Mixture Random Effect Model Based Meta-analysis For Medical Data Mining

Yinglong Xia*, Shifeng Weng*, Changshui Zhang**, and Shao Li

State Key Laboratory of Intelligent Technology and Systems,Department of Automation, Tsinghua University, Beijing, China
xiayl03@mails.tsinghua.edu.cn, wengsf@tsinghua.org.cn,
{zcs,shaoli}@mail.tsinghua.edu.cn

Abstract. As a powerful tool for summarizing the distributed medical information, Meta-analysis has played an important role in medical research in the past decades. In this paper, a more general statistical model for meta-analysis is proposed to integrate heterogeneous medical researches efficiently. The novel model, named mixture random effect model (MREM), is constructed by Gaussian Mixture Model (GMM) and unifies the existing fixed effect model and random effect model. The parameters of the proposed model are estimated by Markov Chain Monte Carlo (MCMC) method. Not only can MREM discover underlying structure and intrinsic heterogeneity of meta datasets, but also can imply reasonable subgroup division. These merits embody the significance of our methods for heterogeneity assessment. Both simulation results and experiments on real medical datasets demonstrate the performance of the proposed model.

1 Introduction

As the great improvement of experimental technologies, the growth of the volume of scientific data relevant to medical experiment researches is getting more and more massively. However, often the results spreading over journals and online database appear inconsistent or even contradict because of variance of the studies. It makes the evaluation of those studies to be difficult. Meta-analysis is statistical technique for assembling to integrate the findings of a large collection of analysis results from individual studies. Many academic papers and books have discussed the application of meta-analysis in medical researches[1].

Meta-analysis employs various statistic models to integrate available individual medical research results. Those models can be divided into fixed effect model and random effect model according to the different assumption of effect size, which is conceptualized as a standardized difference between trials for identical purpose. In the fixed effect model, the studies are assumed all to generate from a fixed underlying effect size; while random effect model further takes into account

[*] These two authors contribute equally to this paper

[**] Corresponding author

the extra variation[2]. The Hierarchical Bayes Linear Model (HBLM) mentioned in[3] is essentially a random effect model cooperated with prior knowledge.

Although existing meta-analysis methods have been used for decades, their intrinsic limitations lead to poor performance on complicated meta datasets. The reason is that the model assumption of those methods is too strong and therefore lack of flexibility. For example, fixed effect model regards the underlying effect size is not influenced by any other factor; while random effect model deems that the influences are centralized. In this paper, we propose a novel meta-analysis model based on Gaussian Mixture Model (GMM). The novel model, which can be viewed as an optimal linear combination of random effect models, is named *Mixture Random Effect Model* (MREM). It will be shown that traditional fixed effect methods and random effect methods are just two special cases of the proposed MREM.

2 Statistic strategies in meta-analysis

In meta-analysis, effect size is defined to represent the standardized performance difference between treatment group and control group in medical studies. There are multiple types of definitions of the effect size, such as risk difference (RD), relative risk (RR), odds ratio (OR), and the logarithm of the them[1]. The treatment group consists of individuals who undergo a certain medical treatment; while the control group is a collection of individuals who keep away from the treatment and just serve as reference.

Since a medical study is effected by many factors, the distribution of effect size approximates to be normal according to *Central Limit Theorem*, that is

$$y_i \sim N(\mu_i, s_i^2), \tag{1}$$

where μ_i is the underlying effect size in the i^{th} study and s_i^2 is the correspondent variance. The fixed effect model assumes the effect sizes of all studies are homogeneous and share the same fixed underlying value:

$$y_i \sim N(\mu, s_i^2), \tag{2}$$

where each study shares the same mean but different variance. Different from the fixed effect model, the random effect model considers the heterogeneity of data and assumes the distribution of underlying effect size is normal:

$$\begin{aligned} y_i &\sim N(\mu_i, s_i^2) \\ \mu_i &\sim N(\mu, \tau^2) \end{aligned} \tag{3}$$

where μ_i and s_i^2 are study-specific mean and variance respectively, that is to say, in the random effect model, μ_i is assumed to arise from a gaussian distribution with mean μ and variance τ^2.

The Hierarchical Bayes Linear Model (HBLM) makes an improvement of random effect model. It replaces μ in Equation (3) with a linear combination of the covariates, $x_i\beta$. Thus, it demands strong support of prior knowledge, such as correct selection and complete covariate information. Even so, HBLM is still poor for arbitrary distribution of effect size.

3 Mixture random effect model

3.1 Mixture random effect model

Fixed effect model and random effect model provide us two approaches to exploit the underlying structure of μ_i in meta data set. For a real meta dataset, μ_i may arise from an arbitrarily complex distribution rather than a constant in fixed effect model or a simple gaussian distribution in random effect model. Random effect model may exhibit poor performance at least in the following two cases:

1. The distribution of effect size has a unique peak but it is not a normal. Modelling μ_i as a gaussian distribution will introduce extra error.
2. The distribution of μ_i complies with a multi-peak distribution. This situation is common in real world dataset.

As mentioned, HBLM makes an effort to deal with a complex distribution by introducing covariate, $x_i\beta$. The three shortages of regression based remedy are that the covariates may not be linear additive, the selection of covariates is not easy, and the covariate information is usually unavailable for some studies in many practical cases .

Therefore, we develop a novel model to describe the characteristic of μ_i, which is expected to handle both of the two special cases listed above. Here, we propose the Mixture Random Effect Model (MREM), which is given by:

$$\begin{array}{l} y_i \sim N(\mu_i, s_i^2) \\ \mu_i \sim \sum_{l=1}^{M} \alpha_l N(\xi_l, \sigma_l^2), \;\; \sum_{l=1}^{M} \alpha_l = 1, \alpha_i > 0 \end{array} \tag{4}$$

Mathematically speaking, the above model utilize a gaussian mixture distribution to describe μ_i, the mean of effect size. When the number of gaussian components, M, is equal to 1, MREM degenerates to the traditional random effect model. Since μ_i in mixture random effect model is learnt unsupervisedly, we do not need any special covariate information from literatures.

After presenting MREM in this section, the problems such as choosing a proper method for parameter estimating, finding an explanation for learnt model, and proceeding subgroup analysis when the learnt gaussian components are well clustered, will be discussed in the following sections.

3.2 Parameter estimation by Gibbs sampling

In this section, we explore the methods to learn the parameters of the proposed MREM model from a meta dataset. As an important task in statistics and data mining, parameter estimation for mixture distribution has been explored for many years. Among these parameter estimation schemes, Expectation Maximization (EM) algorithm [4] and Gibbs sampling [5] are two widely used methods. For the proposed mixture random effect model, we tried EM algorithm and found that there was no close form for estimating θ_ls in each iteration, so a complicated nested iteration should be performed in each EM iteration. Thus, Gibbs sampling is considered for MREM.

Gibbs sampling scheme is one of the widely used Markov Chain Monte Carlo (MCMC) routes [5] and has been applied in multidimensional sampling problems. In those problems, the joint distribution $P(x)$ is assumed to be too difficult to draw samples directly; while conditional distributions $P(x_i|\{x_j\}_{j\neq i})$ are comparatively feasible to be sampled.

In MREM, we need to estimate parameters of M gaussian components, $\Theta = \{\alpha_l, \xi_l, \sigma_l | l = 1, ..., M\}$. Those parameters are assumed to be independent to each other. Let all priors on mixture ratio α_l, location ξ_l and logarithm variance $\log \sigma_l$ be noninformative, that is,

$$\alpha_l \sim U(0,1), \quad \xi_l \sim U(-\infty, \infty), \quad \log \sigma_l \sim U(0, \infty) \tag{5}$$

where the prior on ξ_l, $\log \sigma_l$ are two *improper* priors in statistics. An improper prior is not integrable until it times by a likelihood function, that is, we can obtain a proper posterior distribution from an improper prior.

For each sample, we introduce a latent variable termed component indicator, $z_i \in \{1, 2, \cdots, M\}$, which means that μ_i generates from gaussian component z_i. Therefore, the joint distribution is decomposed as:

$$\begin{aligned} p(\alpha, \sigma, \xi, Y, Z) &= p(\alpha, \sigma, \xi) p(Z|\alpha, \sigma, \xi) p(Y|Z, \alpha, \sigma, \xi) \\ &= p(\alpha) p(\sigma) p(\xi) p(Z|\alpha, \sigma, \xi) p(Y|Z, \alpha, \sigma, \xi) \end{aligned} \tag{6}$$

To apply Gibbs sampler, we need to find the full conditional distribution of each parameter. Since

$$p(\alpha_l|\alpha_{\bar{l}}, \sigma, \xi, Y, Z) \propto p(\alpha) p(Z|\alpha, \xi, \sigma) \propto \prod_{i=1}^{K} \alpha_{z_i = l} \tag{7}$$

where $\bar{l} = \{1, 2, \cdots, M\} \backslash \{l\}$ and the full conditional on α is a Dirichlet distribution,

$$p(\alpha_1, ..., \alpha_M | \sigma, \xi, Y, Z) = Dir\left(n_1, n_2 .., n_M\right) \tag{8}$$

where $n_l = \sum_{i=1}^{K} I(z_i = l)$. From Equation (6), we choose factors containing ξ, thus we have,

$$\begin{aligned} p(\xi_l|\alpha, \sigma, \xi_{\bar{l}}, Y, Z) &\propto p(\xi) p(Y|\alpha, \xi, \sigma, Z) \\ &\propto \prod_{i=1, z_i = l}^{K} \exp\left(-\frac{1}{2} \frac{(y_i - \xi_l)^2}{\sigma_l^2 + s_i^2}\right) \end{aligned} \tag{9}$$

Normalizing the proportion above, we get an explicit gaussian distribution,

$$p(\xi_l|\alpha, \sigma, \xi_{\bar{l}}, Y, Z) = \mathcal{N}\left(\frac{\sum\limits_{i=1, z_i = l}^{K} \frac{y_i}{\sigma_l^2 + s_i^2}}{\sum\limits_{i=1, z_i = l}^{K} \frac{1}{\sigma_l^2 + s_i^2}}, \frac{1}{\sum\limits_{i=1, z_i = l}^{K} \frac{1}{\sigma_l^2 + s_i^2}}\right) \tag{10}$$

Now, we derive the updating formula for variance σ. From Equation (6), we have,

$$\begin{aligned} p(\sigma_l^2|\alpha, \sigma_{\bar{l}}, \xi, Y, Z) &\propto p(\sigma)p(Y|\alpha, \xi, \sigma, Z) \\ &\propto \frac{1}{\sigma_l^2} \prod_{i=1, z_i=l}^{K} \frac{1}{\sqrt{\sigma_l^2 + s_i^2}} \exp\left(-\frac{1}{2}\frac{(y_i - \xi_l)^2}{\sigma_l^2 + s_i^2}\right) \end{aligned} \tag{11}$$

At last, the full conditional for z_i is calculated straight forward, that is,

$$\begin{aligned} p(z_i = l|\alpha, \sigma, \xi, Y, z_{\bar{i}}) &\propto p(z_i = l|\alpha, \xi, \sigma)p(y_i|\alpha, \xi, \sigma, z_i = l) \\ &\propto \alpha_l \frac{1}{\sqrt{\sigma_l^2 + s_i^2}} \exp\left(-\frac{1}{2}\frac{(y_i - \xi_l)^2}{\sigma_l^2 + s_i^2}\right) \end{aligned} \tag{12}$$

Note that z_i is a discrete random variable valued in $\{1, 2, \cdots, M\}$, where M is the total number of components of the gaussian mixture model.

3.3 Implement of variance updating

Once obtaining the full conditional distributions of the parameters (Equation (8), (10), (11) and (12)), we can iteratively apply Gibbs sampler for estimation. However, the full conditional distribution on σ_l is not a standard distribution and can not sample directly. Here, we employ rejection sampling with a uniform proposal function to address this problem. Therefore, the problem turns to determining the upper and lower bounds of the proposal function.

Consider a special case in which all s_i are identical, i.e. $s_i \equiv s$ and the noninformative prior, $p(\sigma_l^2) \propto 1/\sigma_l^2$. Thus, the posterior density turns to an inverse χ^2 density function:

$$p((\sigma_l^2 + s^2)|\alpha, \sigma_{\bar{l}}, \xi, Y, Z) = Inv\text{–}\chi^2(n_l, \sum_{i=1}^{n_l}(y_i - \xi_i)^2/n_l) \tag{13}$$

The above posterior is also represented equivalently as a inverse gamma distribution. Denoting sample variance $v_l = \sum_{i=1}^{n_l}(y_i - \xi_i)^2/n_l$, we obtain confidence interval of σ_l^2 with respect to given precision range, (P_{min}, P_{max}), that is,

$$(\sigma_{min}^2, \sigma_{max}^2) = \left(\frac{n_l v_l^2}{\mathcal{I}\chi^2(P_{max}, n_l)}, \frac{n_l v_l^2}{\mathcal{I}\chi^2(P_{min}, n_l)}\right) \tag{14}$$

where $\mathcal{I}\chi^2(P, n_l)$ is inverse χ^2 cumulative distribution function with freedom degree of n_l and value of P, e.g. 2.5% and 97.5%. Let $s = \min_i(s_i)$, we have a confidence interval $(\sigma_{min}^{(1)}, \sigma_{max}^{(1)})$; and $s = \max_i(s_i)$, we have $(\sigma_{min}^{(2)}, \sigma_{max}^{(2)})$. Therefore, we get a new interval as $(\sigma_{min}, \sigma_{max}) = (\sigma_{min}^{(1)}, \sigma_{max}^{(2)})$, within which σ_l^2 occurs with a high propability. Then, rejection sampling is consequently applied on $(\sigma_{min}, \sigma_{max})$ to generate a sample for updating σ_l^2.

4 Model selection and subgroup division

4.1 Model selection by BIC

The number of gaussian components, M, is the only parameter that should be preset in MREM. Essentially, determining the best value of M is a model selection problem which can be solved with some feasible model selection criteria, such as AIC[6], MDL[6] and BIC[7]. Because of its broadly application, BIC is employed in MREM:

$$BIC = \log p(D|\Theta) - \frac{1}{2} d \log(K) \tag{15}$$

where D is the data, Θ is the ML estimate of the parameters, d is the number of parameters, and K is the number of data points. BIC is quite intuitive, namely, it contains a term measuring how well the parameterized model predicts the data ($\log p(D|\Theta)$) and a term which punishes the complexity of the model ($\frac{1}{2} d \log(K)$). Thus, in our algorithm, the model with the highest BIC score is selected.

4.2 Subgroup division

One merit of MREM proposed in this paper is that it is capable to approximate arbitrary distribution, even if the distribution is very complicated. When significant disequilibrium heterogeneity exists, it is natural to divide samples into several subgroups for further study. There are two approaches for subgroup division.

The first approach is to implement division by directly observing the distributions of the gaussian components estimated in MREM, which are all one-dimensional. This approach is often feasible when the number of the gaussian components is small, or there is enough prior knowledge.

The other subgroup division approach is required when the number of components is somewhat large, and it is lack of sufficient prior knowledge. We adopt hierarchical clustering to unsupervisedly merge adjacent components. Different from an ordinary clustering task, the clustering here is applied on gaussian components. Thus, a proper measurement of dissimilarity between two gaussian components is required. Here, we employ symmetric KL divergence[8]:

$$KL(\theta_i, \theta_j) = \int_x (p(x|\theta_i) - p(x|\theta_j)) \log \frac{p(x|\theta_i)}{p(x|\theta_j)} dx \tag{16}$$

where θ_i is the parameter of the i^{th} component. The hierarchical clustering technique is an unsupervised data analysis method, which dose not demand any prior knowledge. The results of hierarchical clustering not only reveals the proper number of subgroups, but also indicates which components should be merged in most cases.

5 Experiments

5.1 Simulation experiment

In this section, we design an experiment on a simulated dataset to illustrate the performance of MREM. In this experiment (Experiment 1), the underlying distribution of μ_i is a GMM with three gaussian components:

$$\mu_i \sim 0.18N(x|-2.5, 0.7^2) + 0.45N(x|-1, 0.7^2) + 0.36N(x|2.5, 0.6^2) \tag{17}$$

We draw 150 samples from Equation (17) as the means of effect size and denote them as $\mu_1, \mu_2, ..., \mu_{150}$. The correspondent variance s_i^2 is generated from a uniform distribution $U(0.5, 1)$. Then, a effect size y_i is drawn from $y_i \sim N(\mu_i, s_i^2)$. The task in this experiment is to approximate the distribution of μ_i given y_i and s_i for $i = 1, 2, \cdots, K$. Figure 1 shows the results.

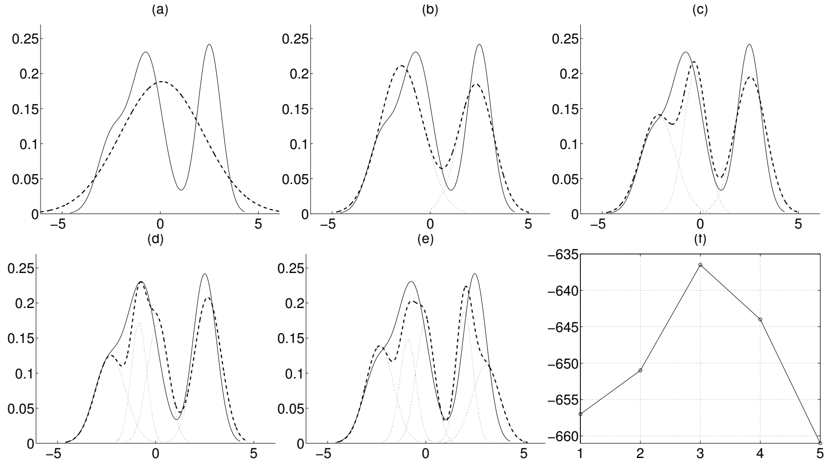

Fig. 1. Results of Experiment 1. (a)∼(e) the estimation results of GMMs with 1,2,· · ·,5 components respectively, where the solid curve indicates the true pdf and the dashed curve represents the estimated pdf. (f) BIC scores versus the number of components

The number of components in GMMs are set from 1 to 5 respectively (see Figure 1(a) to (e)). It is found that all the iterative processes of MREM converge rapidly. The BIC score curve (see Figure 1(f)) suggests the model with 3 components is the best one, which is consistent with that of the true model.

From the selected model in Figure 1(c), we find that the left two components locate closely and they are prone to merging together as a subgroup. Therefore, it is intuitionally reasonable to divide the simulated data into two subgroups. This division is also supported by the KL divergences of the three gaussian components: $KL(\theta_1, \theta_2) = 7.86, KL(\theta_1, \theta_3) = 33.76$ and $KL(\theta_2, \theta_3) = 15.96$, where the three components are denoted as 1, 2 and 3 from left to right.

5.2 Real data experiments

In this section, we give two experiments (Experiment 2 and Experiment 3) on two real medical meta datasets. The first experiment (Experiment 2) concentrates on the level difference of the hormone factor *cortisol* in rheumatoid arthritis (RA) and healthy population. The data summarized from 15 random controlled trials [9–15] (Figure 2 (a)). The experimental results are given in Figure 2 (b) and (c).

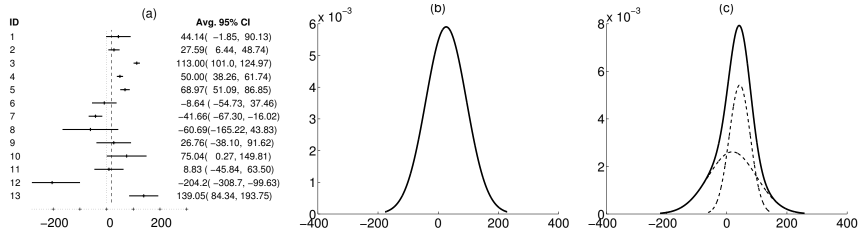

Fig. 2. Figures of Experiment 2. (a) the meta data where the short dash indicates the coordinate axes; the long dash shows the mean of data ;(b) and (c) results of random effect model and MREM respectively

It is found that the MREM with one gaussian component, which is equivalent to the random effect model, has the highest BIC score. The MREMs with multiple components e.g. the one shown in Figure 2(c), get smaller scores. This experiment shows that traditional random effect model is just a particular case of the proposed model. And the results illustrates the level of the hormone factor cortisol is not different between RA and healthy population. This conclusion complies with medical domain knowledge that inflammatory factor such as IL-6 is active in RA patients; while the hormone factor is not significant.

The data of the second real experiment (Experiment 3) is taken from [16] which summaries 90 randomized studies valuating the effect of *Nicotine Replacement Therapy* (NRT) on smoking cessation. Eliminating four incomplete data, we apply MREM to the remaining 86 effect sizes to find out whether the use of NRT successfully stops smoking and what its efficacy to different populations.

The result shown in Figure 3 suggests the best model is two components MREM, which presents more explicit information of underlying effect size compared to random effect model and it implies us to divide those studies into two subgroups for further study. The heterogeneity in each subgroup suggested by MREM is relatively equilibrium and their confidence intervals (CI) are listed in Table 1.

The CIs of two subgroups in Table 1 illustrate that the effect of NRT in Subgroup 2 is quite positive. While in Subgroup 1, the effect is minor, because the interval contains zero. This result of subgroup division obtained by MREM suggests an interesting direction for medical study. In fact, we find that Subgroup 1 mainly corresponds to women patients both in mid- and long-term follow-up and men patients in long-term follow-up; while Subgroup 2 mainly corresponds

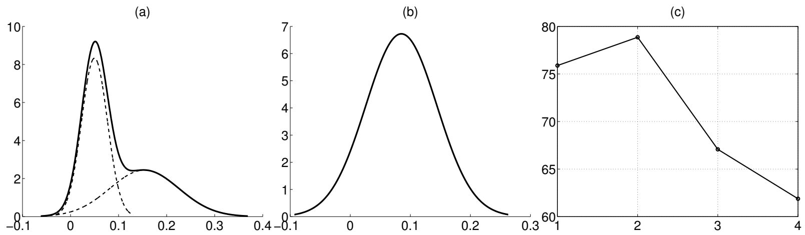

Fig. 3. Results of Experiment 3. (a) MREM with highest BIC scores. (b) estimated random effect model. (c) BIC score of models with different number of components

Table 1. The BIC scores and confidence intervals for Experiment 3

Experiment	BIC Score	95% CI
Whole data (random effect model)	75.9	(-0.0309, 0.201)
Whole data (MREM)	78.9	(-0.00874, 0.253)
Subgroup 1 (MREM)	86.2	(-0.00119, 0.0923)
Subgroup 2 (MREM)	26.8	(0.108, 0.305)

to women patients in short-term follow-up and men patients in both short- and mid-term follow-up. This division demonstrates the NRT effect is promising at short-term follow-up for all population and its efficacy becomes minor with the time lapse. Comparatively speaking, the long-term maintenance of NRT treatment gains decrease more rapidly for women than men according to our subgroup division. We find such phenomenon arising from our subgroup division consists with the medical knwoledge[17][18] that NRT is efficacious both in men and women at short-term follow-up while the abstinence-rate efficacy significantly decline at long-term follow-up especially for women who suffer more from smoking cessation, such as dysphoric or depressed mood, anxiety and weight gain associated with quitting cigarettes.

6 Conclusion

In this paper, we present a novel statistical model, the mixture random effect model, for summarizing distributed heterogeneous medical studies. The proposed model unifies the traditional meta-analysis tools, that is, the fixed effect model and random effect model are just two particular cases of it. The mixture random effect model has the ability to capture arbitrary complex distribution of the effect size and provides useful information for subgroup division without prior knowledge. We construct the model essentially by GMM, the parameters of which are estimated by MCMC approach. The novel model achieves prominent results in experiments on real clinical data, which demonstrate its potentially value for heterogeneous data analysis and medical data mining.

Acknowledgement

This work is supported by the National Natural Science Foundation of China (No. 30200365 and 60475001).

References

1. Whitehead, A.: Meta-Analysis of controlled Clinical Trials. John Wiley & Sons, New York (2002)
2. Sutton, A., Abram, K., Jones, D., Sheldon, T., Song, F.: Methods for Meta analysis in medical Research. John Wiley & Sons, New York (2000)
3. DuMouchel, W.H., Normand, S.T.: Computer modeling strategies for meta-analysis. In Stang, D., Berry, D., eds.: Meta-analysis in medicine and health policy. Marcel Dekker, New York (2000) 127–178
4. Dempster, A.P., Laird, N.M., Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B **39** (1977) 1–38
5. Hastings, W.K.: Monte carlo sampling methods using markov chains and their applications. Biometrika **57** (1970) 97–109
6. Carlin, B.P., Louis, T.A., Carlin, B.: Bayes and Empirical Bayes Methods for Data Analysis. Second edn. Chapman & Hall/CRC, Florida (2000)
7. Schwarz, G.: Estimating the dimension of a model. The Annals of Statistics **2** (1978) 461–464
8. Duda, R., Hart, P., Stork, D.: Pattern Classification. Second edn. John Wiley & Sons (2001)
9. Zolil, A., et al.: Acth, cortisol and prolactin in active rheumatoid arthritis. Clinical Rheumatology **21** (2002) 289–293
10. Rovensky, J., et al.: Cortisol elimination from plasma in premenopausal women with rheumatoid arthritis. Ann. Rheum. Dis. (2003) 674–676
11. Jing, L., et al.: Circadian variation of interleukin26 and cortisol in rheumatoid arthritis. Chin. J. Rheumatol **6** (2002) 252–254
12. Keith, S., et al.: Adrenocorticotropin, glucocorticoid, and androgen secretion in patients with new onset synovitis/rheumatoid arthritis: Relations with indices of inflammation. Journal of Clinical Endocrinology & Metabolism **35** (2000)
13. Dekkers, J., et al.: Experimentally challenged reactivity of the hypothalamic pituitary adrenal axis in patients with recently diagnosed rheumatoid arthritis. J. Rheumatol **28** (2001) 1496–504
14. Harbuz, M., et al.: Hypothalamo- pituitary- adrenal axis dysregulation in patients with rheumatoid arthritis after the dexamethasone/corticotrophin releasing factor test. J. Endocrinol **178** (2003) 55–56
15. Straub, R.H., et al.: Inadequately low serum levels of steroid hormones in relation to interleukin-6 and tumor necrosis factor in untreated patients with early rheumatoid arthritis and reactive arthritis. Arthritis Rheum. **46** (2002) 654–662
16. Cepeda-Benito, A., Reynoso, J., Erath, S.: Meta-analysis of the efficacy of nicotine replacement therapy for smoking cessation: Differences between men and women. Journal of Consulting and Clinical Psychology **72** (2004) 712–722
17. Perkins, K.: Smoking cessation in women: Special considerations. CNS Drugs **15** (2001) 391–411
18. Cepeda-Benito, A., Reig-Ferrer, A.: Smoking consequences questionnairespanish. Psychology of Addictive Behaviors **14** (2000) 219–230