Highly Scalable Attribute Selection for Averaged One-Dependence Estimators

Shenglei Chen[1,2], Ana M. Martinez[2], and Geoffrey I. Webb[2]

[1] College of Information Science,
Nanjing Audit University, Nanjing, China
tristan_chen@126.com
[2] Faculty of Information Technology,
Monash University, VIC 3800, Australia
anam.martinezf@gmail.com, geoff.webb@monash.edu

Abstract. Averaged One-Dependence Estimators (AODE) is a popular and effective approach to Bayesian learning. In this paper, a new attribute selection approach is proposed for AODE. It can search in a large model space, while it requires only a single extra pass through the training data, resulting in a computationally efficient two-pass learning algorithm. The experimental results indicate that the new technique significantly reduces AODE's bias at the cost of a modest increase in training time. Its low bias and computational efficiency make it an attractive algorithm for learning from big data.

Keywords: Classification, Naive Bayes, AODE, Semi-naive Bayes, Attribute Selection

1 Introduction

Naive Bayes (NB) [1] is a simple, computationally efficient probabilistic approach to classification learning. It assumes that all attributes are conditionally independent of each other given the class. As an improvement to NB, Averaged One-Dependence Estimators (AODE) [2] relaxes the attribute independence assumption by averaging all models that assume all attributes are conditionally dependent on the class and one common attribute, known as the super-parent. This often improves the classification performance significantly. An extensive comparative study [3] shows that AODE obtains significant lower error rates than most alternative semi-naive Bayes algorithms with similar computational complexity. One of the attractive features of AODE is that it has complexity linear with respect to data quantity, making it a useful approach for big data.

Attribute selection has been demonstrated to be effective at improving the accuracy of AODE [4, 5]. However, the most effective conventional attribute selection techniques have high computational complexity and hence are not feasible in the context of big data. In this paper we develop an efficient attribute selection algorithm for AODE that is linear with respect to data quantity, and of low polynomial complexity in the number of attributes and hence well suited to

big data. The empirical results show that this technique obtains lower bias than AODE, and thus usually achieves lower error on larger data sets, at the cost of only a modest increase in training time.

2 Background

The classification task can be described as follows, given a training sample $\mathcal{T}$ of t classified objects, we are required to predict the probability $\mathrm{P}(y \mid \mathbf{x})$ that a new example $\mathbf{x} = \langle x_1, \ldots, x_a \rangle$ belongs to some class y, where x_i is the value of the attribute $\mathbf{x}_i$ and $y \in \{c_1, \ldots, c_k\}$.

In the following sections, we describe AODE for this classification task and a number of its key variants.

2.1 AODE

From the definition of conditional probability, we have $\mathrm{P}(y \mid \mathbf{x}) = \mathrm{P}(y, \mathbf{x})/\mathrm{P}(\mathbf{x})$. As $\mathrm{P}(\mathbf{x}) = \sum_{i=1}^{k} \mathrm{P}(c_i, \mathbf{x})$ and $y \in \{c_1, \ldots, c_k\}$, it is reasonable to consider $\mathrm{P}(\mathbf{x})$ as the normalizing constant and estimate only the joint probability $\mathrm{P}(y, \mathbf{x})$ in the remainder of this paper.

Since the example $\mathbf{x}$ does not appear frequently enough in the training data, we cannot directly derive an accurate estimate of $\mathrm{P}(y, \mathbf{x})$ and must extrapolate this estimate from observations of lower-dimensional probabilities in the data [6]. Applying the definition of conditional probabilities again, we have $\mathrm{P}(y, \mathbf{x}) = \mathrm{P}(y)\mathrm{P}(\mathbf{x} \mid y)$. The first term $\mathrm{P}(y)$ on the right side can be sufficiently accurately estimated from the sample frequencies, if the number of classes, k, is not too large. For the second term $\mathrm{P}(\mathbf{x} \mid y)$, AODE assumes every attribute depends on the same parent attribute, the super-parent, thus obtains an one-dependence estimator (ODE), and then averages all eligible ODEs [2]. The joint probability $\mathrm{P}(y, \mathbf{x})$ is estimated as follows,

$$\hat{\mathrm{P}}(y, \mathbf{x}) = \frac{\sum_{i:1 \le i \le a \wedge \mathrm{F}(x_i) \ge m} \hat{\mathrm{P}}(y, x_i) \prod_{j=1}^{a} \hat{\mathrm{P}}(x_j \mid y, x_i)}{|\{i : 1 \le i \le a \wedge \mathrm{F}(x_i) \ge m\}|}, \tag{1}$$

where $|\cdot|$ denotes the cardinality of a set, $\hat{\mathrm{P}}(\cdot)$ represents an estimate of $\mathrm{P}(\cdot)$, $\mathrm{F}(x_i)$ is the frequency of x_i and m is the minimum frequency to accept x_i as a super parent. The current research uses $m = 1$ [7].

2.2 Weightily AODE

In the classification of AODE, each ODE is treated equally, that is, all eligible models are averaged and contribute uniformly to the classification rule. However, in many real world applications, attributes do not play the same role in classification. This observation inspires the weightily AODE [8], in which the joint probability is estimated as,

$$\hat{\mathrm{P}}(y, \mathbf{x}) = \frac{\sum_{i:1 \le i \le a \wedge \mathrm{F}(x_i) \ge m} W_i \hat{\mathrm{P}}(y, x_i) \prod_{j=1}^{a} \hat{\mathrm{P}}(x_j \mid y, x_i)}{\sum_{i:1 \le i \le a \wedge \mathrm{F}(x_i) \ge m} W_i} \ . \tag{2}$$

In practice, mutual information between the super-parent and the class is often used as the weight W_i.

2.3 AODE with Subsumption Resolution

One extreme type of inter-dependence between attributes results in a value of one being a generalization of a value of the other. For example, consider *Gender* and *Pregnant* as two attributes, then $Pregnant = yes$ implies that $Gender = female$. Therefore, $Gender = female$ is a generalization of $Pregnant = yes$. Likewise, $Pregnant = no$ is a generalization of $Gender = male$. Where one value x_i is a generalization of another, x_j, $P(y|x_i, x_j) = P(y|x_j)$. In consequence dropping the more general value from any calculations should not harm any posterior probability estimates, whereas assuming independence between them may.

Motivated by this observation, Subsumption Resolution (SR) [9] identifies pairs of attribute values such that one appears to subsume the other and deletes the generalization. Suppose that the set of indices of the resulting attribute subset is denoted by R, the joint probability is estimated as,

$$\hat{\mathrm{P}}(y, \mathbf{x}) = \frac{\sum_{i:i\in R\wedge \mathrm{F}(x_i)\geq m} \hat{\mathrm{P}}(y, x_i) \prod_{j\in R} \hat{\mathrm{P}}(x_j \mid y, x_i)}{|\{i : i \in R \wedge \mathrm{F}(x_i) \geq m\}|} \ . \tag{3}$$

2.4 Forward and Backward Attribute Selection in AODE

In order to repair harmful inter-dependencies among highly correlated attributes, Zheng et al [5] proposed to select an appropriate attribute subset by hill climbing search. Two different search strategies can be used: FSS begins with the empty attribute set and successively adds attributes [10], while BSE starts with the complete attribute set and successively removes attributes [11]. Both strategies greedily select the attribute whose addition or elimination best reduces the leave-one-out cross validation error on the training set. The process is terminated if there is no error improvement.

To differentiate the selection of parent or child, they introduce the use of a *parent* (p) and a *child* (c) set, each of which contains the set of indices of attributes that can be employed in, respectively, a parent or a child role in AODE. The joint probability is estimated as,

$$\hat{\mathrm{P}}(y, \mathbf{x}) = \frac{\sum_{i:i\in p\wedge \mathrm{F}(x_i)\geq m} \hat{\mathrm{P}}(y, x_i) \prod_{j\in c} \hat{\mathrm{P}}(x_j \mid y, x_i)}{|\{i : i \in p \wedge \mathrm{F}(x_i) \geq m\}|} \ . \tag{4}$$

As indicated in [5], the performance of BSE is better than FSS, so we focus on BSE in this paper. Four types of attribute elimination are considered, *parent elimination* (PE), *child elimination* (CE), *parent and child elimination* (P∧CE), *parent or child elimination* (P∨CE) which performs the former three types of attribute eliminations in each iteration, selecting the option that best reduces the error.

The last strategy allows flexible selection of parents and children, but comes at a high cost, since it needs to scan the training data $2a$ times in the worst case.

2.5 AnDE

The last extension to AODE we review here is AnDE [6], which allows children to depend on not just one super-parent, but a combination of n parents. The joint probability $\mathrm{P}(y, \mathbf{x})$ is estimated as follows,

$$\hat{\mathrm{P}}(y, \mathbf{x}) = \frac{\sum_{s:s\in\binom{\mathcal{A}}{n}\wedge \mathrm{F}(x_s)\geq m} \hat{\mathrm{P}}(y, x_s) \prod_{j=1}^{a} \hat{\mathrm{P}}(x_j \mid y, x_s)}{|\{s : s \in \binom{\mathcal{A}}{n} \wedge \mathrm{F}(x_s) \geq m\}|}, \tag{5}$$

where $\binom{\mathcal{A}}{n}$ indicates the set of all size-n subsets of $\{1, \cdots, a\}$ and x_s means the set of attribute values indexed by the element in s.

Note that AnDE is in fact a superclass of AODE and NB. That is, AODE is AnDE with $n = 1$ (A1DE) and NB is AnDE with $n = 0$ (A0DE).

3 Our Proposal: Attribute Selective AODE

Previous work on attribute selection for AODE through BSE and FSS [4, 5] has demonstrated attribute selection did succeed in reducing the harmful influence of inter-dependencies among attributes. This success may be attributed to their ability to search in a large model space. For P$\vee$CE, the search space is of size 2^{a+1}, as it includes all subsets of attributes in parent role coupled with all subsets of attributes in child role.[1]

Nevertheless, this is achieved at a high computational overhead. The strategy of P$\vee$CE needs to scan the training data $2a$ times, as each time either one child or one parent can be deleted. This is impractical for data sets with a large number of attributes.

In order to explore a large space of models in a single additional pass through the data, we propose a new attribute selection approach for AODE. Our proposal is based on the observation that it is possible to nest a large space of alternative models such that each is a trivial extension to another. Let p and c be the set of indices of parent and child attributes, respectively. For every attribute $\mathbf{x}_i$, the AODE models that use attributes in p as parents and attributes in $c \cup \{i\}$ as children are minor extensions of a model that uses attributes in p as parents and attributes in c as children. The same is true of models that use attributes in $p \cup \{i\}$ as parents and attributes in c as children. Importantly, multiple models that build upon one another in this way can be efficiently evaluated in a single set of computations. Using this observation, we create a space of models that are nested together, and then select the best model using leave-one-out cross validation in single extra pass through the training data.

Step by step information of the algorithm is provided in the following sections.

[1] Note that although the search space is of size 2^{a+1}, the actual number of models evaluated is $\mathcal{O}(a^2)$, which is much smaller.

3.1 Ranking the Attributes

Our method for nesting models depends on a ranking of the attributes. Models containing lower ranked attributes will be built upon models containing higher ranked attributes. The mutual information between an attribute and the class measures how informative this attribute is about the class [12], and thus it is a suitable metric to rank the attributes.

The advantage of using mutual information is that it can be computed very efficiently after one pass through the training data. Although the mutual information between an attribute and the class can help to identify the attributes that are individually most discriminative, it is important to note that it does not directly assess the discriminative power of an attribute in combination with other attributes. Nevertheless, the ranking of attributes based on mutual information with the class will permit the search over a large space of possible models and the deficiencies of this discriminative approach will be mitigated by the richness of the search space that is evaluated in a discriminative fashion.

3.2 Building the Model Space

Without loss of generality, in the following we assume that the attributes are ordered by mutual information. That is, $\mathbf{x}_i$ represents the attribute with the i^{th} greatest mutual information with the class. As the attributes have been ranked, we can create, in total, a^2 nested submodels of attribute subsets. To be more specific, suppose we select top r attributes as parents and top s attributes as children, where $1 \leq r, s \leq a$, the candidate AODE model would be,

$$\hat{\mathrm{P}}(y, \mathbf{x})_{r,s} = \frac{\sum_{i:1\leq i\leq r\wedge \mathrm{F}(x_i)\geq m} \hat{\mathrm{P}}(y, x_i) \prod_{j=1}^{s} \hat{\mathrm{P}}(x_j \mid y, x_i)}{|\{i : 1 \leq i \leq r \wedge \mathrm{F}(x_i) \geq m\}|} \ . \tag{6}$$

Figure 1 gives an example of the model space with 3 attributes. For instance, model m_{21} considers the two attributes $\{\mathbf{x}_1, \mathbf{x}_2\}$ as parents and a single attribute $\{\mathbf{x}_1\}$ as a child. Then, when the attribute $\mathbf{x}_2$ is considered to be added as a child, we obtain a new model m_{22}. When instead the attribute $\mathbf{x}_3$ is considered to be added as a parent, we obtain a new model m_{31}. Both of these models are minor extensions to the existing model m_{21} and all three (and all their extensions) can be applied to a test instance in a single nested computation. Consequently all models can be efficiently evaluated in a single set of nested computations.

		children		
		$\{\mathbf{x}_1\}$	$\{\mathbf{x}_1, \mathbf{x}_2\}$	$\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$
parents	$\{\mathbf{x}_1\}$	m_{11}	m_{12}	m_{13}
	$\{\mathbf{x}_1, \mathbf{x}_2\}$	m_{21}	m_{22}	m_{23}
	$\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$	m_{31}	m_{32}	m_{33}

Fig. 1. An example of the model space with 3 attributes.

3.3 Selecting the Best Model

Once we have built the model space, we can perform model selection within this space. To evaluate the goodness of an alternative model, an evaluation function is required, which commonly measures the discriminative ability of the model among classes.

We use leave-one-out cross validation error to measure the performance of each model. Rather than building a new model for every fold, we use incremental cross validation [13], in which the contribution of the training example being left out in each fold is simply subtracted from the count table, thus producing a model without that training example. This method allows the model to be evaluated quickly, whilst obtaining a good estimate of the generalization error.

There are several loss functions to measure model performance for leave-one-out cross validation, zero-one loss and root mean squared error (RMSE) are among the most common and effective. Zero-one loss simply assigns a loss of '0' to correct classification, and '1' to incorrect classification, treating all misclassifications as equally undesirable. RMSE, however, accumulates for each example the squared error, which is the probability of incorrectly classifying the example, and then computes the root mean of the sum. As RMSE gives a finer grained measure of the calibration of the probability estimates compared to zero-one loss, with the error depending not just on which class is predicted, but also on the probabilities estimated for each class, we use RMSE to evaluate the candidate models in this research.

3.4 Algorithm and Analysis

Based on the methodology presented above, we develop the training algorithm for attribute selective AODE shown in Algorithm 1.

Algorithm 1 Training algorithm for attribute selective AODE.

```
Form the table of joint frequencies of pairwise attribute-values and class
Compute the mutual information
Rank the attributes
for all example in 𝒯 do
    Build all a^2 models while leaving the current example out
    Predict the current example using a^2 models
    Accumulate the squared error for each model
end for
Compute the root mean squared error for each model
Select the model with the lowest RMSE
```

As in AODE, we need to form the table of joint frequencies of pairs of attribute-values and class from which the probability estimates $\hat{\mathrm{P}}(y, x_i)$, $\hat{\mathrm{P}}(x_j \mid y, x_i)$ and the mutual information between the attributes and class are derived.

This is done in one pass through the training data (line 1). Note that this provides all of the information needed to create any selective AODE model with any sets of parent and child attributes.

In the second pass through the training data (line 4-8), the squared error is accumulated for each model. After this pass, the RMSE will be computed and used to select the best model.

At training time, the space complexity of the table of joint frequencies of attribute-values and class is $\mathcal{O}(k(av)^2)$ as in AODE, where v is the average number of values per attribute. Attribute selection will not require more memory. Derivation of the frequencies required to populate this table is of time complexity $\mathcal{O}(ta^2)$. Attribute selection needs one more pass through the training data, the time complexity of which is $\mathcal{O}(tka^2)$, since for each example we need to compute the joint probability in (1) for each class. So the overall time complexity is $\mathcal{O}(t(k+1)a^2)$.

Classification requires the table of probability estimates formed at training time of space complexity $\mathcal{O}(k(av)^2)$. The time complexity of classifying a single example is $\mathcal{O}(ka^2)$ in the worst-case scenario, because some attributes may be omitted after attribute selection.

4 Empirical Comparisons

In this section, we compare the newly proposed attribute selective AODE (ASAODE) with AODE, weightily AODE (WAODE), AODE with subsumption resolution (AODESR), BSE selective AODE (BSEAODE) and A2DE.

Zheng et al [9] discussed three different subsumption resolution techniques, Lazy SR, Eager SR and Near SR. Lazy SR is used in this paper, as it can improve AODE with low training time and modest test time overheads. The minimum frequency for identifying generalizations is set to 100. The results in [5] show that BSE performs better than FSS, and the elimination of a child is more effective than the elimination of a parent. So we select only the children in BSEAODE. However, we do not perform statistical tests in BSEAODE, as we do not do this in ASAODE, either. We also include A2DE in the set of experiments so as to provide a comprehensive comparison.

The experimental system is implemented in C++. In order to deal with numerical data, Minimum Description Length (MDL) discretization [14] is implemented. More specifically, the cut points are computed on 100,000 examples randomly selected from training data or on all training examples if the training data is less than 100,000. These cut points are then used to discretize the training and test data. The base probabilities are estimated using m-estimation ($m = 1$) [15]. Missing values have been considered as a distinct value.

We run the above algorithms on 71 data sets from the UCI repository [16]. Table 1 presents the detailed characteristics of data sets in ascending order on the number of instances. We run the experiments on a single CPU single core virtual Linux machine running on a Sun grid node with dual 6 core Intel Xeon L5640 processors running at 2.27 GHz with 96 GB RAM.

Table 1. Data sets.

No.	Name	Inst	Att	Class
1	contact-lenses	24	4	3
2	lung-cancer	32	56	3
3	labor-negotiations	57	16	2
4	post-operative	90	8	3
5	zoo	101	16	7
6	promoters	106	57	2
7	echocardiogram	131	6	2
8	lymphography	148	18	4
9	iris	150	4	3
10	teaching-ae	151	5	3
11	hepatitis	155	19	2
12	wine	178	13	3
13	autos	205	25	7
14	sonar	208	60	2
15	glass-id	214	9	3
16	new-thyroid	215	5	3
17	audio	226	69	24
18	hungarian	294	13	2
19	heart-disease-c	303	13	2
20	haberman	306	3	2
21	primary-tumor	339	17	22
22	ionosphere	351	34	2
23	dermatology	366	34	6
24	horse-colic	368	21	2
25	house-votes-84	435	16	2
26	cylinder-bands	540	39	2
27	chess	551	39	2
28	syncon	600	60	6
29	balance-scale	625	4	3
30	soybean	683	35	19
31	credit-a	690	15	2
32	breast-cancer-w	699	9	2
33	pima-ind-diabetes	768	8	2
34	vehicle	846	18	4
35	anneal	898	38	6
36	tic-tac-toe	958	9	2
37	vowel	990	13	11
38	german	1000	20	2
39	led	1000	7	10
40	contraceptive-mc	1473	9	3
41	yeast	1484	8	10
42	volcanoes	1520	3	4
43	car	1728	6	4
44	segment	2310	19	7
45	hypothyroid	3163	25	2
46	splice-c4.5	3177	60	3
47	kr-vs-kp	3196	36	2
48	abalone	4177	8	3
49	spambase	4601	57	2
50	phoneme	5438	7	50
51	wall-following	5456	24	4
52	page-blocks	5473	10	5
53	optdigits	5620	64	10
54	satellite	6435	36	6
55	musk2	6598	166	2
56	mushrooms	8124	22	2
57	thyroid	9169	29	20
58	pendigits	10992	16	10
59	sign	12546	8	3
60	nursery	12960	8	5
61	magic	19020	10	2
62	letter-recog	20000	16	26
63	adult	48842	14	2
64	shuttle	58000	9	7
65	connect-4	67557	42	3
66	ipums.la.99	88443	60	19
67	waveform	100000	21	3
68	localization	164860	5	11
69	census-income	299285	41	2
70	poker-hand	1025010	10	10
71	record-linkage	5749132	11	2

4.1 Bias, Variance and RMSE

Because ASAODE explores a larger space of models than AODE and BSEAODE explores a larger space of models than ASAODE, we expect BSEAODE to have the lowest bias, followed by ASAODE then AODE and this order to be reversed for their relative variance. Hence we expect AODE to deliver the lowest error on smaller datasets, ASAODE to dominate at some intermediate data size, and for BSEAODE to deliver the lowest error on very large data. The bias and variance of ASAODE relative to WAODE, AODESR and A2DE can be expected to vary from dataset to dataset as these all embody different learning biases and none of their spaces of models subsumes the other.

In order to assess these expectations, we first perform bias variance decomposition using the experimental method proposed by Kohavi and Wolpert [17]. As this study is more meaningful with more data, we run these experiments only on the largest 28 data sets which have at least 2000 examples. For each data set, 1000 training examples and 1000 test examples are randomly selected. The bias variance decomposition is calculated from the error on the test examples. This process is repeated 10 times to obtain the mean bias and variance.

A summary of pairwise win/draw/loss records, which indicate the number of data sets on which one algorithm has lower, equal or higher outcome relative to the other, is presented in Table 2. Each entry in cell $[i, j]$ compares the algorithm in row i against the algorithm in column j. The p value following each win/draw/loss record is the outcome of a binomial sign test and represents the probability of observing the given number of wins and losses if each were equally likely. The reported p value is the result of a two-tailed test. We consider a difference to be significant if $p \leq 0.05$. All such p values have been changed to boldface in the table.

Table 2 shows that all five variants to AODE achieve significant reductions in bias relative to AODE. While ASAODE achieves lower bias than WAODE and AODESR more often than not, the reverse is true for BSEAODE and A2DE; although these differences are not significant.

Next, we conduct 10-fold cross validation experiments to obtain the error of the alternative algorithms. As attribute selection is based on the RMSE metric, we are inclined to evaluate the error by RMSE. The win/draw/loss records of alternative algorithms for RMSE on 71 data sets are also presented in Table 2.

We can see that all five improvements to AODE have achieved significant reductions in RMSE relative to AODE. ASAODE has also achieved significant reductions in RMSE relative to WAODE and AODESR. The p value (0.807)indicates that ASAODE and BSEAODE have achieved almost the same performance. But the advantages of BSEAODE over WAODE and AODESR are not as significant as those of ASAODE over WAODE and AODESR. While A2DE achieves significant reductions in RMSE relative to AODE, WAODE, AODESR and BSEAODE, its advantage over ASAODE is not significant.

The fact that ASAODE obtains, in general, lower bias and higher variance compared with WAODE and AODESR, indicates that it will perform better on larger datasets, since it will be able to capture more complex relationships from large amount of data [18]. In order to demonstrate this hypothesis, we also compile the win/draw/loss results in terms of RMSE on the 43 smallest data sets and the 28 largest data sets in Table 2. We can see that the performance of ASAODE is better on large data sets than on small data sets. While for even larger data sets BSEAODE and A2DE might outperform ASAODE for the same reason, both have high computational complexity that can be prohibitive for large data, since BSEAODE requires $2a$ pases on the whole training set and A2DE's memory requirements and classification time are very high (see the following Section 4.2).

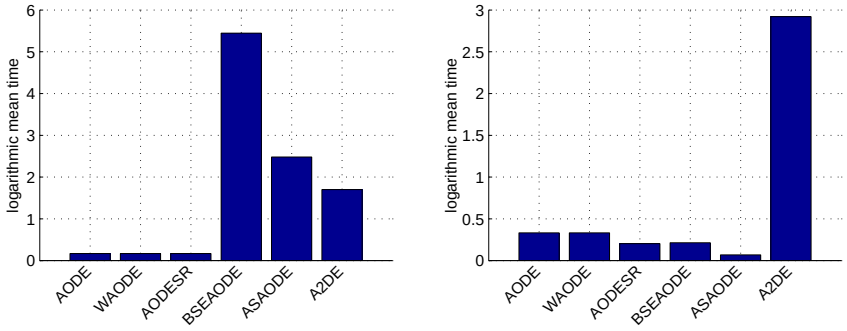
4.2 Computation Time

The logarithmic means of training and classification time on the 71 data sets for all algorithms are shown in Fig. 2. We have added 1 to each mean before computing the logarithm to avoid negative bars. ASAODE requires more training time than such one pass algorithms as AODE, WAODE and AODESR. This is because ASAODE involves two passes through the training data. As BSEAODE needs at most $2a$ passes, it requires significantly more training time than ASAODE.

Table 2. Win/draw/loss records of bias, variance and RMSE with binomial sign test.

		AODE		WAODE		AODESR		BSEAODE		ASAODE	
		W/D/L	p	W/D/L	p	W/D/L	p	W/D/L	p	W/D/L	p
Bias[1]	WAODE	21/2/5	**0.002**								
	AODESR	15/8/5	**0.041**	12/1/15	0.701						
	BSEAODE	19/5/4	**0.003**	17/4/7	0.064	17/2/9	0.169				
	ASAODE	21/3/4	**<0.001**	18/1/9	0.122	16/1/11	0.442	11/2/15	0.557		
	A2DE	23/2/3	**<0.001**	21/1/6	**0.006**	20/3/5	**0.004**	14/1/13	1	17/1/10	0.248
Variance[1]	WAODE	13/1/14	1								
	AODESR	7/8/13	0.263	13/0/15	0.851						
	BSEAODE	11/5/12	1	10/0/18	0.185	12/3/13	1				
	ASAODE	9/1/18	0.122	11/1/16	0.442	13/0/15	0.851	13/2/13	1		
	A2DE	14/1/13	1	14/0/14	1	14/3/11	0.69	14/1/13	1	14/0/14	1
RMSE[2]	WAODE	45/5/21	**0.004**								
	AODESR	32/27/12	**0.004**	28/6/37	0.321						
	BSEAODE	40/20/11	**<0.001**	40/4/27	0.142	35/14/22	0.111				
	ASAODE	43/6/22	**0.013**	42/4/25	**0.05**	42/5/24	**0.036**	35/4/32	0.807		
	A2DE	52/4/15	**<0.001**	47/2/22	**0.004**	48/3/20	**<0.001**	42/4/25	**0.05**	43/2/26	0.053
$RMSE_S$[3]	WAODE	26/3/14	0.081								
	AODESR	20/19/4	**0.002**	18/3/22	0.636						
	BSEAODE	19/14/10	0.136	23/2/18	0.533	16/10/17	1				
	ASAODE	19/5/19	1	19/4/20	1	18/5/20	0.871	20/4/19	1		
	A2DE	27/3/13	**0.038**	24/1/18	0.441	23/2/18	0.533	22/3/18	0.636	25/2/16	0.211
$RMSE_L$[4]	WAODE	19/2/7	**0.029**								
	AODESR	12/8/8	0.503	10/3/15	0.424						
	BSEAODE	21/6/1	**<0.001**	17/2/9	0.169	19/4/5	**0.007**				
	ASAODE	24/1/3	**<0.001**	23/0/5	**<0.001**	24/0/4	**<0.001**	15/0/13	0.851		
	A2DE	25/1/2	**<0.001**	23/1/4	**<0.001**	25/1/2	**<0.001**	20/1/7	**0.019**	18/0/10	0.185

[1] Bias and variance results on the 28 largest data sets.
[2] RMSE results on all the 71 data sets.
[3] $RMSE_S$: RMSE results on the 43 smallest data sets.
[4] $RMSE_L$: RMSE results on the 28 largest data sets.



(a) Training time (b) Classification time

Fig. 2. Computation time comparison of different algorithms (seconds).

As for the classification time, ASAODE, AODESR and BSEAODE require, in general, less time than AODE and WAODE because they might eliminate some attributes. Fig. 2 also shows that ASAODE requires even less classification time than AODESR and BSEAODE.

A2DE requires more training and classification time than AODE, as it needs to compile a more complicated table at training time and requires more computation at classification time.

5 Conclusion

In this paper, a new attribute selection algorithm is proposed for AODE. It is a two-pass algorithm, so compared to AODE, it just requires one more pass through the training data. The alternative attribute selection methods, such as FSA and BSE, need a number of passes that is linear to the number of attributes to obtain similar results.

The empirical results show that the new algorithm is significantly more accurate than AODE, WAODE and AODESR, has comparable error to BSEAODE, and as we expected, worse than A2DE. It requires significantly less training time than BSEAODE, and less classification time than AODE and all other variants, especially than A2DE.

It is worthwhile to note that the technique proposed in this paper is of squared complexity in the number of attributes, so it is not scalable to high dimensional data. On the other hand, it is compatible with weighting, subsumption resolution and higher orders of AnDE. Consequently, it might be possible to further improve the accuracy by combining it with weighting, subsumption resolution and A2DE. This is a promising direction for future research.

Acknowledgments. This research has been supported by the Australian Research Council under grant DP110101427, Asian Office of Aerospace Research and Development, Air Force Office of Scientific Research under contract FA2386-1214030, National Natural Science Foundation of China under grant 71271117, 61202135, Natural Science Foundation of Jiangsu, China under grant BK2011692, BK2012472, Qinglan Project and Priority Academic Program of Audit Science and Technology of Jiangsu, China, Jiangsu Government Scholarship for Overseas Studies, Overseas Studying Scholarship of Nanjing Audit University.

This research has also been supported in part by the Monash e-Research Center and eSolutions-Research Support Services through the use of the Monash Campus HPC Cluster and the LIEF grant. This research was also undertaken on the NCI National Facility in Canberra, Australia, which is supported by the Australian Commonwealth Government.

References

1. Duda, R.O., Hart, P.E.: Pattern Classification and Scene Analysis. 1 edn. John Wiley & Sons Inc (1973)

2. Webb, G.I., Boughton, J.R., Wang, Z.: Not so naive Bayes: Aggregating one-dependence estimators. Machine Learning **58**(1) (2005) 5–24
3. Zheng, F., Webb, G.I.: A comparative study of semi-naive Bayes methods in classification learning. In: AusDM. (2005) 141–156
4. Yang, Y., Webb, G.I., Cerquides, J., Korb, K.B., Boughton, J., Ting, K.M.: 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) (2007) 1652–1665
5. Zheng, F., Webb, G.I.: Finding the right family: parent and child selection for averaged one-dependence estimators. In Kok, J.N., Koronacki, J., de Mantaras, R.L., Matwin, S., Mladeni, D., Skowron, A., eds.: ECML. Springer (2007) 490–501
6. Webb, G.I., Boughton, J.R., Zheng, F., Ting, K.M., Salem, H.: Learning by extrapolation from marginal to full-multivariate probability distributions: decreasingly naive Bayesian classification. Machine Learning **86**(2) (2012) 233–272
7. Cerquides, J., de Mántaras, R.L.: Robust Bayesian linear classifier ensembles. In Gama, J., Camacho, R., Brazdil, P.B., Jorge, A.M., Torgo, L., eds.: ECML, Springer (2005) 72–83
8. Jiang, L., Zhang, H.: Weightily averaged one-dependence estimators. In Yang, Q., Webb, G.I., eds.: PRICAI: trends in artificial intelligence. Springer (2006) 970–974
9. Zheng, F., Webb, G.I., Suraweera, P., Zhu, L.: Subsumption resolution: an efficient and effective technique for semi-naive Bayesian learning. Machine Learning **87**(1) (2012) 93–125
10. Langley, P., Sage, S.: Induction of selective Bayesian classifiers. In: Proceedings of the tenth international conference on uncertainty in artificial intelligence, Morgan Kaufmann Publishers Inc. (1994) 399–406
11. Kittler, J.: Feature selection and extraction. Handbook of pattern recognition and image processing (1986) 59–83
12. MacKay, D.J.: Information theory, inference and learning algorithms. Cambridge university press (2003)
13. Kohavi, R.: The power of decision tables. In Lavrac, N., Wrobel, S., eds.: ECML, Springer (1995) 174–189
14. Fayyad, U.M., Irani, K.B.: Multi-interval discretization of continuous-valued attributes for classification learning. In: IJCAI. (1993) 1022–1027
15. Cestnik, B.: Estimating probabilities: a crucial task in machine learning. In: ECAI. Volume 90. (1990) 147–149
16. Bache, K., Lichman, M.: UCI machine learning repository (2013)
17. Kohavi, R., Wolpert, D.H.: Bias plus variance decomposition for zero-one loss functions. In: ICML. (1996) 275–283
18. Brain, D., Webb, G.I.: The need for low bias algorithms in classification learning from large data sets. In Elomaa, T., Mannila, H., Toivonen, H., eds.: Principles of Data Mining and Knowledge Discovery. Springer (2002) 62–73