Fast and Effective Single Pass Bayesian Learning

Nayyar A. Zaidi and Geoffrey I. Webb

Faculty of Information Technology, Monash University, VIC 3800, Australia
{nayyar.zaidi,geoff.webb}@monash.edu

Abstract. The rapid growth in data makes ever more urgent the quest for highly scalable learning algorithms that can maximize the benefit that can be derived from the information implicit in big data. Where data are too big to reside in core, efficient learning requires minimal data access. Single pass learning accesses each data point once only, providing the most efficient data access possible without resorting to sampling. The AnDE family of classifiers are effective single pass learners. We investigate two extensions to A2DE, subsumption resolution and MI-weighting. Neither of these techniques require additional data access. Both reduce A2DE's learning bias, improving its effectiveness for big data. Furthermore, we demonstrate that the techniques are complementary. The resulting combined technique delivers computationally efficient low-bias learning well suited to learning from big data.

Keywords: Averaged n-Dependence Estimators, Subsumption Resolution, Big Data, Naive Bayes, Bias-Variance Trade-off.

1 Introduction

When data are too big to reside in RAM, machine learning has two options. The first is learn from a sample, thereby potentially losing information implicit in the data as a whole. The second is to process the data out-of-core. In the latter case, data access is very expensive, and single-pass learning becomes very desirable. The Averaged n-Dependence Estimators (AnDE) family of Bayesian learning algorithms provide efficient single pass learning with accuracy competitive with the state-of-the-art in-core learning [1]. In addition, AnDE classifiers

- have time complexity linear with respect to the number of training examples,
- directly handle multiple class problems,
- directly handle missing values, and
- do not require parameter tuning.

These features make them strong contenders for application with big data.

Previous research has shown that as n is increased, the bias of the AnDE algorithms decreases, at the cost of an increase in variance [1]. Variance tends to decrease as data quantity increases, so for big data low bias algorithms tend to have an advantage [2]. Hence, for large data, larger n is desirable. Unfortunately, however, large n has high time and space complexity, especially as the

J. Pei et al. (Eds.): PAKDD 2013, Part I, LNAI 7818, pp. 149–160, 2013.

dimensionality of the data increases. In practice, A2DE has proven effective for moderate dimensional data.

A number of techniques have demonstrated a capacity to lower the bias of A1DE with negligible computational cost. Subsumption Resolution (SR) [3] achieves this with a form of lazy (classification time) feature elimination. Weightily Averaged One-Dependence Estimators (WAODE) [4] achieves it by weighting the sub-models. While previous studies have demonstrated the independent effectiveness of each of these algorithms, their interoperability has not previously been investigated. In this paper we investigate whether they are compatible and the extent to which applying both together reduces bias relative to applying each alone. Further, neither of these techniques has been studied in the context of AnDE with n greater than 1. We herein investigate their effectiveness when applied to A2DE, both severally and jointly. We reveal that they are indeed effective at further reducing A2DE's bias with minimal additional computation.

The rest of this paper is organized as follows. We discuss related work and our proposed improvements to A2DE in section 2. We will discuss experimental results in section 3. We conclude in section 4.

2 Semi-naive Bayes Method - AnDE

We seek to estimate $\mathrm{P}(y \mid \mathbf{x})$, where y is a class label and $\mathbf{x}$ is a vector of attribute values $\mathbf{x} = \langle x_1, \dots x_m \rangle$. For notational convenience we define

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

For example, $x_{\{2,3,5\}} = \langle x_2, x_3, x_5 \rangle$. We use $\hat{\mathrm{P}}(\cdot)$ to denote an estimate of $\mathrm{P}(\cdot)$.

AnDE aims to estimate $\mathrm{P}(y \mid \mathbf{x})$ using $\mathrm{P}(y \mid \mathbf{x}) \propto \mathrm{P}(y, \mathbf{x})$ and hence normalizing each $\hat{\mathrm{P}}(y, \mathbf{x})$ to derive the respective $\hat{\mathrm{P}}(y \mid \mathbf{x})$. The required joint probability is estimated using

$$\hat{\mathrm{P}}_{\mathrm{AnDE}}(y, \mathbf{x}) = \begin{cases} \sum_{s \in \binom{\mathcal{A}}{n}} \delta(x_s) \hat{\mathrm{P}}(y, x_s) \prod_{i=1}^{a} \hat{\mathrm{P}}(x_i \mid y, x_s) / \sum_{s \in \binom{\mathcal{A}}{n}} \delta(x_s) & : \sum_{s \in \binom{\mathcal{A}}{n}} \delta(x_s) > 0 \\ \hat{\mathrm{P}}_{\mathrm{A(n\text{-}1)DE}}(y, \mathbf{x}) & : \text{otherwise} \end{cases} \tag{1}$$

where $\binom{\mathcal{A}}{n}$ indicates the set of all size-n subsets of $\{1, \dots a\}$ and $\delta(x_\alpha)$ is a function that is 1 if the training data contains an object with the value x_α, otherwise 0.

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.

Subsumption resolution [3] is an effective technique for rectifying a specific class of extreme violations of the attribute independence assumption, those where $\mathrm{P}(x_i \mid x_j) = 1.0$. In this case $\mathrm{P}(y \mid \mathbf{x}) = \mathrm{P}(y \mid x_{\{1 \dots i-1, i+1 \dots m\}})$ and hence all inaccuracies introduced into $\hat{\mathrm{P}}(y \mid \mathbf{x})$ by this violation of the attribute independence assumption can be avoided by dropping x_i from (1). For example, when the attribute values include *female* and *pregnant* only the latter should be used, when they include *male* and *not-pregnant* only the former should be used, and when they include *female* and *not-pregnant* both should be used. This requires, however, that one infer whether $\mathrm{P}(x_i \mid y, x_s) = 1$ for each pair of attribute values. In the current research we infer that $\mathrm{P}(x_i \mid x_j) = 1.0$ if $\#(x_j) = \#(x_i, x_j) > 100$, where $\#(x_j)$ is the count of the number of times attribute value x_j occurs in the data and $\#(x_i, x_j)$ is the count of the number of times both x_i and x_j occur together in the data. To prevent both attribute values being deleted if they cover exactly the same data, we delete the one with the higher index if $\#(x_i) = \#(x_j)$.

$$\hat{\mathrm{P}}_{\mathrm{AnDE^{SR}}}(y, \mathbf{x}) = \hat{\mathrm{P}}_{\mathrm{AnDE}}(y, x_{\{i\in\mathbf{x}:\neg\exists j\in\mathbf{x}\#(x_i)=\#(x_i,x_j)>100\wedge[\#(x_j)>\#(x_i)\vee j<i]\}})$$

Subsumption resolution has been shown to be effective at reducing the bias of A1DE [5,3].

Another approach to reducing bias in AnDE that has been shown to be effective for A1DE [6,4,7] is to weight the sub-models, modifying (1) to

$$\hat{\mathrm{P}}_{\mathrm{WAnDE}}(y, \mathbf{x}) = \begin{cases} \sum_{s\in\binom{\mathcal{A}}{n}} \delta(x_s) w_s \hat{\mathrm{P}}(y, x_s) \prod_{i=1}^{a} \hat{\mathrm{P}}(x_i \mid y, x_s) / \sum_{s\in\binom{\mathcal{A}}{n}} \delta(x_s) & : \sum_{s\in\binom{\mathcal{A}}{n}} \delta(x_s) > 0 \\ \hat{\mathrm{P}}_{\mathrm{WA(n\text{-}1)DE}}(y, \mathbf{x}) & : \text{otherwise} \end{cases}$$

WAODE [4] weights A1DE, where s is a single attribute value. It sets w_s to the mutual information of the attribute with the class. WAODE is effective at reducing the bias of A1DE with minimal computational overhead. We here generalize that strategy to MI-weighted AnDE, using $w_s = \mathrm{MI}(S, Y)$,

$$\mathrm{MI}(s, Y) = \sum_{y\in Y} \sum_{x_s\in X_s} \mathrm{P}(x_s, y) \log \frac{\mathrm{P}(x_s, y)}{\mathrm{P}(x_s)\mathrm{P}(y)} \qquad (2)$$

where Y is the set of class labels and X_s is the cross product of values for attributes with indices in s.

While subsumption resolution and weighting have each been shown to reduce the bias of AnDE in isolation, they have not previously been used in conjunction. To assess the effect of doing so we also evaluate MI-weighted AnDESR,

$$\hat{\mathrm{P}}_{\mathrm{WAnDE^{SR}}}(y, \mathbf{x}) = \hat{\mathrm{P}}_{\mathrm{WAnDE}}(y, x_{\{i\in\mathbf{x}:\neg\exists j\in\mathbf{x}\#(x_i)=\#(x_i,x_j)>100\wedge[\#(x_j)>\#(x_i)\vee j<i]\}})$$

2.1 Computational overheads

AnDE has training time complexity of $\mathrm{O}(t\binom{m}{n+1})$ and classification time complexity of $\mathrm{O}(km\binom{m}{n})$ for classifying a single example, where t is the number of training examples.

Subsumption resolution requires no additional training time and at classification time requires $\binom{m}{2}$ comparisons to identify any subsumed attribute values,

Table 1. Data sets

Domain	Case	Att	Class	Domain	Case	Att	Class
Abalone	4177	9	3	Liver Disorders (Bupa)	345	7	2
Adult	48842	15	2	Lung Cancer	32	57	3
Annealing	898	39	6	Lymphography	148	19	4
Audiology	226	70	24	MAGIC Gamma Telescope	19020	11	2
Auto Imports	205	26	7	Mushrooms	8124	23	2
Balance Scale	625	5	3	Nettalk(Phoneme)	5438	8	52
Breast Cancer (Wisconsin)	699	10	2	New-Thyroid	215	6	3
Car Evaluation	1728	8	4	Nursery	12960	9	5
Census-Income (KDD)	299285	40	2	Optical Digits	5620	49	10
Connect-4 Opening	67557	43	3	Page Blocks Classification	5473	11	5
Contact-lenses	24	5	3	Pen Digits	10992	17	10
Contraceptive Method Choice	1473	10	3	Pima Indians Diabetes	768	9	2
Covertype	581012	55	7	Postoperative Patient	90	9	3
Credit Screening	690	16	2	Primary Tumor	339	18	22
Echocardiogram	131	7	2	Promoter Gene Sequences	106	58	2
German	1000	21	2	Segment	2310	20	7
Glass Identification	214	10	3	Sick-euthyroid	3772	30	2
Haberman's Survival	306	4	2	Sign	12546	9	3
Heart Disease (Cleveland)	303	14	2	Sonar Classification	208	61	2
Hepatitis	155	20	2	Splice-junction Gene Sequences	3190	62	3
Horse Colic	368	22	2	Statlog (Shuttle)	58000	10	7
House Votes 84	435	17	2	Syncon	600	61	6
Hungarian	294	14	2	Teaching Assistant Evaluation	151	6	3
Hypothyroid(Garavan)	3772	30	4	Tic-Tac-Toe Endgame	958	10	2
Ionosphere	351	35	2	Vehicle	846	19	4
Iris Classification	150	5	3	Volcanoes	1520	4	4
King-rook-vs-king-pawn	3196	37	2	Vowel	990	14	11
Labor Negotiations	57	17	2	Waveform-5000	5000	41	3
LED	1000	8	10	Wine Recognition	178	14	3
Dermatology	366	35	6	Zoo	101	17	7
Cylinder	540	40	2	Letter Recognition	20000	17	26
Spambase	4601	58	2	Localization	164860	7	3
Wall-following	5456	25	4	Poker-hand	1025010	11	10
yeast	1484	9	10	Thyroid	9169	30	20
Satellite	6435	37	6	Musk1	476	167	2
Chess	551	40	2				

and hence does not increase the classification time complexity so long as $n > 0$. In practice subsumption resolution can substantially reduce classification time by reducing the number combinations of attribute values that must be processed.

MI weighted AnDE requires the calculation of the weights at training time, $O(k\binom{m}{n})$. In practice this is dominated by the training time complexity of regular AnDE and hence does not increase the effective complexity and the additional training time impost is modest. The classification time impact is negligible.

3 Experimental Results

The experiments are conducted in the Weka work-bench (version 3.5.7) on data sets described in table 1. Each algorithm is tested on each data set using 20 rounds of 2-fold cross validation. Probability estimates were smoothed using m-estimation [8] with $m = 1$.

The bias-variance decomposition provides valuable insights into the components of the error of learned classifiers. *Bias* denotes the systematic component of error, which describes how closely the learner is able to describe the decision surfaces for a domain. *Variance* describes the component of error that stems from sampling, which reflects the sensitivity of the learner to variations in the training sample [9,10]. There are a number of different bias-variance decomposition definitions. In this research, we use the bias and variance definitions of [9], together with the repeated cross-validation bias-variance estimation method [10]. When two algorithms are compared, we count the number of data sets for which one algorithm performs better, equally well or worse than the other on a given measure. A standard binomial sign test, assuming that wins and losses are equiprobable, is applied to these records. We assess a difference as significant if the outcome of a two-tailed binomial sign test is less than 0.05. The base probabilities of each algorithm are estimated using m-estimation, since in our initial experiments it leads to more accurate probabilities than Laplace estimation for naive Bayes, A1DE and A2DE. The data sets are divided into four categories. First, consisting of all 71 data sets. Second, large data sets with number of instances $> 10,000$. Third, medium data sets with number of instances > 1000 and $< 10,000$. Fourth, small data sets with number of instances < 1000. The following techniques are compared:

- NB, Standard naive Bayes with m-estimates of probabilities.
- A1DE, $\hat{\mathrm{P}}_{\mathrm{AnDE}}(y, \mathbf{x})$ with $n = 1$.
- A1DE-S, $\hat{\mathrm{P}}_{\mathrm{AnDE^{SR}}}(y, \mathbf{x})$ with $n = 1$.
- A1DE-W, $\hat{\mathrm{P}}_{\mathrm{WAnDE}}(y, \mathbf{x})$ with $n = 1$.
- A1DE-SW, $\hat{\mathrm{P}}_{\mathrm{WAnDE^{SR}}}(y, \mathbf{x})$ with $n = 1$.
- A2DE, $\hat{\mathrm{P}}_{\mathrm{AnDE}}(y, \mathbf{x})$ with $n = 2$.
- A2DE-S, $\hat{\mathrm{P}}_{\mathrm{AnDE^{SR}}}(y, \mathbf{x})$ with $n = 2$.
- A2DE-W, $\hat{\mathrm{P}}_{\mathrm{WAnDE}}(y, \mathbf{x})$ with $n = 2$.
- A2DE-SW, $\hat{\mathrm{P}}_{\mathrm{WAnDE^{SR}}}(y, \mathbf{x})$ with $n = 2$.
- RF10, Random Forest with 10 decision trees.

Numeric attributes are discretized using MDL discretization [11] for all compared techniques except Random Forest. Bias, variance, 0-1 Loss and RMSE results are reported in the following sections.

3.1 Comparison of Bias and Variance

The WDL bias and variance results are shown in Tables 2 and 3 respectively with significant ($\alpha = 0.05$) results shown in bold. We summarize the results as:

- Both weighting and subsumption resolution reduce the bias of both A1DE and A2DE significantly more often than they increase it.
- Jointly applying both weighting and subsumption resolution to either A1DE or A2DE reduces bias significantly more often than it increases it relative to applying either alone.
- Both weighting and subsumption resolution increase the variance of both A1DE and A2DE more often than they decrease it, although these results are not always statistically significant.

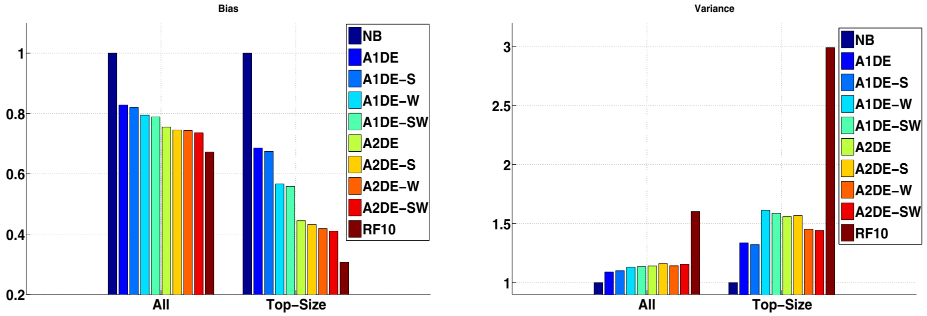

Fig. 1. Averaged Bias (left) and Variance (Right) results normalized with respect to NB. The error-bars are ordered in the same sequence as in the legend.

Table 2. Win/Draw/Loss of Bias Comparison, all data sets

	NB	A1DE	A1DE-S	A1DE-W	A1DE-SW	A2DE	A2DE-S	A2DE-W	A2DE-SW
A1DE	**56/3/12**								
A1DE-S	**56/2/13**	**34/33/4**							
A1DE-W	**56/3/12**	**51/4/16**	**41/4/26**						
A1DE-SW	**59/2/10**	**51/6/14**	**44/5/22**	**25/41/5**					
A2DE	**57/2/12**	**53/3/15**	**47/5/19**	**44/3/24**	**41/4/26**				
A2DE-S	**57/2/12**	**51/3/17**	**50/4/17**	**48/3/20**	**48/3/20**	**31/35/5**			
A2DE-W	**57/2/12**	**54/4/13**	**52/4/15**	**52/5/14**	**49/5/17**	**48/7/16**	36/8/27		
A2DE-SW	**58/2/11**	**54/4/13**	**54/4/13**	**53/4/14**	**52/4/15**	**50/6/15**	**45/7/19**	**32/32/7**	
RF10	**57/1/13**	**54/2/15**	**53/2/16**	**51/3/17**	**51/3/17**	**49/4/18**	**49/3/19**	**49/4/18**	**49/4/18**

- Jointly applying both weighting and subsumption resolution to either A1DE or A2DE increases variance more often than it decrease it relative to applying either alone, but these differences are also not always statistically significant.
- Random Forest has lower bias and higher variance significantly more often than the reverse relative to all AnDE variants.

The average bias and variance results are shown in figure 1. One can see that RF10 has better bias than any member of the AnDE family but worse variance.

3.2 Comparison of the Accuracy - 0-1 Loss and RMSE

The above results show that subsumption resolution and weighting both reduce bias at the cost of an increase in variance. These two techniques have synergistic effect. Used together they further reduce bias at cost of increased variance. If we accept that as data quantity increases, the bias term will increasingly dominate error, we should expect these strategies to be most effective at decreasing error for larger data sets.

Table 3. Win/Draw/Loss of Variance Comparison, all data sets

	NB	A1DE	A1DE-S	A1DE-W	A1DE-SW	A2DE	A2DE-S	A2DE-W	A2DE-SW
A1DE	**23/3/45**								
A1DE-S	**22/2/47**	13/33/25							
A1DE-W	**21/2/48**	**18/6/47**	**17/7/47**						
A1DE-SW	**20/2/49**	**18/6/47**	**17/7/47**	11/43/17					
A2DE	**22/3/46**	28/3/40	**25/3/43**	38/4/29	37/3/31				
A2DE-S	**20/2/49**	**20/2/49**	**22/5/44**	29/3/39	29/2/40	**10/35/26**			
A2DE-W	**20/3/48**	**26/3/42**	26/2/43	30/4/37	30/4/37	22/11/38	36/9/26		
A2DE-SW	**19/3/49**	**23/2/46**	**24/2/45**	26/5/40	28/4/39	**21/7/43**	29/9/33	**9/33/29**	
RF10	**8/1/62**	**8/2/61**	**9/2/60**	**8/4/59**	**9/3/59**	**6/2/63**	**7/2/62**	**6/4/61**	**6/4/61**

The WDL 0-1 Loss and RMSE results are shown in Table 4 and 5 respectively. The significant ($\alpha = 0.05$) results are shown in bold. We summarize the results as:

- Subsumption resolution decreases error more often than not relative to both A1DE and A2DE for both measures of error and for almost all of the different data collections. The exceptions are A1DE, 0-1 loss, medium data and A2DE, 0-1 loss, small data for which there are draws. However, not all these results are statistically significant.
- Subsumption resolution with weighting can decrease error for both measures of error for the first two collections (all and large data sets). As predicted, the effectiveness reduces as data set sizes reduce and for medium data sets, subsumption resolution with weighting can have slightly worst performance relative to weighting in terms of 0-1 loss but better in terms of RMSE. The results, however, are non-significant. The same pattern can be observed in smaller data sets with subsumption resolution and weighting not very effective.
- Subsumption resolution in tandem with weighting can project AnDE to be competitive to RF10, winning significantly on all data sets in terms of the two error measures on all and small data sets. On medium data sets, it results in winning significantly often for A2DE and non-significant often for A1DE over RF10. On large data sets, both A1DE and A2DE lose to RF10. The results are, however, not significant. With five wins and seven losses over RF10, we conjecture, that AnDE with subsumption resolution and weighting, with all desirable properties of learning from big data, is a strong contender for big data learning.

To give an indication of the magnitude of the differences in performance, the average 0-1 Loss results and RMSE results are shown in figures 2 and 3 respectively. It is apparent that A2DE-SW achieves lower average 0-1 loss and RMSE on all of small, medium and large datasets, although this advantage does diminish to being very slight for the largest datasets.

Table 4. Win/Draw/Loss of 0-1 Loss Comparison

All Data Sets

	NB	A1DE	A1DE-S	A1DE-W	A1DE-SW	A2DE	A2DE-S	A2DE-W	A2DE-SW
A1DE	**53/4/14**								
A1DE-S	**51/4/16**	**27/31/13**							
A1DE-W	**50/2/19**	35/8/28	29/8/34						
A1DE-SW	**48/3/20**	38/6/27	32/10/29	20/42/9					
A2DE	**54/3/14**	**50/4/17**	**48/4/19**	**45/8/18**	**41/10/20**				
A2DE-S	**49/3/19**	**46/3/22**	**45/4/22**	**44/5/22**	**43/5/23**	23/34/14			
A2DE-W	**48/2/21**	**46/3/22**	**45/4/22**	**47/6/18**	**46/6/19**	36/8/27	35/9/27		
A2DE-SW	**47/2/22**	**45/2/24**	**42/3/26**	**45/7/19**	**44/6/21**	37/9/25	36/11/24	21/34/16	
RF10	40/1/30	28/2/41	26/5/40	**24/2/45**	**24/2/45**	**22/3/46**	**20/4/47**	**17/3/51**	**17/3/51**

Large Data Sets

	NB	A1DE	A1DE-S	A1DE-W	A1DE-SW	A2DE	A2DE-S	A2DE-W	A2DE-SW
A1DE	**12/0/0**								
A1DE-S	**12/0/0**	7/4/1							
A1DE-W	**12/0/0**	**9/2/1**	7/1/4						
A1DE-SW	**12/0/0**	**10/1/1**	8/2/2	5/6/1					
A2DE	**12/0/0**	**12/0/0**	**12/0/0**	**12/0/0**	**11/0/1**				
A2DE-S	**12/0/0**	**12/0/0**	**12/0/0**	**12/0/0**	**12/0/0**	**7/5/0**			
A2DE-W	**12/0/0**	**12/0/0**	**12/0/0**	**12/0/0**	**12/0/0**	9/1/2	5/1/6		
A2DE-SW	**12/0/0**	**12/0/0**	**12/0/0**	**12/0/0**	**12/0/0**	9/1/2	8/1/3	**6/6/0**	
RF10	**12/0/0**	9/0/3	9/0/3	9/0/3	9/0/3	7/1/4	6/1/5	5/1/6	5/1/6

Medium Data Sets

	NB	A1DE	A1DE-S	A1DE-W	A1DE-SW	A2DE	A2DE-S	A2DE-W	A2DE-SW
A1DE	**18/1/0**								
A1DE-S	**19/0/0**	7/5/7							
A1DE-W	**19/0/0**	13/1/5	10/3/6						
A1DE-SW	**18/1/0**	12/1/6	10/4/5	5/8/6					
A2DE	**19/0/0**	**17/0/2**	**15/1/3**	11/1/7	11/1/7				
A2DE-S	**19/0/0**	**16/0/3**	**14/1/4**	12/1/6	12/1/6	6/9/4			
A2DE-W	**19/0/0**	**17/0/2**	**16/2/1**	**15/2/2**	**14/2/3**	**13/3/3**	**13/3/3**		
A2DE-SW	**19/0/0**	**16/0/3**	**14/1/4**	**14/2/3**	**14/2/3**	11/4/4	11/5/3	5/7/7	
RF10	**15/0/4**	10/0/9	8/3/8	6/1/12	6/1/12	6/1/12	5/2/12	**4/1/14**	**4/1/14**

Small Data Sets

	NB	A1DE	A1DE-S	A1DE-W	A1DE-SW	A2DE	A2DE-S	A2DE-W	A2DE-SW
A1DE	23/3/14								
A1DE-S	20/4/16	13/22/5							
A1DE-W	19/2/19	13/5/22	12/4/24						
A1DE-SW	18/2/20	16/4/20	14/4/22	**10/28/2**					
A2DE	23/3/14	21/4/15	21/3/16	22/7/11	19/9/12				
A2DE-S	18/3/19	18/3/19	19/3/18	20/4/16	19/4/17	10/20/10			
A2DE-W	17/2/21	17/3/20	17/2/21	20/4/16	20/4/16	14/4/22	17/5/18		
A2DE-SW	16/2/22	17/2/21	16/2/22	19/5/16	18/4/18	17/4/19	17/5/18	10/21/9	
RF10	13/1/26	**9/2/29**	**9/2/29**	**9/1/30**	**9/1/30**	**9/1/30**	**9/1/30**	**8/1/31**	**8/1/31**

Table 5. Win/Draw/Loss of RMSE Comparison

All Data Sets

	NB	A1DE	A1DE-S	A1DE-W	A1DE-SW	A2DE	A2DE-S	A2DE-W	A2DE-SW
A1DE	**59/2/10**								
A1DE-S	**59/2/10**	**32/32/7**							
A1DE-W	**58/1/12**	39/5/27	29/5/37						
A1DE-SW	**59/1/11**	**44/4/23**	35/4/32	**24/42/5**					
A2DE	**59/2/10**	**49/3/19**	**41/4/26**	**50/1/20**	**47/1/23**				
A2DE-S	**57/2/12**	**47/1/23**	**45/2/24**	**48/3/20**	**47/3/21**	**28/30/13**			
A2DE-W	**53/2/16**	**44/1/26**	**44/1/26**	**46/4/21**	**45/3/23**	**41/8/22**	28/10/33		
A2DE-SW	**54/1/16**	**44/1/26**	**44/1/26**	**46/3/22**	**46/3/22**	**41/6/24**	35/11/25	**25/34/12**	
RF10	42/0/29	32/0/39	30/0/41	28/2/41	28/1/42	**23/0/48**	**22/1/48**	**19/1/51**	**16/1/54**

Large Data Sets

	NB	A1DE	A1DE-S	A1DE-W	A1DE-SW	A2DE	A2DE-S	A2DE-W	A2DE-SW
A1DE	**12/0/0**								
A1DE-S	**12/0/0**	7/4/1							
A1DE-W	**12/0/0**	8/2/2	6/1/5						
A1DE-SW	**12/0/0**	9/1/2	6/1/5	5/6/1					
A2DE	**12/0/0**	**12/0/0**	**12/0/0**	**12/0/0**	**11/0/1**				
A2DE-S	**12/0/0**	**12/0/0**	**12/0/0**	**12/0/0**	**12/0/0**	7/4/1			
A2DE-W	**12/0/0**	**12/0/0**	**12/0/0**	**12/0/0**	**12/0/0**	9/1/2	4/0/8		
A2DE-SW	**12/0/0**	**12/0/0**	**12/0/0**	**12/0/0**	**12/0/0**	**10/0/2**	8/1/3	7/4/1	
RF10	**12/0/0**	9/0/3	9/0/3	9/0/3	9/0/3	6/0/6	6/0/6	6/0/6	5/0/7

Medium Data Sets

	NB	A1DE	A1DE-S	A1DE-W	A1DE-SW	A2DE	A2DE-S	A2DE-W	A2DE-SW
A1DE	**18/1/0**								
A1DE-S	**18/1/0**	8/7/4							
A1DE-W	**18/1/0**	**15/2/2**	**13/3/3**						
A1DE-SW	**18/1/0**	**14/2/3**	**15/2/2**	7/10/2					
A2DE	**18/1/0**	**15/1/3**	**13/2/4**	10/1/8	10/1/8				
A2DE-S	**17/2/0**	**15/1/3**	**14/2/3**	11/1/7	10/1/8	8/7/4			
A2DE-W	**17/2/0**	**15/1/3**	**16/1/2**	**14/1/4**	13/1/5	**14/4/1**	**12/4/3**		
A2DE-SW	**17/1/1**	**15/1/3**	**15/1/3**	**14/1/4**	**14/1/4**	**12/4/3**	**12/4/3**	6/9/4	
RF10	14/0/5	10/0/9	10/0/9	6/1/12	7/0/12	7/0/12	7/0/12	**3/0/16**	**3/0/16**

Small Data Sets

	NB	A1DE	A1DE-S	A1DE-W	A1DE-SW	A2DE	A2DE-S	A2DE-W	A2DE-SW
A1DE	**29/1/10**								
A1DE-S	**29/1/10**	**17/21/2**							
A1DE-W	**28/0/12**	16/1/23	**10/1/29**						
A1DE-SW	**29/0/11**	21/1/18	14/1/25	**12/26/2**					
A2DE	**29/1/10**	22/2/16	16/2/22	**28/0/12**	26/0/14				
A2DE-S	**28/0/12**	20/0/20	19/0/21	25/2/13	25/2/13	13/19/8			
A2DE-W	24/0/16	17/0/23	16/0/24	20/3/17	20/2/18	18/3/19	12/6/22		
A2DE-SW	25/0/15	17/0/23	17/0/23	20/2/18	20/2/18	19/2/19	15/6/19	12/21/7	
RF10	16/0/24	**13/0/27**	**11/0/29**	13/1/26	**12/1/27**	**10/0/30**	**9/1/30**	**10/1/29**	**8/1/31**

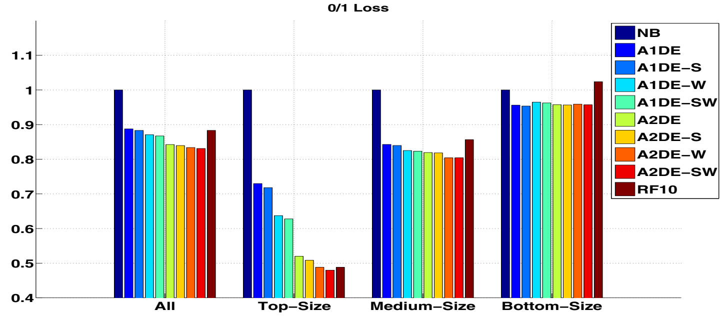

Fig. 2. Average 0-1 Loss results on 4 different collections of data sets normalized with respect to NB. The error-bars are ordered in the same sequence as in the legend.

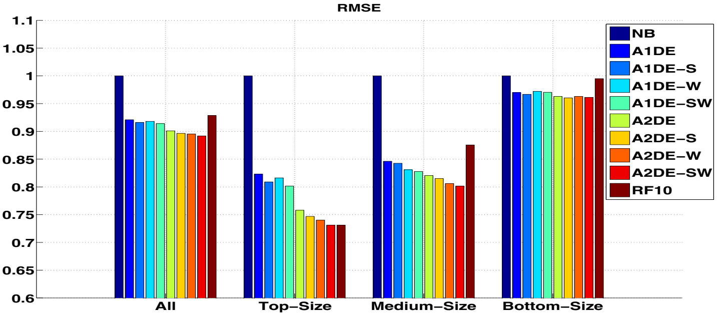

Fig. 3. Average RMSE results on 4 different collections of data sets normalized with respect to NB. The error-bars are ordered in the same sequence as in the legend.

3.3 Analysis of Classification and Learning Time

The average results of classification and learning time for all the compared techniques are shown in figure 4. One can see that subsumption resolution can greatly reduce A2DE's classification time. While A2DE-S and A2DE-SW require only slightly less training time on average than RF10, the training time complexity of AnDE and its variants is linear with respect to data quantity while RF10's is super-linear, as shown by the difference between training times for all data and for large data. The training time advantage would substantially increase if RF10 were applied to data that were too large to maintain in RAM. A2DE and its variants require substantially more classification time than RF10, even with the decreases introduced by subsumption resolution. However, it can be seen that the classification time of RF10 is also super-linear with respect to training set size, whereas AnDE's is not. This is due to the size of the trees increasing as the data quantity increases.

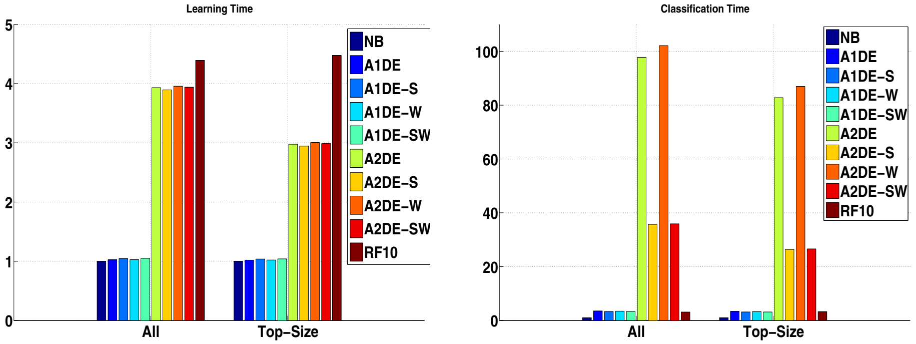

Fig. 4. Averaged Learning and Classification timing results normalized with respect to NB. The error-bars are ordered in the same sequence as in the legend.

3.4 Code

The code of the methods proposed in this work can be obtained from the website, `https://sourceforge.net/projects/averagedndepend/`.

4 Conclusion

AnDE is a strong contender for learning from big data due to its capacity to learn in a single pass through the training data, and consequent training time complexity that is linear with respect to the number of training examples. Weighting using mutual information and subsumption resolution have both previously been demonstrated to be computationally efficient approaches to further reducing the bias of A1DE. As low bias is desirable when learning from large data, it is important to assess the extent to which each of these approaches can reduce the bias of A1DE's lower bias sibling, A2DE. Further, it is important to assess the extent to which these two approaches can augment one another.

The experimental evidence is conclusive. We confirm previous findings that each technique reduces A1DE's bias. We demonstrate that each technique is just as effective at reducing A2DE's bias as it is at reducing A1DE's. We find further that there is strong synergy between the two techniques and that they operate in tandem to reduce the bias of both A1DE and A2DE more effectively than does either in isolation. As is inevitable, these gains in bias come at a cost in increased variance. This bias/variance trade-off can be expected to play out in different error outcomes for different types of data. In particular, for big data, where variance can be expected to be low, low bias can be expected to result in low error [2]. Our experiments demonstrate that this expectation is born out in practice, with both weighting and subsumption resolution reducing error on the largest datasets significantly more often than not relative to standard A2DE and with the two in tandem significantly often further reducing the error relative to MI-weighting alone, and often, but not significantly so, further reducing the error of subsumption resolution alone.

We compared A2DE with MI-weighting and subsumption resolution against the state-of-the-art in-core learning algorithm Random Forest. Random Forest is a lower bias algorithm. However, that bias advantage comes with a considerable variance disadvantage. Even for datasets with 10,000+ training examples Random Forest achieved lower error slightly less often than higher relative to A2DE-SW.

Using only single-pass learning, A2DE with MI-weighting and subsumption resolution achieves accuracy that is very competitive with the state-of-the-art in-core learning, making it a desirable algorithm for learning from very large data.

Acknowledgment. This research has been supported by the Australian Research Council under grant DP110101427 and Asian Office of Aerospace Research and Development, Air Force Office of Scientific Research under contract FA23861214030.

References

1. Webb, G.I., Boughton, J., 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), 233–272 (2012)
2. 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.) PKDD 2002. LNCS (LNAI), vol. 2431, pp. 62–73. Springer, Heidelberg (2002)
3. 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), 93–125 (2012)
4. Jiang, L., Zhang, H.: Weightily averaged one-dependence estimators. In: Yang, Q., Webb, G. (eds.) PRICAI 2006. LNCS (LNAI), vol. 4099, pp. 970–974. Springer, Heidelberg (2006)
5. Zheng, F., Webb, G.I.: Efficient lazy elimination for averaged one-dependence estimators. In: Proceedings of the Twenty-Third International Conference on Machine Learning (ICML 2006), pp. 1113–1120 (2006)
6. 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 2005. LNCS (LNAI), vol. 3720, pp. 72–83. Springer, Heidelberg (2005)
7. Yang, Y., Webb, G., Cerquides, J., Korb, K., Boughton, J., Ting, K.: 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 (2007)
8. Cestnik, B.: Estimating probabilities: A crucial task in machine learning. In: Proceedings of the Ninth European Conference on Artificial Intelligence (ECAI 1990). Pitman, London (1990)
9. Kohavi, R., Wolpert, D.: Bias plus variance decomposition for zero-one loss functions. In: Proceedings of the Thirteenth International Conference on Machine Learning, pp. 275–283. Morgan Kaufmann, San Francisco (1996)
10. Webb, G.I.: Multiboosting: A technique for combining boosting and wagging. Machine Learning 40(2), 159–196 (2000)
11. Fayyad, U.M., Irani, K.B.: On the handling of continuous-valued attributes in decision tree generation. Machine Learning 8(1), 87–102 (1992)