Beyond Independence: Conditions for the Optimality of the Simple Bayesian Classifier

Pedro Domingos
Dept. Information and Computer Science
University of California, Irvine
Irvine, CA 92717, U.S.A.
pedrod@ics.uci.edu

Michael Pazzani
Dept. Information and Computer Science
University of California, Irvine
Irvine, CA 92717, U.S.A.
pazzani@ics.uci.edu

Abstract

The simple Bayesian classifier (SBC) is commonly thought to assume that attributes are independent given the class, but this is apparently contradicted by the surprisingly good performance it exhibits in many domains that contain clear attribute dependences. No explanation for this has been proposed so far. In this paper we show that the SBC does not in fact assume attribute independence, and can be optimal even when this assumption is violated by a wide margin. The key to this finding lies in the distinction between classification and probability estimation: correct classification can be achieved even when the probability estimates used contain large errors. We show that the previously-assumed region of optimality of the SBC is a second-order infinitesimal fraction of the actual one. This is followed by the derivation of several necessary and several sufficient conditions for the optimality of the SBC. For example, the SBC is optimal for learning arbitrary conjunctions and disjunctions, even though they violate the independence assumption. The paper also reports empirical evidence of the SBC's competitive performance in domains containing substantial degrees of attribute dependence.

1 THE SIMPLE BAYESIAN CLASSIFIER

Bayes' theorem tells us how to optimally predict the class of a previously unseen example, given a training sample (Duda & Hart, 1973). The chosen class should be the one which maximizes $P(C_i|E) = P(C_i)\,P(E|C_i)\,/\,P(E)$, where C_i is the ith class, E is the test example, $P(Y|X)$ denotes the conditional probability of Y given X, and probabilities are estimated from the training sample. Let an example be a vector of a attributes. If the attributes are *independent* given the class, $P(E|C_i)$ can be decomposed into the product $P(v_1|C_i)\ldots P(v_a|C_i)$, where v_j is the value of the jth attribute in the example E. Therefore one should predict the class that maximizes:

$$P(C_i|E) = \frac{P(C_i)}{P(E)} \prod_{j=1}^{a} P(v_j|C_i) \qquad (1)$$

This procedure is often called the *naive Bayesian classifier*. Here we will prefer the term *simple*, and abbreviate to *SBC*. The SBC is commonly thought to be optimal, in the sense of achieving the best possible accuracy, only when the "independence assumption" above holds, and perhaps close to optimal when the attributes are only slightly dependent. However, this very restrictive condition seems to be inconsistent with the SBC's surprisingly good performance in a wide variety of domains, including many where there are clear dependencies between the attributes. In a study on 28 datasets from the UCI repository reported in the next section, the SBC was found to be more accurate than C4.5 in 16 domains, and similarly for CN2 and PEBLS. Other authors have made similar observations (e.g., (Clark & Niblett, 1989; Langley, Iba & Thompson, 1992; Rachlin, Kasif, Salzberg & Aha, 1994; Dougherty, Kohavi & Sahami, 1995)), but no interpretation of this has been proposed so far. Several extensions of the SBC have been introduced with the goal of increasing its tolerance of attribute dependences (e.g., (Kononenko, 1991; Langley, 1993; Langley & Sage, 1994; Pazzani, 1995)), usually with moderate success. Here we begin to shed some light on the matter by showing that the SBC is in fact optimal even when the independence assumption is grossly violated, and is thus applicable to a much broader range of domains than previously thought. This is essentially due to the fact that in many cases Eq. 1 may produce poor probability estimates, but the correct class will

still have the highest estimate, leading to correct classification.

After the empirical section, we begin by showing a simple example that illustrates some of the key points to be made. The following section contains the fundamental result of the paper: a derivation of necessary and sufficient conditions for the local optimality of the SBC (i.e., its optimality for any given example). This result is then generalized to a necessary and sufficient condition for the SBC's global optimality (i.e., its optimality for any given dataset). Finally, we show some fundamental limitations of the SBC, and that it is optimal for learning conjunctions and disjunctions.

2 EMPIRICAL EVIDENCE

In order to investigate the SBC's performance compared to that of other classifiers, and relate it to the degree of attribute dependence in the data, an empirical study was carried out on a large and varied selection of datasets from the UCI repository (Murphy & Aha, 1995). For the SBC, numeric values were discretized into ten equal-length intervals (or one per observed value, whichever was least). This has been found to give good results, more so than assuming normal distributions, as is often done in the pattern recognition literature (Dougherty, Kohavi & Sahami, 1995). Missing values were treated as having the value "?", at both training and testing times. This avoids losing potentially useful information. (For example, in medical domains, missing values often indicate that the doctors considered the corresponding tests unnecessary.) Null attribute probabilities $P(v_j|C_i)$ were replaced by $P(C_i)/e$, where e is the number of training examples, as done in (Clark & Niblett, 1989) and elsewhere.

The SBC was compared with state-of-the art representatives of three major approaches to classification learning: decision tree induction (C4.5 (Quinlan, 1993)), instance-based learning (PEBLS, (Cost & Salzberg, 1993)) and rule induction (CN2, (Clark & Boswell, 1991)). The default classifier, which assigns the most frequent class to all test examples, was also included. Twenty runs were conducted for each dataset, randomly selecting $\frac{2}{3}$ of the data for training and the remainder for testing. The accuracies obtained are shown in Table 1.

The results are summarized in Table 2. The first line shows the number of domains in which the SBC was more accurate than the corresponding classifier, versus the number in which it was less. For example, the SBC was more accurate than C4.5 in 16 domains, and less in 12. The second line considers only those domains where the accuracy difference was significant at the 5% level, using a one-tailed paired t test. For example, the SBC was significantly more accurate than C4.5 in 12 datasets. According to both these measures, the SBC wins out over each of the other approaches (with the exception of number of significant wins vs. PEBLS). The last line shows the average rank of each algorithm, computed by, in each domain, assigning rank 1 to the most accurate algorithm, rank 2 to the second best, and so on (including the default classifier). The SBC is the best-ranked of all algorithms, indicating that when it does not win it tends to be the second best.

Overall the SBC is quite competitive with the other approaches. This is a remarkably good result for such a simple and apparently limited classifier. However, it can be due to the datasets themselves representing "easy" concepts (Holte, 1993), and does not by itself disprove the notion that the SBC relies on the assumption of attribute independence. To investigate this, we need to measure the degree of attribute dependence in the data in some way. Measuring high-order dependencies is difficult, because the relevant probabilities are apt to be very small, and not reliably represented in the data. However, a first and feasible approach consists in measuring pairwise dependencies (i.e., dependencies between pairs of attributes given the class). Given attributes A_m and A_n and the class variable C, a possible measure of the degree of pairwise dependence between A_m and A_n given C is (Wan & Wong, 1989; Kononenko, 1991):

$$D(A_m, A_n|C) = H(A_m|C) + H(A_n|C) - H(A_m A_n|C) \quad (2)$$

where $A_m A_n$ represents the Cartesian product of attributes A_m and A_n (i.e., a derived attribute with one possible value corresponding to each combination of values of A_m and A_n), and for all classes i and attribute values k:

$$H(A_j|C) = \sum_i P(C_i) \sum_k -P(C_i \wedge A_j = v_{jk}) \log_2 P(C_i \wedge A_j = v_{jk}) \quad (3)$$

$D(A_m, A_n|C)$ is 0 when A_m and A_n are completely independent given C, and increases with their degree of dependence, with the maximum occurring when the class and one attribute completely determine the other.

D was computed for all classes and attribute pairs in each dataset, using uniform discretization as before, ignoring missing values, and excluding pairings of an attribute with itself. The results are shown in Table 3.[1] For comparison purposes, the first column shows the

[1] The annealing and audiology domains are omitted because some of the relevant entropies $H(A_m A_n|C)$ could not be computed.

Table 1: Empirical results: average accuracies and standard deviations. Superscripts denote significance levels for the difference in accuracy between the SBC and the corresponding algorithm, using a one-tailed paired t test: 1 is 0.005, 2 is 0.01, 3 is 0.025, 4 is 0.05, 5 is 0.1, and 6 is above 0.1.

Domain	SBC	Default	C4.5	PEBLS	CN2
Audiology	73.9±5.3	21.3±2.6 [1]	72.5±5.8 [6]	75.8±5.4 [4]	71.0±5.1 [2]
Annealing	93.5±2.7	76.4±1.8 [1]	91.3±2.3 [3]	98.7±0.9 [1]	81.2±5.4 [1]
Breast cancer	68.7±5.4	67.6±7.6 [6]	70.1±5.6 [4]	65.8±4.7 [3]	67.9±7.1 [6]
Credit screening	85.2±1.7	57.4±3.8 [1]	85.0±2.0 [6]	81.3±2.0 [1]	82.0±2.2 [1]
Chess endgames	88.0±1.4	52.0±1.9 [1]	99.2±0.1 [1]	96.9±0.7 [1]	98.1±1.0 [1]
Pima diabetes	74.4±3.0	66.0±2.3 [1]	72.4±2.8 [4]	71.4±2.4 [1]	73.8±2.7 [6]
Echocardiogram	66.7±7.4	67.8±6.6 [6]	65.8±6.2 [6]	64.1±6.1 [5]	68.2±7.2 [6]
Glass	50.4±15.9	31.7±5.5 [1]	66.1±8.4 [1]	65.8±7.3 [1]	63.8±5.5 [1]
Heart disease	83.1±3.2	55.0±3.4 [1]	74.2±4.2 [1]	79.2±3.8 [1]	79.7±2.9 [1]
Hepatitis	81.2±3.7	78.1±3.1 [2]	78.7±4.7 [4]	79.9±6.6 [6]	80.3±4.2 [6]
Horse colic	77.8±4.2	63.6±3.9 [1]	83.6±4.1 [1]	76.3±4.4 [5]	82.5±4.2 [1]
Thyroid disease	97.3±0.7	95.3±0.6 [1]	99.1±0.2 [1]	97.3±0.4 [6]	98.8±0.4 [1]
Iris	89.0±12.8	26.5±5.2 [1]	93.4±2.4 [5]	91.7±3.7 [6]	93.3±3.6 [5]
Labor neg.	92.6±7.9	65.0±9.5 [1]	79.7±7.1 [1]	91.6±4.3 [6]	82.1±6.9 [1]
Lung cancer	46.4±14.7	26.8±12.3[1]	40.9±16.3[6]	42.3±17.3[6]	38.6±13.5[4]
Liver disease	61.8±6.9	58.1±3.4 [3]	63.7±4.3 [6]	60.1±3.6 [6]	65.0±3.8 [4]
LED	66.8±5.9	8.0±2.7 [1]	61.2±8.4 [2]	55.3±6.1 [1]	58.6±8.1 [1]
Lymphography	81.5±5.6	57.3±5.4 [1]	75.3±4.8 [1]	82.9±5.6 [6]	78.8±4.9 [3]
Post-operative	61.8±9.8	71.2±5.2 [1]	70.2±4.9 [1]	58.8±8.1 [6]	60.8±8.2 [6]
Promoters	87.6±6.0	43.1±4.2 [1]	74.3±7.8 [1]	91.7±5.9 [1]	75.9±8.8 [1]
Primary tumor	44.9±5.4	24.6±3.2 [1]	35.9±5.8 [1]	30.9±4.7 [1]	39.8±5.2 [1]
Solar flare	68.0±3.1	25.2±4.4 [1]	70.6±2.9 [1]	67.6±3.5 [6]	70.4±3.0 [1]
Sonar	24.1±8.7	50.8±7.6 [1]	64.7±7.2 [1]	73.3±7.5 [1]	66.2±7.5 [1]
Soybean	100.0±0.0	30.0±14.3[1]	95.0±9.0 [3]	100.0±0.0 [6]	96.9±5.9 [3]
Splice junctions	95.4±0.6	52.4±1.6 [1]	93.4±0.8 [1]	94.3±0.5 [1]	81.5±5.5 [1]
Voting records	91.2±1.6	60.5±3.1 [1]	96.3±1.3 [1]	94.9±1.2 [1]	95.8±1.6 [1]
Wine	90.9±13.3	36.4±9.9 [1]	91.7±5.6 [6]	96.9±2.2 [4]	90.8±4.7 [6]
Zoology	91.9±3.6	39.4±6.4 [1]	89.6±4.7 [1]	94.6±4.3 [1]	90.6±5.0 [5]

SBC's rank in each domain (i.e., 1 if it was the most accurate algorithm, 2 if it was the second most accurate, etc.) The second column shows the maximum value of D observed in the dataset. The third column shows the percentage of all attributes which exhibited a degree of dependence with some other attribute of at least 0.2.[2] The fourth column shows the average D for all attribute pairs in the dataset.

This table leads to two important observations. One is that the SBC achieves higher accuracy than more sophisticated approaches in many domains where there is substantial attribute dependence, and therefore the reason for its good comparative performance is not that there are no attribute dependences in the data. The other is that the correlation between the average degree of attribute dependence and the difference in accuracy between the SBC and other algorithms is small ($R^2 = 0.13$ for C4.5, 0.27 for PEBLS, and 0.19 for CN2), and therefore attribute dependence is not a good predictor of the SBC's differential performance vs. approaches that can take it into account. Thus the SBC's surprisingly good performance remains unexplained. In the remainder of this paper we begin to shed some light on this matter.

[2]This value is commonly used as a threshold above which attributes are considered to be significantly dependent.

Table 2. Summary of accuracy results.

Measure	SBC	C4.5	PEBLS	CN2
No. wins	-	16-12	15-11	18-10
No. sig. wins	-	12-9	7-9	12-8
Rank	2.32	2.54	2.79	2.68

Table 3: Empirical measures of attribute dependence.

Domain	Rank	D_{Max}	% Hi.	D_{Avg}
Breast cancer	2	0.548	66.7	0.093
Credit screening	1	0.790	46.7	0.060
Chess endgames	4	0.383	25.0	0.015
Pima diabetes	1	0.483	62.5	0.146
Echocardiogram	3	0.769	85.7	0.360
Glass	4	0.836	100.0	0.363
Heart disease	1	0.388	53.8	0.085
Hepatitis	1	0.589	52.6	0.089
Horse colic	3	0.510	95.5	0.157
Thyroid disease	3	0.516	44.0	0.054
Iris	4	0.731	100.0	0.469
Labor neg.	1	1.189	100.0	0.449
Lung cancer	1	1.226	98.2	0.165
Liver disease	3	0.513	100.0	0.243
LED	1	0.060	0.0	0.025
Lymphography	2	0.410	55.6	0.076
Post-operative	3	0.181	0.0	0.065
Promoters	2	0.394	98.2	0.149
Primary tumor	1	0.098	0.0	0.023
Solar flare	3	0.216	16.7	0.041
Sonar	5	1.471	100.0	0.491
Soybean	1	0.726	31.4	0.016
Splice junctions	1	0.084	0.0	0.017
Voting records	4	0.316	25.0	0.052
Wine	3	0.733	100.0	0.459
Zoology	2	0.150	0.0	0.021

3 AN EXAMPLE

Consider a Boolean concept, described by three attributes A, B and C. Let the two classes be + and −, and equiprobable ($P(+) = P(-) = \frac{1}{2}$). Given an example E, let $P(A|+)$ be a shorthand for $P(A = a_E|+)$, a_E being the value of attribute A in the instance, and similarly for the other attributes. Let A and C be independent, and let $A = B$ (i.e., A and B are completely dependent). B should therefore be ignored, and the optimal classification procedure for a test instance is to assign it to class + if $P(A|+)P(C|+)-P(A|-)P(C|-) > 0$, and to class − if the inequality has the opposite sign (the classification is arbitrary if the two sides are equal). On the other hand, the SBC will take B into account as if it was independent from A, and this will be equivalent to counting A twice. Thus, the SBC will assign the instance to class + if $P(A|+)^2P(C|+) - P(A|-)^2P(C|-) > 0$, and to − otherwise.

Applying Bayes' theorem, $P(A|+)$ can be re-expressed as $P(A)P(+|A)/P(+)$, and similarly for the other probabilities. Since $P(+) = P(-)$, after canceling like terms this leads to the equivalent expressions

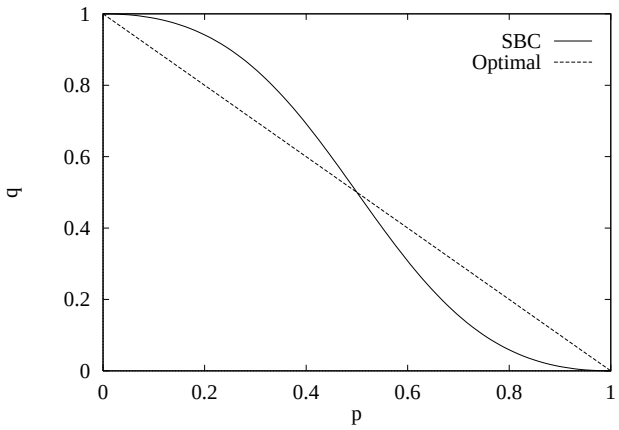

Figure 1: Decision boundaries for the SBC and the optimal classifier.

$P(+|A)P(+|C) - P(-|A)P(-|C) > 0$ for the optimal decision, and $P(+|A)^2P(+|C) - P(-|A)^2P(-|C) > 0$ for the SBC. Let $P(+|A) = p$, and $P(+|C) = q$. Then class + should win when $pq-(1-p)(1-q) > 0$, which is equivalent to $q > 1 - p$. With the SBC, it will win when $p^2q - (1-p)^2(1-q) > 0$, which is equivalent to $q > \frac{(1-p)^2}{p^2+(1-p)^2}$. The two curves are shown in Fig. 1. The remarkable fact is that, even though the independence assumption is decisively violated because $B = A$, the SBC disagrees with the optimal procedure only in the two narrow regions that are above one of the curves and below the other; everywhere else it performs the correct classification. Thus, for all problems where (p, q) does not fall in those two small regions, the SBC is effectively optimal. By contrast, according to the independence assumption it should be optimal only when the two expressions are identical, i.e. at the three isolated points where the curves cross: (0, 1), $(\frac{1}{2}, \frac{1}{2})$ and (1, 0). This shows that the SBC's range of applicability may in fact be much broader than previously thought. In the next section we examine the general case and formalize this result.

4 LOCAL OPTIMALITY

We begin with some necessary definitions.

Definition 1 *The* Bayes rate *for an example is the lowest error rate achievable by any classifier on that example (Duda & Hart, 1973).*

Definition 2 *A classifier is* locally optimal *for a given example iff its error rate on that example is equal to the Bayes rate.*

Definition 3 *A classifier is* globally optimal *for a given sample (dataset) iff it is locally optimal for every example in that sample. A classifier is globally optimal for a given problem (domain) iff it is globally optimal for all possible samples of that problem (i.e., for all datasets extracted from that domain).*

Consider the two-class case in general. Let the classes be + and − as before, $p = P(+|E)$, $r = [P(+)/P(E)] \prod_{j=1}^{a} P(v_j|+)$, and $s = [P(-)/P(E)] \prod_{j=1}^{a} P(v_j|-)$ (refer to Eq.1). In this section we will derive a necessary and sufficient condition for the local optimality of the SBC, and show that the volume of the SBC's region of optimality in the space of valid values of (p, r, s) is half of the total volume of this space.

The key to these results lies in the distinction between classification and probability estimation. Equation 1 yields a correct estimate of the class probabilities only when the independence assumption holds; but for purposes of classification, the class probability estimates can diverge widely from the true values, as long as the maximum estimate still corresponds to the maximum true probability. For example, suppose there are two classes + and −, and let $P(+|E) = 0.51$ and $P(-|E) = 0.49$ be the true class probabilities given example E. The optimal decision is then to assign E to class +. Suppose also that Equation 1 gives the estimates $\hat{P}(+|E) = 0.99$ and $\hat{P}(-|E) = 0.01$. The independence assumption is violated by a wide margin, and yet the SBC still makes the optimal decision.

Theorem 1 *The SBC is locally optimal for an example E iff $(p \geq \frac{1}{2} \wedge r \geq s) \vee (p \leq \frac{1}{2} \wedge r \leq s)$ for E.*

Proof. The SBC is optimal when its error rate is the minimum possible. When $p = P(+|E) > \frac{1}{2}$, the minimum error is $1 - p$, and is obtained by assigning E to class +. The SBC assigns E to class + when $P(+|E) > P(-|E)$ according to Eq. 1, i.e., when $r > s$. Thus if $p > \frac{1}{2} \wedge r > s$ the SBC is optimal. Conversely, when $p = P(+|E) < \frac{1}{2}$, the minimum error is p, it is obtained by assigning E to class −, and the SBC does this when $r < s$. Thus the SBC is optimal when $p < \frac{1}{2} \wedge r < s$. When $p = \frac{1}{2}$, either decision is optimal, so the inequalities can be generalized as shown. □

Corollary 1 *The SBC is locally optimal in half the volume of the space of possible values of (p, r, s).*

Proof. Since p, r and s are probabilities, (p, r, s) only takes values in the unit cube $[0, 1]^3$. The region of this cube satisfying the condition in Theorem 1 is shown shaded in Fig. 2; it can easily be seen to occupy half of the total volume of the cube. However, not all (r, s) pairs correspond to valid probability combinations. Since p is unconstrained, the projection of the space U of valid probability combinations on all planes $p = p_0$ is the same. By Theorem 1, the region of optimality on planes below $p_0 = \frac{1}{2}$ becomes the region of nonoptimality on planes above $p_0 = \frac{1}{2}$, and vice-versa (i.e., the optimal region for projections below $p_0 = \frac{1}{2}$ is the photographic negative of the optimal region for projections above). Thus, if S is the area of U's projection and S_O is the area of the optimal region for $p_0 < \frac{1}{2}$, the area of the optimal region for $p_0 > \frac{1}{2}$ is $S - S_O$, and the total volume of the region of optimality is $\frac{1}{2}S_O + \frac{1}{2}(S - S_O) = \frac{1}{2}S$. (Also, since if (r, s) is a valid probability combination then so is (s, r), the region of optimality is symmetric about $s = r$, and therefore $S_O = \frac{1}{2}S$ both above and below $p_0 = \frac{1}{2}$.) □

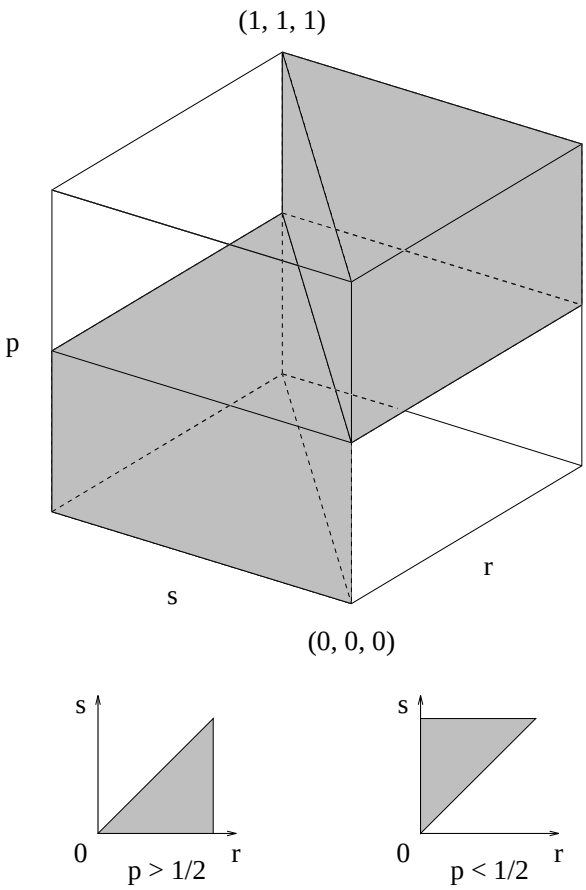

Figure 2. Region of optimality of the SBC.

In contrast, by the independence assumption the SBC would be optimal only on the line where the planes $r = p$ and $s = 1 - p$ intersect. Thus the previously assumed region of optimality of the SBC is a second-order infinitesimal fraction of the actual one.

5 GLOBAL OPTIMALITY

The extension of Theorem 1 to global optimality is immediate. Let p, r and s for example E be indexed as p_E, r_E and s_E.

Theorem 2 *The SBC is globally optimal for a sample (dataset)* Σ *iff* $\forall E \in \Sigma\ (p_E \geq \frac{1}{2} \wedge r_E \geq s_E) \vee (p_E \leq \frac{1}{2} \wedge r_E \leq s_E)$.

Proof. By Definition 3 and Theorem 1. □

However, verifying this condition directly on a test sample will in general not be possible, since it involves finding the true class probabilities for all examples in the sample. Further, verifying it for a given domain (i.e, for all possible samples extracted from that domain) will in general involve a computation of size proportional to the number of possible examples, and therefore exponential in the number of attributes and computationally infeasible. Thus the remainder of this section is dedicated to investigating more concrete conditions for the optimality of the SBC, some necessary and some sufficient.

5.1 NECESSARY CONDITIONS

Let a be the number of attributes, as before, let c be the number of classes, let v be the maximum number of values per attribute, and let d be the number of different numbers representable on the machine implementing the SBC. For example, if numbers are represented using 16 bits, $d = 2^{16} = 65536$.

Theorem 3 *The SBC cannot be globally optimal for more than* $d^{c(av+1)}$ *different problems.*

Proof. Since the SBC's state is composed of $c(av + 1)$ probabilities, and each probability can only have d different values, the SBC can only be in at most $d^{c(av+1)}$ states, and thus it cannot distinguish between more than this number of concepts. □

Even though $d^{c(av+1)}$ can be very large, this is a significant restriction because many concept classes have size doubly exponential in a (e.g., arbitrary DNF formulas in Boolean domains), and due to the extremely rapid growth of this function the SBC's capacity will be exceeded even for commonly-occurring values of a. On the other hand, this restriction is compatible with concept classes whose size grows only exponentially with a (e.g., conjunctions).

This result reflects the SBC's limited information storage capacity, and should be contrasted with the case of classifiers (like instance-based, rule and decision tree learners) whose memory size can be proportional to the sample size. It also shows that the condition in Theorem 2 is satisfied by an exponentially decreasing fraction of all possible domains as a increases. This is consistent with the fact that local optimality has to be verified for every possible combination of attribute values if the SBC is to be globally optimal for a domain (Definition 3), and the probability of this decreases exponentially with a, starting at 100% for $a = 1$. However, a similar statement is true for other learners; it simply reflects the fact that it is very difficult to optimally learn a very wide class of concepts. The SBC's information storage capacity is $O(a)$. If e is the training set size, learners that can memorize all the individual examples (or the equivalent) have a storage capacity of $O(ea)$, and therefore have the ability in principle to converge to optimal when $e \to \infty$. However, for any finite e there is a value of a after which the fraction of problems on which those learners can be optimal also starts to decrease exponentially with a.

Theorem 4 *In symbolic domains, the SBC is globally optimal only for linearly separable problems.*

Proof. Define one Boolean feature b_{jk} for each attribute value, i.e., $b_{jk} = 1$ if $A_j = v_{jk}$ and 0 otherwise, where v_{jk} is the kth value of attribute A_j. Then, by taking the logarithm of Eq. 1, the SBC is equivalent to a linear machine (Duda & Hart, 1973) whose discriminant function for class C_i is $\log P(C_i) + \sum_{j,k} \log P(A_j = v_{jk}|C_i)\, b_{jk}$ (i.e., the weight of each Boolean feature is the log-probability of the corresponding attribute value given the class). □

This is not a sufficient condition, because the SBC cannot learn some linearly separable concepts. For example, it narrowly fails on the concept 3-of-7 (i.e., the concept composed of examples where at least 3 of 7 Boolean attributes are true) (Kohavi, 1995). Thus in Boolean domains the SBC's range of optimality is a subset of the perceptron's (Duda & Hart, 1973). However, in numeric domains the SBC is not restricted to linearly separable problems; for example, if classes are normally distributed, nonlinear boundaries and multiple disconnected regions can arise, and the SBC is able to identify them (see (Duda & Hart, 1973)).

5.2 SUFFICIENT CONDITIONS

In this section we establish the SBC's optimality for some common concept classes.

Theorem 5 *The SBC is globally optimal if, for all classes* C_i *and examples* $E = (v_1, v_2, \ldots, v_a)$, $P(E|C_i) = \prod_{j=1}^{a} P(v_j|C_i)$.

This result is restated here for completeness. The crucial point is that this condition is sufficient, but not necessary.

Theorem 6 *The SBC is globally optimal for learning conjunctions of literals.*

Proof. Suppose there are n literals L_j in the conjunction. A literal may be a Boolean attribute or its negation. In addition, there may be $a - n$ irrelevant attributes; they simply cause each line in the truth table to become 2^{a-n} lines with the same values for the class and all relevant attributes, each of those lines corresponding to a possible combination of the irrelevant attributes. For simplicity, they will be ignored from here on (i.e., $n = a$ will be assumed without loss of generality). Recall that, in the truth table for conjunction, the class C is 0 (false) for all but $L_0 = L_1 = \cdots = L_n = 1$ (true). Thus, using a bar to denote negation, $P(C) = \frac{1}{2^n}$, $P(\overline{C}) = \frac{2^n-1}{2^n}$, $P(L_j|C) = 1$, $P(\overline{L_j}|C) = 0$, $P(\overline{L_j}|\overline{C}) = \frac{2^{n-1}}{2^n-1}$ (the number of times the literal is 0 in the truth table, over the number of times the class is 0), and $P(L_j|\overline{C}) = \frac{2^{n-1}-1}{2^n-1}$ (the number of times the literal is 1 minus the one time it corresponds to C, over the number of times the class is 0). Let E be an arbitrary example, and let m of the conjunction's literals be true in E. For simplicity, the factor $1/P(E)$ will be omitted from all probabilities. Then:

$$\begin{aligned} P(C|E) &= P(C)\,P^m(L_j|C)\,P^{n-m}(\overline{L_j}|C) \\ &= \begin{cases} \frac{1}{2^n} & \text{if } n = m \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

$$\begin{aligned} P(\overline{C}|E) &= P(\overline{C})\,P^m(L_j|\overline{C})\,P^{n-m}(\overline{L_j}|\overline{C}) \\ &= \frac{2^n-1}{2^n}\left(\frac{2^{n-1}-1}{2^n-1}\right)^m\left(\frac{2^{n-1}}{2^n-1}\right)^{n-m} \end{aligned}$$

Notice that $\frac{2^{n-1}-1}{2^n-1} < \frac{1}{2}$ for all n. Thus, for $m = n$, $P(\overline{C}|E) = P(\overline{C})\left(\frac{2^{n-1}-1}{2^n-1}\right)^n < P(\overline{C})(\frac{1}{2})^n < \frac{1}{2^n} = P(C|E)$, and class 1 wins. For all $m < n$, $P(C|E) = 0$ and $P(\overline{C}|E) > 0$, and thus class 0 wins. Therefore the SBC always makes the correct decision, i.e., it is globally optimal. □

Notice that conjunctive concepts verify the independence assumption for class 1, but not for class 0. (For example, if $C = A_0 \wedge A_1$, $P(A_1|\overline{C}) = \frac{1}{3} \neq P(A_1|\overline{C}, A_0) = 0$, by inspection of the truth table.) Thus conjunctions are an example of a class of concepts where the SBC is in fact optimal, but would not be if it required attribute independence.

Theorem 7 *The SBC is globally optimal for learning disjunctions of literals.*

Proof. Similar to that for Theorem 6, letting m be the number of the disjunction's literals that are false in E. □

Conversely, disjunctions verify the independence assumption for class 0, but not for class 1, and are another example of the SBC's optimality even when the independence assumption is violated.

As corollaries, the SBC is also optimal for negated conjunctions and negated disjunctions, and for the identity and negation functions, with any number of irrelevant attributes.

6 CONCLUSIONS AND FUTURE WORK

In this paper we verified that the SBC performs quite well in practice even when strong attribute dependences are present, and showed that this is at least in part due to the fact that, contrary to previous assumptions, the SBC does not depend on attribute independence to be optimal. We then derived a number of necessary and a number of sufficient conditions for the SBC's optimality. In particular, we showed that the SBC is an optimal learner for conjunctive and disjunctive concepts, even though these violate the independence assumption.

Ideally, we would like to have a complete set of necessary and sufficient conditions for the optimality of the SBC, efficiently verifiable on real problems. In the previous section we began the work towards this goal. Also, even if the SBC is optimal in the limit (i.e., given an infinite sample), other classifiers may converge faster to the Bayes rate for certain problems. Thus, investigating the behavior of the SBC when the probability estimates it employs are imperfect due to the finiteness of the sample is also of interest. Another important area of future research concerns finding conditions under which the SBC is not optimal, but comes very close to being so, for example because it makes the wrong prediction on only a small fraction of the examples. Even when it is not optimal, the SBC will perform well relative to other algorithms as long as it is closer to the optimum than they are. This may explain some of the results in the empirical section.

In summary, the work reported here demonstrates that the SBC has a much greater range of applicability than previously thought. Since it also has advantages in terms of learning speed, classification speed, storage space and incrementality, this suggests that its use should be considered more often.

Acknowledgements

This work was partly supported by a JNICT/PRAXIS XXI scholarship. The authors are grateful to all those who provided the datasets used in the empirical study.

References

Clark, P. & Boswell, R. (1991). Rule induction with CN2: Some recent improvements. In *Proceedings of the Sixth European Working Session on Learning*, (pp. 151–163), Porto, Portugal. Springer-Verlag.

Clark, P. & Niblett, T. (1989). The CN2 induction algorithm. *Machine Learning*, *3*, 261–283.

Cost, S. & Salzberg, S. (1993). A weighted nearest neighbor algorithm for learning with symbolic features. *Machine Learning*, *10*, 57–78.

Dougherty, J, Kohavi, R, & Sahami, M. (1995). Supervised and unsupervised discretization of continuous features. In *Proceedings of the Twelfth International Conference on Machine Learning*, (pp. 194–202), Tahoe City, CA. Morgan Kaufmann.

Duda, R. O. & Hart, P. E. (1973). *Pattern Classification and Scene Analysis*. New York, NY: Wiley.

Holte, R. C. (1993). Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, *11*, 63–91.

Kohavi, R. (1995). *Wrappers for Performance Enhancement and Oblivious Decision Graphs*. PhD thesis, Department of Computer Science, Stanford University, Stanford, CA.

Kononenko, I. (1991). Semi-naive Bayesian classifier. In *Proceedings of the Sixth European Working Session on Learning*, (pp. 206–219), Porto, Portugal. Springer-Verlag.

Langley, P. (1993). Induction of recursive Bayesian classifiers. In *Proceedings of the Eighth European Conference on Machine Learning*, (pp. 153–164), Vienna, Austria. Springer-Verlag.

Langley, P, Iba, W, & Thompson, K. (1992). An analysis of Bayesian classifiers. In *Proceedings of the Tenth National Conference on Artificial Intelligence*, (pp. 223–228), San Jose, CA. AAAI Press.

Langley, P. & Sage, S. (1994). Induction of selective Bayesian classifiers. In *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*, (pp. 399–406), Seattle, WA. Morgan Kaufmann.

Murphy, P. M. & Aha, D. W. (1995). UCI repository of machine learning databases. Machine-readable data repository, Department of Information and Computer Science, University of California at Irvine, Irvine, CA.

Pazzani, M. (1995). Searching for attribute dependencies in Bayesian classifiers. In *Preliminary Papers of the Fifth International Workshop on Artificial Intelligence and Statistics*, (pp. 424–429), Fort Lauderdale, FL. Society for Artificial Intelligence and Statistics.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann.

Rachlin, J, Kasif, S, Salzberg, S, & Aha, D. W. (1994). Towards a better understanding of memory-based reasoning systems. In *Proceedings of the Eleventh International Conference on Machine Learning*, (pp. 242–250), New Brunswick, NJ. Morgan Kaufmann.

Wan, S. J. & Wong, S. K. M. (1989). A measure for concept dissimilarity and its applications in machine learning. In *Proceedings of the International Conference on Computing and Information*, (pp. 267–273), Toronto, Canada. North-Holland.