Cognitive Classification

Janet Aisbett and Greg Gibbon

School of Information Technology
The University of Newcastle, University Drive, Callaghan 2308 AUSTRALIA
{mgjea, mgggg}@cc.newcastle.edu.au

Abstract

Classification assigns an entity to a category on the basis of feature values encoded from a stimulus. Provided they are presented with sufficient training data, inductive classifier builders such as C4.5 are limited by encoding deficiencies and noise in the data, rather than by the method of deciding the category. However, such classification techniques do not perform well on the small, dirty /or and dynamic data sets which are all that are available in many decision making domains. Moreover, their computational overhead may not be justified. This paper draws on conjectures about human categorization processes to design a frugal algorithm for use with such data. On presentation of an observation, case-specific rules are derived from a small subset of the stored examples, where the subset is selected on the basis of similarity to the encoded stimulus. Attention is focused on those features that appear to be most useful for distinguishing categories of observations *similar to the current one*. A measure of logical semantic information value is used to discriminate between categories that remain plausible after this. The new classifier is demonstrated against neural net and decision tree classifiers on some standard UCI data sets and shown to perform well.

Introduction

Classification/categorization can be seen as a special case of commonsense or hypothetical reasoning, in which the competing hypotheses are the various possible assignments, and the observations are sets of feature-value pairs, or some other representation of data, relevant to the stimuli. Any classifier requires prior knowledge as to the relationship between the possible categories and the patterns of observational stimuli that will be encountered. The learning phase of classification is concerned with assembling this knowledge, that is, with building a classifier.

Categorization processes divide between those that induce rules from training data or cases, and then apply them to observations, and those which determine the similarity of an observation to stored exemplars of a category and then invoke a probabilistic decision function.

Classification techniques also divide into those which build a fixed classifier, and those that support change. Classifiers may change incrementally upon receipt of new data which relate feature values to categories, or may be recompiled for each observational stimulus, for example by using only those features deemed relevant to the current situation (Shepherd 1961)

The performance of inductive classifier builders developed in machine learning has for some time been limited more by encoding deficiencies and noise than by the method of making the categorization decision. To average out data anomalies, recent work has concentrated on pooling results from classifiers developed using single or multiple base learning algorithms, applied to different subsets of the data set and/or with different learning parameters (Dietterech 1997). As well as this trend toward more complex computational machinery, a popular research area is the analysis of massive data sets: massive in number of attributes, size of fields, and/or number of records. Finally there is renewed interest in defining and computing what is desirable in classification performance (eg. Cameron –Jones and Richards 1998)

Frugal data and frugal processes

In contrast to the categorization problems receiving most attention in machine learning, real-world decision making often involves classification performed on small, incomplete, dynamic data sets, for which there may be little opportunity to review performance. It may not be cost effective to collect the data needed to make accurate categorizations or evaluations, or the data are not available. In such situations, the contribution of the encoding of the stimulus to categorization errors overwhelms the contribution of the method of making the categorization decision. This paper is targeted at this type of classification decision where frugal approaches to processing are appropriate. Examples abound, and many are commercially and/or socially important. They involve rare events (diagnosing an uncommon disease), sparse data (categorising an individual consumer's purchase decisions within a business transaction database), noisy data (predicting how senators will vote on a non-standard matter) and dynamic data (categorising teenage fashion items) as well as situations in which detailed decision making is not warranted (selecting which apples to buy for lunch).

Our method seeks to mimic human categorization processes. It is quite general and quite simple. A frugal set of rules is recompiled on presentation of each stimulus. The rules are formulated from examples retrieved from long-term memory on the basis of their similarity to the stimulus. Attention is then focussed on those features, or more generally, on those parts of the description of the entities in memory, that are expected to convey most information about the categories in this data set. Here, information is calculated as difference in entropy. Generalization of the truncated exemplars provides a rule set. The size of the rule set is further reduced using a logical semantic information measure (Lozinskii 1994), the use of which is motivated by our desire to mimic conscious deductive human reasoning at this stage. The entity is assigned to the category about which there is most logical information brought by the knowledge base consisting of the encoded stimulus together with any other prior knowledge about the entity, and the frugal rule set. If required, ties between categories are broken by acquiring more information from long term memory, in the first instance through looking at further features. This step is again motivated by human reasoning processes.

Unlike conventional commonsense reasoning that starts with a set of (possibly conflicting) rules and facts, we start with a set of facts and derive rules from these in order to answer a question about an entity. Unlike common machine learning algorithms, we design a new classifier for each stimulus, and combine rule and exemplar based approaches. Moreover, the method naturally accommodates prior knowledge in forms other than attribute value pairs, to make it more "commonsense".

The remainder of this paper is organised as follows. The next section looks at work in cognitive science that suggests how human categorization might be performed. This is used in section 3 to develop an algorithm to perform frugal categorization. The algorithm is described in general terms, to allow it to be applied to categorization decisions in which both prior knowledge and observations are presented as logical formulae. Section 4 applies the new cognitively-inspired classifier to some standard data sets from the UCI repository. The method is seen to outperform standard decision tree and neural net classifiers when the training sets are small. Surprisingly, the method performs as well on the difficult UCI data set Cleveland Heart on a training set of 20 as conventional packages which have used more than five times as many training examples. The final section reviews our contribution and flags further work.

Background

Some version of a *principle of frugality* is recognised in any automatic computational area that involves complex computations on noisy data. Thus, in statistical data classification, data are routinely reduced through principal component analysis. Such algorithms rely either upon redundancy in the original data, or on more active search for new data to compensate for limited "memory" (eg Havens 1997). However, the main driver for frugal approaches is the fact that data often cannot support more sophisticated processing. This was the argument put in (Gigerenzer and Goldstein 1996). The authors presented various strategies for problem solving which were frugal in their processing and data requirements compared with conventional approaches such as linear models and multiple regression, and showed that these performed comparably on the types of data sets that would be encountered in real world decision making.

In our setting, the key issue surrounding data encoded from a stimulus is that these may not have the power to fully explain the categorization. Thus, a data set puts an associated "ceiling" on the classification accuracy of any classifier. On the simplest level, this may be due to missing values or to noise on the encoded values. It may also be due to significant interactions between features that have been encoded and features that have not. There may be instability in the underlying relationship between observational stimuli and the categories of the stimuli eg. as in credit application data where different decision makers may employ different rules and where all the knowledge used in the decision making may not be encoded. Stimuli may be encoded using different representation for different classes, unlike the fixed representation used in machine learning data sets. (For example humans appear to use different representations of facial features for different races (Levin 1996).)

Because data collection and storage are expensive, a complete data record may also not be encoded for each and every stimuli even if the data are available. Thus, humans are thought to encode feature data with exemplars of common categories containing the most typical features, and exemplars of rare categories containing the most distinctive features (Krushke 1996). In this particular case common categories may appear to have missing the very data that would enable them to be distinguished. So it is not just redundant information about a class that is dropped.

We want to develop a classification algorithm that can deal with such domains, so will draw on methods inspired by human processing. Part of our algorithm will mimic unconscious processes and will involve data reduction from recalled exemplars, and part will mimic conscious deductive reasoning on a frugal rule set derived from the examples.

Unfortunately, the nature of the interaction between exemplar and rule based approaches remains an open question in the cognitive science literature. Rules derived from small data sets are often claimed to be unstable and likely to be invalid. On the other hand, the performance of exemplar based methods relies on the richness of the example space, and they do not deal well with extrapolation compared with rule or model based methods.

Smith, Patalano and Jonides (1998) showed that the results of many experimental studies could be explained by assuming either a rule based or exemplar based approach, by varying other under-specified elements of the experimental situation (eg. difference in instructions). However, they say cognitive neuroscience research suggests that there exist separate neural circuits for categorising an observation using similarity to exemplars, and using rules. Shanks (1997) claims that categorization is mediated by references to remembered instances represented in a multidimensional psychological space, and supports the view that neural net type models are good descriptions of the process by which instances are encoded. On the other hand, conceptual knowledge is mediated by rules derived from the known members of the conceptual grouping.

If both exemplars and rules are used, then it is plausible that rules are derived as generalizations of at least some of the examples. The connectionist retrieval from memory of exemplars will be mediated by some measure of similarity between the observation and the stored exemplars of competing categories. Different measures may be invoked depending on the context. Nosofsky's Generalise Context Model (McKinley and Nofosky 1996) allows for an *attention weight* on features to model selective attention.

Exemplar and rule based categorization processes have been modeled with triggers for switching between the two (eg, Kalish and Kruschke 1997, Erickson and Kruschke 1998). Such models involve parameters that are fitted through typical feedback techniques which involve probabilities estimated using frequencies of occurrence. While we are going to develop a categorization process which derives rules from examples and does not switch between the two, we will need some notion of probability.

It has been widely held that humans do not reason in accord with a calculus of probability, and famous work has been done in this area by the "heuristics and biases" school of Amos Tversky. This has recently been revisited. A reinterpretation of probability in terms of observed frequencies of occurrence removes much of the evidence about neglect of base rates and other indicators of illogical reasoning (eg. Cosmides and Tooby 1996, Gigerenzer and Hoffrage 1995). The probabilistic model can be replaced by an experiential model where probability reflects the observed proportion, or the recalled proportion of observations. Base rates are less likely to be neglected when learned from experience (Kruschke 1996). It is still not clear to cognitive researchers when or whether probabilities are hard-coded through learning, rather than computed when needed through inquiry of stored instances (Jonides and Jones 1992).

We will in fact use two different notions of probability, one that is assumed to be computed on-the-fly on the basis of frequency, and the other that reflects the language of description of the problem. This is detailed in the next section. We will also invoke the phenomenon of anchoring. The tendency of humans to modify probability in order to anchor on prior beliefs about frequency has been recorded in many laboratory and real world situations. People tend to disregard data that are disconfirming of previously formed opinions, and tend to seek confirming evidence (eg. Garb 1996, Hogarth 1980). Thus, new data that increase the objective probability of an interest have a different impact to data that decrease its probability (Spies 1995). Subjects are also likely to be more confident of their judgement about events which they believe to be either very likely or very unlikely (eg. Lichenstein reported in Kleindorfer et al. 1993, Pulford and Colma 1996) than about those which they are uncertain. The tendency to anchor will be least when prior probability is around one half.

A Frugal Classification Algorithm

This section develops the frugal classification algorithm inspired by possible cognitive categorization processes. The presentation is very general, so some of the equations look reasonably cumbersome. However, the underlying algorithm is straightforward and natural. The design choices available are illustrated throughout by the choices made in the implementation used in section 4. The implementation introduces just two parameters.

Consider a decision maker tasked to assign a category to an observation about an entity y. The decision maker has access to a knowledge base M in which knowledge is encoded as formulae ϑ, which, for representational purposes in this paper, we write as statements in a typed first order language $\boldsymbol{L}$. The language has two distinguished types: the type of the category labels, and the type of the labels identifying the entities. The identifier label is often implicit in machine learning data sets.

On presentation of the observational stimulus, the decision maker encodes the new information as a formula φ. The decision maker also formulates an hypothesis set $\Phi(y) =\{\phi_i: C(y, c_i)\}$ where each hypothesis assigns y to a possible category. The encoded stimulus φ is used along with M to select between the hypotheses in $\Phi(y)$. In general, though not in the standard machine learning scenarios, M may contain formulae that directly refer to the entity y and which may be brought to bear in the decision making. Such prior knowledge is important in human categorization eg (Heit 1998).

(In the standard machine learning classification scenario, observational stimuli are encoded using typed binary predicates $A_j(u, v)$ where without loss of generality the first variable identifies the entity and second variable specifies the jth. feature value. The observation φ is then representable as a conjunction of instantiations of some or all of the feature predicates, $\wedge_j A_j(y, b_j)$. The knowledge base M is a set of exemplars $\vartheta = \vartheta(x)$, each stored as the conjunction of its instantiation of the category and of some or all of the feature predicates, viz. $\vartheta = C(x, c_i) \wedge_j A_j(x, a_{j,k})$. In general $A_j(x, a)$ does not imply $\neg A_j(x, b)$ for $a \neq b$: that is, a feature may take multiple values, because of uncertainty or non-exclusivity.)

Stage 1. Filtering

In deriving case-specific rules from the data in memory M, we model subconscious recall using two types of filters on M. The filters are orthogonal in the sense that one decreases the number of formulae to be considered and one restricts the size of each formula. When M is a conventional table of attribute-values, the filters are respectively a *select* and a *project* operation.

The first filter selects those exemplars that are in some sense "similar enough" to the encoded stimulus. In the following, this is defined in terms of syntactic similarity, on the basis that this type of measure appears to be used in low level recall processes in humans.

The second filter acts to focus attention on those parts of the exemplars that appear to be most relevant to the task of distinguishing between the hypotheses in $\Phi(y)$. Focussing is done in two ways. Firstly, the set of feature values (constants in the language) is reduced. If a car that is to be categorized as a sports car or some other type of car is blue, then the color of other remembered cars is not going to help this categorization unless they too are blue. Because of this, and in the spirit of frugality, we might suppose that the decision maker uses the observation to focus only on the terms that appear in φ. However, a remembered case should not be excluded because it refers to red cars, because there is a chance that the case conclusions may generalize to blue cars. Thus the filtered version of the language will have predicates typed to feature domains with up to two elements, as in "blue" and "a color other than blue" if such is known to exist. This filtering operation is formalized below. The second way that attention is focussed is by reducing the number of predicates under consideration. The method of selecting which predicates to retain will be discussed below.

The following definitions formally present the filtering operations:

Definitions

(i) The *recall triggered by* φ is the set $\boldsymbol{R}_\varphi(M) = \{\vartheta \in M: Sim(\varphi,\vartheta) > threshold\}$ where $Sim(-,-)$ is a measure of similarity between formulae.

In the next section, $Sim(\varphi,\vartheta)$ is defined to be the count of common constants. That is, $Sim(\varphi,\vartheta) = |Const(\varphi) \cap Const(\vartheta)|$ where $Const(\Omega)$ is the set of constants appearing in the formulation of Ω.

(ii) The *focus set triggered by* φ, $\boldsymbol{T}_{\varphi,\Phi}(M) = \{\vartheta|_{L'}: \vartheta \in M\}$, is obtained by restricting formulae in M to the sublanguage $\boldsymbol{L'}$ of $\boldsymbol{L}$ defined as follows:

(a) The constants in $\boldsymbol{L'}$ are either entity identifiers, categories, constants that appear in φ, or constants that represent the "other values" of a type set that do not appear in φ. Formally, for each type predicate *Type* in $\boldsymbol{L}$ other than the distinguished types, $Type \cap \boldsymbol{L'} = \{c, c''\}$ where $c \in Const(\varphi)$ and c'' is some arbitrarily-chosen representative element of $Type \cap \boldsymbol{L} \setminus Const(\varphi)$.

(b) The predicates in $\boldsymbol{L'}$ are those that convey highest relative information about φ. That is, there is a function $Ivalue_\Phi(-)$ from the predicates in $\boldsymbol{L}$ to the real numbers, together with a selection criterion, and $P \in \boldsymbol{L'}$ if and only if $P \in \boldsymbol{L}$ and $IValue_\Phi(P)$ satisfies the selection criterion (an example of this is given below).

Typically, *Ivalue* would be the expected relative entropy; this is the definition we use in the next section. Computing entropy requires a definition of probability, or at least, a frequency of occurrence. Given a set of formulae S, the frequentist probability $p_S(\gamma)$ of any formula γ is the proportion of formulae in S which prove γ. Then the relative entropy of γ given η is $-p_S(\gamma \wedge \eta) \log p_S(\gamma \wedge \eta) + p_S(\eta) \log p_S(\eta)$.

The expected relative entropy brought by a predicate P for a given hypothesis is obtained by summing over all the possible instantiations of P for fixed entity y, weighting each contribution to entropy with the proportion of formulae in S which prove the instantiation. The weighted sum over all hypotheses gives the expected total entropy, where the weight on the summand contributed by a hypothesis ϕ is the proportion of formulae in S which prove ϕ. The relevant set S here is $\boldsymbol{R}_\varphi(M)$. Formally, we are suggesting:

Definition

$$Ivalue_\Phi(P) = \sum_v \sum_\phi \{p_{\mathbf{R}\varphi(M)}(\phi) \{ p_{\mathbf{R}\varphi(M)}(P(v))(p_{\mathbf{R}\varphi(M)}(\phi \wedge P(v)) \log p_{\mathbf{R}\varphi(M)}(\phi \wedge P(v)) - p_{\mathbf{R}\varphi(M)}(\phi) \log p_{\mathbf{R}\varphi(M)}(\phi)) : v \in Domain\ P \cap \{y\}\}, \phi \in \Phi(y)\}$$

The selection criterion that we have used along with $Ivalue_\Phi$ is to take five predicates P such that no predicate that has not been selected has a higher $Ivalue_\Phi$ than a selected predicate. Ties are broken on a first-come-first-in basis. An alternative strategy would be to threshold on $Ivalue_\Phi$..

The enumeration of categories used in forming the hypothesis set is obtained from the categories instantiated in the exemplars recalled, that is, in $\boldsymbol{R}_\varphi(M)$.

The next step is to form rules as generalizations of the formulae in $\boldsymbol{T}_{\varphi, \Phi}\, \boldsymbol{R}_{\varphi}(M)$. A generalization of a formula is obtained by replacing any identifiers with variables. Thus the exemplar $\vartheta = C(x, c_i) \wedge_j A_j(x, a_{j,k})$ generalizes to the rule $\forall x \wedge_j A_j(x, a_{j,k}) \Rightarrow C(x, c_i),\ j = 1, \ldots n.$

In memory M there may be duplicate exemplars, and the likelihood of this increases in $\boldsymbol{T}_{\varphi, \Phi}\, \boldsymbol{R}_{\varphi}(M)$ because of the reduced language of description. It is reasonable to suppose that when duplicate formulae are encountered, their generalization is given more weight in the decision maker's mind. The strength of a formula is therefore important.

Definitions

Retain multiple occurrences of a formula Ω in both $R_{\varphi}(M)$ and $T_{\varphi, \Phi}(M)$, and say that Ω has *strength h* if there are h occurrences of the formulae in the focussed recall set $T_{\varphi, \Phi}\, R_{\varphi}(M)$.

$G\boldsymbol{T}_{\varphi, \Phi}\, \boldsymbol{R}_{\varphi}(M\ )$, the generalization of $\boldsymbol{T}_{\varphi, \Phi}\, \boldsymbol{R}_{\varphi}(M)$, is the set of ordered pairs $(\gamma, h(\gamma))$, where γ is the generalization of a formula in $\boldsymbol{T}_{\varphi, \Phi}\, \boldsymbol{R}_{\varphi}(M\ )$ in the sense described above, and $h(\gamma)$ is its strength.

Stage 2. Deductive Reasoning

Up to now we have been mimicking subconscious recall and focussing. Now we want to bring information in to the "conscious memory", that is, the working knowledge base. There are too many formulae in $G\boldsymbol{T}_{\varphi, \Phi}\, \boldsymbol{R}\varphi(M)$ to keep in the working memory of a frugal decision maker at the one time, if we are to keep within the limits of the "magic number 7" long postulated to constrain human reasoning (Miller 1956). A further cull of the set makes sense in any case as the set will be in general inconsistent with the observation φ which, as "prior knowledge", is already in the working knowledge base.

The decision maker should aim to bring into the working knowledge base those rules that carry most information about the hypotheses (that is, about the categories to which y could be assigned) relative to the prior knowledge. Usually, prior knowledge is taken just to be the encoded stimulus φ. It is possible that M contains other knowledge pertaining to the category of the entity to be classified, and in a general system this could be assembled before the task was attempted. However, here we assume that prior knowledge is just φ.

How should the information that a rule brings about a hypothesis given the observation be measured? It will need to be defined in terms of some difference in relative information (see for example (Blachman 1968)). Restated, it is a measure of how much knowing the rule and the observation reduce the surprise value on learning that the hypothesis is true, compared with the surprise value of finding the hypothesis is true when all you have is the observation.

We claim that the information intrinsic in measuring surprise should not be based on a frequentist view of probability, because humans do not normally consciously count up occurrences. Rather, here it is more appropriate to take a logical semantic definition of probability, as defined for example in (Lozinskii 1994). The logical semantic probability of a formula γ expressed in a language $\boldsymbol{L}$ is the proportion of models of the language that are models of γ. We employ a modified definition of semantic probability that takes account of anchoring. We suppose that in the face of disconfirming data the decision maker anchors on prior belief in whether the categorization is likely or not, with anchoring proportionate to that estimate. Specifically, if $p^*(\phi|\gamma)$ is the logical semantic posterior probability, our posterior probability $p(\phi|\gamma)$ is given by $(1 - p(\phi))p^*(\phi|\gamma) + p(\phi)\, p(\phi))$ if $p(\phi)\underline{>}$ ½, and by $p(\phi)p^*(\phi|\gamma) + (1 - p(\phi))\, p(\phi)$ otherwise. Logical semantic information in a formula γ is then as usual given by -log $p(\gamma)$. Relative information $I(\phi|\ \gamma)$ is $-log(p(\phi|\gamma)) / p(\phi))$.

The value of information about a hypothesis ϕ brought by γ given the observation φ is defined to be the difference of relative information values $I(\phi \mid \gamma \wedge \varphi) - I(\phi \mid \varphi)$. The total information brought by γ can be defined to be the sum of the information brought by each of the hypotheses, multiplied by the strength of γ. That is, the information brought by γ given the observation φ is $h(\gamma) \sum_i I(\phi_i \mid \gamma \wedge \varphi) - I(\phi_i \mid \varphi)$. For a similar formulation, see (Aisbett and Gibbon 1998).

This computation is not complicated in practice, and in most cases reduces to choosing the rules that have most terms in common with the observation. (Remember that while the initial filtering operation chose exemplars with a high degree of commonality, the orthogonal second filtering operation may have left exemplars with little or nothing in common with the observation).

Given a ranking of rules based on the value of information brought by each rule, a strategy is needed to determine which rules to actually bring in to the working knowledge base. For example, a strategy might be to bring in any rule which exceeds a threshold information value, or the J most informative rules for some given $J > 0$, or, most frugally, only those rules which maximise the information value. We adopted this last strategy in deriving the results presented in the next section. Only one or two rules are normally in this set, and refer to only a small number of feature values and one or two categories. This corresponds to a human using only a few key features to determine between a few of the most salient categories.

The final step is to determine which category the instance y belongs to on the basis of the information set consisting of the observation φ and the rule set Δ which maximises Equation (1). This is achieved by selecting the hypothesis that is most supported, that is: assign the entity to the category i for which the hypothesis ϕ_i has maximal logical semantic information relative to $\{\varphi\} \cup \Delta$. If there is more than one such i

then more features will have to be considered. This corresponds to a human reasoner seeking more clues about the classification. Features will be considered in order corresponding to their predicate ranking in the second filtering step. In the runs reported, if this procedure did not break the tie, then the first rule encountered would be given higher priority. More sophisticated tie breaking could be implemented, eg. picking the most probable category, where probability is assessed from the set $\boldsymbol{T}_{\varphi,\, \Phi}\, \boldsymbol{R}_{\varphi}(M)$.

Table 1: Average fractional misclassification rates

	Notes about runs	*Mushroom (2 categories)*	*Cleveland Heart (5 categories)*	*German Credit (2 categories)*
Cognitive classifier	***20 training observations, 5 runs***	***0.08***	***0.47***	***0.34***
CART	*20 training observations unless otherwise stated, 5 runs*	*0.17*	*0.66* *0.47 (100 training observations)*	*0.38*
NeuroShell	*20 training observations, 5 runs*	*0.19*	*0.53*	*0.36*
1 - Probability of most common class	*Calculated on full data set; see text*	*0.48*	*0.46*	*0.30*
C4.5	*Trained on 200 for Mushrooms, 500 for Heart and Credit, average of 50 runs*	*0.01*	*0.48*	*0.29*

Results

The data sets used in initial testing of the cognitively-inspired classifier are from the UCI Repository of Machine Learning Databases (Blake, Keogh, & Merz 1998). These results are presented as indicative of performance, and not as a rigorous comparison between classifiers which are designed to operate in different conditions (Salzberg, 1997). Of the UCI data sets, the Mushroom set was chosen because conventional techniques perform very well on it, whereas the German credit is moderately difficult and the Cleveland Heart produces high fractional error rates.

All data sets were converted to categorical, with any field taking more than 20 values converted to 5 values by simple linear transformation between the minimum and maximum. Five runs were used on each set. Twenty training and 100 test samples were randomly selected for each run. Other runs were done to ensure that 100 tests were sufficient for asymptotic error rates to have been achieved. The recall parameter for the cognitive classifier was fixed so that all 20 training samples were considered, and the focus parameter was fixed to focus on 5 attributes out of the available 13 (Heart) to 24 (German credit). These settings were designed to be cognitively realistic; no other parameter settings were investigated and so there has been no tuning to suit these data sets.

The new classifier was tested against a binary recursive partitioning tree building method, and a neural net. CART was selected as the decision tree builder in part because of its good handling of missing data (Steinberg and Colla 1995). A Probabilistic Neural Network implemented in NeuroShell2 package was selected because of its performance on sparse data sets (Burrascano 1991). Missing data were replaced by average field values when required by these classifiers. For CART, training data had to be enlarged in some of the Heart runs to ensure all categories had at least one representative.

As well, C4.5 error rates are reported, taken from Table 2 of Cameron-Jones and Richards. These rates are for classifiers trained on sets of between 200 and 500 items, so can be taken as representative of good classifier performance on typical machine learning data sets. (Note the Mushroom training set is still smaller than in most experiments reported in the literature -- the classification accuracy is also less than the almost-perfect scores achievable when 8000 or so records are used). One minus the probability of the most common class is the misclassification rate that would be achieved if a “dumb” decision maker always chose the most common category. These probabilities (which were reported by Cameron-Jones) are calculated over the full data sets, and so represent prior information unavailable to the frugal reasoner who only has 20 observations to hand.

Convergence characteristics depend heavily on the characteristics of the data set. On a set like the German Credit in which there are only 2 classes and the attributes do not explain the categorisations well, the cognitive classifier can perform adequately (though not as well as the "dumb" decision maker picking the most common class) even on 5 training/recall samples. Performance improves steadily with size of the recall set. When a recall set pegged to 20 exemplars could be selected from a larger set of exemplars in "long term memory", then improvement was only marginal for this type of data set. On very small training sets, CART will not build a tree, in which case the default is to become a "dumb" decision maker -- on a difficult dataset like German credit this means CART actually improves its performance, as measured by misclassification rate, over its performance when more training data are available.

In contrast, on databases like the UCI dermatology whose 34 attributes can explain all the six classes, having more exemplars in long term memory allows the cognitive classifier to perform better when recalling the most similar 20 exemplars to an observation than it can when it has only limited experience.

Discussion and further work

We have described and implemented a classifier that is frugal in a number of senses. Firstly, it builds a new but simple classifier for each observation, which is frugal behaviour for one-off classification tasks, or for classification using unstable data when the overhead of building a classifier is not justified. Our classifier is also frugal in the sense that it requires little training data to provide reasonable results. It is also frugal in the processing sense that it uses elementary filters to reduce the data. It is frugal in that it uses only a very small data set after the filters are applied. And finally, it is frugal in having few free parameters.

We showed that this classifier performed surprisingly well both on the sort of data sets for which it was designed, and for cleaner data. Thus, on a training set of 20 examples chosen at random from some of the standard UCI machine learning data sets, the frugal classifier outperforms some standard classifiers, and actually has performance comparable with classifiers trained on hundreds of training samples.

It is important to note that the UCI data sets were used to give a comparative feel for the performance of the new classifier, and should not be taken as suggesting that it is “just-another-classifier” in an already crowded field. We reiterate that our classifier is unique in being designed to cope with small training sets.

The classifier needs to be extensively tested on three different types of data. Firstly, it needs to be tested against the sort of data for which it is designed, to ensure that results are accurate enough to be useable in actual applications. Secondly, it needs to be comprehensively comparatively tested on the standard machine learning data sets, to gauge its performance as an all-round classifier. Thirdly, it needs to be tested on more complex data that include first order formulae.

When dealing with larger data sets, a design decision arises as to whether the decision maker should be modelled as retaining a perfect memory of all cases in long term memory. If not, and it is assumed that the decision maker retains the most recently experienced samples in memory, then the classifier will be able to cope with change but may be locally sensitive. If memory retains only the first samples encountered, then the classifier may have anchored on the past and may not be able to cope with change. A sensible and cognitively-supported scheme would be for the decision maker to retain a library of “good” exemplars for each category. This may affect base rate estimations.

Developments of the algorithm include modelling the formation of rules from exemplars, and using weights to focus attention on some attributes. Attributes encode stimuli in quite different ways, and Krushke (1996) suggested humans go further in treating different categories differently, encoding typical features for some categories, and distinguishing features for others. Treating all attributes equally, as our algorithm has done, is not cognitively supported. Alternative possibilities to having the system learn weights through training are to investigate differential encoding or, more generally, to allow users to weight features for their applications (eg. Minka & Picard 1997).

Other research underway is examining the impact on classification performance of some of the more complicating parts of the algorithm, such as anchoring.

References

Aisbett, J. and Gibbon, G. 1999. A practical measure of the information in a logical theory *Journal of Experimental and Theoretical AI.* 11 (1): +1-17.

Blachman, N. 1968. A mathematical theory of communication, *IEEE Transactions on Information Theory* IT-14: 27-31.

Blake, C., Keogh, E. and Merz, C. 1998. *UCI Repository of machine learning databases* Irvine, CA: Uni of California, Dept Information & Comp. Sc. http://www.ics.uci.edu/~mlearn/MLRepository.html

Burrascano, P. 1991. Learning Vector Quantization for the Probabilistic Neural Network. *IEEE Transactions on Neural Networks* 2: 458-461.

Cosmides, L. and Tooby, J. 1996. Are humans good intuitive statisticians after all? *Cognition* 58: 1-73.

Dietterich, T. 1997. Machine learning research. *AI Magazine*. 18: 97-136.

Cameron-Jones, M. and Richards, L. 1998. Repechage bootstrap aggregating for misclassification cost reduction. In *Proceedings of the 5th Pacific Rim Conference on Artificial Intelligence*, ed. H-Y Lee and H. Motoda, *Lecture Notes in Artificial Intelligence* 1531 1-11.: Springer.

Erickson, M. and Kruschke, J. 1998. Rules and exemplars in category learning. *Journal of Experimental Psychology: General* 127(2): 107-140.

Garb, H. 1994. Cognitive heuristics and biases in personality assessment. In Heath, L., Tindale, L, Edwards, J. et al (ed.) *Applications of heuristics and biases to social issues,* Plenum Press.

Hogarth, R. 1980. *Judgement and Choice*, John Wiley and Sons.

Gigerenzer, G. and Goldstein, D. 1996. Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review* 103 (4): 650-669.

Gigerenzer, G. and Hoffrage, U. 1995. How to improve Bayesian reasoning without instruction: frequency formats. *Psychological Review* 102 (4): 684-704.

Havens, W. 1997. Extending dynamic backtracking for distributed constraint satisfaction problems. In, *Proc 10th Aust. Joint Conf. on Artificial Intelligence, AI'97,* ed. A. Sattar, *Lecture Notes in Artificial Intelligence* 1342. 37-46.: Springer.

Heit, E. 1998. Influences of prior knowledge on selective weighting of category members. *Journal of Experimental Psychology: Learning, Memory and Cognition* 24(3): 712-731.

Jonides, J. and Jones, C. 199. Direct coding of frequency of occurrence. *Journal of Experimental Psychology: Learning, Memory and Cognition* 18(2): 368-378.

Kleindorfer, P., Kunreuther, H. and Schoemaker, P. 1993. *Decision Sciences: An Integrative Perspective.*: Cambridge University Press.

Kalish, M. and Kruschke, J. 1997. Decision boundaries in one dimensional categorization. *Journal of Experimental Psychology: Learning, Memory and Cognition* 23(6) :1362-1377.

Kruschke, J. 1996. Base rates in category learning. *Journal of Experimental Psychology: Learning, Memory and Cognition* 22(1): 3-26.

Lassaline, M. 1996. Structural alignments in induction and similarity. *Journal of Experimental Psychology: Learning, Memory and Cognition* 22(3): 754-770.

Levin, D. 1996. Classifying faces by race: the structure of face categories. *Journal of Experimental Psychology: Learning, Memory and Cognition* 22(6): 1364-1382.

Lozinskii, E. 1994. Information and evidence in logic systems. *Journal of Experimental and Theoretical Artificial Intelligence* 6: 163-193.

McKinley, S. and Nosofsky, R. 1996. Selective attention and the formation of linear decision boundaries. *Journal of Experimental Psychology: Human perception and performance*. 22(2): 294-317.

Miller, G. 1956. The magical number seven plus or minus two: some limits on our capacity for processing information. *Psychological Review* 101: 343-352 (1994) (reprinted from *Psychological Review* 63: 81-97).

Minka, T. and Picard, R. 1997. Interactive learning with a 'society of models'. *Pattern Recognition* 30 (4): 565-581.

Pulford, B. and Colman, A. 1996. Overconfidence, base rates and outcome positivity/negativity of predicted events *British J. of Psychology* 87(3): 431-447.

Salzberg, S. 1997. On Comparing Classifiers: Pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery* 1: 317-327.

Shanks, D. 1997. Representation of categories and concepts in memory. In *Cognitive models of memory. Studies in cognition.* Ed. Conway, M. et al. The MIT Press. Cambridge :111-146.

Shepherd, R. and Chang, J. 1961. Stimulus generalization in the learning of classifications. *Journal of Experimental Psychology* 65: 94-102.

Smith, E. Patalano, A. and Jonides, J. 1998. Alternative strategies of categorization. *Cognition* 65(2-3):167-196.

Spies, M. 1995. Uncertainty and Decision Making. In *Contributions to Decision Making* ed. Caverni, J., Bar-Hillel, M. et al, North Holland 51-83.

Steinberg, D. and Colla, P. 1995. CART: Tree-structured non-parametric data analysis. San Diego: Salford Systems.