MEDINFO 2017: Precision Healthcare through Informatics
A.V. Gundlapalli et al. (Eds.)

doi:10.3233/978-1-61499-830-3-653

Research on Ratio of Dosage of Drugs in Traditional Chinese Prescriptions by Data Mining

Yu Xing-wen, Gong Qing-yue, Hu Kong-fa, Mao Wen-jing, Zhang Wei-ming

College of Information Technology, Nanjing University of Chinese Medicine, Nanjing, Jiangsu, China

Abstract

Maximizing the effectiveness of prescriptions and minimizing adverse effects of drugs is a key component of the health care of patients. In the practice of traditional Chinese medicine (TCM), it is important to provide clinicians a reference for dosing of prescribed drugs. The traditional Cheng-Church biclustering algorithm (CC) is optimized and the data of TCM prescription dose is analyzed by using the optimization algorithm. Based on an analysis of 212 prescriptions related to TCM treatment of kidney diseases, the study generated 87 prescription dose quantum matrices and each sub-matrix represents the referential value of the doses of drugs in different recipes. The optimized CC algorithm can effectively eliminate the interference of zero in the original dose matrix of TCM prescriptions and avoid zero appearing in output sub-matrix. This results in the ability to effectively analyze the reference value of drugs in different prescriptions related to kidney diseases, so as to provide valuable reference for clinicians to use drugs rationally.

Keywords:

Medicine, Chinese Traditional; Prescriptions; Data Mining

Introduction

Biclustering algorithms is a field of research that is being developed and its algorithm is widely used in gene expression data analysis [1-3], literature research hotspot analysis [4-6] and biological data analysis [7-8]. Using double clustering algorithms can obtain doses in different traditional chinese medicine TCM prescriptions, elaborate scientifically dose-effect relationship, give a reasonable choice of TCM dose, and lay the foundation for the safe and effective medication [9].

Current biclustering algorithms can be divided into traditional clustering, iterated greedy search biclustering, exhaustive enumeration biclustering and mathematical modeling biclustering [10]. Among them, traditional clustering, such as coupled two-way clustering (CTWC) based on hierarchical clustering [11], cannot avoid the influence of global clustering nor can-generate favorable local submatrices, despite of their simple realization. This paper adopts the improved CC algorithm based on the iterated greedy search for the prescription dose analysis. The computing speed of the algorithm is faster than the exhaustive enumeration biclustering [12], and can effectively avoid global disturbance to obtain valid biclustering submatrices.

Methods

CC biclustering algorithm

Being one of the earliest biclustering algorithms, Cheng-Church biclustering algorithm [13] was first proposed by Cheng and Church in 2001. Sticking to the principle of iterated greedy search, this algorithm can generate a submatrix meeting the regulated threshold value in each search.

Mathematical description of CC algorithm

CC biclustering algorithm introduces the concept of mean squared residue to describe degree of internal similarity or consistency of submatrices. $H(I,J)$ refers to a mean squared residue of the submatrix whose row number is $|I|$ and column number is $|J|$.

$$H(I,J)=\frac{1}{|I||J|}\sum_{i\in I,j\in J}(a_{ij}-a_{iJ}-a_{Ij}+a_{IJ})^2=\frac{1}{|I||J|}\sum_{i\in I,j\in J}RS_{ij}^2$$

$$a_{iJ}=\frac{1}{|J|}\sum_{j\in J}a_{ij}$$

$$a_{Ij}=\frac{1}{|I|}\sum_{i\in I}a_{ij}$$

$$a_{IJ}=\frac{1}{|I||J|}\sum_{i\in I,j\in J}a_{ij}$$

$$RS_{ij}=a_{ij}-a_{iJ}-a_{Ij}+a_{IJ}$$

In a submatrix, RS_{ij} is a residue; a_{iJ} is the mean of Row i, a_{Ij} is the mean of Column j; a_{IJ} is the mean of the submatrix. Where, the larger the value of $H(I,J)$ is, the smaller the degree of similarity within the submatrix is. The smaller the value of $H(I,J)$ is, the more consistent the internal value of the submatrix is. Set ε to be maximal threshold value of the mean squared residue. All matrices satisfying the condition of $H(I,J)\leq\varepsilon$ is adopted as the target output submatrices of CC biclustering algorithm.

To find a submatrix satisfying the condition of $H(I,J)\leq\varepsilon$ in one search. First, the target matrix should undergo row and column deletion. When the mean squared residue of certain row (column) of the submatrix is larger than ε, then:

$$R=\{i\in I;\frac{1}{|J|}\sum_{j\in J}RS_{IJ}(i,j)^2>H(I,J)\}$$

Then, the row (column) is deleted to efficiently reduce the value of its mean squared residue. During the whole column and row deletion process, every iteration deletes the maximum column or row with the maximal mean squared residue until the mean squared residue of the submatrix is smaller than the threshold value, ε. Then, the qualified matrix can be preliminarily obtained:

Since the submatrix obtained by deletion of columns and rows is not the maximal submatrix, it is necessary to increase rows and columns of the submatrix:

If the average of mean squared residue–of the column (row) within the submatrixis smaller than the threshold value, ε, then:

$$R=\{i\notin I;\frac{1}{|J|}\sum_{j\in J} RS_{IJ}(i,j)^2 \le H(I,J)\}$$

Then, substitute the column (row) into the submatrix. The mean squared residue of the newly-generated submatrix is smaller than or equal to ε, so qualified columns and rows are added to generate the maximal submatrix satisfying the output condition.

After the qualified submatrix is output, the random number substitution is conducted before the next round of searching. To be specific, a random number matrix whose size is similar to the submatrix is used to substitute the submatrix to destroy the original consistency of the matrix and to obtain different qualified submatrices in every search.

Applications of CC biclustering algorithm to prescription dose analysis

The research made the dose of the prescription dose into the dose of the original matrix , according to the horizontal axis name and the vertical axis square name. Each element in the matrix corresponds to the dose of a drug in a prescription , as shown in Table 1.

Table 2 is the use of traditional CC algorithm for traditional Chinese medicine prescription data on the double-cluster analysis to output some sub-matrix. The analysis set the random number range out of the prescription dose range (0-500g), excluding the effect of the random number on the resulting sub-matrices. In the analysis results, the sub-matrix contains more 0 values, and the number of iterations increases with the algorithm. The proportion of 0 values gradually increased, seriously affecting the generation of effective results. The reason for the above results is mainly due to the unique nature of TCM prescription dose analysis and the limitations of ordinary CC algorithm, and the specific existent problems are the following two aspects:

Table 1 – Prescription dose original matrix

Prescription	Angelica	Ginseng	Fluorite	Authentic colla corii asini
Radix Morindae Officinalis	0	22.5	0	0
Dragon Tiger Panacea	15	0	0	0
Ten Supplements Pill	60	60	0	0
Jiaotai Pill	0	0	0	0
Dendrobium Powder	0	0	0	0
Guoxia Pill	0	0	0	0

Prescription	Fangfeng root	Cnidium	Chuanxiong	Digupi	Angelica
Keel pill	1.5	0	0	0	0
Eyesight improving water pill	0	0	0	0	0
31 Shenqi Pill	0	0	0	0	0
Kanlijiji Pill	0	0	0	0	0
Suppressing yang wine	0.9	0	0	0	0.9

Prescription	Chinese Photinia leaf	Chrysanthemum	Chinese yam	Huaisheng Rehmannia	Ceylon
Luminous pill	0	0	0	0	0
Tianxiong pill	0	0	0	0	0
Rehmannia pill	0	0	0	0	0
Guben-jianyang-Pill	0	0	0	0	0
Lysol soup	0	0	0	0	0

First, since the number of drugs in every prescription is smaller than the total number of drugs in all prescriptions, 0 appears in many places of the matrix. The submatrix composed of 0 is exactly the output submatrix of CC biclustering algorithm, so the output result using the traditional CC biclustering algorithm is mostly a submatrix composed of 0. However, 0 does not make any sense for the TCM dose analysis. Worse still, it might influence validity of the target submatrix.

Second, in each search, CC biclustering algorithm should conduct the random value substitution. In the traditional CC biclustering analysis, though these random numbers might impose certain inference on the submatrix output, the results obtained are still reliable. However, in the prescription dose matrix, the original matrix has too many 0s. The interference of the random number might seriously influence the output of valid submatrices. Worse still, the random number submatrix might be output, thus seriously influencing reliability of results.

To sum up, the key to optimizing biclustering algorithms is to exclude the interference of 0 on the prescription dose matrix and to seek a method more effective than the random number substitution method.

Optimization of CC algorithm

Improvement of CC in this paper refers to the improvement strategy of the POBA [14]. The punishment strategy is introduced to eliminate 0 from the matrix during the calculation process.

First, a punishment matrix, W, with a size similar to that of the prescription dose matrix is built. Let the initial value of the matrix element be 0. If the original prescription dose matrix element is 0, let the element value at the same position of the punishment matrix be 1.

Then, the calculation process of the mean squared residue by CC biclustering algorithm is optimized:

$$H(I,J)=\frac{1}{|I||J|}\sum_{i\in i,j\in J}(a_{ij}-a_{iJ}-a_{Ij}+a_{IJ})^2+(W_{ij}*\theta)$$

The above equation adds the "punishment formula" to the original calculation equation of the mean squared residue. W_{ij} stands for the element of the punishment matrix in Row i and Column j; θ controls the degree of punishment. In practical applications, 0 makes no sense to TCM prescription dose analysis. On the contrary, it might influence the output of the target submatrix. Therefore, in order to guarantee full cleaning of 0 in the output submatrix, let θ be:

$$\theta = |I|*|J|*\varepsilon$$

Then, there will contain no 0 in the output submatrix. In the following part, it is necessary to improve the column and row deletion and adding process of the submatrix so as to guarantee smooth generation of the target submatrix:

$$R=\{i\in I;\frac{1}{|J|}\sum_{j\in J} RS_{IJ}(i,j)^2+(W_{ij}*\theta) > H(I,J)\}$$

The above equation can delete columns or rows containing more 0s in advance during the column and row deletion process.

$$R=\{i\notin I;\frac{1}{|J|}\sum_{j\in J} RS_{IJ}(i,j)^2+(W_{ij}*\theta) \le H(I,J)\}$$

The above equation can guarantee adding of columns or rows not containing 0s during the column and row adding process.

In practical applications, let θ be:

$\theta = |I|*|J|*\varepsilon$

Finally, after the end of every search, let the element value of the punishment matrix whose position is the same to the output submatrix be 1. Replacing the random number substitution method with this method can guarantee full cleaning of 0s in the output submatrix and satisfy demands of outputting valid target submatrices.

Results

An original matrix was built for relevant TCM prescriptions to treat kidney diseases. The improved CC algorithm is used for biclustering analysis. The prescription dose submatrix is output to obtain referential value of different drugs in different prescriptions.

Building of the prescription dose original matrix

All prescriptions are from TCM Prescription Dictionary[6] with "stranguria", "impotence", "turbid semen", "kidney essence deficiency" and "retention of urine" as search terms. In total, 212 TCM prescriptions to treat kidney diseases are screened out.

The prescription data are further screened. With the prescription name as the row and the drug name as the column, the prescription dose original matrix is built. This is shown in Table 1 above.

Generation of the prescription dose submatrix

Using the improved CC clustering algorithm and setting the iterations to be 120 and ε be 10, the prescription dose original matrix outputs 87 prescription dose submatrices in total. Every submatrix stands for the mean dose of different drugs in different prescriptions. This is shown in Table 2 below:

Table 2 – Prescription dose submatrix

Prescription name	Schisandra chinensis	Cistanche	Fructus lyci
Yeguang Pill	15	15	23
Shihu Mingmu Pill	12.5	12.5	10
Gorgon Euryale Pill	30	30	30

Prescription name	White poria	Astragalus	Radix sileris
Longgu Pill	2.1	2.1	1.5
Rabbit Liver Pill	23	30	23

Prescription name	Shaved cinnamon bark	Notopterygium incisum
Radix Achyranthis Bidentate Pill	22.5	30
Tianxiong Powder	30	22.5

Discussion

Take one last submatrix as an example. Radix Achyranthis Bidentatae Pill and Tianxiong Powder contain shaved cinnamon bark and notopterygium incisum. In the former prescription, the dose of shaved cinnamon bark and the notopterygium incisum is 22.5g and 30g, respectively. The ratio between the two is 3:4. This prescription highlights the function of notopterygium incisum to relieve pains and treat rheumatism. In the latter prescription, the dose of shaved cinnamon bark and notopterygium incisum is 30g and 22.5g. The ratio between the two is 4:3. This prescription highlights the function of shaved cinnamon bark to treat wind-pathogenic kidney diseases. Thus, different doses of different drugs in a prescription follow certain rule. The flexibility of TCM prescriptions lies in its dose. Different doses focus on different treatment functions. Even if the same drugs are included in different prescriptions, different doses can result in different functions. In the clinical medication, the doctor should be based on the actual situation of each patient, according to the results of the analysis, and flexibly choose a different dose ratio to achieve different therapeutic purposes.

The iterated greedy search has its inherent defect, which is prone to get the local optimal solution. Even the optimized CC biclustering algorithm in this paper cannot get rid of the limitation. It is impossible to output high-quality submatrix in every iteration. Thus, it is necessary to increase iterations and further filter the final output submatrix using programming techniques.

However, the optimized CC algorithm can effectively eliminate the interference of 0s in the TCM dose original matrix, thus avoiding outputting meaningless submatrices containing 0s. This can greatly promote applicability of biclustering algorithm to TCM dose analysis. Meanwhile, without using the random number substitution method, the algorithm proposed in this paper can avoid negative influence of the random number on the output submatrices obtained through iterations.

Conclusion

The prescription dose ratio is an important part of the TCM prescription compatibility. A drug can have different doses in different prescriptions, thus playing a different role. Drug dose forms the soul of the prescription compatibility. Moderateness of drug doses can directly influence functions and clinical effects of prescriptions, and even life safety of patients. ZHANG Jiebin (1563~1640), a doctor of the Ming Dynasty, said that, "To create a disease, drug dose should not be too little nor too much. If being too little, the drug dose cannot help treat a disease; if being too much, the drug dose might cause other injuries and diseases." Therefore, doctors should attach great importance to the influence of drug dose on clinical prescriptions under the prerequisite of sticking to the principle of "Jun-Chen-Zuo-Shi". In this study, the improved biclustering algorithm was used to analyze the dose reference values of different drugs associated with nephropathy in different prescriptions. It can help clinicians optimize prescription, maximize the efficacy of prescriptions, and minimize the adverse effects of drugs on patients. To furtherly guide the physician to use drugs rationally based on the analysis results , combine with all aspects of environmental factors and the expected treatment results, flexibly choose a different dose ratio, and achieve optimal treatment.

Acknowledgements

This study was supported by The National Natural Science Foundation (Project number: 81674099).

References

[1] Williams, A. and S. Halappanavar, *Application of biclustering of gene expression data and gene set enrichment analysis methods to identify potentially disease causing nanomaterials* [J]. Beilstein J Nanotechnol, 2015. 6: 2438-2448.

[2] Gonzalez-Calabozo, J.M., F.J. Valverde-Albacete and C. Pelaez-Moreno, *Interactive knowledge discovery and data mining on genomic expression data with numeric formal concept analysis*. BMC Bioinformatics [J], 2016. 17(1): 374.

[3] Fu R.Y., Huang J., Hu B.Q., Pang C.Y., *Application of Alzheimer's Gene Expression Data to Compare Two Kinds of Hierarchical Clustering Methods* [J]. Journal of Sichuan Normal University (Natural Science Edition), 2015, (06): 925-929.

[4] Li, F., et al., *Mapping publication trends and identifying hot spots of research on Internet health information seeking behavior: a quantitative and co-word biclustering analysis. J Med Internet Res* [J], 2015. 17(3): e81.
[5] Jiang J.Z.,*Literature Research Hotspot analysis on Chinese medicine treatment of cerebral infarction* [J]. Science and Technology Information Development and Economy, 2015, (17): 140-142 + 160.
[6] Gong X.C., Zhao Y.G., An X.Y., *Study on the frontier of breast cancer related enzymes based on double clustering method* [J]. Chinese Journal of Medical Library and Information Science, 2016, (02): 69-74.
[7] Backman, T.W., D.S. Evans and T. Girke, *Large-scale bioactivity analysis of the small-molecule assayed proteome* [J]. PLoS One, 2017. 12(2): e0171413.
[8] Zhao Y.N., Wei Z., Wu X.L., Guo Y.Q., Sun C., Zhong S.N., Zhao C.Y., *Double Clustering Analysis Of Drug Resistance Of Gram-Negative Bacteria* [J].Chinese Journal Of Microecology, 2015, (05): 531-533
[9] Wang R. X. *Prescription Dose Pattern Research Based On Bicluster Algorithm* [J]. Liaoning Journal of Traditional Chinese Medicine, 2016, 01: 8-9.
[10] Zhang M. & Ge W. H. *The Research And Advances On Biclustering*[J].Microcomputer & Its Applications,2012,04:4-6+10.
[11] Getz G, Levine E, Domany E. *Coupled Two-Way Clustering Analysis Of Gene Microarray Data*[C]. In Proceedings Of The Natural Academy Of Sciences Usa, 2000:12079-12084.
[12] Tanay A, Sharan R, Et Al. *Revealing Modularity And Organization In The Yeast Molecular Network By Integrated Analysis Of Highly Heterogeneous Genomewide Data*[C]. Proc Natl Acad Sic U S A., 2004: 2981-6.
[13] Cheng Y, Church Gm. *Biclustering of Expression Data*[C]. Proc. Eighth Int'l Conf. Intelligent Systems for Molecular Biology (Ismb'00), 2000:93-103.
[14] Zhou Ch. *Research And Applications Of Bicluster Algorithm Based On High-Dimensional Data* [D]. Nanjing University of Science and Technology, 2009.
[15] Peng H. R. *Tcm Prescription Dictionary* [M]. People's Medical Publishing House Co., Ltd., 1996.

Addresses for correspondence

Gong Qing-yue (corresponding author): qygong@126.com
Yu Xing-wen: 437472167@qq.com