Probabilistic approach over Decision Trees for problems with discrete data, medium number of instances and a small to medium number of attributes 2017 — Norwich, UK

Luke M. Garrigan

Machine Learning, University of East Anglia, UK
l.garrigan@uea.ac.uk

Abstract

Bayesian classifiers are widely known for their optimality when attributes are independent given the class. This paper attempts to prove that small samples of discrete data with arbitrary dependencies are more accurately classified using a probabilistic approach over decision trees.

Index Terms: Probabilistic, Naive Bayes, Decision Trees

1 Introduction

In machine learning a learner algorithm is given a set of training instances with their corresponding class labels, it then produces a classifier. The classifier takes unlabelled testing instances and assigns it to a class. Choosing the best suited algorithm specific to the sample set is not a trivial process.

Probabilistic classification is the application of approximating a joint distribution with a product distribution. Bayes rule is used to approximate the conditional probability of a given class label. Approaches such as *naive* Bayes are among the most popular classifiers used in the machine learning community. Derived from generative probability models they are generally easy to understand and the induction of these classifiers is extremely fast, requiring only a single pass through the data if all attributes are discrete [1]. The *naive* Bayes classifier is the simplest of models in this paper, it assumes that all attributes are independent of each other given the context of the class. Although the *naive* assumption of independence is not true in terms of most sample sets, many papers such as [2] have proven that *naive* Bayes classification accuracy is very competitive when compared with more complex state-of-the-art algorithms.

Decision trees classify instances by sorting them down the tree from the root to some leaf node which represents the classification of the given instance. Nodes specify a test of some attribute of the instance and each branch from that node corresponds to one of the possible values for this attribute. A given instance is classified moving down the tree, the attribute specific to that node is tested. Following down the branch corresponding to the value of the attribute in the given example, this is then repeated until a leaf node is reached and a classification is made. Decision trees are convenient due to their transparency, they explicitly display all possible alternatives and pursues these alternatives to a conclusion. This allows for a comprehensive analysis of the consequences of each decision.

2 Data Description

- *Teaching Assistant* consists of evaluations of teaching performance the scores are divided into 3 roughly equal-size categories (”low”, ”medium,” and ”high”).
- *Seeds* comprises of data belonging to three different varieties of wheat: Kama, Rosa and Canadian. The classification is to determine which wheat it is given its characteristics
- *Breast Cancer* is data to determine whether a tumor has the ability to invade neighbouring tissue. (Benign or Malignant)
- *Ecoli*
- *Glass* classification of types of glass, motivated by criminology investigation.
- *Haberman* is data from the survival of patients who had undergone surgery for breast cancer.
- *Hayes Roth*
- *Heart*
- *Lymphography* provided by the Oncology Institute to determine whether tumours are metastases, malign or fibrous.
- *Promoters* is data for promoter gene sequences (DNA), the class labels are binary, either a promoter or not a promoter.
- *Shuttle Landing* rules for determining the conditions under which an autolanding would be preferable to manual control of the spacecraft the class label is either auto or noauto.
- *Sonar* patterns obtained by bouncing sonar signals off a metal cylinder at various angles and under various conditions. The label associated with each record contains the letter ”R” if the object is a rock and ”M” if it is a mine (metal cylinder).
- *Thyroid*

2.1 Case Study

2.1.1 Dataset Information

Electroencephalography is a domain concerning recording and interpretation of the electroencephalogram (EEG). EEG is a record of the electric signal generated by the cooperative action of brain cells, or more precisely, the time course of extracellular field potentials generated by their synchronous action[3]. EEG detects electrical activity in the brain using small, flat metal discs (electrodes) attached to the scalp. An EEG is used for diagnosing brain disorders, most frequently epilepsy.

For the current study, EEG data was collected 5 times on various days from a healthy right-handed subject of 25 years of age. The data was recorded on a Medelec Profile Digital EEG machine. The settings of high-frequency filter 50 Hz, low frequency filter 1.6 Hz, notch filter 50 Hz, sensitivity 70 micro volts/mm, and a sampling rate of 256 Hz were used for the basic signal processing.

In summary, the subject was asked to lie down in a relaxed position with eyes closed. The EEG recorded for the relaxed state for 5 minutes.Following this, an audible beep of 60 dB for 0.91 seconds was given before and after the subject was asked to mentally plan lifting of the right-hand thumb (no actual movement), after a gap of 5 minutes the same cue is given to repeat the experiment lasting approximately 30 minutes [4].

2.1.2 Attribute Information

By applying wavelet packet analysis on the original signal 12 wavelet coefficients in the 7-13 Hz frequency band were obtained. This is a classification problem to determine whether the subject is relaxing or planning. There are 182 instances; the univariate dataset contains 12 real attributes and a binary class label.

Table 1: Details of the data used. C=Categorical, I=Integer, R=Real

Dataset	Inst/Attr	Attr Type	Class #
Teaching Assistant	151/5	C,I	3
Seeds	210/7	R	3
Planning Relax	182/13	R	2
Breast Cancer	286/9	C	2
Ecoli	223/10	R	8
Glass	141/6	R	2
Haberman	203/4	R	2
Hayes Roth	160/5	C	3
Heart	169/14	R	2
Lymphography	148/18	C	4
Promoters	106/58	C	2
Shuttle Landing	252/7	C	2
Sonar	137/60	R	2
Thyroid	142/6	R	2

3 Classifier Description

3.1 Probabalistic

3.1.1 Naive Bayes

Naive Bayes was selected for testing due to its popularity for classification problems and its presence in the machine learning community. As its name suggests its assumptions are naive and are not generally concordant with the data; it assumes all attributes are independent of each other given the context of the class but has been shown to perform surprisingly well in many classification problems. It is computationally efficient as training is linear in both the number of instances and attributes [5].

Let C represent the classification variable, and let C_k be the value of C. According to Bayes Rule, the probability of an instance X with attributes $X = (x_1, x_2, .., x_n)$ having class label C_k is

$$P(C_k \mid X) = \frac{P(X \mid C_k)\, P(C_k)}{P(X)}$$

where $P(C_k \mid X)$ represents the posterior probability of class c given a predictor X. $P(X \mid C_k)$ is the likelihood; the probability of the predictor given C_k. $P(C_k)$ is the prior probability of C_k; the current knowledge of the class distribution and $P(X)$ is the predictor prior probability.

naive Bayes assumes that all attributes are independent given the value of the class label.

$$P(X|C_k) = P(x_1, x_2, .., x_n \mid C_k) = \prod_{i=1}^{n} P(x_i|C_k)$$

So the conditional distribution over the class variable C can be represented by:

$$P(C_k \mid x_1, ..x_n) = \frac{1}{Z} P(C_k) \prod_{i=1}^{n} P(x_i \mid C_k)$$

where Z is a scaling factor dependent on $x_1, x_2, .., x_n$. In order to build a classifier from this, a decision rule is used by picking the classification which is most probable; for two posterior probabilities $P(C_1|X) = 0.54$ and $P(C_2|X) = 0.46$ using the *maximum a posteriori* (MAP) rule, the larger of the two is chosen [6].

3.1.2 Bayesian Networks

Bayesian networks belong to the family of probabilistic graphical models. Graphical model expresses the conditional dependence structure between random variables and are used to represent knowledge about uncertain domains. Nodes within a Bayesian network represent the random variables, edges represent the conditional dependencies; nodes not connected are conditionally independent of each other. Each node is associated with a probability function that takes as input a particular set of values for the node's parent variables and outputs the probability of the variable represented by the node [7].

In Bayesian networks each node is conditionally independent of any subset of nodes that are not descendants of itself gives its parent, so the value of a node is conditional only on the values of its parent nodes. Let V represent a node in the graph and $par(V_i)$ be the parent of the node:

$$P(V_1, V_2, \ldots V_n) = \prod_{i=1}^{n} P(V_i \mid par(V_i))$$

So to compute the joint probabilities we must calculate the conditional probability between itself and its parent for each node in the graph. The chain rule is then computer to determine the graphs joint probability functions.

$$P(V_1, V_2, \ldots V_n) = P(V_1) P(V_2 \mid V_1) P(V_3 \mid V_1, V_2)$$

Prior probabilities are nodes without parents; they are not conditioned on other random variables [8].

Bayesian networks do not necessarily imply a commitment to Bayesian statistics, it is common to use frequentists methods to estimate the parameters of the CPDS.

3.2 Decision Trees

3.2.1 C4.5

C4.5 is an algorithm used to build decision trees, they are created using a divide-and-conquer approach. The tree is formed using information entropy as found in the ID3 algorithm which is a precursor to C4.5, however, this paper only covers the C4.5 algorithm. C4.5 offers a number of improvements over ID3: handling data with missing values, accepts both continuous and discrete attributes, handling attributes with different costs and solves over-fitting by pruning. Ross Quinlan's latest iteration is the C5.0 algorithm which he states is several orders of magnitude faster than C4.5 [9], unfortunately, this software is proprietary thus not compared.

Entropy is a measure of unpredictability, Shannon entropy calculates the level of uncertainty:

$$Entropy(P) = -\sum_{i=1}^{n} P_i(\log_2 P_i)$$

Information gain measures the expected reduction in entropy caused by partitioning according to the given attribute.

$$Gain(S, A) = Entropy(S) - \sum_{vinValues} \frac{|S_v|}{|S|} Entropy(S)$$

C4.5 is implemented recursively, let T be the set of training instances, the algorithm chooses an attribute that best differentiates the instances contained in T, a tree node is created with the value of the chosen attribute. Child nodes are created each link represents a unique value for the given attribute, the child values are then used to further subdivide the instances into subclasses. The subclasses either satisfy the predefined criteria or the remaining attributes choice for the path is null, this is repeated recursively. The classifications for the testing instances is made by following the decision path.

3.2.2 Random Forests

Meta-algorithms are approaches to combine several machine learning techniques into one predictive model in order to decrease the variance (bagging) or bias (boosting) [10].

Ensemble learning methods generate many classifiers and aggregate their results. Random forest incorporates a supplementary layer of randomness to bagging, in addition to constructing each tree using a different bootstrap sample of the data. Random forests construct the classification distinctly using the best among a subset of predictors randomly chosen at that node [11]. Although the approach is somewhat counter-intuitive it has been shown to outperform many state-of-the-art classifiers.

3.2.3 Logistic Model Trees

Logistic Model Trees (LMT) use a combination of a tree structure and logistic regression models to for a single tree. [12] performed experiments showing that LMT produces more accurate classifiers than C4.5, CART, logistic regression, model trees, functional trees, *naive* Bayes trees and Lots.

LMT is a combination of learners which rely on logistic regression models. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables [13]. LMT uses cost-complexity pruning. The compute time of LMT is much greater than the other algorithms.

4 Preliminaries

4.1 Discretization

Discretization concerns with the process of transferring continuous data into discrete counterparts. Numeric attributes were discretized into ten equal-length intervals unless the number of uniquely observed values for an attribute was less than 10. This approach was compared in [14] with entropy-based and purity-based methods, which are supervised algorithms. An empirical evaluation showed that the *naive* Bayes algorithm significantly improved accuracy when features were discretized using an entropy-based method. However, due to its simplicity, the unsupervised binning discretization method was used for all continuous attributes.

4.2 Methods For Accuracy Estimation

4.2.1 K-Fold Cross-Validation

Cross-validation is a computationally expensive algorithm used to estimate performance, it uses all available instances as both training and testing sets. The dataset is split into k equally sized non-overlapping subsets S. Given a fold S_i a model is trained on $S \setminus S_i$, then S_i is used to create the accuracy estimation.

4.2.2 Leave-One-Out Cross-Validation (LOOCV)

Leave-One-Out Cross-Validation is K-fold cross validation where K is equal to the number of instances in the dataset. The classifier is trained on all data except the one instance being left, this is repeated until all instances in the dataset have been the test instance. An average of the data is collected and used to evaluate the classifier.

4.2.3 Bias And Variance Tradeoff

As k increases the less bias the classification is in overestimating the true expected error as the folds will be closer to the total dataset. However, in doing this it induces variance. To minimise the testing bias a large portion of the dataset must be used for training, meaning not much data is used for testing, this ensures that the model will be as close as possible to the model achievable by training using the entire dataset. Minimising the testing variance would mean quite the opposite, a large amount of data would be used for testing, this ensures a more reliable estimate error of the classifier.

Figure 1 shows the bias and variance of k-fold cross-validation on arbitrary datasets using *naive* bayes. More in-depth experiments have been conducted and state that for real-world data the best method for cross-validation is ten-fold stratified and that even if you have the computational power to use more folds ten-fold is the better option [15]. Ten-fold stratified cross-validation was used for experiments in this paper.

4.3 Missing Values

For scenarios such as $P(X_1 = x_1 \dots X_k = ? \dots X_n = x_n|c)$ where ? represents a missing value, *naive* Bayes skips over the missing value: $P(x_1 \dots X_k \dots x_n \mid c) = \prod_{i \neq k}^{n} P(x_i \mid c)$. The

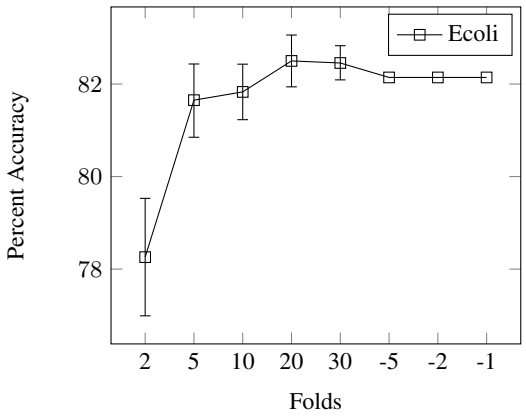

Figure 1: Bias representation of cross-validation. The negative k-folds shows leave-k-out, -1 is LOOCV.

documentation of J48 implementation of C4.5 states that any split on an attribute with missing value will be done with weights proportional to frequencies of the observed non-missing values [16]. In this paper missing values were ignored completely, in an attempt to ensure that certain models don't have an advantage over others.

4.4 Zero Counts

Zero counts are when a class value and attribute value have never occurred together in the training set; this doesn't necessarily mean that this data is never going to be apart of a testing set. Laplace correction was used as a fail safe. Let N_{ijk} represent the number of examples where $X_i = x_k$ and $C = C_j$, let N_j represent the number of instances of class c_j and k the number of possible values of Xi. When computing the conditional probabilities rather than using the estimate $P(X_i = x_k|c_j) = \frac{N_{ijk}}{N_j}$, Laplace correction was used:

$$P(X_i = x_k \mid C_j) = \frac{N_{ijk} + 1}{N_j + k}$$

5 Results

Table 2 displays the average classification accuracies and sample standard deviations over 20 iterations of ten-fold cross-validation. All 14 data-sets were classified using each algorithm.

Planning Relax was specifically chosen as a case study as it goes against the initial hypothesis and rather significantly; all decision tree algorithms output higher classification accuracies than both probability algorithmsF for this dataset. This could be for a number of reasons, it is possible that some of these datasets contain little attribute dependencies, this could result in *naive* Bayesoutperform the other or on the contrary the attributes could have significant dependencies meaning *naive* Bayes would likely perform poorly. To come to any conclusion a measurement of attribute dependence must be carried out.

5.1 Timing Experiment

All classifiers were put to a timed test using the largest dataset(Ecoli). Table 3 shows results for experiments performed on all classifiers, the data was split up into training and testing. The training data was split randomly by a set percentage of its original size, the size of the testing data was kept constant throughout. As expected, as the training data decreases in size so does the time taken to train the classifier, this is consistent through all algorithms. The testing data seemed to follow unexpected trends; when the training set was at its largest for *naive* Bayes, Bayes Net, C4.5 and Random Forest all outputted the smallest evaluation time. The time taken for the testing set continued to increase until 60% where it seemed to plateau.

Figure 2 shows the comparison of *naive* Bayes, Bayes Net and C4.5 showing the relationship between amount of training instances and the increase inevitable increase in time. Figure 3 shows the same relationship between LMT and random forests, a different graph was used because the time taken to build the classifer for LMT and random forests (specifically LMT) are much larger than the other algorithms.

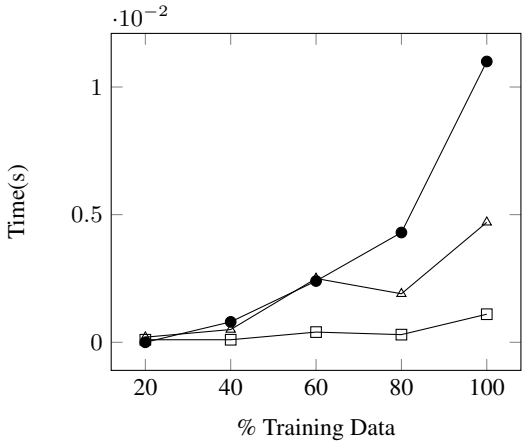

Figure 2: Training timing Experiment, square = *naive* Bayes, triangle = Bayes Net and diamond = C4.5

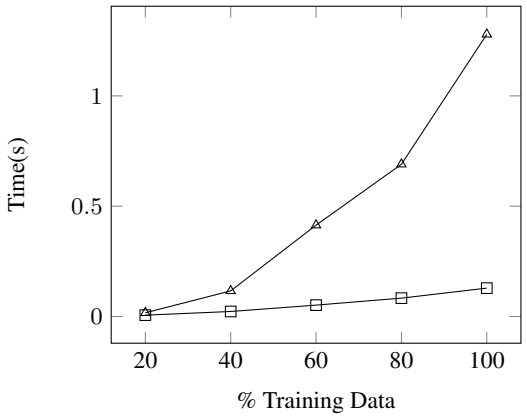

Figure 3: Training timing Experiment, square = Random Forest, triangle = LMT

Table 2: Classification accuracies and sample standard deviation, using 20 iterations of ten-fold cross-validation.

Dataset	*Naive* Bayes	Bayes Net	C4.5	Random Forests	LMT
Teaching Assistant	54.01±11.02	46.32±7.74	57.41±12.64	67.54±11.35	53.27±11.08
Seeds	89.38±7.09	89.76±7.05	89.57±7.09	89.81±6.42	90.38±6.35
Planning Relax	56.59±9.41	55.23±9.72	71.46±1.53	71.72±4.61	71.02±3.15
Breast Cancer	72.88±8.61	72.69±8.87	76.11±5.44	73.51±7.34	74.31±6.69
Ecoli	81.88±7.28	83.22±6.64	71.65±6.90	76.52±6.66	81.30±8.08
Glass	93.98±6.50	94.98±6.48	85.00±7.91	92.44±7.30	89.07±9.04
Haberman	73.54±6.90	73.48±7.09	73.12±2.90	67.82±8.97	72.17±6.43
Hayes Roth	83.83±8.01	83.96±7.97	72.27±8.46	80.90±9.33	85.15±7.79
Heart	81.17±8.16	81.17±7.57	75.06±9.69	77.94±9.62	80.28±8.30
Lymphography	84.04±9.43	85.46±8.78	79.12±9.96	83.33±9.84	81.86±9.82
Promoters	91.14±8.71	91.14±9.12	77.45±12.79	91.05±9.37	90.41±9.15
Shuttle Landing	93.26±4.56	93.26±4.56	96.59±3.63	98.30±2.64	98.41±2.52
Sonar	73.49±9.68	73.27±9.85	62.90±11.07	74.85±11.17	72.09±11.81
Thyroid	91.32±6.49	92.87±6.40	90.13±7.13	91.88±6.18	90.56±7.18
Mean	80.04	79.77	76.99	81.26	80.73

Table 3: Timed Experiment Of All Classifiers, training (top five rows) and testing (bottom five)

% train	*Naive* Bayes	Bayes Net	C4.5	RF	LMT
100	0.0011±0.0003	0.0047±0.0009	0.0110±0.0011	0.1286±0.0067	1.2791±0.2681
80	0.0003±0.0005	0.0019±0.0003	0.0043±0.0005	0.0832±0.0013	0.6901±0.1273
60	0.0004±0.0005	0.0025±0.0015	0.0024±0.0005	0.0516±0.0034	0.4140±0.1904
40	0.0001±0.0003	0.0005±0.0005	0.0008±0.0004	0.0227±0.0007	0.1161±0.0217
20	0.0001±0.0003	0.0002±0.0004	0.0000±0.0000	0.0062±0.0004	0.0168±0.0039
100	0.0029±0.0003	0.0007±0.0005	0.0002±0.0004	0.0042±0.0014	0.0005±0.0005
80	0.0077±0.0005	0.0014±0.0005	0.0004±0.0005	0.0096±0.0007	0.0004±0.0005
60	0.0090±0.0011	0.0017±0.0005	0.0008±0.0004	0.0111±0.0006	0.0007±0.0005
40	0.0090±0.0014	0.0015±0.0005	0.0005±0.0005	0.0085±0.0005	0.0009±0.0003
20	0.0087±0.0009	0.0011±0.0003	0.0008±0.0004	0.0040±0.0000	0.0004±0.0005

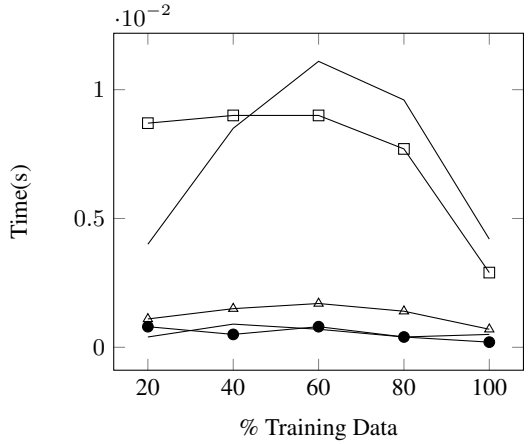

Figure 4: Timing Experiment, square = Random Forest, triangle = LMT

5.2 Supplementary Material

6 Conclusions

6.1 References

References

[1] R. Kohavi, "Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid." in *KDD*, vol. 96. Citeseer, 1996, pp. 202–207.

[2] P. Domingos and M. Pazzani, "On the optimality of the simple bayesian classifier under zero-one loss," *Machine learning*, vol. 29, no. 2, pp. 103–130, 1997.

[3] D. B. Lindsley, "Emotions and the electroencephalogram." 1950.

[4] "Uci machine learning repository: Planning relax data set." [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Planning Relax

[5] E. Frank, M. Hall, and B. Pfahringer, "Locally weighted naive bayes," in *Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc., 2002, pp. 249–256.

[6] H. Zhang, "The optimality of naive bayes," *AA*, vol. 1, no. 2, p. 3, 2004.

[7] D. Heckerman, D. Geiger, and D. M. Chickering, "Learning bayesian networks: The combination of knowledge and statistical data," *Machine learning*, vol. 20, no. 3, pp. 197–243, 1995.

[8] F. V. Jensen, *An introduction to Bayesian networks*. UCL press London, 1996, vol. 210.

[9] R. Quinlan, "Data mining tools see5 and c5. 0," 2004.

[10] A. Galkin, "Bagging, boosting and stacking in machine learning." [Online]. Available: https://stats.stackexchange.com/questions/18891/bagging-boosting-and-stacking-in-machine-learning

[11] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[12] N. Landwehr, M. Hall, and E. Frank, "Logistic model trees," *Machine Learning*, vol. 59, no. 1-2, pp. 161–205, 2005.

[13] "What is logistic regression?" [Online]. Available: http://www.statisticssolutions.com/what-is-logistic-regression/

[14] J. Dougherty, R. Kohavi, M. Sahami *et al.*, "Supervised and unsupervised discretization of continuous features," in *Machine learning: proceedings of the twelfth international conference*, vol. 12, 1995, pp. 194–202.

[15] R. Kohavi *et al.*, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Ijcai*, vol. 14, no. 2. Stanford, CA, 1995, pp. 1137–1145.

[16] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.