REPORT DOCUMENTATION PAGE

Form Approved

OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
19-03-2015	Final	12-03-2012 – 11-03-2015

4. TITLE AND SUBTITLE

Dynamic Dimensionality Selection for Bayesian Classifier Ensembles

5a. CONTRACT NUMBER
FA2386-12-1-4030

5b. GRANT NUMBER
61102F

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Geoffrey Webb and Mark Carman

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Faculty of Information Technology, Monash Uiversity
PO Box 63, Monash University, Victoria 3800
Australia

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

AOARD
UNIT 45002
APO AP 96338-5002

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/AFOSR/IOA(AOARD)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AOARD-124030

12. DISTRIBUTION/AVAILABILITY STATEMENT

Distribution A: Approved for public release. Distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT This project seeked to develop new learning algorithms specifically tailored to be efficent and effective in learning from big data. It exploited the capacity of generative learning to efficently extract useful summary statistics and used discriminative learning to meld them into a highly accurate classifier. Two classes of learning algorithm were developed. The first uses discriminative learning to select a generative model (selective ANDE and selective KDB). Very effective feature selection was achieved with a single pass through the training data for each attribute that is finally selected. The second combines generatively and discriminatively learned parameters (WANBIA, WANBIA-C,WANJE). It uses discriminative learning of weights in an otherwise generatively learned naive Bayes classifier. WANBIA-C is very cometitive to Logistic Regression but much more efficient in learning the model. WNANJE can model higher-order attribute interdependencies.

15. SUBJECT TERMS
Big data, Low bias classifier, Generative learning, Discriminative learning, Naïve Bayes, Feature selection, Logistic regression, higher order attribute independence

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Hiroshi Motoda, Ph. D.
U	U	U	SAR	23	**19b. TELEPHONE NUMBER** *(Include area code)* +81-42-511-2011

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
19 MAR 2015	**Final**	**12-03-2012 to 11-03-2015**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Dynamic Dimensionality Selection for Bayesian Classifier Ensembles	**FA2386-12-1-4030**
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
Geoffrey Webb; Mark Carman	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
Faculty of Information Technology, Monash Uiversity,PO Box 63,, Monash University, Victoria 3800,Australia,AU,3800	**N/A**
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
AOARD, UNIT 45002, APO, AP, 96338-5002	**AFRL/AFOSR/IOA(AOARD)**
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)
	AOARD-124030

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This project seeked to develop new learning algorithms specifically tailored to be efficent and effective in learning from big data. It exploited the capacity of generative learning to efficently extract useful summary statistics and used discriminative learning to meld them into a highly accurate classifier. Two classes of learning algorithm were developed. The first uses discriminative learning to select a generative model (selective ANDE and selective KDB). Very effective feature selection was achieved with a single pass through the training data for each attribute that is finally selected. The second combines generatively and discriminatively learned parameters (WANBIA, WANBIA-C,WANJE). It uses discriminative learning of weights in an otherwise generatively learned naive Bayes classifier. WANBIA-C is very cometitive to Logistic Regression but

15. SUBJECT TERMS

Big data, Low bias classifier, Generative learning, Discriminative learning, Na??ve Bayes, Feature selection, Logistic regression, higher order attribute independence

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**23**	

Final Report for AOARD Grant AOARD-124030 "Dynamic dimensionality selection for Bayesian classifier ensembles"

March 19, 2015

Name of Principal Investigators (PI and Co-PIs): Geoffrey Webb and Mark Carman
E-mail address : geoff.webb@monash.edu
Institution : Monish University
Mailing Address :
Faculty of Information Technology,
P.O. Box 63
Monash University,
Victoria 3800, Australia
Phone : +61 3 990 53296
Fax : +61 3 990 55159

Period of Performance: 4/1/2012 – 3/11/2015

Abstract:

This project seeks to develop new learning algorithms specifically tailored to be efficient and effective in learning from big data. It aims to achieve this by combining generative and discriminative learning, exploiting the capacity of generative learning to efficiently extract useful summary statistics from large data and using discriminative learning to meld these statistics into a highly accurate classifier.

We have developed two classes of learning algorithm that utilize this overarching strategy. The first uses discriminative learning to select a generative model. In this case the generative model is either Averaged n Dependence Estimators (ANDE) or K-Dependence Bayes (KDB) and the discriminative technique is feature selection. We have shown that very effective feature selection can be achieved with a single pass through the training data for each attribute that is finally selected. We have demonstrated two benefits of feature selection. First, feature selection reduces the bias of the classifier, which typically decreases error for larger data. Second, feature selection decreases the memory requirements of the classifier, allowing higher values of n to be used, further decreasing bias and further reducing error on large data.

The second class of learning algorithm combines generatively and discriminatively learned parameters. Weighting attributes to Alleviate Naive Bayes' Independence Assumption (WANBIA) uses discriminative learning of weights in an otherwise generatively learned naïve Bayes classifier. WANBIA-C modifies the WANBIA model in such a way as to enable models that are equivalent to Logistic Regression models to be learned. The resulting learner creates classifiers that are of equivalent accuracy to those learned by Logistic Regression; sometimes the classifiers are slightly better and sometimes slightly

worse, but neither algorithm having a systematic advantage. However, the learning process for WANBIA-C is substantially faster than that for Logistic Regression, sometimes orders of magnitude faster. We have demonstrated that these techniques generalize to models that directly model higher-order attribute interdependencies and developed a proof-of-concept system WANJE.

We have also investigated the purely generative techniques of Subsumption Resolution and Submodel Weighting in ANDE and demonstrated that they are effective with higher orders of n and that they are complementary to one another. We intend these techniques to be incorporated into future combined generative/discriminative approaches.

In the process of creating algorithms for learning from high-throughput data streams, we have developed effective algorithms for discretization of streaming numeric data.

1 Introduction

Machine learning algorithms can be categorized into two families, generative and discriminative learners [12]. Generative learners seek to model the joint distribution between the independent X variables and the dependent Y variable. They classify by calculating the posterior probability $P(Y \mid X)$ from the joint probability. In contrast, discriminative learners seek to directly model the posterior probability, without seeking to model the probabilities of the independent X variables.

The ANDE learners that we have developed [19] are generative learners that are parameterized using maximum likelihood estimates. Maximum likelihood parameterization is extremely efficient and can be performed in a single out-of-core pass through the training data. The resulting models have been shown to be highly accurate for moderate quantities of data. However, discriminative learning is better able to model the desired posterior distribution when sufficient training data are available.

The current project seeks to combine

- the manner in which the efficient learning of valuable information in a generative manner with maximum likelihood estimation; and
- the capacity of discriminative learning to more accurately model the posterior distribution when there are sufficient data to avoid overfitting.

We are pursuing two main strategies. The first uses discriminative learning to select between alternative generative models on the basis of discriminative performance. The second uses maximum likelihood parameterization to augment discriminative learning of discriminative models.

A key theoretical insight that drives our research is derived from the bias/variance trade-off [12]. Most learning algorithms represent a trade-off between *variance* that results from overfitting the data, and *bias* that results from failure to appropriately fit the data. Low variance algorithms achieve relatively low error on small quantities of data, while low bias algorithms achieve relatively low error on larger quantities of data. This is illustrated in Figure 3.

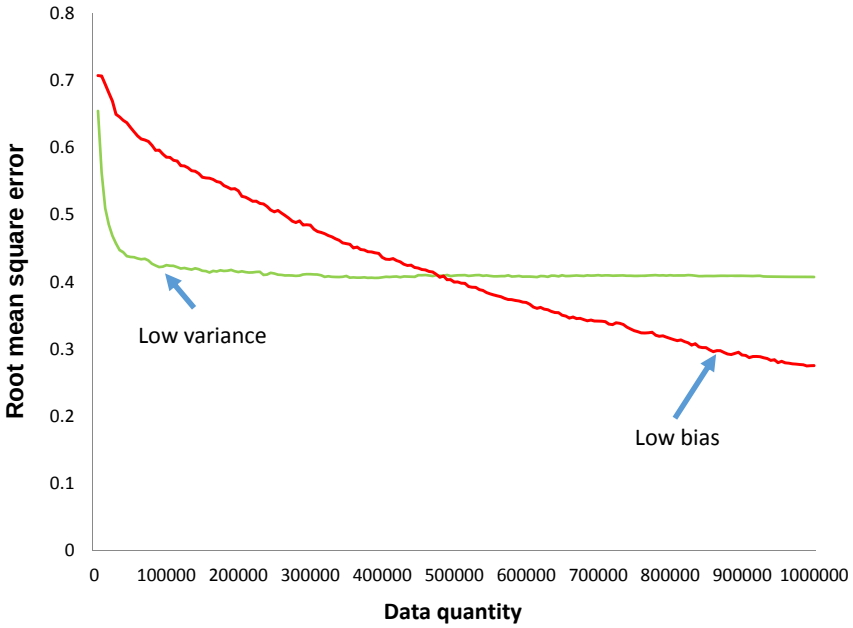


Figure 1: Learning algorithms that closely fit closely fit complex multivariate distributions overfit small data, but achieve lower error on large data

2 Scalable discriminative extensions to ANDE

In this line of research we seek to improve the scalability of the Averaged N-Dependence Estimators (ANDE) family of algorithms. These algorithms are attractive for learning form large data because their training time is linear with respect to data quantity and they can learn out-of-core in a single pass through the data. Hence, they can learn from data that are too large to fit into RAM. Further, their models can be learned incrementally, and hence they are suitable for streaming data. However, for large data quantity it is desirable to have low bias [1, 2], and hence it is desirable to have ANDE with higher values of n. This is problematic because ANDE does not scale well to high n.

2.1 Background

Averaged N-Dependence Estimators (ANDE) classify using the model

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

The parameters to the model ($\hat{\mathrm{P}}(y, x_s)$ and $\hat{\mathrm{P}}(\mathbf{x} \mid y, x_s)$) are learned by maximum likelihood estimation in a single pass through the data that can be implemented out-of-core.

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 $\mathrm{P}(e)$
$\hat{\mathrm{P}}_{\mathrm{AODE}}(e)$	an AODE estimate of $\mathrm{P}(e)$
$\hat{\mathrm{P}}_{\mathrm{AnDE}}(e)$	an AnDE estimate of $\mathrm{P}(e)$
$\hat{\mathrm{P}}_{\mathrm{WAnDE}}(e)$	a weighted AnDE estimate of $\mathrm{P}(e)$
$\hat{\mathrm{P}}_{\mathrm{AnDE}^{\mathrm{SR}}}(e)$	an AnDE with Subsumption Resolution estimate of $\mathrm{P}(e)$
$\hat{\mathrm{P}}^{N}_{\mathrm{WANJE}}(e)$	a WANJE estimate of $\mathrm{P}(e)$
a	the number of attributes
c_i	the i^{th} 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 i^{th} 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
π_i	the parents of attribute x_i in a Bayesian Network Classifier
π_Y	the parents of the class attribute Y in a Bayesian Network Classifier

Averaged One-Dependence Estimators (AODE) refers to ANDE with $n = 1$ [18].

ANDE models are extremely efficient to learn with low n. Increasing n reduces bias, which supports more accurate learning from larger data. However, its training time complexity is $\mathrm{O}\big(t\binom{a}{n+1}\big)$, 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 $\mathrm{O}\big(ka\binom{a}{n}\big)$.

2.2 Weighting and Subsumption Resolution

The first study [26] took two techniques, *subsumption resolution* [28] and *mutual information weighting* [9] that have been developed for AODE (ANDE with $n = 1$), assessed their interaction, and assessed whether they are effective with higher values of n.

Subsumption resolution 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 $\mathrm{P}(y \mid \mathbf{x}) = \mathrm{P}(y \mid x_{\{1\ldots i-1,i+1\ldots m\}})$ 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 $\#(x_i) = \#(x_j)$.

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

Subsumption resolution has been shown to be effective at reducing the bias of A1DE [29, 28].

Another approach to reducing bias in AnDE that has been shown to be effective for A1DE [4, 9, 22] is to weight the sub-models, modifying (1) to

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

WAODE [9] weights A1DE, where s is a single attribute value. It sets w_s to the mutual information [13] 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 $w_s = \mathrm{MI}(S, Y)$,

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

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

2.3 Study

We implemented Average 1 and 2 Dependence Estimators with both subsumption resolution and mutual information weighting in the Weka machine learning workbench.

We compared the two levels of ANDE without either extension, with each in isolation and with both in combination. We also compared the important comparators naive Bayes (NB) and Random Forest with 10 trees (RF10).

These algorithms were all applied to 71 benchmark datasets.

2.4 Results

We found that

- 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.
- 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.
- 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 both RMSE and 0-1 loss 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.

The average results of classification and learning time for all the compared techniques are shown in figure 2. One can see that subsumption resolution can greatly reduce

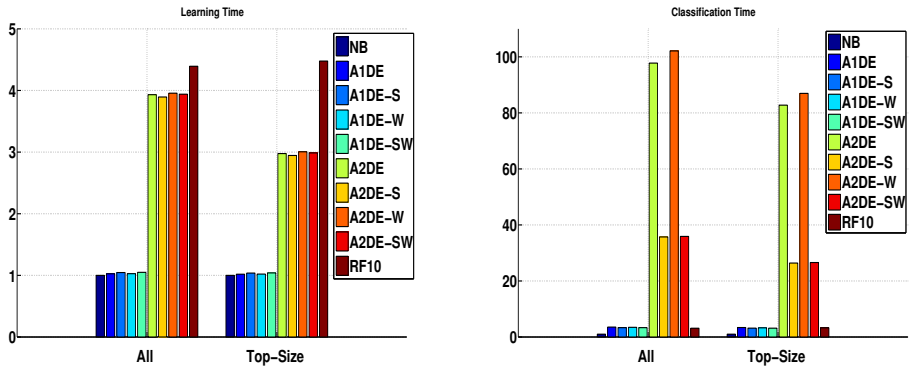


Figure 2: Averaged Learning and Classification timing results normalized with respect to NB.

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.

More detail about this study can be found in our PAKDD paper [26].

2.5 Fast and effective attribute selection

The second study [5] investigated a two-pass approach to attribute selection for AODE. Previous research [22, 23, 30] has shown that attribute selection can be very effective at reducing the bias of AODE. However, these techniques have used computationally expensive wrapper techniques for attribute selection that are infeasible to implement out-of-core. To make attribute selection feasible out-of-core it is necessary to add only a small number of passes through the data. The approach we investigate here is based on the observation that it is possible to nest a large space of alternative models such that each is a trivial extension of another. Let p and c be the set of indices of parent and child attributes, respectively. For every attribute 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.

The mutual information [13] between an attribute and the class measures how informative this attribute is about the class, and thus it is a suitable metric to rank the attributes. It can be computed using the information already gathered by AODE in its first pass through the training data.

We extend AODE by adding a second pass through the training data. The first pass constructs the full AODE model. The attributes are then ranked in descending order of mutual information with the class. This is used to create a two-dimensional space of nested models. In one dimension we have n increasing subsets of the attributes used in the parent role. In the other dimension we have n increasing subsets of the attributes used in the child role. Using incremental cross-validation [10] it is possible to efficiently perform cross-validation on an AODE model in a single pass through the data. As the models being assessed are nested, they can be evaluated in this manner in a single pass using very little more computation than is required to evaluate the full model in isolation.

In this manner we perform discriminative evaluation of all $n \times n$ nested models in a single computationally efficient pass. Any measure of classification performance could potentially be employed. In this study we used root mean squared error (RMSE), which is an effective measure of the calibration of the probability estimates produced by a model.

In an extensive study we compare the proposed attribute selective AODE (ASAODE) with AODE, mutual information weighted AODE (WAODE), AODE with subsumption resolution (AODESR), BSE selective AODE (BSEAODE) and A2DE on the same 71 benchmark datasets used in the first study.

2.6 Results

The empirical results show that the new algorithm is significantly more accurate than AODE, WAODE and AODESR.

It has comparable error to BSEAODE, while requiring substantially less computation and being executed out-of-core.

As we expected, it is not as accurate on average as A2DE. However, in continuing research we are implementing the approach with A2DE and A3DE, with extremely promising initial results.

Our new out-of-core attribute selection algorithm requires significantly less training time than BSEAODE, and less classification time than AODE and all other variants, especially than A2DE.

3 Scalable discriminative extensions to Bayesian Network Classifiers

Standard Bayesian Network Classifiers can also be improved using our approach of single-pass discriminative selection between a large class of nested models. Like the work with ANDE, our motivation is to use efficient out-of-core discriminative learning to improve

models learned using efficient out-of-core generative learning using maximum likelihood parameterization.

Bayesian Network Classifiers are defined by parent relation π and Conditional Probability Tables (CPTs). π encodes conditional independence relationships between the attributes. CPTs encode the specific conditional probabilities.

A Bayesian Network Classifier classifies using

$$\mathrm{P}(y \mid \mathbf{x}) \propto \mathrm{P}(y \mid \pi_Y) \prod \mathrm{P}(x_i \mid \pi_i) \tag{3}$$

The class variable Y is usually set to be a parent of all attributes X_i.

Given π, CPTs can be learned by counting joint frequencies in a single single pass through the data.

K-Dependence Bayes (KDB) [15] learns a restricted Bayesian Network Classifier in two passes through the learning data. In the first [ass it learns the structure. In the second pass it learns the parameters to the CPTs.

To learn the structure it first collect counts for all pairs of attributes with the class. It then orders the attributes based on mutual information with the class. Each attribute is then assigned parents such that

- each has no more than k parents;
- an attributes parents must be earlier in the order than the child;
- within these constraints the parents are selected that maximize mutual information between the parent and child conditioned on the class.

The parameter k controls the bias variance trade-off. Low k has low variance but high bias, and is well suited to small quantities of data, while higher k has higher variance but lower bias and is better suited to larger quantities of data. This is illustrated in Figure 3. Further, spurious attributes may increase error.

KDB models are naturally nested. A model with $k = v$ subsumes a model with $k = v - 1$. A model using the first w attributes as children subsumes a model using the first $w - 1$ attributes as children. Thus it can be seen that a KDB model with $k = v$ and w attributes subsumes $v \times w$ submodels. All of these submodels can be assessed in a single out-of-core pass through the training data using incremental cross-validation. Due to the nested nature of the models, this process takes little more computation than incremental cross validation of the full model.

Our algorithm, Selective KDB performs such evaluation in a third pass through the data, resulting in a very efficient three pass out-of-core algorithm.

Let a be the number of attributes; v be the average number of values; y be the number of classes; a^* be the number of attributes selected ($a^* \leq a$); and k^* be the best value of k found ($k^* \leq k_{max}$). Selective KDB's training space complexity is $\mathrm{O}\big(ya^2v^2 + yav^{k+1}\big)$. Its classification space complexity is $\mathrm{O}\big(ya^*v^{k^*+1}\big)$. Its training time complexity is $\mathrm{O}\big(ta^2 + ya^2v^2 + tayk\big)$. Its classification time complexity is $\mathrm{O}\big(ya^*k^*\big)$.

On large datasets our out-of-core algorithm has very competitive error to state-of-the-art in-core algorithm Random Forests [3] (RF), to state-of-the-art out-of-core algorithm

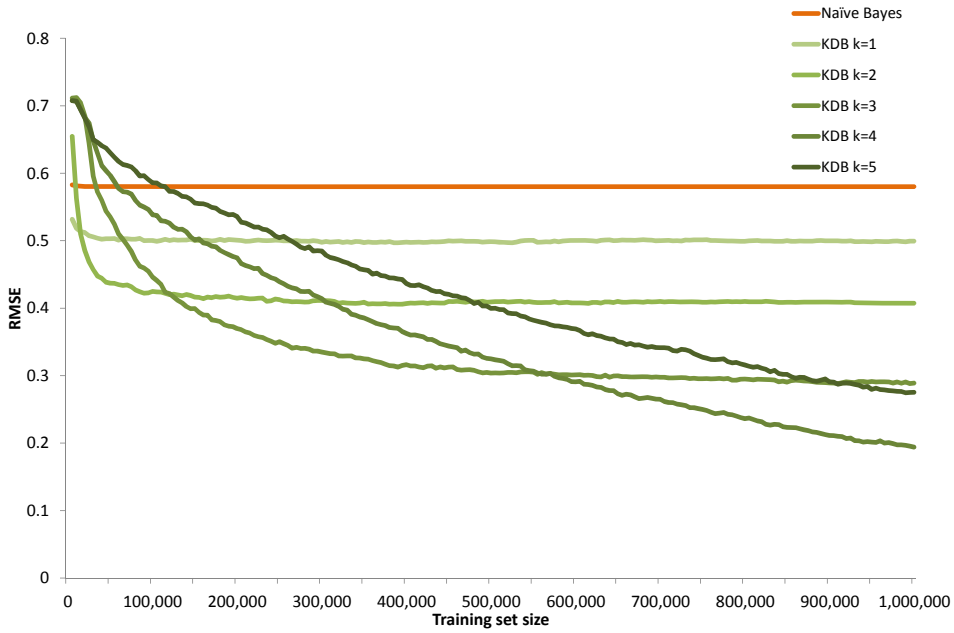


Figure 3: Learning curves for KDB classifiers with varying k on the benchmark poker-hand dataset. Lower values of k (naive Bayes is KDB with $k = 0$) have low variance and have low error on small data, but their high bias means their error asymptotes quickly, while the lower bias of KDB with higher k leads to lower error on large quantities of data.

Vowpal Wabbit (VWLF) [11]. As illustrated in Figure 4, it has lower error significantly more often than not relative to state-of-the-art Bayesian Network Classifiers, including computationally expensive in-core approaches (BayesNet) and out-of-core naive Bayes (NB), AODE and TAN. As illustrated in Figures 5 and 6, Selective KDB is an order of magnitude more efficient in training than in-core Random Forests, and two orders of magnitude more efficient than conventional Bayesian Network Classifiers. It is also substantially more efficient than state-of-the-art out-of-core logistic regression algorithm Vowpal Wabbit under either of its two main objective functions, squared loss (VWSF) or logistic loss (VWLF).

A paper presenting these results in greater detail has been submitted for journal publication.

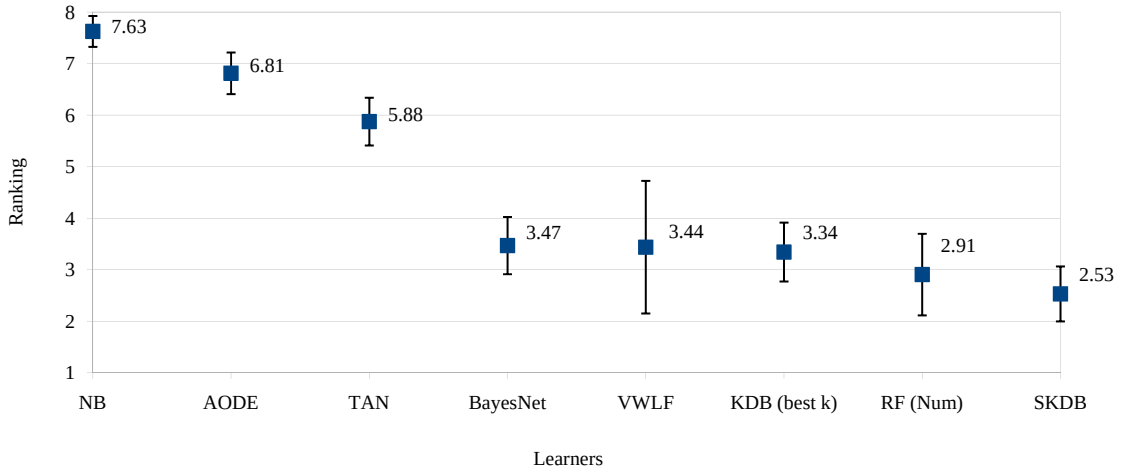


Figure 4: The RMSE rank of alternative algorithms over 16 large datasets. Out–of-core Selective KDB achieves lower error more often than any alternative, including state-of-the-art in-core Random Forests, and significantly more often than any of the Bayesian Network Classifiers.

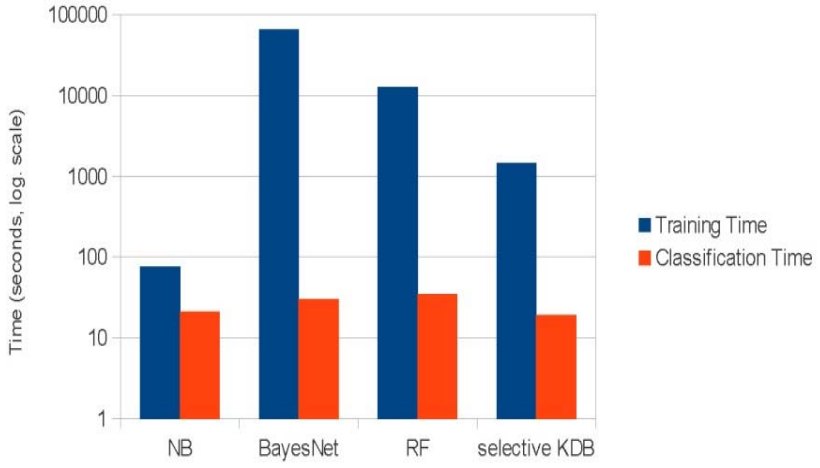
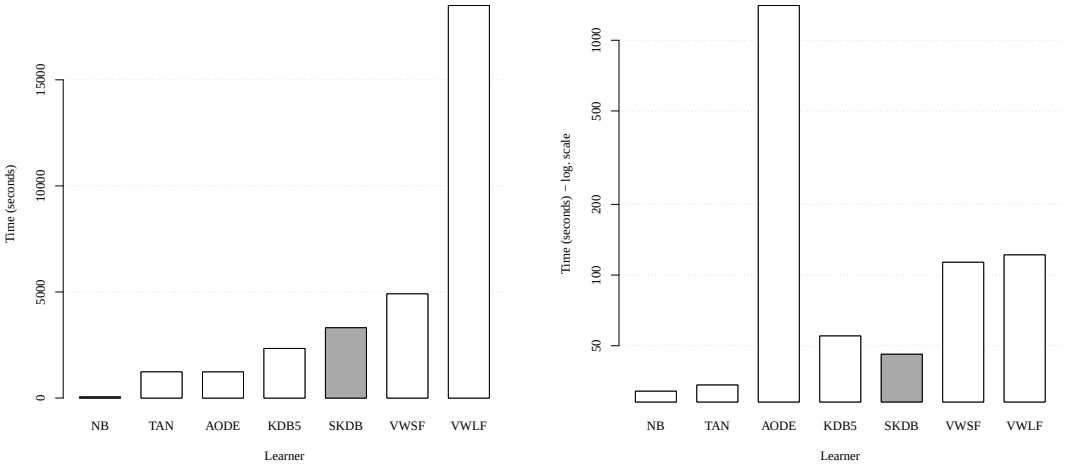


Figure 5: The relative training times of in-core algorithms BayesNet and Random Forests compared with out-of-core naive Bayes and Selective KDB



(a) Training Times

(b) Classification Times

Figure 6: Training and classification time comparisons for the out-of-core classifiers.

4 WANBIA

A number of researchers have previously investigated adding weights to naive Bayes models [8, 6, 27, 7, 21]. Naive Bayes classifies using

$$\hat{\mathrm{P}}_{\mathrm{NB}}(y, \mathbf{x}) \propto \hat{\mathrm{P}}(y) \prod_{i=1}^{n} \hat{\mathrm{P}}(x_i \mid y). \tag{4}$$

There are three main variants of weighted naive Bayes. In the most general case, this weight depends on the attribute value and class:

$$\hat{\mathrm{P}}(y, \mathbf{x}) \propto \hat{\mathrm{P}}(y)^{w_y} \prod_{i=1}^{n} \hat{\mathrm{P}}(x_i \mid y)^{w_{i,x_i,y}}. \tag{5}$$

Doing this results in $\sum_i^a |\mathcal{X}_i|$ weight parameters. A second possibility is to give a single weight per attribute:

$$\hat{\mathrm{P}}(y, \mathbf{x}) \propto \hat{\mathrm{P}}(y)^{w_y} \prod_{i=1}^{a} \hat{\mathrm{P}}(x_i \mid y)^{w_i}. \tag{6}$$

One final possibility is to set all weights to a single value:

$$\hat{\mathrm{P}}(y, \mathbf{x}) = \hat{\mathrm{P}}(y)^{w_y} \left(\prod_{i=1}^{a} \hat{\mathrm{P}}(x_i \mid y) \right)^{w}. \tag{7}$$

Equation 6 is a special case of Equation 5, where $\forall_{i,j,y}\ w_{i,j,y} = w_i$, and Equation 7 is a special case of Equation 6 where $\forall_i\ w_i = w$.

Most previous research into weighting naive Bayes has used the weights to give greater emphasis to more informative attributes. We argue that this is suboptimal, as Bayesian approaches automatically take account of the amount of information about a class that each attribute contributes. Indeed, naive Bayes is a Bayes optimal classifier except insofar as the base probability estimates are incorrect and that the assumption that the attributes are conditionally independent is violated. We argue that it is in this latter respect that weighting has the most to offer and propose to address it through discriminative learning.

Our initial algorithm WANBIA learns weights of the form in Equation 6. A second algorithm, WANBIA-C learns weights of the form in Equation 5.

The approach is inspired by MAPLMG [4], a system that learns weights on the sub-models in AODE by maximizing Conditional Log-Likelihood.

To this end we have derived the gradient of Equations 5 and 6 with respect to Conditional Log-Likelihood. This allows the use of a regular gradient descent optimization procedure.

WANBIA-C learns a model that maps directly onto the standard logistic regression model. By optimizing for the same objective function as used by logistic regression we ensure that the models learned will be equivalent, except insofar as the optimization process fails to find the true optima.

We show that WANBIA substantially reduces the bias of naive Bayes at the cost of a modest increase in variance. For any but extremely small datasets it tends to produce lower error models than naive Bayes. The WANBIA models are more biased than logistic regression, but with fewer free parameters have lower variance. As a result they have very competitive error for small to medium sized datasets. With fewer free parameters the optimization process is also considerably more efficient.

As shown in Figure 7, when using unconstrained optimization WANBIA-C has similar RMSE and 0-1 loss to logistic regression but the models converge much faster due to the naive Bayes probability estimates acting as an effective preconditioner.

Of particular significance for stream learning, as shown in Figure 8, WANBIA-C substantially outperforms logistic regression when optimized using stochastic gradient descent.

Full details can be found in our Journal of Machine Learning Research and IEEE International Conference on Data Mining papers [25, 24].

5 WANJE

One limitation of WANBIA and WANBIA-C is that they use linear models that cannot directly capture high-order multivariate interactions in data.

Our approach of using maximum likelihood parameterization to speed up discriminative parameter optimization using conditional log-likelihood can also be generalized to other forms of Bayesian Network Classifier that directly model higher-order interactions. However, to make the optimization feasible it is important that it form a convex optimization problem, and only moral Bayesian networks do so [14].

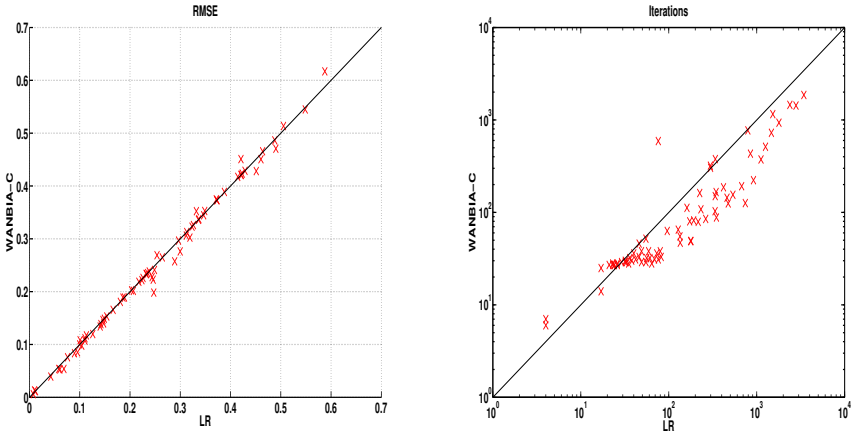


Figure 7: Comparison of RMSE (Left) and number of iterations (Right) of WANBIA-C and LR on 73 datasets using quasi-Newton optimization. LR parameters are initialized to NB MAP estimates. Number of iterations are on log-scale.

The ANDE classifiers form moral networks. However, each combination of $n+1$ attributes and class require

$$k\binom{n+1}{n}\bar{v} \tag{8}$$

parameters to model, where k is the number of classes, n is the number of parents per sub-model (the n from ANDE), and $\bar{v}$ is the average number of values per attribute. This is because for every combination of n parents and class we need to model the CPT for the child attribute.

This results in a very large number of parameters and a very difficult optimization task for high values of n.

To address this issue we have developed a variant of ANDE that uses a single parameter for each combination of attribute and class values.

It is possible to partition the attributes X into any set of sets of attributes $\mathcal{P}$ such that $\bigcup_{\alpha\in\mathcal{P}}\alpha = X$ and $\forall\alpha\in\mathcal{P},\beta\in\mathcal{P},\alpha\cap\beta=\emptyset$ and then by assuming independence only between the different sets of attributes one obtains

$$\mathrm{P}(\mathbf{x}\mid y) = \prod_{\alpha\in\mathcal{P}}\mathrm{P}(\alpha\mid y). \tag{9}$$

For example, if there are four attributes x_1, x_2 , x_3 and x_4 that are partitioned into the sets $\{x_1,x_2\}$ and $\{x_3,x_4\}$ then by assuming conditional independence between the sets we obtain

$$\mathrm{P}(x_1,x_2,x_3,x_4\mid y) = \mathrm{P}(x_1,x_2\mid y)\mathrm{P}(x_3,x_4\mid y). \tag{10}$$

The ANJE model is equivalent to the geometric mean of all such models in which all subsets are of size N.

$$\hat{\mathrm{P}}(\mathbf{x}\mid y) = \prod_{\alpha\in\binom{X}{N}}\mathrm{P}(\alpha\mid y)^{(N-1)!(A-N)!/(A-1)!} \tag{11}$$

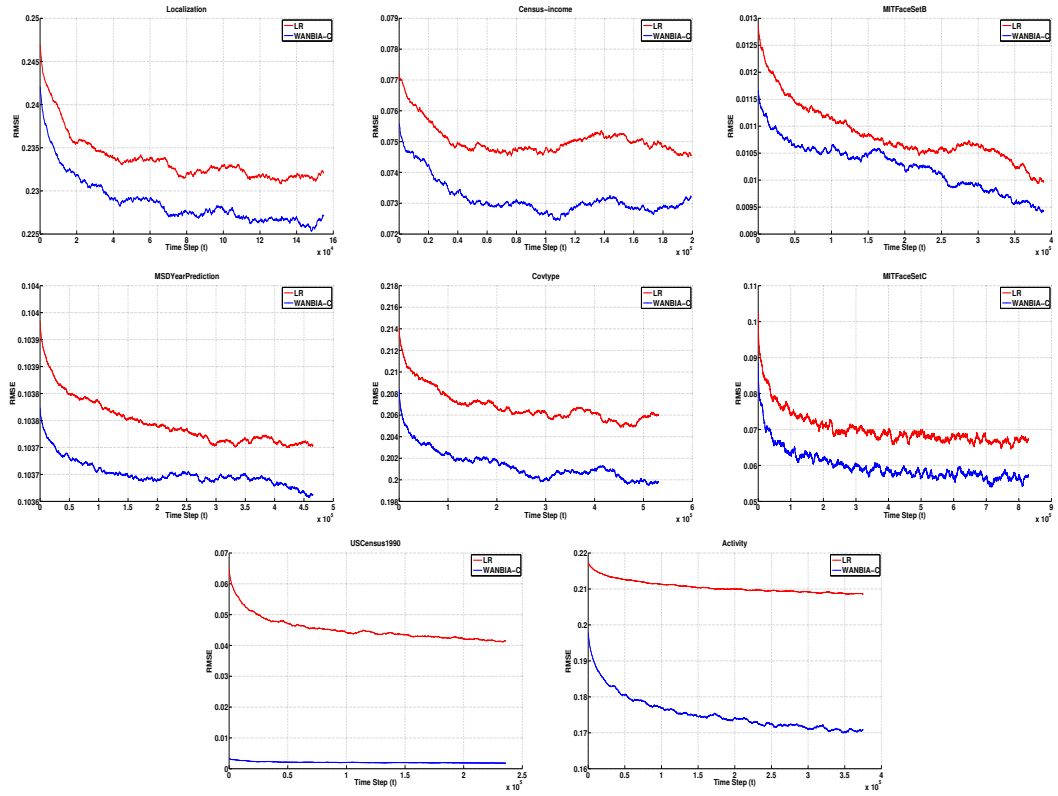


Figure 8: RMSE comparison of LR and WANBIA-C with SGD on `Localization`, `Census-income`, `MITFaceSetB`, `MSDYearPrediction`, `Covtype`, `MITFaceSetC`, `USCensus1990`, and `Activity` datasets.

where $A = |X|$ and $\binom{X}{N}$ is the set of sets of attributes of size N, $\{\alpha \mid \alpha \subset X, |\alpha| = N\}$.

Equation 11 is derived as follows. Let $S = |\mathcal{P}|$, the number of sets of attributes in each model of the form (9). Let T be the total number of these sets, $T = |\binom{X}{N}| = \binom{A}{N}$. Let M be the total number of models of the form (9) of which (11) represents the geometric mean. This differs depending on whether one allows different permutations or only combinations of the different sets of attributes. It turns out that this value cancels out, so we do not need to specify what it is. Let X be the number of models in which each set of attributes appears. As this is identical for all sets of attributes, and as the total number of terms in all models is MS, this equals MS/T. Let $x_{i,j}$ be the j^{th} set of attributes in the i^{th} model. The geometric mean of all models is

$$\mathrm{P}(\mathbf{x} \mid y) = \sqrt[M]{\prod_{i=1}^{M}\prod_{j=1}^{S}\mathrm{P}(x_{i,j} \mid y)} \tag{12}$$

$$= \prod_{i=1}^{M}\prod_{j=1}^{S}\mathrm{P}(x_{i,j} \mid y)^{1/M} \tag{13}$$

$$= \prod_{\alpha \in \binom{X}{N}} \mathrm{P}(\alpha \mid y)^{X/M}. \tag{14}$$

The final exponent X/M can be expanded as follows

$$X/M = MS/(TM) \tag{15}$$

$$= S/T \tag{16}$$

$$= A/\left[N\binom{A}{N}\right] \tag{17}$$

$$= A/(NA!/\left[N!(A-N)!\right]) \tag{18}$$

$$= (N-1)!(A-N)!/(A-1)!. \tag{19}$$

Substituting (19) for the exponent in (14) we get (11). We can then classify using

$$\hat{\mathrm{P}}(y \mid \mathbf{x}) \propto \hat{\mathrm{P}}(y) \prod_{\alpha \in \binom{X}{N}} \mathrm{P}(\alpha \mid y)^{(N-1)!(A-N)!/(A-1)!} \tag{20}$$

An ANJE model with a given N models the same order of interactions between attributes as an ANJE model with $n = N - 1$, because n refers to the number of parent attributes for each child in an ANDE model while N refers to a clique of attributes in an ANJE model.

Our experiments indicate that ANJE has substantially higher bias than ANDE where $n = N - 1$, and that the resulting reduction in variance is rarely being sufficient to lead to lower error. However, the lower number of parameters required by the ANJE models can make it feasible to model higher order interactions than can be modeled by ANDE. It is our belief that usually for large data modeling higher order interaction will be more effective at reducing bias than more detailed modeling of the lower order interactions.

ANJE serves as an effective basis for discriminative weighting of the form used in WANBIA and WANBIA-C.

$$\hat{\mathrm{P}}_{\mathrm{WANJE}^N}(y \mid \mathbf{x}) \propto \hat{\mathrm{P}}(y)^{w_y} \prod_{\alpha \in \binom{X}{N}} \hat{\mathrm{P}}(\alpha \mid y)^{w_{\alpha,y}}. \tag{21}$$

Our experiments show that the resulting models are very competitive with state-of-the-art learning algorithm Random Forest. Of particular significance, we have developed an out-of-core single pass variant, where the maximum likelihood estimates are

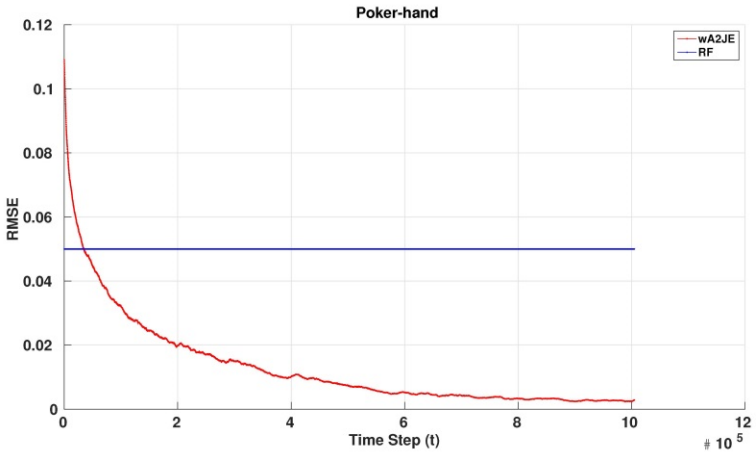


Figure 9: Learning curves for single-pass out-of-core incremental WANJE and multiple-pass in-core batch Random Forests on the Poker-Hand dataset.

updated incrementally and the weights are learned using stochastic gradient descent. As illustrated in Figure 9, our experiments show that for large datasets this single pass out-of-core incremental learner is very competitive with state-of-the-art in-core batch learner Random Forests.

We are in the process of writing a research paper on this work.

6 Incremental discretization

The ANDE, WANBIA and WANJE algorithms require that numeric data be discretised. While investigating the potential to extend our approaches to streaming data we were confronted with the problem of discretizing such data. This is on the face-of-it problematic, because streaming data require cutpoints that vary over time, while many algorithms require the meaning of a cutpoint to remain invariant. In this line of research we investigated this issue and identified that there is a manner to allow cutpoints to drift while maintaining invariant meaning.

This can be achieved by performing quantile-based discretization, where each cutpoint is set at a specific quantile in the current data distribution. This way the meaning of the cutpoint is invariant even while the value changes. For example, the cutpoint for an interval representing high income should grow over time as inflation increases incomes. If it is set at the upper 20 percentile, then it will do so.

We developed an algorithm, IDA, for efficient quantile-based discretization of streaming data. It operates by maintaining a sample of the data encountered in the stream because 1) it is not feasible for high-throughput streams to maintain a complete record of all values observed to date; 2) it is computationally efficient; and 3) it is possible to

place tight bounds on the expected variance of the cut points .

We used the reservoir sampling algorithm [17] to maintain a random sample for each attribute.

We developed an efficient data structure and algorithms for storing and updating the sample using a vector of interval heaps [16]. This data structure ensures that insertion and deletion are of order $O(\log s)$, where s is the sample size, and retrieving a cut point is constant time.

We also developed a variant of IDA, IDAW, that tracks the current distribution by using a window of recent examples in place of a random sample of all examples. This is computationally more intensive, as it requires an update at every time step.

The computational complexity of IDA is dominated by the costs of maintaining the samples and determining the quantiles from those samples. The required operations are to insert a new value (only required while the sample is not yet at full size), to replace a random value with a new value, and to return the required quantiles.

As each bin is maintained as an interval heap [16], finding the quantiles takes constant time and inserting or removing a value from a bin V_i^j takes $O(\log |V_i^j|) = O(\log(s/m))$ time. As replacement requires up to m insertions and deletions, replacement requires order $O(m \log(s/m))$ time.

However, these relatively expensive updates are only required on average once every s/t updates, where t is the current time step or size of the stream to date. Thus the amortized cost is $O([\sum_{i=1}^{s} m \log i/m + \sum_{i=s+1}^{t} \frac{s}{i} m \log s/m]/t)$, where the first term represents the initial s time steps during which the sample is built up to its operating size and the second term represents updates to the sample once it reaches operating size. It is readily apparent that these updates rapidly become very rare and that as the size of the stream becomes very large the amortized cost becomes negligibly small.

The situation is more complex for IDAW, which maintains a window of the s most recent values for each attribute. This requires that the values be maintained in both time and value order. Maintaining an order by time can be achieved very efficiently with a circular buffer, which supports all updates and accesses in constant time. As the elements to be replaced in a replacement operation are no longer selected at random, it is not efficient to maintain the bins as interval heaps, as above. Rather we need to use slightly more expensive balanced binary trees for which the time to identify the location of the value to be removed is $O(\log(s/m))$, which this does not increase the overall complexity of the update operation relative to that for IDA. The major computational penalty, however, is that these updates must be performed for every object encountered in the queue, which makes the maintenance of the discretization a non-trivial ongoing overhead.

6.1 Evaluation

We conducted a series of studies on 5 benchmark stream classification datasets performing classification with Logistic Regression learned with Stochastic Gradient Descent, as this is an effective stream classification learning algorithm.

In the absence of concept drift the use of sampling was demonstrated to result in only negligible loss in accuracy.

On data with concept drift IDA has only negligible loss in accuracy relative to maintaining a complete quantile-based discretization relative to the entire stream encountered to date. This computationally intensive comparator approach would not be feasible in practice.

The IDAW approach delivered very substantial reductions in error for two data streams and substantial increases for two more. The reductions demonstrate that for appropriate forms of data maintaining discretizations based on quantiles as they vary over time can maintain relevant meaning while the cutpoints vary. However, the poor results on some other data streams demonstrate that this is not always appropriate.

Both IDA and IDAW consistently and substantially reduced error relative to Logistic Regression on the undiscretized numeric data.

Full details can be found in our ICDM 2014 paper [20].

References

[1] Damien Brain and G Webb. On the effect of data set size on bias and variance in classification learning. In *Proceedings of the Fourth Australian Knowledge Acquisition Workshop, University of New South Wales*, pages 117–128, 1999.

[2] Damien Brain and Geoffrey I. Webb. The need for low bias algorithms in classification learning from large data sets. In *Proceedings of the Sixth European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2002)*, pages 62–73, Berlin, 2002. Springer-Verlag.

[3] Leo Breiman. Random forests. *Machine Learning*, 45:5 – 32, 2001.

[4] Jesus Cerquides and Ramon Lopez de Mantaras. Robust Bayesian linear classifier ensembles. In *Proceedings of the 16th European Conference on Machine Learning (ECML-05)*, pages 72–83, 2005.

[5] Shenglei Chen, Ana M Martinez, and Geoffrey I Webb. Highly scalable attribute selection for averaged one-dependence estimators. In *Advances in Knowledge Discovery and Data Mining*, pages 86–97. Springer International Publishing, 2014.

[6] J. T. A. S. Ferreira, D. G. T. Denison, and D. J. Hand. Weighted naive Bayes modelling for data mining, 2001.

[7] M. A. Hall. A decision tree-based attribute weighting filter for naive Bayes. *Knowledge-Based Systems*, 20:120–126, March 2007.

[8] J. Hilden and B. Bjerregaard. Computer-aided diagnosis and the atypical case. In *In Decision Making and Medical Care: Can Information Science Help*, pages 365–378. North-Holland Publishing Company, 1976.

[9] L. Jiang and H. Zhang. Weightily averaged one-dependence estimators. In *Proceedings of the Pacific Rim Intenational Conference on Artificial Intelligence (PRICAI 2006)*, pages 970–974. Springer, 2006.

[10] Ron Kohavi. The power of decision tables. In Nada Lavrac and Stefan Wrobel, editors, *Proceedings of the European Conference on Machine Learning ECML-95*, volume 912 of *Lecture Notes in Computer Science*, pages 174–189. Springer Berlin Heidelberg, 1995.

[11] John Langford, L Li, and A Strehl. Vowpal wabbit, 2011. https://github.com/JohnLangford/vowpal_wabbit/wiki.

[12] Bin Liu and Geoffrey I Webb. Generative and discriminative learning. In Claude Sammut and Geoffrey I. Webb, editors, *Encyclopedia of Machine Learning*, pages 454–455. Springer, 2010.

[13] David JC MacKay. *Information theory, inference, and learning algorithms*, volume 7. Citeseer, 2003.

[14] Teemu Roos, Hannes Wettig, Peter Grunwald, Petri Myllymaki, and Henry Tirri. On discriminative bayesian network classifiers and logistic regression. *Machine Learning*, 59(3):267–296, 2005.

[15] M Sahami. Learning limited dependence Bayesian classifiers. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pages 334–338, Menlo Park, CA, 1996. AAAI Press.

[16] Jan van Leeuwen and Derick Wood. Interval heaps. *The Computer Journal*, 36(3):209–216, 1993.

[17] Jeffrey S Vitter. Random sampling with a reservoir. *ACM Transactions on Mathematical Software*, 11(1):37–57, 1985.

[18] Geoffrey I. Webb, Janice Boughton, and Zhihai Wang. Not so naive Bayes: Averaged one-dependence estimators. *Machine Learning*, 58(1):5–24, 2005.

[19] Geoffrey I. Webb, Janice Boughton, Fei Zheng, Kai Ming Ting, and Houssam Salem. Learning by extrapolation from marginal to full-multivariate probability distributions: decreasingly naive Bayesian classification. *Machine Learning*, 86(2):233–272, 2012. 10.1007/s10994-011-5263-6.

[20] G.I. Webb. Contrary to popular belief incremental discretization can be sound, computationally efficient and extremely useful for streaming data. In *Proceedings of the 14th IEEE International Conference on Data Mining*, pages 1031–1036, 2014.

[21] J. Wu and Z. Cai. Attribute weighting via differential evolution algorithm for attribute weighted naive Bayes (WNB). *Journal of Computational Information Systems*, 7(5):1672–1679, 2011.

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

[23] Ying Yang, Kevin Korb, Kai Ming Ting, and Geoffrey I Webb. Ensemble selection for superparent-one-dependence estimators. In *Proceedings of the 18th Australian Conference on Artificial Intelligence (AI 05)*, pages 102–111, Berlin, 2005. Springer-Verlag.

[24] N. Zaidi, M. Carman, J. Cerquides, and G.I. Webb. Naive-Bayes inspired effective pre-conditioner for speeding-up logistic regression. In *Proceedings of the 14th IEEE International Conference on Data Mining*, pages 1097–1102, 2014.

[25] Nayyar A. Zaidi, Jesus Cerquides, Mark J. Carman, and Geoffrey I. Webb. Alleviating naive Bayes attribute independence assumption by attribute weighting. *Journal of Machine Learning Research*, 14:1947–1988, 2013.

[26] Nayyar A Zaidi and Geoffrey I Webb. Fast and effective single pass Bayesian learning. In *Advances in Knowledge Discovery and Data Mining*, pages 149–160. Springer Berlin Heidelberg, 2013.

[27] H. Zhang and Shengli Sheng. Learning weighted naive Bayes with accurate ranking. In *Proceedings of the Fourth IEEE International Conference on Data Mining (ICDM -04)*, pages 567–570, 2004.

[28] F. Zheng, G.I. Webb, P. Suraweera, and L. Zhu. Subsumption resolution: An efficient and effective technique for semi-naive Bayesian learning. *Machine Learning*, 87(1):93–125, 2012.

[29] Fei Zheng and Geoffrey I. Webb. Efficient lazy elimination for averaged one-dependence estimators. In *Proceedings of the Twenty-third International Conference on Machine Learning (ICML-06)*, pages 1113 – 1120, 2006.

[30] Fei Zheng and Geoffrey I Webb. Finding the right family: parent and child selection for averaged one-dependence estimators. In *Machine Learning: ECML 2007*, pages 490–501. Springer, 2007.

7 Funding

The research reported herein has been supported both by AOARD Grant AOARD-124030 and Australian Research Council grant DP110101427.

8 List of Publications and Significant Collaborations that resulted from your AOARD supported project:

a) papers published in peer-reviewed journals

1. NA Zaidi, J Cerquides, MJ Carman, GI Webb (2013). Alleviating Naive Bayes Attribute Independence Assumption by Attribute Weighting. Journal of Machine Learning Research. 14: 1947-1988.

b) papers published in peer-reviewed conference proceedings

2. N Zaidi, M Carman & GI Webb (2014). Naive-Bayes Inspired Effective Pre-Conditioner for Speeding-up Logistic Regression. In Proceedings of the 14th IEEE International Conference on Data Mining, ICDM-14, accepted 20 September 2014.

3. GI Webb (2014). Contrary to Popular Belief Incremental Discretization can be Sound, Efficient and Extremely Useful for Streaming Data. In Proceedings of the 14th IEEE International Conference on Data Mining, ICDM-14, pp. 1031-1036.

4. S Chen, A Martinez, and GI Webb (2014). Highly Scalable Attribute Selection for AODE. In Proceedings of the 18th Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD 2014. Taiwan. Berlin/Heidelberg: Springer, pp. 86-97.

5. NA Zaidi & GI Webb (2013). Fast and Effective Single Pass Bayesian Learning. In Proceedings of the 17th Pacific-Asia Conference, PAKDD 2013. Gold Coast, Australia. Berlin/Heidelberg: Springer, pages 149-160

c) papers published in non-peer-reviewed journals and conference proceedings

d) conference presentations without papers

e) manuscripts submitted but not yet published

6. A Martinez, GI Webb, S Chen & NA Zaidi (submitted). Scalable learning of Bayesian network classifiers. 33 pp.

f) provide a list any interactions with industry or with Air Force Research Laboratory scientists or significant collaborations that resulted from this work