A medical network clustering method with weighted graph structure

Hong Wu, Zijian Fu and Yizhou Wang

Abstract
Today, most of the databases used for drug information mining are derived from the collection of many treatments under a single disease, and some special drug compatibility rules can be found from them. However, researchers’ exploration of medical data is not limited to this. The comparative analysis of drugs for different diseases has become a new research point. In this paper, the drug is used as a node, the relationship is the edge connecting the two nodes, the co-occurrence frequency of the drug is used as the weight of the edge to establish a network graph. We use the clustering algorithm of the weighted network graph center diffusion method combining the network topology and the edge weights to divide the network graph into communities. Then we proposed the Structural Clustering Algorithm on Weighted Networks (SCW), it helps to study the prescription of medical prescriptions and provides more scientific recommendations for auxiliary prescriptions. In the experiment, SCW is compared with the classic community discovery algorithm CPM, the network function modular analysis algorithm MCODE and the hierarchical network graph structure analysis algorithm BGLL. We analyze the results according to NMI, ARI and F-Measure. Finally, a case study of real data was conducted to ensure the correctness and effectiveness of the algorithm, and to obtain the potential drug combination in the medical prescription.

Keywords
Weighted network graph, medical network, dense subgraph

Date received: 10 June 2020; accepted: 6 July 2020

Introduction
Traditional association rule mining has some drawbacks in the process of exploring drug compatibility. It can only analyze the medication rules for a disease based on a large number of medical prescriptions for a certain disease, and cannot take into account the use of drugs between related diseases. Exploring the common use of medicines between different diseases helps experimental researchers to explore the core medicines prescribed, and also provides clues and goals for clinical researchers to explore further. In this way, not only can we more objectively confirm and reproduce the existing disease medication laws, but also discover new knowledge and clues, and make more contributions to the research of medical data.

The development of the times has not only caused a sharp increase in the number of data, but also showed a diversified trend in the storage form, which includes text data, traditional relational data and network graph data. Compared with relational databases, graph databases have obvious advantages in performance, flexibility, and excavability. These advantages are mainly reflected in: (1) Performance. When the number and depth of the graph database increase, the query time is basically unchanged. But the data of the relational database is stored in a two-dimensional table structure in the form of records. When the number and depth of the relationships between the data increase, the time used by the relational database to query the relationship increases sharply, or even collapses. For example, in the communication network, querying the combination of two points whose maximum depth is a specified value. (2) Flexibility. The graph database data is extensible, and new nodes can be added at any time according to the needs. This will not affect current functionality. However, the relational database needs to consider the existing table structure when adding relationships, and the scalability is relatively low. (3) Excavability. The graph database focuses on the...
correlation between the stored data, which helps to discover and mine the potential relationship between nodes in the huge and complex relationship network. These relationships are difficult to find in traditional relational databases.

In recent years, research on various complex networks is one of the hotspots in many fields. For example, biological networks, social networks, e-mail networks, citation networks, etc. have become the main research objects of many scholars. Among them, medical data mining has always been a hot issue in the field of complex network graph data research. Cluster analysis is a commonly used data mining method that can extract unknown but potentially valid information from a huge database. However, the explosion of data volume makes network graph data extremely complicated. A series of nodes and edges are connected to each other to form a dense or sparse community structure. The nodes within these communities are closely related, while the nodes between communities are relatively sparsely related. Designing a reasonable algorithm to classify the nodes in the complex network graph according to the close relationship between the nodes helps us to better discover and utilize the potential functions and values in the network graph. Therefore, it is of great theoretical significance and application value to classify communities in complex networks.

At present, the research on complex networks is more oriented to undirected and unweighted networks. Saito et al. use k-dense algorithm to divide community networks. Newman et al. divided clusters based on graphs so that closely connected nodes were assigned to the same group. Then use spectral analysis clustering to express the relationship between the nodes as a distance. The closer the distance is, the higher the similarity between two nodes. When cutting graphs, cut the points that are closer to the same subgraph as much as possible, and ensure that the distance between different subgraphs is as far as possible. Lin et al. used hierarchical clustering. First treat each node as a cluster, then merge the two clusters with the closest distance into a new cluster, and repeat this process to form a dendrogram. Through this dendrogram, clusters can be quickly divided according to demand. Bai et al. and Ren et al. use the DBSCAN density clustering algorithm to classify clustering categories according to the closeness of node distribution.

But in the real application background, the structure of the nodes in the network graph can be used as the prior information of clustering, and the edges between the vertices also have important information that characterizes the clustering of the network graph. For example, in the social network graph, the edges between objects are used to indicate the closeness between two nodes in the protein interaction network, confidence is used as the weight of protein edges. These show that we have an urgent need for the clustering method of network graphs with edge weights. The hierarchical greedy discovery algorithm proposed by Blondel et al.: fast modularity optimization (BGLL) algorithm. Each node is regarded as a cluster, and the merger of clusters is judged by the criterion of maximizing modularity increment. This process is iterated until the modularity no longer increases. Liu et al. used the CMC algorithm to search for maximal groups in all knowledge network graphs and calculated the weighted density as the basis for judging the closeness between clusters. However, these methods only refer to edge weights or topological structure, do not comprehensively consider the edge weights and topological structure.

Edge weights and topological characteristics of the network graphs are both important reference factors for community classification of complex network graphs. If only one factor is considered, the result of the community classification will be unsatisfactory and deviate from the actual situation, which is not conducive to people’s accurate understanding and analysis of the actual situation. Therefore, clustering of edge-weighted network graphs combined with topology is an important research direction today. In this paper, we used medicine as the node, and the relationship between medicines is used as the edge to build a network graph. The processed drug co-occurrence frequency is used as the edge weight of the network graph, and then the network graph is used to divide the communities by the clustering algorithm of the weighted network graph center diffusion method combining the network topology and the network graph edge weight. To improve and optimize the idea of high-quality clustering algorithm, considering edge weights and topological characteristics of the network graphs, comparative experiments are performed on multiple data sets, and the experimental results are analyzed according to the evaluation indicators. After ensuring the correctness and effectiveness of the algorithm, we applied the algorithm to the analysis of medical prescriptions. It is helpful for us to study the medication rules of diseases in medical prescriptions and discover new knowledge and clues, so as to provide more scientific auxiliary prescription recommendations.

Related definitions

The clustering algorithm refers to dividing the data set into several subgraphs according to certain rules, so that the data with lower similarity is allocated to different subgraphs, and the data with higher similarity are collected into the same subgraph.

Definition 1: (Graph) The network graph is composed of a limited number of non-empty sets and a finite number of vertices. It is usually defined as $G = (V, E)$, where: $G$ represents a network graph, $V = \{v_1, v_2, \ldots, v_N\}$ represents the set of vertices in $G$, and $E \subset V \times V$ represents the set of edges between
vertices \( V \). If \((v_i, v_j) \in E\), then the edge existing between nodes and \( v_i \) can be represented by \( e_{ij} \). If graph \( G \) is an undirected graph, then \( e_{ij} \in E \iff e_{ji} \in E \), it means that the edge set \( E \) is symmetrical.

**Definition 2:** (Subgraph) If there are graph \( G = (V, E) \), graph \( G' = (V', E') \), and at the same time meet the two conditions of \( V' \subseteq V \) and \( E' \subseteq E \), then \( G' \) is called a subgraph of graph \( G \). In particular, if \( V' = V \) and \( E' \subseteq E \), the \( G' \) is called the generated subgraph of graph \( G \).

**Definition 3:** (Undirected weighted graph) On the basis of Definition 1, let \( w_{ij} \) denote the weight of the edge connecting the nodes \( v_i \) and \( v_j \), if \( e_{ij} \in E \), then \( w_{ij} \geq 0 \). Otherwise, if \( (v_i, v_j) \notin E \), then \( w_{ij} = 0 \).

**Definition 4:** (Structural similarity) The similarity between nodes is a crucial indicator to measure the closeness. The more common nodes, the more similar and tighter the two nodes are in the network structure.

Jaccard defines structural similarity as shown in formula (1):

\[
\text{Jaccard}(v_i, v_j) = \frac{|U(v_i) \cap U(v_j)|}{|U(v_i) \cup U(v_j)|}
\]  

Jaccard’s definition of similarity is based on the structure of the network graph only; however, in the actual application background, there will be other parameters to characterize the network graph, such as the weight or confidence of edges. Many descriptions of the network graph parameters make Jaccard’s definition of the similarity of network nodes have certain limitations. Therefore, Hub Promoted and Hub Depressed considered the influence of the edge weight of connected nodes on the similarity, and proposed the definitions of formulas (2) and (3):

\[
\text{HP}(v_i, v_j) = \frac{|U(v_i) \cap U(v_j)|}{\min\{|U(v_i)|, |U(v_j)|\}}
\]

\[
\text{HD}(v_i, v_j) = \frac{|U(v_i) \cap U(v_j)|}{\max\{|U(v_i)|, |U(v_j)|\}}
\]

**Structural clustering algorithm on weighted networks**

**Center-diffusion strategy**

In complex network graphs, there are usually several clusters of specific function types that are related to each other. Among them, the functions of the nodes in the cluster are highly similar and closely related, so as to ensure the synergy of different types of clusters between graph data networks, and then excavating the cluster structure in the network, it is beneficial to explore the characteristics of clusters of different functional types in the network graph in a deeper level. Because the nodes in the cluster have a stronger correlation than other nodes, when we initially determine the cluster, we first determine several nodes with weaker association relationships as different initial clusters according to the characteristics of the node. Then take the initial cluster as the center, use the “center-diffusion” strategy to divide the closely related nodes into clusters, and iterate this process to form cluster sets with similar functional characteristics. Therefore, we propose a weighted network graph structure center diffusion clustering algorithm to realize the classification of different clusters.

The Structural Clustering Algorithm on Weighted Networks (SCW) combines the graph data network topology and the construction process of Kruskal’s minimum spanning tree. SCW first selects the data node with the largest weight as the initial cluster, and then uses the diffusion method of the predefined cluster evaluation function to iteratively update the weights to obtain a weighted dense subgraph. Finally, the weighted dense subgraph with higher repetition rate is fused, and the nodes that are not divided into the dense subgraph are assigned to the weighted dense subgraph by using the node’s degree of belonging to the cluster \( l(v, C) \). The algorithm is roughly divided into four steps:

The first step is to give the definition of the edge weight in the network graph based on the application background of the data graph;

In the second step, the node with the larger weight is selected as the central node according to the value of the edge weight in the network graph, the weight of the nodes with the same area and higher density of the central node is reduced. Iteratively this method selects the most suitable central node in turn;

The third step is to diffuse the initial central node into multiple dense subgraphs in accordance with the evaluation criteria of the subgraph density evaluation function and the difference degree function;

In the fourth step, the dense subgraphs with high coincidence degree are fused, and the unassigned nodes are allocated to the existing dense subgraphs according to the attribution function.

**Definition of graph data node weights**

The properties of edges in graph data networks can be expressed using edge weights. Edge weight is an important factor for describing the degree of association between nodes, and it has a very important influence on the effect of later graph clustering. The accurate description of the edge weights helps us to assign nodes to the correct cluster. The weight of the edge in the social network represents the frequency of contact and interaction between each individual node. In protein interaction networks, edge weights are often used to
express the credibility of node interaction edges. When constructing a graph data network, its practical significance needs to be considered, and the edge weights are defined in conjunction with the network topology and other special factors.

In order to more accurately define the correlation between nodes, we refer to the definition of node similarity proposed by Jaccard, and combine Hub Promoted and Hub Depressed to improve Jaccard’s method, using the similarity of network graph structure judgment and the initial definition of edge weights. We define the weight of the network, as shown in formula (4):

\[
w(v_i, v_j) = \alpha \cdot \frac{|U(v_i) \cap U(v_j)|^2}{\min\{|U(v_i)|, |U(v_j)|\}^2} + \beta \cdot \frac{|U(v_i) \cap U(v_j)|^2}{\max\{|U(v_i)|, |U(v_j)|\}^2} + (1 - \alpha - \beta) \cdot w^0(v_i, v_j)
\]

where \(\alpha\) is the weight of the similarity of the HP node and \(\beta\) is the weight of the similarity of the HD node. \(w(v_i, v_j)\) is the initial edge weight in the real network. \(U(v_i)\) is the set of nodes adjacent to node \(v_i\). Similarly, \(U(v_j)\) is the set of nodes adjacent to node \(v_j\). \(U(v_i) \cap U(v_j)\) is the set of nodes adjacent to node \(v_i\) and node \(v_j\) in common, the similarity of the two nodes topology network structure.

**Central node selection**

For complex network graphs, the central node of the subgraph is crucial. It is the topological center of the subgraph, has a degree higher than other nodes, and is in a dense area. In order to refer to the characteristics of its network topology, most of the nodes in the network graph are included in the subgraph diffused by the central node through as few iterations as possible. When selecting the central node, we should consider the size of the node degree and the degree of association with other surrounding nodes. The greater the node degree of node \(v_i\) and the greater the sum of the weights of adjacent nodes, the more important node \(v_i\) is in the network topology. Formula (5) shows numerical calculation method for the selection of the central node:

\[
w_d(v_i) = d_i \cdot \sum_{v_j \in U(v_i)} w(v_i, v_j)
\]

where \(U(v_i)\) is a set of nodes adjacent to the node \(v_i\), \(d_i\) is the value of the degree the node \(v_i\), \(w(v_i, v_j)\) is the weight value of the edge after node \(v_i\) and node \(v_j\) reconcile HP node similarity and HD node similarity.

According to the calculation method above, we select the node with the largest value of \(w_d(v_i)\) as the first central node. Since the network topology diagram shows that nodes have a relatively high density in the region and a relatively low density in the relative region between regions. When we select node \(v_i\) as the central node, we try to reduce the weight \(w_d\) of the node in the same area which has a higher degree of association with \(v_i\). Therefore, after finding another subgraph, we need to reduce the edge weight \(w(v_i, v_j)\) of the subgraph node, recalculate the edge weight of the node, and then select the center node in turn.

**The central diffusion rule of the initial cluster**

In the process of expanding the subgraphs of other surrounding nodes to the central node, not all nodes with associated weights can be divided into the central node. If we do not formulate the boundary in advance, it may cause the nodes with lower correlation to be allocated to unrelated subgraphs, and it may also cause multiple subgraphs of different types to be considered as the same cluster, which will make the classification of clusters unclear. Therefore, we need a subgraph evaluation function as the evaluation metric, and the evaluation function of the density of the subgraph is shown in formula (6). This function could help us to determine whether the new node can be extended to the node in the subgraph.

\[
\text{Density}(S) = \sum_{v_i, v_j \in S} w(v_i, v_j) - \frac{1}{2|E(G)|} [(p - 1) \cdot p - 2q] \sum_{v_i, v_j \in S} w(v_i, v_j)
\]

where \(S\) is the connected subgraph of graph \(G\), \(V_s\) is the set of all vertices of subgraph \(S\), \(E_s\) is the set of all edges of subgraph \(S\), and \(p = |V_s|\), \(q = |E_s|\), \(w(v_i, v_j)\) is the weight of the edge between the node \(v_i\) and the node \(v_j\) in the subgraph \(S\).

When we judge whether the node \(v_k (v_k \notin V_s)\) should be allocated into the subgraph \(S\), we only need to judge whether the calculation result of the difference function \(\Delta \text{Density}(S) = \text{Density}(S \cup v_k) - \text{Density}(S)\) is > 0. If the result of \(\Delta \text{Density}(S)\) is < 0, the node \(v_k\) is not allocated to the subgraph \(S\), otherwise it is allocated. We iterate this evaluation process and operation mode until we cannot find the nodes that satisfy the conditions can be assigned to the subgraph \(S\), then we can obtain the subgraph \(S\) according to the current central node diffusion.

**Optimization of cluster partition**

The SCW algorithm we proposed optimizes subgraphs in two aspects:

On the one hand, according to the algorithm of node assignment to subgraphs shown in the last chapter, we can get a larger number of dense subgraphs. However, these dense subgraphs may have the phenomenon that the number of common nodes in the subgraph is too repetitive, which means that the divided subgraph types are too similar. For this phenomenon, we need to merge the subgraphs with a large number of overlapping nodes, so that the resulting clusters satisfy the principles that internal nodes are tightly connected and nodes between clusters are relatively distant. When (7),

\[
\]
**Algorithm 1.** Structural clustering algorithm on weighted networks

**Input:** Network graph $G = (V, E)$, where $G$ is a network graph, $V$ is the set of vertices in $G$, and $V = \{v_1, v_2, \ldots, v_N\}$, $E$ is the set of edges between vertices $V$, $E \subset V \times V$, and $W$ is the weight of the edge in the application background.

**Output:** The set of clusters $\Omega = \{C_1, C_2, \ldots, C_k\}$.

**Process:**
1. Let $p = |V|$, $q = |E|$, $T = V$.
2. Calculate the initial weight $W_{p \times p}(G)$ of the edge weight matrix between nodes in the network graph $G$ according to formula (4).
3. Calculate the adjusted weight $w_k(v)$ in the vertex set $T$ according to formula (5).
4. If $T$ is an empty set, jump to Step 5; otherwise, let $S = S + 1$, $C_0 = \emptyset$.
   4.1. Select the node $v$ with the largest weight $w_k(v)$ in set $T$, $T = T - \{v\}$, $C_0 = C_0 \cup \{v\}$, and let the set of adjacent points of node $v$ be $\text{Adjacent}(v)$.
   4.2. Calculate the subgraph density evaluation function $\text{Density}(S)$ of the nodes in set $\text{Adjacent}(v)$ according to formula (6).
   4.3. Find a node $u$ that satisfies $\Delta r_u(v) = \text{Density}(C_0 \cup v) - \text{Density}(C_0) > 0$ and difference function $\Delta r_v(u) = \max_{v \in \text{Adjacent}(v)} G_i(y)$.
   4.4. If node $u$ exists, then let $C_0 = C_0 \cup \{u\}$ and jump to Step 4.3.
   4.5. If $|C_0|/p \leq 0.8$, jump to Step 4.
5. Calculate the number of overlapping nodes between each $C$ and other dense subgraphs in the set. If $C_m \cap C_j = \min\{|S_m, S_j|\} > 0.6$, then let $C_m = C_m \cup C_j$.
6. Renumber the merged dense subgraphs to obtain $\{C_1, C_2, \ldots, C_k\}$.
7. Let $T = T - \cup C_i$.
8. Let $x = x + 1$ and the set of adjacent points of node $x \in L$ is $\text{Adjacent}(x)$.
   8.1. Calculate the attribution function $l(x, C)$ of node $x$ and each adjacent point, and select the node $y = \max_{y \in \text{Adjacent}(x)} l(y, C)$ with the largest attribution. If $y \in C_k$, then $C_y = C_y \cup \{x\}$.
8.2. If the set $L \neq \emptyset$, jump to Step 8.
9. Output the set of clusters $\Omega = \{C_1, C_2, \ldots, C_k\}$.

**Experiments**

**Dataset**

In order to verify the performance of our clustering algorithm, we experiment on three classic real-world datasets with labels, which are the American college football network, Zachary’s karate club, and Books about US politics(http://www.orgnet.com/). The relevant information of our experimental datasets is shown in Table 1.

**Evaluation metrics**

In order to evaluate the performance of the clustering algorithm effect, we need to use some evaluation metrics, such as Normalized Mutual Information (NMI) and Adjusted Rand Index (ARI) to describe the clustering effect. We use these metrics to perform numerical calculations on the clustering results to obtain a score. A higher score indicates that the clustering result of our algorithm is more accurate, otherwise, a low score indicates that we need to improve our algorithm. First, we represent the verification label results on the original dataset as $O = \{O_1, O_2, \ldots, O_n\}$, and the clustering results as $\Omega = \{C_1, C_2, \ldots, C_m\}$. For $C_i$ in set $\Omega$ and $O_j$ in set $O$, we give $T_{i,j} = |C_i \cap O_j|$, $b_i = \sum_{j=1}^{n} T_{i,j}$ and $d_j = \sum_{i=1}^{m} T_{i,j}$, where $1 \leq i \leq m$, $1 \leq j \leq n$.

\[
\begin{align*}
\text{NMI} &= \frac{2 \sum_{i=1}^{m} \sum_{j=1}^{n} T_{i,j} \log \frac{n T_{i,j}}{b_i d_j}}{\sum_{i=1}^{m} b_i \log \frac{b_i}{n} + \sum_{j=1}^{n} d_j \log \frac{d_j}{n}} \quad (9) \\
\text{ARI} &= \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} T_{i,j} (\frac{h_{ij}}{2} + \frac{d_{ij}}{2} - \sum_{i=1}^{m} \frac{h_{ij}}{2} - \sum_{j=1}^{n} \frac{d_{ij}}{2})}{\sum_{i=1}^{m} \sum_{j=1}^{n} \frac{h_{ij}}{2} + \sum_{j=1}^{n} \frac{d_{ij}}{2} - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{h_{ij}}{2} - \frac{1}{2} \sum_{j=1}^{n} \sum_{i=1}^{m} \frac{d_{ij}}{2}} \quad (10)
\end{align*}
\]
We can also use the F-measure defined by the average of Precision and Recall as the evaluation metric to evaluate the similarity between the clustering results and the original dataset labels.\(^{31}\)

\[
F - \text{measure} = \frac{2 \times \text{Prec} \times \text{Rec}}{\text{Prec} + \text{Rec}} \quad (11)
\]

where

\[
\text{Prec} = \frac{|\{V_i \in \Omega : \exists O_j \in O \land OS(V_i, O_j) \geq \theta\}|}{|\Omega|}
\]

\[
\text{Rec} = \frac{|\{O_j \in O : \exists C_i \in C \land OS(C_i, V_j) \geq \theta\}|}{|O|}
\]

\[
OS(V_i, O_j) = \frac{(T_{ij})^2}{|V_i| \times |O_j|}, \theta = 0.2
\]

### Analysis of results

In the experiment, we compare our SCW algorithm with the classic community discovery algorithm CPM, MCODE, and BGLL. According to the evaluation metrics of the clustering algorithm, we can make a correct evaluation of the clustering results. To evaluate the performance of the clustering effect, we use the previously mentioned NMI, ARI, and F-measure as metrics to evaluate the similarity between the clustering results and the original dataset labels. The comparison of clustering results of various experimental algorithms is shown in Table 2:

In order to verify the effectiveness of our SCW clustering algorithm for weighted network graphs, we also use the classic algorithms CPM, MCODE, and BGLL on the three datasets of Karate, Books, and Football to experiment and compare the performance. And we use NMI, ARI and F-measure as metrics to evaluate clustering results. From the comparison in Figure 1, we can see that our proposed SCW algorithm performs well in the comparison experiment of community division effect.

### Medical case analysis

In order to verify the effectiveness and correctness of the SCW algorithm, the effect of the dense subgraph classification method combining edge weights and topology in real weighted network graphs. In the experiment, we use the medicines and diseases in the Compendium of Materia Medica as nodes, and use the relationship as an edge to construct a network graph, and the contribution rate of the drug as the weight of the edge. All formulas are constructed and connected according to the same rules, so as to obtain the network graph of medicines relationship.

Compendium of Materia Medica classifies crude medicines into three categories: plant medicines, mineral medicines, and animal medicines. The plant medicines include grass part, grain part, vegetable part, fruit part, and wood part; The mineral medicines include gold part, jade part, stone part, and halogen part. The animal medicines include insect part, scale part, bird part, beast part, and human part. Among

![Figure 1](image-url)
them, the type, smell and treatment of each medicinal material are structured in a fixed format, while the name of the disease, the use of medicines, dosage and processing methods are piled together in the form of text, the format is relatively scattered and inconsistent. For the traditional Chinese medicine Glycyrrhiza uralensis, its type, smell and treatment have fixed forms. In the “treatment” column, information such as drugs, dosage, and processing in the corresponding treatment methods for typhoid fever, sore throat, pulmonary heat sore throat and lung atrophy are all mixed together. In order to facilitate the in-depth exploration of the prescription law, it is necessary to extract the disease and medicine names, dosage units and processing methods contained in the string, and convert it into structured data to model its ontology, attributes and relationships. Specific steps are as follows:

In the first step, we build a Chinese medicine noun database.

In the second step, we use the JIEBA word segmentation tool for word segmentation, and repeated corrections to separate the disease names, medicine names, dosage units, and processing methods.

In the third step, we create a structured traditional Chinese medicine prescription and disease database, and get the medicine table, disease table, and treatment table.

Under the real background, the medicinal materials contained in the separated medical formulas were sorted out, and a total of 1892 medicinal materials and 11,096 cases of formulas were obtained to verify the effect of community division. Table 3 shows the list of the most frequent medicines. Table 4 shows partial prescription combination relationships.

The first step in community classification is to build a complete medical network graph based on the drug combination relationship of medical prescriptions, which helps us analyze drug data and association relationships. We use algorithms to build medical weighted network diagrams, and each medicinal material has a corresponding overview and weight. The weight comes from the combination of drugs contained in the medical formula, which represents the closeness of the association between the two medicinal materials. The core idea of the algorithm is as follows: we use the drug combination contained in these formulas as data, and associate the drug nodes in the formula two by two in the network graph to ensure that there are edges between any two nodes. If there is already an edge between the two drug nodes, the weight is increased by one. The core code is shown in Algorithm 2, which contains the entire process of establishing a medical weighted network graph.

Similarly, we define the initial weight $w'(v_i, v_j)$ of the node in the graph data as the value of the edge weight according to the degree of association between drugs. Then we combine Jaccard’s definition of structural similarity with Hub Promoted and Hub Depressed’s definition of edge weight and node similarity to obtain the weight of medical prescription network $w(v_i, v_j)$. The local drug co-citation network graphs after optimization of central node selection, diffusion, and community division are shown in Figure 2.

Figure 2(a) shows that Ephedra, Pinellia ternata, Glycyrrhiza uralensis, Apricot kernel, Asarum, Belamcanda and Pheretima are the main drugs for the treatment of asthma, where ephedra is the core drug.

| Medicine                   | Frequency | Medicine                | Frequency | Medicine       | Frequency |
|----------------------------|-----------|-------------------------|-----------|----------------|-----------|
| Glycyrrhiza uralensis      | 178       | Wolfiporia cocos       | 55        | Gleditsia sinensis | 35     |
| Coptis chinensis           | 99        | Rosa banksiae          | 49        | Astragalus     | 32       |
| Aconitum carmichaelii     | 89        | Cyperus rotundus       | 46        | Ephedra       | 30       |
| Angelica sinensis         | 83        | Apricot kernel         | 43        | Syringa       | 28       |
| Rehmannia glutinosa       | 71        | Atractylodes macrocephala | 43    | Foeniculum vulgare | 27     |
| Rheum hybridum            | 61        | Musk                   | 41        | Arctium lappa | 26       |

Table 3. Frequency statistics of some medicines.
These drugs have the effect of purging lung qi and reducing phlegm and activating blood, which is consistent with the actual clinical medicine.

Figure 2(b) shows that we can conclude that Cuscuta is the core medicinal material for high-frequency renal diseases through community classification. Cuscuta can be combined with Rehmannia glutinosa, Fructus lycii, Angelica sinensis, Cistanche deserticola, Salvia miltiorrhiza, Atractylodes macrocephala, where the Salvia miltiorrhiza is a medicine for promoting blood circulation and the rest are all tonic medicines. It is consistent with the clinical medicine.

Figure 2(c) shows that Angelica sinensis, Astragalus, Glycyrrhiza uralensis, Ligusticum chuanxiong, Cynanchum otophyllum, Cinnamomum cassia, Commiphora myrrha, Boswellia carterii, and Saposhnikovia divaricate are divided into the same community, where Angelica sinensis is the core medicinal material, Angelica sinensis, Astragalus, Glycyrrhiza uralensis, Ligusticum chuanxiong are tonics and drugs for invigorating blood circulation and eliminating stasis, Cinnamomum cassia, Commiphora myrrha, and Boswellia carterii are drugs for invigorating blood circulation and eliminating stasis, Cinnamomum cassia and Saposhnikovia divaricate are wind-cold-dispersing medicinal, which are consistent with the clinical medicine.

**Conclusion**

In existing network graph clustering methods, many researchers have realized the importance of graph topology and node attributes. They proposed some graph clustering methods based on graph structure and attributes, but these methods have some shortcomings such as inaccurate graph model or unsatisfactory clustering effect. In this paper, we briefly describe the reasons for the poor community classification results only considering the network graph edge weights or network graph topology on real datasets. And we propose the Structural Clustering Algorithm on Weighted Networks (SCW) combining network topology and network graph edge weights. The clustering algorithm draws on the idea of Kruskal’s algorithm, and first determines several nodes with weaker association relationships as different initial clusters according to the characteristics. Then we take the initial cluster as the center and use the “center-diffusion” strategy to divide closely related nodes into clusters. Iterating this process will eventually form cluster sets with dissimilar functional characteristics. The experiment compared SCW with CPM, MCODE, and BGLL. After analysis on three real network data sets, the results show that the algorithm SCW shows good performance in the evaluation indicators such as NMI, ARI, and F-measure. Finally, the SCW community classification algorithm is used to analyze the association relationship between treatment and medicine on medical case data. We get the commonly used medicine combination rules which are conducive to assisting medical diagnosis.

**Declaration of conflicting interests**

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

---

**Algorithm 2. Establishment of medical weighted network graph based on drug combination**

1. set treatmentList, w[x][y], x, y;
2. for (n=0, n < treatmentList.length, n ++ )
3. {
4.   set medicineList;
5.   for (i=0, i < medicineList.length-1, i ++ )
6.     for (j=i+1, j < medicineList.length, j ++ )
7.       x = medicineList[i];
8.       y = medicineList[j];
9.       w[x][y] ++ ;
10.   }  
11. end for
12. end for
13. end for

---

**Table 4. Partial prescription medicines combination relationship.**

| No  | Diseases               | Medicines combination                                                                 |
|-----|------------------------|---------------------------------------------------------------------------------------|
| 3047| Postpartum hemorrhage  | Ligusticum chuanxiong Angelica sinensis -  Paonia lactiflora -  E jiao -  Rehmannia glutinosa -  Artemisia argyi -  Glycyrrhiza uralensis |
| 3588| Chest pain             | Panax ginseng -  Atractylodes macrocephala -  Zingiber officinale -  Glycyrrhiza uralensis |
| 3589| Qi deficiency          | Panax ginseng -  Atractylodes macrocephala -  Wolfiporia cocos -  Glycyrrhiza uralensis -  Zingiber officinale -  Ziziphus jujuba |
| 8429| Hemostasis             | Kalimeris indica -  Eclipta prostrata -  Colophony -  Gleditsia sinensis leaves         |
| 9195| Typhoid bloating       | Platycodon grandiflorus -  Pinellia ternata -  Orange peel                           |
| 9341| Heat                   | Forsythia suspensa -  Astragalus -  Aquilaria agallocha -  Rheum hybridum              |
References

1. Gallagher R, Passmore MJ and Baldwin C. Hastened death due to disease burden and distress that has not received timely, quality palliative care is a medical error. *Med Hypotheses* 2020; 142: 109727.

2. Panicker RO and Sabu MK. Factors influencing the adoption of computerized medical diagnosing system for tuberculosis. *Int J Inf Technol* 2020; 12(2): 503–512.

3. Li Z, Zhang P, Xie N, et al. A novel three-way decision method in a hybrid information system with images and its application in medical diagnosis. *Eng Appl Intell* 2020; 92: 103651.

4. Yu W, Jia M, Fang X, et al. Modeling and analysis of medical resource allocation based on Timed Colored Petri net. *Future Gener Comput Syst* 2020; 111: 368–374.

5. Wang C-H, Lee T-Y, Hui K-C, et al. Mental disorders and medical comorbidities: association rule mining approach. *Perspect Psychiatr Care* 2019; 55(3): 517–526.

6. Shih W-F, Lin C-W, Wang W-F, et al. Association rule mining of care targets from hospitalized dementia patients from a medical center in Taiwan. *J Stat Manag Syst* 2018; 21(7): 1299–1310.

7. Kim S-Y, Hong SH, Park J-W, et al. Analysis of diagnostic decision in acupuncture from the actual functional dyspepsia patient’s clinical information. *Integr Med Res* 2020; 9(4): 100419.

8. Angles R, Arenas M, Barceló P, et al. Foundations of modern query languages for graph databases. *ACM Comput Surv* 2017; 50(5): 1–40.

9. Angles R and Gutiérrez C. Survey of graph database models. *ACM Comput Surv* 2008; 40(1): 1–39.

10. Estrada E. *The structure of complex networks: theory and applications*. Oxford: Oxford University Press, Inc, 2011.

11. Liu R, Feng S, Shi R, et al. Weighted graph clustering for community detection of large social networks. *Procedia Comput Sci* 2014; 31: 85–94.

12. A. Criado-Alonso, E. Battaner-Moro, D. Aleja, et al. Using complex networks to identify patterns in specialty mathematical language: a new approach. 2020, 10(1): 56–61.

13. Malliaros FD and Vaziriannis M. Clustering and community detection in directed networks: a survey. *Physica A* 2013; 393(4): 95–142.

14. Noy N, Gao Y, Jain A, et al. Industry-scale knowledge graphs: lessons and challenges. *ACM Queue* 2019; 17(2): 48–75.

15. Saito K, Yamada T and Kazama K. Extracting communities from complex networks by the k-dense method. In: *Proceeding of the 6th IEEE International Conference on Data Mining Workshops (ICDMW 2006)* (ed Patist JP), Hong Kong, China, 18–22 December 2006, pp. 300–304, Piscataway, NJ: IEEE Press.

16. Newman MEJ. Spectral methods for community detection and graph partitioning. *Phys Rev E* 2013; 88(4): 042822.

17. Lin CC, Kang JR and Chen J-Y. An integer programming approach and visual analysis for detecting hierarchical community structures in social networks. *Inf Sci* 2015; 299: 296–311.

18. Bai XY, Yang PL and Shi XH. An overlapping community detection algorithm based on density peaks. *Neurocomputing* 2017; 226: 7–15.

19. Ren J, Wang JX, Li M, et al. Identifying protein complexes based on density and modularity in protein-protein interaction network. *BMC Syst Biol* 2013; 7(suppl 4): S12.

20. Li D, Wu H, Gao J, et al. Uncertain knowledge reasoning based on the fuzzy multi entity bayesian networks. *CMC Comput Mater Con* 2019; 61(1): 301–321.

21. Blondel VD, Guillaume J-L, Lambiotte R, et al. Fast unfolding of communities in large networks. *J Stat Mech* 2008; 2008(10): 10008–100019.

22. Liu GM, Wong L and Chua HN. Complex discovery from weighted PPI networks. *Bioinformatics* 2009; 25(15): 1891–1897.

23. Hogan A, Blomqvist E, Cochez M, et al. Knowledge graphs. arXiv:2003.02320 [cs.AI], 2020.

24. Nascimento MCV and Carvalho ACP. A graph clustering algorithm based on a clustering coefficient for weighted graphs. *J Braz Comput Soc* 2011; 17(1): 19–29.

25. Lü LY and Zhou T. Link prediction in complex networks: a survey. *Physica A* 2011; 390(6): 1150–1170.

26. Aggarwal CC and Wang H. A survey of clustering algorithms for graph data. In: C Aggarwal and H Wang (eds.) *Managing and mining graph data*. Advances in database systems. Boston, MA: Springer.

27. Newman MEJ. Finding community structure in networks using the eigenvectors of matrices. *Phys Rev E* 2006; 74(3): 036104.

28. Zachary WW. An information flow model for conflict and fission in small groups. *J Anthropol Res* 1977; 33(4): 452–473.

29. Danon L, Blomqvist E, Cochez M, et al. Knowledge graphs. *Bioinformatics* 2009; 25(15): 1891–1897.

30. Zachary WW. An information flow model for conflict and fission in small groups. *J Anthropol Res* 1977; 33(4): 452–473.

31. Danon L, Diaz-Guilera A, Duch J, et al. Comparing community structure identification. *J Stat Mech* 2005; 2005(9): P09008.

32. Rand WM. Objective criteria for the evaluation of clustering methods. *J Am Stat Assoc* 1971; 66(336): 846–850.

33. Li XL, Wu M, Kwok CK, et al. Computational approaches for detecting protein complexes from protein interaction networks: a survey. *BMC Genomics* 2010; 11(Suppl 1): S3.