Novel Two-Stage Analytic Approach in Extraction of Strong Herb-Herb Interactions in TCM Clinical Treatment of Insomnia

Xuezhong Zhou[1], Josiah Poon[2], Paul Kwan[3], Runshun Zhang[4], Yinhui Wang[4],
Simon Poon[2], Baoyan Liu[5], and Daniel Sze[6]

[1] School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China
[2] School of Information Technologies, University of Sydney, Sydney, Australia
{josiah.poon,simon.poon}@sydney.edu.au
[3] School of Science and Technology, University of New England, Armidale, Australia
[4] Guanganmen Hospital, China Academy of Chinese Medical Sciences, Beijing, China
[5] China Academy of Chinese Medical Sciences, Beijing, China
[6] Department of Health Technology and Informatics,
Hong Kong Polytechnic University, Hong Kong

Abstract. In this paper, we aim to investigate strong herb-herb interactions in TCM for effective treatment of insomnia. Given that extraction of herb interactions is quite similar to gene epistasis study due to non-linear interactions among their study factors, we propose to apply Multifactor Dimensionality Reduction (MDR) that has shown useful in discovering hidden interaction patterns in biomedical domains. However, MDR suffers from high computational overhead incurred in its exhaustive enumeration of factors combinations in its processing. To address this drawback, we introduce a two-stage analytical approach which first uses hierarchical core sub-network analysis to pre-select the subset of herbs that have high probability in participating in herb-herb interactions, which is followed by applying MDR to detect strong attribute interactions in the pre-selected subset. Experimental evaluation confirms that this approach is able to detect effective high order herb-herb interaction models in high dimensional TCM insomnia dataset that also has high predictive accuracies.

Keywords: Multifactor Dimensionality Reduction, Features Selection, Traditional Chinese Medicine, Insomnia, Network Analysis, Herb Interactions.

1 Introduction

Insomnia is a “nightmare” to many people. It refers to the poor quality of sleep or insufficient sleep duration; it can lead to adverse daytime consequences. Studies have shown that rates of insomnia with associated daytime dysfunction in attendees of general medical practices are high in both the western countries [1] and Asia region [2-4]. It is also common in many different morbid conditions [5] and it could be a risk factor to other diseases, such as depression and anxiety and further studies also find that there are genetic link between insomnia and depression/anxiety [6]. Therefore, insomnia is a

D. Zhang and M. Sonka (Eds.): ICMB 2010, LNCS 6165, pp. 258–267, 2010.

popular, important and vital disorder that needs to be addressed. However, the treatment of insomnia still remains particularly challenging for clinicians because of the lack of guidelines and the small number of studies conducted in patient populations with behavioural and pharmacologic therapies [7]. In the world especially in Asian region, a significant proportion of patients with insomnia consume herbal hypnotics regularly [8, 9]. Also alternative therapies, herbal products and other agents with sedative hypnotic effects are being increasingly sought after by the general population in the world [10-12].

Traditional Chinese medicine (TCM) has a long history and has been accepted as one of the main medical approaches in China [13]. It has also been successfully applied to the treatment of insomnia. Some of these well-known Chinese formulae include Baixia Tang [14], Suan Zaoren Tang [15] and Wendan Tang [16]. There are clinical studies have shown that the TCM herbal treatment for insomnia is effective [17, 18]. Since a classical TCM formula may be modified according to individual patient, it is necessary to find the common and true effective herb patterns (e.g. herb combinations) that actually facilitate the insomnia treatment. Herb prescription (also named formula in TCM field) is one of the main TCM therapies for various disease treatments. A specific formula constitutes of multiple herbs with appropriate dosages. Therefore, when huge number of clinical encounters occurred, they could generate large-scale formulae in the clinical practice. The herb combination patterns used in the herb prescriptions are both theoretically focused and important for effective TCM treatment. It would be significant to find the effective herb combination patterns; we called them *strong herb-herb interactions*, from the large-scale clinical data of herb prescriptions.

We propose in this paper a two-stage analytical procedure by which (1) the subset of herbs that have a high probability in participating in herb-herb interactions is preselected using hierarchical core network analysis (HCNA) method [19], and (2) the set of pre-selected herbs will constitute the input into multifactor dimensionality reduction (MDR) [20] to detect attribute interactions of order even up to 9, 10. The benefits of our approach are two-fold. First, it overcomes the computational bottleneck in applying MDR to detect attribute interactions of a higher order such as 9, 10, against a high dimensional dataset such as those appearing in TCM. Second, by combining the analysis of the hierarchical core sub-network and the result of MDR, "core" attribute interactions that would otherwise be unable to reveal due to the flat nature of the output of MDR can be made known.

2 Methods

2.1 Two-Step Approach

To study interaction, MDR has been demonstrated to be an effective tool to identify/understand the gene-gene and gene-environment epistasis [21]. It is based on the general idea of attribute induction in data mining research to discover possible non-linear interactions among genes. However, using MDR, it is still time-consuming in analyzing high-dimensionality dataset to discover higher order interactions using exhaustive search method. One way to address this drawback would be to reduce the dimensionality of the dataset prior to the analysis using MDR.

In TCM clinical datasets such as the insomnia that we are studying in this paper, the dimensionality is often high, above 100 dimensions (in our case, we have 261 dimensions). Also, the TCM physicians often prescribe formulae with 10-20 herbs. This means that the higher order interaction may exist in the prescriptions. We need to run over 4-5 months to perform such analysis (9 variable model with 261 dimensions) to get the best models on HP Proliant ML350 server with Intel Xeon(R) CPU (2.33GHz) and 8GB Memory.

Although MDR is the method of choice for detecting higher order interactions (i.e., beyond main effects and pairwise interaction) among herb-herb combinations in the analysis of the insomnia dataset (or in general, TCM dataset), the high dimensionality of the dataset makes it prohibitive to analyze beyond 4-5 attributes interaction. On the other hand, our recent result [19, 22] on analyzing common herb pairs frequently appearing in regular TCM herb prescriptions using HCNA confirmed that there often exists a core herb network, which comprises a subset of herbs appearing in prescriptions treating the disease concerned in the dataset. Herbs that appear in this core sub-network are expected to have a high probability of participating in herb-herb interactions.

MDR is a still a time consuming approach to find the combination patterns with outcomes. The high dimensionality of herbs (e.g. several hundreds) and the lengthy resulting pattern (e.g. more than ten herbs) of herb combination are still hurdles to apply this tool. Since HCNA can generate the important herb combinations, which are mostly smaller than the dimensionality of the whole herbs dataset, we propose to use HCNA as a filter before MDR analysis.

In order to minimize the computational overhead to extract the strong herb-herb interactions, the high dimensionality of herbs in the original prescriptions is first reduced by a hierarchical core sub-network. This analysis can generate a subset of important herbs (20-30 herbs), which is substantially smaller than the original dimensionality. The result from this extraction step will then serve as input to the MDR software in the next step, which can report a list of strong interacting herbs combination. Hence, in our proposal, it is a two-stage algorithm that uses HCNA to pre-select the important herbs and uses MDR to find the strong herb-herb interactions. Figure 1 is the system overview of this data mining process.

2.2 Hierarchical Core Sub-network Analysis

Social network analysis (or complex network analysis) has penetrated to various scientific fields, especially the biology and medical sciences [23]. Scale-free network is one of the distinguished characteristics of complex networks [24], and of which the degree distribution follows as the power law. This further implies that there are nodes that serve as hubs in a large-scale scale-free network. We have shown that from the related work on TCM herb network, the weight distribution of the herb network follows the power law. This indicates that there are common herb pairs frequently used in the regular TCM herb prescriptions. We have developed a complex network analysis system to model and analyze TCM clinical data [25]. A hierarchical core sub-network analysis method is deployed to retrieve the multi-level core herb combination patterns for insomnia treatment.

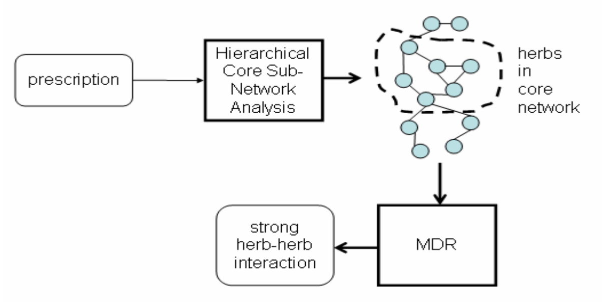
prescription
Hierarchical Core Sub-Network Analysis
herbs in core network
MDR
strong herb-herb interaction

Fig. 1. System overview

The basic formula is usually modified to accommodate individual differences. The main modifications make up of a core herb combination structures with hierarchical levels. In our HCNA, the hierarchical core herb combination structures are discovered based on degree distribution feature of scale-free network. The value of α is assigned near the power law exponent of node weight distribution in herb combination network. The number of hierarchical levels can be manually controlled by researchers. Usually the core sub-networks with 2-4 hierarchical levels will be clinical significant and easily be interpreted. We set the default value of the number of hierarchical levels as 3 to extract the three levels of core herb combination sub-networks.

2.3 Multifactor Dimensionality Reduction (MDR)

MDR [20, 21] has 7 steps and was initially designed to detect and to interpret gene-gene interactions. It utilizes information theory for selecting interesting single nucleotide polymorphisms (SNPs). It is a constructive induction approach to collapse high-dimensional genetic data into a single dimension.

3 Data Set

A clinical data warehouse was developed [22] to integrate and to manage large-scale real-world TCM clinical data. This data warehouse consists of structured electronic medical record from all TCM clinical encounters, including both inpatient and outpatient encounters. There are about 20,000 outpatient encounters of the TCM expert physicians. These encounters included clinical prescriptions for the treatment of various diseases, in which insomnia is a frequently treated disorder.

We have selected and labeled 460 insomnia outpatient encounters. The outcome of each encounter was annotated by TCM clinical experts who went through the changes of the insomnia-related variables over consecutive consultation; these include the sleep time per day, sleep quality and difficulty in falling asleep. The outcomes are then classified into two categories: good and bad. When a treatment was effective, which means that if the patient recovered completely or partly from insomnia in the next encounter, then the prescription of the current encounter would be categorized as ‘good’; otherwise, the herb prescription would be categorized as ‘bad’.

After labeling these 460 outpatient encounters, there are 68 encounters with bad outcomes in this dataset; in other words, it is an imbalanced dataset to the advantage of the target class. The average good outcome rate (GOR) of the whole data set is 392/460=85.21%. There are 261 distinct herbs in the dataset and there are on average 14 herbs in a formula. The insomnia data set is then transformed into the formats for hierarchical core sub-network analysis as well as for MDR methods. In the transformed dataset for MDR analysis, the value of a herb variable is set to '1' if the current encounter includes the corresponding herb, else it is set as '0'. Also, the value of outcome variable is set to '1' for good outcome and '0' for bad outcome.

4 Results

4.1 Extraction of the Hierarchical Core Herbs

We use the 460 herb prescriptions to construct a herb network with the herbs as nodes and the herb combinations as edges. The result herb network has 8,335 edges. Using the complex network analysis system [25], the HCNA method was applied to extract the top three level core sub-networks. Figure 2 shows the top three levels of core herb combinations. The degree coefficient was chosen as 2.0 to generate these results. A total of 39 core herbs and their corresponding combinations are shown in these figures. It shows that the herbs like *stir-frying spine date seed*, *grassleaf sweetflag rhizome*, *prepared thinleaf milkwort root*, *Chinese angelica*, *oyster shell* and *dragon bone* are used frequently in a combinatorial way for insomnia treatment. The extracted herbs are commonly found in the classical formulae and are frequently used by TCM physicians in insomnia treatment [14, 15, 17, 18]. The result herbs become a filter set with clinical meanings.

4.2 Finding the Strong Herb-Herb Interaction by MDR

We filtered the herb dataset with only 39 herb variables generated by the HCNA step. Then we use the MDR software (MDR 2.0 Beta 6) to find the strong herb-herb interactions for insomnia treatment. Using 10-fold cross evaluation, we select the exhaustive search type and track one top model to get the best model with 9 herb variables. The task took about 83 hours on a PC with Intel Core(TM) 2 Quad 2.66GHz CPU and 4G memory. A tremendous saving in computation cost was obtained with this two-stage approach as it is less than 1/1000th of the original effort, i.e. $C(39,9)$ as compared to $C(260,9)$. The training balanced accuracy of the 9 herb variables model is 90.89%. The accuracy with the test instances is 63.13% and the cross-validation (CV) consistency is 3/10. The best model includes the herb variables: VAR33 (*indian bread*), VAR34 (*golden thread*), VAR40 (*grassleaf sweetflag rhizome*), VAR113 (*fresh rehmannia*), VAR117 (*dried tangerine peel*), VAR174 (*Chinese angelica*), VAR196 (*Chinese date*), VAR203 (*white peony root*) and VAR237 (*stir-frying spine date seed*) as the main prediction factors.

Figure 3 depicts the interaction map of the herb variables. The regular boxes are the herbs while the number in each box is its relationship to the outcome; in other words, a positive percentage indicates a positive correlation with the outcome. The edges in this figure represent the interaction, and the percentage on the line is the interaction effect. The thickness of the line is a graphical representation of this effect; the thicker the line, the higher the interaction it is. The figure shows that 5 herbs (enclosed in black circles): VAR235 (*prepared thinleaf milkwort root*), VAR237, VAR196, VAR79 (*cassia bark*) and VAR40 have the independent main effects on the outcome. Furthermore, it shows significant synergic interaction effects between several herb variables which are denoted by the dark thick dotted lines. For example, significant synergic interaction effects between VAR235 and VAR117, VAR76 (*bamboo shavings*) and VAR117, VAR235 and VAR196, and VAR210 (*prepared pinellia tuber*) and VAR117. Also there are strong redundant interactions between several herb pairs like VAR237 and VAR235. The entropy-based interaction map proposes a good approach to show the significant main effect and interaction effect of herbs for outcomes. The discovered strong herb interactions have a significant support for the necessity of multi-herb prescriptions in TCM treatment.

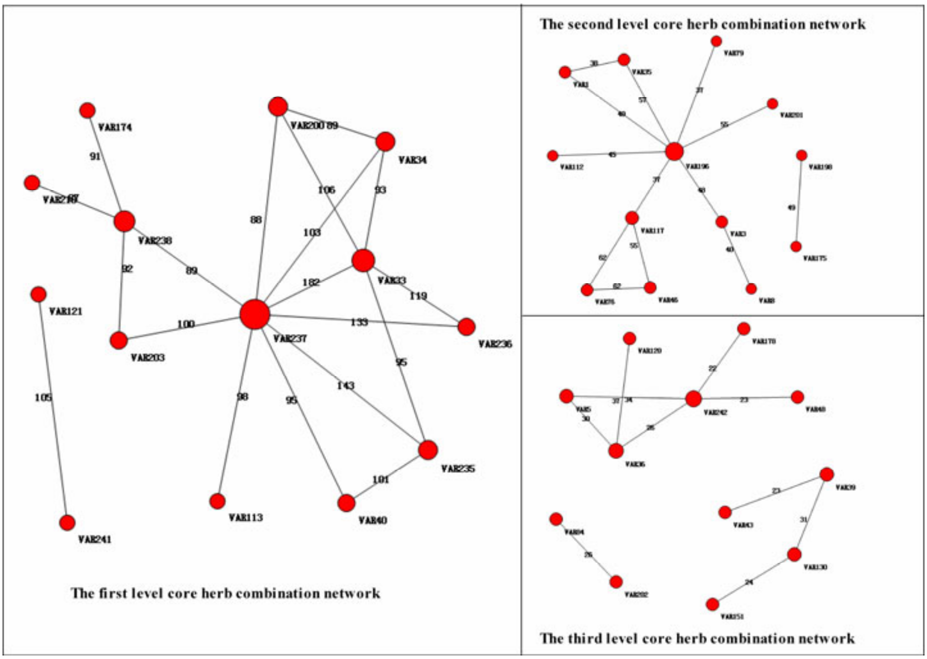

Fig. 2. The top three levels of core herb combination network

To find the final strong herb-herb interaction with good outcome or bad outcome, we could query the whole data set using the significant herb interactions suggested by the result model. The herb pairs with synergic interactions are more useful for finding the final strong effective or *non*-effective herb-herb interactions. Based on the synergic interactions suggested by the interaction map, we have tested the interesting herb combinations with GORs. Table 1 displays the GORs of some of the interesting herb combinations. It clearly shows that there are nonlinear interaction (like 'XOR' function) between VAR235 and VAR196. The single use of each herb has high GOR

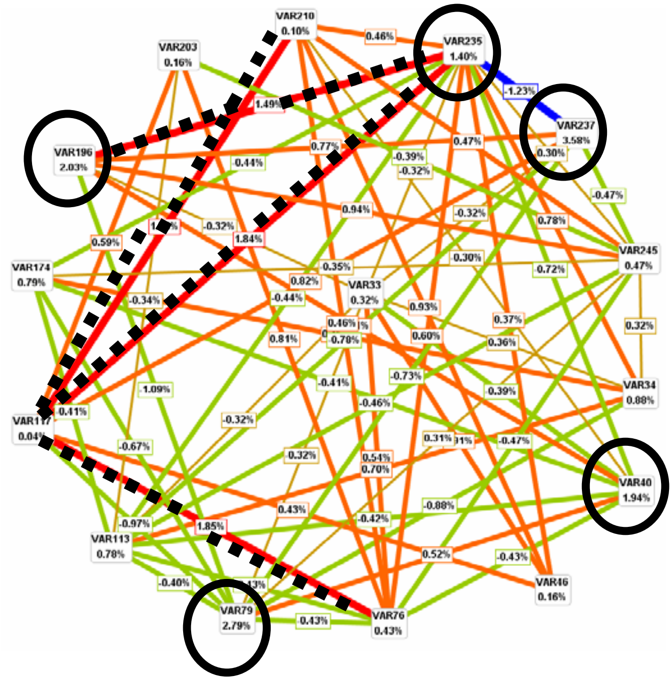

Fig. 3. The interaction graph for the herb variables based on entropy-based measures of information gain. To have a clear view of the graph, we filter out the edges and nodes whose information gain value is smaller than 0.3%. The red lines show positive interaction effects between herb variables. While the blue lines show negative interaction effects between herb variables. The larger of the edge sizes the stronger the interaction between the associated two variables of the edges.

(VAR235:92.80%; VAR196:96.70%), however, the GOR (87.80%) of the combined use is only little higher than the average GOR (85.21%) of the whole data set. Also the GOR is low (74.87%) while both of them are not used. This means that both of the two herbs are significant for effective insomnia treatment, however, there may exist negative interaction between them since while the GOR of combined usage is clearly lower than the GOR of mutually exclusively use of them. It is similar with the interaction between VAR117 and VAR76 except that this pair is a combination to avoid since the GOR of combined usage is largely lower than the average GOR. There are redundant interactions between the herbs, e.g. (VAR235, VAR237) where the second herb covers most of the outcome contributions. This kind of interaction information is both useful for theoretical research and clinical practice of herb combinations for insomnia treatment. From Table 1 (the other records are not listed owing to the page limit), we could find the significant (effective or *non*-effective) herb combinations with strong herb-herb interactions for insomnia treatment, e.g. some herb combinations like (VAR237, VAR203, VAR34) has 97.87% GORs with 46 frequency. However, the herb combination: (VAR46 [*stir-frying immature orange fruit*], VAR76, VAR117, VAR210) has bad outcome, which has a 72.73% GOR with 44 frequency.

Table 1. The example herb combinations with clinical usefulness. The value '01' in the combinations field means the condition that is not use of the former variables in the herb variables field and use of the latter variables.

Herb variables	Combinations	Total frequency	Good outcome frequency	Good outcome rate
VAR235,VAR196	11	41	36	87.80
	01	91	88	**96.70**
	10	125	116	**92.80**
	00	203	152	74.87
VAR76,VAR117	11	62	48	77.42
	01	44	44	**100.00**
	10	25	21	84.00
	00	329	279	84.80
VAR235,VAR237	11	143	132	**92.31**
	01	114	105	**92.11**
	10	23	20	86.96
	00	180	135	75.00

4.3 Comparison of the Different Filters

The HCNA method was included as a filter for herb interaction detection by the MDR. The training balanced accuracy (87.01%) of HCNA filter is slightly better than the other filters (SURF*TuRF, OddsRatioF, ReliefF, χ^2F, TuRF, SURF). Although it has lower testing balanced accuracy than several filters, HCNA has the advantages of maintaining the multi-level core herbs in the filtered attribute set. This makes sure that the different levels of core herb variables are retained in the result, which is practically used in TCM herb prescriptions according to the organizational principle of TCM formula theories, such as master-deputy-assistant-envoy (君臣佐使) [26], which indicate that the herbs were used as different roles to form a systematic formula for disease treatment. Thus, HCNA proposes a feasible choice as filter to select the herb variables from the original high dimensional variable set.

5 Discussion and Conclusion

The herb prescriptions in the TCM clinical data consist of multiple herbs and have complicated interactions between different herbs for disease treatment. There are TCM herb combinatorial theories like seven features of herb compatibility (SFHC) [27] to illuminate the combinatorial use of herbs. The SFHC theory says that there are seven kinds of herbs used in prescriptions, namely "going alone", "mutual reinforcement", "assistant", "detoxication", "restraint", "inhibition" and "antagonism". For example, the "inhibition" interaction between two herbs should be avoided in the clinical prescriptions since one herb would inhibit the effect of another herb. Thus, it is significant to find the strong non-linear herb-herb interactions (corresponding to the seven kinds of herb compatibility) in the clinical herb prescriptions, as it would be both useful for drug development and clinical guideline generation.

In this paper, we have proposed a two-stage analytical approach to obtain strong herb-herb interactions of order up to 9 from a TCM clinical dataset of insomnia. Stage 1 pre-selects a subset of the original herbs appeared in the dataset that has a high probability of participating in herb-herb interactions using HCNA method, while stage 2 applies MDR on the pre-selected set of herbs to detect the required attribute interactions. The results show that there exist non-linear herb-herb interactions in the herb prescription for insomnia treatment. Furthermore, the effective or non-effective herb combinations could be found by the proposed two-stage approach and by querying the whole data set. The entropy-based herb interaction map proposes a feasible visualization of the herb compatibility. It is interesting that some types of the herb compatibility, such as the "inhibition", could be recognized as synergic interaction in the model. These kinds of results have very high value to be investigated by further clinical studies.

Experimental results confirmed that our approach is able to detect a 9-order herb interactions with acceptable accuracy, while having the distinct advantage of maintaining the multi-level core herbs in the interaction result when compared to other attribute filtering methods.

Although this paper introduces a computing approach for finding the strong herb-herb interactions and the useful herb combinations in the context of clinical outcomes. Several key issues still exist to be addressed in the future work. Firstly, a straightforward approach to detect the strong herb interactions is needed as the querying of the whole data set could not systematically find the top *K* strong herb combinations. Secondly, the computing cost of MDR to find the optimal or hyper-optimal herb variable models should be further studied. Thirdly, two important information components, namely herb dosage and clinical manifestation (e.g. symptoms, co-morbid conditions), of the clinical data could be included in the data set. Because it is widely recognized in medical field that the herb dosage has an important effect for treatment and also performs a significant role for herb-herb interactions.

Acknowledgement

This work is partially supported by Program of Beijing Municipal S&T Commission, China (D08050703020803, D08050703020804), China Key Technologies R&D Programme (2007BA110B06-01), China 973 project (2006CB504601), National Science and Technology Major Project of the Ministry of Science and Technology of China (2009ZX10005-019), and the S&T Foundation of Beijing Jiaotong University (2007RC072).

References

1. Sateia, M.J., Nowell, P.D.: Insomnia. Lancet 364, 1959–1973 (2004)
2. Xiang, Y., Ma, X., Cai, Z.: The Prevalence of Insomnia, Its Sociodemographic and Clinical Correlates, and Treatment in Rural and Urban Regions of Beijing, China. A General Population-Based Survey. Sleep 31, 1655–1662 (2008)
3. Cho, Y.W., Shin, W.C., Yun, C.H.: Epidemiology of insomnia in Korean adults: prevalence and associated factors. Journal of Clinical Neurology 5, 20–23 (2009)
4. Doi, Y.: Epidemiologic research on insomnia in the general Japanese populations. Nippon Rinsho. 67, 1463–1467 (2009)
5. Pien, G.W., Schwab, R.J.: Sleep disorders during pregnancy. Sleep 27, 1405–1417 (2004)

6. Gehrman, P.: Heritability of insomnia in adolescents: How much is just depression and anxiety? Sleep 32, A264 (2009)
7. Benca, R.M.: Diagnosis and treatment of chronic insomnia: a review. Psychiatry Service 56, 332–343 (2005)
8. Wing, Y.K.: Herbal treatment of insomnia. Hong Kong Medical Journal 7, 392–402 (2001)
9. Chen, L.C., Chen, I.C., Wang, B.R., Shao, C.H.: Drug-use pattern of Chinese herbal medicines in insomnia: a 4-year survey in Taiwan. Journal of Clinical Pharm Therapy 43, 555–560 (2009)
10. Attele, A.S., Xie, J.-T., Yuan, C.S.: Treatment of insomnia: An alternative approach. Alternative Medicine Review 5, 249–259 (2000)
11. Cuellar, N.G., Roger, A.E., Hisghman, V.: Evidenced based research of complementary and alternative medicine (CAM) for sleep in the community dwelling older adult. Geriatric Nursing 28, 46–52 (2007)
12. Gooneratne, N.S.: Complimentary and alternative medicine for sleep disturbances in older adults. Clinical Geriatric Medicine 24, 121–viii (2008)
13. Tang, J.-L., Liu, B., Ma, K.-W.: Traditional Chinese Medicine. The Lancet 372, 1938–1940 (2008)
14. Sun, H., Yan, J.: Discussion of the ‘BuMei’ syndrome in Inner Canon of Huangdi. Zhongyi Zazhi, 7 (2004)
15. Sun, H., Yan, J.: Discussion of the ’BuMei’ syndrome in Synopsis of Golden Chamber. Lishizhen medicine and material medical research 16, 182–183 (2005) (in Chinese)
16. Fei, Z.D., Wu, X.Q.: Discussion of the effective formula for insomnia: wendang tang. Jiangsu traditional Chinese medicine 26, 39 (2005) (in Chinese)
17. Yu, Y., Ruan, S.: Using cassia bark for insomnia treatment. Zhongguo Zhongyao Zazhi 23, 309–310 (1998) (in Chinese)
18. Cui, Y., Wang, S., Liu, W.: The advancement of Chinese medicine diagnosis and treatment to insomnia. Journal of Henan University of Chinese Medicine 23, 102–104 (2008) (in Chinese)
19. Zhou, X., Chen, S., Liu, B., et al.: Extraction of hierarchical core structures from traditional Chinese medicine herb combination network. In: ICAI 2008, Beijing, China (2008)
20. Ritchie, M.D., Hahn, L.W., Roodi, N., et al.: Multifactor dimensionality reduction reveals high-order interactions among estrogen metabolism genes in sporadic breast cancer. American Journal of Human Genetics 69, 138–147 (2001)
21. Hanh, L.W., Ritchie, M.D., Moore, J.H.: Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. Bioinformatics 19, 376–382 (2003)
22. Zhou, X., Chen, S., Liu, B., et al.: Development of Traditional Chinese Medicine Clinical Data Warehouse for Medical Knowledge Discovery and Decision Support. Artificial Intelligence in Medicine 48(2-3), 139–152 (2010)
23. Wasserman, S., Faust, K.: Social network analysis: Methods and application (1994)
24. Barabasi, A.-L., Reka, A.: Emergence of scaling in random networks. Science 286, 509–512 (1999)
25. Zhou, X., Liu, B.: Network Analysis System for Traditional Chinese Medicine Clinical Data. In: Proceedings of BMEI 2009, Tianjin, China, vol. 3, pp. 1621–1625 (2009)
26. Shi, A., Lin, S.S., Caldwell, L.: Essentials of Chinese Medicine, vol. 3. Springer, Heidelberg (2009)
27. Wang, J., Guo, L., Wang, Y.: Methodology and prospects of study on theory of compatibility of prescriptions in traditional Chinese medicine. World science and technology-modernization of TCM and materia medica. 8(1), 1–4 (2006)