# Enhancing MML Clustering Using Context Data with Climate Applications

Gerhard Visser, David L. Dowe, and Petteri Uotila

Monash University, VIC 3800, Melbourne Australia
gerhardus.visser@infotech.monash.edu.au

**Abstract.** In Minimum Message Length (MML) clustering (unsupervised classification, mixture modelling) the aim is to infer a set of classes that best explains the observed data items. There are cases where parts of the observed data do not need to be explained by the inferred classes but can be used to improve the inference and resulting predictions. Our main contribution is to provide a simple and flexible way of using such context data in MML clustering. This is done by replacing the traditional mixing proportion vector with a new context matrix. We show how our method can be used to give evidence regarding the presence of apparent long-term trends in climate-related atmospheric pressure records. Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) solutions for our model have also been implemented to compare with the MML solution.

## 1 Introduction

### 1.1 Minimum Message Length

The Minimum Message Length (MML) [16, 17, 21] principle states that the best explanation for observed data D is the one that minimises the optimal coding length (according to information theory) of a two-part message. The first part encodes the hypothesis $H$ (this is known as the assertion) from Bayesian priors while the second part encodes the data $D$ given the $H$ (this is known as the detail). In practice, we do not actually construct any message but rather strive to infer a hypothesis which minimises some approximation to that code length.

The MML principle can be thought of as a quantitative version of Ockham's razor [5, footnotes 18 and 181-182] and is compared to Kolmogorov complexity and algorithmic complexity [2, 12, 15] in [19]. For a contrast with the much later Minimum Description Length (MDL) principle [14], see [4, sec. 11.4] and [16, chap. 10]. MML inference is statistically invariant (inference is preserved under 1-to-1 transformations of the parameter space) and is in general statistically consistent [5, 6, 16]. Where many methods have been shown to be statistically inconsistent on misspecified models [9], there is as yet no known example of MML having this failing [9, sec. 7.1.5][5, sec. 0.2.5].

MML is capable of selecting between models with varying numbers of parameters without overfitting [21, sec. 6], and outperforms Maximum Likelihood (ML) even when aided by Akaike's Information Criterion (AIC) [6].

A. Nicholson and X. Li (Eds.): AI 2009, LNAI 5866, pp. 350–359, 2009.
© Springer-Verlag Berlin Heidelberg 2009

MML is a general model selection criterion and as such is intended to replace traditional hypothesis tests and confidence intervals. Instead, message lengths are compared for different hypotheses. Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC) are two comparable and popular model selection criteria. We have implemented both AIC and BIC versions of our method for comparison with the MML solution (section 3).

### 1.2 Clustering with Minimum Message Length

Given a set of observed items $y = (y_1, y_2, ..., y_N)$ the aim of clustering is to find a set of $C$ classes such that each item can be assigned to a class. The number, $C$, is often assumed known and the properties of the classes are inferred from the data. These inferred properties of a class describe its typical items.

MML clustering as described in [17, 18, 20] and [16, sec. 6.8] is an unsupervised mixture modelling method which will also select the number of classes present.

The Expectation Maximisation (EM) algorithm is used to infer the class parameters for a fixed number of classes. This is repeated assuming different numbers of classes. For each such EM run a message length (section 1.1) is calculated. The solution with the smallest message length is selected as the best.

Given an inferred hypothesis, the message length is calculated as the length of an optimal code described as follows.

1. The *Assertion* encodes the hypothesis in the following order.
   (a) The number of classes used, $C$.
   (b) The relative frequencies of all classes.
   (c) The inferred parameters defining each class.
   (d) The partial assignments of items to classes.
2. The *Detail* encodes the data given the hypotheses.
   (a) The observed attributes of each item.

### 1.3 Our Extension and Some Motivations

Since this work was developed with atmospheric time-series data in mind we will use that as an example throughout this paper but our methods are intended to be general purpose.

MML clustering attempts to capture all regularities that are present in the data. This means that in practice as one adds more attributes to each data item, the number of classes inferred tends to increase. As there is more data to compress the first part of the message can become larger and more complex. Too many classes can be hard to interpret - which in some cases may be undesirable.

In our data set each data item $y_i$ is a set of atmospheric pressure values from several weather stations for a single day. These pressure values $y_{i,j}$ (where $i$ indices a day and $j$ indexes a weather station) are the attributes that we wish to cluster. There may be other attributes that can be associated with each day (data item) which might help with the clustering but which we do not wish to model or explain with the classes inferred. Examples include seasons, extreme weather conditions and global indexes such as those relating to the El Niño cycle.

It helps to notice that with clustering there are often two types of attributes. One can think of them as target attributes and context (known) attributes. We are not interested in discovering regularities in the context attributes and it is not desirable that the number of classes and the complexity of the classes increase to explain those regularities. On the other hand the inferred hypothesis should mention these attributes if they help explain the target attributes.

Our aim is to explain the target attributes while using the context attributes to discriminate and (this aim) is therefore similar to what Jebara [10] describes as combining Discriminative and Generative learning.

Our work provides a simple yet flexible way of dealing with these context attributes differently from target attributes while adhering to the well established MML clustering framework of [17, 20] and [16, sec. 6.8]. By doing that our method inherits the features of MML clustering which has made it successful which includes the ability to select the number of classes without over-fitting.

## 2 Methods

### 2.1 A Clustering Model with Context Data

Let $y$ be a set of observed items where $y_{i,j}$ is the value of attribute $j$ for data item $i$. Let $x$ be the corresponding class assignments where $x_i \in \{1, 2, ..., C\}$ and $C$ is the number of classes. In our atmospheric time-series example $y_{i,j}$ is the measurement on day $i$ at weather station $j$.

There are other context attributes associated with each day that we can use to improve the clustering but do not wish to model. For our climate example this could include time of year (season) or global indices such as those relating to the El Niño cycle. For this we introduce a context value $z_{i,k}$ where $i$ indexes the item (day) and $k \in \{1, 2, ..., K\}$. Here there are $K$ different *contexts*. Each item $i$ belongs to each context to some degree $z_{i,k}$. The context data $z$ is given as prior knowledge. Each context vector $z_i$ is used much like a fuzzy indicator, however, we interpret them strictly as probability distributions, hence we require that for all $i$, $\sum_{k=1}^{k=K} z_{i,k} = 1$ and that all $z_{i,k} \geq 0$.

As an example we can divide the days of each year into four seasons, so $K = 4$. A day in the middle of summer (context $k = 1$) can be assigned completely to that season $z_{i,1} = 1$ while a day between summer and autumn can be assigned partially to those two seasons $z_{i,1} = 0.5, z_{i,2} = 0.5$.

In our model we replace the mixing proportion parameter vector with a $K \times C$ mixing proportion matrix $S$. Now the probability of item $i$ belonging to class $c$ is defined as,

$$\Pr(x_i = c) = \sum_{k=1}^{K} z_{i,k} S_{k,c}. \quad (1)$$

Each of the $K$ rows of matrix $S$ is a relative frequency vector associated with a context. In our example $S_{k,c}$ is the probability of day $i$ belonging to class $c$ if the season is $z_{i,k} = 1$. It follows that $\sum_{c=1}^{C} S_{k,c} = 1$ and all $S_{k,c} \geq 0$. This mixing proportion matrix $S$ will be inferred from the data. In our season example this means that the effective mixing proportions will change gradually according to time of year and we avoid having to use hard boundaries when specifying $z$.

The matrix $S$ allows the context vector $z$ to be used to provide information about the class assignments $x$ prior to seeing the data $y$. This means $x$ can be encoded more efficiently given $S$ and $z$ but only if that saving is not outweighed by the cost of stating $S$, which increases with $K$ and $C$. Effectively $S$ allows $z$ to inform the classification and inferred model. We are not encoding $z$ at all, one could imagine a separate message fragment encoding $z$ preceding the rest of the message. This imaginary message fragment would be unaffected by $y$, $x$, $S$, $C$ and the class parameters. The idea is that how $z$ is modelled or encoded does not affect the rest of the message.

Each class defines a distribution $\Pr(y_i|x_i)$ for the data items assigned to it. These distributions have parameters associated with them which must be inferred. For our climate example we will consider each weather station to have an independent Gaussian distribution. For details on how these parameters are inferred with MML, for this and other distributions, see [16, 18, 20].

### 2.2 Coding Approximation and Optimisation Algorithm

In MML inference one usually creates an approximation to the message length of the two-part code described in section 1.1, and then infers a hypothesis which optimises that approximation. We first describe the form of the hypothetical message, then how it is approximated and then the optimisation algorithm. Given a hypothesis the message is made up of the same message fragments as in the list given at the end of section 1.2. For our model part 1b of that list states the matrix $S$ instead of a single mixing proportion vector.

In accordance with MML convention our message lengths are calculated in nits where 1 nit $= \log_2 e$ bits. For item 1a we use the prior distribution $2^{-C}$ over the number of classes, this message fragment has a length of $C \log_e 2$ nits.

For part 1b the rows of matrix $S$ can be stated using a standard MML multistate distribution solution (see [20]).

The code length for the class parameters (1c) can be approximated using standard MML solutions for the distributions used (see [20]). Because the order of the classes is arbitrary, $\log_e C!$ nits can be subtracted from this length.

For part 1d the coding length for stating each class assignment $x_i$ precisely is the negative logarithm of the conditional probability $\Pr(x_i|z_i, S)$. Since an optimal code would not state these parameters ($x$) precisely, a coding trick (see [16, sec. 6.8]) can be used to calculate the message length improvement that can be achieved through imprecisely encoding $x$. The result of this is that one can subtract form the above described message length the entropy of $x$ given everything else ($z$,$y$,$S$ and the class parameters).

Finally the length of the detail (part 2) is simply the negative log likelihood of $y$ given the inferred assignments $x$ and the inferred class parameters.

Because of these imprecise encodings of $x$ one can interpret their assignments as partial (uncertain) and the expectations (over the partial assignments of $x$) of the code length described above (parts 1b, 1c, 1d and 2) is used.

Now that we have an approximation to the code length for a given hypothesis and data set, a search algorithm which finds an optimal hypothesis is needed. The Expectation Maximisation (EM) algorithm is used.

```
1: initialise partial assignments for x;
2: initialise values for S;
while(not(some termination condition))
{
  3: update class parameters to their optimal values given x and y;
  4: update partial assignments of x given S and y;
  5: update the matrix S given x and z;
}
```

Step 3 is done as with standard MML clustering, see [20] or [16, sec. 6.8]. In step 4 the optimal degree of assignment of item $i$ to class $c$ is equal to its posterior probability $\Pr(x_i = c|S, z, y)$. This type of estimate for discrete parameters like $x$ is discussed in [16, sec. 6]. Step 5 uses the same multi-state distribution solution used in standard MML clustering for the rows of $S$, however the contribution of item $i$ to the parameter row vector $S_k$ is weighted according to,

$$w_{i,k} = \frac{z_{i,k} \sum_{c=1}^{C} S_{k,c} \Pr\left(y_i | x_i = c\right)}{\sum_{t=1}^{K} z_{i,t} \sum_{c=1}^{C} S_{t,c} \Pr\left(y_i | x_i = c\right)}. \tag{2}$$

These weights are also used in the message length calculations for $S$. The individual reassignments (steps 3, 4 and 5) each decrease the overall message length in every iteration and the result is that the solution as a whole moves to a local optimum.

# 3 Data and Results

## 3.1 Tests on Artificially Generated Data

Because MML is a Bayesian method, our first claim is that if a *true* hypothesis is generated from the assumed model then our method will on average tend to be good at inferring back that true hypothesis. The hypotheses that were generated for these tests were intended to roughly imitate those one would expect to infer for our atmospheric pressure data.

For the first test the true model has 5 classes with 5 pressure values generated for each day over a 50 year period. The context variable $z$ has been used to divide the 50 years into 4 long term divisions. This simulates how the relative frequencies of our 5 classes change over the long run. Each day is assigned partially to two of these divisions so that the change in relative frequencies occurs slowly and smoothly over time (as with the seasons example in section 2.1).

Data was generated from the assumed hypothesis as described above. Half the attributes from this data were randomly removed as a validation set. From the training set our algorithm was used to infer the number of long term divisions while knowing the true number of classes. This test was repeated for ten such data sets. The results are summarised in figs. 1 and 2. Aside from the MML criterion we have also optimised the Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC) for all results. In Fig. 1 the lines titled MML, AIC and BIC show the average resulting criterion values belonging to the left vertical axis. Here we can see all three criteria have their average optimal value at the correct number of divisions ($K = 4$). The predictive performance is measured as the average negative log likelihood of the validation set given the chosen hypotheses, divided by the validation set size. This measure is titled *score* in Fig. 1 and belongs to the right vertical axis. It can be seen that the predictive score reaches its optimal value for $K = 4$ and extra divisions do not improve this.

In Fig. 2 we can see the number of data sets (out of ten) for which MML, AIC and BIC preferred $K$ divisions. The results show that MML and BIC tended to be similarly conservative while AIC sometimes prefers more divisions than the true number $K = 4$.

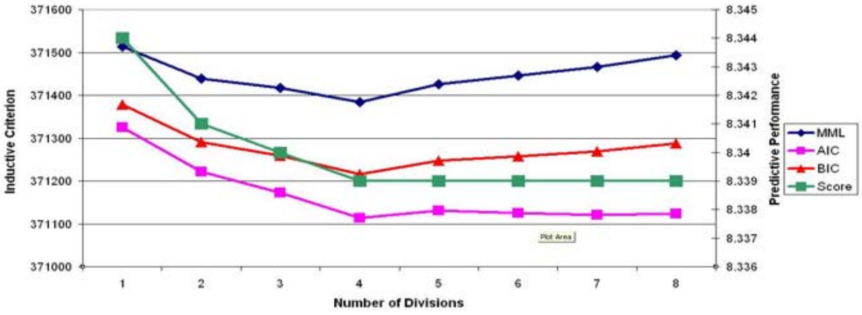

**Fig. 1.** Average MML, AIC and BIC values inferred for different values of $K$ belong to the left axis. The predictive performance *score* belongs to the right axis. The true value is $K = 4$.

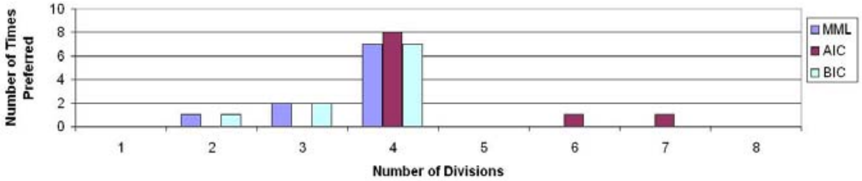

**Fig. 2.** The number of data sets (out of ten) for which MML, AIC and BIC preferred $K$ divisions, with the true value being $K = 4$

In the next test we have generated data as with the first test but now the algorithm knows the number of long-term divisions ($K = 4$) while the number of classes ($C = 5$) must be inferred. For this test MML and BIC performed similarly and well (preferring either $C = 5$ or $C = 4$) while AIC tended to over-fit (preferring $6 \leq C \leq 8$).

In the final artificial data test we have generated data using only one long term division $K = 1$ (equivalent to no context variable). All three methods were used to infer $K$ as before. Here both MML and BIC chose the correct number $K = 1$ all ten times while AIC chose $K = 1$ seven times but also made estimates as high as $K = 5$.

Our conclusion from these three tests is that both MML and BIC can be expected to either choose the true values for $K$ and $C$ or to choose more conservatively, while AIC will occasionally overestimate these values.

### 3.2 Atmospheric Time-Series Data and MML Clustering

The meteorological data was derived from historical sub-daily station mean sea level air pressure observations digitised by the Australian Bureau of Meteorology. The air pressure was observed in approximately 50 weather stations across Australia with earliest observations dating back to 1859. The data has been quality controlled. This processing included removal of errors in the observations by mistakes made when digitising observations or when observers incorrectly recorded air pressure values.

Clustering such data both from real world observations or from climate model output is valuable as it allows for large and complex data sets to be interpreted more easily. This can then be used to look for variations in pattern frequencies over time and to link these variations to other climate/weather related events. Self Organising Maps (SOM) [11] have been successfully used for this purpose in the past [1, 13]. The work we present in this paper is an early step in continuing work aimed at providing alternative tools to SOM and k-means clustering specific to atmospheric time-series data.

The existence of multiple atmospheric circulation regimes (classes) in the extratropics is an important, but a controversial, hypothesis in meteorology [3]. Many conflicting results exist and are critically discussed in [3].

We hope that by refining our probabilistic models to fit this problem domain, MML can with its resistance to overfitting provide important evidence regarding this issue. In this paper for this data set our primary goal is to measure and analyse the link between context information and the atmospheric data. Sections 3.3 and 3.4 demonstrate this.

### 3.3 Dividing the Year into Seasons

It is known that atmospheric pressure states are dependent on time of year (seasons). We demonstrate how our method can be used to determine into how many seasons each year can be divided. We test our method's ability to choose the best predictive model by dividing the data into training and validation sets and comparing predictive performance with the message lengths.

The data from 1865 to 1915 was used. Half the observed measurements were randomly removed as a validation set. We have assumed here that the number of classes is $C = 20$. The algorithm was used to infer the number of seasonal divisions $K$. Each day is assigned partially to two seasonal divisions as described in section 2.1. In this way we model how the relative frequencies of classes cycle smoothly over time. Fig. 3 compares the resulting MML message lengths and BIC and AIC values for different numbers of seasonal divisions $K$. For each value of $K$ MML, BIC and AIC inference were repeated 10 times and the solution for which each criterion performed best was selected and is shown on Fig. 3.

For this test, both MML and BIC preferred 8 seasonal divisions while AIC preferred 16. It can be seen that there is no significant improvement on the validation set for more than 8 divisions.

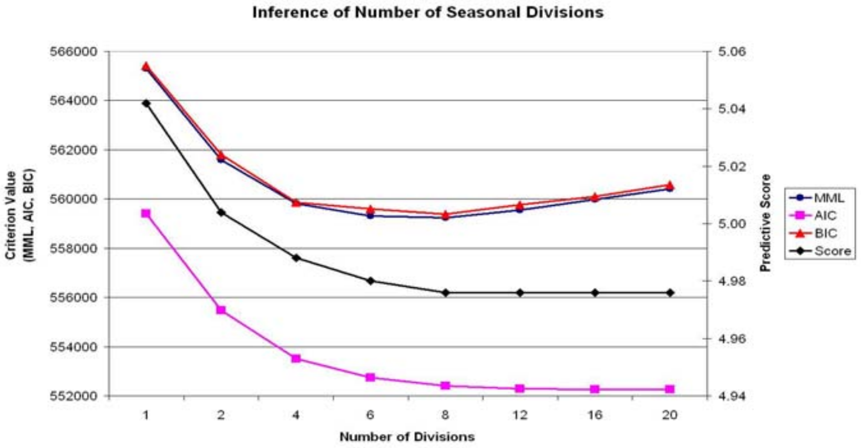

**Fig. 3.** MML, BIC and AIC values for different numbers of seasonal divisions $K$ belong to the left axis. The *score* measures the performance of the validation set and belongs to the right axis.

### 3.4 Identifying Long-Term Trends in Atmospheric Time-Series Data

Finally we have used our method with the data from 1865 to 1965 to see how many long-term trends can be justified when assuming $C = 20$ classes. Again half the data was randomly removed as a validation set. The algorithm was used to infer the correct number of long-term divisions $K$ as defined in section 3.1. For each value of $K$ both MML and BIC inference were repeated 10 times and the solution for which each criterion performed best was selected. MML had a clear preference for 4 long term divisions while BIC preferred 7. Both the 4-term and 7-term solutions had the same predictive performance.

## 4 Conclusion and Further Work

In clustering there is often additional information that can be used to improve the inference but which should not be included as target attributes (attributes to be clustered). The context clustering method that we have presented provides a flexible yet simple extension to standard MML clustering which achieves our goal of using context data. Our results on artificial data show that we can detect the presence of such context divisions and estimate their number if the assumed model is correct. An implementation of our algorithm which uses BIC instead of MML performs similarly while AIC tends to overfit. With the atmospheric pressure time-series data we have demonstrated how our method can be used to give evidence regarding the presence of apparent long term trends in atmospheric pressure patterns and to determine the number of seasonal divisions that can be justified.

It is known for this data set that the class of each day is highly dependant on the class of the previous day and that this can be modelled using a hidden Markov unit model as in [7]. We are currently working on combining our context variable model with that hidden Markov unit model.

Instead of using the context variable for long term divisions or seasons one could use it to try and link global weather indexes, like those measuring the El Niño cycle, to atmospheric pressure patterns. This would require that the context variable have two possible assignments, one for El Niño and one for La Niña, where each day would be (partially) assigned to both with some degree based on the Southern Oscillation Index (SOI).

Other uses for the context variable could include weather extremes such as storms, unusual rainfall, cyclones and hurricanes. Another simple extension of this work will be to allow multiple context variables to be use, this would allow for example for both season and long term trend information to be used.

With clustering real world data the difference between model and reality can lead to an excessive number of classes. One way to address this is to remove the assumption that the attributes within each class can be modelled as independent Gaussian distributions. It should be possible to allow for inter-attribute relations such as latent factors, which have been used in MML clustering in [8].

## References

1. Cassano, J.J., Uotila, P., Lynch, A.: Changes in synoptic weather patterns in the polar regions in the twentieth and twenty-first centuries, part 1: Arctic. International Journal of Climatology 26(8), 1027–1049 (2006)
2. Chaitin, G.J.: On the length of programs for computing finite binary sequences. Journal of the Association of Computing Machinery 13, 547–569 (1966)
3. Christainsen, B.: Atmospheric Circulation Regimes: Can Cluster Analysis Provide the Number? Climate Journal 20(10), 2229–2250 (2007)
4. Comley, J.W., Dowe, D.L.: Minimum message length and generalized Bayesian nets with asymmetric languages. In: Grünwald, P., Pitt, M.A., Myung, I.J. (eds.) Advances in Minimum Description Length: Theory and Applications, pp. 265–294. MIT Press, Cambridge (2005)

5. Dowe, D.L.: Foreword re C. S. Wallace. Computer Journal 51(5), 523–560 (2008)
6. Dowe, D.L., Gardner, S., Oppy, G.R.: Bayes not bust! Why simplicity is no problem for Bayesians. British J. Philosophy of Science, 709–754 (December 2007)
7. Edgoose, T., Allison, L.: MML Markov classification of sequential data. Statistics and Computing 9, 269–278 (1999)
8. Edwards, R.T., Dowe, D.L.: Single factor analysis in MML mixture modeling. In: Wu, X., Kotagiri, R., Korb, K.B. (eds.) PAKDD 1998. LNCS (LNAI), vol. 1394, pp. 96–109. Springer, Heidelberg (1998)
9. Grunwald, P., Langford, J.: Suboptimal behavior of Bayes and MDL in classification under misspecification. Machine Learning 66(2-3), 119–149 (2007)
10. Jebara, T.: Discriminative, Generative and Imitative learning. PhD thesis, MIT (2001)
11. Kohonen, T.: Self-Organizing Maps, vol. 30. Springer, Heidelberg (2001)
12. Kolmogorov, A.N.: Three approaches to the quantitative definition of information. Problems of Information Transmission 1, 1–17 (1965)
13. Reusch, D.B., Alley, R.B.: Relative performance of Self-Organizing Maps and Principal Component Analysis in pattern extraction from synthetic climatological data. Polar Geography 29(3), 188–212 (2005)
14. Rissanen, J.: Modeling by the shortest data description. Automatica 14, 465–471 (1978)
15. Solomonoff, R.J.: A formal theory of inductive inference. Information and Control 7, 1–22, 224–254 (1964)
16. Wallace, C.S.: Statistical and Inductive Inference by Minimum Message Length. Springer, Heidelberg (2005)
17. Wallace, C.S., Boulton, D.M.: An information measure for classification. Computer Journal 11, 185–194 (1968)
18. Wallace, C.S., Dowe, D.L.: Intrinsic classification by MML - the Snob program. In: Proc. 7th Australian Joint Conf. on Artificial Intelligence, pp. 37–44. World Scientific, Singapore (1994)
19. Wallace, C.S., Dowe, D.L.: Minimum message length and Kolmogorov complexity. Computer Journal 42(4), 270–283 (1999)
20. Wallace, C.S., Dowe, D.L.: MML clustering of multi-state, Poisson, von Mises circular and Gaussian distributions. Statistics and Computing 10, 73–83 (2000)
21. Wallace, C.S., Freeman, P.R.: Estimation and inference by compact coding. J. Royal Statistical Society B 49, 240–252 (1987)