

Received 17 August 2023, accepted 3 September 2023, date of publication 11 September 2023,
date of current version 19 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3313943

RESEARCH ARTICLE

A Multi-Objective Hyper-Heuristic Clustering Algorithm for Formulas in Traditional Chinese Medicine

WEN SHI¹, JINGYU ZHANG², BIN YU¹, YIBO LI¹, AND SHIHUI CHENG¹

¹School of Information Engineering, Tianjin University of Commerce, Tianjin 300134, China

²Tianjin Nankai Hospital, Tianjin 300100, China

Corresponding author: Wen Shi (shiwen@tjcu.edu.cn)


This work was supported by the National Student Training Program for Innovation and Entrepreneurship of China under Grant 202210069016 and Grant 202310069063.

ABSTRACT Syndrome types are important for diagnosis and treatment in traditional Chinese medicine. Syndrome types can be summarized by domain experts as formula clusters. In this paper, we propose seven feature models for the formula clustering problem based on categories, subcategories, functional tendencies and names of Chinese materia medica. A novel multi-objective clustering hyper-heuristic algorithm is obtained. In our proposed algorithm, 12 low-level heuristics are used for clustering solution perturbation by merging clusters, dividing clusters or moving points between clusters based on received solutions from the high-level heuristic. The high-level heuristic evaluates the received solutions from low-level heuristics, updates the solution pool, and selects initial solutions for the next iteration via roulette wheel selection on the Pareto front. Experimental results demonstrate that the proposed algorithm outperforms other clustering algorithms in most datasets. The initial number of clusters has less influence on the final clustering solutions for our proposed algorithm than for other clustering algorithms. For most datasets, the roulette wheel selection mechanism on the Pareto front shows higher convergence rates and accuracy than a random selection mechanism. Accuracy was higher for feature models based on functional tendencies than for the other feature models.

INDEX TERMS Clustering, data mining, hyper-heuristic, syndrome differentiation.

I. INTRODUCTION

Traditional Chinese medicine (TCM) has played a crucial role in health care in China for 2000 years. In TCM, a syndrome type (*Zheng*) is an important concept and represents a combination of signs and symptoms. It is the generalization of the process of a disease at a certain stage. Since it involves the location, cause and nature of the disease, and the relation between pathogenic factors and healthy *Qi*, a syndrome type can comprehensively and accurately reveal the nature of a disease. In the diagnosis stage, four methodological aspects of diagnosis are applied for syndrome differentiation (*Bian Zheng*) and are called the “Four pillar”: observation (inspection), auscultation, olfaction, and inquiry including

The associate editor coordinating the review of this manuscript and approving it for publication was Yiming Tang .

pulse-taking, and palpation. Syndrome differentiation implies that the clinical data of a patient are analyzed and generalized to identify the pathological mechanism of the disease. In the treatment stage, the doctor of TCM selects the corresponding therapy, such as TCM formulas according to the syndrome types.

Many approaches have been proposed to determine the relationships between syndrome types and diagnostic information [1], [2], metabonomics [3], [4], multi-omics [5] and so on. However, few studies have focused on the relationship between syndrome types and ancient formulas. There are thousands of formulas recorded in the ancient medical books of TCM. Regrettably, a significant portion of ancient formulas have not made any mention of the corresponding syndrome types. This hinders the clinical application of these formulas. TCM experts have to manually infer

the corresponding syndrome types according to the formula information.

A formula contains various Chinese materia medica (CMMs) with certain dosages that are used as the treatment for a particular syndrome type. Typically, different formulas used as treatments for the same syndrome type contain CMMs of the same category or with the same functional tendencies. In this paper, we will cluster formulas according to their CMMs, with each formula cluster expected to correspond to a syndrome type. Note that the accuracy of some clustering algorithms, such as k -means [6] and k -medoids [7], strongly depend on the initial number of clusters. However, for the formula clustering problem, it is very difficult to determine the number of syndrome types. Therefore, we would like to develop a clustering algorithm that can adjust the number of clusters during the process of this algorithm in order to reduce the influence of the initial number of clusters on the clustering results.

For the purpose of this work, we first propose seven feature models for the formula clustering based on categories, subcategories, functional tendencies, and names of CMMs. The reason is that, as unstructured texts, formulas cannot be directly clustered. Based on the feature model, formulas can be transformed into numerical features for clustering. Clustering algorithms can evaluate the distance between formulas based on the features. Since this is the first study of the formula clustering problem for syndrome types, we hope to find out which factors in the formula are more effective for clustering. Hence, we select the categories, subcategories, functional tendencies, and names of CMMs for feature extraction for formula clustering. Then a novel clustering algorithm based on a multi-objective hyper-heuristic (MOCHH) for formula clustering in TCM is proposed. In MOCHH, one high-level heuristic (HLH) and 12 low-level heuristics (LLHs) are proposed. Unlike the other hyper-heuristics reviewed in Section II, LLHs in MOCHH are divided into three classes: merging clusters, dividing clusters, and perturbing clusters, which dynamically adjust the number of clusters. Firstly, MOCHH generates an initial k -means clustering solution and inserts it into the solution pool. Secondly, the HLH sends the initial clustering solutions to all LLHs. The LLHs search for clustering solutions by their own heuristic rules and send the new clustering solutions back to the HLH. Thirdly, the HLH evaluates all the received solutions according to a multi-objective optimization and updates the solution pool in the HLH. Finally, the HLH selects initial solutions for the next iteration by roulette wheel selection on the Pareto front and sends them to all LLHs. Experimental results demonstrate that (a) MOCHH can obtain better clustering solutions than other clustering algorithms such as k -means, k -medoids, DBSCAN, and GA; (b) the final clustering solutions are less sensitive to the initial number of clusters in MOCHH than in other clustering algorithms; (c) the proposed selection mechanism in MOCHH exhibits higher convergence rates and accuracy than a random selection mechanism for most datasets; and (d) accuracy is

higher for feature models based on functional tendencies than for the other feature models.

The contributions of this work are presented as follows: First, to the best of our knowledge, this is the first study of the formula clustering problem for syndrome types. Seven feature models based on categories, subcategories, functional tendencies, and names are proposed for the formula clustering problem. Formulas can be transformed into numerical features for clustering based on the feature model. From our experiments evaluating the performance of MOCHH, we find that the functional tendencies of CMMs are more suitable for summarizing syndrome types than the categories, subcategories, and names of CMMs. This discovery is of great significance for future studies of formula clustering. Second, a novel clustering algorithm based on a multi-objective hyper-heuristic for formula clustering in TCM is proposed. The MOCHH can adjust the number of clusters during the process of this algorithm in order to reduce the influence of the initial number of clusters on the clustering results. Hence, MOCHH can generate high accuracy results regardless of the initial number of clusters. TCM experts can summarize the syndrome types conveniently based on the high accuracy automated clustering of TCM formulas by MOCHH instead of manually inferring the corresponding syndrome types according to the formula information.

The remainder of this paper is organized as follows. Section II reviews the current research about the syndrome types, clustering algorithms and hyper-heuristic. Section III offers the definitions of feature models for formula clustering in TCM. Section IV provides the details of the proposed MOCHH algorithm. Section V reports the results from a simulation evaluating MOCHH formula clustering performance. Section VI concludes the paper and presents recommendations for future work.

II. LITERATURE REVIEW

1) SYNDROME TYPES

In TCM, many approaches have been proposed to determine the relationships between common syndrome types and diagnostic information for a given disease. Huang et al. [1] constructed an algorithm model compatible for the multidimensional, highly sparse, and multiclassification task of TCM syndrome differentiation. The model was based on the classification of different symptoms and physical signs according to the four diagnostic examinations in TCM diagnosis. Wang et al. [2] identified the clusters of TCM syndrome types in type 2 diabetes mellitus patients and explored the association between those TCM syndrome types clusters and health-related behaviors, including smoking, alcohol use, tea drinking, the intensity of physical activity, sleep quality, and sleep duration. Zhou et al. [8] proposed the discovery and assessment of potential biometabolic markers for various syndrome types of coronary heart disease. Yang et al. [9] explored the association rules of breast cancer and TCM syndromes based on the clinical manifestations and body parameters of breast cancer patients. Song et al. [10] studied

the diagnoses of TCM syndromes for irritable bowel syndrome. Xia et al. [11] conducted factor analysis for syndrome element extraction and k-means cluster analysis for syndrome type classification in the prevention and treatment of metabolic syndrome. Wang et al. [12] explored the possible correlations between soluble ST2, IL-33, IL-10, and IL-17 levels, and syndromes in patients with rheumatoid arthritis.

Metabonomics is very important for the studying on the differentiation, material basis, metabolic pathways, and efficacy on syndrome types. Lyu et al. [3] reviewed the application of metabonomics in the study of TCM syndrome types. He et al. [4] explored the mechanism of Taohong Siwu Decoction in the treatment of blood deficiency and blood stasis syndrome by gut microbiota combined with metabolomics. Liu et al. [5] summarized relevant studies in genomics, transcriptomics, proteomics, and metabolomics. Some studies found that gene polymorphisms, differential lncRNAs, mRNAs, miRNAs, proteins, and metabolites may be associated with TCM syndrome types of stroke.

2) CLUSTERING ALGORITHMS

Many approaches have been proposed for clustering. The most well-known clustering algorithm is *k*-means. Reference [6] *K*-means allows the division of objects into *k* partitions based on object attributes. In *k*-means, the centroid point of each cluster is the average of all points in the cluster but not necessarily an object point. In *k*-medoids [7], another clustering algorithm, the main process is the same as for *k*-means, but the final centroids in *k*-medoids are object points. One of the disadvantages of *k*-means and *k*-medoids is the requirement of a predefined number of clusters.

The density-based DBSCAN [13] algorithm is also commonly used for clustering. DBSCAN groups together points that are close to each other based on a distance metric and a minimum number of points per cluster. However, DBSCAN cannot cluster data well when there are large heterogeneities in density.

Swarm intelligence methods have also been adopted to solve the data clustering problem. For example, a GA based on a clustering coefficient [14] was proposed for detecting communities in social and complex networks. An ant colony optimization-based [15] clustering algorithm was proposed for a bus-based vehicular *ad hoc* network. Artificial bee colony [16] was also adopted as a clustering algorithm to minimize the execution time and to optimize the clustering according to dataset size. Lin et al. [17] proposed a new algorithm for the clustering problem, the fruit-fly optimization *k*-means algorithm and designed a distribution centre location problem and three clustering indicators to evaluate the performance of the algorithm. Tawhid and Ibrahim [18] developed a new hybrid swarm intelligence optimization algorithm called monarch butterfly optimization algorithm with cuckoo search algorithm for optimization problems.

3) HYPER-HEURISTIC

Hyper-heuristics [19] was introduced to describe the idea of “heuristics to choose heuristics”. A hyper-heuristic selects

appropriate heuristics from among several candidate heuristics or generates new heuristics automatically. A hyper-heuristic usually contains one HLH and several LLHs. The HLH is in charge of managing LLHs by selecting from among LLHs and updating solutions for the next search iteration. LLHs operate directly on the search space of solutions.

The hyper-heuristic algorithm has already been used to solve clustering problems. Costa et al. [20] proposed a cluster-based hyper-heuristic for large-scale vehicle routing problems. The proposed hyper-heuristic used 11 LLHs to find the optimal routes. Tsai et al. [21] made use of four heuristic algorithms, including tabu search, GA, ant colony optimization, and particle swarm optimization, as LLHs for a hyper-heuristic clustering algorithm for wireless sensor networks. In another study, a simulated annealing algorithm was used as the HLH to select and evaluate performance for 19 LLHs for various clustering problems [22]. The cost function was the summed distance between each instance and the cluster center. Kumari and Srinivas [23] proposed a hyper-heuristic algorithm for multi-objective software module clustering. The proposed algorithm included 12 LLHs, classified according to their selection, recombination, and mutation mechanisms. Alshareef and Maashi [24] applied a multi-objective hyper-heuristic method to solve the multi-objective software module clustering problem with three objectives: minimize coupling, maximize cohesion, and ensure high modularization quality.

III. FEATURE MODELS FOR FORMULA CLUSTERING

Typically, a single traditional Chinese medical formula contains several CMMs for the treatment of a particular syndrome type. For example, the classic formula Four Gentlemen Decoction (*Sijunzi decoction*) is applied to treat the syndrome type of *Qi deficiency in the spleen* by strengthening the spleen and replenishing *Qi*. *Sijunzi decoction* has been used for the treatment of diseases such as disorders of gastrointestinal function [25], accompanied by poor appetite, reduced food intake, and loose stools [26]. In modern pharmacological studies, *Sijunzi decoction* has also been found to strengthen the immune system [27]. There are four CMMs in the formula *Sijunzi decoction*: the root of *Panax ginseng* C.A. Mey. (*renshen*), the rhizome of *Atractylodes macrocephala* Koidz. (*baizhu*), *Poria cocos* (Schw.) Wolf (*fuling*), and the root, and rhizome of *Glycyrrhiza uralensis* Fisch. (*gancao*) [28].

The effectiveness of the formula is based on the main functional tendencies of the CMMs in this formula. For this example, the functional tendencies of the CMMs in the formula *Sijunzi decoction* are shown in Table 1. The functional tendency shared by all four CMMs strengthens the spleen. Meanwhile, the shared functional tendency of *renshen*, *baizhu* and *gancao* replenishes *Qi*. Since the main effectiveness of *Sijunzi decoction* is in *replenishing Qi* and *strengthening the spleen*, it is used for the syndrome type *Qi deficiency in the spleen*.

Each CMM belongs to one medicinal category based on its functional tendencies. In our example, *renshen*, *baizhu*,

TABLE 1. Functional tendencies of Chinese materia medica in *Sjunzi decoction*.

CMM	Functional Tendencies
<i>baizhu</i>	strengthening the spleen, replenishing Qi, and promoting urination
<i>renshen</i>	strengthening the spleen, replenishing Qi, and tranquilizing
<i>fuling</i>	strengthening the spleen, promoting urination, and tranquilizing
<i>gancao</i>	strengthening the spleen, replenishing Qi, and relieving cough

and *fuling* belong to the medicinal category of deficiency-tonifying (*Buxuyao*). *Baizhu* belongs to the medicinal category of dampness-draining diuretic (*Lishuishenshiyao*). Some but not all medicinal categories can be divided further into medicinal subcategories. For example, *Buxuyao* includes 4 subcategories: Qi-tonifying (*Buqi Yao*), yang-tonifying (*Buyang Yao*), blood-tonifying (*Buxue Yao*), and yin-tonifying (*Buyin Yao*). *Lishuishenshiyao* includes 3 subcategories: water-draining and swelling-dispersing (*lishuixiaozhong Yao*), strangury-relieving (*Liniaotonglin Yao*), and bile-draining anti-icteric (*Lishituihuang Yao*). In our example, *renshen*, *baizhu*, and *fuling* belong to the medicinal subcategory of *Buqi Yao*. *Baizhu* belongs to the medicinal subcategory of *Lishuixiaozhong Yao*. There are 21 medicinal categories and 48 medicinal subcategories [29] shown in Table 2. In our study, some medicinal categories cannot be divided into subcategories. These medicinal categories are set to include only one medicinal subcategories whose names are the same as the medicinal categories.

Many formulas in ancient medical texts are very useful for clinical treatment. However, these texts often did not record the corresponding syndrome types. Clustering of formulas by medicinal category, medicinal subcategory, functional tendency, or CMM name may allow the common characteristics of each formula cluster to be summarized. The corresponding syndrome types could then be extracted by TCM experts for clinical treatment.

Therefore, in this paper, we propose seven formula feature models for clustering. Some of the variables and parameters in the models are defined in Table 3.

1) MC-B MODEL

In the MC-B model, the feature vector of the formula is defined as

$$f_{MC-B} = [B_1^{MC}, B_2^{MC}, \dots, B_{|MC|}^{MC}], \quad (1)$$

where the Boolean variable B_i^{MC} equals 1 when the formula includes at least one CMM that belongs to the medicinal category mc_i .

2) MC-I MODEL

In the MC-I model, the feature vector of the formula is defined as

$$f_{MC-I} = [I_1^{MC}, I_2^{MC}, \dots, I_{|MC|}^{MC}], \quad (2)$$

where the integer variable I_i^{MC} equals the total number of CMMs included in the formula that belong to the medicinal category mc_i .

3) MSC-B MODEL

In the MSC-B model, the feature vector of the formula is defined as

$$f_{MSC-B} = [B_1^{MSC}, B_2^{MSC}, \dots, B_{|MSC|}^{MSC}], \quad (3)$$

where the Boolean variable B_i^{MSC} equals 1 when the formula includes at least one CMM that belongs to the medicinal subcategory msc_i .

4) MSC-I MODEL

In the MSC-I model, the feature vector of the formula is defined as

$$f_{MSC-I} = [I_1^{MSC}, I_2^{MSC}, \dots, I_{|MSC|}^{MSC}], \quad (4)$$

where the integer variable I_i^{MSC} equals the total number of CMMs included in the formula that belong to the medicinal subcategory msc_i .

5) FT-B MODEL

In the FT-B model, the feature vector of the formula is defined as

$$f_{FT-B} = [B_1^{FT}, B_2^{FT}, \dots, B_{|FT|}^{FT}], \quad (5)$$

where the Boolean variable B_i^{FT} equals 1 when the formula includes at least one CMM that involves the functional tendency ft_i .

6) FT-I MODEL

In the FT-I model, the feature vector of the formula is defined as

$$f_{FT-I} = [I_1^{FT}, I_2^{FT}, \dots, I_{|FT|}^{FT}], \quad (6)$$

where the integer variable I_i^{FT} equals the total number of CMMs in the formula that involve the functional tendency ft_i .

7) CMM-B MODEL

In the CMM-B model, the feature vector of the formula is defined as

$$f_{CMM-B} = [B_1^{CMM}, B_2^{CMM}, \dots, B_{|CMM|}^{CMM}], \quad (7)$$

where the Boolean variable B_i^{CMM} equals 1 when m_i (i.e., the i th CMM in M) is included in the formula.

IV. MOCHH ALGORITHM

In this paper, a multi-objective hyper-heuristic algorithm for clustering called MOCHH is proposed. The MOCHH algorithm contains two parts: one HLH and 12 LLHs. LLHs are used to receive the initial clustering solutions and perturb them by merging clusters, dividing clusters, or moving points between clusters. All LLHs then send the new clustering solutions back to the HLH. The HLH evaluates the received solutions and updates the solution pool. Finally, the HLH selects initial solutions for the next iteration by roulette wheel selection on the Pareto front and sends them to the LLHs. A flowchart for MOCHH is shown in Fig. 1.

TABLE 2. Medicinal categories and subcategories.

Category	Subcategory
Exterior-releasing medicinal (<i>Jiebiaoyao</i>)	Wind cold-dispersing medicinal (<i>Fasanfenghanyao</i>)
	Wind heat-dispersing medicinal (<i>Fasanfengreyao</i>)
Heat-clearing medicinal (<i>Qingreyao</i>)	Heating-clearing and fire-purging medicinal (<i>Qingrexiehuoyao</i>)
	Heating-clearing and dampness-drying medicinal (<i>Qingrezhaoshiyao</i>)
	Heating-clearing and detoxifying medicinal (<i>Qingrejieduyao</i>)
	Heating-clearing and blood-cooling medicinal (<i>Qingreliangxieyao</i>)
	Deficiency heat-clearing medicinal (<i>Qingxureyao</i>)
Purgative medicinal (<i>Xiexiyao</i>)	Offensive purgative medicinal (<i>Gongxiyao</i>)
	Laxative medicinal (<i>Runxiyao</i>)
	Drastic (purgative) water-expelling medicinal (<i>Junxiazhushuiyao</i>)
Wind dampness-dispelling medicinal (<i>Qufengshiyao</i>)	Wind dampness-dispelling and cold-dispersing medicinal (<i>Qufenghanshiyao</i>)
	Wind dampness-dispelling and heat-clearing medicinal (<i>Qufengshireyao</i>)
	Wind dampness-dispelling and bone and muscle-strengthening medicinal (<i>Qufeng-shiqiangjinguyao</i>)
Dampness-resolving medicinal (<i>Huashiyao</i>)	Dampness-resolving medicinal (<i>Huashiyao</i>)
Dampness-draining diuretic medicinal (<i>Lishuishenshiyao</i>)	Water-draining and swelling-dispersing medicinal (<i>Lishuixiaozhongyao</i>)
	Strangury-relieving medicinal (<i>Liniaotonglinyao</i>)
	Dampness-draining anti-icteric medicinal (<i>Lishituihuangyao</i>)
Interior-warming medicinal (<i>WenLiyaoyao</i>)	Interior-warming medicinal (<i>WenLiyaoyao</i>)
Qi-regulating medicinal (<i>Liqiyao</i>)	Qi-regulating medicinal (<i>Liqiyao</i>)
Digestant medicinal (<i>Xiaoshiyao</i>)	Digestant medicinal (<i>Xiaoshiyao</i>)
Worm-expelling medicinal (<i>Quchongyao</i>)	Worm-expelling medicinal (<i>Quchongyao</i>)
Hemostatic medicinal (<i>Zhixueyao</i>)	Blood-cooling hemostatic medicinal (<i>Liangxuezhixueyao</i>)
	Stasis-resolving hemostatic medicinal (<i>Huayuzhixueyao</i>)
	Astringent hemostatic medicinal (<i>Shoulianzhixueyao</i>)
	Meridian-warming hemostatic medicinal (<i>Wenjingzhixueyao</i>)
Blood-activating and stasis-resolving medicinal (<i>Huoxuehuayuyao</i>)	Blood-activating analgesic medicinal (<i>Huoxuezhitongyao</i>)
	Blood-activating menstruation-regulation medicinal (<i>Huoxuetiaojingyao</i>)
	Blood-activating trauma-curing medicinal (<i>Huoxueliaoshangyao</i>)
	Blood-breaking mass-eliminating medicinal (<i>Poxuexiaoweiyao</i>)
Phlegm-resolving, cough-suppressing and panting-calming medicinal (<i>Huatanzhikepingchuanayao</i>)	Cold phlegm-warming medicinal (<i>Wenhuhantanyao</i>)
	Heat phlegm-resolving medicinal (<i>Qinghuaretanyao</i>)
	Cough-suppressing and panting-calming medicinal (<i>Zhikepingchuanayao</i>)
Tranquilizing medicinal (<i>Anshenyao</i>)	Settling tranquilizing medicinal (<i>Zhongchenanshenyao</i>)
Liver-pacifying and wind-extinguishing medicinal (<i>Pingganxifengyao</i>)	Heart-nourishing tranquilizing medicinal (<i>Yangxinanshenyao</i>)
	Liver-pacifying and Yang-suppressing medicinal (<i>Pingyiganyangyao</i>)
Orifice-opening medicinal (<i>Kaiqiaoyao</i>)	Wind-extinguishing and spasmolytic medicinal (<i>Xifengzhijingyao</i>)
	Orifice-opening medicinal (<i>Kaiqiaoyao</i>)
Deficiency-tonifying medicinal (<i>Buxuyao</i>)	Qi-tonifying (<i>Buqiyao</i>)
	Yang-tonifying (<i>Buyangyao</i>)
	Blood-tonifying (<i>Buxueyao</i>)
	Yin-tonifying (<i>Buyinyao</i>)
Astringent medicinal (<i>Shouseyao</i>)	Exterior-securing anhidrotic medicinal (<i>Gubiaoazhihanayao</i>)
	Lung-intestine astringent medicinal (<i>Lianfeisechangyao</i>)
	Urine-tightening and leucorrhea-stopping medicinal (<i>Gujingsuoniaozhidaiyao</i>)
Emetic medicinal (<i>Yongtuyao</i>)	Emetic medicinal (<i>Yongtuyao</i>)
Poison-attacking, insecticidal, and antipruritic medicinal (<i>Gongdushachongzhiyangyao</i>)	Poison-attacking, insecticidal, and antipruritic medicinal (<i>Gongdushachongzhiyangyao</i>)
Detoxification-, putrescence- and granulation-promoting medicinal (<i>Baduhua-fushengjiyao</i>)	Detoxification-, putrescence- and granulation-promoting medicinal (<i>Baduhua-fushengjiyao</i>)

TABLE 3. Functional tendencies of chinese materia medica in *Sjuzi decoction*.

Symbol	Quantity
P	formulas set to be clustered
p_i	i th formula in P
M	CMM set including all CMMs included in P
m_i	i th CMM in M
MC	medicinal category set to which CMMs in set M belong
mc_i	i th medicinal category in MC
MSC	medicinal subcategory set to which CMMs in set M belong
msc_i	i th medicinal subcategory in MSC
FT	functional tendency set of CMMs in M
ft_i	i th functional tendency in FT

A. HIGH-LEVEL HEURISTIC

1) SOLUTION STRUCTURE

In MOCHH, a solution s is a set that contains several formula clusters. A MOCHH solution is described as

$$s = [c_1, c_2, \dots, c_N], \tag{8}$$

where c_i is the i th cluster in solution s and N is the number of clusters. Note that, as a formula set, no cluster can be empty, *i.e.*,

$$c_k \neq \emptyset. \tag{9}$$

Moreover, no one formula belongs to two clusters in any given solution. That is, the intersection between each pair of clusters in a solution is empty, *i.e.*,

$$c_i \cap c_j = \emptyset (i \neq j). \tag{10}$$

Finally, in our study, one formula is correspond to one syndrome type. No two clusters are correspond to the same syndrome type. Hence, every formula must belong to one cluster.

$$\bigcup_{i=1}^N c_i = P. \tag{11}$$

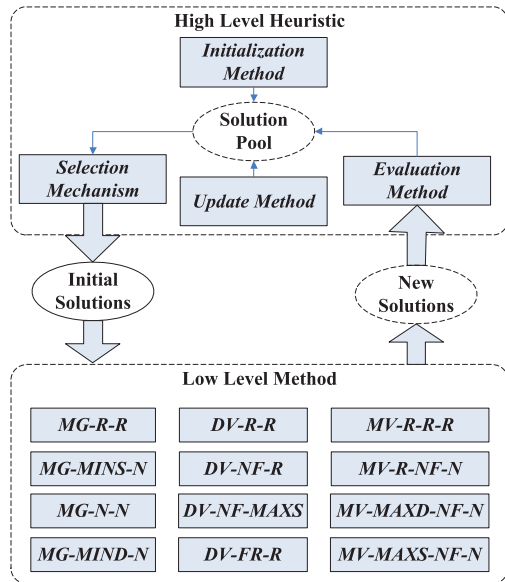


FIGURE 1. Flowchart for the MOCHH algorithm.

2) INITIALIZATION METHOD

LLHs in MOCHH are perturbation heuristics. That is to say, LLHs operate local searches based on a complete initial solution on the solution space. Hence, at the beginning of the process of MOCHH, an initialization method in the HLH was used to build an initial solution. In our proposed algorithm, k -means was adopted as the initialization method.

3) EVALUATION METHOD

As a multi-objective algorithm, the HLH makes use of two complementary objective functions to evaluate the new clustering solutions. One objective is based on the compactness of clusters, while the other objective is based on the connectedness of clusters.

We make use of the overall deviation of partitioning to express cluster compactness. The overall deviation is the sum, over all formulas, between the formula, and its cluster center.

$$Dev(s) = \sum_{c_k \in C} \sum_{p \in c_k} \delta(p, \mu_k). \quad (12)$$

Here C is the set of all clusters, μ_k is the centroid of cluster c_k , and the function δ calculates the distance between the formula p and the centroid μ_k . In our proposed algorithm, the Euclidean distance function is used for δ .

$$\delta(p_i, p_j) = \sqrt{\sum_{q=1}^r |x_{iq} - x_{jq}|^2}. \quad (13)$$

As an optimization of compactness, the overall deviation $Dev(s)$ should be minimized.

The other objective, connectivity, reflects the connectedness of clusters. The connectivity metric evaluates the degree to which neighboring data points have been placed in the

same cluster.

$$Conn(C) = \sum_{i=1}^N \left(\sum_{j=1}^h w_{i,nn_{ij}} \right) \quad (14)$$

where

$$w_{a,b} = \begin{cases} \frac{1}{j}, & \text{if } \nexists C_k : a \in C_k \wedge b \in C_k \\ 0, & \text{otherwise} \end{cases}. \quad (15)$$

Here nn_{ij} is the j th nearest neighbor of formula i , N is the number of formulas to be clustered, and h is a parameter quantifying the number of neighbors that contribute to the connectivity measure. As an optimization of connectedness, the connectivity objective metric $Conn(s)$ should be maximized.

4) UPDATE METHOD

After the compactness and connectivity are calculated, the solution pool is updated with new solutions based on Pareto optimality. For two solutions s_1 and s_2 , s_1 dominates s_2 (denoted as $s_1 < s_2$) if and only if

$$Dev(s_1) \leq Dev(s_2) \text{ and } Conn(s_1) > Conn(s_2) \quad (16)$$

or

$$Dev(s_1) < Dev(s_2) \text{ and } Conn(s_1) \geq Conn(s_2). \quad (17)$$

In the update method, the size of the solution pool S_p is N_s . All new solutions, and solutions in the solution pool S_p are placed into a new solution set S_t . The solution pool S_p is then cleared. In S_t , if solution s_1 dominates solution s_2 , then solution s_2 is placed in a new empty solution set S_d . If no solution in S_t dominates s_1 then s_1 is placed in the solution pool S_p . When S_t is empty, if the size of S_p is larger than N_s , then solutions are randomly selected for removal from S_p until the size of S_p equals N_s . On the other hand, if the size of S_p is smaller than N_s , then all solutions in S_d are placed in S_t . The solution pool S_d is then cleared, and the full set of steps above is repeated.

5) SELECTION MECHANISM

The HLH selects initial solutions and sends them to all LLHs. In this paper, we propose a roulette selection from solutions in the Pareto front (RouSPF). The set $conn_{sum}$ is the sum of the connectivities of the solutions in the Pareto front. The dev_{sum} is the sum of the overall deviations of the solutions in the Pareto front. Notably, merging of clusters by an LLH tends to increase both the connectivity and the deviation. The reason for this is that merging clusters reduces the number of clusters, automatically increasing both the connectivity and deviation. Accordingly, dividing clusters increases the number of clusters, automatically reducing both the connectivity, and deviation. In MOCHH, we stipulated that if a cluster contains only one formula, the formula cannot be moved to another cluster. Therefore, moving one formula from one cluster to another cluster never changes the number of clusters and thus has no intrinsic effects on connectivity and deviation.

For these reasons, the probability p_c of selection of solution s by the HLH for sending to the LLHs that merge clusters, is

$$p_c(s) = 1 - \frac{1}{2} \left(\frac{Conn(s)}{conn_{sum}} + \frac{Dev(s)}{dev_{sum}} \right). \quad (18)$$

On the other hand, the probability p_d of selection of solution s by the HLH for sending to the LLHs that divide clusters is

$$p_d(s) = \frac{1}{2} \left(\frac{Conn(s)}{conn_{sum}} + \frac{Dev(s)}{dev_{sum}} \right). \quad (19)$$

LLHs that move formulas from one cluster to another, act with less bias on the connectivity and the deviation than do those LLHs that merge or divide clusters. The HLH selects one solution at random from among the solutions in the Pareto front and sends it to the LLHs that move formulas from one cluster to another.

B. LOW-LEVEL HEURISTIC

As stated above, three kinds of LLHs were explored: those that merge two clusters into one cluster, those that divide one cluster into two clusters, and those that move formulas from one cluster to another cluster. Four LLHs of each type were explored. These 12 LLHs are described as follows:

- *MG-R-R*: Select two clusters at random and merge them into a single cluster.
- *MG-MINS-N*: Merge the cluster with minimum size and the cluster nearest to this cluster into a single cluster.
- *MG-N-N*: Merge the two nearest clusters into a single cluster.
- *MG-MIND-N*: Merge the cluster with minimum deviation and the cluster nearest to this cluster into a single cluster.
- *DV-R-R*: Select a cluster at random and randomly select a random number of formulas from this cluster and create a new cluster containing only these formulas.
- *DV-NF-R*: Select a cluster at random and find the formula farthest from the center of this cluster. Select a random number of formulas within this cluster that are nearest to this outlier and create a new cluster containing only these formulas.
- *DV-NF-MAXS*: Select the cluster with the maximum size and find the formula farthest from the centre of this cluster. Select a random number of formulas that are nearest to this outlier and create a new cluster containing only these formulas.
- *DV-FR-R*: Select a cluster at random and select a formula at random from within this cluster. Select all those formulas farther from the center of this cluster than the randomly selected formula and create a new cluster containing only these formulas.
- *MV-R-R-R*: Select two clusters at random to serve as a source and a destination. Move a random number of formulas from the source cluster to the destination cluster.
- *MV-R-NF-N*: Select a cluster at random and, within it, find the formula farthest from the center of this cluster.

TABLE 4. UCI benchmark dataset.

Datasets	No. of samples	Dimensions	No. of labels
iris	150	4	3
glass	214	9	6
zoo	101	17	7
machine	209	8	8

TABLE 5. Feature datasets of ATOH formulas.

Datasets	No. of samples	Dimensions	No. of labels
MC-B	214	21	13
MC-I	214	21	13
MSC-B	214	48	13
MSC-I	214	48	13
FT-B	214	194	13
FT-I	214	194	13
CMM-B	214	193	13

Select a random number of formulas that are nearest to this outlier and move them to the nearest cluster.

- *MV-MAXD-NF-N*: Select the cluster of maximum deviation and, within it, find the formula farthest from the center of this cluster. Select a random number of formulas that are nearest to this outlier and move them to the nearest cluster.
- *MV-MAXS-NF-N*: Select the cluster of maximum size and, within it, find the formula farthest from the center of this cluster. Select a random number of formulas that are nearest to this outlier and move them to the nearest cluster.

V. EXPERIMENTAL RESULTS

A. COMPUTATIONAL ENVIRONMENT

To evaluate the performance of the MOCHH algorithm, a series of experiments were conducted. The comparison algorithms included k -means and k -medoids as partitioning methods, DBSCAN as a density-based method, and GA as an evolutionary clustering algorithm.

All experiments were performed on an Intel Core i7-6500U running Microsoft Windows 7 at 2.50 GHz with 8 GB RAM.

B. DATASET DESCRIPTION

1) BENCHMARK DATASETS

In the experiments, two kinds of datasets were used to evaluate the performance of the MOCHH algorithm. The first was a set of UCI (University of California, Irvine) benchmark datasets: iris, glass, zoo, and machine. [30] The specific features of the datasets are shown in Table 4.

2) TCM FORMULA DATASET

We collected 214 formulas related to ancient thoracic obstruction and heartache (ATOH) from the ‘‘Dictionary of Traditional Chinese Medicine Formula’’ [31] and the ‘‘Chinese Medicine Encyclopedia.’’ [32] Seven datasets of ATOH formulas were generated based on the feature models described above. The ATOH formula datasets are shown in Table 5.

C. CLUSTERS VALIDATION INDICES

In this paper, the adjusted Rand index (ARI) was used as an external validation to measure the clustering accuracy of each

algorithm.

$$ARI = \frac{RI - E[RI]}{\max(RI) - E[EI]} \tag{20}$$

where RI is the Rand Index, $E[RI]$ is the Expected Rand Index, and $\max[RI]$ is the Maximum Rand Index.

Silhouette Coefficient (SC) was used as an internal validation indice. It takes a data point and checks how similar it is to its cluster in comparison to the other clusters.

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{21}$$

where a is the mean distance between a point and the points in that cluster. The value of a reflects the compactness of the cluster to which i belongs. And b is the mean distance between a point and the points in the next nearest cluster. The value of b captures the degree to which i is separated from other clusters.

D. PERFORMANCE EVALUATION ON UCI DATASET

We next evaluated the performance of the proposed MOCHH algorithm by comparing it with that of other algorithms, including k -means, k -medoids, DBSCAN, and GA on UCI datasets. The best results of ARI and SC over 30 replications of each algorithm for each of the four UCI datasets are shown in Figs. 2 and 3, respectively.

In terms of ARI and SC, the performance of k -means was better than k -medoids. DBSCAN achieved the lowest ARI and SC values among the five algorithms on the zoo dataset. The reason was that the zoo dataset has the highest dimensions among the four UCI datasets. DBSCAN is suitable for low-dimensional datasets, and the parameters of DBSCAN need to be tuned based on the characteristics of the dataset. As an evolutionary clustering algorithm, GA did not perform as well as MOCHH on all UCI datasets. The reason was that GA fell into the trap of a local optimal solution earlier than MOCHH. MOCHH outperformed the other four algorithms for all four of the UCI datasets. The SC values for MOCHH ranged from 0.691 to 0.815. That implied that the clusters generated by MOCHH are dense and well-separated. The ARI values of MOCHH on the iris and zoo datasets were close to each other at 0.729 and 0.721, respectively. However, MOCHH had a lower ARI of 0.273 on the glass dataset. We believe that the reason is that the range of values for different attributes is quite different in the glass dataset. For example, the value range of Fe is $[0,0.51]$ and the value range of Si is $[69.81,75.41]$. The Euclidean distance metric of MOCHH does not represent the sample discrepancy well.

E. EVALUATION OF PERFORMANCE ON THE ATOH DATASETS

The best SC and ARI values over 30 replications are shown in Figs. 4 and 5, respectively, for various algorithms operating on the ATOH datasets, including MC-B, MC-I, MSC-B, MSC-I, FT-B, FT-I, and CMM-B.

Similar to the UCI dataset, the performance of k -means was better than k -medoids in terms of ARI and SC, and the

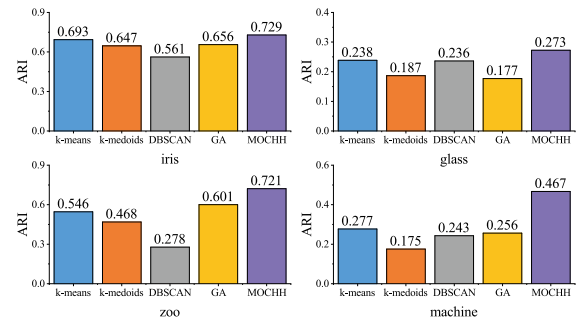


FIGURE 2. Best ARI values for various for the UCI datasets.

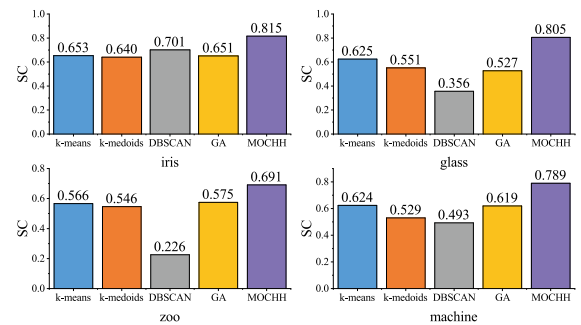


FIGURE 3. Best SC values for various for the UCI datasets.

values of ARI and SC were significantly lower on the ATOH datasets than the UCI datasets. This is because the ATOH datasets have a higher dimension than the four UCI datasets. Due to the high dimension of the ATOH datasets, the results generated by DBSCAN are worse than those generated by the other algorithms. GA did not perform as well as MOCHH on the ATOH datasets, except for the value of ARI on MC-B.

We note that MOCHH outperformed all other algorithms in terms of ARI for all ATOH datasets except MC-B. For the MC-B dataset, the ARI of 0.534 for MOCHH is very close to the best ARI observed, namely that of GA (0.537). The SCs of MOCHH are higher than those of the other four algorithms. Fig. 4 shows that the clustering accuracy is higher for the datasets generated based on the functional tendencies of CMMs, such as FT-B and FT-I, than for the other datasets. Our interpretation is that the functional tendencies of CMMs are more suitable for summarizing syndrome types than the categories, subcategories, and names of CMMs. Fig. 5 shows that the SCs of MOCHH on the ATOH datasets are lower than those on the UCI datasets, due to the high dimension of the ATOH datasets.

F. PERFORMANCE EVALUATION OVER DIFFERENT INITIAL NUMBERS OF CLUSTERS

For many clustering algorithms, the number of clusters strongly influences the output. However, LLHs in MOCHH can merge and divide clusters, so the number of clusters changes dynamically during the operation of the algorithm. Fig. 6 represents the ARIs generated by the various algo-

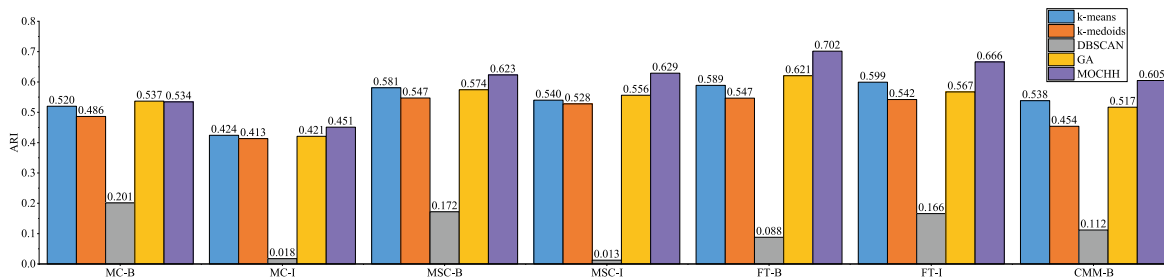


FIGURE 4. Best ARI values for various algorithms on the ATOH dataset.

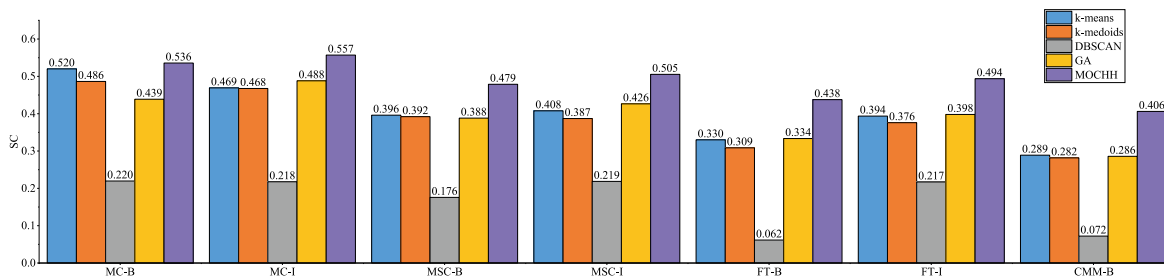


FIGURE 5. Best SC values for various algorithms on the ATOH dataset.

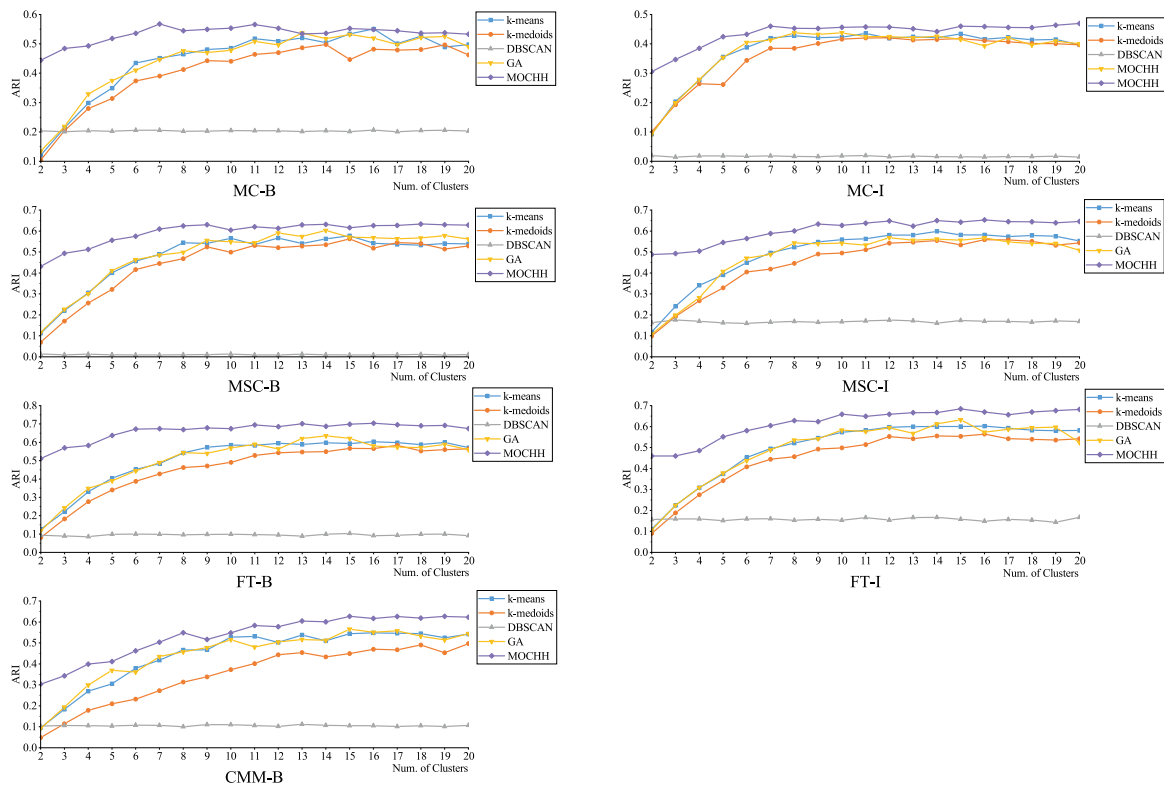


FIGURE 6. Best ARI values for various algorithms over various numbers of clusters.

gorithms for the initial cluster count ranging from 2 to 20. The number of labels for the ATOH datasets is 13. It can be clearly seen that ARIs generated by *k*-means, *k*-medoids, and GA form curves. “Elbows” in the curve of MC-I and MSC-B

are near 8-9. “Elbows” in the curve of MC-B, MSC-I, FT-B, FT-I, and CMM-B are near 12-13. ARI increases rapidly when the number of clusters is smaller than the elbow. The reason is that in this “under-fitting” region, more clusters are

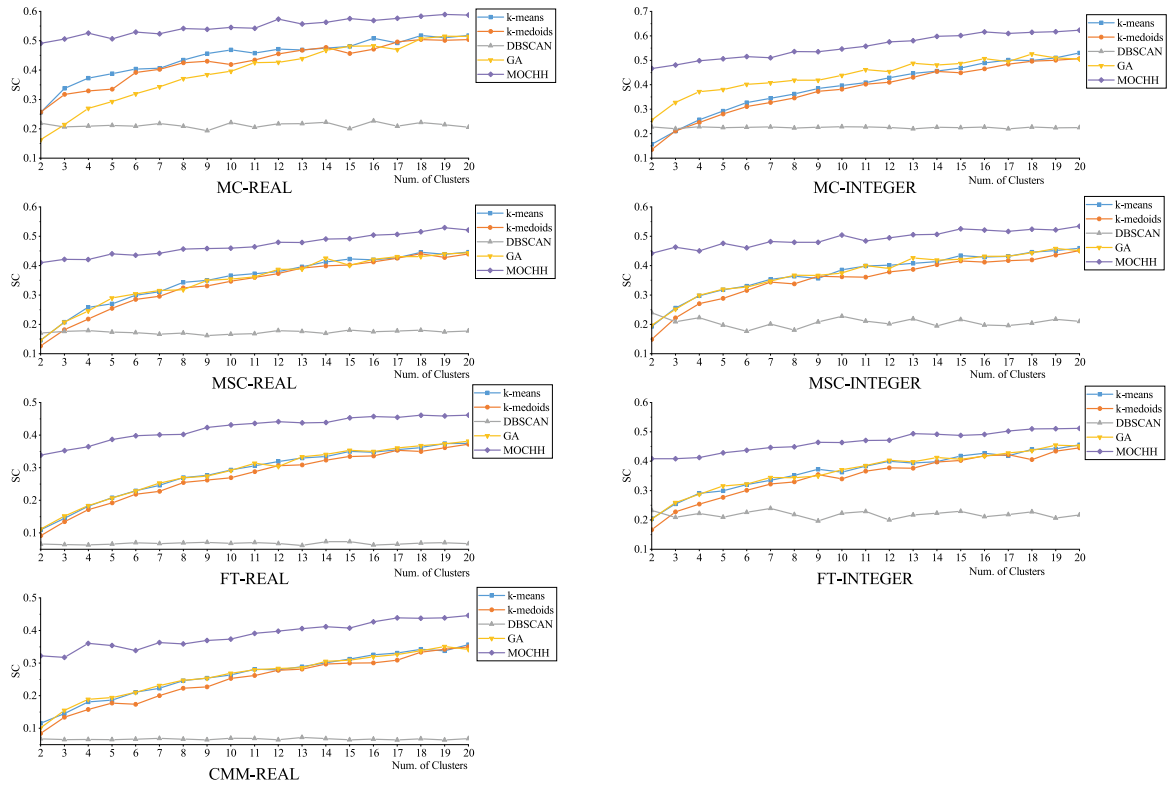


FIGURE 7. Best SC values for various algorithms over various numbers of clusters.

necessary. On the other hand, when the number of clusters is larger than the elbow, the ARI increases slowly. Because it is just subdividing the actual groups in the “over-fitting” region. We think the optimal number of clusters is near to the “elbow” in the curve.

The influence of the number of clusters on the results was stronger under *k*-means, *k*-medoids, and GA than that under MOCHH. Performance was less sensitive to the number of clusters under DBSCAN than under *k*-means, *k*-medoids, or GA. Nonetheless, the DBSCAN results were not as good the MOCHH results. When in the “under-fitting” region, ARIs generated by MOCHH are much higher than those generated by the other four algorithms. The reason is that even if the initial number of clusters is two, MOCHH can divide the clusters and generate more clusters to optimize the overall deviation and connectivity. Thus, the number of clusters generated by MOCHH is greater than two when the initial number of clusters is two.

Fig. 7 represents the SCs generated by all five algorithms for the initial cluster count ranging from 2 to 20. As a density based clustering algorithm, the performance of DBSCAN is significantly less sensitive to the initial number of clusters. It can be seen that the values of SC generated by all algorithms except DBSCAN rise as the number of clusters increases. The reason is that the fewer clusters in ATOH datasets, the more formulas are included in one cluster. In Eq. 21, *a* will increase and *b* will decrease, then *s* will decrease. On the other hand, if the number of clusters increases, the value of SC increases.

When the number of clusters equals the number of formulas, in other words, when each cluster includes only one formula, then *a* equals 0 and *s* equals 1 in Eq. 21. The SC values generated by MOCHH are higher than those generated by the other algorithms and increase slowly. It also indicates that MOCHH can merge and divide clusters, so that the number of clusters changes dynamically, even if the initial number of clusters is too small or too large.

G. RESULTS OF MULTI-OBJECTIVE CLUSTERING

Fig. 8 shows the ARIs for 30 runs based on two objectives: connectivity and overall deviation. This visualization makes clear that higher ARIs are associated with smaller overall deviations and larger connectivity for all of the ATOH datasets. The Pareto front can be seen in the plane defined by connectivity and overall deviation.

Figs. 9 and 10 show the SCs for 30 runs based on the two objectives: connectivity and overall deviation, respectively. We use 2D scatter plots to represent the relationship between SCs and connectivities and overall deviations, respectively, but not 3D scatter plots as used for ARIs and the two objectives. The reason is that the trends of SCs vary based on the two objectives. Specifically, the SC increases rapidly as the connectivity decreases below the elbow. When the connectivity is above the elbow, the SC increases slowly as the connectivity decreases. We think it is because when the connectivity is high, for most formulas, the formula and its *n* nearest neighbours are in the same cluster, so the number

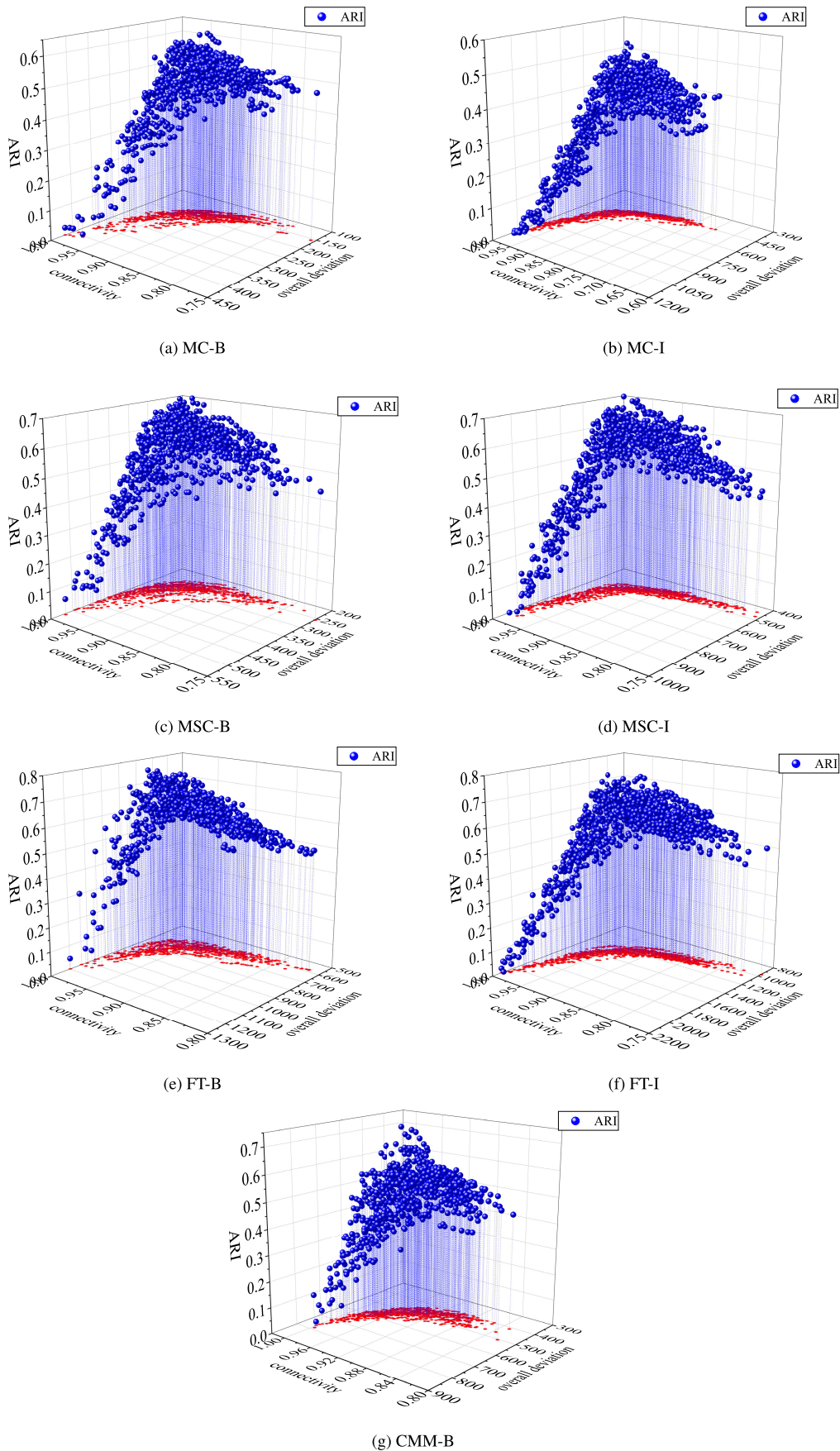


FIGURE 8. ARI values generated by multi-objective clustering.

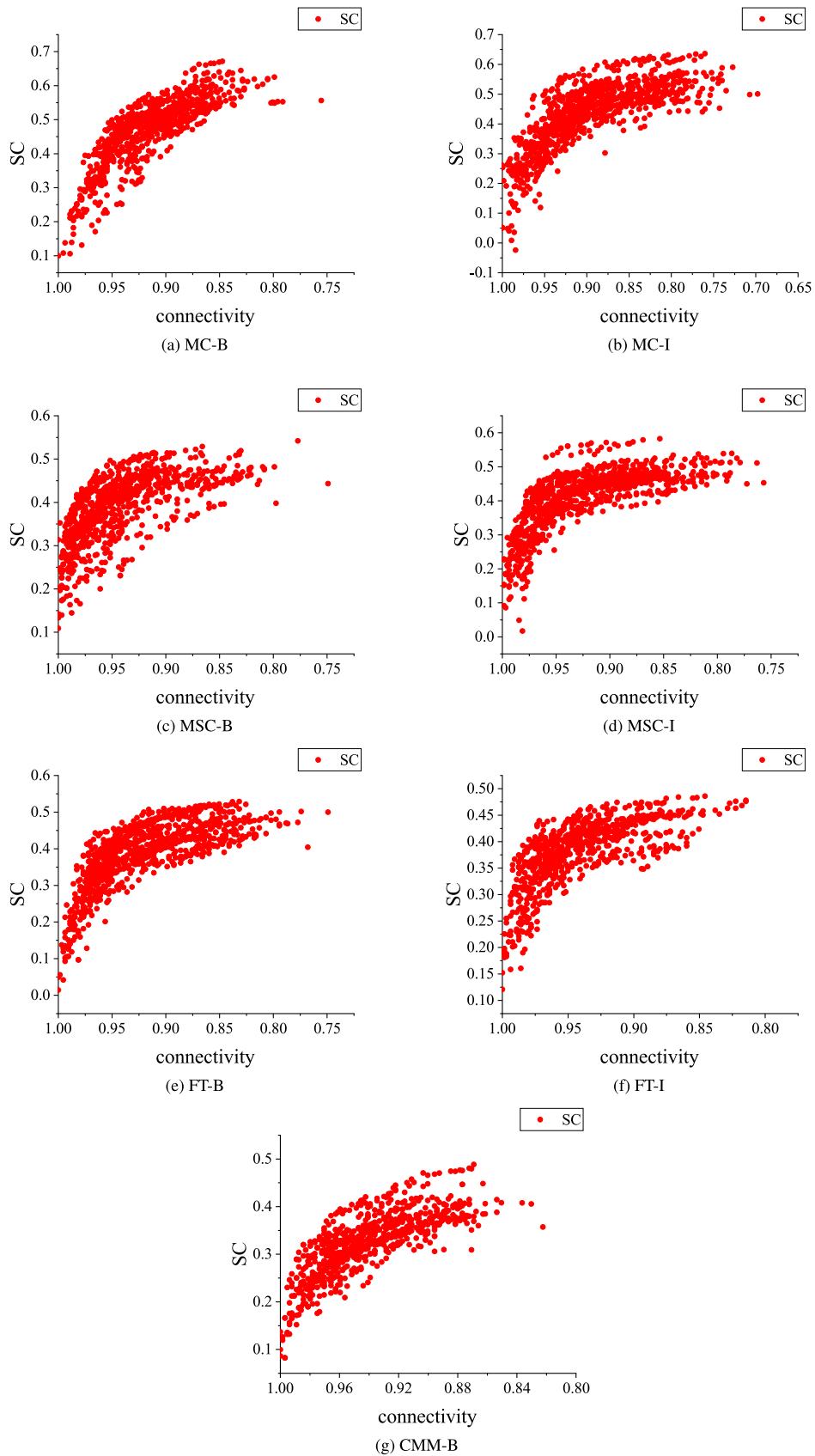


FIGURE 9. SC values generated with different connectivities.

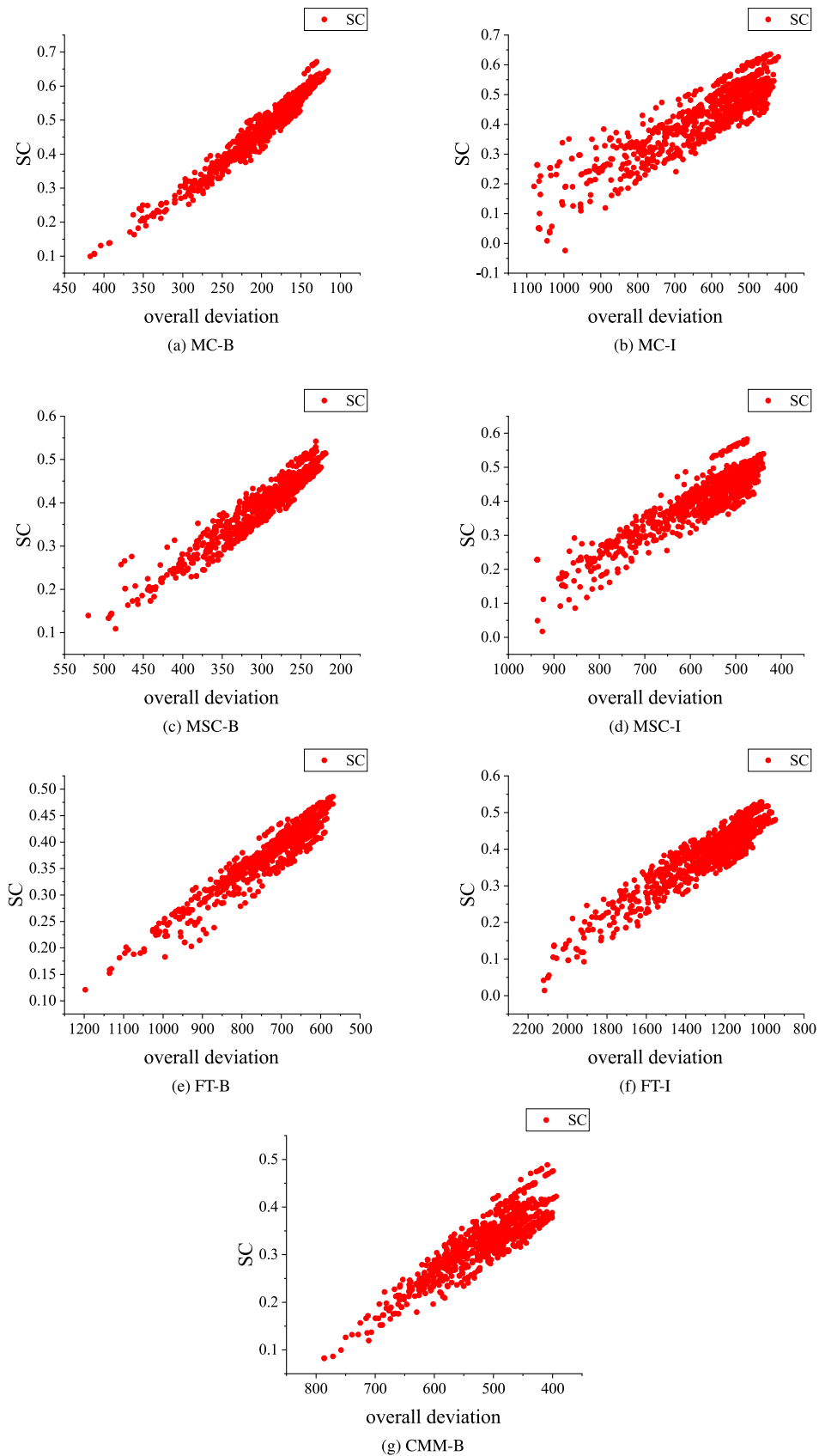


FIGURE 10. SC values generated with different total deviations.

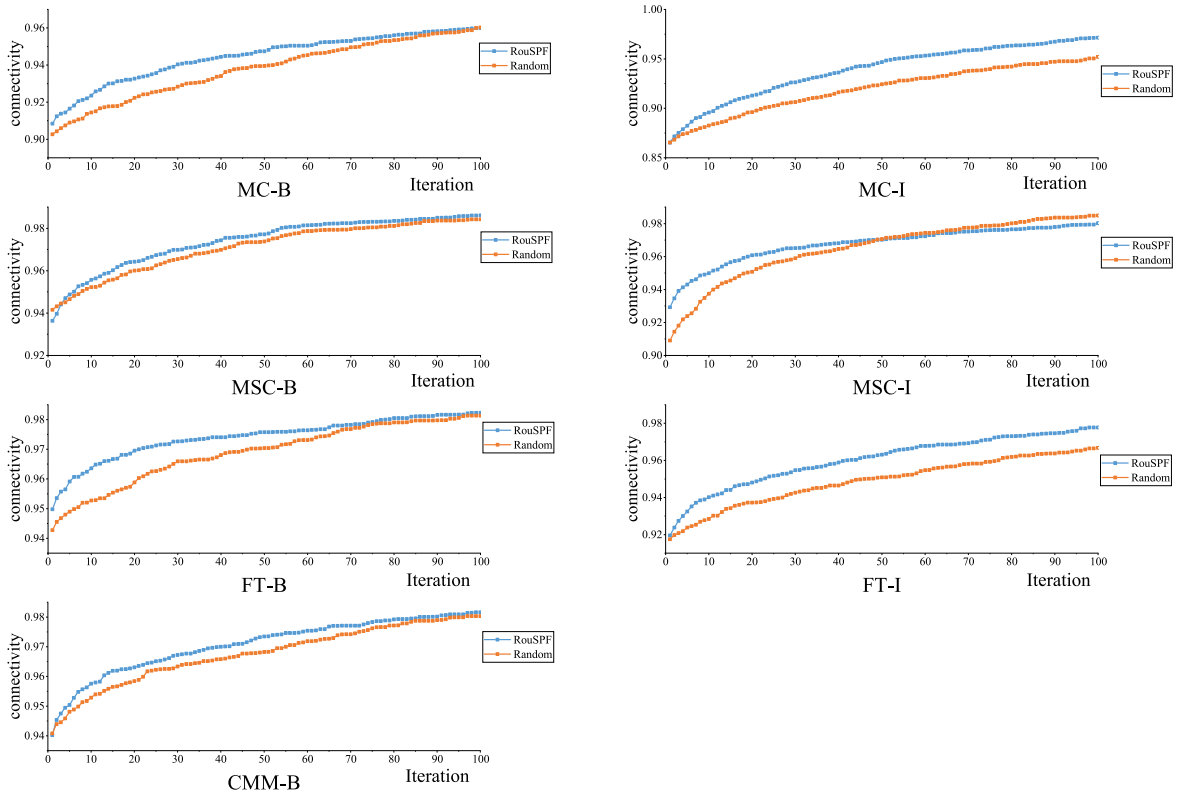


FIGURE 11. Connectivity in each iteration for each selection mechanism.

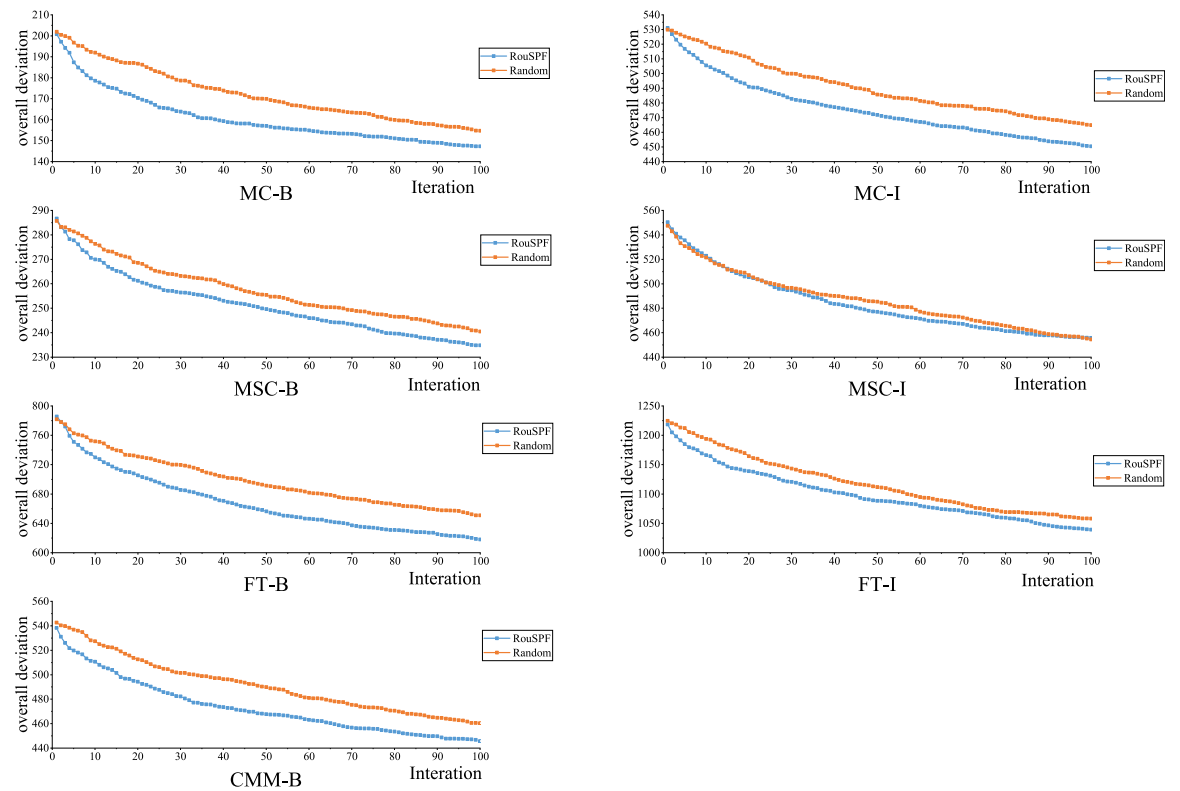


FIGURE 12. Overall deviation in each iteration for each selection mechanism.

of clusters is small. Then the less the number of clusters, the smaller SC is. When the connectivity is above the elbow, some formulas and their n nearest neighbours are not in the same cluster, so a decrease of the connectivity will only affect the other formulas. Therefore, the influence of the decrease of the connectivity on the value of SC is less than before the elbow. Fig. 10 shows that the SC increases linearly as in the overall deviation. The smaller the overall deviation, the higher the density clusters are, then the lower a in Eq.21 is, the higher the SC is.

H. PERFORMANCE EVALUATION OF SELECTION MECHANISMS

In this paper, we proposed a novel initial solution selection mechanism called RouSPF. To evaluate the performance of RouSPF, we compared RouSPF to a selection mechanism that selects a solution at random from the solution pool. In particular, we compared the best connectivity and overall deviation values produced by RouSPF and by the random selection mechanism during each iteration. Fig. 11 shows that higher connectivity values were found by RouSPF for several datasets, including MC-I, and FT-I. For the datasets MC-B, MSC-B, FT-B, and CMM-B, the results of the two selection mechanisms were very close. RouSPF exhibited a higher convergence rates for most datasets. However, RouSPF likely fell into local optima, since it selected solutions more likely to be improved by LLHs. For this reason, the best connectivity value found by RouSPF was worse than the best value found by the random selection mechanism for dataset MSC-I. Similarly, RouSPF showed higher convergence rates and better overall deviation results than the random selection mechanism for all datasets except MSC-I (Fig. 12).

VI. CONCLUSION

Syndrome differentiation is an important concept in TCM. Syndrome types can be summarized from not only the signs and symptoms of patients but also formula clusters as identified by TCM experts. In this paper, seven feature models of TCM formulas and a multi-objective hyper-heuristic clustering algorithm were proposed for the formula clustering problem. Experimental results demonstrate that MOCHH outperforms other clustering algorithms in most datasets. The number of clusters has less influence on the results for MOCHH than for other clustering algorithms. RouSPF shows higher convergence rates and accuracy than a random selection mechanism for most datasets. Accuracy was higher for feature models based on functional tendencies than for the other feature models. TCM experts can summarize syndrome types conveniently based on high accuracy automated clustering of TCM formulas.

In our study, the dosages of CMMs in formulas were not considered. The dosage of each CMM is recorded in most ancient formulas. Dosages can be seen as the weights of CMMs in the formulas. However, the weights of the CMMs are the same as in the case of neglecting dosages in our study. Therefore, further development may focus on the dosages of

CMMs in formulas when calculating the distance between two formulas to improve the clustering accuracy of MOCHH. Furthermore, as shown in Figs. 11 and 12, the RouSPF of MOCHH generates better results and exhibits higher convergence rates than a random selection mechanism on most datasets. However, RouSPF likely falls into local optima, since it selects solutions more likely to be improved by LLHs. A better selection mechanism for MOCHH is expected to be developed in the future works.

REFERENCES

- [1] Z. Huang, J. Miao, J. Chen, Y. Zhong, S. Yang, Y. Ma, and C. Wen, "A traditional Chinese medicine syndrome classification model based on cross-feature generation by convolution neural network: Model development and validation," *JMIR Med. Informat.*, vol. 10, no. 4, Apr. 2022, Art. no. e29290, doi: [10.2196/29290](https://doi.org/10.2196/29290).
- [2] L. Wang, J. Li, Y. Dang, R. Pan, and Y. Niu, "The association between health-related behaviors and traditional Chinese medicine syndromes in type 2 diabetes mellitus patients," *Diabetes, Metabolic Syndrome Obesity*, vol. 16, pp. 1977–1985, Jun. 2023, doi: [10.2147/DMSO.S409179](https://doi.org/10.2147/DMSO.S409179).
- [3] Y.-H. Lyu, L. Xie, W. Chen, J. Wang, X.-T. Wei, Y.-P. Wei, X.-P. Zu, and J.-X. He, "Application of metabonomics in study of traditional Chinese medicine syndrome: A review," *Zhongguo Zhong Yao Za Zhi*, vol. 47, no. 2, pp. 367–375, Jan. 2022, doi: [10.19540/j.cnki.cjcm.20210817.602](https://doi.org/10.19540/j.cnki.cjcm.20210817.602).
- [4] Y. He, H. Jiang, K. Du, S. Wang, M. Li, C. Ma, F. Liu, Y. Dong, and C. Fu, "Exploring the mechanism of Taohong Siwu decoction on the treatment of blood deficiency and blood stasis syndrome by gut microbiota combined with metabolomics," *Chin. Med.*, vol. 18, no. 1, p. 44, Apr. 2023, doi: [10.1186/s13020-023-00734-8](https://doi.org/10.1186/s13020-023-00734-8).
- [5] T. Liu, M. Qin, X. Xiong, X. Lai, and Y. Gao, "Multi-omics approaches for deciphering the complexity of traditional Chinese medicine syndromes in stroke: A systematic review," *Frontiers Pharmacol.*, vol. 13, Sep. 2022, Art. no. 980650. [Online]. Available: <https://www.frontiersin.org/articles/10.3389/fphar.2022.980650>
- [6] G. Hamerly and C. Elkan, "Learning the k in k -means," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 16, 2003, pp. 1–8, Accessed: Feb. 4, 2023. [Online]. Available: <https://proceedings.neurips.cc/paper/2003/hash/234833147b97bb6aed53a8f4f1c7a7d8-Abstract.html>
- [7] H.-S. Park and C.-H. Jun, "A simple and fast algorithm for K-medoids clustering," *Expert Syst. Appl.*, vol. 36, no. 2, pp. 3336–3341, Mar. 2009, doi: [10.1016/j.eswa.2008.01.039](https://doi.org/10.1016/j.eswa.2008.01.039).
- [8] H. Zhou, L. Li, H. Zhao, Y. Wang, J. Du, P. Zhang, C. Li, X. Wang, Y. Liu, Q. Xu, T. Zhang, Y. Song, C. Yu, and Y. Li, "A large-scale, multicenter urine biomarkers identification of coronary heart disease in TCM syndrome differentiation," *J. Proteome Res.*, vol. 18, no. 5, pp. 1994–2003, May 2019, doi: [10.1021/acs.jproteome.8b00799](https://doi.org/10.1021/acs.jproteome.8b00799).
- [9] J. Yang, Y. Wen, G. Zhao, and J. Duan, "Research on association rules of breast cancer and TCM : Syndrome based on data mining," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Dec. 2019, pp. 2788–2792, doi: [10.1109/SSCI44817.2019.9003024](https://doi.org/10.1109/SSCI44817.2019.9003024).
- [10] J. Song, L. Yang, S. Su, M.-Y. Piao, B.-L. Li, L.-X. Liang, G.-W. Zuo, Z.-M. Tang, Y.-Q. Long, X.-L. Chen, N. Dai, J.-L. Mo, Y. Yu, W.-Y. Yu, M. Zhang, R.-Q. Wang, J. Chen, and X.-H. Hou, "The diagnosis performance of the TCM syndromes of irritable bowel syndrome by gastroenterologists based on modified simple criteria compared to TCM practitioners: A prospective, multicenter preliminary study," *Evidence-Based Complementary Alternative Med.*, vol. 2020, Jul. 2020, Art. no. 9507674, doi: [10.1155/2020/9507674](https://doi.org/10.1155/2020/9507674).
- [11] S. Xia, J. Cai, J. Chen, X. Lin, S. Chen, B. Gao, and C. Li, "Factor and cluster analysis for TCM syndromes of real-world metabolic syndrome at different age stage," *Evidence-Based Complementary Alternative Med.*, vol. 2020, Jul. 2020, Art. no. 7854325, doi: [10.1155/2020/7854325](https://doi.org/10.1155/2020/7854325).
- [12] Y. Wang, Z. Chen, Y. Huang, L. Yafei, and S. Tu, "Prognostic significance of Serum interleukins and soluble ST2 in Traditional Chinese Medicine (TCM) syndrome-differentiated rheumatoid arthritis," *Med. Sci. Monitor*, vol. 24, pp. 3472–3478, May 2018, doi: [10.12659/MSM.907540](https://doi.org/10.12659/MSM.907540).

- [13] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. 2nd Int. Conf. Knowl. Discovery Data Mining*, Portland, OR, USA, Aug. 1996, pp. 226–231.
- [14] A. Said, R. A. Abbasi, O. Maqbool, A. Daud, and N. R. Aljohani, "CC-GA: A clustering coefficient based genetic algorithm for detecting communities in social networks," *Appl. Soft Comput.*, vol. 63, pp. 59–70, Feb. 2018, doi: [10.1016/j.asoc.2017.11.014](https://doi.org/10.1016/j.asoc.2017.11.014).
- [15] Z. Khan, S. Fang, A. Koubaa, P. Fan, F. Abbas, and H. Farman, "Street-centric routing scheme using ant colony optimization-based clustering for bus-based vehicular ad-hoc network," *Comput. Electr. Eng.*, vol. 86, Sep. 2020, Art. no. 106736, doi: [10.1016/j.compeleceng.2020.106736](https://doi.org/10.1016/j.compeleceng.2020.106736).
- [16] S. S. Ilango, S. Vimal, M. Kaliappan, and P. Subbulakshmi, "Optimization using artificial bee colony based clustering approach for big data," *Cluster Comput.*, vol. 22, no. S5, pp. 12169–12177, Sep. 2019, doi: [10.1007/s10586-017-1571-3](https://doi.org/10.1007/s10586-017-1571-3).
- [17] T.-X. Lin, Z.-H. Wu, and W.-T. Pan, "Optimal location of logistics distribution centres with swarm intelligent clustering algorithms," *PLoS One*, vol. 17, no. 8, Aug. 2022, Art. no. e0271928, doi: [10.1371/journal.pone.0271928](https://doi.org/10.1371/journal.pone.0271928).
- [18] M. A. Tawhid and A. M. Ibrahim, "An efficient hybrid swarm intelligence optimization algorithm for solving nonlinear systems and clustering problems," *Soft Comput.*, vol. 27, no. 13, pp. 8867–8895, Jul. 2023, doi: [10.1007/s00500-022-07780-8](https://doi.org/10.1007/s00500-022-07780-8).
- [19] E. K. Burke, M. Gendreau, M. Hyde, G. Kendall, G. Ochoa, E. Özcan, and R. Qu, "Hyper-heuristics: A survey of the state of the art," *J. Oper. Res. Soc.*, vol. 64, no. 12, pp. 1695–1724, Dec. 2013, doi: [10.1057/jors.2013.71](https://doi.org/10.1057/jors.2013.71).
- [20] J. G. C. Costa, Y. Mei, and M. Zhang, "Cluster-based hyper-heuristic for large-scale vehicle routing problem," in *Proc. IEEE Congr. Evol. Comput. (CEC)*, Jul. 2020, pp. 1–8, doi: [10.1109/CEC48606.2020.9185831](https://doi.org/10.1109/CEC48606.2020.9185831).
- [21] C.-W. Tsai, W.-L. Chang, K.-C. Hu, and M.-C. Chiang, "An improved hyper-heuristic clustering algorithm for wireless sensor networks," *Mobile Netw. Appl.*, vol. 22, no. 5, pp. 943–958, Oct. 2017, doi: [10.1007/s11036-017-0854-5](https://doi.org/10.1007/s11036-017-0854-5).
- [22] M. B. Bonab, S. Z. M. Hashim, T. Y. Haur, and G. Y. Kheng, "A new swarm-based simulated annealing hyper-heuristic algorithm for clustering problem," *Proc. Comput. Sci.*, vol. 163, pp. 228–236, Jan. 2019, doi: [10.1016/j.procs.2019.12.104](https://doi.org/10.1016/j.procs.2019.12.104).
- [23] A. C. Kumari and K. Srinivas, "Hyper-heuristic approach for multi-objective software module clustering," *J. Syst. Softw.*, vol. 117, pp. 384–401, Jul. 2016, doi: [10.1016/j.jss.2016.04.007](https://doi.org/10.1016/j.jss.2016.04.007).
- [24] H. Alshareef and M. Maashi, "Application of multi-objective hyper-heuristics to solve the multi-objective software module clustering problem," *Appl. Sci.*, vol. 12, no. 11, p. 5649, Jun. 2022, doi: [10.3390/app12115649](https://doi.org/10.3390/app12115649).
- [25] L. Liu, L. Han, D. Y. L. Wong, P. Y. K. Yue, W. Y. Ha, Y. H. Hu, P. X. Wang, and R. N. S. Wong, "Effects of Si-Jun-Zi decoction polysaccharides on cell migration and gene expression in wounded rat intestinal epithelial cells," *Brit. J. Nutrition*, vol. 93, no. 1, pp. 21–29, Jan. 2005, doi: [10.1079/BJN20041295](https://doi.org/10.1079/BJN20041295).
- [26] The Pharmacopoeia Commission of PRC, "Formula and single preparation," in *Pharmacopoeia People's Republic China*, vol. 1. Beijing, China: China Medical Science Press, 2015, pp. 782–783.
- [27] C. Liang, S. Zhang, and Z. Cai, "Effects of early intestinal application of Sijunzi decoction on immune function in post-operational patients of gastrointestinal tumor," *Zhongguo Zhong Xi Yi Jie He Za Zhi*, vol. 25, no. 12, pp. 1070–1073, Dec. 2005.
- [28] B. Gao, R. Wang, Y. Peng, and X. Li, "Effects of a homogeneous polysaccharide from Sijunzi decoction on human intestinal microbes and short chain fatty acids in vitro," *J. Ethnopharmacology*, vol. 224, pp. 465–473, Oct. 2018, doi: [10.1016/j.jep.2018.06.006](https://doi.org/10.1016/j.jep.2018.06.006).
- [29] Z. X. Zhou and D. C. Tang, "Efficacy of Chinese Medicine," in *Chinese pharmaceuticals*. Beijing, China: China Press of Traditional Chinese, 2020, pp. 17–19.
- [30] D. Dua and C. Graff. (2017). *UCI Machine Learning Repository*. [Online]. Available: <http://archive.ics.uci.edu/ml>
- [31] H. R. Peng, *Dictionary of Traditional Chinese Medicine Formula*. Beijing, China: People's Medical Publishing House, 1993.
- [32] *Chinese Medicine Encyclopedia*, Hunan Electronic and Audiovisual Publishing House, Hunan, China, 2006.



data mining for medicine.

WEN SHI received the B.S. degree in computer science and technology from Tianjin University, in 2006, the M.S. degree in traffic information engineering and control from the Civil Aviation University of China, and the Ph.D. degree in computer science and technology from Tianjin University, in 2015. He is currently a Lecturer with the School of Information Engineering, Tianjin University of Commerce, China. His current research interests include heuristics, machine learning, and



JINGYU ZHANG received the B.S., M.S., and Ph.D. degrees from the Tianjin University of Traditional Chinese Medicine. She was a Postdoctoral Researcher with Tianjin Medical University. She is currently a Physician with the Tianjin Nankai Hospital. Her current research interest includes gynecology.



BIN YU received the M.S. degree in computer science and technology from the Hebei University of Technology, in 2004. He is currently a Lecturer with the School of Information Engineering, Tianjin University of Commerce, China. His current research interests include machine learning and digital image processing.



YIBO LI is currently pursuing the degree with the Tianjin University of Commerce.



SHIHUI CHENG is currently pursuing the degree with the Tianjin University of Commerce.

• • •