Minimum Message Length Hidden Markov Modelling

Timothy Edgoose, Lloyd Allison
Computer Science Department, Monash University, Clayton, VIC 3168, Australia.
eMail: {time,lloyd}@cs.monash.edu.au
phone: +61-3-9905-5779

This paper describes a Minimum Message Length (MML) approach to finding the most appropriate Hidden Markov Model (HMM) to describe a given sequence of observations. A MML estimate for the expected length of a two-part message stating a specific HMM and the observations given this model is presented along with an effective search strategy for finding the best number of states for the model. The information estimate enables two models with different numbers of states to be fairly compared which is neccessary if the search of this complex model space is to avoid the worst locally optimal solutions. The general purpose MML classifier 'Snob' has been extended and the new program 'tSnob' is tested on 'synthetic' data and a large 'real world' dataset. The MML measure is found to be an improvement on the Bayesian Information Criteria (BIC) and the un-supervised search strategy effective.

1 Introduction

Classification, also known as mixture modelling or clustering, is the building of models from sets of observations where each observation is assumed to have been generated from one of a finite number of classes. A classification model specifies the number of such classes and a distribution over the observations expected for each. Un-supervised classification programs attempt to find the most likely class structure and parameterisation given a set of observations. This task requires that a balance be struck between model complexity and explanatory power. The best model will be sufficiently complex as to avoid discarding information implicit in the observed data, but not so complex as to be fitting noise in the observed data (over fitting).

A partial solution to this problem was presented in earlier un-supervised classification work by Wallace and Boulton (1968) and subsequently generalised by Wallace (1987,1990). A Minimum Message Length (MML) information measure was proposed that would estimate the length of an optimal two-part message stating a model and a set of observations given the model stated. Such a message length gives a fair measure by which any two competing classification models can be compared.

This earlier MML classification work was designed to model randomly sampled data and hence assumed independence between observations in a dataset. In this paper we present a MML based approach to the un-supervised classification of a sequence of observations which takes advantage of some of the extra information available in such data. Specifically the data is modeled as if it were generated from a first order Markov process with as many states as there are classes of observation. The state of such a process at any point in the sequence determines the class from

which the corresponding observation is generated. Such a model is commonly referred to as a Hidden Markov Model (HMM) which although not appropriate for all types of sequential data is none the less of significant practical interest. For a good introduction to these models which are rich in mathematical structure and have been used extensively in the area of speech recognition refer Rabiner (1989). An iterative solution for these models was first proposed by Baum et al. (1970). The technique applied was an Expectation Maximisation (EM) method later generalised by Dempster et al. (1977). Later work by Leroux and Puterman (1992) improved on the work by Baum by using a Bayesian Information Criteria (BIC) to estimate the complexity of a given HMM model and hence they were able to compare two HMMs with a different number of states. However, these works suffered from the lack of a suitable search method for larger model spaces and also from the surprising notion that there can ever be enough observational evidence to justify a probability of zero (or one) when estimating model parameters. This in turn led to zeros being preserved in the transition matrix and the possibility of a search being trapped in such a solution.

We extend this earlier work by deriving a MML information estimate for such a model, an improvement on the approximate BIC estimate, and we specify a effective search method of this complex model space which is guided by this measure.

The MML classification program 'Snob' of Wallace (1990) has been re-implemented and extended in order to model first order Markov processes. The new program, tSnob, is a more portable implementation of the MML classifier written in the C programming language. The program is designed to model multi-variate data with a fixed number of attributes. The type of these attributes can be discrete, continuous or angular these being modeled by multi-state, Gaussian or von Mises distributions respectively. Attribute values are assumed to be independently distributed within a class and the model correctly handles observations with missing attribute values.

The MML modelling approach taken is Bayesian in nature and strong parallels exist between Snob and the Bayesian classifier Autoclass produced by Cheeseman (1988). The two methods are contrasted in Wallace (1990).

2 MML Basics

Within the MML paradigm, models are judged by their ability to reduce the expected length of a message sending our model (the hypothesis) to an optimal precision and our observations given this model (the evidence) to a receiver who initially only shares our prior beliefs. The best model will minimise the length of this two-part message. No model that fails to compress the evidence can be considered superior to the empty model (null hypothesis).

These two message parts are used as estimates for $P(H)$ and $P(D|H)$ in Bayes Theorem.

$$P(H\&D) \;=\; P(H).P(D|H) \;=\; P(D).P(H|D)$$

For any one particular dataset $P(D)$ is constant, so maximizing $P(H\&D)$ also maximizes $P(H|D)$. This gives an fair criterion which we can use to compare any two hypotheses based on a given set of observations.

Wallace and Freeman (1987) gives the general form of such a MML estimate (given an appropriate likelihood function and a stable prior) as

$$ML(H\&D) \approx -\ln h(H) + \tfrac{1}{2}\ln det(F(H)) - \ln f(D|H) + g(n_p)$$

where $h(H)$ is a prior distribution over parameter values, $F(H)$ is the Fisher Information matrix, $f(D|H)$ is the likelihood function for the model, $g(n_p)$ is a function of the number of parameters being estimated, and the unit of the result is natural bits or nits (divide by $\ln 2$ to convert to bits).

The objective function for tSnob is constructed using three such MML expected optimal code length estimates. They predate the derivation of the general form, but are close approximations.

The first of these is the multi-state distribution where observed values are discrete and come from a finite unordered set of possibilities. From Wallace and Boulton (1968) the optimal MML code length to transmit K values from an M state multi-state distribution (assuming a uniform prior over the possible combinations for the frequencies of the observed values) is:

$$ML(H\&D) \approx \tfrac{M-1}{2}(\ln \tfrac{K}{12} + 1) - \ln(M-1)! - \textstyle\sum_{m=1}^{M}(n[m] + \tfrac{1}{2})\ln p[m]$$

where $n[m]$ is the number of values in state m and $p[m]$ is the probability stated for state m and is re-estimated as $p[m] = \frac{n[m]+\frac{1}{2}}{K+\frac{M}{2}}$.

The second is the Normal distribution where values are continuous reals stated to a specified accuracy. From Wallace and Boulton (1968) the optimal MML code length to transmit K values from a normal distribution with mean, μ, standard deviation, σ, and measurement accuracy, ε, from a global distribution with mean, μ_p, and standard deviation, σ_p, (assuming μ has a uniform prior in the range $\mu_p \pm 2\sigma_p$ and $\ln\sigma$ has a uniform prior in the range $\ln\varepsilon$ to $\ln\sigma_p\sqrt{2\pi}$) is:

$$ML(H\&D) \approx -\ln(4\sqrt{\tfrac{K}{12}}\tfrac{\sigma_p}{\sigma}) - \ln(\ln(\tfrac{\sqrt{2\pi}\sigma_p}{\varepsilon})\sqrt{\tfrac{K-1}{6}}) - K\ln(\tfrac{\sigma\sqrt{2\pi}}{\varepsilon} + \tfrac{1}{2}) + \tfrac{1}{2}$$

and σ is re-estimated as $\sqrt{\frac{v}{K-1}}$ where v is the sample variance.

Finally, we consider the von Mises distribution detailed in Fisher (1993) for modelling angular values stated to a known accuracy. It has mean direction μ and concentration parameter κ. Letting $I_0(\kappa)$ be the relevant normalisation constant, it has probability density function

$$f(x|\mu,\kappa) = \tfrac{1}{2\pi I_0(\kappa)} e^{\kappa.\cos(x-\mu)}$$

which for small κ tends to a uniform distribution and for large κ tend toward a Normal distribution with variance $1/\kappa$. This is a circular analogue of the Normal distribution - both being maximum entropy distributions. The MML estimate for the von Mises distribution is less compact and can be found in Wallace and Dowe (1993).

3 Calculating the Message Length of Model and Data

In this section we define the calculation for the expected length of a near optimal message encoding a specific parameterisation of our Markov classification model (stated to an appropriate accuracy) and a sequence of observations given this model. We derive the length of such a message by first stating the length of a non-optimal encoding and then deriving the length of the optimal encoding by argument.

Our encoding consists of four parts. Part (a) states the number of classes in the model. Part (b) states the relative abundance of these classes. Part (c) states, for each class, the distribution parameters and the relative abundance of each class conditional upon being proceeded by this class. Finally part (d) states, for each observation, an assigned class and the attribute values given this class. Parts (a),(b) and (c) constitute our hypothesis, H, and part (d) is one possible encoding of our data, D, given this hypothesis.

In part (a) of our message all values for the number of classes, N, are considered equally likely so stating N is assumed to have some unknown constant cost. As we only use this calculation to compare models we can safely omit part (a).

The length of part (b) of our message is the cost of sending the description of a multi-state distribution which could be used to assign each observation (K in all) to a particular class (N possibilities).

The code length required to describe a class, c_i, can be closely approximated as the sum of the optimal code lengths required to state the parameters describing each attribute. The message length of part (c) of our message is the sum of these individual class message lengths and additionally another N multi-state distributions specifying the class distribution for next observation given the class of this observation. Each of these additional multi-state distributions encodes a proportion of the total number of observations as specified by the class relative frequency stated in part (b).

One caveat of note in this current implementation of tSnob is that this calculation is a conservative encoding of the N^2 transition matrix (the optimal message is slightly shorter). An optimal encoding of this transition matrix is not known to the authors. However, one should be able to save something like one row of the matrix. Specifically our estimate is calculated assuming independence between rows in the transition matrix. This is clearly not the case (there are fewer degrees of freedom), however, this assumption yields a close approximation that is only slightly biased toward more conservative models. In turns out experimentally that part (b) of the message can be omitted with out any over-fitting of data and so the experimental results stated use this more aggressive measure.

Once parts (a),(b) and (c) which constitute our hypothesis, H, have been transmitted we can transmit part (d), the actual observations we have, by selecting a class for each observation and encoding each observation accordingly. An observation is coded as the sum of the optimal encoding for each attribute value using the stated class with missing attribute values coding as zero length messages (i.e. the receiver is assumed to know a-priori which attributes are missing). The class of the first observation is specified using the un-conditional multi-state distribution stated in part (b) of the message and the class of each successive observation is specified using the

appropriate conditional multi-state distribution as stated in part (c) of the message (i.e. based on the class of the preceding observation). In this way it is possible to calculate the length of one possible decodable message. We can consider any one such assignment of observations to classes as a path through our data and note that with N classes and K observations there are N^K such paths. A message stating any one such path will not be an optimal encoding of the data given the model. However, we can now calculate the probability (and hence the length) of the optimal message by summing over all of these N^K sub-optimal encodings.

Summing these N^K probabilities appears to be a formidable task. However, if we consider our encoding process as a state machine, we find that our model is left in only one of N possible states after the encoding of any observation. So for any observation, we can calculate the sum over all paths that lead to one of our N states based on the N sums calculated for the preceding observation.

As the only messages we consider are prefix codes (ie. uniquely decodable) we can, for notational convenience, define a mapping from message lengths to probabilities as $P_{ml}(x) = e^{-ML(x)}$ where $P_{ml}(x)$ is the probability that we will send a message, x, of length $ML(x)$ nits.

We define $P_{ml}(c_i|c_j)$ to be the probability associated with a message stating that an observation from class i follows an observation from class j and $P_{ml}(o_k|c_i)$ to be the probability associated with a message stating the attribute values associated with observation k using class i. We can now define $F(o_k \in c_i)$ to be the sum over all paths (messages) that lead to and include an encoding of observation k as a member of class i as

$$\begin{aligned} F(o_1 \in c_i) &= P_{ml}(c_i).P_{ml}(o_1|c_i) \\ F(o_k \in c_i) &= \textstyle\sum_{j=1}^{N} F(o_{k-1} \in c_j).P_{ml}(c_i|c_j).P_{ml}(o_k|c_i) \quad , 1 < k \leq K \end{aligned}$$

and finally

$$P(D|H) = \textstyle\sum_{i=1}^{N} F(o_K \in c_i)$$

which is the sum over all the possible encodings of our data given our model.

The message length of parts (a), (b) and (c) give us P(H), and we can calculate $P(D|H)$ so we now have $P(H\&D)$. This is the objective function that the tSnob program maximises by minimising it as a message length.

4 Searching the Model Space

The complexity of this first order Markov classification model space increases dramatically with the number of classes as does the probability of finding locally optimal solutions. Our search of this model space attempts to avoid local optima by limiting the complexity of our model (the number of classes) at any one time to that justified by our MML information measure. To this end we divide the search into two sub-problems. Searching for the best parameterisation of a model with N classes, and finding the best value for N.

4.1 Improving the Parameter Estimates

We improve the model parameterisation given a particular class structure by the repeated application of an EM re-estimation step. In order to apply the EM algorithm on this problem it is necessary to consider the optimal assignment of each observation to the N classes independently of the assignment of any of the other observations in the sequence. In fact we wish to calculate the sum over all the N^{K-1} possible encodings of our dataset that specify any one of the N states for any one observation.

To achieve this we can define a backward sum over all possible paths that lead from a classification of class i for observation k to the end of the data sequence as

$$\begin{aligned} B(o_K \in c_i) &= 1 \\ B(o_k \in c_i) &= \textstyle\sum_{j=1}^{N} P_{ml}(c_j|c_i).P_{ml}(o_{k+1}|c_j).B(o_{k+1} \in c_j) \quad ,0 < k < K \end{aligned}$$

We can now define the contribution of the class i for observation k to the final $P(D|H)$ to be

$$P(D|H, o_k \in c_i) = F(o_k \in c_i).B(o_k \in c_i)$$

and note that

$$P(D|H) = \textstyle\sum_{i=1}^{N} P(D|H, o_k \in c_i) \quad ,\forall k \in [1, K]$$

Once we have calculated these N sums for any particular observation we can calculate the relative contribution of each of the N states to the encoding of the entire sequence, $P(D|H)$. With this information we can correctly re-estimate all the class distribution and transition parameters. This calculation differs from the usual forward-backward maximum log likelihood calculation in that the appropriate MML message lengths used may also include small penalty terms which depend on the accuracy to which the corresponding distribution is specified in the hypothesis.

4.2 Selecting the Best Number of Classes

In order to select the best number of classes we employ a variation on the class splitting and merging search procedure implemented in the original snob program described in Wallace (1990). At any one time we consider a specific N class model and we move toward the best solution in this model space. However, it turns out that by calculating this we can also easily search a useful subset of the $N+1$ and $N-1$ class models. If a model in either subset turns out to be more likely than our current N class model then we switch our focus, N, to the better model space. In this way a simple model will shift to the more complex hypothesis spaces only when this is justified by the MML objective function. Having an accurate information measure to guide this shift in focus is essential to get good initial parameter estimates in the more complex model spaces and thus avoid the worst of the locally optimal solutions.

We only consider one model in each alternate model space at any particular time. These models are constructed from the current N class model and given a limited number of improvement cycles in which to yield a better solution than the current model. If a better solution is not soon found then alternative $N-1$ and $N+1$ models

are constructed and the process repeated. The selection of candidate models in these other model spaces is guided by message length estimates based on the current N class model. These estimates give an upper bound for the true message length in the alternate model spaces. The most promising change that has not been recently evaluated is selected in each case.

We calculate N estimates in the $N+1$ model space. These being where any of the current N classes is split to form two new classes while the other $N-1$ classes are kept the same. This is achieved by maintaining a hidden two class split model within each of the current N classes. These split models are initialised by random assignment from the corresponding model class and then re-estimated on each pass of the dataset. To speed up the re-estimation process the observations are assigned to one split class or the other for the first three cycles and thereafter probabilistically by EM. The split models are periodically re-initialised in order to search for different asymmetries in the data.

We calculate ${}^{N}C_{2}$ estimates in the $N-1$ class model space. These being where any two classes are combined into one class while all the other classes remain unchanged. These estimates are derived by adding the observation statistics for candidate merge classes and then calculating the revised expected message length for the new model.

This class splitting and merging search differs from that of the original Snob program in that the message length estimates are only used to select candidate split or merge models. The complete model message length evaluation is still required before such a model can be selected as the new focus of the search. Naturally when we split or merge classes to generate a new model, care must be taken that all the starting values for the class transition probabilities are reasonable.

Practically speaking, the repeated application of these two model search methods is an effective search strategy (the EM algorithm may of course converge to a local minimum).

5 Experimental Results

In this section we compare MML and BIC classification models on a variety of difficult 'synthetic' datasets. We also consider the MML model on a difficult 'real-world' dataset.

5.1 Generating Synthetic Data

The general aim here has been to generate some tough multivariate testing datasets using a generator with a minimal number of parameters. The Data generated has two continuous attributes generated from Gaussian distributions. As the number of classes varies the class attribute means are chosen so that the classes are evenly spaced around the circumference of a unit circle with the standard deviations fixed at 0.5. The class transition matrix used to generate the data has probabilities of 0.8 for the diagonal elements with all other probabilities being equal (i.e. $\frac{0.2}{N-1}$ for N class data).

Table 1: Model Selection Comparison

Model		Best MML model (%)							Best BIC model (%)						
N	K	1	2	3	4	5	6	7	1	2	3	4	5	6	7
5	100	6	58	36	-	-	-	-	4	60	36	-	-	-	-
	177	-	4	72	24	-	-	-	-	10	78	12	-	-	-
	316	-	2	20	66	12	-	-	-	-	32	66	2	-	-
	562	-	-	2	22	76	-	-	-	-	10	34	56	-	-
	1000	-	-	-	-	100	-	-	-	-	-	15	85	-	-
	1778	-	-	-	-	100	-	-	-	-	-	4	96	-	-
	3162	-	-	-	-	100	-	-	-	-	-	-	100	-	-
6	100	5	45	50	-	-	-	-	-	60	40	-	-	-	-
	177	-	16	70	14	-	-	-	-	16	80	4	-	-	-
	316	-	-	35	65	-	-	-	-	-	45	55	-	-	-
	562	-	-	6	54	36	4	-	-	-	10	82	8	-	-
	1000	-	-	-	10	20	70	-	-	-	-	25	55	20	-
	1778	-	-	-	2	4	94	-	-	-	-	4	2	94	-
	3162	-	-	-	-	-	100	-	-	-	-	-	-	100	-
7	100	10	65	25	-	-	-	-	10	60	30	-	-	-	-
	177	-	18	76	6	-	-	-	-	20	76	4	-	-	-
	316	-	-	50	50	-	-	-	-	-	75	25	-	-	-
	562	-	-	2	84	14	-	-	-	-	10	88	2	-	-
	1000	-	-	-	5	65	30	-	-	-	-	40	60	-	-
	1778	-	-	-	5	10	55	30	-	-	-	-	34	66	-
	3162	-	-	-	-	5	-	95	-	-	-	-	5	10	85

5.2 Comparing Model Selection Criteria

The MML model was compared against a BIC model for datasets with between 1 and 7 classes with dataset sizes varying between $\log_{10} 1$ and $\log_{10} 3.5$. The results for 4,5 and 6 class data are presented in table 1. The BIC measure used was defined as $\frac{N*(N-1)+N*4}{2} \log K - L$, where N is the number of states, K is the dataset size, and L the log-likelihood of the data.

Both model types were evaluated on the same 100 datasets of each type in order to fairly determine the distribution of models selected by each. It was observed that MML was more conservative for small datasets (less than 30 items) with the BIC criteria often over-fitting (about 10% of the time) for very small datasets (10 items). However, MML out performs the BIC criteria for the more complex models (more than 4 classes) with moderate numbers of observations (between 100 and 1000). Except for the problems that the BIC criteria has with small datasets both models rarely if ever over-fit this data. As the dataset size increases both methods converge to the correct generating model with the MML model converging more rapidly in the more complex model spaces.

5.3 Real World Data

The 'real-world' dataset selected consists of 41731 pairs of protein dihedral angles (ϕ,ψ). Secondary structure classification of such data is of significant interest in the area of protein modelling. The angle pairs are constructed from approximately 230

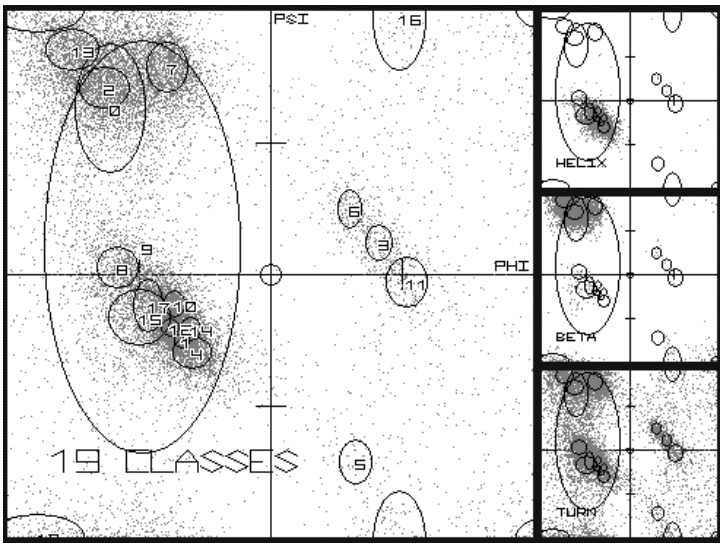

Figure 1: 19 class protein structure model

separate proteins as detailed in Edgoose et al. (1998) and the program was modified to encode each protein segment independently. The von Mises distribution was used to model both the ϕ and ψ angle attributes.

The best Markov classification model found had 19 classes and a message length of 266973 bits. This 19 class structure is shown in figure 1 with each class depicted by an ellipse with centre (μ_ϕ,μ_ψ) and dimensions $(\frac{1}{\sqrt{\kappa_\phi}}, \frac{1}{\sqrt{\kappa_\psi}})$. The actual observations are overlaid to create a scatter plot which is a square depiction of the surface of a torus and hence wraps around the edges. The class model found correlates well with known biological structures as well as pointing to other statistically significant relationships some of which are sequence related.

The search procedure for the Markov model space was found to be effective and consistent on this large and complex 'real-world' dataset.

6 Conclusion

We have extended the MML un-supervised classifier Snob to model ordered datasets where the best classification of an observation need not be independent of the classification of neighbouring observations. Specifically we model the data as if it had been generated from a first order Markov process with the state at any point specifying the class of the corresponding observation. Such a model is commonly referred to as a Hidden Markov Model.

We define a near optimal information measure for the cost of stating such a model and a set of observations given the stated model. This gives an objective criteria by which we can judge two competing models which differ in the numbers of classes they contain given a specific dataset. This measure is used to guide a robust un-supervised search of the Markov classification model space that correctly balances model complexity against explanatory power.

Experimentally it has been shown that the MML information measure for the Markov classification model yields improved class model selection results when compared with the more commonly used BIC criteria.

The Markov classification model has been used with consistent success on a large and difficult 'real-world' protein dataset indicating that the un-supervised search heuristics are effective and the model search robust.

Acknowledgments

Special thanks to Rohan Baxter for discussions and thanks also to David Dowe for assistance with the implementation of the von Mises distribution.

References

Baum L.E., Soules G., Petrie T. and Weiss N. (1970) A maximisation technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, 41:164–171.

Cheeseman P. C. (1988) Autoclass II conceptual clustering system. *Proceedings Machine Learning Conference*, pages 54–64.

Dempster A.P., Laird N.M. and Rubin D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society (Series B)*, 39:1–22.

Edgoose T., Allison L. and Dowe D.L. (1998) An MML Classification of Protein Structure that knows about Angles and Sequence *Proceedings of the 3rd Pacific Symposium on Biocomputing*

Fisher N.I. (1993) em Statistical Analysis of Circular Data. Cambridge University Press, Cambridge.

Leroux B.G. and Puterman M.L. (1992) Maximum-Penalized-Likelihood Estimation for Independent and Markov Dependent Mixture Models. *Biometrics*, 48:545–558.

Rabiner L. R. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77:257–286.

Wallace C. S. (1990) Classification by minimum length inference. *AAAI Spring Symposium on the Theory and Application of Minimum Length Encoding, Standford*, pages 5–9.

Wallace C. S. and Boulton D.M. (1968) An information measure for classification. *Computer Journal*, 11:185–194.

Wallace C. S. and Freeman P.R. (1987) Estimation and inference by compact coding. *Journal of the Royal Statistical Society (Series B)*, 49:240–252.

Wallace C. S. and Dowe D.L. (1993) MML estimation of the von Mises concentration parameter. Technical report TR 93/193, Dept. of Comp. Sci., Monash Univ., Clayton, Vic. 3168, Australia. prov. accepted, Aust. J. Stat.