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: eivind.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 combined Latent Semantic indexing like approach 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". 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 subtask of identifying relevant structure in the combined data set of sun habit diaries and daily UV dose measurements. We aim at identifying relevant structure using hierarchical probabilistic clustering. 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 [4, 5, 6, 7, 9] mining framework for analysis and understanding of behavioral data.

SUN EXPOSURE STUDY

A specially designed device, measuring received sun radiation (*PID*), was given to 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

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 patterns for the presented set of questions equals 20736, however, only a small fraction of 423 patterns actually exist in the investigated data set.

PREPROCESSING

Latent Semantic Indexing (LSI) [2] was developed for text and multimedia mining, see [4, 5]. In this study we pursue a similar idea, which enables to combining different types of data into a common framework. Figure 1 presents the general framework of preprocessing, clustering and data post-processing. In the first step, data is windowed creating vectors that contain data from consecutive days. The optimal size of the window is an issue 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. In the experiments a window of size 7

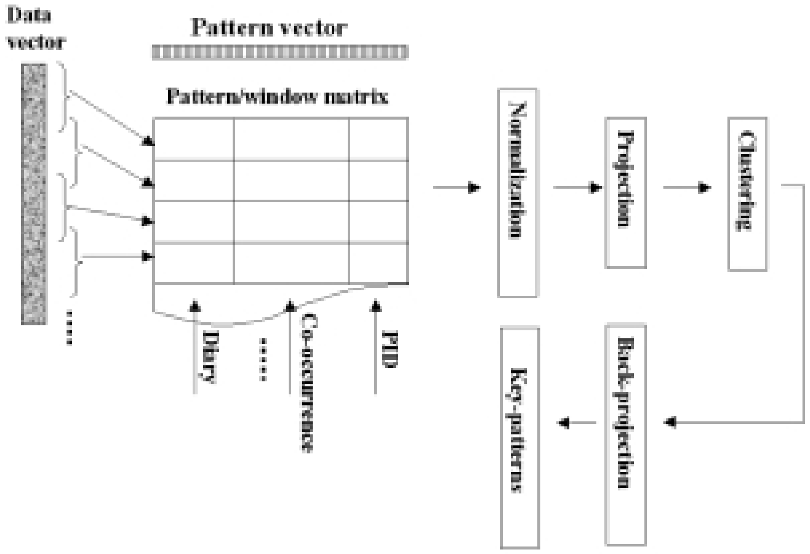

Figure 1: Framework for data clustering: 1) the data is windowed into several histogram vectors and together with the co-occurrence matrix and the *PID* histogram forms a pattern/window matrix. 2) data is then normalized and projected onto the orthogonal singular value decomposition space. 3) the Gaussian 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-patterns are identified.

is used. This was decided after several experiments, taking into account stationarity of the clustering and complexity level. In the final paper we plan to invoke the concept of generalization for optimal window selection, see further [3].

Originally, the pattern/window matrix is formed from the histogram vectors achieved by counting occurrences of every found pattern in the window. However, the histograms does not convey time ordering information. It is possible to include time information by considering the co-occurrence matrix of joint occurrences of neighbor patterns in the window. There are 20736^2 possible co-occurences but only 1509 were present in the actual data set. The continuous sun radiation measurements were quantized in order to fit the presented framework. Both diary histograms, the co-occurrence matrix and sun radiation are screened against rare patterns by removing patterns which have occurrence below a certain threshold.

The next step involves normalization of the pattern/window matrix. Two types of normalization are performed. First, each window vector is scaled to unity length, and then, pattern vectors are scaled to zero mean and unit variance over training samples. The three component matrices (diary-window histograms, co-occurrence and *PID* histograms) are then separately projected onto the few principal component directions found by singular value decomposition (SVD). Finally, the generalizable Gaussian mixture model is used for clustering in the subspace.

UNSUPERVISED GAUSSIAN MIXTURE MODEL

The Gaussian mixture model was previously addressed in [4, 6, 8]. The K component mixture of Gaussian densities of the d-dimensional feature vector $\boldsymbol{x}$ is defined as:

$$p(\boldsymbol{x}) = \sum_k p(\boldsymbol{x}|k) \cdot p(k) \tag{1}$$

where $p(\boldsymbol{x}|k) \equiv \mathcal{N}(\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k})$ are Gaussian densities and $p(k)$ are nonnegative mixture proportions such that $\sum_k p(k) = 1$. The parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are estimated from the training data set $\mathcal{D} = \{\boldsymbol{x}_n, n = 1 \ldots N\}$ by minimizing negative log-likelihood cost function of the form: $\mathcal{L} = -\sum_n log(p(\boldsymbol{x}_n|k))$ through expectation-maximization method. In order to ensure generalizability, parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are estimated from the disjoint sets of observations and the optimal number of mixture components is found by the AIC-criterion [1, 3]. The complete algorithm for generalizable Gaussian mixture model (GGM) can be found in [4, 7].

OUTLIER DETECTION

The Gaussian mixture model models the data density In order to spot an outlier, which indicate non-stationarity in data, the cumulative probability [6, 7] is computed $Q(t) = \text{Prob}(x \in \mathcal{R}), \mathcal{R} = \{\boldsymbol{x} : p(\boldsymbol{x} < t)\}$ for all thresholds t. Thus, the outliers occupy lower part of the cumulative curve.

PROTOTYPES

In order to find key-patterns corresponding to each of the clusters, centers $\boldsymbol{\mu}$ need to be back-projected to the original space of normalized histograms[1]. Furthermore, 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-patterns. The confidence of assigning the person into the given cluster k can be expressed by the posterior probability:

$$p(k|Per) = \frac{1}{N} \sum_i p(k|Per, \boldsymbol{x}_i) \cdot p(\boldsymbol{x}_i), \tag{2}$$

where $\boldsymbol{x}_i$ is a feature vector of the size d and $i = 1, 2, \ldots, N$. The number of feature vectors N is different for every person and depends on the number of returned diary records and the window size.

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

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 testing. Sun exposure measurements were quantized into 4 values. The slicing window of size 7 was applied forming 2580 training and 158 test feature vectors. Each feature vector consist of the diary histogram, the co-occurrence matrix and the *PID* histogram. The diary histogram is reduced from 423 to 97 patterns by removing rare patterns. In a similar way, the co-occurrence matrix is reduced from 1509 to the 80 most often occurring pairs of patterns. Each of these matrices are projected separately on the orthogonal directions found by SVD. For both diary and co-occurrence the 9 largest eigenvalues is used and 3 for *PID* data[2].

The investigation was performed of the importance of the co-occurrence matrix and the *PID* histograms for the clustering. The results of the experiments are collected in the tables 1, 2, 3 and 4.

In the experiments hard assignment GGM model [4] is used, i.e., the parameters of the clusters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ were 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 [9, 7].

In the tables 1, 2, 3 and 4 the results of back-propagation are shown. The key-patterns, associated probabilities and description of the clusters are provided. In the first column the cluster number is displayed. Second column contains the most probable patterns for the cluster. 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.

In table 1 the results of clustering of the diary histograms are shown. The presented patterns are equivalent to the set of questions given in section: *Sun Exposure Study*. For example: pattern 10111 describes 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", or pattern 0: all the questions where answered "no" or pattern 1: 1. holiday - "yes" and the rest of the questions from 2 to 8 - "no". This rule for describing patterns hold as well in the case of table 2, 3 and 4.

Table 2 presents key-patterns for clustering diary histograms combined with *PID* histograms. Eight clusters were found. Diary key-patterns are explained in table 1. Patterns corresponding to the *PID* histograms are marked with the subscript "*PID*". Four different values of *PID* from 0 to 3 are observed: 0 corresponds to the very low sun radiation and 3 describes very high one. This rule for describing *PID*-patterns hold as well in the case of table 4.

[2] The decision was made based on the shape of the eigenvalue curve but a more elaborate selection can be invoked using the concept of generalization [4].

#.	Key-Pattern	Probability.	Description
1.	10001,11,10111	0.33,0.32,0.19	holiday, on the beach, sun bathing
2.	0	0.98	working - no sun
3.	1	0.9	on holiday - no sun
4.	0,0001,1	0.4,0.27,0.18	working naked shoulders - no sun
5.	1,1101	0.67,0.17	holiday, naked shoulders
6.	1011,1001,10011	0.47,0.17,0.16	holiday , sun bathing
7.	11	0.5	holiday abroad - no sun
8.	10111,0001,1001	0.45,0.17,0.13	holiday, sun bathing, naked shoulders
9.	0000001	0.05	no sun, sunburned - red
10.	0	0.99	working - no sun

Table 1: Key-patterns for clustering diary histograms. In the first column the cluster number is shown. Second column contains the most probable patterns for the cluster. The presented pattern numbers are equivalent to the set of questions given in section: *Sun Exposure Study.* For example: pattern 10111 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, or pattern 0 means that all the questions where answered "no". Third column gives the probabilities for the key-patterns, and fourth column presents a general description of cluster.

In table 3 the key-patterns for clustering diary histograms combined with co-occurrence matrix are presented. The diary key-patterns are explained in table 1. The co-occurring patterns are shown with the dash between them e.g., "0-1" means that a pattern working is followed by pattern holiday, pattern "1-10011" means that holiday without sun was followed by holiday spent on the beach. This rule for describing co-occurrence patterns hold as well in the case of table 4.

Table 4 shows the key-patterns for clustering diary histograms combined with co-occurrence matrix and *PID* histograms. The diary key-patterns are explained in table 1. The co-occurred patterns are explained in table 3 and the *PID* patterns in table 2. Both the *PID* values and the co-occurrence pairs are likely to appear as key-patterns. This could suggest that joining time information and the sun exposure measurements are important for the clustering. Moreover, the description of the clusters is more explicit.

In figure 2 the probability of observing certain groups of behaviors in the clusters together with registered sun exposure values are presented. Clustering was done using full pattern/window matrix for which keywords are displayed in table 4. Five behaviors are specified: *working - no sun exposure*, *holiday - no sun exposure*, *sun exposure* describes mild sun behaviors often on the beach or naked shoulders without sun-screen and without sunburns, *using sun-block* and diary records with reported *sunburns.* In the bottom figure the observed sun exposure measurements are presented. For example cluster number 6 groups behaviors marked as *working - no sun* and corresponding *PID* values are low. Opposite, cluster no. 5 contains records with reported sunburns, sun exposure and using sun-block and consequently *PID* values are high.

For the same clustering setting the cluster probabilities were calculated Eq. (2) for 10 test subjects. Together with key-patterns presented in table 4

#	Key-Pattern	Probability.	Description
1.	1001,1000, 1_{PID},10011	0.31,0.26, 0.16,0.11	holiday, naked shoulders, small *PID*
2.	11,0001,0, 2_{PID},0_{PID}	0.29,0.2,0.17, 0.16,0.15	holiday abroad, working
3.	1,11	0.39,0.12,	holiday
4.	1011,2_{PID}, 3_{PID},0001	0.0.31,0.25, 0.14,0.13	naked shoulders, high sun radiation
5.	1,2_{PID},3_{PID},10001	0.2,0.17,0.16,0.14,0.12,0.1	holidays, high *PID*
6.	1_{PID},0_{PID}	0.14,0.13	low *PID*
7.	3_{PID},1001, 2_{PID},10011	0.22,0.22, 0.16,0.15	holiday, naked, shoulders, high *PID*
8.	0,0_{PID}	0.6,0.4	no sun

Table 2: Key-patterns for clustering diary histograms combined with *PID* histograms. In the first column the cluster number is displayed. Second column contains the most probable patterns for the cluster. The diary key-patterns are explained in table 1. Patterns corresponding to the *PID* histograms are marked with the subscript "*PID*". Four different values of *PID* are observed: 0 corresponds to very low sun radiation and 3 describes very high one. Third column gives the probabilities for the key-patterns and fourth column presents general description of cluster based on the key-patterns.

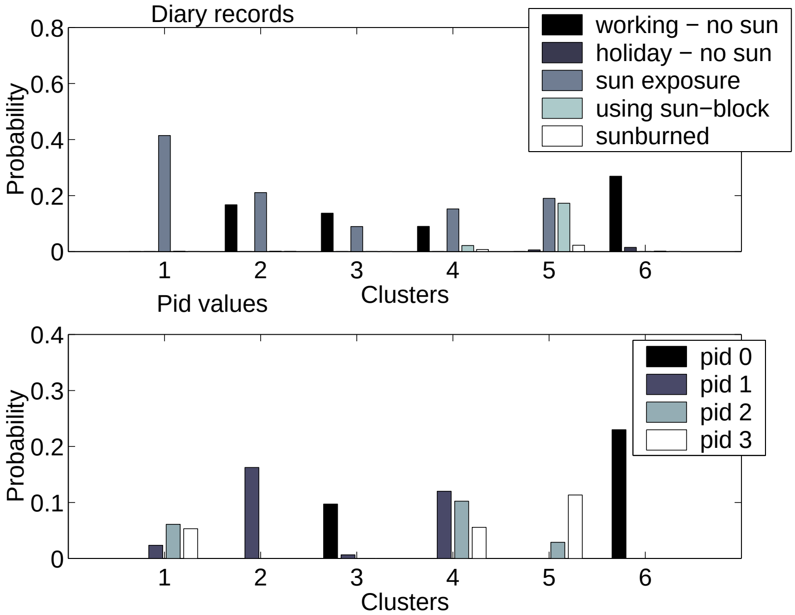

Figure 2: The probability of observing certain groups of behaviors in the clusters together with registered sun exposure values. Key-patterns for the clusters are presented in the table 4. For each cluster grouped behaviors from diary records are presented on the upper plot and corresponding *PID* is shown on the lower figure.

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

#	Pattern	Probability.	Description
1.	1001,1101-1101	0.27,0.13	holiday,naked sholders
2.	1001,1101-1101,1	0.26,0.21,0.1	holiday,naked sholders
3.	0001,1001-0, 10111,0-1001	0.17,0.12, 0.11,0.1	working, naked shoulders
4.	11,11-11	0.14,0.11	holiday, abroad
5.	0001,1001-0, 1001,0-1001	0.27,0.14, 0.13,0.1	holiday or working, naked shoulders
6.	1001,0,0-1,1-0,1,1-1	0.29,0.19,0.16,0.14,0.1,0.09	work - holiday, no sun
7.	10011	0.19	holiday, on the beach
8.	10001,1-10011	0.21,0.12	holiday, on the beach
9.	0-0,0,0-1,1-0	0.36,0.35,0.12,0.12	working - no sun
10.	1001,1-1101,1101-1, 1101-1101,1011	0.26,0.16,0.12, 0.12,0.11	holiday, naked shoulders

Table 3: Key-patterns for clustering diary histograms combined with co-occurrence matrix. In the first column the cluster number is displayed. Second column contains the most probable patterns for the cluster. The diary key-patterns are explained in table 1. The co-occurring patterns are shown with the dash between them e.g., "0-1" means that a pattern working is followed by pattern holiday. Third column gives the probabilities for the key-patterns and fourth column presents general description of cluster based on the key-patterns.

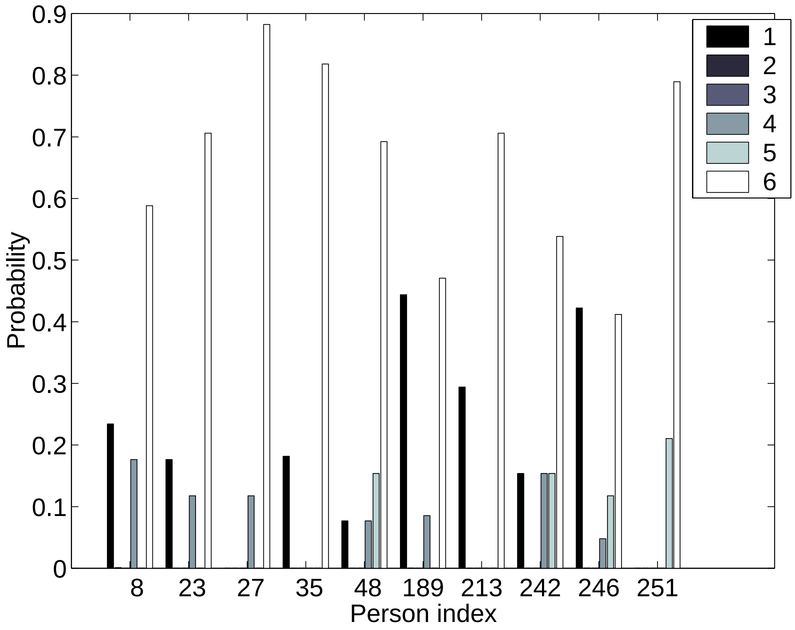

Figure 3: Cluster probabilities calculated for the 10 test persons Eq. (2). Person index is shown on the x-axes and different grey level colors corresponds to six clusters. Key-patterns are given in table 4.

probability of the cluster no. 6 that describes working and no sun exposure. However, some of the periods are described by other behaviors. For example for person no. 251 there is high probability component for cluster no. 5

#	Pattern	Probability	Description
1.	1001,0001,1101-1101	0.15,0.1,0.09	naked shoulders
2.	0,1_{PID},0-0,0001	0.17,0.16,0.14,0.13	working, low sun radiation
3.	1001-0,0-1001, 0,1-0,0-1,0_{PID}	0.17,0.14 ,0.14,0.12,0.11,0.1	no sun radiation, holiday-work
4.	1_{PID},2_{PID}	0.12,0.1	medium sun exposure
5.	3_{PID},11,11-11	0.11,0.11,0.09	holiday, high sun radiation
6.	0-0,0,0_{PID}	0.29,0.27,0.23	working, no sun

Table 4: Key-patterns for clustering the diary histograms combined with the co-occurrence matrix and the *PID* histograms. In the first column the cluster number is displayed. Second column contains the most probable patterns for the cluster. The diary key-patterns are explained in table 1. The co-occurred patterns are explained in table 3 and the *PID* patterns in table 2. Third column gives the probabilities for the key-patterns and fourth column presents general description of cluster based on the key-patterns.

describing holidays with high sun radiation. Persons no. 213 and 35 can be well described by clusters 6 (working, no sun) and 1 (naked shoulders) while person no. 23 by clusters 6, 1 and 4 (medium sun exposure).

CONCLUSION

This paper discusses using an Latent Semantic Indexing like method for processing and clustering categorical data. Moreover, it provides the possibility for combining multiple date types into a common vector space framework. We applied the method to analysis a combination of categorical diary data and real valued sun radiation measurements. Using the analogy to textmining we proposed methods for interpretation of the identified clusters. This scheme allows for evaluating the significance of various feature representations. For the specific data set we addressed the role of different representations. Preliminary results indicate that the sequence information and UV dose measurements contribute to stabilizing the clustering model and its interpretation.

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," 2002, **submitted for NNSP2002**.

[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.