Bayesian Classification with Correlation and Inheritance

Robin Hanson
Sterling Software

John Stutz
NASA

Peter Cheeseman
RIACS*

Artificial Intelligence Research Branch
NASA Ames Research Center, Mail Stop 244-17
Moffett Field, CA 94035, USA
Internet: <last-name>@ptolemy.arc.nasa.gov

Abstract

The task of inferring a set of classes and class descriptions most likely to explain a given data set can be placed on a firm theoretical foundation using Bayesian statistics. Within this framework, and using various mathematical and algorithmic approximations, the AutoClass system searches for the most probable classifications, automatically choosing the number of classes and complexity of class descriptions. Simpler versions of AutoClass have been applied to many large real data sets, have discovered new independently-verified phenomena, and have been released as a robust software package. Recent extensions allow attributes to be selectively correlated within particular classes, and allow classes to inherit, or share, model parameters though a class hierarchy.

1 Introduction

The task of *supervised* classification - i.e., learning to predict class memberships of test cases given labeled training cases - is a familiar machine learning problem. A related problem is *unsupervised* classification, where training cases are also unlabeled. Here one tries to predict all features of new cases; the best classification is the least "surprised" by new cases. This type of classification, related to clustering, is often very useful in exploratory data analysis, where one has few preconceptions about what structures new data may hold.

We have previously developed and reported on AutoClass [Cheeseman *et al.,* 1988a; Cheeseman *et* a/., 1988b], an unsupervised classification system based on Bayesian theory. Rather than just partitioning cases, as most clustering techniques do, the Bayesian approach searches in a model space for the "best" class descriptions. A best classification optimally trades off predictive accuracy against the complexity of the classes, and so does not "overfit" the data. Such classes are also "fuzzy"; instead of each case being assigned to a class, a case has a probability of being a member of each of the different classes.

Autocla88 III, the most recent released version, combines real and discrete data, allows some data to be missing, and automatically chooses the number of classes from first principles. Extensive testing has indicated that it generally produces significant and useful results, but is primarily limited by the simplicity of the models it uses, rather than, for example, inadequate search heuristics. AutoClass III assumes that all attributes are relevant, that they are independent of each other within each class, and that classes are mutually exclusive. Recent extensions, embodied in Autoclass IV, let us relax two of these assumptions, allowing attributes to be selectively correlated and to have more or less relevance via a class hierarchy.

We begin by describing the Bayesian theory of learning, and then apply it to increasingly complex classification problems, from various single class models up to hierarchical class mixtures. Finally, we report empirical results from an implementation of these extensions.

2 Bayesian Learning

Bayesian theory gives a mathematical calculus of degrees of belief, describing what it means for beliefs to be consistent and how they should change with evidence. This section briefly reviews that theory, describes an approach to making it tractable, and comments on the resulting tradeoffs. In general, a Bayesian agent uses a single real number to describe its degree of belief in each proposition of interest.

2.1 Theory

Let *E* denote some evidence that is known or could potentially be known to an agent; let *H* denote a hypothesis specifying that the world is in some particular state; and let the sets of possible evidence *E* and possible states of the world *H* each be mutually exclusive and exhaustive sets.

In general, *P(ab\cd)* denotes a real number describing an agent's degree of belief in the conjunction of propositions *a* and 6, conditional on the assumption that propositions *c* and *d* are true. More specifically, $\pi(H)$ is a "prior" describing the agent's belief in *H before,* or in the absence of, seeing evidence *E,* $\pi(H|E)$ is a "posterior" describing the agent's belief *after* observing some particular evidence E, and *L(E\H)* is a "likelihood" em-

•Research Institute for Advanced Computer Science

bodying the agent's theory of how likely it would be to see each possible evidence combination E in each possible world H.

To be consistent, beliefs must be non-negative, $P(a|b) \geq 0$, and normalized, so that $\sum_H \pi(H) = 1$ and $\sum_E L(E|H) = 1$. The likelihood and the prior together give a "joint" probability $J(EH) \equiv L(E|H)\pi(H)$ of both E and H. Normalizing the joint gives Bayes' rule, which tells how beliefs should change with evidence.

$$\pi(H|E) = \frac{J(EH)}{\sum_H J(EH)} = \frac{L(E|H)\pi(H)}{\sum_H L(E|H)\pi(H)}$$

When the set of possible Hs is continuous, the prior $\pi(H)$ becomes a differential $d\pi(H)$, and the sums over H are replaced by integrals. Similarly, continuous Es have a differential likelihood $dL(E|H)$, though any real evidence ΔE will have a finite probability $\Delta L(E|H) \approx dL(E|H)\frac{\Delta E}{dE}$.

In theory, all an agent needs to do in any given situation is to choose a set of states H, an associated likelihood function describing what evidence is expected to be observed in those states, a set of prior expectations on the states, and then collect some evidence. Bayes' rule then specifies the appropriate posterior beliefs about the state of the world, which can be used to answer most questions of interest.

2.2 Practice

In practice this theory can be difficult to apply, as the sums and integrals involved are often mathematically intractable. Here is our approach.

Rather than consider all possible *states* of the world, we focus on some smaller space of *models,* and do all of our analysis conditional on an assumption *S* that the world really is described by one of the models in our space. This assumption is almost certainly false, but it makes the analysis tractable.

The parameters which specify a particular model are split into two sets. First, a set of discrete parameters *T* describe the general form of the model, usually by specifying some functional form for the likelihood function. For example, *T* might specify whether two variables are correlated or not, or how many classes are present in a classification. Second, free variables in this general form, such as the magnitude of the correlation or the relative sizes of the classes, constitute the remaining continuous model parameters *V.*

We generally prefer a likelihood[1] $L(E|VTS)$ which is mathematically simple and yet still embodies the kinds of complexity relevant in some context.

Similarly, we prefer a simple prior distribution *dπ(VT\S)* over the model space, allowing the resulting *V* integrals, described below, to be at least approximated. We also usually prefer a relatively broad and uninformative prior, and one that gives nearly equal weight to different levels of model complexity, resulting in a "significance test". Adding more parameters to a model then induces a cost, which must be paid for by a significantly better fit to the data before the more complex model can be preferred.

The joint can now be written as $dJ(EVT|S) = L(E|VTS)\,d\pi(VT|S)$ and, for a reasonably-complex problem, is usually a very rugged distribution in VT, with an immense number of sharp peaks distributed widely over a huge high-dimensional space. Because of this we despair of directly normalizing the joint, as required by Bayes' rule, or of communicating the detailed shape of the posterior distribution.

Instead we break the continuous *V* space into regions *R* surrounding each sharp peak, and search until we tire for combinations *RT* for which the "marginal" joint

$$M(ERT|S) \equiv \int_{V\in R} dJ(EVT|S)$$

is as large as possible. The best few such "models" *RT* found are then reported, even though it is usually almost certain that more probable models remain to be found.

Each model *RT* is reported by describing its marginal joint *M{ERT\S),* its discrete parameters T, and estimates of typical values of *V* in the region R, such as the mean estimate of *V:*

$$\mathcal{E}(V|ERTS) \equiv \frac{\int_{V\in R} V\, dJ(EVT|S)}{M(ERT|S)}$$

or the *V* for which $dJ(EVT|S)$ is maximum in *R.* While these estimates are not invariant under reparameterizations of the *V* space, and hence depend on the syntax with which the likelihood was expressed, the peak is usually sharp enough that such differences don't matter.

A weighted average of the best few models found is used to make predictions. Almost all of the weight is usually in the best few, justifying the neglect of the rest.

Even though the sums and integrals can be difficult, and large spaces must be searched, Bayesian theory offers the advantages of being theoretically well-founded and empirically well-tested [Berger, 1985]; one can almost "turn the crank", modulo doing integrals and search[2], to deal with any new problem. Disadvantages include being forced to be explicit about the space of models one is searching in, and occasional ambiguities regarding what an appropriate prior is. Also, it is not clear how one can take the computational cost of doing a Bayesian analysis into account without a crippling infinite regress.

We will now illustrate this general approach by applying it to the problem of unsupervised classification.

3 Single Class Models

For all the models to be considered in this paper, the evidence *E* will consist of a set of I cases, an associated set *K* of attributes, of size[3] *K,* and case attribute values[4]

[1] *A* variable like *V* in a probability expression stands for the proposition that the variable has a particular value.

[2] The joint probability provides a good local evaluation function for searching though.

[3] We use script letters like *K* to denote sets, and matching ordinary letters *K* to denote their size.

[4] Nothing in principle prevents a Bayesian analysis of more complex model spaces for relational data.

X_{ik}, which can include "unknown."[5] For example, medical case number 8, described as *(age* = 23, *blood-type* — *A*, ...), would have $X_{8.1} = 23, X_{8.2} = A$, etc.

In this section and the next we will describe applications of Bayesian learning theory to various kinds of models which could explain this evidence, beginning with simple model spaces and building more complex spaces from them. We begin in this section with a single class. First, a single attribute is considered, then multiple independent attributes, then fully covariant attributes, and finally selective covariance. In the next section we combine these single classes into class mixtures, first flat then tree-based.

Space	*Description*	V	T	R
S_{D1}	Single Discrete	q_l		
S_{R1}	Single Real	$\mu\sigma$		
S_I	Independent Attrs	V_k		
S_D	Covariant Discrete	$q_{l_1 l_2 \ldots}$		
S_R	Covariant Real	$\mu_k \Sigma_{kk'}$		
S_V	Block Covariance	V_b	BK_b	
S_M	Flat Class Mixture	$\alpha_c V_c$	C	R
S_H	Tree Class Mixture	$\alpha_c V_c$	$J_c K_c T_c$	R

Table 1: Model Spaces Described

For each space *S* we will describe the continuous parameters *V,* any discrete model parameters T, normalized likelihoods *L(E\VTS),* and priors $\pi(VT|S)$ A8 most spaces have no discrete parameters T, and only one region R, we can usually ignore these parameters. Approximations to the resulting marginals *M(ERT\S)* and estimates $\mathcal{E}(V|ERTS)$ will be given, but not derived. These will often be given in terms of general functions *F* which are common to various models. As appropriate, comments will be made about algorithms and computational complexity.

All of the likelihood functions considered here assume the cases are independent, i.e., *L(E\VTS)* = II, *L(E_i\VTS)* so we need only give *L{E_i\VTS)* for each space, where $E_i \equiv \{X_{i1}, X_{i2}, \ldots, X_{iK}\}$.

4 Single Class Models

4.1 Single Discrete Attribute - *SD1*

A discrete attribute *k* allows only a finite number of possible values $l \in [1, 2, ..., L]$ for any X_i. A set of independent coin tosses, for example, might have *L* = 3 with /i = heads, l_2 = tails, and /3 = "unknown*. If we make the assumption *SD* that there is only one discrete attribute, then the only parameters are the continuous *V* = *qi... qL,* consisting of the likelihoods $L(X_i|VS_{D1}) = q_{(l=X_i)}$ for each possible value *l.* In the coin example, *q1* = .7 would say that the coin had a 70% of coming up heads each time.

There are only $L - 1$ free parameters since normalization requires $\sum_l q_l = 1$. For this likelihood, all that matters from the data for the total $L(E|VTS_{d1})$ are the number of cases with each value[6] $I_l = \sum_i \delta_{X_i l}$. In the coin example, I_1 would be the number of heads. Such sums are called "sufficient statistics" since they summarize all the information relevant to a model.

We choose a prior[7]

$$d\pi(V|S_{D1}) = dB(q_1 \ldots q_L|L) \equiv \frac{\Gamma(aL)}{\Gamma(a)^L} \prod_l q_l^{a-1} dq_l$$

which for $a > 0$ is a special case of a beta distribution [Berger, 1985]. a is a "hyperparameter" which can be set to different values to specify different priors. Here we set $a = 1/L$. This simple problem has only one maximum, whose marginal is given by

$$M(E|S_{D1}) = F_1(I_1, \ldots, I_L, I, L) \equiv \frac{\Gamma(aL) \prod_l \Gamma(I_l + a)}{\Gamma(aL + I)\Gamma(a)^L}$$

The prior above was chosen because it scales nicely, and to sirnplify the mean estimate of q_l

$$\mathcal{E}(q_l|ES_{D1}) = F_2(I_l, I, L) \equiv \frac{I_l + a}{I + aL} = \frac{I_l + \frac{1}{L}}{I + 1}$$

for $a = 1/L$. Using a hash table, these results can be computed in order I numerical steps, independent of L.

4.2 Single Real Attribute - S_{R1}

Real attribute values X_i specify a small range of the real line, with a center x_i and a precision, Δx_i, assumed to be much smaller than other scales of interest. For example, someone's weight might be measured as 70±1 kilograms.

For S_{R1}, where there is only one real attribute, we assume the likelihood is a standard normal distribution, where the sufficient statistics are the data mean $\bar{x} = \frac{1}{I}\sum_i^I x_i$, the geometric mean precision $\widehat{\Delta x} = (\prod_i^I \Delta x_i)^{\frac{1}{I}}$ and the standard deviation s given by $s^2 = \frac{1}{I}\sum_i (x_i - \bar{x})^2$. V consists of a model mean μ and standard deviation σ. For example, people's weight might be distributed[8] with a mean of 80 kilograms and a deviation of 15.

For brevity we here give only the resulting marginal

$$M(E|S_{R1}) = \frac{\sqrt{\pi}}{2} \frac{\Gamma(\frac{I-1}{2})}{(\pi I)^{\frac{1}{2}}} \frac{1}{log(\Delta\mu/\min \Delta x_i)} \frac{\widehat{\Delta x}^I}{s^{I-1}\Delta\mu},$$

where $\Delta\mu = \max x_i - \min x_i$, and estimates, which are simply $\mathcal{E}(\mu|ES_{R1}) = \bar{x}$, and $\mathcal{E}(\sigma|E) = \sqrt{\frac{I}{I+1}}s$. See [Hanson *et al.*, 1991] for more details. Computation here takes order I steps, used to compute the sufficient statistics.

4.3 Independent Attributes - S_I

We now introduce some notation for collecting sets of indexed terms like X_{ik}. A single such term inside a {} will denote the set of all such indexed terms collected

[5] If the fact that a data value is unknown might be informative, one can model "unknown" as just another possible (discrete) data value; otherwise the likelihood for an unknown value is just a sum over the possible known values.

[6] δ_{uv} denotes 1 when *u* — *v* and 0 otherwise.

[7] *T(y)* is the Gamma function [Spiegel, 1968].

[8] Actually, log(weight) is more normally distributed.

across all of the indices, like i and k in $E = \{X_{ik}\} \equiv \{X_{ik}$ such that $i \in [1, \ldots, I], k \in \mathcal{K}\}$. To collect across only some of the indices we use $\bigcup_k$ as in $E_i = \bigcup_k X_{ik} \equiv \{X_{i1}, X_{i2}, \ldots\}$, all the evidence for a single case i.

The simplest way to deal with cases having multiple attributes is to assume S_I that they are all independent, i.e., treating each attribute as if it were a separate problem. In this case, the parameter set V partitions into parameter sets $V_k = \bigcup_{l_k} q_{l_k}$ or $[\mu_k, \sigma_k]$, depending on whether that k is discrete or real. The likelihood, prior, and joint for multiple attributes are all simple products of the results above for one attribute: $S_1 = S_{D1}$ or S_{R1} — i.e., $L(E_i|VS_I) = \prod_k L(X_{ik}|V_k S_1)$, $d\pi(V|S_I) = \prod_k d\pi(V_k|S_1)$, and $M(E|S_I) = \prod_k M(E(k)|S_1)$ where $E(k) \equiv \bigcup_i X_{ik}$, all the evidence associated with attribute k. The estimates $\mathcal{E}(V_k|ES_I) = \mathcal{E}(V_k|E(k)S_1)$ are exactly the same. Computation takes order IK steps here.

4.4 Fully Covariant Discretes - S_D

A model space S_D which allows a set $\mathcal{K}$ of *discrete* attributes to fully covary (i.e, contribute to a likelihood in non-trivial combinations) can be obtained by treating all combinations of base attribute values as particular values of one super attribute, which then has $L' = \prod_k L_k$ values[9] V consists of terms like $q_{l_1 l_2 \ldots l_K}$, indexed by all the attributes. I_l generalizes to $I_{l_1 l_2 \ldots l_K} = \sum_i \prod_k \delta_{x_{ik} l_k}$. Given this transformation, the likelihoods, etc. look the same as before: $L(E_i \mid VS_D) = q_{l_1 l_2 \ldots l_K}$, where each $l_k = X_{ik}$, $d\pi(V|S_D) = dB(\{q_{l_1 l_2 \ldots l_K}\} \mid L')$, $M(E|S_D) = F_1(\{I_{l_1 l_2 \ldots l_K}\}, I, L')$, and [10] $\mathcal{E}(q_{l_1 l_2 \ldots l_K}|ES_D) = F_2(I_{l_1 l_2 \ldots l_K}, I, L')$ Computation takes order IK steps here. This model could, for example, use a single combined hair-color eye-color attribute to allow a correlation between people being blond and blue-eyed.

4.5 Fully Covariant Reals - S_R

If we assume S_R that a set $\mathcal{K}$ of real-valued attributes follow the standard multivariate normal distribution, we replace the σ_k^2 above with a model covariance matrix $\Sigma_{kk'}$ and s_k^2 with a data covariance matrix $\mathbf{S}_{kk'} = \frac{1}{I}\sum_i(x_{ik} - \overline{x}_k)(x_{ik'} - \overline{x}_{k'})$. The inverse Wishart distribution [Mardia *et al.*, 1979] gives an integrable prior on $\Sigma_{kk'}$. We again give only the marginal joint[11]

$$M(E|S_R) = \frac{\prod_{a=1}^{K} \frac{\Gamma(\frac{I+K-a}{2})}{\Gamma(\frac{1+K-a}{2})}}{I^{\frac{K}{2}} \pi^{\frac{K(I-1)}{2}}} \frac{\prod_k \mathbf{S}_{kk}^{\frac{K}{2}} \prod_k \widehat{\frac{\Delta\sigma_k}{\Delta\mu_k}}^{I}}{|(I + \delta_{kk'})\mathbf{S}_{kk'}|^{\frac{I+K-1}{2}}},$$

and estimates[12]

$$\mathcal{E}(\Sigma_{kk'}|ES_R) = \frac{I+\delta_{kk'}}{I-2}\mathbf{S}_{kk'}.$$

See [Hanson *et al.*, 1991] for details. Computation here takes order $(I + K)K^2$ steps.

4.6 Block Covariance - *Sv*

Rather than just having either full independence or full dependence of attributes, we prefer a model space *Sv* where some combinations of attributes may covary while others remain independent, with full or no dependence as special limiting cases. This allows us to avoid paying the cost of specifying covariance parameters when they cannot buy us a significantly better fit to the data.

Our approach to partial dependence is simply to partition the attributes *K* into *D* blocks *Kb,* each of size K_b, with full covariance within each block and full independence between blocks.[13] We also currently prohibit reals and discretes in the same covariant block.

The evidence E partitions block-wise into $E(\mathcal{K}_b)$ (using $E_i(\mathcal{A}) \equiv \bigcup_{k \in \mathcal{A}} X_{ik}$ and $E(\mathcal{A}) \equiv \{E_i(\mathcal{A})\}$), each with its own sufficient statistics; and the parameters V partition into parameters $V_b = \{q_{l_1 l_2 \ldots l_K}\}$ or $[\{\Sigma_{kk'}\}, \{\mu_k\}]$. Each block is treated as a different problem, except that we now also have discrete parameters $T = [B, \{\mathcal{K}_b\}]$ to specify. Thus the likelihood $L(E_i|VTS_V) = \prod_b^B L(E_i(\mathcal{K}_b)|V_b S_B)$ is a simple product of block terms $S_B = S_D$ or S_R assuming full covariance within each block, and the estimates $\mathcal{E}(V_b|ETS_V) = \mathcal{E}(V_b|E(\mathcal{K}_b)S_B)$ are the same as before.

We choose a prior $d\pi(VT|S_V) = \pi(B\{\mathcal{K}_b\}|S_V)\prod_b d\pi(V_b|S_B)$ which predicts the block structure $B\{\mathcal{K}_b\}$ independently of the parameters V_b within each independent block resulting in a similarly decomposed marginal $M(ET|S_V) = \pi(B\{\mathcal{K}_b\}|S_V)\prod_b M(E(\mathcal{K}_b)|S_B)$. We choose a block structure prior

$$\pi(B\{\mathcal{K}_b\}|S_V) = 1/K_R Z(K_R, B_R) K_D Z(K_D, B_D),$$

where $\mathcal{K}_R$ is the set of real attributes and B_R is the number of real blocks (and similarly for $\mathcal{K}_D$ and B_D). It is normalized using $Z(A, U) \equiv \sum_{u=1}^{U}(-1)^{u-1}\frac{(U-u+1)^A}{(U-u+1)!(u-1)!}$, the number of ways one can partition a set with A elements into U subsets. This prior prefers the special cases of full covariance and full independence, and thus includes a significance test. For example, in comparing the hypothesis that each attribute is in a separate block (i.e., all independent) with the hypothesis that only one particular pair of attributes covary together in a block of size two, this prior penalizes the covariance hypothesis in proportion to the number of such pairs possible.

Computation here takes order $NK(I\overline{K_b} + \overline{K_b^2})$ steps, where N is the number of search trials done before quitting, which would be around $(K-1)!$ for a complete search of the space. $\overline{K_b}$ is an average, over both the search trials and the attributes, of the block size of real attributes (and unity for discrete attributes).

[9] *L'* can be a very large number!

[10] *F₁* and F_2 are defined in Section 4.1.

[11] At present, we lack a satisfactory way to approximate this marginal when some values are unknown.

[12] To obtain these formulas, certain free hyperparameters in the Wishart prior have been set using simple statistics from the data. This is more robust and simplifies the math, but is "cheating" because priors are supposed to be independent of the data. Similar cheating was also done in Section 4.2.

[13] We choose it because it is easy, fits well with our model of class hierarchy, and allows full dependence.

5 Class Mixtures

5.1 Flat Mixtures - *SM*

The above model spaces *Sc* = *Sy* or s/can be thought of as describing a single class, and so can be extended by considering a space *SM* of simple mixtures of such classes [D.M.Titterington *et al,* 1985]. With S_c = ST, this is the model space of AutoClass III, and Figure 1 shows how it can fit a set of artificial real-valued data in five dimensions.

In this model space the likelihood $L(E_i|VTS_M) = \sum_c^C \alpha_c L(E_i|V_cT_cS_C)$ sums over products of "class weights" «$_c$, each giving the probability that any case would belong to class *c* of the *C* classes, and class likelihoods describing how members of each class are distributed. In the limit of large *C* this model space is general enough to be able to fit any distribution arbitrarily closely, and hence is "asymtotically correct".

The parameters $T = [C, \{T_c\}]$ and $V = [\{\alpha_c\}, \{V_c\}]$ combine parameters for each class and parameters describing the mixture. The prior is similarly broken down

$$d\pi(VT|S_M) = F_3(C)C!\, dB(\{\alpha_c\}|C) \prod_c d\pi(V_cT_c|S_C),$$

where $F_3(C) \equiv \frac{6}{\pi^2 C^2}$ for $C > 0$..The a is treated as if the choice of class were another discrete attribute, except that a *C* is added because classes are not distinguishable a priori.

Except in very simple problems, the resulting joint *dJ(EVT\S)* has many local maxima, and so we must now distinguish regions *R* of the *V* space. To find a local maxima we use the "EM" algorithm [Dempster *et al.$_y$* 1977] which is based on the fact that at a maxima the class parameters V_c can be estimated from weighted sufficient statistics. Relative likelihood weights Wi_C = $\alpha_c L(E_i|V_cT_cS_C)/L(E_i|VTS_M)$, satisfying $\sum_c w_{ic} = 1$ give the probability that a particular case *i* is a member of class c. Using these weights we can break each case into "fractional cases", assign these to their respective classes, and create new "class data" $E^c = \bigcup_{ik} [X_{ik}, w_{ic}]$ with new weighted class sufficient statistics obtained by using weighted sums $\sum_i w_{ic}$ instead of sums £V. For example $I_c = \sum_i w_{ic}$ and $\bar{x}_{kc} = \frac{1}{I_c} \sum_i w_{ic} x_{ik}$. Substituting these statistics into any previous class likelihood function *L(E\V$_c$T$_c$Sc)* gives a weighted likelihood $L'(E^c|V_cT_cS_C)$ and associated new estimates and marginals.

At the maxima, the weights w_{ic} should be consistent with estimates of $V = \{[\alpha_c, C_c]\}$ from $\mathcal{E}(V_c|ERS_M) = \mathcal{E}'(V_c|E^cS_C)$ and $\mathcal{E}(\alpha_c|ERS_M) = F_2(I_c, I, C)$. To reach a maxima we start out at a random seed and repeatedly use our current best estimates of *V* to compute the w_{ic}, and then use the w_{ic} to re-estimate the V, stopping after 10 — 100 iterations when they both predict each other.

Integrating the joint in *R* can't be done directly because the product of a sum in the full likelihood is hard to decompose, but by using fractional cases to approximate the likelihood $L(E_i|VTRS_m) =$

$$\sum_c^C \alpha_c L(E_i|V_cT_cS_C) \cong \prod_c (\alpha_c L(E_i|V_cT_cS_C))^{w_{ic}}$$

while holding the *wu* fixed, we get an approximate marginal:

$$jM(ERT|S_M) \cong F_3(C)C!\, F_1(\{I_c\}, I, C) \prod_c M'(E^cT|S_C)$$

Our standard search procedure begins each converging trial from classes built around *C* random case pairs. The number of classes *C* is chosen randomly from a log-normal distribution fit to the *Cs* of the 6 — 10 best trials seen so far, after trying a fixed range of *Cs* to start. We also have developed alternative search procedures which selectively merge and split classes according to various heuristics. While these usually do better, they sometimes do much worse.

The marginal joints of the different trials generally follow a log-normal distribution, allowing us to estimate during the search how much longer it will take on average to find a better peak, and how much better it is likely to be.

In the simpler model space *SMI* where *Sc* = *Si* the computation is order $NI\overline{C}K$, where $\overline{C}$ averages over the search trials. *N* is the number of possible peaks, out of the immense number usually present, that a computation actually examines. In the covariant space *SMV* where *Sc* = *Sv* this becomes $NK\overline{C}(I\overline{K_b} + \overline{K_b^2})$

5.2 Class Hierarchy and Inheritance - SH

When there are many attributes and each class must have its own set of parameters for each of these attributes, multiple classes are strongly penalized. Attributes which are irrelevant to the whole classification, like a medical patient's favorite color, can be particularly costly. To reduce this cost, one can allow classes to share the specification of parameters associated with some of their independent blocks.

Rather than allow arbitrary combinations of classes to share blocks, it is simpler to organize the classes as leaves of a tree. Each block can be placed at some node in this tree, to be shared by all the leaves below (farther from the root than) that node. In this way different attributes can be explained at different levels of an abstraction hierarchy. For medical patients the tree might have *viral-infectionsneai* the root, predicting *fevers,* and some more specific viral disease near the leaves, predicting more disease specific symptoms. Irrelevant attributes like *favorite-color* would go at the root. To make predictions about a member of a leaf class, one first inherits down the attribute descriptions from classes above it.

Therefore the above class mixture model space *SM CAN* be generalized to a hierarchical space *SH* by replacing the above set of classes with a tree of classes, and using the tree to inherit specifications of class parameters. From the view of the parameters specified at a class, all of the classes below that class pool their weight into one big class. Figure 3 shows some sample trees, and Figure 2 shows how a class tree, this time with *Sc* — *Sv,* can better fit the same data as in Figure 1.

A tree of classes has one root class r. Every other class *c* has one parent class P_c, and every class has J_c child classes given by *C$_c$j*, where the index *j* ranges over the children of a class. Each child class has a weight

α_{cj} relative to its siblings, with $\sum_j^{J_c} \alpha_{cj} = 1$, and an absolute weight $\alpha_{C_{cj}} = \alpha_{cj}\alpha_c$, with $\alpha_r = 1$.

Each class has an associated set of attributes $\mathcal{K}_c$, which it predicts independently through a likelihood $L(E_i(\mathcal{K}_c)|V_cT_cS_C)$ and which no class above or below it predicts. To avoid having redundant trees which describe the same likelihood function, only $\mathcal{K}_r$ can be empty, and non-leaves must have $J_c \geq 2$

We need to ensure that all attributes are predicted somewhere at or above each leaf class. So we call $\mathcal{A}_c$ the set of attributes which are predicted at or below each class, start with $\mathcal{A}_r = K$, and then recursively partition each A_c into attributes K_c "kept" at that class, and hence predicted directly by it, and the remaining attributes to be predicted at or below each child $\mathcal{A}_{C_{cj}}$. For leaves $A_c = K_c$.

Expressed in terms of the leaves the likelihood is again a mixture $L(E_i|VTS_M) =$

$$\sum_{c:J_c=0} \alpha_c \prod_{c'=c,P_c,P_{P_c},\ldots,r} L(E_i(\mathcal{K}_{c'})|V_{c'}T_{c'}S_C)$$

allowing the same EM procedure as before to find local maximas The case weights here *wci* $\sum_j^{J_c} w_{C_{cj}i}$ (with *w* *r i* 1) sum like in the flat mixture case and define class statistics $E^c(\mathcal{K}_c) = \bigcup_{k\in\mathcal{K}_c,i} [X_{ik}, w_{ci}]$.

We also choose a similar prior, though it must now specify the K_c as well: $d\pi(VT|S_H) =$

$$\prod_c d\pi(J_c\mathcal{K}_c \mid \mathcal{A}_cS_H)J_c!\, dB(\bigcup_j \alpha_{cj}|J_c)\, d\pi(V_cT_c \mid \mathcal{K}_cS_C)$$

$$d\pi(J_c\mathcal{K}_c \mid \mathcal{A}_cS_H) = F_3(J_c - 1)\frac{\mathcal{K}_c!\,(\mathcal{A}_c - \mathcal{K}_c)!}{(\mathcal{A}_c + \delta_{rc})\mathcal{A}_c!}$$

for all subsets K_c of *Ac* of size in the range $[1 - \delta_{cr}, \mathcal{A}_c]$, except that $F_3(J_c - 1)$ is replaced b δ_{0J_c} h e n A_c = K_c. Note that this prior is recursive, as tfte prior for each class depends on the what attributes have been chosen for its parent class.

This prior says that each possible number of attributes kept is equally likely, and given the number to be kept each particular combination is equally likely. This prior prefers the simpler cases of K_c = *Ac* and K_c = 1 and so again offers a significance test. In comparing the hypothesis that all attributes are kept at a class with the hypothesis that all but one particular attribute is kept at that class, this prior penalizes the all-but-one hypothesis in proportion to the number of attributes that could have been kept instead.

The marginal joint becomes $M(ERT|S_H) \cong$

$$\prod_c d\pi(J_c\mathcal{K}_c \mid \mathcal{A}_cS_H)J_c!\, F_1(\bigcup_j I_{C_{cj}}, I_c, J_c)M'(E^c(\mathcal{K}_c)T_c|S_C)$$

and estimates are $\mathcal{E}(V_c|ERS_H) = \mathcal{E}'(V_c|E^c(\mathcal{K}_c)S_C)$ and $\mathcal{E}(\alpha_{cj}|ERS_H) = F_2(I_{cj}, I_c, J_c)$ again.

In the general case of *SHV*, where *Sc* = *Sv*, computation again takes $NK\overline{C}(I\overline{K_b} + K_b^2)$, except that the $\overline{J}$ is now also an average of, for each k the number of classes in **the hierarchy which** use **that** *k* (i.e., have $k \in \mathcal{K}_c$). Since this is **usually less than** the number of leaves, the model S_H is typically cheaper to compute than S_M for the same number of leaves.

Searching in this most complex space *SHV* is challenging. There are a great many search dimensions where one can trade off simplicity and fit to the data, and we have only begun to explore possible heuristics. Blocks can be merged or split, classes can be merged or split, blocks can be promoted or demoted in the class tree, EM iterations can be continued farther, and one can try a random restart to seek a new peak. But even the simplest approaches to searching a more general model space seem to do better than smarter searches of simpler spaces.

6 Results

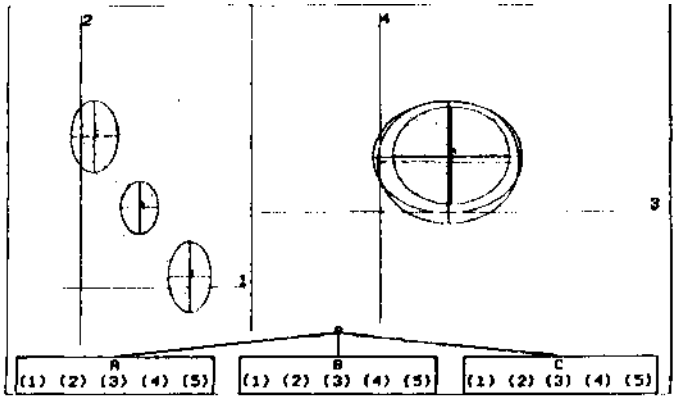

Figure 1: AutoClass III Finds Three Classes

We plot attributes 1 vs. 2, and 3 vs. 4 for an artificial data set. One σ deviation ovals are drawn around the centers of the three classes.

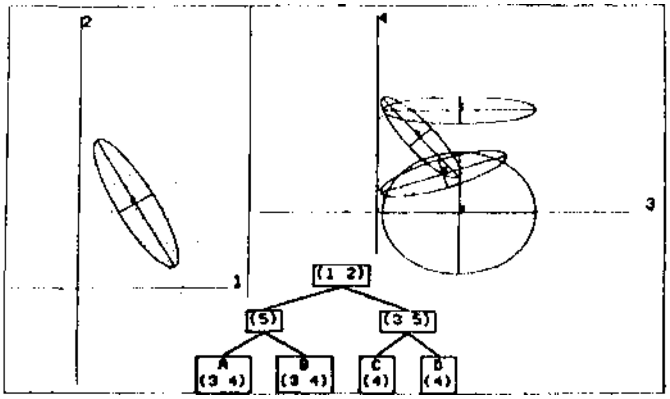

Figure 2: AutoClass IV Finds Class Tree x 10^{120} Better Lists of attribute numbers denote covariant blocks within each class, and the ovals now indicate the leaf classes.

We have built a robust software package, AutoClass III, around the flat independent model *SMI*, with utilities for reading data, controlling search, and viewing results, and have released it for general use. It is written in CommonLisp, and runs on many different machines. We, and others, have applied this system to many large and real databases. When applied to infrared stellar data we found new, independently verified, phenomena [Goebel *et* a/., 1989]. Others have successfully applied it to protein structure [Hunter and States, 1991].

We axe now developing Autoclass IV around the full hierarchical block covariant model space *SHV* • As hoped, it gives a significant improvement over the previous version when applied to real data, though we have only tried it on small problems so far. On the standard IRIS flowers data (150 cases, 4 attributes) we find a model over 10^{68} times more probable than the independent model, i.e., with a marginal joint (absolute value 10~^{913}) that much larger.[14] This compares with typical improvements of 10^2 for doubling the search time or for using a smarter merging search.

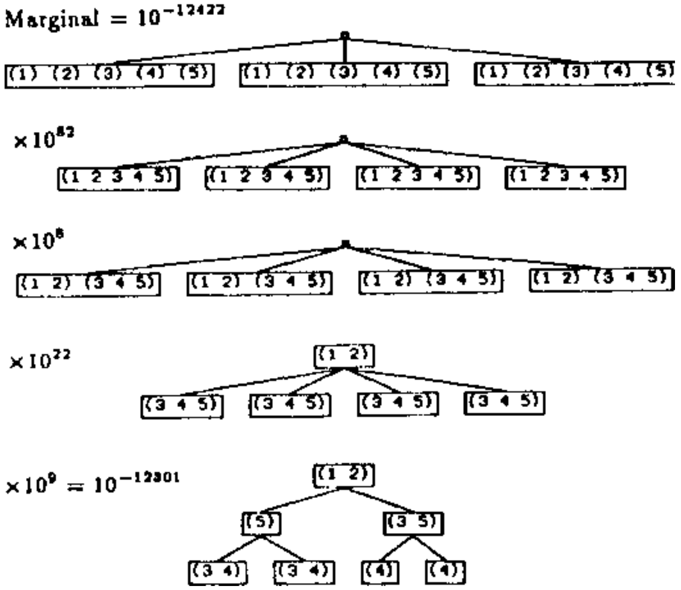

Figure 3: Each New Feature Allows a Better Model
The marginal probabilities improve by large factors as we fit each new class tree.

Figures 1, 2, and 3 illustrate a similar gain on an artificial data set with 400 cases and 5 real attributes. Figure 1 shows how the data is distributed in attributes #1 and #2, and in attributes #3 and #4 (attribute #5 is not shown). Superimposed upon this is the best result found with Autoclass III.

Since the data is dominated by a covariance between attributes #1 and #2, the independent model tries to model this by stringing classes along the 1 — 2 covariance axis. There is too little data to justify any more structure in this model class, so the #3, #4 and #5 axes are basically ignored. The

Figure 3 shows a progression of models between this model and the current best answer from Autoclass IV, given in Figure 2. Each new model adding one new feature and gains an improved marginal joint. Moving from a flat independent model to a flat fully covariant model, the best model found is over 10^{81} times more probable, with 4 classes each of which have significant covariances in 2 pairs of attributes. This large gain comes from combining a much better fit to the data, despite extra costs paid for the added class and covariance parameters. But, on inspection one finds that the four covariant classes are virtually the same in the 1 — 2 projection and show little correlation with the other attributes. We therefore split the covariance blocks and raise the (12) block to the common root for another relative increase of 10^{30}. This process can be repeated on pairs of the (3 4 5) blocks to get the 3 level tree shown in Figure 2, gaining another 10^9 in relative probability. We have thus gained a total factor of over 10^{120} in relative marginal probability over the best classification found using the independent model. Of that total, about 10^{39} comes from the fact that the tree now requires fewer parameters to specify a similar likelihood.

These preliminary results support previous indications that the ability to represent various kinds of structures in the data is the major limiting factor of such a system.

[14] Cross-validation would probably be a better test here, since our priors are fairly crude.

References

[Berger, 1985] J. O. Berger. *Statistical Decision Theory and Dayesian Analysis.* Springer-Verlag, New York, 1985.

[Cheeseman *et* a/., 1988a] P. Cheeseman, J. Kelly, M. Self, J. Stutz, W. Taylor, and D. Freeman. Autoclass: a Bayesian classification system. In *Proceedings of the Fifth International Conference on Machine Learning,* 1988.

[Cheeseman *et* a/., 1988b] P. Cheeseman, M. Self, J. Kelly, J. Stutz, W. Taylor, and D. Freeman. Bayesian classification. In *Seventh National Conference on Artificial Intelligence,* pages 607-611, Saint Paul, Minnesota, 1988.

[Dempster *et* a/., 1977] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J, Roy. Statist. Soc. B,* 39:1-38, 1977.

[D.M.Titterington *et* a/., 1985] D.M.Titterington, A.F.M. Smith, and U.E. Makov. *Statistical Analysis of Finite Mixture Distributions.* John Wiley & Sons, New York, 1985.

[Goebel *et* a/., 1989] J. Goebel, K. Volk, H. Walker, F. Gerbault, P. Cheeseman, M. Self, J. Stutz, and W. Taylor. A Bayesian classification of the IRAS LRS atlas. *Astron. Astrophys.,* 222:L5- L8, 1989.

[Hanson *et* a/., 1991] Robin Hanson, John Stutz, and Peter Cheeseman. Bayesian classification theory. Technical Report FIA-90-12-7-01, NASA Ames Research Center, Artificial Intelligence Branch, 1991.

[Hunter and States, 1991] Lawrence Hunter and David States. Applying Bayesian classification to protein structure. In *IEEE Conference on Applications of AI,* 1991.

[Mardia *et* a/., 1979] K.V. Mardia, J.T. Kent, and J.M. Bibby. *Multivariate Analysis.* Academic Press Inc., New York, 1979.

[Spiegel, 1968] Murray Spiegel. *Mathematical Handbook of Formulas and Tables.* McGraw-Hill Book Company, New York, 1968.