Model-Based Clustering of Sequential Data

Suzannah Molloy[1], David Albrecht[1], David Dowe[1], and Kai Ming Ting[2]

[1]School of Computer Science and Software Engineering
[2]Gippsland School of Computer Science and Software Engineering
Monash University, Vic. 3800, Australia
{Suzannah.Molloy,David.Albrecht,David.Dowe,KaiMing.Ting}@infotech.monash.edu.au

Abstract. Discrete, sequential data consists of multiple sequences of states, possibly containing some underlying structure or pattern. We develop two clustering approaches based on the following information theoretic criteria: Akaike's Information Criterion (AIC) and Minimum Message Length (MML), as a means of searching for any underlying structure. We compare the performance of our approaches with the method described in Cadez et al. (2000, 2001) by varying sequence length, number of states, and number of true classes within the data. The criteria are also compared using data describing navigation paths of web site users. It was observed that a penalty term is necessary to prevent overfitting of the data, and in the case of the AIC adaption, it was also necessary to incorporate prior information into the parameter estimates to ensure the criterion could handle previously unseen cases.
The number of clusters inferred and Kullback-Leibler (KL) distances indicated that Cadez et al.'s method was not particularly suitable for this type of modelling problem. Between MML and AIC, results suggested that MML achieved lower KL distances in general for both the synthetic and real-world data. Overall, the MML criterion proved to be the most promising of the three model selection criteria.

1 Introduction

Sequential data is very common in a wide range of fields including bioinformatics, astronomy, Web and telecommunications environments. Data can accumulate rapidly and has led to increased popularity of automated techniques for modelling sequential data, such as machine learning. In this investigation a standard approach to modelling sequential data is adopted, which assumes that the data has been generated by a Markov, or mixture of Markov processes. Each sequence is assumed to have been generated by one of these Markov processes. Sequences that have been produced by the same source form a group where the sequences within this group have similar characteristics to each other, while remaining distinctly different in nature to sequences from other groups. Modelling the sequential data, as presented in this work, involves identifying these groups within the sequential data set.
Clustering is a data mining method useful for identifying groups or clusters within a data set, such that members of the clusters exhibit similar behaviour

and characteristics to other members of the same cluster. Given an assumed model for an unknown number of clusters, samples are grouped together based on their statistical properties. These clusters do not necessarily represent the true groupings of the underlying data source or sources, but do identify groups of samples with significantly similar characteristics. There exists for each cluster, a statistical model whose parameters describe the characteristics of the cluster. This model can be used for probabilitstically assigning additional samples to the clusters with characteristics most statistically similar to those of the new samples, predicting the future behaviour of sequences in the case of time series or other types of sequential data, and various other applications.

Clustering of data requires the learning of three main features: 1) the number of clusters, 2) the proportion of data in each cluster, and 3) the parameters describing the statistical properties of the data in each cluster. A traditional approach to clustering represents these features in terms of a mixture model, where each cluster corresponds directly to a component of the mixture model.

Given this equivalence between clustering and mixture modelling, a standard approach to clustering is to learn a mixture model with K components, allowing the model components to represent the data clusters. The distributions describing a particular component of the mixture model are then also assumed to describe the statistical properties of the corresponding cluster. Data items can be assigned to the cluster they most likely belong to by comparing the likelihood of the data items belonging to each of the model components. Better clustering will be achieved when the components of the mixture model more closely match the statistical properties of the clusters they represent. First-order Markov models allow some of the dynamic properties of the data to be included and are used here in preference to higher order models, whose complexity increases rapidly as the order increases. Each component of the mixture model is assumed to be first-order Markov in nature, hence the underlying data source is also assumed to be a mixture of first-order Markov processes.

This approach to clustering is used in this investigation for the task of clustering sequential data. Competing mixture models learnt in this way are compared using a model selection criterion, of which there are many currently in use for various data types. This investigation adapts two penalty-based, information-theoretic criteria. These are compared with a method due to Cadez et al. [5, 6] that was developed for the purpose of clustering sequential web data. Cadez et al. also used a mixture model approach as described here.

The first penalty-based criterion we compare is adapted from Akaike's Information Criterion (AIC) [1]. This criterion is well known outside of computer science and related fields, which is more than can be said for many other model selection criteria. It has found widespread usage in a broad range of fields such as ecology [18], astronomy [25], biogeography [26], and biology [24] as some more recent examples of it's usage. However, in many of these other fields of research, very few other model selection criteria have been tested. Given that this is one of the best known and most commonly used model selection criteria in such a broad range of fields, it was deemed appropriate for this criterion to be included in the current study. As far as we are aware, there has been no prior use of AIC in this problem domain.

The second penalty-based method is an adaption of a version of the Minimum Message Length approach [28, 32, 29, 30, 27]. This criterion has been used for a variety of clustering purposes, but has not previously been used for clustering sequences in the manner described here. Previous investigations using MML have shown that the criterion is very versatile and performs well across a range of modelling problems. In addition, the MML approach has compared favourably with AIC [13]. Where AIC takes a simplistic approach to incorporating model complexity into the selection process, the MML criterion not only incorporates the complexity of the mixture model, but also takes into account the precision the parameter estimates can reasonably be stated to.

The three criteria as implemented here share the same basic components, while still remaining distinctly different from each other. These components include the log likelihood, priors, and in the case of AIC and MML, terms penalizing the complexity of the model. Differences between the criteria are largely due to these penalty terms, and the manner in which prior information is incorporated. The results presented here are intended to provide an overview of the suitability of these criteria for sequential learning problems. The techniques and criteria are quite general in nature and could be applied to a range of sequential learning applications. Cadez et al. [5, 6] chose to apply their method to modelling the behaviour of users of a web site, and this investigation also applies the techniques to similar data sets describing web user behaviour. Modelling the behaviour of humans is a recognized area of research and has numerous uses such as the identification of users belonging to a particular group [12] and prediction of customer needs [2]. Outside the realms of user modelling, other well established uses of sequence learning techniques include classification of DNA sequences and proteins [11], robotics [22] and speech recognition [21]. It has also been shown that clustering time series data can produce meaningful results [19, 16, 15, 17]. Sets of real-world, continuous valued sequences were simplified to sequences of discrete states, similar in nature to those under investigation here, and a variety of data mining tasks, including clustering, were carried out.

The remainder of the paper is organised as follows: section 2 gives details of the criteria, section 3 describes the manner in which the criteria are used to identify the best model, section 4 details the methodology, section 5 presents results and discussion, and concluding remarks are given in section 6.

2 Model Selection Criteria

The approach we take to learn a mixture of first-order Markov models is to find the set of clusters and parameter estimates that minimize a selection criterion. This section describes the three criteria we compared, and highlights significant differences between them.

2.1 Notation

The following notation is used throughout this report:

- S: the number of states
- K: the number of clusters, $\hat{K}$ indicates the number of clusters in the inferred model
- N: the number of data samples
- $\mathbf{X}$: a data set
- $\mathbf{x}$: a single data item or sequence
- L: sequence length
- θ: a Markov mixture model, distribution or parameter estimate. $\hat{\theta}$, indicates an inferred model, distribution or parameter estimate
- π: the vector of mixing proportions of the clusters
- $f(A|B)$: a conditional probability, or the likelihood of A given B

2.2 Log Likelihood Equation

The log likelihood of the mixture model is a core component of each of the criteria, and is given as follows:

$$\ln f(\mathbf{X}|\theta) = \ln \sum_{k=1}^{K} \pi_k f(\mathbf{X}|\theta_k) \tag{1}$$

$$\text{where } f(\mathbf{X}|\theta_k) = \prod_{i=1}^{N} f(x_{i,1}|\theta_k^I) \prod_{j=2}^{L} f(x_{i,j}|x_{i,j-1}, \theta_k^T) \tag{2}$$

and $f(x_{i,1}|\theta_k^I)$ and $f(x_{i,j}|x_{i,j-1}, \theta_k^T)$ are, respectively, the mutinomial distributions of the initial states and the transitions between states.

2.3 Number of Free Parameters

The penalty-based criteria, AIC and the MML approximation, both incorporate the number of free parameters into their penalty terms. Adding an additional cluster to the inferred model increases the number of free parameters and the complexity of the model. For both penalty-based criteria, an additional cluster will only be added if the resulting saving in the encoding of the data is sufficiently large to make up for the additional cost in the penalty term due to the more complex model.
For a mixture of K first-order Markov models, the following free parameters exist:

- A single parameter for the number of clusters, K
- The mixing proportions or mixture weights of the clusters, giving $K-1$ free parameters
- For each of the K clusters, $S-1$ free parameters for the initial state distributions

- For each of the K clusters, a transition table with $S(S-1)$ free parameters

The total number of free parameters for a mixture of K first-order Markov models with S states is the sum of the items listed above. This results in a total of KS^2 free parameters.

2.4 Cadez et al.'s criterion

Cadez et al.'s criterion applies no penalty for model complexity, and for sequences of a fixed length, reduces simply to the negative log likelihood and a Dirichlet prior:

$$Cadez = -\ln h(\hat{\theta}) - \ln f(\mathbf{X}|\hat{\theta}) \tag{3}$$

In equation (3), $h(\theta)$ refers to the Dirichlet prior used by Cadez et al. to smooth out the parameter estimates, preventing zero probabilities occuring.

In carrying out their investigation, Cadez et al. [5, 6] divided their data into a training and test set. The implementation required an estimate of the number of clusters, $\hat{K}$, to be set at initialization, then the best model with this number of clusters was inferred from the training set. Many different values of $\hat{K}$ were used, creating a collection of models which could then be compared to select the best model for the data set, and hence find the best value of $\hat{K}$. The models were compared using the out-of-sample predictive log score, calculated by equation (3), and the best model was considered to be the one which minimised this score on what Cadez et al. reffered to as the test data. Strictly speaking, this data set is not really a test set, as it is used in the model selection process. Rather, it is a secondary training set which is used to "tweak" the parameters, in this case the value of $\hat{K}$. A genuine test set should be kept completely separate from the model selection process and used only for evaluation of the final model. There appears to be no test set for evaluation in Cadez et al.'s work and it is assumed that subjective evaluation of the results through the visualization procedure determines the success of the clustering technique.

When splitting data into training and test sets and designing the experimental procedure, care must to taken to avoid a situation described by Russell and Norvig [23](ch. 18, sec. 18.3) as "peeking". When peeking occurs, the learning algorithm is allowed to see the test data and use this information in the selection of the best model. Essentially, the test data is used to "tweak" the parameter values, optimising them for the current data set, and selecting the best model by the performance on the test data. This is a situation to be avoided since it does not allow proper evaluation of the results unless an additional, previously unseen test set is used for evaluation. Depending on the purpose for which the clustering is being employed, this type of evaluation may be suitable. However, for a general approach to clustering sequential data, a more objective approach to model selection is required.

For the implementation of Cadez et al.'s criterion tested here, the model selection criterion is taken to be a criterion based on minimizing the negative log likelihood of a data set combined with a Dirichlet prior.

2.5 Akaike's Information Criterion

The AIC criterion was originally derived from an approximation to the expected Kullback-Leibler distance between the true model and the inferred model [1]. A simple penalty term based on the number of free parameters in the inferred model is included:

$$\mathrm{AIC} = -\ln h(\hat{\theta}) - \ln f(\mathbf{X}|\hat{\theta}) + \hat{K}S^2 \tag{4}$$

The penalty term is equivalent to the total number of free parameters to be estimated given that there are S states and $\hat{K}$ clusters found in the data. Preliminary work found that the AIC criterion was unable to handle previously unseen cases in the test sets. For this reason a Dirichlet prior, $h(\hat{\theta})$, identical to that used by Cadez et al. [5, 6] was included.

2.6 The Minimum Message Length Criterion

The Minimum Message Length (MML) [28, 32, 30, 27] criterion has been shown to perform better than AIC in previous investigations [13], and is a more generalised approach [8, 9] than AIC. This criterion is based on information theory, viewing the problem as one of minimizing the length of a two part message. The first part of the message describes the model, and the second part describes the data given the model. The description must be sufficiently detailed to allow a receiver with minimal prior knowledge to unambiguously decode the message and retrieve the model and data. This approach assumes the best model is the one that results in the minimum message length. Depending on the assumptions made about the prior knowledge of sender and receiver, messages may be constructed in a number of ways. The MML variant used in this investigation assumes that both sender and receiver know the value of N, the number of attributes and the type of distribution describing each attribute. In this scenario, the attributes consist of the relative abundances, distributions for initial states and transition tables for S states. All attributes are described by multinomial distributions. Given the assumed prior knowledge, the first part of the message describes four main parts:

1. The number of clusters
2. The mixing proportions or mixture weights
3. The initial state distributions for each cluster
4. The transition distributions for each cluster

There are a variety of ways a message can be constructed using this framework, leading to different estimates of the expected message length. In this investigation we will use an approximation to the Minimum Message Length, which we will call the MML I_1 criterion, as described by Wallace [27]. Preliminary investigations indicated that this version of the MML criterion was the most suitable for this type of problem. MML I_1 combines the lengths of the components listed earlier and the second part of the message describing the data. The general form of the MML criterion, using the derivation given in [32] is as follows:

$$MsgLen = -\ln h(\hat{\theta}) - \ln f(\mathbf{x}|\hat{\theta}) + \frac{1}{2}\ln F(\hat{\theta}) + c \tag{5}$$

where $h(\hat{\theta})$ is the prior probability density function of the model $\hat{\theta}$ and is assumed to be uniform, $F(\hat{\theta})$ is the expected Fisher Information and c is a constant depending on the number of states, S. This constant is a part of the penalty that MML applies for model complexity.

For a mixture model, equation (5) can be approximated by the following:

$$\text{MML I}_1 \approx \sum_{k=1}^{\hat{K}}(MsgLen_k) - \ln \hat{K}! + MsgLen_\pi + \hat{K}\ln 2 \tag{6}$$

$$\text{where } MsgLen_k \approx -\ln h(\hat{\theta}_k) - \sum_{s=1}^{S}\ln \Gamma(n_{ks}+1) + \ln \Gamma(N_k + S) + \frac{S-1}{2} \\ -\frac{S-1}{2}\ln 12 + \frac{S-1}{2}\ln 2\pi \tag{7}$$

Where n_{ks} refers to the number of occurances of state s in class k, and N_k refers to the total number of observations in class k. For the mixture model case given in equation (6), the individual classes are encoded as a set of multinomial distributions, where $MsgLen_k$ is the length of the message encoding class k, and is calculated using equation (7). The mixing proportions are also encoded as a multinomial distribution with data $\{N_1, N_2, \ldots, N_K\}$. The message length, $MsgLen_\pi$, is calculated using equation (7). The numbering of classes is arbitrary, hence the $\ln \hat{K}!$ term, and $\hat{K}\ln 2$ is the length of encoding the number of classes. These additional terms maintain consistency with the sender/receiver framework, ensuring that the hypothetical message could be uniquely decoded by a receiver.

Mixture modelling as carried out here uses partial assignment of samples to classes, as opposed to total assignment. As a result of this, the number of samples assigned to a class, and the number of occurrences of states in a class can take real values. For more detailed information regarding the derivation of equations (6) and (7), see [27].

Preliminary investigations showed that other methods of calculating the message length were inconsistent when applied to multinomial data. Some versions of MML incorporate simplifying assumptions, and when certain conditions are not met, the criterion begins to break down. Small sample sizes can be a contributing factor to this problem, and can occur frequently in mixture modelling of sequential data, particularly as the number of parameters increases. MML I_1, equation (6), is a variant of MML that behaves in a consistent manner with multinomial distributions in the presence of small sample sizes.

3 Minimising the Criteria

For each criterion the best mixture model describing the clustering of the sequential data corresponds to the model which minimizes the relevant criteria. The main steps in the process of learning the best model are 1) determining the optimal number of clusters in the inferred model, $\hat{K}$, 2) learning their mixing proportions, π, and 3) inferring parameter estimates for each cluster. The three criteria use a search algorithm adapted from the clustering programme Snob [28, 3, 31], which carries out these two tasks concurrently. An estimate of K is determined by a random process of splitting and merging the clusters.

3.1 Learning Parameter Estimates via the EM Algorithm

The parameters are learned using a standard technique based on the Expectation Maximization (EM) algorithm [10]. The EM algorithm is extrememly versatile and has been used in a great many previous applications and studies. The EM algorithm uses an iterative approach to infer a set of parameter estimates and mixing proportions from a data set given an estimate for the true number of clusters K. This estimate, $\hat{K}$, remains fixed throughout the process. The data is considered to be composed of two components:

- Observed data - the data available to infer the model from. In our case, the set of sequences of discrete states
- Unobserved data - the label of the class to which each sample, or in this case, each sequence belongs to. This data is to be inferred from the observed data set.

The ith sequence, $\mathbf{x}_i$, therefore consists of the observed sequence of discrete values, x_i, and the set of unobserved data, $\mathbf{z}_i$, describing the assignment of $\mathbf{x}_i$ to each of the $\hat{K}$ clusters:

$$\mathbf{x}_i = (\mathbf{z}_i, \{x_{i,1}, \ldots, x_{i,L}\})$$

In the case of partial assignment of sequences to clusters, as used in this investigation, $\mathbf{z}_i = \{z_{i,1}, \ldots, z_{i,K}\}$.

At initialization sequences are randomly assigned to clusters and initial estimates of the model parameters and unobserved data values are calculated. From this point, a two-step process is repeated until the algorithm converges to a set of stable parameter estimates. These two steps are the Expectation (or E) step, where the expected values of the unknown variables, or the unobserved data set, are calculated, and the Maximization (or M) step, where sequences are assigned to the cluster they most likely belong to.

While the EM algorithm is guaranteed to converge on stationary values, it must be noted that these are not necessarily the optimum global values, and may only be local optima.

3.2 Finding the Optimum Number of Clusters

As mentioned at the beginning of this section, an adaption of the clustering programme Snob [28, 3, 31] is used to search for and identify the model giving the best fit for the data. The EM algorithm, which requires the number of clusters to be constant throughout, is embedded into the search procedure. At initialization, an estimate of the number of clusters, $\hat{K}$, is chosen randomly. Samples are randomly assigned to these clusters and initial parameter estimates and mixing proportions are calculated. The search algorithm then alternates between two main procedures:

- Running the EM algorithm using the current value of $\hat{K}$ and the current model as the initial start point. The EM algorithm runs through the iterative two-step process, converging towards stable estimates of the parameters.
- Selecting one of three procedures at random, intended to trial different values of $\hat{K}$ and help avoid local optima. These procedures include 1) selecting a single cluster and splitting it into two separate clusters, 2) selecting two clusters and merging them into one cluster, and 3) reverting to the current best model. Clusters are selected at random.

In general, splitting, merging and reverting to the best model are carried out when there has been no significant improvement in the value of the selection criterion from one iteration of the EM algorithm to the next. Occaisionally however, these procedures will be carried out regardless.

As mentioned in section 3.1, the EM algorithm can become trapped in a local minima. The random processes described above partially address this problem by allowing the EM algorithm to restart from an altered clustering of the samples from time to time, but this still does not guarantee always finding the global optimum.

4 Methodology

We wanted to compare the three criteria under conditions similar to those found by Cadez et al. in [5, 6], to the known conditions in the real-world data set used here, and under conditions that enabled us to see the behaviour patterns resulting from changes to various data attributes. The synthetic data we generated was based around the following three attributes: the number of states, S, the length of the sequences, L, and the number of true classes, K. The values 5 and 20 were used as the low and high values respectively in a 2^3 factorial design [4], resulting in eight different (S, L, K) tuples.

Factorial designs have been discussed extensively in experimental design literature [4, 20, 7, 14] for quite some time. This approach to experimentation allows the effects of multiple factors to be investigated. In our case, the factors are the attributes S, L, and K, and we use the factorial design to observe the effects these factors have on the ability of a model selection criterion to select a good, statistical model of a data set. By allowing the factors to take on values at different levels, in our case a high level of 20 and a low level of 5, it becomes possible to quantitatively observe the effect of each factor on the results, and in addition identify dependencies or interactions that may exist between the factors themselves. Despite providing a very structured and powerful approach to experimental investigation, factorial designs seem to be absent from much of the literature within computer science and related fields.

For each tuple, 20 first-order Markov models with the required properties were generated. These models correspond to the true models. From each of these, a training set of 800 sequences was generated, and 10 test sets of 200 sequences each. Each Markov model describing a class k consists of a multinomial distribution, $f(s_1|\theta_k^I)$, describing the probability of the initial state s_1, and a set of S multinomial distributions, $\{f(s_j|s_i, \theta_k^T)\}_{i=1}^S$, describing the transition from state s_i to state s_j. All these distributions were generated randomly. Moreover, each sequence in the synthetic data was generated by first randomly selecting a component, then selecting an initial state according to $f(s_1|\theta_k^I)$, then generating the remaining transitions using the distributions $\{f(s_j|s_i, \theta_k^T)\}_{i=1}^S$.

Models learned, using each of the three criteria, from each of the training sets, for each of the different (S, L, K) combinations were tested on each of the 10 corresponding test sets. 20 training sets were run to completion, resulting in 200 sets of results for each of the criteria, for each (S, L, K) combination.

The clustering criteria were allowed to run for 500 iterations of the EM algorithm on each training set. Partial assignment was used throughout, except at initialization of the EM algorithm when sequences were randomly segmented into initial classes.

5 Results and Discussion

The performance of the criteria was measured by: 1) the number of clusters found and 2) an approximation to the Kullback-Leibler (KL) distances from the true to the inferred models. The behaviour of the selection criteria is affected by the amount of data available to infer the parameter estimates, and discussion of this is included in the analysis of the results. Finally, the model selection criteria are tested on a real-world data set, similar in nature to Cadez et al.'s, consisting of sequences of discrete states representing the paths that users of a web site have followed while moving through the site. The performance of the criteria on the real-world data set is analysed predominantly using the mean difference in bit costs.

5.1 Amount of Data Per Free Parameter

Given the changing values of S, L, and K in the true source models, the average amount of data available per free parameter varies considerably, and can have a significant effect on the behaviour of the penalty-based selection criteria. For this reason, the average number of observations per free parameter was taken into account during analysis of the results presented here. In general, a data set generated by a more complex mixture of sources or models, has a larger number of free parameters. If good parameter estimates are to be inferred, a larger data set is needed. For the synthetic data, the number of sources generating each data set is known, as is the number of states and the length of the sequences. With this information, the total number of free parameters in the true model, as described in section 2.3, and the average number of observations per parameter can be calculated. Table 1 gives the average number of observations per parameter in the true model, for each of the eight experimental set-ups in the factorial design. In some instances the amount of data per parameter is very low, e.g. 0.5 observations per parameter for the $(20, 5, 20)$ scenario. These values are based on the true number of classes, K, rather than the number of clusters inferred by the model selection criteria, $\hat{K}$.

Table 1. Average number of observations per parameter

S	L	K	No. parameters	No. observations	Av. no. observations per parameter
5	5	5	125	4000	32
5	5	20	500	4000	8
5	20	5	125	16 000	128
5	20	20	500	16 000	32
20	5	5	2000	4000	2
20	5	20	8000	4000	0.5
20	20	5	2000	16 000	8
20	20	20	8000	16 000	2

5.2 Number of Clusters Found

Table 2 gives the median number of clusters found by the methods for each combination of the number of states, S, the length of the sequences, L, and the number of true classes present in the data, K. The minimum and maximum number of clusters found over all the training sets is also included in brackets. There is clear disagreement between the number of clusters found by Cadez et al.'s method compared to AIC and MML I_1. Cadez et al.'s criterion showed a very strong tendency to overfit the data, and find too many clusters. In contrast, the penalty-based methods, AIC and MML I_1, tended towards more conservative estimates of K, and in most cases, found fewer clusters than there were true classes present in the data sources. The tendency for Cadez et al.'s method to overfit the data is easily explained by the lack of penalty term. Penalty terms effectively add an additional cost whenever a new cluster is added to the mixture model. The cluster will only be retained if it provides a decreased cost in the likelihood component that more than compensates for the increased cost of the additional cluster resulting in an overall reduction of the criterion's score. If no

Table 2. Median number of clusters found. Minimum and Maximum clusters found included as (Min, Max).

S	L	K	MML I_1	AIC	Cadez et al.
			Med. (Min, Max)	Med. (Min, Max)	Med. (Min, Max)
5	5	5	1, (1,1)	3, (2,4)	28, (23,36)
5	5	20	1, (1,3)	4, (2,4)	25, (19,34)
5	20	5	4, (3,5)	5, (3,5)	29, (19,40)
5	20	20	7, (1,22)	15, (12,17)	45, (37,53)
20	5	5	1, (1,1)	1, (1,1)	31, (17,42)
20	5	20	1, (1,1)	1, (1,1)	29, (17,38)
20	20	5	3, (2,5)	3, (3,5)	30, (11,35)
20	20	20	1, (1,2)	4, (2,4)	27, (17,37)

penalty term is included, there is no reason not to continue adding new clusters to the inferred mixture model which leads to a model that is too specific to the training data set.

For almost all tests, both MML I_1 and AIC found fewer clusters than there were true classes present in the data. In a small number of cases, larger values of $\hat{K}$ are inferred. The cause of this appears to be the presence of either local optima, or the splitting and merging algorithm choosing to try a split rather than a merge. However, the general tendency was to underfit the data, and is somewhat more difficult to explain than the overfitting of Cadez et al.'s criterion. A number of possible reasons exist which may contribute to this behaviour. Firstly, the size of the training sets may not have been sufficiently large enough to identify all the true classes within the data. In addition, the mixture weights of the classes were determined randomly from a uniform distribution. Particularly for large values of K, the number of examples from each class was probably relatively small, and given the length of the sequences, certainly not enough samples to learn the characteristics of the class and identify it as a separate cluster. A much larger data set that contained many examples from all classes could have provided sufficient examples from each class to allow identification of that class within the data. Similarly, longer sequence lengths would have provided more data from which to learn parameter estimates for each class. AIC and MML I_1, due to their penalty terms, will be sensitive to the amount of data that is available for identifying classes and inferring parameter estimates for the corresponding inferred clusters. Table 1 lists the number of parameters in the true models, half of which are in the order of thousands. Increasing the number of classes in the true model from the low level to the high level causes the number of parameters to increase by a factor of four. Even more importantly, increasing the number of states in the model causes a sixteen-fold increase in the number of parameters. Adding additional clusters to an inferred model has serious consequences for both AIC and MML I_1 in terms of the number of additional parameters that must be estimated and hence the number of clusters that will be found. Increasing K and particularly S led to greater complexity in the data and models and a more difficult learning problem. Each additional cluster became a costly addition requiring significant savings to be justified.

MML I_1 also requires that there be sufficient data to justify the accuracy of the parameter estimates, which will also contribute to a more conservative estimate of the number of clusters when there is little data available.

In addition to the points already given, there is another possible contributing factor to the tendency for AIC and MML I_1 to underfit. Given the manner in which the synthetic data was generated, i.e. constructing each first-order Markov model from a set of multinomial distributions generated randomly from a uniform distribution, the true models may be very similar to each other, particularly as the values of K and S increase. While there may technically be K true classes, the statistical characteristics of samples from some of these classes may be similar enough that there really is no reason for these classes to exist as separate clusters in the inferred model. Figures 1 and 2 are intended to illustrate the similarity in the multinomial distributions describing the mixture models. The two examples given, (5, 5, 5) and (20, 20, 20), were selected as they show the two extremes in the number of parameters requiring estimation. As described in section 4, each class in the true model is described by a first-order Markov model. This Markov model can itself be described as a set of multinomial distributions, one for the initial state probabilities and one for each row of the transition table. To generate a multinomial distribution with S states, a total of $S - 1$ values

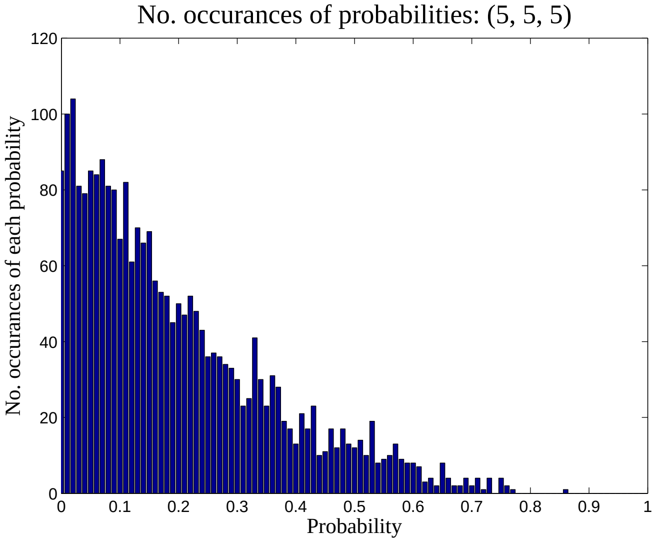

Fig. 1. Histogram showing the number of transition probabilities in the M = 5, L = 5, K = 5 mixture models.

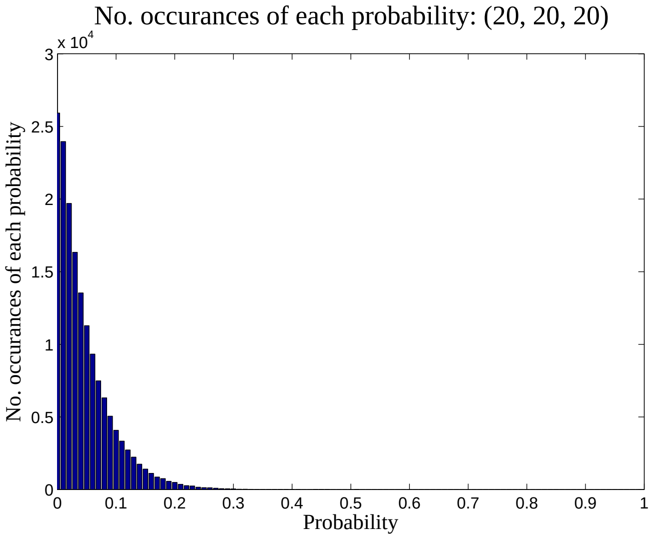

Fig. 2. Histogram showing the number of transition probabilities in the M = 20, L = 20, K = 20 mixture models.

in the range $[0, 1]$ are generated randomly from a uniform distribution. These values are ordered, and the size of the intervals between consecutive pairs represent the probability in the multinomial distribution of each individual state, s_i, where $i = 1, \ldots, S$. This process is repeated for every multinomial distribution required to describe the mixure model. Figures 1 and 2 show the total number of probabilities, generated in this way, that fall within a certain range. It is clear from the two histograms, that there is a definite skew towards the lower values, and this becomes all the more obvious when S is large, i.e. 20 as in Figure 2. This shows us that the probabilities in the multinomial distributions rarely have large values, and tend to be quite uniform from one state to the next, and from one distribution to the next. Particularly for $S = 20$, there is little to distinguish one multinomial distribution from another, and similarly when combined to form first-order Markov models, there is little to distinguish one class model from another. The result of this is that the penalty based methods are unlikely to find much justification to separate the sequences into separate clusters.

Increasing L effectively increased the amount of data available for learning, clearly seen by an increase in the amount of data per parameter in Table 1. The result was that both MML I_1 and AIC were more likely to increase the number of clusters found. This was not always the case for Cadez et al.'s method, which showed varied behaviour as L increased, although the number of clusters always remained high. Increasing the length of the sequences provides additional data for learning, without increasing the number of sequences. The additional data gained by increasing the length of a sequence is known to be produced by the same source as the rest of the data in that sequence and hence provides information about a particular source. In contrast, additional information is available if the number of sequences in the data set is increased, but these new sequences may belong to one or more separate classes of which no prior examples have been seen. Therefore, in terms of learning good estimates for the model parameters, it is quite likely that having a data set containing long sequences will be more useful than having a data set consisting of many short sequences. Preliminary testing of this theory has begun, and tests so far have indicated that this is the case.

The choice of priors may also impact the number of clusters found. The prior used in MML approximations is a harsher prior than the Dirichlet prior used for Cadez et al.'s method and also AIC as implemented here. The priors flatten the parameter estimates and a certain amount of data is required before the effects of the priors are effectively swamped. In comparison to the Dirichlet prior, the MML prior requires a greater quantity of data for this swamping to occur. This will again contribute to a more conservative estimate of the number of clusters present in the data particularly when the data is from very uniform sources such as it is here.

5.3 Comparison of Kullback-Leibler Distances

The KL distance gives an asymmetric value indicating the expected inefficiency of a code constructed with an inferred model compared to the true model. The lower the KL distance, the more efficient is the encoding of the data by the inferred model. The closer the KL distance is to zero, the closer the inferred model is considered to be to the true model. Given that this investigation uses a numerical approximation to the average KL distance per state symbol (from true to inferred model), negative values can occur. This indicates that the inferred model gives a more efficient encoding of the data than the true model. Equation (8) gives the approximate KL distance used here:

$$KL(\theta, \hat{\theta}) \approx \frac{1}{LN_{test}} \sum_{i=1}^{N_{test}} \ln \left(\frac{f(\mathbf{x}_i|\theta)}{f(\mathbf{x}_i|\hat{\theta})} \right) \tag{8}$$

Table 3 gives the mean KL distances for each of the three criteria and Table 4 gives the main effects and the level to which the factors S, L and K interact with each other (two and three-factor interactions). As discussed in section 4, a factorial design such as this allows the behaviour of the three selection criteria to be compared while the values of multiple factors or variables are adjusted. Rather than investigating the effects of one factor at a time, a 2^3 factorial design allows the effects of three factors, in this case S, L and K, to be compared simultaneously. A high level and a low level, in our case 20 and 5 respectively, are used to investigate the effects of each factor on the behaviour of the model selection criteria. In addition, the factorial design gives an indication of dependencies between the factors themselves. The main effects given in Table 4 indicate the average change in the average KL distance that occurs when the values of the factors S, L and K are increased from the low value, 5, to the high value, 20. A positive value indicates an overall increase in KL distance, whereas a negative value indicates a decrease. Interaction effects tell us if the values of the KL distances that result from changing the value of a main factor, are significantly influenced by the values of the other factors, and what the resulting effect of this influence will be on the average KL distance.

MML I_1 wins three tests, AIC wins four and the remaining one test out of the total eight is won by Cadez et al.'s criterion. Overall, Table 4 shows MML I_1 achieves the lowest average KL distance across all tests, despite not acheiving the lowest KL distance for the majority of the eight tests. MML I_1 is followed by AIC, then Cadez et al.'s criterion, which is significantly larger than the penalty-based criteria. A lack of penalty term is the main difference between Cadez et al.'s criterion and the other two, leading to severe overfitting of the data. It is reasonable to assume that this is one of the most significant contributing factors to the poor performance. Tables 3 and 4 show that the KL distances for Cadez et al.'s criterion always increase as the level of S is increased. Table 4 also shows that all other main effects and interactions lead, on average, to a decrease in the average KL distance (from true to inferred model), with one exception seen in Table 3 when $S = 20$ and $L = 5$. This exception corresponds to the scenario where the inference problem is at it's most difficult, with a very low quantity of data per parameter with which to infer estimates. Despite this, the results show that the performance of Cadez et al.'s criterion is not comparable to either MML I_1 or AIC. The behaviour and performance of Cadez et al.'s criterion strongly suggests that some tradeoff must be made between model complexity and ability to provide an efficient encoding of the data.

Table 3. Mean KL distances. Bold face indicates the criteria that performed significantly better than the others.

S	L	K	MML I_1		AIC		Cadez et al.	
			mean	S.D.	mean	S.D.	mean	S.D.
5	5	5	**0.0047**	**0.011**	0.0334	0.011	0.2168	0.030
5	5	20	0.0433	0.004	**0.0330**	**0.003**	0.1563	0.011
5	20	5	0.0068	0.001	**0.0057**	**0.001**	0.0266	0.002
5	20	20	0.0481	0.012	0.0221	0.004	**0.0144**	**0.007**
20	5	5	0.1239	0.013	**0.1126**	**0.010**	1.3133	0.047
20	5	20	**-0.0648**	**0.005**	0.0245	0.007	1.3603	0.049
20	20	5	**-0.0120**	**0.010**	0.1916	0.049	0.6482	0.069
20	20	20	0.2338	0.006	**0.2211**	**0.009**	0.4120	0.034

Increasing the level of K results in very similar behaviour for MML I_1 and AIC. This change leads to an increase in the KL distance, with the exception of the $S = 20$ and $L = 5$ case for both MML I_1 and AIC, and for AIC only,

Table 4. Effects and interactions of factors for KL distances in Table 3.

Effect on average KL distance	Estimate ± standard error		
	MML I_1	AIC	Cadez et al.
Average	0.0480 ± 0.0020	0.0805 ± 0.0046	0.5185 ± 0.0092
Main effects			
No. states S	0.0222 ± 0.0020	0.0569 ± 0.0046	0.4149 ± 0.0092
Length L	0.0212 ± 0.0020	0.0296 ± 0.0046	-0.2432 ± 0.0092
No. true classes K	0.0171 ± 0.0020	-0.0053 ± 0.0046	-0.0327 ± 0.0092
Two factor interactions			
$S \times L$	0.0195 ± 0.0020	0.0393 ± 0.0046	-0.1602 ± 0.0092
$S \times K$	-0.0028 ± 0.0020	-0.0093 ± 0.0046	-0.0146 ± 0.0092
$L \times K$	0.0546 ± 0.0020	0.0168 ± 0.0046	-0.0294 ± 0.0092
Three factor interaction			
$S \times L \times K$	0.0540 ± 0.0020	0.0126 ± 0.0046	-0.0415 ± 0.0092

the $S = 5$ and $L = 5$ case. The $S = 20$ and $L = 5$ scenario was also an exception for Cadez et al.'s criterion, and as mentioned earlier, corresponds to the most difficult learning problem. The more conservative approach to inferring a model taken by MML I_1 and AIC may be beneficial in situations where there is little data per parameter. For the $(20, 5, 20)$ and $(20, 20, 5)$ cases, the average KL distance for MML I_1 was found to be negative. Both correspond to scenarios where there is little data per parameter, but as the behaviour is not consistent for other cases where there is little data per parameter, i.e. $(20, 5, 5)$ and $(20, 20, 20)$, it cannot be entirely due to the complexity of the modelling problem.

The behaviour of AIC as the level of L increases also seems dependent on S. If S is at a low level, increasing L results in reduced KL distance, whereas high levels of S lead to increased values for the KL distance. Increasing S rapidly increases the number of parameter estimates to be inferred, and while increasing L provides more data, this increase in information to infer estimates may not be enough to allow for the increased complexity of the problem. It may be expected that the KL distances should decrease as more data becomes available, but in the case of MML I_1, at least under these conditions, increasing L led to greater KL distances. The exception in this case was when $S = 20$ and $K = 5$.

The main effects in Table 4 indicate that both MML I_1 and AIC are significantly effected by the level of S, and an increase in this factor leads on average to an increase in the KL distance. However, significant two-factor interactions exist, and in the case of MML I_1, significant three-factor interactions. It is difficult to identify a consistent pattern of behaviour for the criteria, indicating that further tests are required. The behaviour of AIC is a little clearer than that of MML I_1, most likely due to the lack of three-factor interaction. In general for AIC, increasing S leads to an increase in the KL distance, with the exception of the case when $L = 5$ and $K = 20$.

A number of points discussed in section 5.2 explaining the underfitting of the penalty-based methods may also help to explain the behaviour observed in the

KL distances. The uniform manner in which the synthetic models and data sets were generated and the quantity of data available to infer parameter estimates makes it difficult for the criteria to distinguish the different groups within the data sets. This led to underfitting of the data. There is also a strong likelihood that the true models are all very similar in nature, again due to the random generation of the models from a uniform distribution. This uniformity of the true models and underfitting of the data may be obscuring any patterns in the behaviour of the criteria caused by the changing factor values, which results in irregular trends in the KL distances. To some degree, larger data sets could help address these issues.

5.4 Clustering Web Navigation Path Data

The web data used here was very similar in nature to that used by Cadez et al. [5, 6]. The sequences consisted of a number of states, where each state indicated the type of page the user had moved to. A total of 13 page categories existed, and sequence lengths ranged from a single state to a maximum length of 897, with the average at around 21 states per sequence. In total, 10, 756 sequences were randomly divided into 10 approximately equally sized sets. Given that the values of S and L are known to be 13 and 21 respectively, we can use the results from Table 3 to select which of the three criteria we would expect to perform better. The values of S and L that most closely match those of the web data are $L = 20$ and either $S = 20$ or $S = 5$, with more bias towards $S = 20$. Of the eight (S, L, K) scenarios, the KL distances from Table 3 suggest that AIC might be expected to perform a little better than MML I_1. This is expected as AIC performs better than MML I_1 in three of the four scenarios with $L = 20$ and either $S = 20$ or $S = 5$.

Table 5. Results summary for web data clustering: Median no. of clusters found, mean difference in bit costs of the criteria compared to the cheapest (MML I_1).

Method	MML I_1	AIC	Cadez et al.
Median (Min, Max)	1, (1,2)	3, (2,3)	32, (22,36)
Mean bit cost difference	0.000	1.525	16.976

Table 5 gives a summary of the results of clustering the web data and shows that MML I_1 actually out-performs both Cadez et al.'s criterion and AIC. The data in Table 5 includes the median, minimum, and maximum number of clusters found by each of the criteria and the mean difference in bit costs [9, sec. 11.4.2] using the cheapest mehod (MML I_1) as the base case. The mean difference in bit costs is similar to the KL distance, but the data in question is not assumed to have been generated by either of the models involved in the calculation.

As with the synthetic data, the number of clusters estimated by Cadez et al.'s method is significantly higher than MML I_1 and AIC. Given the estimated number of clusters found by MML I_1 and AIC, and the known values of S and L, the test sets that the web data most closely resemble can be narrowed down further, and are most similar to those with values (20, 20, 5) and (20, 20, 20). As MML I_1 acheives a more efficient encoding that AIC, holding more in common with the (20, 20, 5) synthetic data scenario, it suggests that the web data contains only a small number of true classes. In addition, the more efficient encoding due to MML I_1 may indicate that the web data sets contain sequences that have been generated by less uniform distributions such as those described by the histograms in Figures 1 and 2.

The value of the mean bit cost difference gives an indication of the mean difference in coding efficiency of a model compared to the base case. As with KL distance, a positive value indicates the base case is able to encode the data more efficiently than the other model, whereas negative values indicate the opposite. Over all 10 web data sets, the mean bit cost differences were positive, indicating that MML I_1 inferred models that more efficiently encoded the data using a smaller number of clusters than AIC, and models inferred by both AIC and MML I_1 are far more efficient than Cadez et al.'s criterion, and less complex.

6 Conclusion

Lack of a penalty term in the criterion used by Cadez et al. resulted in overfitting, and generally larger estimates of the number of clusters than what was actually present. This criterion also produced significantly larger KL distances when compared with MML I_1 and AIC. This indicates that the criterion used by Cadez et al. could be greatly improved by the use of a penalty term.

Significant multi-factor interactions for AIC and MML I_1 may partially explain the lack of any clear pattern emerging within the results. Further investigations are necessary to enable accurate interpretation of these interations. The low estimates of the number of clusters by MML I_1 indicates there was insufficient evidence to suggest that a greater number of distinct clusters existed within either the synthetic or web data. The true class models were generated using uniform random distributions, and particularly for cases where S or K was high, there may have been little difference between the true models.

For sequence learning problems it may be possible to determine, prior to clustering, what the number of states and average sequence lengths are, and selection of a criterion can then be based on known performance on synthetic data for specific values of S and L. Given that the general trend is for Cadez et al.'s criterion to overfit, the best choice of criterion is between AIC and MML I_1. From Table 3, we can see that when $S = 5$ and $L = 20$, AIC seems to be a better choice, but for other combinations, either MML I_1 or AIC could be considered. Given the performance of MML I_1 on the web data, and the possiblility that a data set may not be generated from such uniform sources as used here in the synthetic data, there is stronger evidence that MML I_1 would be a better choice of clustering criterion.

The behaviour observed on synthetic data was reflected in the clustering of the web data. Overall, MML I_1 resulted in a set of models providing the most efficient encoding of the data, indicated by the mean bit cost differences in Table 5. Therefore, the models inferred by MML I_1 were a closer fit to the data than those of either AIC or Cadez et al.'s criterion.

While MML I_1 did not win as many tests outright as AIC on the KL distances for the synthetic data, overall it did show a lower average KL distance on the synthetic data sets, as well as a more efficient encoding of the real world data. Both MML I_1 and AIC show some promise as clustering criteria in this domain. Future work includes further investigation into variations of the Minimum Message Length criterion and application of the model selection criteria to the data set used by Cadez et al [5,6]. Sequences of amino acids describing the primary structure of proteins consist of long sequences of discrete states. Application of these techniques to modelling protein secondary structure or function from the primary sequences is under consideration, possibly extending the work of Edgoose et al. [11]. There is potential for time series clustering applications by combining the methods described here with a suitable technique for discretising continuous valued sequences such as the procedure used in SAX (Symbolic Aggregate ApproXimation) [19,16,15,17].

The uniformity of the synthetic data sources used in this investigation favoured the use of AIC. The flattening constants, in the form of Dirichlet priors, were far less harsh in their effect on the parameter estimates than the flattening constants used for the MML I_1 criterion. It is expected that sources showing a greater level of statistical variation between each other will be better modelled using the MML I_1 criterion, as the harsher flattening constants will help to smooth some of the more extreme estimates that may be inferred from a given data set. In addition, preliminary findings indicate that a small set of long sequences produces better results when learning model parameters compared to learning from a set of many short sequences. Further investigations into the effects of sequence length will also be carried out.

The first author was supported by Australian Research Council (ARC) Linkage Grant LP0219725 with Industry partner Telstra.

References

1. H. Akaike. Information Theory and an Extension of the Maximum Likelihood Principle. In *Proc. of the 2nd Intl. Symp. on Information Theory*, pages 267–281. Akademiai Kiado, Budapest, 1973. Reproduced in *Breakthroughs in Statistics*, Vol.I, Foundations and Basic Theory, S. Kotz and A.L.Johnson eds. Springer-Verlag, New York, 1992, pp610-624.
2. D.W. Albrecht, I. Zukerman, and A.E. Nicholson. Pre-sending Documents on the WWW: A Comparative Study. In *IJCAI - The Sixteenth Joint Conference on Artificial Intelligence*, pages 1274–1279, Stockholm, Sweden, 1999.
3. D.M. Boulton and C.S. Wallace. A Program for Numerical Classification. *Computer Journal*, 13(1):63–69, 1970.

4. G.E.P. Box, W.G. Hunter, and J.S. Hunter. *Statistics for Experimenters.* John Wiley & Sons, 1978.
5. I. Cadez, D. Heckerman, C. Meek, P. Smyth, and S. White. Visualization of navigation patterns on a web site using model-based clustering. In R. Ramakrishnan and S. Stolfo, editors, *Proc. of the 6th ACM SIGKDD Intl. Conf. on Know. Dis. and Data Mining.* New York: Assoc. for Computing Machinery, 2000.
6. I. Cadez, D. Heckerman, C. Meek, P. Smyth, and S. White. Model-Based Clustering and Visualization of Navigation Patterns on a Web Site. Microsoft Research Tech. Report: MSR-TR-00-18, September 2001. Redmond, WA 98052, United States.
7. W.G. Cochran and G.M. Cox. *Experimental Designs.* Jossey-Bass, 2nd edition, 1957.
8. J. W. Comley and D.L. Dowe. General Bayesian Networks and Asymmetric Languages. In *Proc. 2nd Hawaii Intl. Conf. on Stats. and Related Fields*, June 2003.
9. J.W. Comley and D.L. Dowe. Minimum Message Length and Generalised Bayesian Nets with Assymetric Languages. *Advances in Minimum Description Length: Theory and Applications*, pages 265–294, April 2005. ISBN: 0-262-07262-9.
10. A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. of the Royal Stat. Soc. (Series B)*, 1977.
11. T. Edgoose, L. Allison, and D.L. Dowe. An MML Classification of Protein Structure that knows about Angles and Sequence. In *Pacific Symposium on Biocomputing '98*, pages 585–596, Singapore, 1998. World Scientific Publishing.
12. T. Fawcett and F. Provost. Adaptive fraud detection. *Knowledge Discovery and Data Mining*, 1:291–316, 1997.
13. L.J. Fitzgibbon, D.L. Dowe, and F. Vahid. Minimum message length autoregressive model order selection. In *Intl. Conf. on Intelligent Sensing and Information Processing (ICISIP)*, January 2004. M. Palanaswami et al. (eds.), Chennai, India.
14. N.L. Frigon and D. Mathews. *Practical Guide to Experimental Design.* Wiley, 1996.
15. E. Keogh, J. Lin, and A. Fu. HOT SAX: Efficiently Finding the Most Unusual Time Series Subsequence. In *The Fifth IEEE International Conference on Data Mining*, 2005.
16. E. Keogh, J. Lin, and W. Truppel. Clustering of Time Series Subsequences is Meaningless: Implications for Past and Future Research. In *Proceedings of the 3rd IEEE International Conference on Data Mining, Melbourne, FL*, pages 115–122, Nov 19-22, 2003.
17. N. Kumar, Lolla N., E. Keogh, S. Lonardi, C. A. Ratanamahatana, and L. Wei. Time-series Bitmaps: A Practical Visualization Tool for working with Large Time Series Databases. In *Proceedings of SIAM International Conference on Data Mining (SDM '05), Newport Beach, CA*, pages 531–535, April 21-23, 2005.
18. Diane L. Larson, Patrick J. Anderson, and Wesley Newton. Alien Plant Invasion in Mixed-Grass Prairie: Effects of vegetation type and anthropogenic disturbance. *Ecological Applications*, 11(1):128–141, 2001.
19. J. Lin, E. Keogh, S. Lonardi, and B. Chiu. A Symbolic Representation of Time Series, with Implications for Streaming Algorithms. In *Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, San Diego, CA*, June 13, 2003.
20. R.L. Mason, R.F. Gunst, and J.L. Hess. *Statistical Design and Analysis of Experiments: With Applications to Engineering and Science.* Wiley, 2nd edition, 2003.
21. L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, volume 77, pages 257–286, 1989.

22. Marco Ramoni, Paola Sebastiani, and Paul Cohen. Bayesian Clustering by Dynamics. *Machine Learning*, 47(1):91–121, 2001.
23. Stuart J. Russell and Peter Norvig. *Artificial Intelligence, A Modern Approach.* Prentice Hall, Pearson Education, Inc., 2nd edition, 2003.
24. K. Shimo-onoda, T. Tanaka, K. Furushima, T. Nakajima, S. Toh, S. Harata, K. Yone, S. Komiya, H. Adachi, E. Nakamura, H. Fujimiya, and I. Inoue. Akaike's Information Criterion for a Measure of Linkage Disequilibrium. *Journal of Human Genetics*, 47(12):649–655, 2002.
25. T.T. Takeuchi. Application of the Information Criterion to the Estimation of Galaxy Luminosity Function. *Astrophysics and Space Science*, 271(3):213–226, 2000.
26. K. A. Triantis, M. Mylonas, M. D. Weiser, K. Lika, and K. Vardinoyannis. Species richness, environmental heterogeneity and area: a case study based on land snails in Skyros archipelago (Aegean Sea, Greece). *Journal of Biogeography*, 32(10):1727–1735, 2005.
27. C.S. Wallace. *Statistical and Inductive Inference by Minimum Message Length.* Springer-Verlag, 2005. ISBN: 0-387-23795-X.
28. C.S. Wallace and D.M. Boulton. An Information Measure for Classification. *Computer Journal*, 11(2):185–194, 1968.
29. C.S. Wallace and D.L. Dowe. Intrinsic classification by MML - the Snob program. In *Proc. 7th Australian Joint Conf. on Aritficial Intelligence*, pages 37–44, Armidale, Australia, World Scientific, 1994.
30. C.S. Wallace and D.L. Dowe. Minimum Message Length and Kolmogorov Complexity. *Comp. Jour., Special Issue - Kolmogorov Complexity*, 42(4):270–283, 1999.
31. C.S. Wallace and D.L. Dowe. MML Clustering of Multi-State, Poisson, von Mises Circular and Gaussian Distributions. *Stats. and Comp.*, 10(1):73–83, Jan. 2000.
32. C.S. Wallace and P.R. Freeman. Estimation and Inference by Compact Coding. *Journal of the Royal Statistical Society (Series B)*, 49:240–252, 1987.