Universal Clustering with Regularization in Probabilistic Space *

Vladimir Nikulin[1] and Alex J. Smola[2]

[1] Computer Science Laboratory, Australian National University,
Canberra, ACT 0200, Australia
vladimir.nikulin@anu.edu.au
[2] NICTA, Canberra, ACT 0200, Australia
alex.smola@nicta.com.au

Abstract. We propose universal clustering in line with the concepts of universal estimation. In order to illustrate above model we introduce family of power loss functions in probabilistic space which is marginally linked to the Kullback-Leibler divergence. Above model proved to be effective in application to the synthetic data. Also, we consider large web-traffic dataset. The aim of the experiment is to explain and understand the way people interact with web sites.
The paper proposes special regularization in order to ensure consistency of the corresponding clustering model.

1 Introduction

Clustering algorithms group empirical data according to the given criteria into several clusters with relatively stable and uniform statistical characteristics.

In this paper we consider prototype-based or distance-based clustering model. The corresponding solution may be effectively approximated using k-means algorithm within *Clustering-Minimization* (CM) framework [1] which may regarded as an analog of the EM (*Expectation-Maximization*) framework for soft clustering or segmentation.

Recently, the Divisive Information-Theoretic Feature Clustering algorithm in probabilistic space $\mathcal{P}^m$ was proposed by [2]. It provides an attractive approach based on the Kullback-Leibler divergence. According to [3], the probabilistic model can be extremely useful in many applications including information retrieval and filtering, natural language processing, machine learning from text and in related areas.

As it is outlined in [4] and [5], in practice, however, an exact form of a loss function is difficult to specify. Hence, it is important to study the domination criterion simultaneously under a class of loss functions. Respectively, we introduce the family of power loss functions in probabilistic space with KL-divergence as a marginal limit.

Pollard [6] demonstrated that distance-based clustering model in $\mathbb{R}^m$ is consistent under some conditions of general nature. Further, [7] introduced definition

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

of trimmed or robustified k-means and proved consistency of the corresponding model, [8] extended result of [6] to the clustering model with Projection Pursuit which is regarded as a common technique in data analysis with such main advantage as to reduce dimensionality of the data in order to improve its visualization.

We propose definition of α-regularized KL-divergence. On the one hand, in most cases, the corresponding α-regularized clustering model may be made close to the original model with KL-divergence according to the given requirements. On the other hand, α-regularized model will be always consistent.

2 Prototype-based Approach

Suppose that $\mathbf{X} := \{x_1, \ldots, x_n\}$ is a sample of i.i.d. observations drawn from probability space $(\mathcal{X}, \mathcal{A}, \mathbb{P})$ where probability measure $\mathbb{P}$ is assumed to be unknown.

We denote by $\mathcal{Q} \in \mathcal{X}^k$ a codebook as a set of *prototypes* $q(c)$ indexed by the code $c = 1..k$ where k is a *clustering size*.

Following [6] we estimate actual distortion error

$$\Re^{(k)}[\mathcal{Q}, \Phi] := \mathbf{E}\, \Phi(x \| \mathcal{Q})$$

by the empirical error

$$\Re^{(k)}_{\text{emp}}[\mathcal{Q}, \Phi] := \frac{1}{n} \sum_{t=1}^{n} \Phi(x_t \| \mathcal{Q}) \tag{1}$$

where $\Phi(x \| \mathcal{Q}) := \Phi(x, q(c(x)))$, $\Phi(\cdot, \cdot)$ is a loss function, and

$$c(x) := \operatorname*{argmin}_{c \in \{1..k\}} \Phi(x, q(c)). \tag{2}$$

Above rule will split the given sample $\mathbf{X}$ into k empirical clusters: $\mathbf{X}_c := \{x_t : c(x_t) = c\}$, $\mathbf{X} = \cup_{c=1}^{k} \mathbf{X}_c, \mathbf{X}_i \cap \mathbf{X}_c = \emptyset, i \neq c$. Similarly, we can define set of k actual clusters $\mathcal{X}_c, c = 1..k$.

Definition 1. *We will call $\overline{\mathcal{Q}}$ as an optimal actual codebook if*

$$\Re^{(k)}[\overline{\mathcal{Q}}, \Phi] := \inf_{\mathcal{Q} \in \mathcal{X}^k} \Re^{(k)}[\mathcal{Q}, \Phi]. \tag{3}$$

We will call $\mathcal{Q}_n$ as an optimal empirical codebook if

$$\Re^{(k)}_{\text{emp}}[\mathcal{Q}_n, \Phi] := \inf_{\mathcal{Q} \in \mathcal{X}^k} \Re^{(k)}_{\text{emp}}[\mathcal{Q}, \Phi]. \tag{4}$$

Note that an outcome of the k-means algorithm is not necessarily $\mathcal{Q}_n$ as it is defined in (4).

2.1 CM Framework

The algorithm 1 represents a typical structure of an algorithm within CM-framework.

Algorithm 1. *CM*

1: `Clustering`: encode any observation x_t according to the rule (2).
2: `Minimization`: re-compute centroids specifically for any particular empirical cluster

$$q(c) := \operatorname*{arginf}_{a \in \mathcal{X}} \sum_{x_t \in \mathbf{X}_c} \Phi(x_t, a). \tag{5}$$

3: `Test:` compare previous and current codebooks $\mathcal{Q}$. Go to the step 1 if convergence test is not fulfilled, alternatively, stop the algorithm.

The following Proposition 1, which may be proved similarly to the Theorems 4 and 5 of [2], formulates the most important descending and convergence properties of the CM-algorithm.

Proposition 1. *The algorithm 1*
1) monotonically decreases the value of the objective function (1);
2) converges to the local minimum in a finite number of steps if equation (5) has unique solution.

3 Probabilistic Framework

Let $\mathcal{P}^m$ be the m-dimensional probability simplex or probabilistic space of all m-dimensional probability vectors. Following [2] we assume that the probabilities $p_{it} = P(i|x_t), \sum_{i=1}^{m} p_{it} = 1, t = 1..n$, represent relations between observations x_t and attributes or classes $i = 1..m, m \geq 2$. Accordingly, we define the clustering model $(\mathcal{P}^m, KL)$ with *Kullback-Leibler* divergence:

$$KL(\mathbf{v}, \mathbf{u}) := \sum_{i=1}^{m} v_i \cdot \log \frac{v_i}{u_i} = \langle \mathbf{v}, \log \frac{\mathbf{v}}{\mathbf{u}} \rangle, \mathbf{v}, \mathbf{u} \in \mathcal{P}^m. \tag{6}$$

The following notations will be used below $p(x_t) = \{p_{1t}, \cdots, p_{mt}\}, q(c) = \{q_{1c}, \cdots, q_{mc}\}$.

3.1 Power Loss Functions in Probabilistic Space

Let us consider 2 families of loss functions

$$L\Phi_\gamma(\mathbf{v}, \mathbf{u}) := \sum_{i=1}^{m} v_i^{1+\gamma} u_i^{-\gamma} - 1, \quad 0 < \gamma < \infty; \tag{7}$$

$$R\Phi_\gamma(\mathbf{v}, \mathbf{u}) := 1 - \sum_{i=1}^{m} v_i^{1-\gamma} u_i^{\gamma}, \quad 0 < \gamma < 1. \tag{8}$$

Proposition 2. *The loss functions (7) and (8) are non-negative and equal to* 0 *if and only if* $\mathbf{u} = \mathbf{v}$.

Above statement may be proved using the method of mathematical induction.

Proposition 3. *Suppose that* $\mathbf{v} \neq \mathbf{u}$, $\min\{v_i\} > 0$ *and* $\min\{u_i\} > 0$. *Then, the loss function (7) is convex and strictly increasing as a function of* γ.

Proof. The required result follows from the structure of the corresponding first

$$\frac{\partial L\Phi_\gamma(\mathbf{v},\mathbf{u})}{\partial\gamma} = \sum_{i=1}^{m} v_i \log\left(\frac{v_i}{u_i}\right)\left(\frac{v_i}{u_i}\right)^\gamma \tag{9}$$

and second derivatives where the first derivative is strictly positive for $\gamma = 0$ and is strictly increasing for all $\gamma > 0$. □

Proposition 4. *Suppose that* $\mathbf{v} \neq \mathbf{u}$, $\min\{v_i\} > 0$ *and* $\min\{u_i\} > 0$. *Then, the loss function (8) is concave and strictly increasing locally as a function of* γ *at the point of origin* 0*:*

$$\exists\varepsilon > 0 : R\Phi_\alpha(\mathbf{v},\mathbf{u}) < R\Phi_\gamma(\mathbf{v},\mathbf{u}) \quad \forall\alpha,\gamma : 0 \leq \alpha < \gamma \leq \varepsilon.$$

Proof. The required result follows from the structure of the corresponding derivative

$$\frac{\partial R\Phi_\gamma(\mathbf{v},\mathbf{u})}{\partial\gamma} = -\sum_{i=1}^{m} v_i \log\left(\frac{u_i}{v_i}\right)\left(\frac{u_i}{v_i}\right)^\gamma \tag{10}$$

and

$$\frac{\partial^2 R\Phi_\gamma(\mathbf{v},\mathbf{u})}{\partial\gamma^2} = -\sum v_i \log^2\left(\frac{u_i}{v_i}\right)\left(\frac{u_i}{v_i}\right)^\gamma < 0 \tag{11}$$

where the first derivative is strictly positive for $\gamma = 0$ and is strictly decreasing for all $0 < \gamma \leq 1$. Respectively, $\exists\varepsilon > 0$ so that the first derivative is strictly positive for $0 < \gamma \leq \varepsilon$ as a continuous function of γ. □

We can compute centroids for the loss functions (7) and (8) in analytical form similar to (12). For example, the following formula represents centroids for (7)

$$q_i(c) \propto \sqrt[1+\gamma]{A_{ic}(\gamma)}, \quad 0 \leq \gamma < \infty, \tag{12}$$

where $A_{ic}(\gamma) = \sum_{x_t \in \mathbf{X}_c} p_{it}^{1+\gamma}$.

Using result of the Proposition 2 we can define a new family of loss functions as an average of (7) and (8)

$$\Phi_\gamma(\mathbf{v},\mathbf{u}) := \frac{1}{2}\left(L\Phi_\gamma(\mathbf{v},\mathbf{u}) + R\Phi_\gamma(\mathbf{v},\mathbf{u})\right), 0 < \gamma < 1. \tag{13}$$

The following result demonstrates that the KL-divergence may be regarded as a marginal limit in relation to the family of loss functions (13).

Proposition 5. *The family of power loss functions (13) is marginally linked to the* KL*-divergence:* $\lim_{\gamma\to 0} \frac{\Phi_\gamma(\mathbf{v},\mathbf{u})}{\gamma} = KL(\mathbf{v},\mathbf{u})$.

Proof. The statement of the proposition follows from the structure of the derivative:

$$\frac{\partial \Phi_\gamma(\mathbf{v},\mathbf{u})}{\partial \gamma} = \frac{1}{2}\sum_{i=1}^{m} v_i \log\frac{v_i}{u_i}\left[\left(\frac{v_i}{u_i}\right)^{\gamma} + \left(\frac{u_i}{v_i}\right)^{\gamma}\right]. \tag{14}$$

In the case if $\gamma = 0$ the right part of (14) equals to the KL-divergence. □

Proposition 6. *Suppose that* $\mathbf{v} \neq \mathbf{u}$, $\min\{v_i\} > 0$ *and* $\min\{u_i\} > 0$. *Then, the loss function* Φ_γ *defined in (13) is strictly increasing locally as a function of* γ *at the point* 0 $\exists \varepsilon > 0 : \Phi_\alpha(\mathbf{v},\mathbf{u}) < \Phi_\gamma(\mathbf{v},\mathbf{u}) \;\; \forall \alpha,\gamma : 0 \leq \alpha < \gamma \leq \varepsilon$.

Proof follows from above Propositions 3 and 4.

Remark 1. The results of the Propositions 4 and 6 may not necessarily take place for $\varepsilon = 1$, because $KL(\mathbf{u},\mathbf{v}) \to \infty$ if $v_1 \to 0$ and $\min\{u_i\} \geq \delta > 0$. As a consequence, the derivative (9) is limited. At the same time derivative (10) tends to $-\infty$ if $\gamma \to 1$ (see Figure 1(d)).

Minimizing $\sum_{x_t \in \mathbf{X}_c} \Phi_\gamma(p(x_t), q) = \sum_{i=1}^{m} \left(A_{ic}(\gamma) q_i^{-\gamma} - A_{ic}(-\gamma) q_i^{\gamma}\right)$ as a function of $q \in \mathcal{P}^m$ we will formulate iterative algorithm for the computation of centroids in the sense of the loss function (13) with fixed value of the parameter $\gamma > 0$

$$q_i(c, j+1) \propto \sqrt[1+\gamma]{A_{ic}(\gamma) + A_{ic}(-\gamma) q_i^{2\gamma}(c,j)} \tag{15}$$

where j is a sequential number of iteration, initial values of $q(c,1)$ may be computed using (12).

Remark 2. According to [5], it seems rather natural to investigate the situation where the estimator is the same for every loss from a certain set of loss functions under consideration. In line with Propositions 3, 4 and 6 we can use parameter γ in order to increase differentiation between observations. Comparing clustering results for different input parameters γ we can make assessment of the stability of clustering: the smaller fluctuation of the centroids will indicate the higher quality of clustering (see Figure 1).

3.2 Consistency of the Clustering Model

According to [9], p. 33, it is extremely important to use concepts that describe necessary and sufficient conditions for consistency. This guarantees that the constructed theory is general and cannot be improved from the conceptual point of view.

Definition 2. *We say [9] that the clustering model* $(\mathcal{X}, \Phi)$ *is consistent if*

$$\Re^{(k)}_{\text{emp}}[\mathcal{Q}_n, \Phi] \underset{n\to\infty}{\Rightarrow} \Re^{(k)}[\overline{\mathcal{Q}}, \Phi] \;\; a.s. \tag{16}$$

We say [6] that the clustering model $(\mathcal{X}, \Phi)$ *is* ν*-strongly consistent if*

$$\nu(\mathcal{Q}_n, \overline{\mathcal{Q}}) \underset{n\to\infty}{\longrightarrow} 0 \;\; a.s. \tag{17}$$

where ν *is a distance in* $\mathcal{X}^k$.

Definition 3. *We will call the element* $\mathbf{v} \in \mathcal{P}^m$ *as 1) an* ***uniform vector*** *if* $v_i = \frac{1}{m}, i = 1..m$*; and 2) as* ***i-margin*** *if* $v_i = 0$.

Definition 4. *We will call* $KL_\alpha(\mathbf{v}, \mathbf{u}) := KL(\mathbf{v}_\alpha, \mathbf{u}_\alpha)$ *as* α*-regularized* KL*-divergence where* $\mathbf{v}_\alpha = \alpha\mathbf{v} + (1-\alpha)\mathbf{v}_0$ *and* $\mathbf{u}_\alpha = \alpha\mathbf{u} + (1-\alpha)\mathbf{v}_0$, $\mathbf{v}_0$ *is an uniform vector and* $0 < \alpha \leq 1$.

The following result represents an essential generalization of the Lemma 3 [2].

Proposition 7. *Centroids* $q(c)$ *in* $(\mathcal{P}^m, KL_\alpha)$ *are not dependent on* $0 < \alpha \leq 1$ *and must be computed using* k*-means (12).*

Corollary 1. *Suppose that* $q(c) \in \mathcal{Q}_n$ *and* $q_i(c) = 0$. *Then,* $P(i|x_t) = 0 \forall x_t \in \mathbf{X}_c$. *Suppose that* $q(c) \in \overline{\mathcal{Q}}$ *and* $q_i(c) = 0$. *Then,* $P(i|x) = 0 \forall x \in \mathcal{X}_c$ *a.s.*

Theorem 1 *Suppose that the clustering size* k *and parameter* $0 < \alpha < 1$ *are fixed. Then, the model* $(\mathcal{P}^m, KL_\alpha)$ *is consistent.*

Proof. The required result
$\Re^{(k)}_{\text{emp}}[\mathcal{Q}^{(\alpha)}_n, KL_\alpha] \underset{n\to\infty}{\Rightarrow} \Re^{(k)}[\overline{\mathcal{Q}}^{(\alpha)}, KL_\alpha]$ *a.s.*
follows from uniform continuity of the $KL_\alpha(\mathbf{v}, \mathbf{u})$ as a function of both arguments if $0 < \alpha < 1$ where $\mathcal{Q}^{(\alpha)}_n$ and $\overline{\mathcal{Q}}^{(\alpha)}$ are optimal empirical and actual codebooks which correspond to KL_α. □

Corollary 2. *Suppose that the optimal actual codebook* $\overline{\mathcal{Q}}^{(\alpha)}$ *is unique. Then, the model* $(\mathcal{P}^m, KL_\alpha)$ *is* ν*-strongly consistent where a distance* ν *may be defined as* $\max_{c=1}^k \min_{j=1}^k KL(q_n(c), \overline{q}(j))$ *where* $q_n(c) \in \mathcal{Q}^{(\alpha)}_n$ *and* $\overline{q}(j) \in \overline{\mathcal{Q}}^{(\alpha)}$.

3.3 Extension to the Euclidean Space

Monograph [10], pp. 255-258, discusses characterization of families of distributions for which the Pitman estimator of the location parameter in $\mathbb{R}$ does not depend on the loss function. Generally speaking, for the same distribution function F, the Pitman estimator differs from loss function to loss function. However, if F is a normal distribution function, then it is easy to see that, for quadratic trigonometrical and following below exponential loss functions (18), the Pitman estimator is one and the same, namely, sample mean.

The G-means algorithm [11] which is based on the *Gaussian* fit of the data within particular cluster is relevant here. The G-means algorithm is based on a statistical test for the hypothesis that a subset of data follows a Gaussian distribution. G-means runs k-means with increasing k in a hierarchical fashion until the test accepts the hypothesis that the data assigned to each centroid are Gaussian.

Similar to the Sect. 3.1 we can define model of universal clustering in $\mathbb{R}^m$ with the following family of exponential loss functions: $\Phi_\gamma(\mathbf{v}, \mathbf{u}) := \varphi_\gamma(\mathbf{v} - \mathbf{u})$ where $\mathbf{v}, \mathbf{u} \in \mathbb{R}^m$, and $\gamma \in \mathbb{R}^m_+$ is m-dimensional regulation parameter,

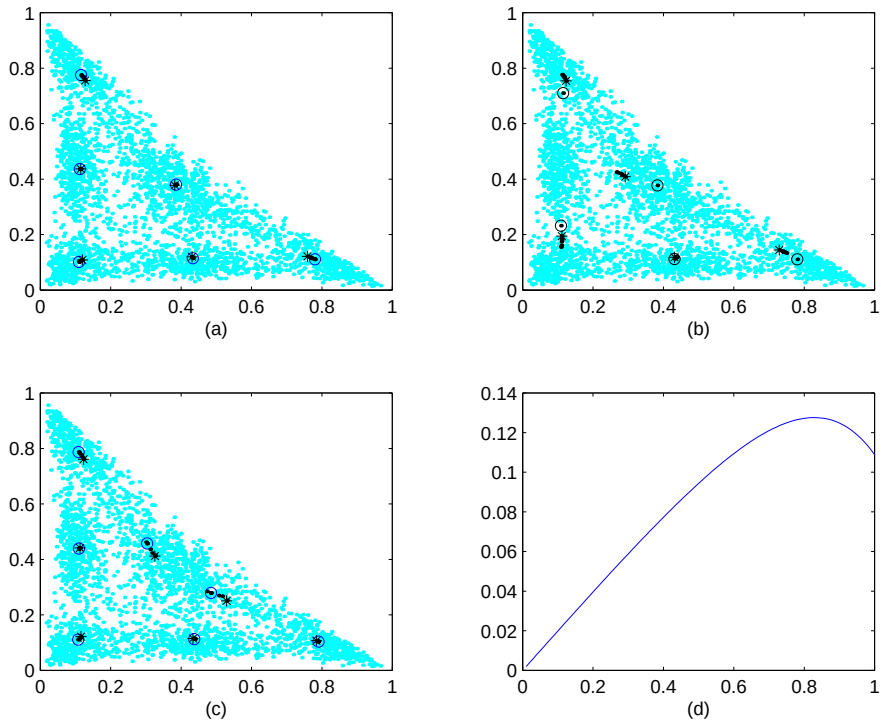

Fig. 1. 3D-synthetic data, $n = 3000$ with 6 clusters, (a): **k=6**: random selection of the cluster seeds; centroids were re-computed using loss function (13) with $\gamma = 0.09 + 0.13 \cdot (i-1), i = 1..8$; symbol $\odot$ marks centroids which corresponds to $\gamma = 0.09$; $*$ marks centroids which corresponds to $\gamma = 1.0$, other centroids are marked by bold black dots $\cdot$; (b): **k=5**; (c): **k=7**; (d): loss (13) as a function of γ where $m = 10$, $u_i = \frac{1}{m}, i = 1..m$, $v_1 = \varepsilon$, $v_i = \frac{1-\varepsilon}{m-1}, i = 2..m, \varepsilon = 0.001$

$$\varphi_\gamma(\mathbf{v}) := \sum_{i=1}^{m} \cosh(\gamma_i \cdot v_i) - m, \tag{18}$$

and corresponding centroids:

$$q_i^{(\gamma)}(c) = \frac{1}{2\gamma_i} \log \frac{\sum_{x_t \in \mathbf{X}_c} e^{\gamma_i x_{ti}}}{\sum_{x_t \in \mathbf{X}_c} e^{-\gamma_i x_{ti}}}$$

which represent a unique k-means solution for the loss function (18).

4 Experiments

The sample of the $3D$-probability data, which is displayed in the Figures 1 was generated using the following procedure.

Table 1. Simulation coefficients for the 3D-synthetic data, see Figure 1.

Cluster	Coefficients				Probabilities
c	b_1	b_2	b_3	e	p
1	1	-1	-1	0.5	0.15
2	-1	1	-1	0.5	0.15
3	-1	-1	1	0.5	0.15
4	-0.4	-0.4	-0.8	0.4	0.25
5	-0.4	-1.9	-0.4	0.3	0.15
6	-1.9	-0.4	-0.4	0.3	0.15

Firstly, the cluster code c was drawn randomly according to the probabilities p, see Table 1, using standard uniform random variable. Secondly, we used the multinomial logit model in order to generate coordinates of the $3D$-probability data: $v_i \propto \exp\{b_{ci} + e_c r\}, \sum_{i=1}^{3} v_i = 1$, where r is a standard normal random variable.

By definition, the family of power loss functions (13) is marginally linked to the KL-divergence if $\gamma \to 0$. By the increase of γ we will increase the power of diversification. Respectively, any centroid, which corresponds to a non significant empirical cluster will move around. Figure 1 illustrates that centroids of the "strong" empirical clusters are stable as a consequence of correct selection of the number of clusters $k = 6$.

Algorithm 2. (Universal Clustering)

1: Order number of clusters k, and select randomly initial codebook with k probability vectors which will be used for all $\tau \geq 2$ runs of the CM algorithm in the next step.
2: Run CM-algorithm using loss function (13) with $\gamma = \gamma_0 + (j-1)\cdot\delta, j = 1..\tau$, where $0 < \gamma_0 < 1$ and $0 < \delta \leq \frac{1-\gamma_0}{\tau-1}$. As an outcome we obtain a set of $k \cdot \tau$ probability vectors $\{\tilde{q}(j,c), j = 1..\tau, c = 1..k\}$.
3: Compute maximum distance between first and other codebooks

$$D := C \cdot \max_{c=1..k} \max_{j=2..\tau} KL(\tilde{q}(1,c), \tilde{q}(j,c)) \tag{19}$$

where $C > 0$ is a constant.

The second experiment was conducted using a large Web navigation **msnbc** dataset. This dataset comes from Internet Information Server **msn.com** for the entire day of *September, 28, 1999* [12]. The dataset [13] includes $n = 989818$ sequences of events with lengths ranging from 1 to 12000.

Table 2. 3D-probabilistic synthetic data: determination of the clustering size k where D is defined in (19), used parameters: $\gamma_0 = 0.002, \delta = 0.01, \tau = 20, C = 1000$.

k:	3	4	5	6	7	8	9
D:	0.6478	0.0263	0.0045	**0.0011**	0.8535	0.9264	2.7150
k:	10	11	12	13	14	15	16
D:	0.8041	1.9056	0.1474	0.3063	0.9377	5.0651	12.1121

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. The number of pages per category ranges from 10 to 5000.

Analysis of the **msnbc** data had revealed the following general properties: 1) users have tendency to stay within particular category; 2) transitions from one category to another are relatively rare.

Respectively, we considered an ultimate simplification of the model by ignoring 1) dependencies between subsequent events and 2) length of the sequence of events for any particular user. As a result, we reduced the given variable-length data to the fixed length data where any user is represented by the m-dimensional probability vector of the frequencies of m categories.

The aim of this experiment is to explain and understand the way people interact with web sites, explore human behavior within internet environment. Briefly, we observed that the table of centroids in the case of $k = 8$ demonstrates clearly user's preferences. Detailed numerical and graphical illustrations may be found in [1].

Also, the paper [1] introduced clustering regularisation based on the balanced complex of two conditions: 1) significance of any particular cluster; 2) difference between any 2 clusters. Subject to some input regulation parameters the corresponding system detected the interval $34 \leq k \leq 47$ as the most likely range for the number of significant clusters in **msnbc**. Another solution for the same task may be found using principles of universal clustering.

A Pentium 4, 2.8GHz, 512MB RAM, computer was used for the computations. The overall complexity of a CM cycle is $O(k \cdot n \cdot m)$. The computer conducted computations according to the special program written in C. The computation time for one CM cycle in the case of 51 clusters was 110 seconds.

5 Concluding Remarks

Experiments on the real and synthetic data had confirmed fast convergence of the CM-algorithm [1]. Unfortunately, the final results of the CM-algorithm depend essentially on initial settings, because the algorithm may be trapped in

local minimum. In this regard, the proposed in the Section 3.2 α-regularization is significant because it will guarantee consistency of the corresponding clustering model. On the other hand, the proposed in the paper universal clustering represents a promising direction. We can make an assessment of quality of clustering using set of codebooks as a function of regulation parameter. The quality function may be computed as a decreasing function of the fluctuation of codebooks.

Acknowledgments. We are grateful to Peter Hall for the consideration and very valuable support. Our thanks go also to anonymous referees for the helpful comments and suggestions.

References

[1] Nikulin, V., Smola, A.: Parametric model-based clustering. In Dasarathy, B., ed.: Data Mining, Intrusion Detection, Information Assurance, and Data Network Security, 28-29 March 2005, Orlando, Florida, USA. Volume 5812., SPIE (2005) 190–201

[2] Dhillon, I., Mallela, S., Kumar, R.: Divisive information-theoretic feature clustering algorithm for text classification. Journal of Machine Learning Research **3** (2003) 1265–1287

[3] Cohn, D., Hofmann, T.: The missing link - a probabilistic model of document content and hypertext connectivity. In: 13th Conference on Neural Information Processing Systems. (2001)

[4] Hwang, J.T.: Universal domination and stochastic domination: Estimation simultaneously under a broad class of loss functions. The Annals of Statistics **13** (1985) 295–314

[5] Rukhin, A.: Universal Bayes estimators. The Annals of Statistics **6** (1978) 1345–1351

[6] Pollard, D.: Strong consistency of k-means clustering. The Annals of Statistics **10** (1981) 135–140

[7] Cuesta-Albertos, J., Gordaliza, A., Matran, C.: Trimmed k-means: an attempt to robustify quantizers. The Annals of Statistics **25** (1997) 553–576

[8] Stute, W., Zhu, L.: Asymptotics of k-means clustering based on projection pursuit. Sankhya **57** (1995) 462–471

[9] Vapnik, V.: The Nature of Statistical Learning Theory. Springer (1995)

[10] Kagan, A., Linnik, Y., Rao, C.: Characterization Problems in Mathematical Statistics. John Wiley Sons (1973)

[11] Hamerly, G., Elkan, C.: Learning the k in k-means. In: 16th Conference on Neural Information Processing Systems. (2003)

[12] Cadez, I., Heckerman, D., Meek, C., Smyth, P., White, S.: Model-based clustering and visualization of navigation patterns on a web site. Data Mining and Knowledge Discovery **7** (2003) 399–424

[13] Msnbc: msnbc.com anonymous web data. In: UCI Knowledge Discovery in Databases Archive: http://kdd.ics.uci.edu/summary.data.type.html. (1999)