	0.0182	0.0179	0.0206	0.0162	0.0452	0.0383
Tic-Tac-Toe Endgame	0.0407	0.0508	0.0536	0.0562	0.0599	0.0787	0.0541	0.0809	0.0490
Vowel	0.2268	0.1637	0.1451	0.1398	0.1444	0.2153	0.1528	0.1424	0.1183
LED	0.0367	0.0371	0.0397	0.0398	0.0392	0.0465	0.0376	0.0551	0.0509
German	0.0481	0.0540	0.0602	0.0519	0.0507	0.0969	0.0546	0.0993	0.0689
Contraceptive Method Choice	0.0719	0.1036	0.1315	0.1190	0.1258	0.1592	0.1132	0.1982	0.1875
Volcanoes	0.0829	0.1333	0.1415	0.1415	0.1413	0.1413	0.1305	0.1469	0.1427
Car Evaluation	0.0506	0.0472	0.0436	0.0405	0.0452	0.0335	0.0456	0.0517	0.0389
Segment	0.0198	0.0184	0.0166	0.0184	0.0189	0.0217	0.0178	0.0269	0.0245
Splice-junction Gene Sequences	0.0083	0.0097	0.0126	0.0175	0.0172	0.0232	0.0098	0.0824	0.0222
King-rook-vs-king-pawn	0.0199	0.0204	0.0200	0.0179	0.0188	0.0155	0.0128	0.0112	0.0067
Hypothyroid (Garavan)	0.0008	0.0006	0.0007	0.0006	0.0007	0.0016	0.0005	0.0076	0.0063
Sick-euthyroid	0.0145	0.0034	0.0028	0.0029	0.0033	0.0053	0.0033	0.0085	0.0073
Abalone	0.0234	0.0837	0.0866	0.0917	0.0859	0.0864	0.0839	0.1052	0.1032
SPAM E-mail	0.0051	0.0050	0.0050	0.0050	0.0051	0.0074	0.0050	0.0067	0.0061
Waveform-5000	0.0132	0.0292	0.0330	0.0166	0.0161	0.0587	0.0328	0.1071	0.0475

Table 13 (*Continued*)

Dataset	NB	AODE	A2DE	PA2DE	FA2DE	TAN	MAPLMG	RF10	RF100
Nettalk (Phoneme)	0.1076	0.1167	0.1295	0.1132	0.1246	0.1602	0.0985	0.0865	0.0686
Page Blocks	0.0177	0.0076	0.0063	0.0070	0.0057	0.0043	0.0061	0.0051	0.0049
Optical Digits	0.0099	0.0086	0.0076	0.0090	0.0091	0.0142	0.0089	0.0419	0.0118
Mushrooms	0.0029	0.0001	0.0000	0.0001	0.0001	0.0001	0.0001	0.0001	0.0000
Pen Digits	0.0079	0.0064	0.0059	0.0054	0.0064	0.0170	0.0068	0.0186	0.0131
Sign	0.0113	0.0100	0.0160	0.0123	0.0189	0.0136	0.0174	0.0193	0.0176
Nursery	0.0077	0.0088	0.0113	0.0104	0.0117	0.0146	0.0123	0.0231	0.0160
MAGIC Gamma Telescope	0.0095	0.0095	0.0182	0.0080	0.0117	0.0079	0.0071	0.0098	0.0089
Letter Recognition	0.0520	0.0478	0.0473	0.0439	0.0469	0.0768	0.0477	0.1000	0.0917
Adult	0.0051	0.0085	0.0105	0.0079	0.0071	0.0140	0.0097	0.0370	0.0331
Connect-4 Opening	0.0146	0.0189	0.0212	0.0146	0.0145	0.0144	0.0196	0.0831	0.0449
Census-Income (KDD)	0.0033	0.0121	0.0110	0.0055	0.0064	0.0077	0.0130	0.0139	0.0102
Covertype	0.0029	0.0028	0.0038	0.0031	0.0024	0.0035	0.0045	0.0149	0.0141
Mean	0.0469	0.0519	0.0551	0.0519	0.0536	0.0664	0.0519	0.0786	0.0626

Table 14 Zero-one loss

Dataset	NB	AODE	A2DE	PA2DE	FA2DE	TAN	MAPLMG	RF10	RF100
Contact-lenses	0.3317	0.3442	0.3642	0.3433	0.4100	0.4217	0.3292	0.3583	0.3375
Lung Cancer	0.5206	0.5344	0.5406	0.5669	0.5694	0.5331	0.5344	0.6050	0.5956
Labor negotiations	0.0947	0.1091	0.1119	0.1077	0.1088	0.1709	0.1095	0.2011	0.1839
Postoperative Patient	0.3589	0.3676	0.3824	0.3773	0.3744	0.3744	0.3689	0.3900	0.3869
Zoo	0.0725	0.0616	0.0614	0.0620	0.0640	0.0990	0.0616	0.0780	0.0715
Promoter Gene Sequences	0.1385	0.2021	0.3479	0.2528	0.2634	0.2370	0.2074	0.2562	0.1485
Echocardiogram	0.3289	0.3376	0.3450	0.3472	0.3495	0.3284	0.3359	0.3609	0.3586
Lymphography	0.1696	0.1646	0.1643	0.1574	0.1624	0.2297	0.1642	0.2204	0.1995
Iris Classification	0.0420	0.0396	0.0415	0.0437	0.0445	0.0441	0.0404	0.0432	0.0435
Teaching Assistant Evaluation	0.5216	0.5033	0.4942	0.4979	0.4977	0.5278	0.5074	0.5119	0.5109
Hepatitis	0.1605	0.1588	0.1604	0.1609	0.1605	0.1630	0.1578	0.1912	0.1791
Wine Recognition	0.0604	0.0570	0.0612	0.0590	0.0608	0.0770	0.0572	0.0824	0.0638
Auto Imports	0.3986	0.3149	0.3059	0.3049	0.3051	0.3268	0.3143	0.2931	0.2651
Sonar Classification	0.2983	0.2493	0.2204	0.2344	0.2355	0.2637	0.2458	0.2601	0.2248
Glass Identification	0.2617	0.2524	0.2559	0.2541	0.2560	0.2707	0.2554	0.2659	0.2690
New-Thyroid	0.1190	0.1202	0.1229	0.1180	0.1169	0.1365	0.1204	0.1279	0.1285
Audiology	0.3206	0.3191	0.3179	0.3169	0.3188	0.3404	0.3193	0.3188	0.2816
Hungarian	0.1482	0.1516	0.1614	0.1610	0.1599	0.1858	0.1532	0.2167	0.2064
Heart Disease (Cleveland)	0.1699	0.1769	0.1849	0.1832	0.1826	0.2032	0.1784	0.2381	0.2238
Haberman's Survival	0.2827	0.2977	0.3052	0.3071	0.3014	0.3051	0.2923	0.3195	0.3178
Primary Tumor	0.5405	0.5401	0.5417	0.5403	0.5405	0.5765	0.5412	0.6079	0.5853
Liver Disorders (Bupa)	0.4237	0.4283	0.4305	0.4323	0.4334	0.4323	0.4288	0.4386	0.4397
Ionosphere	0.2084	0.0806	0.0782	0.0729	0.0733	0.0674	0.0803	0.0925	0.0762
Dermatology	0.0203	0.0231	0.0235	0.0219	0.0218	0.0418	0.0230	0.0584	0.0374
Horse Colic	0.1979	0.1897	0.1893	0.1781	0.1780	0.1998	0.1898	0.1782	0.1656

Table 14 (*Continued*)

Dataset	NB	AODE	A2DE	PA2DE	FA2DE	TAN	MAPLMG	RF10	RF100
House Votes 84	0.0978	0.0567	0.0503	0.0534	0.0521	0.0654	0.0527	0.0471	0.0423
Cylinder Bands	0.2613	0.2430	0.2388	0.2331	0.2334	0.4029	0.2392	0.3134	0.2917
Syncon	0.1110	0.0735	0.0633	0.0894	0.0904	0.0559	0.0684	0.0966	0.0636
Balance Scale	0.1936	0.2059	0.2217	0.2228	0.2226	0.2104	0.2076	0.2258	0.2244
Credit Screening	0.1523	0.1467	0.1473	0.1482	0.1482	0.1632	0.1463	0.1714	0.1606
Breast Cancer (Wisconsin)	0.0271	0.0371	0.0390	0.0337	0.0329	0.0509	0.0377	0.0445	0.0390
Pima Indians Diabetes	0.2752	0.2754	0.2784	0.2772	0.2772	0.2765	0.2763	0.2879	0.2851
Vehicle	0.4707	0.3933	0.3870	0.3873	0.3878	0.3911	0.3924	0.3865	0.3764
Annealing	0.1025	0.0850	0.0840	0.0829	0.0833	0.0831	0.0834	0.1420	0.1369
Tic-Tac-Toe Endgame	0.2916	0.2519	0.1255	0.1566	0.1713	0.2480	0.2502	0.1209	0.0766
Vowel	0.5577	0.3007	0.2645	0.2598	0.2687	0.3827	0.2717	0.2477	0.2239
LED	0.2627	0.2639	0.2667	0.2670	0.2659	0.2706	0.2644	0.2811	0.2786
German	0.2581	0.2550	0.2573	0.2542	0.2542	0.2809	0.2559	0.2819	0.2666
Contraceptive Method Choice	0.5402	0.5375	0.5389	0.5358	0.5363	0.5422	0.5379	0.5664	0.5630
Volcanoes	0.5863	0.5963	0.5992	0.5992	0.5995	0.5970	0.5962	0.6007	0.6004
Car Evaluation	0.1567	0.1063	0.0846	0.0785	0.0869	0.0818	0.0976	0.0912	0.0776
Segment	0.1803	0.1394	0.1328	0.1361	0.1367	0.1321	0.1334	0.1194	0.1178
Splice-junction Gene Sequences	0.0466	0.0401	0.0392	0.0755	0.0741	0.0608	0.0400	0.1179	0.0480
King-rook-vs-king-pawn	0.1273	0.0920	0.0741	0.0901	0.1046	0.0780	0.0586	0.0191	0.0129
Hypothyroid (Garavan)	0.0751	0.0744	0.0744	0.0746	0.0749	0.0761	0.0745	0.0815	0.0804
Sick-euthyroid	0.0808	0.0657	0.0641	0.0647	0.0656	0.0643	0.0636	0.0675	0.0664
Abalone	0.4971	0.4762	0.4769	0.4777	0.4762	0.4770	0.4760	0.4796	0.4801
SPAM E-mail	0.3373	0.3373	0.3373	0.3372	0.3373	0.3384	0.3373	0.3375	0.3373
Waveform-5000	0.2203	0.1984	0.1955	0.2020	0.2040	0.2155	0.1957	0.2383	0.2006
Nettalk (Phoneme)	0.3032	0.2643	0.2807	0.2454	0.2618	0.3480	0.2346	0.1927	0.1784

Table 14 (*Continued*)

Dataset	NB	AODE	A2DE	PA2DE	FA2DE	TAN	MAPLMG	RF10	RF100
Page Blocks	0.0898	0.0847	0.0836	0.0837	0.0837	0.0839	0.0838	0.0826	0.0825
Optical Digits	0.0903	0.0372	0.0305	0.0500	0.0512	0.0524	0.0367	0.0662	0.0339
Mushrooms	0.0234	0.0003	0.0000	0.0001	0.0001	0.0003	0.0003	0.0001	0.0000
Pen Digits	0.1641	0.0665	0.0534	0.0570	0.0572	0.0940	0.0599	0.0517	0.0472
Sign	0.4643	0.4101	0.3821	0.3808	0.3867	0.3995	0.4006	0.3571	0.3567
Nursery	0.0979	0.0743	0.0546	0.0558	0.0544	0.0692	0.0791	0.0336	0.0244
MAGIC Gamma Telescope	0.2482	0.2265	0.2174	0.2153	0.2174	0.2215	0.2237	0.2143	0.2139
Letter Recognition	0.4969	0.4217	0.3946	0.4007	0.4104	0.4436	0.4173	0.3564	0.3508
Adult	0.1828	0.1753	0.1684	0.1731	0.1732	0.1692	0.1698	0.1767	0.1747
Connect-4 Opening	0.2792	0.2440	0.2319	0.2567	0.2729	0.2368	0.2400	0.2135	0.1876
Census-Income (KDD)	0.2391	0.0974	0.0714	0.1373	0.1289	0.0706	0.0936	0.0569	0.0550
Covertype	0.3473	0.3208	0.3113	0.3366	0.3468	0.3268	0.3159	0.2512	0.2510
Mean	0.2440	0.2209	0.2183	0.2199	0.2229	0.2358	0.2182	0.2280	0.2146

Table 15 RMSE

Dataset	NB	AODE	A2DE	PA2DE	FA2DE	TAN	MAPLMG	RF10	RF100
Contact-lenses	0.3782	0.3907	0.3972	0.3924	0.4109	0.4462	0.3837	0.4098	0.3941
Lung Cancer	0.5623	0.5705	0.5736	0.4966	0.4991	0.5169	0.5706	0.4875	0.4736
Labor negotiations	0.2684	0.2820	0.2840	0.2924	0.2857	0.3516	0.2822	0.3708	0.3592
Postoperative Patient	0.4205	0.4266	0.4355	0.4274	0.4510	0.4213	0.4274	0.4448	0.4377
Zoo	0.1245	0.1163	0.1140	0.1125	0.1149	0.1408	0.1160	0.1312	0.1248
Promoter Gene Sequences	0.3342	0.3987	0.4265	0.4159	0.4168	0.4455	0.4019	0.4237	0.3946
Echocardiogram	0.4645	0.4743	0.4841	0.4795	0.5276	0.4769	0.4745	0.5041	0.4993
Lymphography	0.2601	0.2542	0.2557	0.2421	0.2462	0.2895	0.2546	0.2814	0.2710
Iris Classification	0.1450	0.1415	0.1441	0.1472	0.1526	0.1472	0.1424	0.1564	0.1559
Teaching Assistant Evaluation	0.5157	0.5060	0.5078	0.4876	0.5164	0.4875	0.5038	0.4868	0.4825
Hepatitis	0.3654	0.3548	0.3561	0.3427	0.3492	0.3543	0.3538	0.3642	0.3516
Wine Recognition	0.1775	0.1721	0.1784	0.1773	0.1767	0.1952	0.1723	0.2100	0.1965
Auto Imports	0.3053	0.2710	0.2686	0.2490	0.2515	0.2711	0.2706	0.2441	0.2330
Sonar Classification	0.4954	0.4425	0.4176	0.3911	0.3932	0.4490	0.4384	0.4164	0.3935
Glass Identification	0.3618	0.3568	0.3573	0.3541	0.3811	0.3603	0.3575	0.3678	0.3645
New-Thyroid	0.2427	0.2440	0.2463	0.2444	0.2662	0.2592	0.2438	0.2535	0.2518
Audiology	0.1463	0.1462	0.1461	0.1412	0.1498	0.1409	0.1462	0.1398	0.1333
Hungarian	0.3403	0.3374	0.3431	0.3420	0.3545	0.3710	0.3380	0.3911	0.3799
Heart Disease (Cleveland)	0.3714	0.3679	0.3718	0.3685	0.3827	0.3871	0.3687	0.4118	0.3973
Haberman's Survival	0.4552	0.4644	0.4801	0.4814	0.5422	0.4703	0.4626	0.4887	0.4853
Primary Tumor	0.1810	0.1801	0.1801	0.1792	0.1935	0.1813	0.1802	0.1934	0.1890
Liver Disorders (Bupa)	0.4962	0.4962	0.4986	0.4994	0.6126	0.5024	0.4965	0.5193	0.5168
Ionosphere	0.4260	0.2673	0.2591	0.2606	0.2611	0.2369	0.2662	0.2822	0.2672
Dermatology	0.0725	0.0755	0.0767	0.0797	0.0786	0.1009	0.0754	0.1417	0.1306
Horse Colic	0.4067	0.3881	0.3877	0.3740	0.3774	0.4057	0.3885	0.3753	0.3667

Table 15 (*Continued*)

Dataset	NB	AODE	A2DE	PA2DE	FA2DE	TAN	MAPLMG	RF10	RF100
House Votes 84	0.2988	0.2092	0.1947	0.2008	0.2016	0.2277	0.2019	0.1936	0.1842
Cylinder Bands	0.4655	0.4437	0.4400	0.3974	0.4023	0.4900	0.4416	0.4440	0.4353
Syncon	0.1831	0.1373	0.1275	0.1384	0.1399	0.1203	0.1318	0.1708	0.1565
Balance Scale	0.3155	0.3108	0.3099	0.3103	0.3416	0.3148	0.3117	0.3177	0.3150
Credit Screening	0.3414	0.3358	0.3365	0.3326	0.3463	0.3572	0.3346	0.3628	0.3518
Breast Cancer (Wisconsin)	0.1597	0.1777	0.1780	0.1778	0.1715	0.2002	0.1783	0.1915	0.1834
Pima Indians Diabetes	0.4329	0.4309	0.4315	0.4297	0.4794	0.4309	0.4309	0.4542	0.4513
Vehicle	0.4314	0.3560	0.3506	0.3483	0.3684	0.3555	0.3542	0.3654	0.3583
Annealing	0.1535	0.1429	0.1411	0.1417	0.1514	0.1373	0.1416	0.1863	0.1843
Tic-Tac-Toe Endgame	0.4336	0.4044	0.3176	0.3455	0.3385	0.4108	0.4029	0.3132	0.2921
Vowel	0.2556	0.1945	0.1838	0.1855	0.1871	0.2178	0.1869	0.1832	0.1747
LED	0.1986	0.1993	0.2003	0.2007	0.2193	0.2033	0.1994	0.2099	0.2088
German	0.4204	0.4185	0.4223	0.4128	0.4283	0.4470	0.4193	0.4377	0.4218
Contraceptive Method Choice	0.4647	0.4536	0.4540	0.4537	0.5108	0.4586	0.4533	0.5029	0.4970
Volcanoes	0.4116	0.4125	0.4147	0.4149	0.5430	0.4135	0.4125	0.4162	0.4158
Car Evaluation	0.2292	0.2083	0.1922	0.1905	0.1786	0.1848	0.1963	0.1884	0.1782
Segment	0.1922	0.1682	0.1629	0.1664	0.1821	0.1628	0.1630	0.1538	0.1525
Splice-junction Gene Sequences	0.1539	0.1430	0.1428	0.1980	0.1951	0.1751	0.1427	0.2827	0.2601
King-rook-vs-king-pawn	0.3049	0.2719	0.2541	0.2712	0.2604	0.2392	0.2354	0.1449	0.1265
Hypothyroid (Garavan)	0.1865	0.1846	0.1843	0.1848	0.1921	0.1866	0.1848	0.1917	0.1909
Sick-euthyroid	0.2385	0.2302	0.2266	0.2281	0.2441	0.2264	0.2247	0.2317	0.2305
Abalone	0.4801	0.4347	0.4318	0.4325	0.5161	0.4328	0.4314	0.4366	0.4362
SPAM E-mail	0.4524	0.4523	0.4522	0.4521	0.5800	0.4534	0.4523	0.4529	0.4526
Waveform-5000	0.3415	0.3058	0.3022	0.3072	0.3160	0.3230	0.3023	0.3321	0.3146
Nettalk (Phoneme)	0.0953	0.0884	0.0909	0.0883	0.0886	0.0986	0.0844	0.0760	0.0731

Table 15 (*Continued*)

Dataset	NB	AODE	A2DE	PA2DE	FA2DE	TAN	MAPLMG	RF10	RF100
Page Blocks	0.1614	0.1598	0.1593	0.1591	0.1753	0.1593	0.1595	0.1590	0.1588
Optical Digits	0.1227	0.0767	0.0699	0.0853	0.0852	0.0921	0.0762	0.1301	0.1172
Mushrooms	0.1322	0.0153	0.0064	0.0295	0.0288	0.0165	0.0142	0.0123	0.0086
Pen Digits	0.1652	0.1001	0.0895	0.0964	0.0966	0.1186	0.0958	0.0922	0.0886
Sign	0.4369	0.4169	0.4062	0.4050	0.4741	0.4150	0.4114	0.3911	0.3908
Nursery	0.1770	0.1583	0.1424	0.1428	0.1301	0.1425	0.1482	0.1092	0.1009
MAGIC Gamma Telescope	0.4083	0.3956	0.3917	0.3914	0.4184	0.3949	0.3942	0.3896	0.3893
Letter Recognition	0.1573	0.1453	0.1404	0.1421	0.1517	0.1489	0.1444	0.1359	0.1344
Adult	0.3664	0.3519	0.3442	0.3467	0.3649	0.3422	0.3460	0.3542	0.3516
Connect-4 Opening	0.3591	0.3381	0.3303	0.3458	0.3783	0.3322	0.3351	0.3215	0.3057
Census-Income (KDD)	0.4628	0.2731	0.2308	0.3058	0.3085	0.2275	0.2660	0.2098	0.2048
Covertype	0.2604	0.2482	0.2445	0.2545	0.2940	0.2546	0.2460	0.2205	0.2204
Mean	0.3092	0.2890	0.2853	0.2854	0.3045	0.2955	0.2866	0.2945	0.2866

Table 16 Per instance training time

Dataset	NB	AODE	A2DE	PA2DE	FA2DE	TAN	MAPLMG	RF10	RF100
Contact-lenses	0.00571	0.00471	0.00554	0.00521	0.01246	0.00683	0.24567	0.01429	0.04650
Lung Cancer	0.00409	0.00553	0.71631	0.64828	0.69800	0.02534	0.43744	0.03013	0.25059
Labor negotiations	0.00212	0.00219	0.00349	0.00333	0.00295	0.00447	0.10225	0.01247	0.09011
Postoperative Patient	0.00151	0.00146	0.00166	0.00160	0.00124	0.00220	0.05910	0.00791	0.05682
Zoo	0.00131	0.00129	0.00216	0.00223	0.00196	0.00287	0.09247	0.00739	0.04889
Promoter Gene Sequences	0.00192	0.00180	0.18512	0.15537	0.17721	0.01249	0.22453	0.01630	0.14082
Echocardiogram	0.00088	0.00096	0.00104	0.00111	0.00088	0.00157	0.04179	0.00626	0.05076
Lymphography	0.00091	0.00093	0.00205	0.00186	0.00355	0.00228	0.06530	0.00889	0.07238
Iris Classification	0.00102	0.00070	0.00099	0.00083	0.00071	0.00121	0.03442	0.00291	0.01227
Teaching Assistant Evaluation	0.00100	0.00083	0.00125	0.00104	0.00077	0.00427	0.03600	0.00742	0.07446
Hepatitis	0.00100	0.00079	0.00283	0.00174	0.00160	0.00192	0.05664	0.00864	0.07545
Wine Recognition	0.00074	0.00070	0.00266	0.00106	0.00097	0.00146	0.04089	0.00509	0.04271
Auto Imports	0.00074	0.00082	0.01486	0.00898	0.01964	0.00272	0.12048	0.01619	0.14782
Sonar Classification	0.00075	0.00113	0.05417	0.05225	0.06712	0.00514	0.20219	0.01983	0.26942
Glass Identification	0.00064	0.00053	0.00072	0.00072	0.00062	0.00100	0.02960	0.00514	0.04545
New-Thyroid	0.00057	0.00051	0.00066	0.00062	0.00052	0.00251	0.02459	0.00275	0.01636
Audiology	0.00068	0.00552	0.63403	0.58634	0.55959	0.01365	2.24021	0.04293	0.45907
Hungarian	0.00043	0.00041	0.00074	0.00079	0.00065	0.00121	0.02706	0.00712	0.06659
Heart Disease (Cleveland)	0.00045	0.00041	0.00074	0.00092	0.00071	0.00087	0.02678	0.00631	0.06091
Haberman's Survival	0.00041	0.00038	0.00048	0.00041	0.00039	0.00058	0.01670	0.00245	0.01966
Primary Tumor	0.00039	0.00047	0.00130	0.00183	0.00239	0.00119	0.13380	0.01654	0.16519
Liver Disorders (Bupa)	0.00035	0.00030	0.00041	0.00080	0.00037	0.00059	0.01695	0.00328	0.02885
Ionosphere	0.00040	0.00056	0.00463	0.00556	0.00674	0.00175	0.07723	0.01104	0.10150
Dermatology	0.00040	0.00073	0.02099	0.01990	0.02824	0.00230	0.15539	0.01026	0.09502
Horse Colic	0.00037	0.00040	0.00162	0.00161	0.00214	0.00100	0.04104	0.02723	0.25749

Table 16 (*Continued*)

Dataset	NB	AODE	A2DE	PA2DE	FA2DE	TAN	MAPLMG	RF10	RF100
House Votes 84	0.00029	0.00032	0.00086	0.00080	0.00075	0.00070	0.02766	0.00395	0.03675
Cylinder Bands	0.00030	0.00266	0.52135	0.47019	0.15990	0.01125	0.16161	0.16264	1.39439
Syncon	0.00029	0.00087	0.12126	0.11658	0.11387	0.00397	0.47564	0.01881	0.18113
Balance Scale	0.00023	0.00018	0.00023	0.00021	0.00020	0.00029	0.00980	0.00203	0.01860
Credit Screening	0.00023	0.00022	0.00065	0.00065	0.00063	0.00048	0.02263	0.00878	0.08624
Breast Cancer (Wisconsin)	0.00019	0.00019	0.00036	0.00036	0.00084	0.00042	0.01505	0.00252	0.02392
Pima Indians Diabetes	0.00019	0.00016	0.00024	0.00024	0.00023	0.00030	0.01179	0.00446	0.04201
Vehicle	0.00017	0.00020	0.00093	0.00092	0.00127	0.00052	0.03882	0.01148	0.11048
Annealing	0.00019	0.00036	0.01070	0.01043	0.01415	0.00082	0.18188	0.05706	0.62783
Tic-Tac-Toe Endgame	0.00014	0.00013	0.00024	0.00024	0.00022	0.00025	0.01190	0.00430	0.03970
Vowel	0.00014	0.00016	0.00062	0.00058	0.00071	0.00044	0.04711	0.00907	0.09508
LED	0.00012	0.00011	0.00017	0.00017	0.00018	0.00022	0.01648	0.00380	0.03551
German	0.00017	0.00019	0.00114	0.00113	0.00139	0.00051	0.03143	0.00958	0.09459
Contraceptive Method Choice	0.00010	0.00010	0.00020	0.00020	0.00018	0.00029	0.01211	0.00636	0.06713
Volcanoes	0.00009	0.00008	0.00009	0.00008	0.00008	0.00011	0.00551	0.00093	0.00834
Car Evaluation	0.00008	0.00008	0.00011	0.00012	0.00011	0.00014	0.01192	0.00244	0.02308
Segment	0.00008	0.00013	0.00125	0.00094	0.00111	0.00030	0.06186	0.00608	0.06088
Splice-junction Gene Sequences	0.00015	0.00058	0.06743	0.06842	0.06433	0.00200	0.23156	0.02216	0.21648
King-rook-vs-king-pawn	0.00010	0.00024	0.00440	0.00446	0.00463	0.00062	0.10376	0.01554	0.15649
Hypothyroid (Garavan)	0.00009	0.00017	0.00290	0.00265	0.00277	0.00041	0.09119	0.02219	0.22663
Sick-euthyroid	0.00009	0.00016	0.00244	0.00248	0.00254	0.00043	0.05123	0.01783	0.17659
Abalone	0.00005	0.00005	0.00012	0.00012	0.00011	0.00008	0.01016	0.00293	0.02821
SPAM E-mail	0.00013	0.00049	0.01767	0.01750	0.01689	0.00141	0.14233	0.11317	1.15056
Waveform-5000	0.00010	0.00028	0.00724	0.00687	0.00710	0.00076	0.11678	0.02298	0.21661
Nettalk (Phoneme)	0.00004	0.00031	0.01817	0.01616	0.01508	0.00169	0.05430	0.01762	0.17862

Table 16 (*Continued*)

Dataset	NB	AODE	A2DE	PA2DE	FA2DE	TAN	MAPLMG	RF10	RF100
Page Blocks	0.00004	0.00005	0.00017	0.00017	0.00017	0.00009	0.01504	0.00271	0.02672
Optical Digits	0.00012	0.00040	0.04919	0.04899	0.04632	0.00126	0.45907	0.02529	0.21022
Mushrooms	0.00006	0.00012	0.00138	0.00148	0.00149	0.00028	0.02641	0.00457	0.04698
Pen Digits	0.00004	0.00008	0.00059	0.00059	0.00070	0.00016	0.05931	0.00776	0.07362
Sign	0.00003	0.00003	0.00010	0.00010	0.00010	0.00006	0.01297	0.00360	0.03670
Nursery	0.00003	0.00003	0.00011	0.00010	0.00010	0.00006	0.01743	0.00307	0.02825
MAGIC Gamma Telescope	0.00003	0.00004	0.00015	0.00015	0.00015	0.00007	0.01042	0.00494	0.04798
Letter Recognition	0.00005	0.00008	0.00067	0.00063	0.00070	0.00017	0.13907	0.01397	0.13481
Adult	0.00004	0.00007	0.00043	0.00040	0.00042	0.00012	0.02277	0.01949	0.19112
Connect-4 Opening	0.00010	0.00029	0.00722	0.00742	0.00725	0.00076	0.17049	0.03772	0.37281
Census-Income (KDD)	0.00010	0.00033	0.01304	0.01326	0.01279	0.00078	0.11065	0.06108	—
Covertype	0.00013	0.00044	0.01723	0.01756	0.01707	0.00118	0.44194	0.10864	0.87340
Mean (excluding Census-Income)	0.00054	0.00072	0.04129	0.03778	0.03402	0.00220	0.12898	0.01769	0.16448

Table 17 Per instance classification time

Dataset	NB	AODE	A2DE	PA2DE	FA2DE	TAN	MAPLMG	RF10	RF100
Contact-lenses	0.00025	0.00029	0.00071	0.00025	0.00033	0.00033	0.00063	0.00008	0.00021
Lung Cancer	0.00022	0.00416	0.14209	0.14331	0.14019	0.00041	0.00459	0.00009	0.00025
Labor negotiations	0.00009	0.00026	0.00072	0.00100	0.00072	0.00012	0.00065	0.00004	0.00074
Postoperative Patient	0.00014	0.00017	0.00058	0.00050	0.00050	0.00023	0.00044	0.00012	0.00058
Zoo	0.00018	0.00093	0.00446	0.00466	0.00437	0.00018	0.00166	0.00010	0.00039
Promoter Gene Sequences	0.00013	0.00281	0.15011	0.13926	0.14086	0.00025	0.00318	0.00010	0.00034
Echocardiogram	0.00010	0.00015	0.00016	0.00015	0.00016	0.00008	0.00023	0.00009	0.00062
Lymphography	0.00023	0.00076	0.00473	0.00459	0.00416	0.00018	0.00112	0.00016	0.00049
Iris Classification	0.00015	0.00015	0.00023	0.00013	0.00011	0.00009	0.00023	0.00007	0.00029
Teaching Assistant Evaluation	0.00026	0.00013	0.00019	0.00025	0.00014	0.00017	0.00028	0.00015	0.00047
Hepatitis	0.00015	0.00046	0.00554	0.00317	0.00307	0.00012	0.00068	0.00012	0.00057
Wine Recognition	0.00010	0.00042	0.00324	0.00184	0.00164	0.00012	0.00058	0.00013	0.00042
Auto Imports	0.00019	0.00227	0.04120	0.02235	0.02201	0.00026	0.00296	0.00009	0.00068
Sonar Classification	0.00016	0.00378	0.16124	0.16418	0.15268	0.00019	0.00338	0.00008	0.00067
Glass Identification	0.00008	0.00027	0.00068	0.00068	0.00062	0.00008	0.00037	0.00012	0.00054
New-Thyroid	0.00009	0.00011	0.00017	0.00017	0.00018	0.00010	0.00021	0.00008	0.00036
Audiology	0.00073	0.04319	1.25924	1.32199	1.26469	0.00192	0.04573	0.00020	0.00169
Hungarian	0.00008	0.00022	0.00070	0.00079	0.00067	0.00009	0.00038	0.00006	0.00066
Heart Disease (Cleveland)	0.00006	0.00026	0.00127	0.00142	0.00126	0.00007	0.00040	0.00008	0.00054
Haberman's Survival	0.00005	0.00004	0.00006	0.00007	0.00005	0.00005	0.00014	0.00006	0.00030
Primary Tumor	0.00024	0.00300	0.01350	0.02424	0.01361	0.00045	0.00479	0.00022	0.00217
Liver Disorders (Bupa)	0.00008	0.00012	0.00018	0.00028	0.00017	0.00005	0.00017	0.00009	0.00054
Ionosphere	0.00013	0.00132	0.02340	0.02868	0.02342	0.00012	0.00140	0.00011	0.00058
Dermatology	0.00017	0.00313	0.04434	0.04540	0.04470	0.00026	0.00373	0.00008	0.00052
Horse Colic	0.00008	0.00041	0.00276	0.00279	0.00271	0.00011	0.00062	0.00030	0.00399

Table 17 (*Continued*)

Dataset	NB	AODE	A2DE	PA2DE	FA2DE	TAN	MAPLMG	RF10	RF100
House Votes 84	0.00009	0.00039	0.00245	0.00255	0.00243	0.00010	0.00053	0.00007	0.00040
Cylinder Bands	0.00010	0.00154	0.04088	0.04253	0.04065	0.00015	0.00203	0.00008	0.00076
Syncon	0.00025	0.01067	0.33385	0.34002	0.34039	0.00044	0.01024	0.00010	0.00069
Balance Scale	0.00003	0.00008	0.00008	0.00008	0.00007	0.00004	0.00016	0.00006	0.00042
Credit Screening	0.00006	0.00031	0.00192	0.00198	0.00192	0.00008	0.00046	0.00008	0.00078
Breast Cancer (Wisconsin)	0.00005	0.00017	0.00042	0.00045	0.00041	0.00007	0.00028	0.00004	0.00028
Pima Indians Diabetes	0.00004	0.00015	0.00035	0.00036	0.00032	0.00007	0.00021	0.00007	0.00067
Vehicle	0.00009	0.00085	0.00577	0.00607	0.00581	0.00011	0.00105	0.00009	0.00122
Annealing	0.00014	0.00109	0.00335	0.00349	0.00337	0.00016	0.00237	0.00050	0.01097
Tic-Tac-Toe Endgame	0.00005	0.00016	0.00045	0.00048	0.00045	0.00005	0.00026	0.00006	0.00059
Vowel	0.00013	0.00120	0.00484	0.00505	0.00482	0.00019	0.00175	0.00011	0.00144
LED	0.00009	0.00037	0.00058	0.00064	0.00057	0.00012	0.00077	0.00010	0.00076
German	0.00005	0.00049	0.00463	0.00476	0.00468	0.00007	0.00063	0.00007	0.00078
Contraceptive Method Choice	0.00004	0.00018	0.00053	0.00057	0.00053	0.00005	0.00036	0.00008	0.00134
Volcanoes	0.00004	0.00007	0.00005	0.00006	0.00005	0.00004	0.00018	0.00004	0.00035
Car Evaluation	0.00004	0.00014	0.00020	0.00021	0.00019	0.00005	0.00030	0.00005	0.00051
Segment	0.00010	0.00149	0.01394	0.01096	0.01072	0.00017	0.00178	0.00009	0.00069
Splice-junction Gene Sequences	0.00011	0.00464	0.20874	0.21927	0.21233	0.00024	0.00489	0.00008	0.00128
King-rook-vs-king-pawn	0.00006	0.00120	0.02491	0.02568	0.02505	0.00011	0.00145	0.00009	0.00106
Hypothyroid (Garavan)	0.00008	0.00153	0.01972	0.01805	0.01806	0.00014	0.00185	0.00016	0.00280
Sick-euthyroid	0.00005	0.00080	0.01135	0.01198	0.01144	0.00009	0.00098	0.00012	0.00128
Abalone	0.00003	0.00017	0.00042	0.00044	0.00042	0.00005	0.00029	0.00007	0.00064
SPAM E-mail	0.00010	0.00343	0.11847	0.12060	0.11844	0.00016	0.00300	0.00031	0.00310
Waveform-5000	0.00009	0.00263	0.05807	0.05722	0.05632	0.00015	0.00248	0.00011	0.00293
Nettalk (Phoneme)	0.00030	0.00220	0.00313	0.00337	0.00297	0.00061	0.00626	0.00024	0.00239

Table 17 (*Continued*)

Dataset	NB	AODE	A2DE	PA2DE	FA2DE	TAN	MAPLMG	RF10	RF100
Page Blocks	0.00005	0.00038	0.00118	0.00125	0.00118	0.00008	0.00057	0.00006	0.00053
Optical Digits	0.00030	0.01246	0.28798	0.29286	0.28091	0.00055	0.01034	0.00015	0.00249
Mushrooms	0.00004	0.00054	0.00597	0.00601	0.00581	0.00007	0.00071	0.00004	0.00034
Pen Digits	0.00012	0.00164	0.00876	0.00923	0.00881	0.00020	0.00204	0.00012	0.00240
Sign	0.00003	0.00016	0.00042	0.00044	0.00042	0.00004	0.00028	0.00007	0.00084
Nursery	0.00004	0.00024	0.00054	0.00058	0.00055	0.00006	0.00045	0.00007	0.00121
MAGIC Gamma Telescope	0.00003	0.00016	0.00061	0.00063	0.00060	0.00004	0.00025	0.00008	0.00094
Letter Recognition	0.00029	0.00417	0.02229	0.02370	0.02235	0.00048	0.00496	0.00032	0.00595
Adult	0.00004	0.00027	0.00169	0.00167	0.00160	0.00005	0.00037	0.00042	0.00803
Connect-4 Opening	0.00011	0.00289	0.06617	0.06875	0.06519	0.00017	0.00256	0.00039	0.00682
Census-Income (KDD)	0.00009	0.00177	0.05312	0.05313	0.05193	0.00013	0.00180	0.00060	—
Covertype	0.00022	0.00926	0.24854	0.24937	0.24217	0.00045	0.00889	0.00040	0.01561
Mean (excluding Census-Income)	0.00012	0.00224	0.05516	0.05645	0.05434	0.00019	0.00253	0.00013	0.00167

References

Affendey, L., Paris, I., Mustapha, N., Sulaiman, M., & Muda, Z. (2010). Ranking of influencing factors in predicting students' academic performance. *Information Technology Journal*, *9*(4), 832–837.

Birzele, F., & Kramer, S. (2006). A new representation for protein secondary structure prediction based on frequent patterns. *Bioinformatics*, *22*(21), 2628–2634.

Brain, D., & Webb, G. I. (2002). The need for low bias algorithms in classification learning from large data sets. In *Proceedings of the sixth European conference on principles of data mining and knowledge discovery (PKDD)* (pp. 62–73). Berlin: Springer.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Camporelli, M. (2006). *Using a Bayesian classifier for probability estimation: analysis of the AMIS score for risk stratification in myocardial infarction*. Diploma thesis, Department of Informatics, University of Zurich.

Cerquides, J., & Mántaras, R. L. D. (2005). Robust Bayesian linear classifier ensembles. In *Proceedings of the sixteenth European conference on machine learning* (pp. 70–81).

Cestnik, B. (1990). Estimating probabilities: a crucial task in machine learning. In *Proceedings of the ninth European conference on artificial intelligence* (pp. 147–149). London: Pitman.

Domingos, P., & Pazzani, M. J. (1996). Beyond independence: conditions for the optimality of the simple Bayesian classifier. In *Proceedings of the thirteenth international conference on machine learning* (pp. 105–112). San Mateo: Morgan Kaufmann.

Fayyad, U. M., & Irani, K. B. (1993). Multi-interval discretization of continuous-valued attributes for classification learning. In *Proceedings of the thirteenth international joint conference on artificial intelligence* (pp. 1022–1029). San Mateo: Morgan Kaufmann.

Ferrari, L. D., & Aitken, S. (2006). Mining housekeeping genes with a naive Bayes classifier. *BMC Genomics*, *7*(1), 277.

Flikka, K., Martens, L., Vandekerckhove, J., Gevaert, K., & Eidhammer, I. (2006). Improving the reliability and throughput of mass spectrometry-based proteomics by spectrum quality filtering. *Proteomics*, *6*(7), 2086–2094.

Flores, M., Gámez, J., Martínez, A., & Puerta, J. (2009). GAODE and HAODE: two proposals based on AODE to deal with continuous variables. In *Proceedings of the 26th annual international conference on machine learning* (pp. 313–320). New York: ACM.

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. *Machine Learning*, *29*(2), 131–163.

Garcia, B., Aler, R., Ledezma, A., & Sanchis, A. (2008). Protein-protein functional association prediction using genetic programming. In *Proceedings of the tenth annual conference on genetic and evolutionary computation* (pp. 347–348). New York: ACM.

García-Jiménez, B., Juan, D., Ezkurdia, I., Andrés-León, E., & Valencia, A. (2010). Inference of functional relations in predicted protein networks with a machine learning approach. *PLoS ONE*, *4*, e9969.

Hopfgartner, F., Urruty, T., Lopez, P., Villa, R., & Jose, J. (2010). Simulated evaluation of faceted browsing based on feature selection. *Multimedia Tools and Applications*, *47*(3), 631–662.

Hunt, K. (2006). *Evaluation of novel algorithms to optimize risk stratification scores in myocardial infarction*. PhD thesis, Department of Informatics, University of Zurich.

Jiang, L., & Zhang, H. (2006). Weightily averaged one-dependence estimators. In *PRICAI 2006: trends in artificial intelligence* (pp. 970–974).

Kunchevaa, L. I., Vilas, V. J. D. R., & Rodríguezc, J. J. (2007). Diagnosing scrapie in sheep: a classification experiment. *Computers in Biology and Medicine*, *37*(8), 1194–1202.

Kurz, D., Bernstein, A., Hunt, K., Radovanovic, D., Erne, P., Siudak, Z., & Bertel, O. (2009). Simple point-of-care risk stratification in acute coronary syndromes: the AMIS model. *British Medical Journal*, *95*(8), 662.

Langley, P., & Sage, S. (1994). Induction of selective Bayesian classifiers. In *Proceedings of the tenth conference on uncertainty in artificial intelligence* (pp. 399–406). San Mateo: Morgan Kaufmann.

Lasko, T. A., Atlas, S. J., Barry, M. J., & Chueh, K. H. C. (2006). Automated identification of a physician's primary patients. *Journal of the American Medical Informatics Association*, *13*(1), 74–79.

Lau, Q. P., Hsu, W., Lee, M. L., Mao, Y., & Chen, L. (2007). Prediction of cerebral aneurysm rupture. In *Proceedings of the nineteenth IEEE international conference on tools with artificial intelligence* (pp. 350–357). Washington: IEEE Computer Society.

Leon, A., et al. (2009). EcID. A database for the inference of functional interactions in *E. coli*. *Nucleic Acids Research*, *37*, D629 (Database issue).

Liew, C., Ma, X., & Yap, C. (2010). Consensus model for identification of novel PI3K inhibitors in large chemical library. *Journal of Computer-Aided Molecular Design*, *24*(2), 131–141.

Masegosa, A., Joho, H., & Jose, J. (2007). Evaluating query-independent object features for relevancy prediction. In *Advances in information retrieval* (pp. 283–294).

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

Nikora, A. P. (2005). Classifying requirements: towards a more rigorous analysis of natural-language specifications. In *Proceedings of the sixteenth IEEE international symposium on software reliability engineering* (pp. 291–300). Washington: IEEE Computer Society.

Orhan, Z., & Altan, Z. (2006). Impact of feature selection for corpus-based WSD in Turkish. In *Proceedings of the fifth Mexican international conference on artificial intelligence* (pp. 868–878). Berlin: Springer.

Pazzani, M. J. (1996). Constructive induction of Cartesian product attributes. In *ISIS: information, statistics and induction in science* (pp. 66–77).

Sahami, M. (1996). Learning limited dependence Bayesian classifiers. In *Proceedings of the second international conference on knowledge discovery in databases* (pp. 334–338). Menlo Park: AAAI Press.

Shahri, S., & Jamil, H. (2009). An extendable meta-learning algorithm for ontology mapping. In *Flexible query answering systems* (pp. 418–430).

Simpson, M., Demner-Fushman, D., Sneiderman, C., Antani, S., & Thoma, G. (2009). Using non-lexical features to identify effective indexing terms for biomedical illustrations. In *Proceedings of the 12th conference of the European chapter of the association for computational linguistics* (pp. 737–744). Association for Computational Linguistics.

Tian, Y., Chen, C., & Zhang, C. (2008). Aode for source code metrics for improved software maintainability. In *Fourth international conference on semantics, knowledge and grid* (pp. 330–335).

Ting, K. M., Wells, J. R., Tan, S. C., Teng, S. W., & Webb, G. I. (2011). Feature-subspace aggregating: ensembles for stable and unstable learners. *Machine Learning*, *82*(3), 375–397.

Wang, H., Klinginsmith, J., Dong, X., Lee, A., Guha, R., Wu, Y., Crippen, G., & Wild, D. (2007). Chemical data mining of the NCI human tumor cell line database. *Journal of Chemical Information and Modeling*, *47*(6), 2063–2076.

Webb, G. I. (2000). Multiboosting: a technique for combining boosting and wagging. *Machine Learning*, *40*(2), 159–196.

Webb, G. I., Boughton, J., & Wang, Z. (2005). Not so naive Bayes: aggregating one-dependence estimators. *Machine Learning*, *58*(1), 5–24.

Witten, I. H., & Frank, E. (2005). *Data mining: practical machine learning tools and techniques*. San Mateo: Morgan Kaufmann.

Yang, Y., Webb, G., Cerquides, J., Korb, K., Boughton, J., & Ting, K. M. (2006). To select or to weigh: a comparative study of model selection and model weighing for SPODE ensembles. In *Proceedings of the seventeenth European conference on machine learning* (pp. 533–544). Berlin: Springer.

Yang, Y., Webb, G. I., Cerquides, J., Korb, K. B., Boughton, J., & Ting, K. M. (2007). To select or to weigh: a comparative study of linear combination schemes for superparent-one-dependence estimators. *IEEE Transactions on Knowledge and Data Engineering*, *19*(12), 1652–1665.

Yang, Y., Webb, G. I., Korb, K., & Ting, K.-M. (2007). Classifying under computational resource constraints: anytime classification using probabilistic estimators. *Machine Learning*, *69*(1), 35–53.

Zheng, Z., & Webb, G. I. (2000). Lazy learning of Bayesian rules. *Machine Learning*, *41*(1), 53–84.

Zheng, F., & Webb, G. I. (2006). Efficient lazy elimination for averaged-one dependence estimators. In *Proceedings of the twenty-third international conference on machine learning* (pp. 1113–1120). New York: ACM.

Zheng, F., & Webb, G. I. (2007). Finding the right family: parent and child selection for averaged one-dependence estimators. In *Proceedings of the eighteenth European conference on machine learning* (pp. 490–501). Berlin: Springer.