journal homepage: www.intl.elsevierhealth.com/journals/cmpb

A note on relevance of diagnostic classification and rating scales used in psychiatry

Yiqun Gan [a], *Tamar Kakiashvili* [b], *Waldemar W. Koczkodaj* [c,*], *Feng Li* [c]

[a] *Peking University, China*
[b] *Sudbury Therapy, Canada*
[c] *Laurentian University, Computer Science, Canada*

ARTICLE INFO

Article history:
Received 26 November 2012
Received in revised form
13 April 2013
Accepted 23 May 2013

Keywords:
Scale
Relevance
Cluster analysis
Machine learning
Unsupervised learning

ABSTRACT

In the clinical practice of psychology and psychiatry, presence or absence of particular disorder or syndromes is based on the subjective interpretation of mental and behavioral descriptions offered by the patient. This is often done by questionnaires (also called instruments or scales) or by interviews. This subjectivity of the diagnostic decision-making process may limit the reliability of diagnosis. In the present study, a new method of scale relevance, based on double cluster analysis, is proposed as it is important to verify what we are trying to find with the proposed scale. If two data sets cluster differently, we must consider them as different.

1. Introduction

The relevance of diagnostic classification and rating scales used in psychiatry (called the *scale* in this study unless there is an ambiguity problem) is the process of measuring quantitative attributes or traits. It is widely used in psychiatry and psychology. The lack of tangible medical evidence (such as blood tests or X-rays) in psychology often forces clinicians to rely on scales. Such diagnostic uncertainty is perhaps no better demonstrated than in burnout. There is ongoing effort to develop methods for accuracy improvement and the reduction of uncertainty in the diagnosis of mental syndromes and disorders. In this study, we use burnout since it is one of the most uncertain mental conditions (not even recognized as a mental disorder).

According to [13], the term ‘burnout’ was first introduced over three decades ago [5]. Burnout is assumed to be an outcome of chronic stress. It is defined as a state of exhaustion combined with doubts about the value of one’s own work and competence [19]. As with most psychiatric disorders for modern society, burnout cannot be traced back to one single cause. It is linked to a variety of risk factors. Feelings of inadequate control over one’s work, frustrated hopes and expectations, and low levels of satisfaction seem to be independent contributors to job burnout.

Burnout is specified by the International Classification of Diseases (ICD-10, version 2007) under “Factors influencing health status and contact with health services (Z00 Z99)” as a “State of vital exhaustion” (Z73.0) and identified as a problem related to “life-management difficulty.” Burnout is not a recognized disorder in the Diagnostic and Statistical Manual

* *Corresponding author*. Tel.: +1 705 675 1151; fax: +1 705 673 6591; alphabetical order implies equal contribution.
E-mail addresses: wkoczkodaj@cs.laurentian.ca, wkoczkodaj@gmail.com (W.W. Koczkodaj).
0169-2607/$ – see front matter
http://dx.doi.org/10.1016/j.cmpb.2013.05.016

of Mental Disorders (DSM IV TR, 2000). The lack of reference to burnout in the DSM IV TR has important implications.

Burnout adversely influences normal functioning. According to [10], emotional exhaustion in teachers occurs when emotional resources are depleted and they feel that they can no longer give psychologically of themselves. To cope with stress and emotional exhaustion [6], they withdraw from their work and develop negative, cynical, and indifferent attitudes and feelings about their students (depersonalization) [7]. Burnout is considered as a psychological syndrome. To complicate a classification problem even more, burnout and depression have been shown to be strongly related, and this relationship raises the question of conceptual overlap and redundancy.

2. Background

Burnout is of particular significance because of its lack of pathognomonic features and vaguely defined diagnostic characteristics. Many mental and behavioral disorders are diagnosed according to core sets of criteria set out in the DSM-IV and ICD-10. For the Autism spectrum disorders (ASDs), there are three 'domains' (social interaction, communication, and repetitive behaviours), and each domain is further expanded to include four exemplary behaviours. One such behaviour in the social domain is 'difficulties forming and maintaining peer relationships', and within the communication impairment 'difficulties with to and from conversation' without any further guidance on how to decide whether such impairments are present or not. This may lead to significant diagnostic uncertainty, and is exemplified not only among the ASDs but for many other disorders described in similar terms in the DSM and ICD nosological systems (such as schizophrenia, anxiety disorders, mood and personality disorders). In [20], the author used the K-means cluster analysis to divide the observations into two groups. The author mentioned that using K-means cluster analysis identified a high burnout subgroup ($n = 46$, 47%) and a low burnout subgroup ($n = 52$, 53%).

3. The "Army tank" case

According to [17]:

> "...a neural network pattern recognition system that was being developed for the army to identify the presence of enemy tanks. Once trained, the system appeared to work perfectly, able to identify tanks in the testing samples and in a completely separate data set. When taken to the field, however, the system failed. After analysis, it was discovered that the system was actually identifying qualities of the pictures it was being presented with: every photo in the test set that had a tank hidden within it was taken on a cloudy day; coincidentally, every photo without a tank was taken on a clear day. The system had learned to identify cloudy skies and not tanks."

In our case, designing a scale for a psychiatric disorder or syndrome (say, burnout) is our true intention but it could be more relevant for diagnosing another psychiatric disorder or syndrome (say, depression). We propose the use of double cluster analysis for distinguishing two potentially "close cases". It is addressed in the next section.

Table 1 – The two indexes of MBI and BDI.

	Dunn index	Davies–Bouldin index
MBI	1.075065	1.84965
BDI	0.4548356	4.103642

4. Scale relevance

Let us assume that we have designed a new scale to detect a specific disorder or syndrome (say, the proverbial flu). However, this new scale may be more relevant for pneumonia. So, it is not really only a reliability problem although the low reliability would be the first and most important indication that the instrument is not doing what it is intended to do. It is also important to note that if we have clinical external validation or tests (e.g., blood), we can conduct the statistical analysis to verify the reliability of the instrument. In our case, we do not have external validation for burnout. Maslach Burnout Inventory General Survey (MBI-GS but used as MBI in the presentation) was introduced in [5]. MBI proponent assumed an arbitrary threshold but it is based on statistics, not an external validation since we are really unable to verify who has burnout or not. In fact, even distinguishing burnout from depression is not easy.

Cronbach's alpha statistic is widely used in the social sciences, business, nursing, and other disciplines. It is a measure of internal consistency, or reliability of a psychometric test score for a sample of examinees. In other words, Cronbach's alpha statistic is used for showing how closely related a set of items is as a group. We cannot use it since we have two groups. So, we propose a new and unique method to investigate how these two groups overlap, a double cluster analysis. There are two internal validation indexes: Dunn index and Davies–Bouldin index. In [8], Dunn index is defined as:

$$DI_m = \min_{1 \leqslant i \leqslant m} \left\{ \min_{1 \leqslant j \leqslant m} \left\{ \frac{d(c_i, c_j)}{\max_{1 \leqslant k \leqslant m} \Delta_k} \right\} \right\}$$

where $d(c_i, c_j)$ defines the intercluster distance between cluster X_i and X_j; $d(X_k)$ represents the intracluster distance of cluster (X_k) and c is the number of cluster of dataset. According to [8], Davies–Bouldin index is defined as:

$$DB = \frac{1}{c} \sum_{i=1}^{c} \max_{i \neq j} \left\{ \frac{d(X_i) + d(X_j)}{d(c_i, c_j)} \right\}$$

where c denotes the number of clusters, i and j are cluster labels, then $d(X_i)$ and $d(X_j)$ are all samples in clusters i and j to their respective cluster centroids, $d(c_i, c_j)$ is the distance between these centroids.

Table 1 shows Dunn index and Davies–Bouldin index in our case. External validation is traditionally done with the help of a confusion matrix. There are terminologies and derivations from confusion matrix (Table 2).

- Sensitivity or true positive rate (TPR) TPR = TP/P = TP/(TP + FN).

Table 2 – Confusion matrix

	Positive	Negative
Positive	True Positive (TP)	False Positive (MBI count: 0)
Negative	False Negative	True Negative (MBI count: 0)

- False positive rate (FPR) FPR = FP/N = FP/(FP + TN).
- Accuracy (ACC) ACC = (TP + TN)/(TP + TN + FP + FN).
- Specificity (SPC) or true negative rate SPC = TN/N = TN/(TN + FP) = 1 − FPR.
- Positive predictive value (precision) PPV = TP/(TP + FP).
- Negative predictive value (NPV) NPV = TN/(TN + FN).
- False discovery rate (FDR) FDR = FP/(FP + TP).

However, we do not have external validation for burnout. In our case, ROC graphs are another way besides confusion matrices to examine the performance of classifiers. A ROC graph is a plot with the false positive rate on the X-axis and the true positive rate on the Y-axis. In most practical cases, a classifier has a parameter that can be adjusted to increase TP at the cost of an increased FP or decrease FP at the cost of a decrease in TP. Each parameter setting provides a (FP, TP) pair and a series of such pairs can be used to plot an ROC curve. The parameter setting is usually done by setting the cut off (threshold) point for true/false when the decision attribute (in our case: 'has' or 'has not' a particular disorder or syndrome) exists but burnout data lack it. If we set it to false at a given point and to 1 below it, the AUC (area under the curve) value of ROC is 1 which is unrealistic. In case of the MBI, the arbitrary threshold results in both false positives and false negatives are zero hence the AUC is always 1. This is not how scales work in clinical practice. We know that there is no magical "cut off" point for Beck Depression Inventory (BDI) [2]; some patients with the high score do not have depression and some patients with the low BDI the score have depression.

5. Measures

We propose the use of scale relevance as a new way for psychiatric measures. The MBI is the most recognized scale for burnout. According to [13], the three-factorial validity of the complete measure has been reported as reliable for different occupations [4,14,15,22]. The items were scored on a 7-point frequency rating scale ranging from 0 (never) to 6 (daily). High scores on exhaustion and cynicism and low scores on professional efficacy are indicative of burnout. The items of professional efficacy were reversed (lack of professional efficacy) [11,12,18]. According to [13], the original Beck Depression Inventory was used in [2,3] for measuring depression.

6. Data set

Participants included 413 faculty members from all colleges and departments in one university in Mainland China. The age ranged from 22 to 79 years (M age = 41.73 years, SD = 9.304). Each participant completed the two questionnaires (MBI and BDI) at the same time. During the data preparation process, 99 records from the 512 records had to be removed, because these records had one missing value in both the MBI scale and the BDI scale. Finally, a total of 413 records were included in the analysis. According to [13], the BDI consists of 21 items that are scored from 0 to 3. At least 14 items must be answered.

7. Description of the method used

Supervised learning can be used for discovering patterns in the data that relate data attributes to a class attribute. It means that the class attribute (or decision attribute) must be present. These patterns are then utilized to predict the values of the class attribute of new data [16]. For example, the improved scoring system [21] was used to predict the autism spectrum disorders in children based on the data with the class attribute 'has' or 'has not'. However, this approach cannot be used for the data without the class attribute. Clustering can be used for data without class attribute. It organizes data into similarity groups which are called clusters. The data in the same cluster are somehow similar to each other and data in different clusters are as different as possible from each other. Clustering is often called unsupervised learning, because unlike supervised learning, class values denoting an a priori partition or grouping of the data are not given [16].

According to [1], cluster analysis divides data into groups (clusters) that are meaningful and useful. If the goal is to create meaningful groups, then the clusters should capture the natural structure of the data. As the name clustering indicates, discovering a structure takes place by exploring similarities or differences (such as distances) between individual data points in a data set under consideration. This highly intuitive and appealing guideline sounds deceptively simple: we place two points in the same cluster and keep doing it by examining the distances between newly formed clusters and the remaining data points [9]. Clustering techniques can be divided into three main categories [9]:

1. Partition-based clustering, sometimes referred to as objective function-based clustering.
2. Hierarchical clustering.
3. Model-based (a mixture of probabilities) clustering.

One of the best-known partition clustering algorithm or objective function-based clustering algorithm is the K-means algorithm [9,16]. It is also the most widely used among all clustering algorithms due to its simplicity and efficiency. For a given set of data points and the required number of k clusters (k is specified by the user or computed), this algorithm iteratively partitions the data into k clusters based on a distance function [16] by using:

1. distance measures: Euclidean, MWSS, gamma, Pearson, R-squared, Minkowski, chi-square, phi-square, absolute, Mahalanobis and
2. additional options to specify the covariance matrix for computing the Mahalanobis distance.

The initial seeds can be specified from: none, first, last or random k, random or hierarchical segmentation, principal component, partition variable, from file.

The double cluster analysis could be summarized as:

1. clustering each dataset D1 and D2 separately,
2. comparing the obtained cluster numbers of D1 with cluster numbers of D2, and
3. analyzing the overlap of the clusters from the former action.

The double cluster analysis is very efficient and will be posted on the Internet as an open source.

8. Status report

In this study, we have successfully tested the overlapping of burnout (Maslach Burnout Inventory) and depression (Beck Depression Inventory) collected by Zhou, You and Gan [12]. Both MBI and BDI are reliable and valid scales. There are 413 records hence the statistical accuracy is better than 95% so the results are reliable. A K-means cluster analysis is forming 2 clusters on all 413 subjects. The MBI cluster 1 consists of 204 participants, and the MBI cluster 2 consists of 209 participants (Fig. 1). The BDI cluster 1 consists of 280 participants and the BDI cluster 2 consists of 133 participants (Fig. 2). The overlapping cardinality is 160 for both clusters. The cardinality of rest of observations is 253 (Fig. 3). The percentage of two clusters in MBI and BDI is demonstrated by Fig. 1. The overlapping between MBI and BDI is shown in Fig. 3.

In a K-means cluster analysis, MBI required 7 iterations to find the closest centroid(the center of each item in the cluster), while the BDI required 14 iterations. While the iteration was processing, the centroid was changing as well. For the MBI and BDI scales, using K-means cluster analysis divides the data into two clusters as illustrated by (Figs. 4 and 5). After the K-means cluster analysis, for each cluster, average value is calculated for each scale items. As expected, the graph (Figs. 6 and 7) shows different average values for most scale items in separate clusters. The graph in Fig. 6 shows the each item score average number in MBI. As far as we are aware, this is the first study to examine the difference between MBI and BDI using the double cluster analysis. Our results confirm what was reported as overlapping between burnout and depressive disorders.

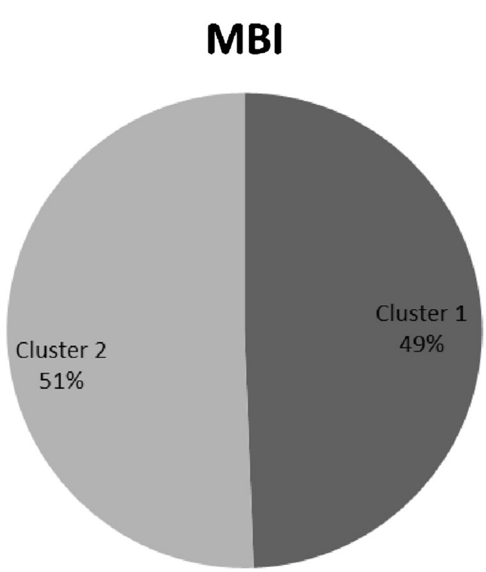

Fig. 1 – MBI cluster.

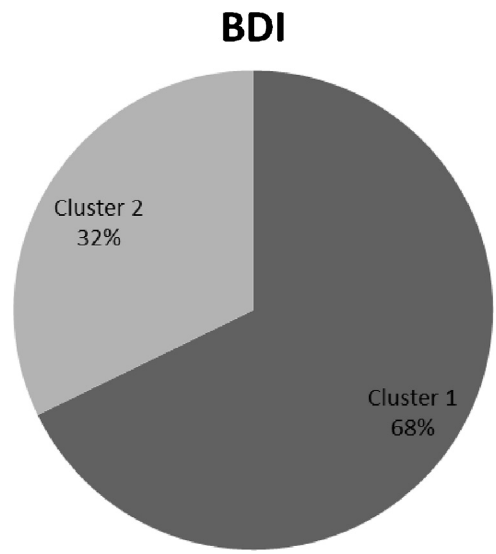

Fig. 2 – BDI cluster.

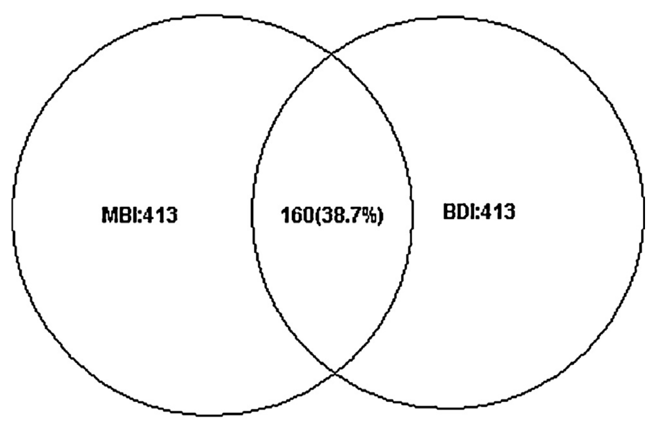

Fig. 3 – MBI–BDI overlap.

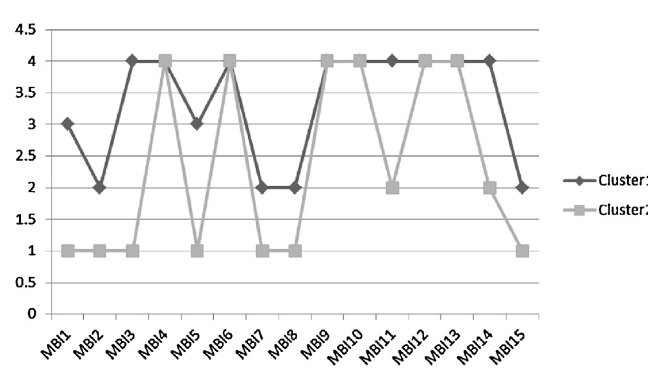

Fig. 4 – MBI cluster centriod.

Fig. 8.

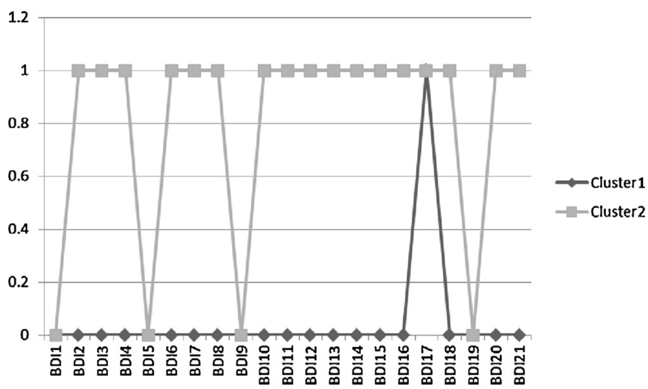

Fig. 5 – BDI cluster centriod.

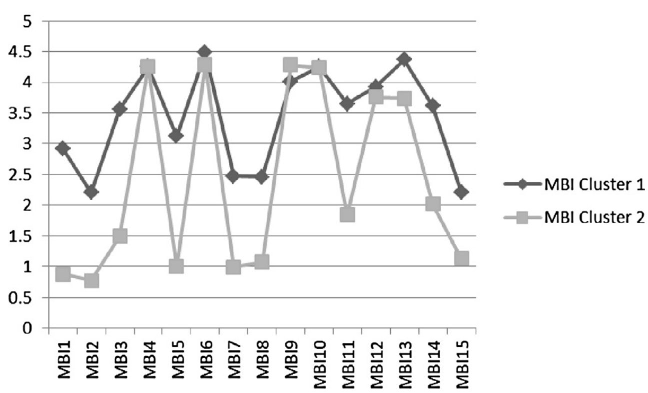

Fig. 6 – Cluster mean MBI.

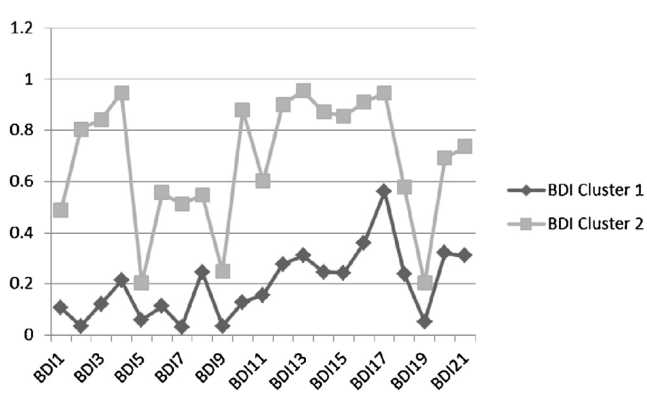

Fig. 7 – Cluster mean BDI.

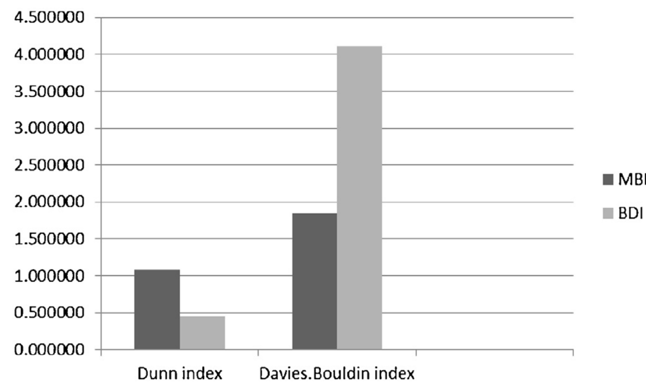

Fig. 8 – Two indices.

9. Lessons learned

Distinguishing scales is not easy when we are unable to validate at least one of them. In our case, it is the burnout scale. It is so since we are unable to reliably establish whether or not a person has burnout. Relying on the MBI alone is certainly inappropriate since we know that no medical or psychiatric scale may claim 100% accuracy and that only external (clinical) validation can decide (also with error but hopefully smaller) whether or not a patient has a certain syndrome or disorder. In case of depression, clinicians (psychiatrists, psychologists, and general practitioners) can reliably diagnose patients for depression. In fact, often a glance at a person is sufficient to conclude whether or not he or she has depression, although no one really does it this way.

10. Future plans

It is important to stress that the presented method can be applied to other medical and psychological scales. Clinicians often struggle with the far-from-perfect diagnostic criteria to diagnose mental and behavioral syndromes. We know that some psychiatric disorders may be similar to each other. We plan to use double cluster analysis for PTSD (post-traumatic stress disorder) and depression.

Certainly, the double-cluster analysis method needs further validation on more clinical data before its wider clinical applicability can be recommended. We plan to use it on data from the Peking University (coming soon but publication clearance is needed). In short, more studies for other scales such as the one described here are urgently needed. The diagnostic process of developmental disorders, depression, bipolar disorder, early schizophrenia, and many other mental disorders relies heavily on medical scales and most of them can be improved by the presented method. We also plan to develop Java programs for the double cluster analysis and post it on the Internet in the nearest future.

Conflict of interest

None of the authors has a conflict of interest.

REFERENCES

[1] http://www-01.ibm.com/software/analytics/spss/ (accessed 16.11.12).

[2] A.T. Beck, C.H. Ward, M. Mendelson, J. Mock, J. Erbaugh, An inventory for measuring depression, Archives of General Psychiatry 4 (1961) 561–571.

[3] A.T. Beck, R.A. Steer, M.G. Garbin, Psychometric properties of the Beck depression inventory: twenty years of evaluation, Clinical Psychology Review 8 (1988) 77–100.

[4] M.P. Leiter, W.B. Schaufeli, Consistency of the burnout construct across occupations, Anxiety Stress Coping 9 (1996) 229–243.
[5] C. Maslach, W.B. Schaufeli, M.P. Leiter, Job burnout, Annual Review of Psychology 52 (2001) 397–422.
[6] C. Cherniss, Staff Burnout: Job Stress in the Human Services, Sage Publications, Beverly Hills, CA, 1980.
[7] C. Cherniss, Stress burnout and the special services provider, Special Services in the Schools 2 (1) (1985) 45–61.
[8] E. Rendn, I. Abundez, A. Arizmendi, E.M. Quiroz, Internal versus External cluster validation indexes, International Journal of Computers and Communications 5 (2011) 27–34.
[9] G. Hinton, Unsupervised Learning: Foundations of Neural Computation, MIT Press, Cambridge, Massachusetts, London, England, 1999, ISBN 0-262-58168-X.
[10] H.K. Jennett, S.L. Harris, G.B. Mesibov, Commitment to philosophy, teacher efficacy, and burnout among teachers of children with autism, Journal of Autism and Developmental Disorders 33 (2003) 583–593.
[11] J. Zhong, J.N. You, Y.Q. Gan, Y.W. Zhang, C.Q. Lu, H.B. Wang, Job stress, burnout, depression symptoms and physical health among Chinese University teachers, Psychological Reports 105 (3) (2009) 1248–1254.
[12] J. Zhong, J.N. You, Y.Q. Gan, Y.W. Zhang, C.Q. Lu, H.B. Wang, The relationship of job stress, burnout, depression and Physical health among Chinese university teachers, Psychological Reports (2009).
[13] K. Ahola, T. Honkonen, M. Kivimaki, M. Virtanen, E. Isometsa, A. Aromaa, J. Lonnqvist, Contribution of burnout to the association between job strain and depression: the Health 2000 study, Journal of Occupational and Environmental Medicine 48 (2006) 1023–1029.
[14] N. Schutte, S. Toppinen, R. Kalimo, W. Schaufeli, The factorial validity of the Maslach burnout inventory-general survey (MBI-GS) across occupational groups and nations, Journal of Occupational and Organizational Psychology 73 (2000) 53–66.
[15] T.W. Taris, P.J.G. Schreurs, W.B. Schaufeli, Construct validity of the Maslach burnout inventory-general survey: two sample examination of its factor structure and correlates, Work Stress 13 (1999) 223–237.
[16] D.M. Skapura, Unsupervised Learning and Clustering, Wiley, New York, 1999, pp. 223–237, ISBN 0-471-05669-3 13.
[17] D.M. Skapura, Building Neural Networks, Addison-Wesley Publishing Co., New York, USA, 1995.
[18] Y.Q. Gan, Y.W. Zhang, X. Shen, The coping flexibility of neurasthenia and depressive patients, Personality and Individual Difference 40 (5) (2006) 859–871.
[19] C. Maslach, S.E. Jackson, M.P. Leiter, Maslach Burnout Inventory Manual, 3rd ed., Consulting Psychologist Press, Palo Alto, 1996, pp. 19–26.
[20] M.E. Dahlin, B. Runeson, Burnout and psychiatric morbidity among medical students entering clinical training: a three year prospective questionnaire and interview-based study, BMC Medical Education 2007 (2007), 7:6.
[21] T. Kakiashvili, W.W. Koczkodaj, M. Woodbury-Smith, Improving the medical scale predictability by the pairwise comparisons method: evidence from a clinical data study, Computer Methods and Programs in Biomedicine (2011).
[22] P.J.G. Schreurs, T.W. Taris, Construct validity and the demand control model: a double cross validation approach, Work Stress 12 (2011) 66–84.