Comparing Naive Bayes, Decision Trees, and SVM with AUC and Accuracy

Jin Huang Jingjing Lu* Charles X. Ling
Department of Computer Science
The University of Western Ontario
London, Ontario, Canada N6A 5B7
{jhuang, jlu, cling}@csd.uwo.ca
(* Research Visitor from Yan Cheng, Jiang Su Province, China)

(To appear in The Third IEEE International Conference on Data Mining, 2003)

Abstract

Predictive accuracy has often been used as the main and often only evaluation criterion for the predictive performance of classification or data mining algorithms. In recent years, the area under the ROC (Receiver Operating Characteristics) curve, or simply AUC, has been proposed as an alternative single-number measure for evaluating performance of learning algorithms. In our previous work, we proved that AUC is, in general, a better measure (defined precisely) than accuracy. Many popular data mining algorithms should then be re-evaluated in terms of AUC. For example, it is well accepted that Naive Bayes and decision trees are very similar in accuracy. How do they compare in AUC? Also, how does the recently developed SVM (Support Vector Machine) compare to traditional learning algorithms in accuracy and AUC? We will answer these questions in this paper. Our conclusions will provide important guidelines in data mining applications on real-world datasets.

1 Introduction

The predictive ability of a classification algorithm is typically measured by its predictive accuracy (or error rate, which is 1 minus the accuracy) on the testing examples. Most classifiers (including C4.5 and Naive Bayes) can also produce probability estimations or "confidence" of the class prediction. Unfortunately, this information is completely ignored in accuracy.

In recent years, the ROC (Receiver Operating Characteristics) curve [9, 5], which is plotted with the probability of the class prediction, has been introduced to evaluate performance of machine learning algorithms [19, 20]. Bradley [2] compared popular machine learning algorithms using AUC (area under the curve) of ROC, [1] and found that AUC exhibits several desirable properties compared to accuracy. An additional benefit of AUC is that it is a way of measuring ranking, which is very useful in many data mining applications. However, no formal arguments or criteria were established for comparing the two measures. Recently, other researchers used AUC to construct learning algorithms [8, 15]. But it is not clear if and why AUC is a better measure than accuracy. In general, how can we compare any two evaluation measures for learning algorithms? How can we establish that one measure is "better" than another for any two measures?

In our recent work [14], we gave formal definitions on the (strict) consistency and discriminancy to compare any two measures. However, for AUC and accuracy, we found counter examples which show that AUC and accuracy are not strictly consistent, and AUC is not strictly discriminant than accuracy. We then extended the definitions to the degree of consistency and degree of discriminancy, and we defined that a measure is better than the other based on the degree of consistency and degree of discriminancy. Then we applied these definitions to AUC and accuracy, and verified empirically and proved theoretically that AUC is a better measure than accuracy. That is, AUC is indeed statistically consistent and more discriminant than accuracy. Details can be found in [14].

However, most previous work only focussed on comparing the learning algorithms in accuracy. A well-accepted conclusion in the machine learning community is that the popular decision tree learning algorithm C4.5 [21] and Naive Bayes are very similar in predictive accuracy [11, 12, 6]. How do popular learning algorithms, such as decision trees and Naive Bayes, compare in terms of the better measure AUC? How does recent Support Vector Machine (SVM) compare to traditional learning algorithms such as

[1] AUC of ROC is simply called AUC in our paper.

Naive Bayes and decision trees? In this paper, we will answer these questions experimentally.

2 Empirical Comparison of Naive Bayes, Decision trees, and SVM

We first compare Naive Bayes and decision trees with AUC and accuracy in Section 2.1, and we then add SVM in our comparison in Section 2.2. This is because we only use binary datasets in comparisons with SVM (see Section 2.2 for details).

2.1 Comparing of Naive Bayes and Decision Trees

The popular decision tree learning algorithm C4.5 have been recently observed to produce poor probability estimations on AUC [22, 20, 18]. Several improvements have been proposed, and we want to include a recent improvement, C4.4 [18], in our comparison.

We conduct our experiments to compare Naive Bayes, C4.5, and its recent improvement C4.4, using both accuracy and AUC as the evaluation criterion. We use 18 datasets with a relatively large number of examples from the UCI repository [1].

Our experiments follow the procedure below:

1. The continuous attributes in all datasets are discretized by the entropy-based method described in [7].
2. For each dataset, create 10 pairs of training and testing sets with 10-fold cross-validation, and run Naive Bayes, C4.5, and C4.4 on the *same* training sets and test them on the *same* testing sets to obtain the testing accuracy and AUC scores.

The averaged results on accuracy are shown in Table 2.1, and on AUC in Table 2.1. As we can see from Table 2.1, the three algorithms have very similar predictive accuracy. The two tailed, paired t-test with 95% confidence level (same for other t-tests in the rest of the paper) shows that there is no statistical difference in accuracy between Naive Bayes and C4.4, Naive Bayes and C4.5, and C4.4 and C4.5. This verifies results of previous publications [11, 12, 6].

When we analyze the table for AUC (see Table 2.1), we get some very interesting and surprising results. The average predictive AUC score of Naive Bayes is slightly higher than that of C4.4, and much higher than that of C4.5. The paired t-test shows that the difference between Naive Bayes and C4.4 is not significant, but the difference between Naive Bayes and C4.5 is significant. (The difference between C4.4 and C4.5 is also significant, as observed by [18]). That is, Naive Bayes outperforms C4.5 in AUC with significant difference. This verifies our second hypothesis mentioned above.

Dataset	NB	C4.4	C4.5
breast	97.5±2.9	92.9±3.0	92.8±1.2
cars	86.4±3.7	88.9±4.0	85.1±3.8
credit	85.8±3.0	88.1±2.8	88.8±3.1
dermatology	98.4±1.9	94.0±3.5	94.0±4.2
echocardio	71.9±1.8	73.6±1.8	73.6±1.8
ecoli	96.7±2.2	96.4±3.1	95.5±3.9
glass	71.8±2.4	73.3±3.9	73.3±3.0
heart	80.8±7.3	78.9±7.6	81.2±5.6
hepatitis	83.0±6.2	81.3±4.4	84.02±4.0
import	96.1±3.9	100.0±0.0	100.0±0.0
iris	95.3±4.5	95.3±4.5	95.3±4.5
liver	62.3±5.7	60.5±4.8	61.1±4.9
mushroom	97.2±0.8	100.0±0.0	100.0±0.0
pima	71.4±5.8	71.9±7.1	71.7±6.8
solar	74.0±3.2	73.0±3.1	73.9±2.1
thyroid	95.7±1.1	96.0±1.1	96.6±1.1
voting	91.4±5.6	95.7±4.6	96.6±3.9
wine	98.9±2.4	95.0±4.9	95.5±5.1
Average	**86.4**	**86.4**	**86.6**

Table 1. Predictive accuracy values of Naive Bayes, C4.4, and C4.5

This conclusion is quite significant to the machine learning and data mining community. Previous research concluded that Naive Bayes and C4.5 are very similar in prediction measured by accuracy [11, 12, 6]. As we have established in this paper, AUC is a better measure than accuracy. Further, our results show that Naive Bayes and C4.4 outperform the most popular decision tree algorithm C4.5 in terms of AUC. This indicates that Naive Bayes (and C4.4) should be favoured over C4.5 in machine learning and data mining applications, especially when ranking is important.

2.2 Comparing Naive Bayes, Decision Trees, and SVM

In this section we compare accuracy and AUC of Naive Bayes, C4.4, and C4.5 to the recently developed SVM on the datasets from the UCI repository. Such an extensive comparison with a large number of benchmark datasets is still rare [17]; most previous work limited to only a few comparisons, with the exception of [17]. SVM is essentially a binary classifier, and although extensions have been made to multiclass classification [23, 10] there is no consensus which is the best. Therefore, we take the 13 binary-class datasets from the 18 datasets in the experiments involving SVM. [17] also only used binary datasets for the classification for the same reason.

For SVM we use the software package LIBSVM [4] modified to directly output the evaluation of the hyperplane target function as scores for ranking. We used the Gaussian

Dataset	NB	C4.4	C4.5
breast	97.5±0.9	96.9±0.9	95.1±2.4
cars	92.8±3.3	94.1±3.2	91.4±3.5
credit	91.9±3.0	90.4±3.2	88.0±4.1
dermatology	98.6±0.1	97.5±1.1	94.6±3.3
echocardio	63.8±2.1	69.4±2.2	68.9±2.3
ecoli	97.0±1.1	97.0±1.0	94.3±3.6
glass	76.1±2.4	73.1±2.6	71.3±3.3
heart	82.7±6.1	80.1±7.8	76.2±7.0
hepatitis	76.5±4.4	62.9±8.2	59.2±6.8
import	91.7±4.5	94.4±2.0	95.1±2.6
iris	94.2±3.4	91.8±3.8	92.4±4.6
liver	61.5±5.9	59.6±5.7	60.5±5.0
mushroom	99.7±0.1	99.9±0.0	99.9±0.0
pima	75.9±4.2	73.4±7.3	72.4±7.4
solar	88.7±1.7	87.7±1.9	85.2±2.8
thyroid	94.9±1.8	94.3±2.6	92.1±5.5
voting	91.4±3.7	95.2±2.2	93.4±3.7
wine	95.3±1.8	94.4±1.2	91.6±4.0
Average	**87.2**	**86.2**	**84.5**

Table 2. Predictive AUC values of Naive Bayes, C4.4, and C4.5

Kernel for all the experiments. The parameters C (penalty for misclassification) and gamma (function of the deviation of the Gaussian Kernel) were determined by searching for the maximum accuracy in the two-dimensional grid formed by different values of C and gamma in the 3-fold cross-validation on the training set (so the testing set in the original 10-fold cross-validation is not used in tuning SVM). C was sampled at 2^{-5}, 2^{-3}, 2^{-1}, ..., 2^{15}, and gamma at 2^{-15}, 2^{-13}, 2^{-11}, ..., 2^{3}. Other parameters are set default values by the software. This experiment setting is similar to the one used in [17]. The experiment procedure is the same as discussed earlier.

The predictive accuracy and AUC of SVM on the testing sets of the 13 binary datasets are listed in Table 2.2. As we can see, the average predictive accuracy of SVM on the 13 binary datasets is 87.8%, and the average predictive AUC is 86.0%. From Table 2.1 we can obtain the average predictive accuracy of Naive Bayes, C4.4, and C4.5 on the 13 binary datasets is 85.9%, 86.5%, and 86.7%, respectively. Similarly, from Table 2.1 we can obtain the average predictive AUC of Naive Bayes, C4.4, and C4.5 on the 13 binary datasets is 86.0%, 85.2%, and 83.6%, respectively.

Some interesting conclusions can be drawn. First, the average predictive accuracy of SVM is slightly higher than other algorithms in comparison. However, the paired t-test shows that the difference is *not* statistically significant. Secondly, the average predictive AUC scores showed that SVM, Naive Bayes, and C4.4 are very similar. In fact, there is no statistical difference among them. However, SVM does have significantly higher AUC than C4.5, so does Naive Bayes and C4.4 (as observed in the early comparison in Section 2.1). Our results on SVM may be inconsistent with some other comparisons involving SVM which showed superiority of SVM over other learning algorithms [17, 13, 3]. We think that one major difference is data pre-processing: we have discretized all numerical attributes (see Section 2.1) as Naive Bayes requires all attributes to be discrete. Discretization is also an important pre-processing step in data mining [16]. The discretized attributes are named 1, 2, 3, and so on. Decision trees and Naive Bayes then take discrete attributes directly. For SVM, those values are taken as numerical attributes after normalization. In most previous comparisons, numerical attributes are used directly in SVM. However, we think that our comparisons are fair since all algorithms use the same training and testing datasets after discretization. If there is loss of information during discretization, the decision trees, Naive Bayes, and SVM would suffer equally from it. The other difference is that we did not seek for problem-specific, best kernels for SVM. This is fair as Naive Bayes, C4.5, and C4.4, are run automatically in the default, problem-independent parameter settings.

Dataset	Accuracy	AUC
breast	96.5±2.3	97.3±1.3
cars	97.0±1.3	98.6±0.4
credit	86.4±2.9	90.4±3.0
echocardio	73.6±1.8	71.5±2.0
ecoli	96.4±3.1	95.0±2.8
heart	79.7±8.2	82.1±8.3
hepatitis	85.8±4.2	64.2±8.7
import	100.0±0.0	93.8±0.6
liver	60.5±4.8	61.6±5.6
mushroom	99.9±0.1	99.9±0.0
pima	72.2±6.3	72.2±7.5
thyroid	96.7±1.3	95.8±3.3
voting	97.0±3.5	95.3±0.7
Average	**87.8**	**86.0**

Table 3. Predictive accuracy and AUC of SVM on the 13 binary datasets

To summarize, our extensive experiments in this section allows us to draw the following conclusions:

- The average predictive accuracy of the four learning algorithms compared: Naive Bayes, C4.5, C4.4, and SVM, are very similar. There is no statistical difference between them. The recent SVM does produce slightly higher average accuracy but the difference on the 13 binary datasets is not statistically significant.

- The average predictive AUC values of Naive Bayes, C4.4, and SVM are very similar (no statistical difference), and they are all higher with significant difference than C4.5.

Our conclusions will provide important guidelines in data mining applications on real-world datasets.

3 Conclusions

In our previous work, we proved that AUC is, in general, a better measure than accuracy. Many popular data mining algorithms should then be re-evaluated in terms of AUC. We compare experimentally Naive Bayes, C4.5, C4.4, and SVM in both accuracy and AUC, and conclude that they have very a similar predictive accuracy. In addition, Naive Bayes, C4.4, and SVM produce similar AUC scores, and they all outperform C4.5 in AUC with significant difference. Our conclusions will provide important guidelines in data mining applications on real-world datasets.

Acknowledgements

We gratefully thank Foster Provost for kindly providing us with the source codes of C4.4, which is a great help to us in the comparison of C4.5 and C4.4 to other algorithms. Jianning Wang, Dansi Qian, and Huajie Zhang also helped us at various stages of the experiments.

References

[1] C. Blake and C. Merz. UCI repository of machine learning databases. http://www.ics.uci.edu/∼mlearn/MLRepository.html, 1998. University of California, Irvine, Dept. of Information and Computer Sciences.

[2] A. P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30:1145–1159, 1997.

[3] M. Brown, W. Grundy, D. Lin, and N. C. et al. Knowledge-based analysis of microarray gene expression data using support vector machines. In *Proceedings of the National Academy of Sciences*, pages 262–267, 2000.

[4] C. C. Chang and C. Lin. Libsvm: A library for support vector machines (version 2.4), 2003.

[5] C.W.Therrien. *Decision Estimation and Classification: An Introduction to Pattern Recognition and Related Topics*. Wiley, New York, 1989.

[6] P. Domingos and M. Pazzani. Beyond independence: conditions for the optimality of the simple Bayesian classifier. In *Proceedings of the Thirteenth International Conference on Machine Learning*, pages 105 – 112, 1996.

[7] U. Fayyad and K. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In *Proceedings of Thirteenth International Joint Conference on Artificial Intelligence*, pages 1022–1027. Morgan Kaufmann, 1993.

[8] C. Ferri, P. A. Flach, and J. Hernandez-Orallo. Learning decision trees using the area under the ROC curve. In *Proceedings of the Nineteenth International Conference on Machine Learning (ICML 2002)*, pages 139–146, 2002.

[9] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, second edition, 1990.

[10] C. Hsu and C. Lin. A comparison on methods for multiclass support vector machines. Technical report, Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan, 2001.

[11] I. Kononenko. Comparison of inductive and naive Bayesian learning approaches to automatic knowledge acquisition. In B. Wielinga, editor, *Current Trends in Knowledge Acquisition*. IOS Press, 1990.

[12] P. Langley, W. Iba, and K. Thomas. An analysis of Bayesian classifiers. In *Proceedings of the Tenth National Conference of Artificial Intelligence*, pages 223–228. AAAI Press, 1992.

[13] Y. Lin. Support vector machines and the bayes rule in classification. *Data Mining and Knowledge Discovery*, 6(3):259–275, 2002.

[14] C. X. Ling, J. Huang, and H. Zhang. AUC: a statistically consistent and more discriminating measure than accuracy. In *Proceedings of 18th International Conference on Artificial Intelligence (IJCAI-2003)*, pages 329–341, 2003.

[15] C. X. Ling and H. Zhang. Toward Bayesian classifiers with accurate probabilities. In *Proceedings of the Sixth Pacific-Asia Conference on KDD*, pages 123–134. Springer, 2002.

[16] H. Liu, F. Hussain, C. L. Tan, and M. Dash. Discretization: An enabling technique. *Data Mining and Knowledge Discovery*, 6(4):393–423, 2002.

[17] D. Meyer, F. Leisch, and K. Hornik. Benchmarking support vector machines. Technical report, Vienna University of Economics and Business Administration, 2002.

[18] F. Provost and P. Domingos. Tree induction for probability-based ranking. *Machine Learning*, 2003. To appear.

[19] F. Provost and T. Fawcett. Analysis and visualization of classifier performance: comparison under imprecise class and cost distribution. In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*, pages 43–48. AAAI Press, 1997.

[20] F. Provost, T. Fawcett, and R. Kohavi. The case against accuracy estimation for comparing induction algorithms. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 445–453. Morgan Kaufmann, 1998.

[21] J. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann: San Mateo, CA, 1993.

[22] P. Smyth, A. Gray, and U. Fayyad. Retrofitting decision tree classifiers using kernel density estimation. In *Proceedings of the 12th International Conference on machine Learning*, pages 506–514, 1995.

[23] J. A. K. Suykens and J. Vandewalle. Multiclass least squares support vector machines. In *IJCNN'99 International Joint Conference on Neural Networks*, Washington, DC, 1999.