Parametric Model-based Clustering[§]

Vladimir Nikulin[a] and Alex J. Smola[b]

[a]Computer Sciences Laboratory, RSISE, Australian National University, Canberra, Australia;
[b]NICTA, Canberra, ACT 0200, Australia

ABSTRACT

Parametric, model-based algorithms learn generative models from the data, with each model corresponding to one particular cluster. Accordingly, the model-based partitional algorithm will select the most suitable model for any data object (*Clustering step*), and will recompute parametric models using data specifically from the corresponding clusters (*Maximization step*). This Clustering-Maximization framework have been widely used and have shown promising results in many applications including complex variable-length data.

The paper proposes *Experience-Innovation* (EI) method as a natural extension of the *Clustering-Maximization* framework. This method includes 3 components: 1) keep the best past experience and make empirical likelihood trajectory monotonical as a result; 2) find a new model as a function of existing models so that the corresponding cluster will split existing clusters with bigger number of elements and smaller uniformity; 3) heuristical innovations, for example, several trials with random initial settings.

Also, we introduce clustering regularisation based on the balanced complex of two conditions: 1) significance of any particular cluster; 2) difference between any 2 clusters.

We illustrate effectiveness of the proposed methods using first-order Markov model in application to the large web-traffic dataset. The aim of the experiment is to explain and understand the way people interact with web sites.

Keywords: model-based clustering, Markov models, fuzzy c-means, web-traffic data

1. INTRODUCTION

Existing papers on the model-based clustering largely concentrate on a specific models or applications. [1] A notable exception is the work of Ref.2 who proposed a maximum likelihood (ML) model within Expectation-Maximization (EM) framework[3] for partitional soft clustering. The model of Ref.2 may be particularly useful in the cases of complex variable-length data such as web-traffic data. This approach provides a natural and consistent mechanism for handling the problems of modelling and clustering sequences of different lengths.

The problem of predicting user's behavior on a web-site has gained importance due to the rapid growth of the world-wide-web and the need to personalize and influence a user's browsing experience. [4] Markov models and their variations have been found well suited for addressing this problem. In general, the input for these problems is the sequence of web-pages that were accessed by a user and the goal is to build Markov models that can be used to model and predict the web-page that the user will most likely access next. This study will help to explore and understand human behavior within internet environment.[5]

In the Sect. 2 we formulate likelihood-based hard clustering approach within general Clustering-Maximization (CM) framework. The popularity of the CM framework stems from monotonical and convergence properties. [6,7] Note that CM framework for a hard clustering may be viewed as an analog of the EM framework for a soft clustering.

The unified framework for parametric, model-based algorithms was presented in Ref.1. Accordingly, any cluster is represented by the model as a set of parameters. Using these models we will compare different observations indirectly by comparing relations between observations and models. In this view, clusters are represented as a probabilistic models in a model space that is conceptually separate from the data space. Model-based methods offer an effective interpretability since the resulting model for each cluster directly characterizes that cluster. For example, in the case of Markov model, Sect. 3, the set of required parameters will include: 1) probabilities of first (start) and last (exit) states; 2) transitional probabilities between states.

[§]Email: vladimir.nikulin@anu.edu.au; Phone: +61-02-612-55908; Web: http://rsise.anu.edu.au/csl/

Data Mining, Intrusion Detection, Information Assurance, and Data Networks Security 2005, edited by Belur V. Dasarathy, Proceedings of SPIE Vol. 5812 (SPIE, Bellingham, WA, 2005) · 0277-786X/05/$15 · doi: 10.1117/12.603199

The performance of the CM-algorithms depends essentially on the selection of initial settings if number of clusters is bigger than one. Accordingly, we propose an iterative procedure to compute a new model as a function of the existing models. A new model will split existing clusters with bigger number of elements (weight) and higher non-uniformity (divergency). This procedure will keep monotonical property and may be regarded as a natural extension of the standard CM-framework.

Determination of the number of clusters k represents an important problem. For example, Ref.8 proposed G-means algorithm which is based on the $Gaussian$ fit of the data within particular cluster. Usually, [1] attempts to estimate the number of Gaussian clusters lead to a very high value of k. Most simple criterions such as AIC (*Akaike Information Criterion*[9]) and BIC (*Bayesian Information Criterion*[10, 11]) either overestimate or underestimate the number of clusters, which severely limit their practical usability. We introduce in the Sect. 4 special clustering regularization. This regularization restricts creation of a new cluster which is not big enough or which is not sufficiently different comparing with existing clusters.

Section 7 presents detailed experimental results on a large web-traffic **msnbc** dataset. [12] Note that all numerical and graphical illustrations in this paper were produced using **msnbc** dataset.

In the Sect. 5 we consider modification of the fuzzy c-means algorithm [13] which is marginally linked to the EM-algorithm.

According to Ref.14, mixture regression methods have received a great deal of attention in marketing research. However, the use of $GLiMMix$ (*Generalized Linear Model for Mixture Distributions*) is not limited to marketing research, but extends to business and economic research, psychology, sociology, anthropology, political science, etc.

In the Sect. 6 we consider *Generalised Linear Clustering* ($GLiC$) and *Stochastic Unfolding Clustering* ($StUnC$) algorithms as a distance-based clustering examples within parametric, model-based framework.

2. LIKELIHOOD-BASED CLUSTERING

Suppose $\mathbf{X} := \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ is a sample of independent observations with variable-lengths. Any particular observation $\mathbf{x}_i$ represents a vector of N_i i.i.d. components with density $f(x, \theta)$ from known family of probability distributions where parameter $\theta \in \Lambda$ (generally, θ may be regarded as a set of parameters) is unknown.

We will denote by Θ a codebook as a set of k models $\Theta(c)$ indexed by the code $c = 1..k$ where k is a number of clusters or *a clustering size*.

It may not be easy to compare different observations directly (using distance-based approach) in the case of variable-length dimension. Respectively, we will employ model-based partitional clustering approach. The key element of this approach is the idea to compare different observations indirectly through the corresponding likelihoods computed according to k different models.

The aims of the model $\Theta(c)$ are different depending on the particular step of the CM cycle: $\Theta(c)$ may be viewed as a generalized prototype [15] of the cluster c at the *Clustering* step or as a generalized center of cluster c (or centroid) at the *Maximization* step.

REMARK 2.1. The algorithm 1 represents just a particular illustration and we will understand further under notation CM-algorithm an arbitrary algorithm within CM-framework.

Let us define the empirical log-likelihood:

$$L_{\text{emp}}[\Theta] := \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{N_i} \log f\left(x_{ij}, \Theta(c(\mathbf{x}_i))\right). \tag{3}$$

The following Proposition 2.1, that may be proved similarly to the Propositions 1 and 2 of Ref.7, formulates the most important ascending and convergence properties of the *Clustering Maximization* framework.

PROPOSITION 2.1. The CM-Algorithm 1
1) monotonically increases the value of the objective function (3);
2) converges to a local maximum in a finite number of steps if Maximization step has a unique solution.

In the following sub-section we consider example in order to illustrate and support the last condition of the Proposition 2.1.

Algorithm 1 CM.

1: `Clustering`: encode any observation $\mathbf{x}_i$ according to the most likely model:

$$c(\mathbf{x}_i) := \underset{c\in\{1..k\}}{\operatorname{argmax}} \prod_{j=1}^{N_i} f(x_{ij}, \Theta(c)). \tag{1}$$

2: `Maximization`: re-compute models specifically for any particular empirical cluster

$$\Theta(c) := \underset{\lambda\in\Lambda}{\operatorname{argsup}} \sum_{c(\mathbf{x}_i)=c} \sum_{j=1}^{N_i} \log f(x_{ij}, \lambda). \tag{2}$$

3: `Test.`: compare previous and current codebooks Θ. Go to the step 1 if convergence test is not fulfilled, alternatively, stop the algorithm.

2.1. Exponential families.

Let f be an exponential probability density $f(x,\theta) := \exp\{\langle\phi(x),\theta\rangle - g(\theta)\}$ where $\theta \in \mathbb{R}^m$ is an unknown parameter, $\phi(x)$ is a sufficient statistics and $g(\theta)$ is the corresponding log-partition function, $\langle\cdot,\cdot\rangle$ denotes an operation of the scalar product in a Hilbert space.

Then, the expression in the right side of (2) may be re-written as follows $\sum_{c(\mathbf{x}_i)=c} \sum_{j=1}^{N_i} \langle\phi(x_{ij}),\theta\rangle - K_c g(\theta)$. Maximizing this function by θ we will obtain equation

$$\bigtriangledown g(\theta) = \frac{\partial g(\theta)}{\partial\theta} = \frac{1}{K_c} \sum_{c(\mathbf{x}_i)=c} \sum_{j=1}^{N_i} \phi(x_{ij}) \tag{4}$$

where $K_c = \sum_{\mathbf{x}_i\in\mathbf{X}_c} N_i$. It follows from convexity of the function $g(\theta)$ that (4) has unique solution.

2.2. EI-method and computation of a next model as a function of the existing models.

The CM-algorithm will split the data from the sample $\mathbf{X}$ into k empirical clusters $\mathbf{X}_c = \{\mathbf{x}_i : c(\mathbf{x}_i) = c\}$.

Then, we will use empirical KL-divergence[16] $\mathcal{L}_{cc}$ in order to measure uniformity within cluster c and compute the distance between models $\Theta(c)$ and $\Theta(a)$:

$$\Psi(c,a) = \mathcal{L}_{ca} - \mathcal{L}_{cc} \tag{5}$$

where

$$\mathcal{L}_{ca} = -\frac{1}{K_c} \sum_{\mathbf{x}_i\in\mathbf{X}_c} \sum_{j=1}^{N_i} \log f(x_{ij}, \Theta(a)). \tag{6}$$

Besides, we will compute prior empirical probabilities of clusters $\pi_c \propto K_c$.

According to Fig. 1(a)-(b) there is no clear correspondence between π_c and $\mathcal{L}_{cc}$, but $\max_c \pi_c$ and $\max_c \mathcal{L}_c$ may be viewed as a decreasing functions of k.

The following two classes of approaches are typically used for the selection of the cluster to split[17]:

A1) the cluster with largest number of elements;

A2) the cluster with higher divergency.

Accordingly, we form a new model as a linear combination of the existing models

$$\Theta(k+1) = \sum_{c=1}^{k} w_c \cdot \Theta(c), k \geq 2, \tag{7}$$

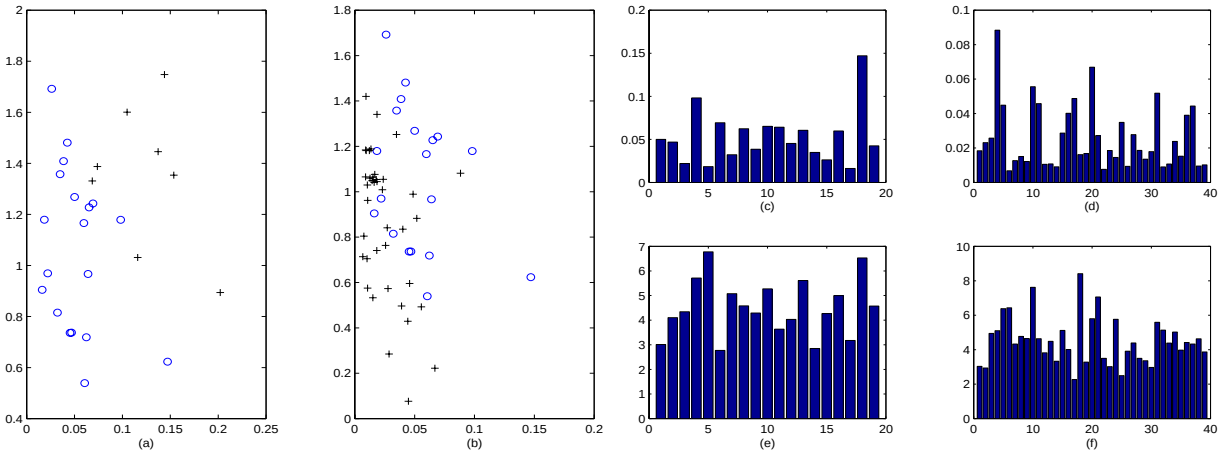

Figure 1. a) $\mathcal{L}_{cc}$ as a function of π_c, cases of 8 and 19 clusters marked by "+" and "o"; b) cases of 19 and 39 clusters marked by "o" and "+", respectively; c) and d) prior empirical probabilities of clusters; e) and f) average numbers of events in the cases of $k = 19$ and $k = 39$.

where $w_c \propto \exp\{\pi_c(\lambda + \mu\mathcal{L}_{cc})\}, \sum_{c=1}^{k} w_c = 1, \lambda > 0$ and $\mu > 0$ are regulation parameters.

The definition of optimal model in case of one cluster is unique and will be found immediately after the first Maximization step. Unfortunately, one model will not give us a direction in order to define a next, second model in (7). In order to overcome this problem and as it was proposed in Ref.17, we will generate randomly a small deviation as a difference between first and second models. Third and following models may be calculated as a linear combinations (7) of the existing models with weight coefficients w_c to be defined according to the multinomial logit (MNL) model.

The coefficients w_c represent the balanced compromise between criterions $A1$ and $A2$: on the one hand, we are interested to create a new cluster closer to the existing clusters with bigger weight (controlled by π_c), on the other hand, we are interested to split clusters with higher divergency (controlled by $\mathcal{L}_{cc}$).

Note that the definition (7) is just one example of many possible designs motivated by the particular Markov model of the following section. The definition (7) represents a basic idea which may be modified for a specific application.

DEFINITION 2.1. The EI-method includes 3 components:

B1) keep the best past experience and make trajectory monotonical as a result (see Proposition 2.2);

B2) find a new prototype as a function of the set of the existing prototypes (7);

B3) heuristical innovations: for example, it may be several trials with random initial settings (see Fig. 2(a)-(b)).

PROPOSITION 2.2. The following relation is valid

$$L_{\text{emp}}[\Theta_+] \geq L_{\text{emp}}[\Theta] \tag{8}$$

where Θ_+ represents a union of Θ and arbitrary model.

Proof. By definition, any observation will be directed to the model which represents higher likelihood. Accordingly, the data will be attracted to the new cluster as a consequence of a higher likelihood. Respectively, the combined likelihood of the data within new cluster may not be smaller. ■

REMARK 2.2. The condition $\Theta \subset \Theta_+$ is a very essential for (8). As an alternative, we can observe the situation where an outcome of CM-algorithm in the case of $k + 1$ clusters is worse comparing with result corresponding to k clusters. Fig. 2(a)-(b) illustrate above facts.

3. MARKOV HARD CLUSTERING

Suppose we have a dataset of n records of dynamic behavior of individuals visiting web sites classified into m different categories (states). The length N_i of any particular record $\mathbf{x}_i, i = 1..n$, is not fixed, and represents a vector of integer indexes (events)

$$1 \le s_{ij} \le m, j = 1..N_i, i = 1..n,$$

$N = \sum_{i=1}^{n} N_i = \sum_{c=1}^{k} K_c$ is the total number of events.

Under above assumptions we can simplify (6) $\mathcal{L}_{ac} = -\sum_{i=1}^{m} p_{ai} \log \theta_{ci}, \quad p_{ai} = \frac{q_{ai}}{K_a}$, q_{ai} is a number of the event i in the cluster a, θ_{ci} is the probability of the event i within the cluster c.

As a next step, we consider a first-order Markov model $\Theta(c)$ with

- vectors of marginal probabilities $\theta_{ci}^{(v)} \ge 0, c = 1..k, i = 1..m, \ \sum_{i=1}^{m} \theta_{ci}^{(v)} = 1$ where index $v = 1$ indicates "start", and index $v = 2$ indicates "exit";
- matrix of first-order-Markov probabilities: $\theta_{cij} \ge 0, c = 1..k, i, j = 1..m, \ \sum_{j=1}^{m} \theta_{cij} = 1.$

REMARK 3.1. The case $N_i = 1$ is a very specific: on the one hand, $\mathbf{x}_i$ will not add any information to the statistics of the Markov probabilities, on the other hand, we will count $\mathbf{x}_i$ twice as "start" and "exit".

The following definition corresponds directly to (1)

$$c(\mathbf{x}_i) = \operatorname*{argmax}_a L(a, \mathbf{x}_i, \Theta) \tag{9}$$

where

$$L(a, \mathbf{x}_i, \Theta) = \log \theta_{as_{i1}}^{(1)} + \log \theta_{as_{iN_i}}^{(2)} + \frac{n}{N-n} \sum_{j=2}^{N_i} \log \theta_{as_{ij-1}s_{ij}}. \tag{10}$$

DEFINITION 3.1. The following empirical probabilities will be used bellow

$$p_c = \frac{n_c}{n}, \ p_{ci}^{(v)} = \frac{q_{ci}^{(v)}}{n_c}, \ p_{cij} = \frac{q_{cij}}{M_{ci}}, \ \varphi_{ci} = \frac{M_{ci}}{M_c}, \ \pi_c = \frac{M_c}{N-n} \tag{11}$$

where $q_{ci}^{(1)}$ is the number of users with "start" state i; $q_{ci}^{(2)}$ is the number of users with "exit" state i; q_{cij} is the number of the consecutive states i and j in the cluster c; $n_c = \sum_{i=1}^{m} q_{ci}^{(v)}, v = 1..2$; $M_{ci} = \sum_{j=1}^{m} q_{cij}, M_c = \sum_{i=1}^{m} M_{ci}, N = \sum_{c=1}^{k} M_c + n$.

PROPOSITION 3.1. The solution of the equation $\Theta(c) = \operatorname{argmax}_{\mathcal{Q}} \sum_{\mathbf{x}_i \in \mathbf{X}_c} L(c, \mathbf{x}_i, \mathcal{Q})$ is unique and is defined as follows

$$\theta_{ci}^{(v)} = p_{ci}^{(v)}, \ \theta_{cij} = p_{cij} \tag{12}$$

where $v = 1..2, c = 1..k; i, j = 1..m$.

Proof. The target of the `Maximization` step is to maximize the log-likelihood:

$$\frac{1}{n} \sum_{i=1}^{n} L(c(\mathbf{x}_i), \mathbf{x}_i, \Theta). \tag{13}$$

We can re-write (13) using different terms:

$$\sum_{c=1}^{k} \sum_{i=1}^{m} \left[\frac{\sum_{v=1}^{2} q_{ci}^{(v)} \log \theta_{ci}^{(v)}}{n} + \sum_{j=1}^{m} \frac{q_{cij} \log \theta_{cij}}{N-n} \right] = \sum_{c=1}^{k} p_c \sum_{i=1}^{m} \sum_{v=1}^{2} p_{ci}^{(v)} \log \theta_{ci}^{(v)} + \sum_{c=1}^{k} \pi_c \sum_{i=1}^{m} \varphi_{ci} \sum_{j=1}^{m} p_{cij} \log \theta_{cij}. \tag{14}$$

Required solutions (12) follow directly from above formula (14). ■

REMARK 3.2. Table 1 illustrates monotonical ascending property of the Markov model-based algorithm defined in the Proposition 3.1 under general result of the Proposition 2.1.

Table 1. The normalized empirical log-likelihood $L_{\rm norm}$ defined in (17) as a function of iteration (CM-cycle) τ in the case $k = 19$; 2) Clustering step; 3) Maximization step; 4) the distance (30).

τ	Clustering	Maximization	$\Delta(\tau)$	τ	Clustering	Maximization	$\Delta(\tau)$
1	-2.233340	-1.197748	21.9437	7	-0.895848	-0.895805	0.1307
2	-1.058968	-0.962400	7.0320	8	-0.895695	-0.895493	0.0495
3	-0.940187	-0.917314	3.0885	9	-0.895428	-0.895099	0.0452
4	-0.908967	-0.901721	1.4871	10	-0.895005	-0.894712	0.0314
5	-0.898962	-0.896614	0.6298	11	-0.894674	-0.894662	0.0266
6	-0.896350	-0.895910	0.2339	12	-0.894660	-0.894655	0.0205

4. CLUSTERING REGULARIZATION

In this section we formulate an approach in terms of the following 2 conditions:

C1) $p_c \geq \alpha_1 > 0; \pi_c \geq \alpha_2 > 0, c = 1..k$ (significance of any particular cluster);

C2) $\mathcal{D}_v(a,b) \geq \beta_v(p_a + p_b)\log 2 > 0, v = 1..2;\ \ \mathcal{D}_3(a,b) \geq \beta_3(\pi_a + \pi_b)\log(2m) > 0$ (difference between any 2 clusters a and $b, a \neq b$)

where

$$\mathcal{D}_v(a,b) = \sum_{c\in\{a,b\}} p_c \sum_{i=1}^{m} p_{ci}^{(v)} \log \frac{p_{ci}^{(v)}}{\widehat{p}_i^{(v)}};\ \ \mathcal{D}_3(a,b) = \sum_{c\in\{a,b\}} \pi_c \sum_{i=1}^{m} \varphi_{ci} \sum_{j=1}^{m} p_{cij} \log \frac{p_{cij}}{\widehat{p}_{ij}};$$

$$\widehat{p}_i^{(v)} = \frac{p_a p_{ai}^{(v)} + p_b p_{bi}^{(v)}}{p_a + p_b}, \widehat{p}_{ij} = \frac{\pi_a \varphi_{ai} p_{aij} + \pi_b \varphi_{bi} p_{bij}}{\pi_a \varphi_{ai} + \pi_b \varphi_{bi}}.$$

PROPOSITION 4.1. The following relations are valid

$$\mathcal{D}_v(a,b) \leq \left(p_a^{(v)} + p_b^{(v)}\right)\log 2, v = 1..2; \mathcal{D}_3(a,b) \leq (\pi_a + \pi_b)\log(2m). \tag{15}$$

Proof. We will consider a more complex case of $\mathcal{D}_3(a,b)$ where $m \geq 2$. (the case $m = 1$ is trivial)

Maximizing $\mathcal{D}_3(a,b)$ we will draw an immediate conclusion that the involved vectors of probabilities must be orthogonal: $\sum_{j=1}^{m} p_{aij} p_{bij} = 0, \quad \forall i = 1..m$.

Therefore,

$$\mathcal{D}_3(a,b) \leq \sum_{i=1}^{m} \sum_{c\in\{a,b\}} \pi_c \varphi_{ci} \log \frac{\pi_a \varphi_{ai} + \pi_b \varphi_{bi}}{\pi_c \varphi_{ci}} \leq -\sum_{c\in\{a,b\}} \pi_c \sum_{i=1}^{m} \varphi_{ci} \log(\pi_c \varphi_{ci}) \leq \log(2m)$$

where the final bound was obtained using uniform probabilities: $\pi_c = 0.5, \varphi_{ci} = \frac{1}{m}, \quad \forall c \in \{a,b\}, i = 1..m$. The upper bounds for $\mathcal{D}_v, v = 1..2$, may be proved using similar methods. ■

Let us define an optimal empirical log-likelihood $L_{\rm emp}^{(k)} = \sup_\Theta \frac{1}{n}\sum_{i=1}^{n} L(c(\mathbf{x}_i), \mathbf{x}_i, \Theta)$ assuming that the codebook Θ contains k prototypes, and the code $c(\mathbf{x}_i)$ is defined in (9).

As it was noticed in Ref.15, if more prototypes are used for the k-means clustering, the algorithm splits clusters, which means that it represents a single cluster by more than one prototype. The following Proposition 4.1 will consider clustering procedure in inverse direction.

Table 2. 1) number of clusters k; 2-3) innovative trajectories of the normalized empirical log-likelihood $L_{\rm norm}$ (19) with random initial settings; 4) $T_{\rm exp}$ was produced according to the EI-method, and in a complex with T_1; 5) $T_{\rm reg}$ is a trajectory of (17).

k	T_1	T_2	$T_{\rm exp}$	$T_{\rm reg}$	k	T_1	T_2	$T_{\rm exp}$	$T_{\rm reg}$
30	-0.7396	-0.7456	-0.7389	-0.9354	40	-0.6557	-0.649	-0.6479	-0.9099
31	-0.7374	-0.7374	-0.731	-0.9341	41	-0.6398	-0.6398	-0.6446	-0.9131
32	-0.722	-0.722	-0.7251	-0.9347	42	-0.6437	-0.6437	-0.6284	**-0.9035**
33	-0.6969	-0.6968	-0.7162	-0.9323	43	-0.6347	-0.6321	-0.6255	-0.9071
34	-0.7054	-0.7054	-0.6862	-0.9089	44	-0.6158	-0.617	-0.6212	-0.9094
35	-0.6914	-0.6914	-0.6776	-0.9069	45	-0.6302	-0.6301	-0.6109	-0.9056
36	-0.7073	-0.6845	-0.6703	-0.9061	46	-0.649	-0.6155	-0.6041	-0.9054
37	-0.6776	-0.6679	-0.6664	-0.9087	47	-0.6145	-0.6296	-0.6009	-0.9087
38	-0.6849	-0.6886	-0.663	-0.9119	48	-0.6277	-0.6007	-0.5956	-0.91
39	-0.6664	-0.6734	-0.6571	-0.9126	49	-0.5991	-0.5987	-0.5902	-0.9112

PROPOSITION 4.2. Assuming that we merge clusters a and b, the following relation is valid

$$L_{\rm emp}^{(k)} - L_{\rm emp}^{(k-1)} \leq \sum_{v=1}^{3} \mathcal{D}_v(a,b) \tag{16}$$

where a strict equality will take place if and only if the merged cluster equal to a union of the input clusters a and b, $k \geq 2$.

Proof. Required bound follows from (12) and (14). $L_{\rm emp}^{(k-1)}$ may become bigger in 2 cases: 1) some data from the merged cluster will be attracted by the other models; 2) the model of the merged cluster will attract data from the other clusters. Respectively, the difference between log-likelihoods (16) will become smaller. ■

According to the complex of the conditions $C1$ and $C2$ and bound (16) we define regularized empirical log-likelihood:

$$L_{\rm reg}\left[\Theta^\star\right] = L_{\rm norm}\left[\Theta^\star\right] - C_k \tag{17}$$

where

$$C_k = \frac{2k}{3}\left(\log(2)\alpha_1(\beta_1+\beta_2) + \log(2m)\alpha_2\beta_3\right); \tag{18}$$

$$L_{\rm norm}\left[\Theta^\star\right] = \frac{L_{\rm emp}\left[\Theta^\star\right]}{3}, \tag{19}$$

and $\Theta^\star$ is an outcome of the CM-procedure: $\Theta^\star = CM(\Theta^\star)$.

As a corollary of the Proposition 4.1 we obtain ranges $0 \leq \alpha_i, \beta_j \leq 1, i = 1..2, j = 1..3$, for the regulation parameters in (18).

The trajectory $T_{\rm reg}$ of (17) may be viewed in the Figure 2(a). Maximizing (17) we will make required selection of the number of clusters k.

REMARK 4.1. Empirical log-likelihood (13) and cost term (18) includes 3 independent components corresponding to the 1) "*start*", 2) "*exit*" and 3) *Markov* probabilities. Respectively, we can apply in (13) three weight linear coefficients and these coefficients will not affect formulas (12).

REMARK 4.2. Note, a structural similarity between (18) and AIC (*Akaike Information Criterion* [9]). Although, we used different assumptions and targets in order to derive (18).

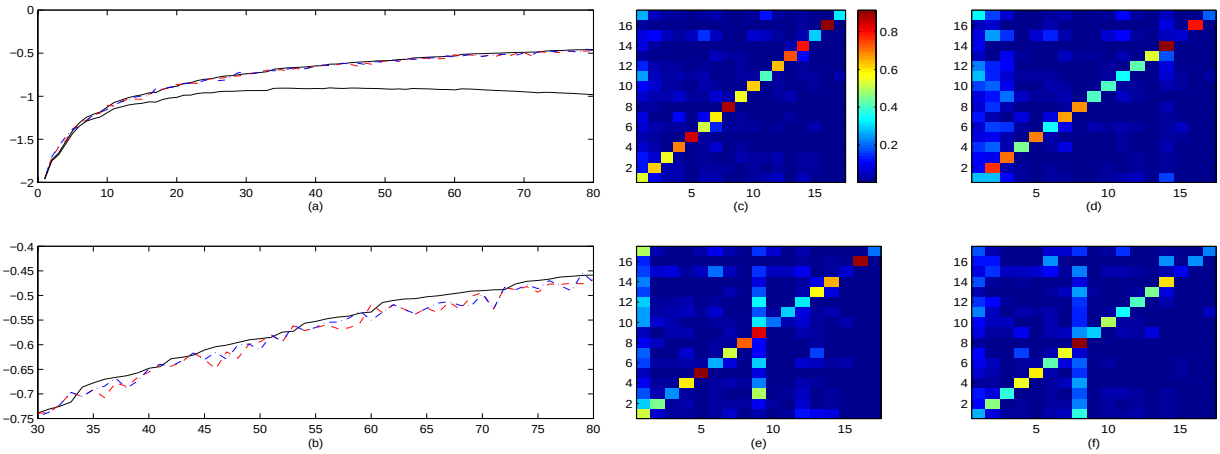

Figure 2. a) Trajectories T_1 (blue dashed line), T_2 (red dashdot line), $T_{\exp}$ (black solid line) and T_{reg}, $k = 1..80$; b) $T_i, i = 1..2$, $T_{\exp}$, $k = 30..80$, where $T_i, i = 1..2$, are innovative trajectories with random initial settings; $T_{\exp}$ was produced according to the EI-method and in a complex with T_1, T_{reg} is a trajectory of (17); c) Markov probabilities in the case of 1 cluster; d)-f) Markov probabilities for the biggest, 20th and 30th biggest clusters in the case 39 clusters, see Fig. 1(d).

Table 3. 1-2) web category; 3-4) "start" and "exit" general probabilities for the whole dataset (case of 1 cluster); 5-12) "start" and "exit" probabilities for four biggest clusters in the case of 39 clusters, see Fig. 1(d).

No	Web	Start	Exit	Start	Exit	Start	Exit	Start	Exit	Start	Exit
1	front	0.2802	0.1933	0.7186	0	0.9971	0.9999	0.2847	0	0	0
2	news	0.0779	0.0989	0	0.455	0	0	0	0	0	0
3	tech	0.0672	0.061	0	0.2643	0	0	0.2967	0.3473	0	0
4	local	0.0516	0.08	0	0	0	0	0.4186	0.6527	0	0
5	opinion	0.0043	0.0187	0	0	0	0	0	0	0	0.0605
6	onair	0.1648	0.0875	0.2814	0	0	0	0	0	0.9766	0.9395
7	misc	0.0036	0.0201	0	0	0	0	0	0	0	0
8	weather	0.0726	0.0914	0	0	0	0	0	0	0	0
9	health	0.066	0.0536	0	0	0	0	0	0	0	0
10	living	0.0129	0.0279	0	0	0	0	0	0	0.0234	0
11	business	0.0144	0.0244	0	0.042	0	0	0	0	0	0
12	sports	0.0573	0.0638	0	0	0	0	0	0	0	0
13	summary	0.0636	0.0622	0	0	0	0	0	0	0	0
14	bbs	0.0542	0.0968	0	0.2387	0	0	0	0	0	0
15	travel	0.0075	0.0145	0	0	0	0	0	0	0	0
16	msn-news	0.0004	0.0016	0	0	0	0	0	0	0	0
17	msn-sport	0.0015	0.0043	0	0	0.0029	0	0	0	0	0

5. MARKOV SOFT CLUSTERING

The *fuzzy c-means* (FCM) algorithm is commonly used for "soft" clustering. [13] We will modify the algorithm so that the results of the EM algorithm[3] may be obtained as a marginal limits if *fuzzy* parameter γ tends to 1 under condition: $\gamma > 1$.

Let us consider minimization of the following target function

$$\sum_{i=1}^{n}\sum_{c=1}^{k} \psi_{ci}^{\gamma} \pi_c^{1-\gamma} \left[1 + (1-\gamma) L(c, \mathbf{x}_i, \Theta)\right] \tag{20}$$

where $\sum_c \pi_c = 1, \sum_c \psi_{ci} = 1$ and the log-likelihood function L is defined in (10) as a function of probabilities of

membership ψ_{ci}, *prior* probabilities of clusters π_c and codebook Θ.

REMARK 5.1. The target function (20) includes regularization term $\sum_{i=1}^{n}\sum_{c=1}^{k}\psi_{ci}^{\gamma}\pi_c^{1-\gamma}$ which will award smoothed solution.

Using standard technique we will formulate iterative algorithm.

Algorithm 2 Modified FCM.

1: Re-compute probabilities of membership and prior probabilities:

$$\psi_{ci} \propto \pi_c \left(1+(1-\gamma)L(c,\mathbf{x}_i,\Theta)\right)^{\frac{1}{1-\gamma}}; \tag{21a}$$

$$\pi_c \propto \left(\sum_{i=1}^{n}\psi_{ci}^{\gamma}\left(1+(1-\gamma)L(c,\mathbf{x}_i,\Theta)\right)\right)^{\frac{1}{\gamma}}. \tag{21b}$$

2: Re-compute codebook Θ according to (23).

3: `Test.`

PROPOSITION 5.1. The solution of the equation

$$\Theta(c) = \underset{\mathcal{Q}}{\operatorname{argmax}}\sum_{i=1}^{n}\psi_{ci}^{\gamma}L(c,\mathbf{x}_i,\mathcal{Q}) \tag{22}$$

is unique and is defined as follows

$$\theta_{cs}^{(v)} \propto \sum_{i=1}^{n} I(i,v,s)\psi_{ci}^{\gamma};\quad \theta_{csz} \propto \sum_{i=1}^{n} N(i,s,z)\psi_{ci}^{\gamma} \tag{23}$$

where $I(i,1,s)$ is an indicator of the event: $x_{i1}=s$; $I(i,2,s)$ is an indicator of the event: $x_{iN_i}=s$; $N(i,s,z)$ is the number of consecutive events s and z in $\mathbf{x}_i$.

Initially, we generate randomly required input parameters ψ_{ci}, π_c and Θ. Then, using (21a) we recompute ψ_{ci}, which will be used consequently in (21b) and (23).

Algorithm 3 EM

1: `Expectation`:

$$\varphi_{ci} = \frac{\pi_c \cdot \exp L(c,\mathbf{x}_i,\Theta)}{\sum_{j=1}^{k}\pi_j \cdot \exp L(j,\mathbf{x}_i,\Theta)},\quad \pi_c = \frac{1}{n}\sum_{i=1}^{n}\varphi_{ci}. \tag{24}$$

2: `Maximization`:

$$\Theta(c) = \underset{\mathcal{Q}}{\operatorname{argmax}}\sum_{i=1}^{n}\varphi_{ci}\cdot L(c,\mathbf{x}_i,\mathcal{Q}). \tag{25}$$

3: `Test.`

Note that the restriction $\gamma > 1$ is a very essential for membership update (21a). The EM-Algorithm 3 is marginally linked to the FCM-Algorithm 2: Bayesian formula (24) within the Expectation step may be obtained from (21a) if $\gamma \to 1$.

REMARK 5.2. In line with the concepts of universal estimation[18] we introduce universal clustering with the target function (20). Considering behavior of the (π, ψ, Θ)-solution as a function of fuzzy parameter γ we can test stability of the clustering configuration.

COROLLARY 5.1. The solution of the equation (25) is unique and is defined as follows

$$\theta_{cs}^{(v)} \propto \sum_{i=1}^{n} I(i,v,s)\varphi_{ci}; \qquad \theta_{csz} \propto \sum_{i=1}^{n} N(i,s,z)\varphi_{ci}.$$

6. DISTANCE-BASED CLUSTERING

The $GLiC$ algorithm 4 uses as an input a sample of observations $\mathbf{x}_i$ with 2 components each: 1) a scalar target variable y_i and 2) a vector of explanatory variables $\mathbf{z}_i$ with fixed dimension m. By definition, any two observations may be compared *indirectly* using k vectors of linear coefficients $\theta(c)$, which have the same dimension as $\mathbf{z}_i$. Initially, we will select randomly matrix with k rows and m columns where rows $\theta(c), c = 1..k$, represent corresponding clusters. The $GLiC$ algorithm will assign any observation to the particular cluster depending on how close are the scalar products of the corresponding vectors of linear coefficients and explanatory variables to the target variable. Then, the algorithm will re-compute vectors of linear coefficients using least squared estimation (LSE) method applied specifically to the data from particular clusters.

Algorithm 4 $GLiC$.

1: `Clustering`:

$$c(\mathbf{x}_i) := \underset{c\in\{1..k\}}{\operatorname{argmin}} \ \|y_i - \langle \theta(c), \mathbf{z}_i \rangle\|^2. \tag{26}$$

2: `Minimization`:

$$\theta(c) := \underset{\lambda\in\mathbb{R}^m}{\operatorname{arginf}} \sum_{c(\mathbf{x}_i)=c} \|y_i - \langle \lambda, \mathbf{z}_i \rangle\|^2. \tag{27}$$

3: `Test.`

The $GLiC$ algorithm may be regarded as a generalization of the LSE method. As a final outcome, the $GLiC$ will produce a set of k vectors of linear regression coefficients specifically for k clusters. Essentially, any particular vector of linear coefficients will be the most suitable for any observation from the corresponding cluster comparing with other vectors of linear coefficients.

Algorithm 5 $StUnC$.

1: `Clustering`:

$$c(\mathbf{x}_i) := \underset{c\in\{1..k\}}{\operatorname{argmin}} \ \|\mathbf{x}_i - \Gamma \cdot \theta(c)\|^2 \tag{28}$$

where Γ is a design matrix with b rows (*brands*) and m columns (*attributes*) $\mathbf{x}_i$ is b-dimensional target variable, $\theta(c)$ is m-dimensional vector-column.

2: `Minimization`:

$$\theta(c) := \underset{\lambda\in\mathbb{R}^m}{\operatorname{arginf}} \sum_{c(\mathbf{x}_i)=c} \|\mathbf{x}_i - \Gamma \cdot \lambda\|^2. \tag{29}$$

3: `Test.`

REMARK 6.1. There is an essential difference between GLiC and StUnC Models: in the case of GLiC explanatory variables are used for the description of observations or respondents; in the case of StUnC (see Algorithm 5) explanatory variables are the same for any particular brand and vector of explanatory variables is defined by the corresponding row of the design matrix Γ. Note, that the number of brands $b \geq 1$ may be different (variable-length) for any particular observation.

REMARK 6.2. Similar to the above models we can define Support Vector Clustering (SVC) as an important example within CM framework. A standard SV algorithm was introduced in Ref.19, and SVC may be regarded as a further development. The main motivation for the SVC is to speed up the training process as a result of split of the whole dataset into several subsets assuming that statistical characteristics of the data may be characterized spatially.

7. EXPERIMENTS ON THE MSNBC DATASET

A large Web navigation **msnbc** dataset comes from the Internet Information Server **msn.com** for the entire day of *September, 28, 1999.*[2] The dataset[12] includes $n = 989818$ sequences of events with lengths ranging from 1 to 12000.

Each sequence in the dataset corresponds to page views of a user during that twenty-four hour period. Each event in the sequence corresponds to a user's request for a page. In total, there are 4698794 events.

The page categories were developed prior to investigation. There are $m = 17$ particular web categories, see Table 3. The number of pages per category ranges from 10 to 5000.

We used a first-order Markov model-based CM-algorithm, section 3, in order to produce most of the experimental results.

Current and previous codebooks were compared using the distance:

$$\Delta(\tau) = \|\Theta(\tau+1) - \Theta(\tau)\| = \frac{1}{k}\sum_{c=1}^{k}\left[\sum_{i=1}^{m}\left(\sum_{v=1}^{2}\delta_{ci}^{(v)}(\tau) + \sum_{j=1}^{m}\delta_{cij}(\tau)\right)\right]. \tag{30}$$

where $\delta_{ci}^{(v)}(\tau) = |\theta_{ci}^{(v)}(\tau+1) - \theta_{ci}^{(v)}(\tau)|$, $\delta_{cij}(\tau) = |\theta_{cij}(\tau+1) - \theta_{cij}(\tau)|$ and τ is a sequential number of iteration. Usually, it was enough to conduct less than 20 CM cycles in order to fulfill the convergence test: $\Delta(\tau) < 0.01$.

A Pentium 4, 2.8GHz, computer was used for the computations. The overall complexity of the CM cycle is $O(k+1)(2n+N))$. The computer conducted computations according to the special program written in C. The computation time for one CM cycle in the case of 49 clusters was 29 seconds.

The columns 5 and 6 of the Table 3 represent a typical structure of the start and exit probabilities for any particular cluster if number of clusters k is large enough. These probabilities clearly reflect user's preferences. As a next step, we can generate graphical model[5] using corresponding matrix of Markov probabilities. Some examples of these matrices may be seen in Fig. 2(c)-(f). Figure 2(c) includes a colorbar which may be helpful in order to make a visual assessment of the values. It may be seen that a user has the tendency to stay within a current web-category. Also, the non-diagonal elements of Fig. 2(c) essentially more uniform (and less informative) comparing with Fig. 2(d-f).

Based on the experimental results we can make a conclusion that empirical KL-divergence (5) is a more informative comparing with direct distance (30).

Figure 2(a) demonstrates 2 independent trajectories of (19) marked by T_1 and T_2. Another trajectory $T_{\exp}$ was developed in a complex with T_1 according to the EI-method of the sub-section 2.2. Regularized graph T_{reg} may be viewed in Fig. 2(a). The following parameters were used in order to compute MNL weight coefficients in (7): $\lambda = 0.5, \mu = 10$. It is interesting to note that the behavior of $T_{\exp}$ is monotonical, but may be worse comparing with $T_i, i = 1..2,$ for any k.

According to the given requirements $\alpha_i = 0.04, i = 1..2, \beta_i = 0.05, i = 1..3,$ of the Conditions $C1$-$C2$, Sect. 4, the system detected the range of $34 \le k \le 47$ for the clustering size k.

8. CONCLUDING REMARKS

We considered likelihood-based and distance-based models within general parametric, model-based framework. Experiments on real and synthetic data has confirmed fast convergence of the CM-algorithm. Unfortunately, the algorithm has a heuristic nature and can not give a guarantee of absolute optimum: it may be trapped in the local optimum depending on the initial settings.

In this regard, the EI-method as an extension of the CM framework is significant. It represents a complex of very important components: experience and innovation. Beginning with 2 clusters, we can make several independent trials ("innovations") of the CM-algorithm. The best results will give us an initial setting for the "experience" model, which can not produce worse results in the case if we will increase number of clusters. The proposed in Sect. 2.2 method is not an ideal and may be developed further. For example, as it was noted in Ref.17 criterions $A1$-$A2$ completely ignore a shape of the cluster, and it will be very important to extend method of Ref.17 to the Markov-model based clustering.

Besides, it will be important to consider a cyclic procedure: 1) create a new cluster in the most promising direction (forward move); 2) make an assessment of any particular cluster as a component within the complex of $k+1$ clusters; 3) remove the cluster which is the most insignificant (backward move); 4) compare previous and current codebooks. Go to the first step if convergence test is not fulfilled, alternatively, stop the algorithm. As a result, we will test stability of the existing clustering configuration.

ACKNOWLEDGMENTS

This work was supported by the grants of the Australian Research Council. National ICT Australia is funded through the Australian Government's *Backing Australia's Ability* initiative.

REFERENCES

1. S. Zhong and J. Ghosh, "A unified framework for model-based clustering," *Journal of Machine Learning Research* **4**, pp. 1001–1037, 2003.
2. I. Cadez, D. Heckerman, C. Meek, P. Smyth, and S. White, "Model-based clustering and visualization of navigation patterns on a web site," *Data Mining and Knowledge Discovery* **7**, pp. 399–424, 2003.
3. A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society* **39**, pp. 1–38, 1977.
4. M. Deshpande and G. Karypis, "Selective markov models for predicting web-page accesses," in *Proceedings of the First SIAM International Conference on Data Mining, Chicago, USA, April 5-7, 2001*, SIAM, 2001.
5. P. Giudici and R. Castelo, "Association models for web mining," *Data Mining and Knowledge Discovery* **5**, pp. 183–196, 2001.
6. I. Dhillon, S. Mallela, and R. Kumar, "Divisive information-theoretic feature clustering algorithm for text classification," *Journal of Machine Learning Research* **3**, pp. 1265–1287, 2003.
7. A. Banerjee, S. Merugu, I. Dhillon, and J. Ghost, "Clustering with bregman divergence," in *SDM: Proceedings of the Fourth SIAM International Conference on Data Mining, Lake Buena Vista, Florida, USA, April 22-24, 2004*, SIAM, 2004.
8. G. Hamerly and C. Elkan, "Learning the k in k-means," in *16th Conference on Neural Information Processing Systems*, 2003.
9. H. Akaike, "On the likelihood of a time series model," *The Statistician* **27**, pp. 217–235, 1978.
10. G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics* **6**(2), pp. 461–464, 1978.
11. C. Fraley and A. Raftery, "How many clusters? which clustering method? answers via model-based cluster analysis," *The Computer Journal* **41**(8), pp. 578–588, 1998.
12. Msnbc, "msnbc.com anonymous web data," in *UCI Knowledge Discovery in Databases Archive: http://kdd.ics.uci.edu/summary.data.type.html*, 1999.
13. M. Hung and D. Yang, "An efficient fuzzy c-means clustering algorithm," in *First IEEE International Conference on Data Mining*, pp. 225–232, 2001.
14. M. Wedel and W. Kamakura, *Market Segmentation Conceptual and Methodological Foundations*, Kluwer Academic Publishers, Massachusetts, USA, second ed., 2000.
15. A. Hinneburg and D. Keim, "A general approach to clustering in large databases with noise," *Knowledge and Information Systems* **5**, pp. 387–415, 2003.
16. S. Zhong and J. Ghosh, "Scalable, balanced model-based clustering," in *Proceedings of the Third SIAM International Conference on Data Mining, San Francisco, California, USA, May 1-3, 2003*, SIAM, 2003.
17. S. Savaresi, D. Boley, S. Bittanti, and G. Gazzaniga, "Cluster selection in divisive clustering algorithms," in *Proceedings of the First SIAM International Conference on Data Mining, Arlington, USA, April 11-13, 2002*, SIAM, 2002.
18. A. Rukhin, "Universal bayes estimators," *The Annals of Statistics* **6**(6), pp. 1345–1351, 1978.
19. B. Boser, I. Guyon, and V. Vapnik, "An training algorithm for optimal margin classifiers," in *Fifth COLT, Pittsburgh, USA*, pp. 144–152, 1992.