CLUSTERING OF SUN EXPOSURE MEASUREMENTS

A. Szymkowiak-Have[1], J. Larsen[1], L.K. Hansen[1],
P.A. Philipsen[2], E. Thieden[2], H.C. Wulf[2]

[1]Informatics and Mathematical Modelling, Building 321
Technical University of Denmark, DK-2800 Lyngby, Denmark
Phone: +45 4525 3900,3923,3889 Fax: +45 4587 2599
E-mail: asz,jl,lkh@imm.dtu.dk Web: isp.imm.dtu.dk

[2]Department of Dermatology, Bispebjerg Hospital
University of Copenhagen, Bispebjerg Bakke 23
DK-2400 Copenhagen, Denmark

Abstract. In a medically motivated sun-exposure study, questionnaires concerning sun-habits were collected from a number of subjects together with UV radiation measurements. This paper focuses on identifying clusters in the heterogeneous set of data for the purpose of understanding possible relations between sun-habits exposure and eventually assessing the risk of skin cancer. A general probabilistic framework originally developed for text and web mining is demonstrated to be useful for clustering of behavioral data. The framework combines principal component subspace projection with probabilistic clustering based on the generalizable Gaussian mixture model.

INTRODUCTION

In the studied sun-exposure experiment, questionnaires concerning sun-habits were collected from 187 subjects. In addition, daily UV radiation were measured at a 10 minute sampling rate using a specially designed “sun-watch” worn by the subjects. The ultimate objective is to relate the heterogeneous data of sun-habits, UV dose and other data (e.g., medical records) with the purpose of assessing the risk of skin cancer for individual subjects. This paper focuses on the sub-task of identifying relevant structure in the combined data set of categorical sun habit diaries and real valued daily UV dose measurements. We aim at identifying relevant structure using hierarchical probabilistic clustering, which allows for interpretation of various features representations, e.g., the role of sequence information. Although the method presented in [7] can be invoked for hierarchical clustering, we resort to simple probabilistic clustering in this work. The diary records can be viewed as a vector of categorical data, whereas the daily UV dose is a continuous measurement which is measured for different persons during 138 days. The long-term theoretical aim is to identify a hierarchical probabilistic clustering model which efficiently handles combinations of categorical and continuous data. However, the idea of the present paper is to study the capabilities of our flexible multimedia text and images data mining framework [4, 5, 6, 7, 9] for analysis and understanding of behavioral data.

MODELING FRAMEWORK

Suppose that we have M different feature modalities which are represented by feature (column) vectors, $\boldsymbol{z}_m = [z_{m1}, \cdots, z_{mL_m}]$, $m \in [1; M]$ with dimensions L_m, and let $\boldsymbol{z} = [\boldsymbol{z}_1, \cdots, \boldsymbol{z}_M]$ be the complete data vector. Behavioral features will be represented by a number of feature modalities and the UV dose measurement by yet another. Probabilistic clustering aims at modeling the probability density of the complete data vector, $p(\boldsymbol{z})$, with the purpose of identifying meaningful cluster structure.

Preprocessing

Conditioning. Let $\mathcal{B}_m = \{1, 2, \cdots, L_m\}$ be the L_m dimensional event space for feature modality m. In this work, z_{mi}, $i \in \mathcal{B}_m$ will represent the likelihood that event i occurred, as given by $\boldsymbol{z}_m = \widetilde{\boldsymbol{z}}_m / \|\widetilde{\boldsymbol{z}}_m\|_2$ where $\widetilde{z}_{mi}$ is the number of occurrences of event i, i.e., a histogram. This representation is motivated by previous work on multimedia mining [4, 5, 7, 9] as it provides insensitivity to the length of the record, and further, unit length normalization $\|\cdot\|_2$ provides a more uniform spherical distribution of features than, e.g., frequency normalization $\sum_i z_{mi} = 1$.

Consider a training data set $\mathcal{D} = \{\boldsymbol{z}(n)\}_{n=1}^{N_{\mathrm{tr}}}$ of N_{tr} examples. In order not to eliminate an arbitrary scaling between feature modalities the feature vectors are further studentized as $\boldsymbol{x}_m(n) = (\boldsymbol{z}_m(n) - \boldsymbol{u}_m)/s_m$, where $\boldsymbol{u}_m = N_{\mathrm{tr}}^{-1} \sum_{n \in \mathcal{D}} \boldsymbol{z}_m(n)$, $s_m^2 = (L_m)^{-1} \sum_i (N_{\mathrm{tr}} - 1)^{-1} \sum_{n \in \mathcal{D}} (z_{mi}(n) - \mu_{mi}))^2$.

Subspace Projection. In order to achieve models with high generalization ability we pursue the idea of Latent Semantic Indexing (LSI) [2], which was developed and successfully applied for text and multimedia mining, see e.g., [4, 5]. For each modality the feature space is projected onto a latent eigenspace. Define the $N_{\mathrm{tr}} \times L_m$ training data matrix $\boldsymbol{X}_m = [\boldsymbol{x}_m^\top(1); \cdots; \boldsymbol{x}_m^\top(N_{\mathrm{tr}})]$ and perform a singular value decomposition $\boldsymbol{X}_m = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top$. Assuming $N_{\mathrm{tr}} > L_m$ then $\boldsymbol{U}$ is the $N_{\mathrm{tr}} \times L_m$ matrix of left eigenvectors as columns, $\boldsymbol{S}$ is the $L_m \times L_m$ matrix of singular values arranged in decreasing order, and $\boldsymbol{V}$ is the $L_m \times L_m$ matrix of right eigenvectors as columns. The selection of subspace dimensions, $d_m < L_m$ is done separately for each modality using the probabilistic PCA formulation described in [3] in which there is d_m dimensional signal space with full covariance structure and a $L_m - d_m$ dimensional noise space with diagonal covariance structure $\sigma^2_{\mathrm{noise,m}} \cdot \boldsymbol{I}$. The subspace dimension is selected to minimize the generalization error associated with the probabilistic PCA model. The feature vector is thus projected onto the d_m dimensional signal subspace $\boldsymbol{y} = \boldsymbol{V}_{\mathrm{sig}}^{\top} \boldsymbol{x}$ and the noise subspace is $\boldsymbol{\nu} = \boldsymbol{V}_{\mathrm{noise}} \boldsymbol{x}$ where $\boldsymbol{V} = [\boldsymbol{V}_{\mathrm{sig}}, \boldsymbol{V}_{\mathrm{noise}}]$ and $\boldsymbol{V}_{\mathrm{sig}}$ corresponds to the d_m largest singular values.

Window size selection

The analysis of sun exposure data is performed during a period of 138 days by using an analysis window of a number of days, and the optimal size of the window is an important issue which needs to be addressed. For example, taking the full set of records belonging to a given person will produce a set of points in the space that will not form any particular clusters, since each of them will contain most of the observed patterns. On the other hand, taking one diary record at the time will significantly increase the computational complexity and preclude the possibility of analyzing time effects. We invoke the generalization scheme is used to select the optimum window size, see [3].

For all modality histogram feature vectors $\boldsymbol{x}_m^w$ calculated over the different window sizes w, the common feature spaces L_m are selected by removing low occurring terms, which enables the comparison of unsupervised generalization errors. For each window size, the modalities are projected onto optimal d_m dimensional signal and $L_m - d_m$ dimensional noise-subspace using the method described in the previous paragraph. The resulting density for $\boldsymbol{x}$ is then given by

$$p(\boldsymbol{x}) = p(\boldsymbol{y}) \cdot \prod_{m=1}^{M} p(\boldsymbol{\nu}_m) \tag{1}$$

where $p(\boldsymbol{y})$ is the density of the Guassian mixture model described below, and $p(\boldsymbol{\nu}_m) \equiv \mathcal{N}(\boldsymbol{0}, \sigma^2_{\mathrm{noise},m} \cdot \boldsymbol{I})$. The window size is selected by minimizing the generalization error estimated from a test set of N_{test} samples, that is, $G = -N_{\mathrm{test}}^{-1} \sum_{n=1}^{N_{\mathrm{test}}} \log p(\boldsymbol{x}(n))$.

Unsupervised Gaussian Mixture Model

In order to pursue probabilistic clustering we deploy the Gaussian mixture model is the joint signal subspace $\boldsymbol{y} = [\boldsymbol{y}_1, \cdots, \boldsymbol{y}_M]$ defined in [4, 6, 8].

$$p(\boldsymbol{y}) = \sum_{k=1}^{K} p(\boldsymbol{y}|k) \cdot p(k) \tag{2}$$

where $p(\boldsymbol{y}|k) \equiv \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ are Gaussian densities, and $p(k)$ are nonnegative mixture proportions with $\sum_k p(k) = 1$. The parameters $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are estimated from the training data set $\mathcal{D} = \{\boldsymbol{y}(n)\}_{n=1}^{N_{\mathrm{tr}}}$ by minimizing negative log-likelihood cost function $\mathcal{L} = -N_{\text{tr}}^{-1} \sum_n \log(p(\boldsymbol{y}(n)))$ through a modified expectation-maximization method. In order to ensure generalizability $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are estimated from the disjoint sets of observations and the optimal number of mixture components is found by the AIC-criterion [1]. The complete generalizable Gaussian mixture algorithnm (GGM) is described in [4, 7].

Model Interpretation

In order to find key-features corresponding to each of the clusters, centers $\boldsymbol{\mu}_k$ are back-projected to the original space of normalized feature histograms[1]. In the case when projection does not have positivity constraint the resulting normalized histogram vector contains negative values which can be believed to be of no importance, hence are set to zero. The key-features are then selected as the features which explain the majority cumulative feature distribution mass.

The used framework makes it possible to describe the behavior of every new person in the experiment by using both cluster assignment and associated key-features. The confidence of assigning the person *Per* into the given cluster k can be expressed by the posterior probability:

$$p(k|Per) = \frac{1}{N_{Per}} \sum_n p(k|\boldsymbol{y}(n)) \cdot p(\boldsymbol{y}(n)), \tag{3}$$

$$p(k|\boldsymbol{y}(n)) = \frac{p(k|\boldsymbol{y}(n))P(k)}{p(\boldsymbol{y}(n))} \tag{4}$$

where $\boldsymbol{y}(n)$ is a feature vector for combined modalities of the size d, $n = 1, 2, \ldots, N_{Per}$, where N_{Per} is the number of samples for person Per.

SUN EXPOSURE STUDY

A specially designed device, measuring received sun radiation (*PID*), was worn by the group of subjects. In addition, subjects were requested to fill out a diary concerning their sun behaviors during each day of the study (for more details, see [10]). Eight selected questions are presented here:

Variable	**Values**
1. Holiday	yes/no
2. Abroad	yes/no
3. Sun Bathing	yes/yes-solarium/no
4. Naked Shoulders	yes/no
5. On the Beach/Water	yes/no
6. Sun Factor Number	no/26 values in range 1-60
7. Sunburned	no/red/hurts/blisters
8. Size of Sunburned Area	no/little/medium/large

[1] Another way would be to project the most probable feature vectors from each of the clusters identified, e.g. by Monte Carlo sampling.

Thus, two types of data were collected: continuous measurements of the sun UV radiation (*PID*) and categorical diary records. Each diary record is represented by an 8 dimensional vector and describes a specific behavior of the particular person during the particular day. The total number of possible events for the presented set of questions equals 20736, however, only a small fraction of 423 events actually exist in the investigated data set.

CLUSTERING OF THE SUN EXPOSURE DATA

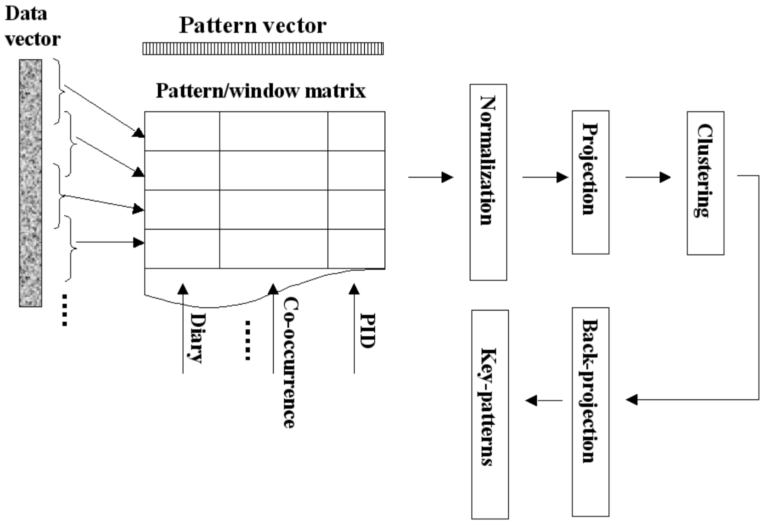

Figure 1: Framework for clustering: 1) the data is windowed into several histogram vectors and together with the co-occurrence matrix and the *PID* (UV dose) histogram forms a pattern/window matrix. 2) data is then normalized and projected onto the orthogonal d_m dimensional signal spaces. 3) the Gaussian mixture algorithm is used to cluster the data. 4) In order to interpret the results, cluster centers are back-projected to the original space where key-features are identified.

Figure 1 presents the general framework of preprocessing, clustering and data postprocessing. In the first step, data is windowed creating vectors that contain data from consecutive days.

In the study we define three different modalities. Diary and sun exposure histograms are defined as $z_m = \sum_{l=1}^{w} \delta(z_{m,l} - i)$, $i \in \mathcal{B}$. The sun exposure measurements were quantized into 100 bins before computing the histogram. Further, we used the co-occurrence histogram of diary records, $z_{m,\tau} = \sum_{l=1}^{w-\tau} \delta(z_{m,l} - i) \cdot \delta(z_{m,l+\tau} - j)$, where τ is an lag and $i, j \in \mathcal{B}$ are co-occurring events. We merely used lag $\tau = 1$ co-occurrence feature. There are 20736^2 possible co-occurrences and again only small fraction is present in the actual data set.

All modalities are screened against rare patterns by removing those which have occurrence below a certain threshold.

RESULTS

The set of 19171 diary records and corresponding *PID* values were selected for the clustering experiments. Data are complete i.e., there is no missing records or *PID* values. The missing record problem for the current data set was partly addressed in [10]. The sun behaviors of 187 subjects during summer period were collected. Of this 10 persons were hold out for evaluation.

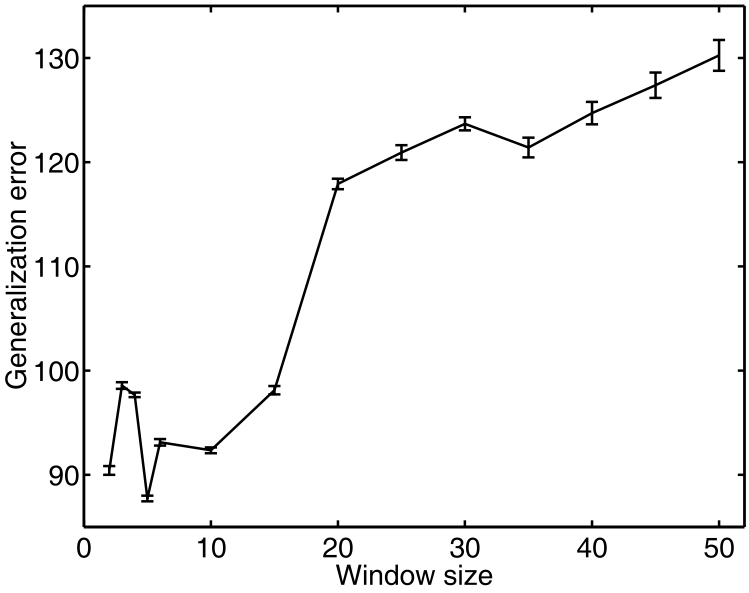

Figure 2: The generalization error for different window sizes. The mean curve is an average over 50 experiments with the cross-validation 80% for the training set and 20% for the test set. Errorbars show the deviation from the mean curve.

Figure 2 presents the generalization error computed for different window sizes. The curve is averaged over 50 experiments in which training was performed on the random chosen 80% of the data set and the generalization error was estimated on the remaining part. Window of the size 5 gives the lowest generalization error, however, any choice in the range 5–10 also will give reasonable results.

For the window of size 5 the training and the test data set contains 1346 and 337 samples, respectively. Each sample consist of the diary histogram, the co-occurrence matrix and the *PID* histogram. The diary histogram is reduced from 423 to the 25 features that contain 90% of the total mass. In a similar way, the *PID* histograms are reduces from 100 to 75 features and the co-occurrence matrix is reduced from 1451 to the 36 most often occurring pairs of patterns[2].

Figure 3 shows the generalization error for individual modalities calculated with respect to different number of principal components. The mean curve is an average over 15 experiments and 80% of the set was used for the training set and 20% for the test set. Errorbars show the deviation from the mean curve. The minimum

[2] In the case of the co-occurrence histogram, the threshold for removing patterns is 80% of the mass.

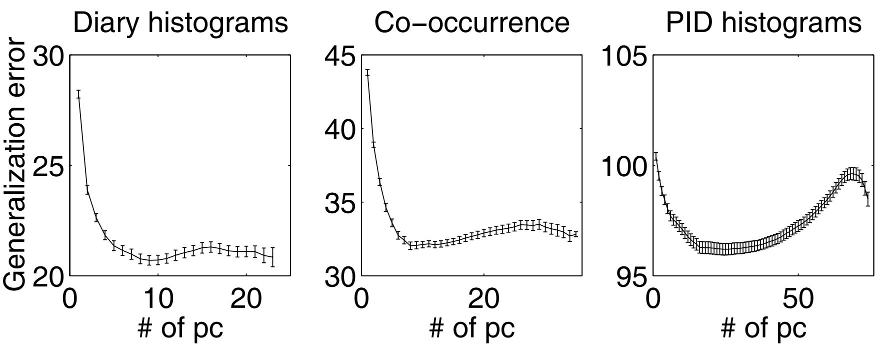

Figure 3: Subspace selection. The generalization error for the individual modalities calculated for different number of principal components. The mean curve is an average over 15 experiments with the cross-validation 80% for the training set and 20% for the test set. Errorbars show the deviation from the mean curve. The minimum of each of the curve shows optimum size of the signal subspace.

of each of the curves show optimum size for the signal subspace. For the diary histograms 9 principal components is used, for the co-occurrence 8 and for the *PID* histograms 25 principal components are used.

In the experiments the hard assignment GGM model [4] is used, i.e., the parameters of the clusters $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are estimated from the set of samples assigned to each of the clusters. In order to achieve a more detailed cluster structure one could use soft GGM [7, 9].

#.	Key-Feature	Probability	Description
1.	11000000,10000000-10010000, 10010000,11000000-11000000	0.08,0.07, 0.06,0.04	holiday
2.	10011000,10110000, 10111000,$< 05.75 >_{PID}$,	0.07,0,04, 0.03,0.02	holiday,sun bathing, naked shoulders,high sun exposure
3.	10001000,$< 0.19 >_{PID}$, 10010000,10111000	0.11,0.08 0.07,0.06	holiday on the beach, low sun radiation
4.	10010000,$< 0.19 >_{PID}$, 00010000, 10000000-10010000	0.09,0.08 0.04,0.04	holiday and working, naked shoulders, low sun radiation
5.	10000000-10010000,10010000 10010000-10000000,$< 6.8 >_{PID}$,	0.12,0.11, 0.08,0.04	holiday, naked shoulders, high sun radiation
6.	00000000-00000000, 00000000,$< 0 >_{PID}$,	0.21, 0.21,0.13	working, no sun

Table 1: Key-features. In the first column the cluster number is displayed. Second column contains the most probable patterns for the cluster. The presented feature numbers are equivalent to the set of questions given in the section *Sun Exposure Study.* For example: feature 10111000 gives the following set of answers: holiday - yes, abroad - no, sun bathing - yes, naked shoulders - yes, on the beach - yes, remaining questions 6,7,and 8 - no. Patterns corresponding to the *PID* histograms are marked with the subscript "*PID*". The average value corresponding to the quantized number is given here. The lowest observed value is 0 and the highest 7.5. The co-occurring features are shown with the dash between them e.g., "00000000-10000000" means that a feature - working is followed by feature - holiday. Third column gives the probabilities for the key-features and fourth column presents general description of cluster based on the key-features.

The optimal model has 6 clusters described by key-features given in the tale 1. The key-features, associated probabilities and description of the clusters are provided. In the first column the cluster number is displayed. Second column contains the most probable features for the cluster. The presented feature numbers are equivalent to the set of questions given in the section *Sun Exposure Study.* For example: feature 10111000 gives the following set of answers: 1. holiday - "yes", 2.

abroad - "no", 3. sun bathing - "yes", 4. naked shoulders - "yes", 5. on the beach - "yes", remaining questions 6,7,and 8 - no. Patterns corresponding to the *PID* histograms are marked with the subscript "*PID*". The quantized number is given here. The lowest observed value is 0 and the highest 7.5[3]. The co-occurring features are shown with the dash between them e.g., "00000000-10000000" means that a feature - working is followed by feature - holiday. The third gives the probabilities for the key-patterns and the fourth column presents a general description of the cluster based on the key-patterns.

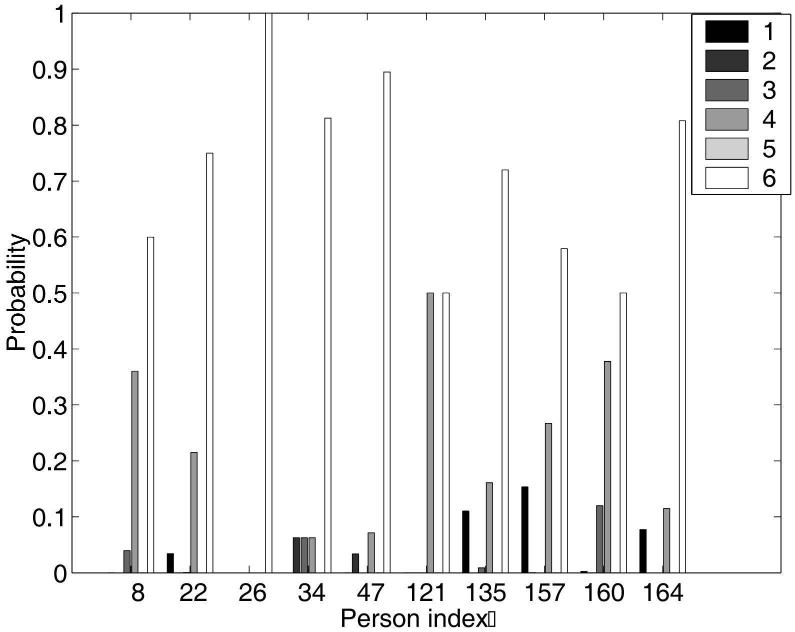

Figure 4: Cluster probabilities calculated for the 10 test persons Eq. (3). Person index is shown on the x-axes and different grey level colors corresponds to six clusters. Associated key-patterns are given in table 1.

Figure 4 presents the cluster probabilities calculated according to the Eq. 3 for the 10 test subjects. The data (consisting of 5 consecutive days of diaries returned by particular subjects) from each person is assigned to the one of the 6 clusters based on the posterior probability $p(k|Per)$, Eq. (3). Together with key-features presented in table 1 it gives a good description of the behavior of the particular persons during the whole period of the experiment. For all test persons there is a large probability for the cluster no. 6, i.e., the person is at work and do not get sun exposure. Also others behaviours are identified. For example, person no. 160 has a high probability component for cluster no. 3 which is holidays on the beach but with low sun radiation. Persons no. 34 has high content of cluster no. 2 that means holiday abroad and high sun radiation, and cluster no. 3 and 4 - holidays with low sun radiation.

[3] The highest observed quantized *PID* value in the data is equal to 30, however, due to the low occurrence, values higher than 7.5 were removed.

CONCLUSION

This paper discusses using the Latent Semantic Indexing framework combined with the Gaussian Mixture Model for processing and clustering categorical data. Moreover, it provides the possibility for combining multiple data types into a common vector space framework. We successfully applied the method to analyze a combination of categorical diary data and real valued sun radiation measurements. All the used sources of information contribute to the final quality of the clustering. Using the analogy to textmining, we proposed methods for interpretation of the identified clusters. This framework also allows interpretation of the behaviour of the new subjects in the experiment.

REFERENCES

[1] H. Akaike, “Fitting Autoregresive Models for Predition,” **Ann. of the Ins. of Stat. Math.**, vol. 21, pp. 243–247, 1969.

[2] S. Deerwester, S. Dumais, G. Furnas, T. Landauer and R. Harshman, “Indexing by Latent Semantic Analysis,” **J. Amer. Soc. for Inf. Science**, vol. 41, pp. 391–407, 1990.

[3] L. Hansen and J. Larsen, “Unsupervised Learning and Generalization,” in **Proceedings of the 1996 IEEE International Conference on Neural Networks**, Washington DC, USA, 1996, pp. 25–30.

[4] L. Hansen, S. Sigurdsson, T. Kolenda, F. Nielsen, U. Kjems and J. Larsen, “Modeling text with generalizable gaussian mixtures,” in **Proceedings of IEEE ICASSP’2000**, 2000, vol. VI, pp. 3494–3497.

[5] T. Kolenda, L. K. Hansen, J. Larsen and O. Winther, “Independent component analysis for understanding multimedia content,” in H. Bourlard, T. Adali, S. Bengio, J. Larsen and S. Douglas (eds.), **Proceedings of IEEE Workshop on Neural Networks for Signal Processing XII**, 2002.

[6] J. Larsen, L. Hansen, A. Szymkowiak-Have, T. Christiansen and T. Kolenda, “Webmining: Learning from the World Wide Web,” **Computational statistics and data analysis**, vol. 38, pp. 517–532, 2002.

[7] J. Larsen, A. Szymkowiak and L. Hansen, “Probabilistic Hierarchical Clustering with Labeled and Unlabeled Data,” **International Journal of Knowledge-Based Intelligent Engineering Systems**, vol. 6, no. 1, pp. 56–62, 2002.

[8] B. Ripley, **Pattern Recognition and Neural Networks**, Cambridge University Press, 1996.

[9] A. Szymkowiak, J. Larsen and L. Hansen, “Hierarchical Clustering for datamining,” in N. Babs, L. Jain and R. Howlett (eds.), **Proc. 5th Int. Conf. on Knowledge-Based Intelligent Information Engineering Systems and Allied Technologies**, 2001, pp. 261–265.

[10] A. Szymkowiak, P. Philipsen, J. Larsen, L. Hansen, E. Thieden and H. Wulf, “Impuating missing values in diary records of sun-exposure study,” in D. Miller, T. Adali, J. Larsen, M. V. Hulle and S. Douglas (eds.), **Proceedings of IEEE Workshop on Neural Networks for Signal Processing XI**, Falmouth, Massachusetts, 2001, pp. 489–498.