Sasang typology classification with data reduction and SOM algorithm

Han Chae †, Byoung Chul Kim‡, Young In Kim‡, Il Kyu Eom*

† School of Korean Medicine, Pusan National University, Gyeongnam, 626-870, Korea
‡ Applied Information Technology and Engineering Department, Pusan National University, Gyeongnam, 627-706, Korea
* School of Electrical Engineering, Pusan National University, Busan, 609-735, Korea

***Abstract*— Personalized medicine has been the major interest for the Western and traditional medicine for a long time since Hippocrates and Yellow Emperor. The purpose of this study was to utilize data reduction and artificial network algorithm for the classification of Sasang types with clinical data. The results showed acceptable type-specific sensitivity and specificity considering its limitation with data. Multidisciplinary approach for the personalized medicine is needed for upcoming integrative medicine.**

***Index Terms*—Sasang Typology, classification, SOM, data reduction**

I. Introduction

The Personalized Medicine, which can increase the efficacy and safety with custom-fit diagnosis and treatment, has been an interest subject for a long time throughout the history of medicine [2]. Even the Hippocrates wrote in his book that one man's cure are another's poison, and suggested four temperaments including Blood, Yellow bile, Black bile and Phlegm as a foundation of his physiology. In the 2nd century AD the Greek physician Galen followed this tradition and suggested four temperament/constitutions; Sanguine, Choleric, Melancholic and phlegmatic [1, 3].

The Personalized Medicine has got its attention with the development of genetics these days. The research on the gene-based tailored medicine such as Whole Genome Association Study (WGAS) and epigenetics is widely used for the examination of patho-physiological examination of many diseases even with metabolic and psychiatric diseases after the completion of Human Genome Project [4]. The Systems biology, a novel approach in the field of life science focusing on all the constituents including DNA and proteins and its interactions in biological system, collects huge data using high-throughput Omics technology and analyzes vast volumes of data through cutting-edge mathematical techniques for the fishing-out of new clue for developing tailored medication [5].

Interestingly the data-mining technique for this new trend of medicine which needs multidisciplinary approach including clinicians, mathematics and engineering plays pivotal role not only for the Western Personalized Medicine but also the traditional medicine (TRM) studies including Yin-Yang theory and the essence of phlegm and blood stasis.

In present days, there are several forms of Personalized Medicine from TRM, and it is actively used in India, Korea, China and Japan. Since the personalized medication has been the major concern of many TRM, they have actively developed medical typology, such as Ayurveda, Sasang medicine, Chinese constitution and Ikkando medicine, based on unique bio-psychological pathophysiology, treatment and prevention for their clinical application with the influence of cultural heritage [6].

Since the basic elements, diagnostic methods of TRM and algorithm and analysis modeling for TRM clinical practice have several challenges to overcome [7, 8], the advances were not sufficient for the actual TRM clinical situation until now. To tackle this issue, we reviewed the clinical perspectives of traditional medical typology from Korea and tried to look at how the data extraction and clinical classification can be achieved more efficiently with clinical data.

II. Sasang typology and artificial neural network

The Sasang typology, foremost interest of this study has several typical features. This traditional Korean medical typology was systematically theorized hundred years ago with the quaternary nature of Neo-Confucianism (sadness, anger, gladness and enjoyment) which makes the four Sasang types with typical temperament profile, logical thinking, behavioral patterns, status of organ system, patho-physiological features, predisposition to a specific illness, physical characteristics and response to a certain treatment [1, 9-11].

Human beings are classified into four Sasang types, such as Tae-Yang (TY), So-Yang (SY), Tae-Eum (TE) and So-Eum (SE), with their distinctive biopsychological temperaments and constitutions, and each of Sasang types has type-specific guideline for safe and effective medical herb and acupuncture use (Table 1).

There were also several studies for the purpose of new objective diagnostic instrument [12] and diagnostic index [13] for the clinical use, but the insufficiency of study on the classification algorithm [8, 14] and its validation [15] have been major challenges to overcome.

In this study, we recognize the need of a new research method that considers the statistical analysis of biological data and its usability to Sasang medicine simultaneously. Therefore, we examine the use of artificial neural network to classify Sasang types with collected clinical data.

H. Chae (Corresponding author email: han@chaelab.org) is with the School of Korean Medicine, Pusan National University, Gyeongnam, 626-870, Korea

Byoung Chul Kim (bckim@pusan.ac.kr), Young In Kim are with the Applied Information Technology and Engineering Department, Pusan National University, Gyeongnam, 627-706, Korea (kimyi@pusan.ac.kr)

Il Kyu Eom is with the School of Electrical Engineering, Pusan National University, Busan, 609-735, Korea (hikeom@pusan.ac.kr)

978-1-61284-696-5/11/$26.00 ©2011 IEEE

Table 1. Key characteristics of the Sasang typology (modified from studies of Chae et al.[1] and others)

Type (prevalence)	*Tae-Yang (太陽人) (<0.1%)*	*So-Yang (少陽人) (20%)*	*Tae-Eum (太陰人) (50%)*	*So-Eum (少陰人) (30%)*
Natural temperament	Sorrow	Anger	Gladness	Enjoyment
organ system	Developed Lung, undeveloped Liver	developed Spleen, undeveloped Kidney	Developed Liver, undeveloped Lung	Developed Kidney, undeveloped Spleen
representative features	Masculine, move forward, originative	Active, external oriented, talented for business, short and small	Feminine, stay retracted, conservative, tall and big	Still, internal oriented, self-directed, short and small
Character	Creative, positive, progressive, charismatic, heroic, rash mind	Unstable, easily get bored, sacrificing, righteous, easily acceptable, hot tempered, anxious mind	Gentle, commercial, endurable, humorous, look foolish, coward, fearful mind	Neat, mild, negative, intelligent, organized, selfish, jealous, persistent, nervous mind
Body shape	Developed nape of the neck, slender waist	Developed chest, small hips	Thick waist, weak nape of the neck	Developed hip, weak chest
Healthy sign & unhealthy sign	Urination Bbbles in mouth, emesis	Bowel movement Constipation	Perspiration No perspiration	Digestion Indigestion
type-specific acupuncture	Diagnosis with HT8. Treatment with LR3(+)/ LU9(-)	Diagnosis with HT3. Treatment with KI3(+)/ SP3(-)	Diagnosis with HT4. Treatment with LU9(+)/ LR3(-)	Diagnosis with HT7. Treatment with SP3(+)/ LI4(-)
type-specific medical herbs	Chaenomelis Fructus, Acanthopanacis Cortex, Phragmitis Rihizoma	Rehmanniae Radi, Corni Fructus, Hoeoen, Alismatis Rhizoma, Osterici Radix, Angelicae Pubescentis Radix	Ephedrae Herba, Liriopis Tuber, Schisandrae Fructus, Dioscoreae Rhizoma, Platycodi Radix, Coicis Semen, Puerariae Radix	Ginseng Radix, Atractylodis Rhizoma Alba, Glycyrrhizae Radix, Cinnamomi Cortex, Citri Pericarpium, Zingiberis Rhizoma Crudus
Significant current studies		High Extraversion and low Neuroticism	High Body Mass Index and high width-to-height ratio of face. Bigger neck circumference	Low Extraversion and high Neuroticism

Self-organizing feature map (SOM) is an unsupervised learning-based neural network introduced by T. Kohonen [19]. SOM is suitable for data mining or classification because it creates a set of prototype vectors representing the input data set. Therefore, in this paper, we present a typical model of Sasang type classification using artificial neural network and examine the usability of SOM to the TRM through the consideration of the reproducibility of input data profiles.

III. Method

2.1. Data Collection

To evaluate the classification ability of artificial neural network, the conventional data used by Chae et al.[1] are applied. Study subjects are 79 students between the ages of 19 and 43 (69 males, 10 females) enrolled in the oriental medical physiology class at the College of Oriental Medicine, Kyung Hee University, Seoul, Korea in 2000. For study subjects, Questionnaire for the Sasang Constitution Classification II (QSCCII) test, Myers-Briggs Type Indicator (MBTI) test, and Bio-Impedance Analysis (BIA) test are used.

QSCC is a *Sasang* typology-based inventory, which was developed by the Department of Sasang Medicine at Kyung Hee Medical Center (Seoul, Korea) in 1993 and revised in 1996 [16], and has been widely used in clinical studies. The QSCC is composed of 121 forced-choice items. The internal consistency (Cronbachs α) of this inventory is as follows: TY type is 0.57, SY type is 0.57, TE type is 0.59, and SE type is 0.63.

The MBTI is a self-report questionnaire using 95 forced choice items developed by Myers and Briggs and translated into Korean in 1990. It is a psychometric instrument designed to assess normal personality traits [17]. This inventory assesses differences of the way people perceive information and preference for using that information. Individuals fall into four dichotomous personality dimensions based on their scores. Thus, there are eight categorical personality types: Introversion/Extroversion, Sensing/Intuition, Thinking/Feeling, and Judging/Perceiving. The categorical dimensions of MBTI individual (i.e., Extroversion/Introversion) were also presented as continuous scores (i.e., below 100 is Extroversion and above 100 is Introversion). The preference scores (MBTI scores, hereafter) of the four dichotomies were used for the analysis in the present study.

The BIA is an electrical method for measuring anthropometric data in epidemic or clinical. It is simple and noninvasive and provides reliable results for estimating total body water and lean body mass (LBM) [18]. Assuming that LBM is hydrated in a constant and uniform manner, the BIA can be used to estimate body fat mass (BFM), the non-hydrated portion of the body, by subtracting LBM from the weight.

The notations and meanings of the data field used in this paper are follows; Sex, Age, MBIT EI (M_EI), MBTI SN (M_SN), MBTI TF (M_TF), MBTI JP (M_JP), Intracellular water (ICF), Extracellular water (ECF), Protein mass (PM), Amino mass (AMM), Body fat mass (BFM), Height, Weight, Percent body fat (PBF), Waist-hip ratio (WHR), Fluid of right arm (FRA), Fluid of left arm (FLA), Fluid of trunk (FT), Fluid of right leg (FRL), Fluid of left leg (FLL), Development of body, Total body water (TBW), Body mass index (BMI), arm leg ratio (ALR), Proper body weight (PBW), average body weight (ABW), average body fat (ABF), average muscle weight (AMW), right and left ratio of the upper limb (RLU), and right and left ratio of the lower limb (RLL).

2.2. Data Reduction

The acquired data using MBIT and BIA etc. have 30 fields and 79 records including age and sex. From this, the number of the input node for SOM is 30. The input dimension for the SOM can be reduced by removing redundancy using proper data reduction methods. Firstly, we perform data field reduction by considering existing functional relationships between input data. Generally, the experimental data can have the functional relations such as, linear or nonlinear relations. Since SOM has self-organizing capability for the complex functional relationship between input data, we can eliminate the number of input fields using the analysis of functional relations. Table 1 shows the eliminated data filed by analyzing latent relationship. As shown in Table 2, 9 data fields are composed of the other fields.

Table 2. Data elimination by analyzing latent relationship

Eliminated data fields	Latent relationship between data fields
TBW	TBW = ICF + ECF
BMI	BMI = Weight / (Height*Height)
ALR	ALR = (FRA + FLA)/(FRL + FLL)
ABW	ABW = f(Height)
BWC	BWC = adequate weight – weight
FRA/FLA	FRA/ FLL = RLU
FRL/FLL	FRL/FLL = RLL
Body fat control	=f(BFM, Height, Weight)
Body muscle control	

Since the correlation coefficient is the numeric measure of the linear dependency, the higher correlation coefficient reflects the higher linear dependency. In this study, we reduce remaining 21 data fields using the cross-correlation analysis. For the data set d_m and d_n with N dimension ($m,n = 1,2,\cdots,N$), the correlation coefficients ρ_{mn} is presented as follows;

$$\rho_{mn} = \frac{E[(d_m - \mu_m)(d_n - \mu_n)]}{STD(d_m)STD(d_n)}, \tag{1}$$

where $E[\bullet]$ is expectation value, $STD[\bullet]$ is standard deviation, and $\mu_k (k = m,n)$ is the average value of d_k, respectively. The range of ρ_{mn} is $-1 \le \rho_{mn} \le 1$. The large absolute value of ρ_{mn}, that is $|\rho_{mn}|$, means high linear dependency. Thus, data fields with high correlation coefficient can be eliminated. In this study, we remove 8 data fields with $|\rho_{mn}| > 0.9$.

To this end, we remove 9 data fields using functional relationship between data fields, and remove 8 data fields using correlation coefficients. Under the assumption that Age field do not affect the Sasang type classification, we remove Age filed. Finally, only 12 data fields of 30 total data fields are used to classify Sasang type. That is, the remaining 12 data fields are input vectors to SOM neural network.

2.3 Sasang type classification using SOM

SOM is composed of input layer and competition (output) layer as shown in Figure 1. The input layer distributes input patterns to each node in the output layer through the weight matrix. In Figure 1, $\mathbf{p} = [p_1, p_2, \cdots, p_K]$ is input vector and the weight vector that connect *i*-th node of input layer and *j*-th node of output layer $w_{ij} (j = 1,2,\cdots,L)$ is presented by

$$\mathbf{w}_j = [w_{j1}, w_j, \cdots, w_{jK}]. \tag{2}$$

SOM has a feed-forward structure with a single computational layer arranged in rows and columns. Each neuron is fully connected to all the input nodes in the input layer. The goal of SOM is to transform an input pattern with arbitrary dimension into a one or two dimensional discrete map. It performs this transformation in a topologically ordered fashion adaptively. The weight vectors of the neuron become selectively updated by various input patterns during the competitive learning.

The update rule of the weight vector $\mathbf{w}_v(t)$ at time *t* is

$$\mathbf{w}_v(t+1) = \mathbf{w}_v(t) + \alpha(t)(\mathbf{p} - \mathbf{w}_v(t)), \tag{3}$$

where $\alpha(t)$ is a monotonically decreasing learning coefficient at time *t*. At each training step, a sample input data vectors $\mathbf{p}$ is randomly presented from the training data sets, and the distance between the data and all the weight vectors of the SOM is calculated. The node whose weight vector is closest to the input vector is called the best-matching unit, and this node will be one single *winning* node. This process is repeated for a given number of cycles or satisfies a given stop criteria.

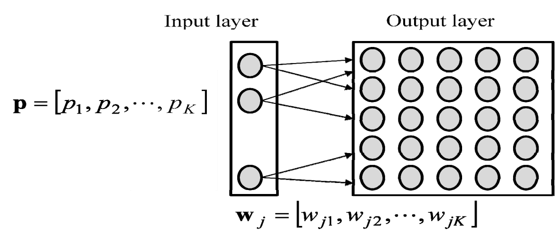

Figure 1. The structure of Self-organizing feature Map(SOM).

In this study, for the reduced 12 data fields, we perform Sasang type classification using SOM as following steps.

Step 1] Initialize weight vectors with small random numbers.
Step 2] Present input vector to input layer.
Step 3] Use Euclidean distance formula to find similarity between the input vector and the weight vector. Track the node that produces the smallest distance.
Step 4] Update the nodes in the neighborhood of best matching unit by pulling them closer to the input vector using (3).
Step 5] Increase *t* and repeat from Step 2 while a given stop criterions is satisfied.

IV. RESULTS

The characteristics of MBTI and BIA of each Sasang types by SOM show similar pattern to those of conventional results. Especially, the MBTI profile by SOM is very similar to the actual measures. EI, SN, TF and JP values for each Sasang type is presented Table 3.

As shown in Table 3, the results of artificial neural network for SY and SE are similar to those of actual measurement. In the case of TE type, the MBIT score profiles of SOM results have higher scores than actual measurements.

There are no significant differences between SOM results and actual measurements of each Sasang types. Table 3 shows BFM (Body Fat Mass) for each Sasang types.

Table 3. MBTI and BFM (mean (standard error)) of each *Sasang* types using artificial neural network and actual measurement

		EI	*SN*	*TF*	*JP*	*BFM*
SY	SOM	93.8 (3.8)	90.3 (4.2)	80.4 (4.1)	97.4 (3.7)	12.5 (0.8)
	Actual	101.6 (4.8)	100.4 (4.7)	88.2 (4.7)	106.4 (5.3)	12.4 (0.7)
SE	SOM	137.4 (2.1)	86.6 (4.0)	88.0 (4.4)	74.8 (3.0)	12.7 (0.6)
	Actual	133.9 (3.0)	91.8 (3.8)	95.1 (4.2)	88.4 (4.3)	12.3 (0.6)
TE	SOM	122.5 (4.0)	110.3 (3.4)	106.0 (3.8)	127.1 (3.1)	14.3 (1.1)
	Actual	118.0 (4.7)	96.3 (4.7)	91.4 (5.0)	107.1 (6.0)	15.2 (1.3)

The comparison of classification by SOM and actual measurement is shown in Table 4, and the Sasang type-specific sensitivity and specificity were calculated as a clinical validation of statistical agreement [15]. The type-specific sensitivity for SY, TE and SE types are 56.0%, 47.8% and 61.3%, and the type-specific specificity for SY, TE and SE types are 81.5%, 71.4%, and 81.3%.

Table 4. Classification by SOM and QSCCII

		Classified with QSCCII			
		SY	TE	SE	Total
Classification by SOM	SY	14	6	4	24
	TE	8	11	8	27
	SE	3	6	19	28
	Total	25	23	31	79

V. Discussions

This study presented the usefulness of SOM in TRM diagnosis with actual clinical data especially in Sasang typology, the traditional Korean personalized medicine with acupuncture and medical herbs.

Considering only 12 fields out of 30 complex clinical data selected by data reduction methods were used for the analysis and plain SOM was applied for this study, the MBTI score profile and BFM value by artificial neural network is quite acceptable. The results of this study can be a good illustration for the need of data mining tools for life science [5, 8].

However the sensitivity (48% to 61%), especially TE type, was not satisfactory compared to the specificity (71 to 82%). This can be improved with inclusion of more clinical measures useful for the differential diagnosis of TE type from others along with the modification of classification algorithm more fit to the clinical data.

For the better understanding of clinical data and efficient diagnosis, the multi-disciplinary approach incorporating engineering and mathematics along side with clinicians can be a prerequisite for the personalized medicine of 21st century Integrative Medicine.

References

[1] H. Chae, I. K. Lyoo, S. J. Lee, S. Cho, H. Bae, M. Hong, and M. Shin, "An alternative way to individualized medicine: psychological and physical traits of Sasang typology," *J Altern Complement Med,* vol. 9, pp. 519-28, Aug 2003.

[2] J. K. Nicholson, "Global systems biology, personalized medicine and molecular epidemiology," *Mol Syst Biol,* vol. 2, p. 52, 2006.

[3] G. Pailhez and A. Bulbena, "Body shpae and psychiatric diagnosis revisited," *International Journal of Psychiatry in Clinical Practice,* vol. 14, pp. 236-243, 2010.

[4] "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls," *Nature,* vol. 447, pp. 661-78, Jun 7 2007.

[5] P. Li and L. P. Yang, "[Application of systems biology method in the research of traditional Chinese medicine]," *Zhong Xi Yi Jie He Xue Bao,* vol. 6, pp. 454-7, May 2008.

[6] K. R. Kang, S. M. Hwang, S. J. Park, and H. Chae, "A Comparative Study on Traditional Constitutional Medicine in the World," *Korean Journal of Oriental Medicine,* vol. 15, pp. 35-43, 2009.

[7] S. Lukman, Y. He, and S. C. Hui, "Computational methods for Traditional Chinese Medicine: a survey," *Comput Methods Programs Biomed,* vol. 88, pp. 283-94, Dec 2007.

[8] J. W. Hong, Y. I. Kim, S. J. Park, B. C. Kim, I. K. Eom, M. W. Hwang, S. W. Shin, B. J. Kim, Y. K. Kwon, and H. Chae, "Data mining algorithms for the development of Sasang type dianosis," *KJOPP,* vol. 23, pp. 1234-1240, 2009.

[9] H. Chae, S. H. Park, S. J. Lee, and K. C. Koh, "Sasang typology from a personality perspective," *J Korean Oriental Med,* vol. 25, pp. 151-164, 2004.

[10] H. Chae, S. H. Park, S. J. Lee, M. G. Kim, D. Wedding, and Y. K. Kwon, "Psychological profile of sasang typology: a systematic review," *Evid Based Complement Alternat Med,* vol. 6 Suppl 1, pp. 21-9, Sep 2009.

[11] S. H. Park, M. G. Kim, S. J. Lee, J. Y. Kim, and H. Chae, "Temperament and Character Profiles of Sasang Typology in an Adult Clinical Sample," *Evid Based Complement Alternat Med,* Apr 20 2009 Advance Access published April 20, 2009 doi:10.1093/ecam/nep034.

[12] S. W. Lee, E. S. Jang, J. Lee, and J. Y. Kim, "Current researches on the methods of diagnosing sasang constitution: an overview," *Evid Based Complement Alternat Med,* vol. 6 Suppl 1, pp. 43-9, Sep 2009.

[13] J. H. Yoo, J. W. Kim, K. K. Kim, J. Y. Kim, B. H. Koh, and E. J. Lee, "Sasangin diagnosis questionnaire: test of reliability," *J Altern Complement Med,* vol. 13, pp. 111-22, Jan-Feb 2007.

[14] W. O. Sung, M. J. Song, and K.-w. Chung, "K-mean and Double Cross-Validation Algorithm for LS-SVM in Sasang Typology Classificaiton," *in Proceedings of the IEEE International Conference on Automation and Logistics, August 20007, Jinan, China,* pp. 426-430, 2007.

[15] S. M. Hwang, S. J. Park, K.-R. Gang, Y. K. Kwon, and H. Chae, "Generalized analysis on validity of Sasnag type diagnosis," *KJOPP,* vol. 23, pp. 950-957, 2009.

[16] S. Kim, B. Ko, and I. Song, "A study on the standardization of QSCC (Questionnaire for the Sasang Constitution Classification)2," *J Korean Orient Med Soc,* vol. 17, pp. 337-393, 1996.

[17] J. Kim, H. Sim, and S. Je, *Development and use of MBTI.* Seoul, Korea: KPTI, 1995.

[18] K. Cha, G. M. Chertow, J. Gonzalez, J. M. Lazarus, and D. W. Wilmore, "Multifrequency bioelectrical impedance estimates the distribution of body water," *J Appl Physiol,* vol. 79, pp. 1316-9, Oct 1995.

[19] T. Kohonen. *Self-Organizing Maps*. Springer, Berlin, Heidelberg, 1995.