A Mixture Model Based Markov Random Field for Discovering Patterns in Sequences

Konstantinos Blekas

Department of Computer Science,
University of Ioannina, 45110 Ioannina, Greece
kblekas@cs.uoi.gr

Abstract. In this paper a new maximum a posteriori (MAP) approach based on mixtures of multinomials is proposed for discovering probabilistic patterns in sequences. The main advantage of the method is the ability to bypass the problem of overlapping patterns in neighboring positions of sequences by using a Markov random field (MRF) prior. This model consists of two components, the first models the pattern and the second the background. The Expectation-Maximization (EM) algorithm is used to estimate the model parameters and provides closed form updates. Special care is also taken to overcome the known dependence of the EM algorithm to initialization. This is done by applying an adaptive clustering scheme based on the k-means algorithm in order to produce good initial values for the pattern multinomial model. Experiments with artificial sets of sequences show that the proposed approach discovers qualitatively better patterns, in comparison with maximum likelihood (ML) and Gibbs sampling (GS) approaches.

Keywords: Pattern discovering, Markov random field, mixture of multinomials model, Expectation-Maximization (EM) algorithm.

1 Introduction

Discovering patterns in sequences is an important problem in many application areas, such as bioinformatics, web mining, etc. Given a set of sequences a pattern (or *motif*) can be represented as a common substring that is repeated in the set. Sequence patterns are focused on highly conserved residues present in active sites of sequences and can be further used for generating rules for classification purposes [1, 2].

Various methods have been introduced for solving this problem that are classified based on the model of the pattern. Under the Bayesian framework, a pattern can be modeled using independent multinomial distributions for its positions. The Gibbs sampling [3, 4], the MEME [5], the SAM [6], the BioProspector [7], the Greedy EM [8] and the LOGOS [9] represent statistical methods for discovering shared patterns in a set of sequences. They all formulate the problem using either mixture models or hidden Markov models, and use the Expectation-Maximization (EM) algorithm [10, 11] or variational EM schemes to estimate the model parameters.

G. Antoniou et al. (Eds.): SETN 2006, LNAI 3955, pp. 25–34, 2006.

The application of statistical methods to discovering sequence patterns usually forces the assumption that all the possible starting positions in sequences are independent. Nevertheless, the problem has the particular characteristic that spatial information should be taken into account. That is, apart from the content of a subsequence, its location must be also used in order to determine its posterior probability for matching it as pattern. In other words, it is not desired to identify overlapping patterns. In most of these methods, the common framework used is the maximum likelihood (ML) where the pattern model parameters are estimated by maximizing the likelihood of the observations, while the spatial constraints are indirectly enforced to the model. Therefore, in a sense, there is an inconsistency between the computed pattern distribution and the one defined by the model [9].

In this paper we present a *maximum a posteriori* (MAP) approach that provides a direct method to implement these ideas. The basic scheme is a two-component mixture of multinomials model, where one component models the pattern and the other the remaining non-pattern regions (background). Following this framework, a likelihood term is used to capture the content information of the data, while a bias term is also used to capture the spatial information of the neighborhood locations. This is accomplished by considering the pattern labels of each starting position of sequences through a Markov random field (MRF) model [12, 13]. This constrains the local characteristics of the sequences and thus provides useful information to the pattern estimation process. The EM algorithm is used to estimate the model parameters which provides closed form update equations for all parameters. Since the EM algorithm is very sensitive to the initial parameter values, we also present a clustering scheme based on the well-known k-means algorithm for initializing properly the pattern model. Finally, multiple patterns are discovered by iteratively applying the two-component mixture model after erasing old pattern occurrences. As will be demonstrated in the experimental study, in contrast to the classical unconstrained mixture model and the Gibbs sampling approach, the proposed one overcomes the problem of overlapping subsequences and also estimates qualitatively better pattern models.

Section 2 presents the two-components mixture of multinomials model that is used for discovering a single pattern in two methods: the classical ML and the proposed MAP approach. Experimental results are given in section 3 using artificial sets of sequences, while section 4 presents conclusions and discussion.

2 Mixture Models for Discovering Patterns

Consider a finite set $\Sigma = \{c_1, \ldots, c_\Omega\}$ consisting of Ω individual characters. An arbitrary string over the set Σ is any sequence $S_j = \{s_{jk}\}_{k=1}^{L_j}$ of length L_j, where $s_{jk} \in \Sigma$ denotes the character at the k-th position of the j-th sequence. Now, let $S = \{S_1, \ldots, S_N\}$ be a set of N strings of length $L_1, \ldots, L_N$, respectively. The pattern discovery problem deals with finding a common subsequence of length K that is repeated at different sites among the sequences of set S. In order to deal with this, we collect all the possible substrings of set S having

length equal to K. This can be done by sliding a window of size K in every sequence S_j, obtaining a set of $L_j - K + 1$ substrings. Each substring indicates the starting position of a possible pattern occurrence in sequences. Therefore, we construct a set of n substrings $X = \{x_i\}_{i=1}^{n}$, $n = \sum_{j=1}^{N}(L_j - K + 1)$, that constitute the observation data. In the next subsections two mixture model based approaches will be presented: the classical maximum likelihood without any constraint, as well as, a new proposed maximum a posteriori approach that uses spatial information.

2.1 The ML Approach

Lets assume that the set X has been generated from a two-component mixture of multinomials, i.e. we assume that each substring x_i belongs to either a pattern class ($y_i = 1$) or a background class ($y_i = 0$) which is indexed by the hidden (binary) variable y_i. The first component models the pattern with a *common* prior probability of $\pi = p(y_i = 1)$, while the second one models the background and represents all the subsequences which do not contribute to the pattern, with a prior probability equal to $1 - \pi = p(y_i = 0)$. The density function $f(x_i|\pi, \Theta)$ of the model for an observation x_i is given by

$$f(x_i|\pi, \Theta) = \pi p(x_i|\theta) + (1 - \pi)p(x_i|b) \ , \tag{1}$$

where $\Theta = \{\theta, b\}$ is the set of parameters for the two multinomial densities. To parameterize the pattern we use a position weight matrix $\theta = [\theta_{kl}]$ of size $\Omega \times K$, where each element θ_{kl} denotes the probability that character $c_l \in \Sigma$ is at the k-th position of the pattern. For each position k it holds $\sum_l \theta_{kl} = 1$. The background distribution is represented with an Ω-vector of probabilities $b = [b_1, \ldots, b_\Omega]$ common for each substring position ($\sum_l b_l = 1$). Following the multinomial distribution and assuming independence among positions, the probability densities function of the pattern and the background model are

$$p(x_i|y_i = 1, \theta) = \prod_{k=1}^{K}\prod_{l=1}^{\Omega} \theta_{kl}^{\delta_{ikl}} \quad , \quad p(x_i|y_i = 0, b) = \prod_{l=1}^{\Omega} b_l^{\sum_{k=1}^{K}\delta_{ikl}} \ , \tag{2}$$

where δ_{ikl} is the Kronecker delta (1 if character c_l is at the k-th position of substring x_i, 0 otherwise).

Based on the above formulation, the model parameters can be estimated through maximum likelihood (ML). The log-likelihood function is then given by

$$L(X|\pi, \Theta) = \sum_{i=1}^{n} \log f(x_i|\pi, \Theta) \ . \tag{3}$$

The EM algorithm [10, 11] is an efficient framework to estimate the model parameters π, $\{\theta_{kl}\}$ and $\{b_l\}$. It requires the computation of conditional expectation z_i of the hidden variables y_i at the E-step, which are given by

$$z_i^{(t)} = p(y_i = 1|x_i, \pi^{(t)}, \Theta^{(t)}) = \frac{\pi^{(t)} p(x_i|y_i = 1, \theta^{(t)})}{\pi^{(t)} p(x_i|y_i = 1, \theta^{(t)}) + (1 - \pi^{(t)}) p(x_i|y_i = 0, b^{(t)})} \ , \tag{4}$$

while at the M-step the complete log-likelihood is maximized over the model parameters. This gives the following update equations

$$\pi^{(t+1)} = \frac{\sum_{i=1}^{n} z_i^{(t)}}{n} \ ,$$
$$b_l^{(t+1)} = \frac{\sum_{i=1}^{n}(1-z_i^{(t)})\sum_{k=1}^{K}\delta_{ikl}}{K\sum_{i=1}^{n}(1-z_i^{(t)})} \ , \ \theta_{kl}^{(t+1)} = \frac{\sum_{i=1}^{n} z_i^{(t)}\delta_{ikl}}{\sum_{i=1}^{n} z_i^{(t)}\sum_{l=1}^{\Omega}\delta_{ikl}} \ . \qquad (5)$$

EM algorithm guarantees the convergence of the likelihood fuction to a local maximum and also satisfies all the constraints of the parameters.

Nevertheless, a significant drawback of the ML approach, arising from the assumption that the observations x_i are i.i.d., is the fact that the spatial information of the subsequences is not taken into account. This results in the estimation of overlapping subsequences as pattern occurrences of the set X, especially in cases where a pattern consists of one or more repeated characters. However, by enforcing spatial constraints one can avoid this problem and estimates better patterns. In [5] for example, a normalization of the posterior value z_i of the adjacent sequences is performed so that guarantees in any window of length K the sum of z_i values remains less than or equal to 1. In this study, we introduce a new approach that deals with this problem in a more systematic way by modeling the spatial arrangements of a pattern.

2.2 The Proposed Spatially-Constrained Mixture Model

In the proposed model, the pattern label priors $\mathit{\Pi} = \{\pi_i = p(y_i = 1)\}_{i=1}^{n}$ are considered as random variables that satisfy the constraint $0 \leq \pi_i \leq 1$. Since they are spatially dependent, we assume that they form a Markov random field (MRF) being sampled by a Gibbs distribution function [12, 13]

$$p(\mathit{\Pi}|\beta) = \frac{1}{Z}\exp(-U(\mathit{\Pi}|\beta)) \ . \qquad (6)$$

The normalization constant Z is called the partition function, while the β is a regularization parameter. The energy function $U(\mathit{\Pi}|\beta)$ is decomposed into a sum of *clique potentials* $V_{\mathcal{N}_i}$

$$U(\mathit{\Pi}|\beta) = \beta\sum_{i=1}^{n} V_{\mathcal{N}_i}(\mathit{\Pi}) \ , \qquad (7)$$

that involves neighboring sites $\mathcal{N}_i$ in the proposed sequential field. A similar in principle spatially-constrained model has been also used for the image segmentation problem [14]. In this study, we consider as neighborhood $\mathcal{N}_i$ all the m positions around the position i whose corresponding subsequences x_m overlaps with the subsequence x_i. In the general case, there are $2(K-1)$ such sites which are mutually dependent. When a pattern is found at position i ($\pi_i \approx 1$), it is desired that none substring $x_m \in \mathcal{N}_i$ to be also labeled as pattern ($\pi_m \approx 0$). An

appropriated potential function that meets this behavior is the following simple inner product of pattern labels

$$V_{\mathcal{N}_i}(\Pi) = \sum_{m \in \mathcal{N}_i} \pi_i \pi_m = \pi_i \pi_{\mathcal{N}_i} \ . \tag{8}$$

Since Dirichlet densities are conjugate to multinomial densities, it is convenient to use them in order to introduce priors for the pattern parameters θ. Thus, for every pattern position k we consider a Dirichlet prior of the form

$$p(\theta_k|\alpha_k) = \frac{\Gamma(\sum_{l=1}^{\Omega} \alpha_{kl})}{\prod_{l=1}^{\Omega} \Gamma(\alpha_{kl})} \prod_{l=1}^{\Omega} \theta_{kl}^{\alpha_{kl}-1} \ , \tag{9}$$

where the parameter α_k is a Ω-vector with components $\alpha_{kl} > 0$ and $\Gamma(\alpha)$ is the Gamma function. Adding Dirichlet priors in effect introduces pseudo-counts at each pattern position. During the experimental study, the Dirichlet prior parameters were the same for every position and set equal to $1+\epsilon_l$, where ϵ_l was some low percentage (e.g. 10%) of the total predefined frequency of character c_l.

Given the above prior densities (Eqs. (6)-(9)) for the model parameters Π and θ, we can formulate the problem as a *maximum a posteriori* (MAP) problem, i.e. maximize the following posteriori log-density function

$$p(\Pi, \Theta|X) \propto \sum_{i=1}^{n} \log f(x_i|\Pi, \Theta) + \log p(\Pi|\beta) + \sum_{k=1}^{K} \log p(\theta_k|\alpha_k) \ . \tag{10}$$

The use of EM algorithm for MAP estimation requires at each step the computation of the conditional expectation values $z_i^{(t)}$ of the hidden parameters y_i at the E-step, which is the same as ML approach (Eq. 4) by substituting the common prior π with the label parameter π_i. During the M-step the maximization of the following complete-data log-likelihood function is performed

$$\begin{aligned} Q(\Pi, \Theta \ | \ \Pi^{(t)}, \Theta^{(t)}) = \sum_{i=1}^{n} z_i^{(t)} \{\log(\pi_i) + \log(p(x_i|\theta))\} + (1 - z_i^{(t)})\{\log(1-\pi_i) + \\ + \log(p(x_i|b))\} - \beta \sum_{i=1}^{n} \pi_i \pi_{\mathcal{N}_i} + \sum_{k=1}^{K} \sum_{l=1}^{\Omega} (\alpha_{kl} - 1) \log(\theta_{kl}) \ , \end{aligned} \tag{11}$$

independently for each parameter. This gives the following update equation for the pattern multinomial parameters:

$$\theta_{kl}^{(t+1)} = \frac{\sum_{i=1}^{n} z_i^{(t)} \delta_{ikl} + (\alpha_{kl} - 1)}{\sum_{i=1}^{n} z_i^{(t)} \sum_{l=1}^{\Omega} \delta_{ikl} + \sum_{l=1}^{\Omega} (\alpha_{kl} - 1)} \ , \tag{12}$$

while for the background model the update rules are the same as in the case of the ML approach (Eq. 5).

The maximization of the function Q with respect to the label parameters π_i reduces to the next quadratic expression

$$\beta\pi_{\mathcal{N}_i}(\pi_i^{(t+1)})^2 - (1+\beta\pi_{\mathcal{N}_i})(\pi_i^{(t+1)}) + z_i^{(t)} = 0 \ , \tag{13}$$

where the summation term $\pi_{\mathcal{N}_i}$ can include both updated labels ($\pi_m^{(t+1)}$) and not yet updated ($\pi_m^{(t)}$). The above equation has two roots

$$\pi_i^{(t+1)} = \frac{(1+\beta\pi_{\mathcal{N}_i}) \pm \sqrt{(1+\beta\pi_{\mathcal{N}_i})^2 - 4\beta\pi_{\mathcal{N}_i} z_i^{(t)}}}{2\beta\pi_{\mathcal{N}_i}} \ . \tag{14}$$

It can be easily shown that only the root with the negative sign is valid, since the other one is discarded due to the constraint $0 \leq \pi_i \leq 1$. Therefore, the above equation provides a simple update rule for the label parameters π_i which ensures the uniqueness of the solution and satisfies the constraint. Looking carefully at Eq. 14 we can make some useful observations. In the case where a substring x_i has high posterior probability value ($z_i^{(t)} \approx 1$), one of the following two scenarios will occur in the neighborhood $\mathcal{N}_i$:

- None of sites $m \in \mathcal{N}_i$ is labeled as a pattern, i.e. $\pi_{\mathcal{N}_i} \lessapprox 1$, and thus, following Eq. 14, this site will be labeled as pattern ($\pi_i^{(t+1)} \approx 1$).
- There is at least one site labeled as pattern in $\mathcal{N}_i$, i.e. $\pi_{\mathcal{N}_i} \gtrapprox 1$. Then, according to Eq. 14, the new label value will be approximately $\pi_i^{(t+1)} \approx \frac{1}{\beta\pi_{\mathcal{N}_i}}$. The larger the value of $\pi_{\mathcal{N}_i}$, the smaller the update label values of π_i. In this "overlapping" neighborhood only one pattern occurrence will be the most probable to survive, the one having the higher posterior value z_i.

On the other hand, when a substring x_i has small posterior value of being a pattern ($z_i^{(t)} \approx 0$) it will continue to be labeled as background ($\pi_i^{(t+1)} \approx 0$), independently of its neighborhood $\mathcal{N}_i$.

From the above analysis it is clear that the regularization parameter β of the Gibbs distribution function plays a significant role. Only large values of this parameter ($\beta \gg 1$) are acceptable in order to discourage overlapping substrings being labeled as pattern. However, in our experiments, a large range of values of β seems to yield a satisfactory behavior, which implies that the proposed method is not sensitive to this parameter. A typical value that has been used successfully during experiments is $\beta = 100$.

Discovering multiple patterns. This can be accomplished by iteratively apply the two-component mixture of multinomials model, after erasing from the set of sequences S the patterns that have been already found. In particular, after convergence of the EM algorithm all substrings x_i whose label parameters π_i surpass a threshold value T (e.g. $T = 0.9$) are deleted from the S. A new set S' is then created, and the initial model is sequentially applied to it to discover another pattern.

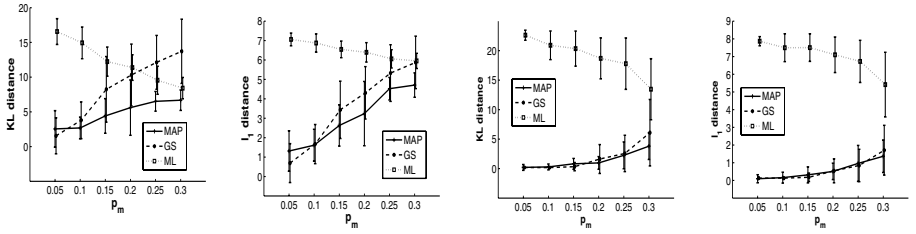
2.3 Initializing Pattern Multinomial Models

The major drawback of the EM algorithm is the dependence on the initial values of the model parameters that may cause it to get stuck in local maxima of the likelihood function [11]. In our study, this problem is mainly concentrated on initializing the pattern multinomial θ, since for the background density b we can use the total relative frequencies of characters in sequences. To overcome this weakness we present here a clustering scheme that is based on the classical k-means algorithm and its recent extensions to categorical data [15].

In the general case the k-means algorithm aims at finding a partition of M disjoint clusters to a set of n objects, so as the overall sum of distances between cluster centers μ_j and objects is minimized. Depending on the type of objects, one must determine an appropriate distance function and also a method for estimating cluster centers. Since we are dealing with discrete data, we define the simple Hamming distance $d(x_i, \mu_j) = \sum_{k=1}^{K}(1 - \delta(x_{ik}, \mu_{jk}))$ to measure similarities among the samples and the cluster centers. Moreover, the cluster center μ_j $j = 1, \ldots, M$ is determined as the *median* substring of the cluster members, i.e. $\mu_j = x_{i^*}$, where $i^* = \arg\min_i \sum_{x_k} d(x_i, x_k)$. When finishing, we initialize the pattern model with the relative frequencies of characters from the cluster j^* that has the minimum average distance (*intracluster* distance) between all cluster members and its center μ_{j^*}. It must be also noted that, in order to avoid selection of outliers, we isolate our search over clusters whose size (number of members) is above a threshold value, e.g. $N/2$. The experimental study has shown that this clustering scheme provides excellent initial values of parameter θ in a very fast way.

3 Experimental Results

Several experiments were performed in an attempt to study the effectiveness of the proposed MAP approach. During all experiments the proposed clustering scheme of k-means was first applied to generate an initial pattern multinomial model, and then the EM algorithm was used for MAP estimation of the model parameters. The pattern labels π_i were all initialized to $\pi_i = 0.5$. Comparative results have been also obtained using the ML approach without spatial constraints (initialized identically to the MAP approach), as well as the Gibbs sampling (GS) method [3, 4]. Starting by an initial (random) estimation of the positions of the patterns ($\theta^{(0)}$), the GS method performs iteratively two steps until likelihood convergence: first, it randomly selects a sequence S_i and re-estimates the pattern model $\theta^{(t+1)}$ using the current pattern positions of all sequences but S_i, and then, a new pattern position is selected in S_i by sampling from the posterior distribution over positions. Obviously, this version of the GS assumes that each sequence has a unique occurrence of the pattern. However, this is not true for our model that permits an arbitrary number of pattern copies in each sequence. Thus, in order to provide fair comparisons, we created sets with a single copy of any pattern to every sequence. Moreover, we execute the GS method 10 times and keep the best solution found.



(a) seed pattern: TATATATATATATTTT (b) seed pattern: HIERHIERVIEWVIEW

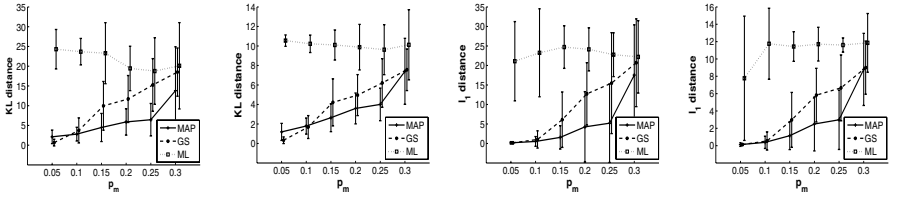
Fig. 1. Mean values and stds of KL and l_1 as calculated by the three methods in two single-pattern discovery problems (a), (b) using six values of noise parameter p_m.

The artificial sets used in the experiments were generated as follows: Using a number of *seed* patterns (one or two in our cases), every sequence contained a noisy copy of any of these seeds, according to a probability p_m common to every pattern position. The rest (non-pattern) positions were filled arbitrarily with characters following a uniform distribution over the alphabet Σ. Six different values for the noise parameter were used $p_m = [0.05, 0.1, 0.15, 0.2, 0.25, 0.3]$, and for each value we generated 40 different sets of $N = 20$ sequences with a mean length of 100 characters (totally 6×40 sets for each problem). From the above it is clear that we are aware of each true pattern density ϑ (estimated from the relative frequencies of characters of each noisy pattern copies). Following this scheme, four such pattern discovery problems were constructed using the DNA alphabet ($\Omega = 4$) and the protein alphabet ($\Omega = 20$), while the pattern length was always $K = 16$. To evaluate the discovered patterns by each method two information content criteria were used: the Kullback-Libler (KL) distance and the sum of the absolute differences (l_1 distance) between the estimated $\hat{\theta}$ and the true density $\vartheta = [\vartheta_{kl}]$, given by

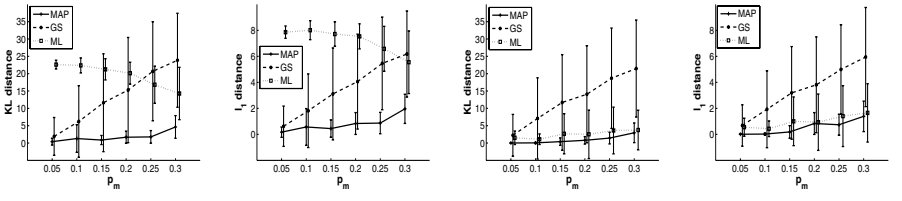
$$KL(\hat{\theta}||\vartheta) = \sum_{k=1}^{K}\sum_{l=1}^{\Omega} \hat{\theta}_{kl} \log \frac{\hat{\theta}_{kl}}{\vartheta kl} \quad , \quad l_1(\hat{\theta}||\vartheta) = \sum_{k=1}^{K}\sum_{l=1}^{\Omega} |\hat{\theta}_{kl} - \vartheta kl| \ . \tag{15}$$

At first we examined the capability of the proposed MAP approach to clearly identify a single pattern without estimating overlapping copies of it. The results of the three comparative methods are shown in Fig. 1, for two such problems (a), (b). These diagrams illustrate the average values and standard deviations of the two evaluation measurements (KL and l_1) obtained by each method to the created 6×40 different set of sequences. As it is obvious, the proposed MAP approach achieves properly the identification of the patterns in all noisy environments, while the GS method maintains satisfactory performance only to low levels of noise rate. On the other hand, as expected, the weakness of ML approach to distinguish overlapping copies of patterns leads to lower discrimation ability.

We have also tested our method to problems with two sequence patterns. In an attempt to increase the difficulty of their discovery, half of the sites in both seed patterns presented identical characters. The results (mean values and stds of



(a) seed pattern1: {TATATATATATATTTT} seed pattern2: {CGCGTATATATACGCG}



(b) seed pattern1: {HIERHIERVIEWVIEW} seed pattern2: {HIERARCHICALVIEW}

Fig. 2. Comparative results taken by applying the three methods in two discovery problems with two seed patterns. Again, calculated mean values and stds of KL and l_1 are illustrated in those diagrams.

KL and l_1) for two such problems (one from each alphabet) are demonstrated in Fig. 2 as obtained by each method. The weakness of the GS method to separate them is apparent, especially in sets with low homology patterns (large values of p_m). On the other hand, the MAP method exhibits almost perfect distinguishing capability by estimating properly the true model of both patterns. Finally, the ML approach presents the tendency to discover only one complex pattern obtained from the synthesis of both, and has shown good performance only in the case of the 2nd pattern of problem (b), where there are not repeated characters.

4 Conclusions

This paper presents a new spatially-constrained approach for discovering probabilistic patterns in sequences. The method uses a mixture of multinomials model with two components for modeling the pattern and the background of sequences. The spatial information is embodied in the model by treating the pattern labels as random variables that form a MRF to modeling their dependencies. The EM algorithm is then used to the reduced MAP problem for estimating the model parameters, after initializing with a clustering scheme that hires properties from the popular k-means algorithm. Experiments, conducted on a variety of categorical time-series, have shown the ability of the MAP method to identify qualitatively better patterns with repeated characters in comparison with the ML approach without constraints and the GS method. Further research can be focused on designing more complex pattern models that can also take into account gaps among sites, as well as on considering patterns of variable length.

References

1. Brāzma A., Jonasses I., Eidhammer I., and Gilbert D. Approaches to the automatic discovery of patterns in biosequences. *Journal of Computational Biology*, 5(2): 277–303, 1998.
2. Bréjova B., DiMarco C., Vinař T., Hidalgo S.R., Holguin G., and Patten C. Finding patterns in biological sequences. Project Report for CS798g, University of Waterloo, 2000.
3. Lawrence C.E., Altschul S.F., Boguski M.S., Liu J.S., Neuwland A.F., and Wootton J.C. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, 226:208–214, 1993.
4. Liu J.S., Neuwald A.F., and Lawrence C.E. Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *J. Amer. Statistical Assoc*, 90:1156–1169, 1995.
5. Bailey T.L. and C. Elkan C. Unsupervised learning of multiple motifs in Biopolymers using Expectation Maximization. *Machine Learning*, 21:51–83, 1995.
6. Hughey R. and Krogh A. Hidden Markov models for sequence analysis: Extension and analysis of the basic method. *CABIOS*, 12(2):95–107, 1996.
7. Liu X., Brutlag D.L., and Liu J.S. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. In *Pac. Symp. Biocomput*, pages 127–138, 2001.
8. Blekas K., Fotiadis D.I., and Likas A. Greedy mixture learning for multiple motif discovering in biological sequences. *Bioinformatics*, 19(5):607–617, 2003.
9. Xing E.P., Wu W., Jordan M.I., and Karp R.M. LOGOS: A modular Bayesian model for *de novo* motif detection. *Journal of Bioinformatics and Computational Biology*, 2(1):127–154, 2004.
10. Dempster A.P., Laird N.M., and Rubin D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. B*, 39:1–38, 1977.
11. McLachlan G.M. and Peel D. *Finite Mixture Models*. New York: John Wiley & Sons, Inc., 2001.
12. Besag J. Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. Roy. Stat. Soc., ser. B*, 36(2):192–326, 1975.
13. Geman S. and Geman D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.
14. Blekas K., Likas A., Galatsanos N.P., and Lagaris I.E. A Spatially-Constrained Mixture Model for Image Segmentation. *IEEE Trans. on Neural Networks*, 62(2):494–498, 2005.
15. Huang Z. Extensions to the k-Means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, 2(3):283–304, 1998.