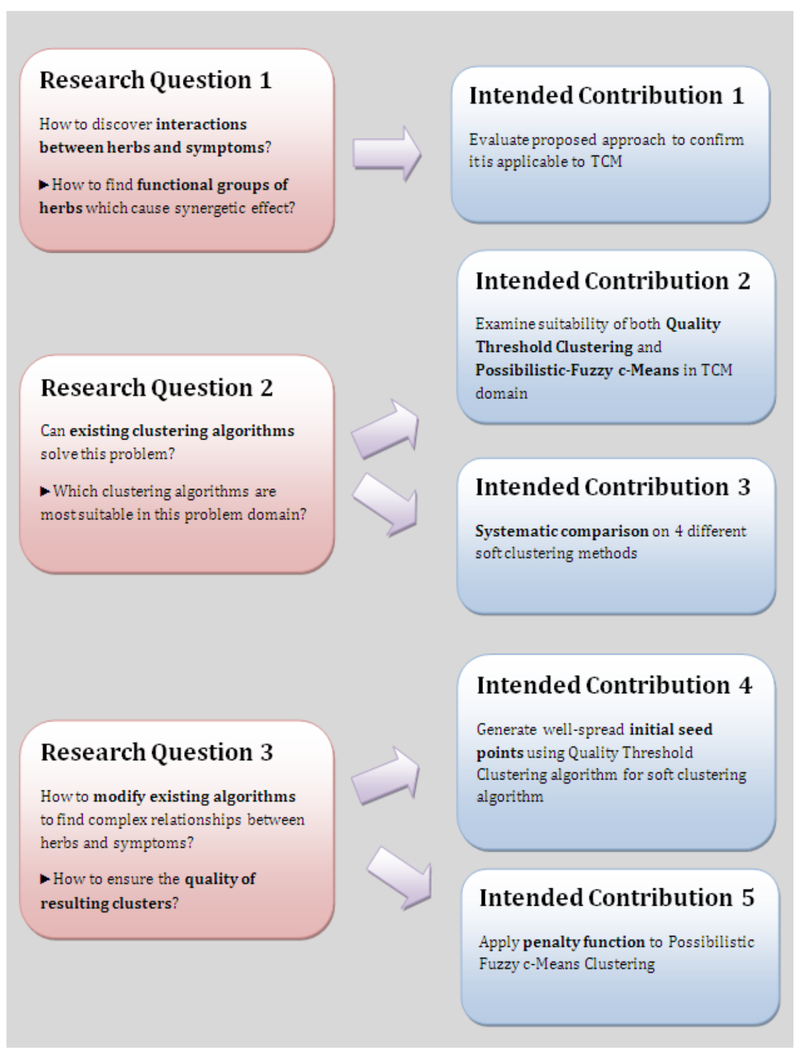
GENERATING SEEDS POINTS FOR SOFT-CLUSTER SYMPTOMS-HERBS RELATIONSHIP

Hwani Park and Josiah Poon (supervisor)
School of Information Technology
FACULTY OF ENGINEERING & INFORMATION TECHNOLOGIES

Introduction

- Traditional Chinese Medicine (TCM) has been beneficial over 4000 years
- Principal of Western Medicine and TCM are different in many aspects
- Yin-Yang Theory: Prescriptions in TCM highly depend on status of each patient
- Synergetic effect of TCM herbs are hidden, therefore there is a great need of systematic approach to discover the relationship between herbs.

Research Questions & Intended Contributions



Existing Clustering Techniques

- Clustering techniques are used to assign a set of objects into groups so that the objects in the same cluster are more similar to each other than to those in other clusters
- **Hard Clustering:** a set of data is divided into distinct clusters, where each data element belongs to exactly one cluster
- **Soft Clustering:** data elements can belong to more than one cluster, and associated with each element is a set of membership levels
- **Biclustering**: simultaneous clustering of rows and columns of a set of data

Quality Threshold Clustering [2]

- An algorithm that groups genes into high quality clusters.
- Quality is ensured by finding large cluster whose diameter does not exceed a given user-defined diameter threshold.
- This method prevents dissimilar genes from being forced under the same cluster and ensures that only good quality clusters will be formed.
- Similarity measure: Jackknife Correlation which is robust to single outliers
- A resulting cluster from QTC contains genes with Jackknife Correlation greater than a given threshold

For an ORF pair i, j, let denote the correlation of the pair i, j;

Also, let $p_{ij}^{(l)}$ denote the correlation of the pair i, j computed with the lth observation deleted.

For a data set with t observations, we defined the jackknife correlation

$J_{ij} = \min \{p_{ij}^{(1)}, p_{ij}^{(2)}, p_{ij}^{(3)} \ldots p_{ij}\}$

Figure 1: Jackknife Correlation

- **Advantages**: Quality guarantee, Number of clusters is not specified a priori, All possible clusters are considered
- **Disadvantages**: Computationally intensive, Time consuming, Hard clustering algorithm

Soft Clustering algorithms

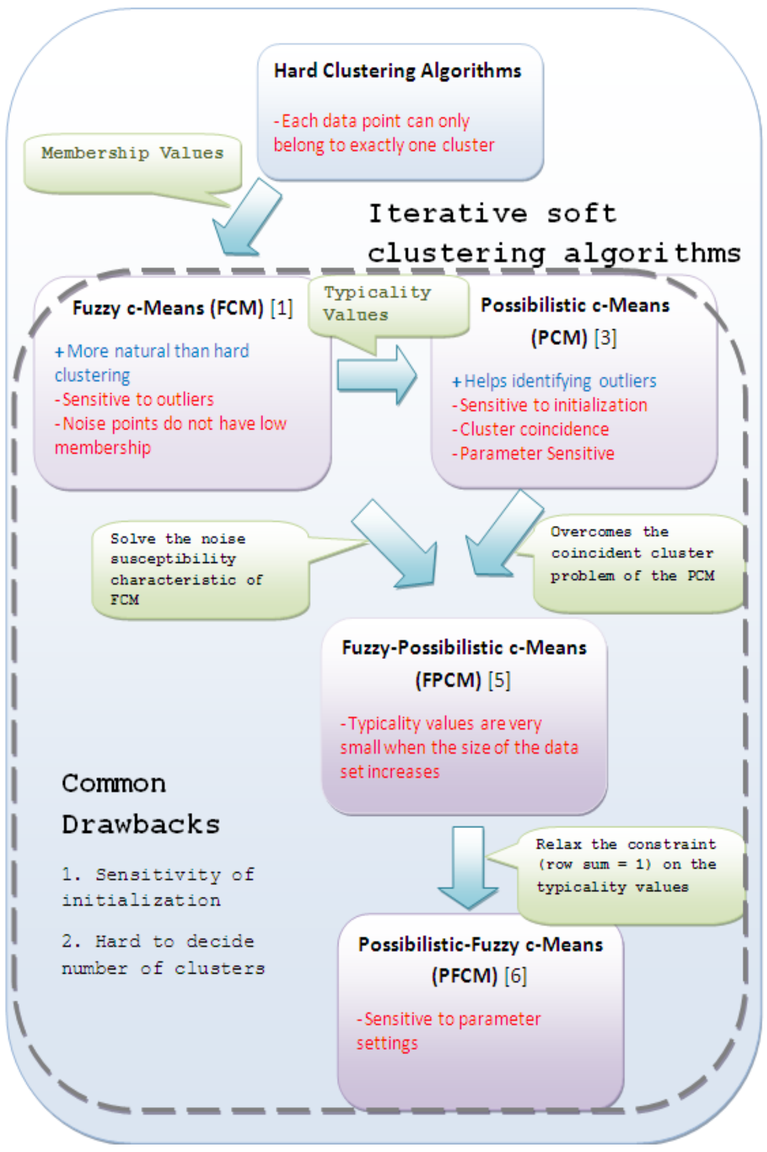


Figure 2: 4 Soft Clustering Algorithms

Combined QTC-PFCM

- In data preprocessing, use Modified Relative Success Ratio to create a table with herbs (rows) and symptoms (columns) ► Enables obtaining biclusters
- To overcome drawbacks of soft clustering, use Quality Threshold Clustering to determine number of clusters and initial seed points ► **Research Question 2 can be resolved**
- Apply penalty on distance values, which exceed user specified correlation threshold, used in determining typicality values. Therefore impact of outliers can be reduced ► **Research Question 3 can be resolved**
- Examining both membership values and typicality values obtained from proposed algorithm can define clusters of herbs and symptoms ► **Research Question 1 can be resolved**

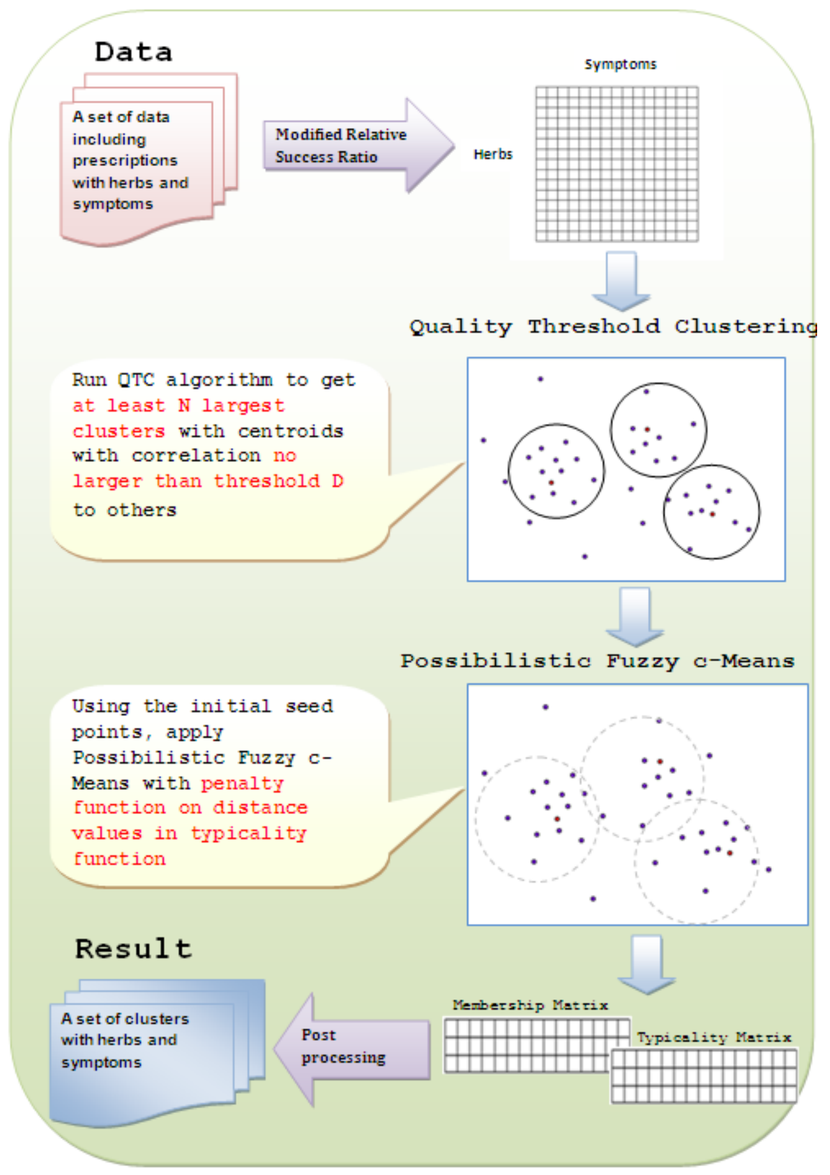


Figure 3: Process of QTC-PFCM

Experiments

- **On TCM data**: Using insomnia dataset, obtain results from the proposed approach and evaluate them to show QTC-PFCM can achieve good quality clusters of herbs and symptoms
- **On Non-TCM data**: Using non-TCM data (e.g. gene expression data) , obtain result from the proposed approach and evaluate them to prove the approach can solve other soft clustering problem

Discussions and Limitations

- Resulting centroids using QTC have correlation values smaller than given threshold value ►well-spread initial seed points
- By applying penalty function on distance values used in typicality function, outliers have less impact on computing centroids
- Iteratively computing Jackknife Correlation increases total computational cost
- Parameter sensitivity might be an issue

References

[1] Bezdek, J. (1981). Pattern Recognition with Fuzzy Objective Function Algorithms (Advanced Applications in Pattern Recognition), Springer.

[2] Heyer, L., S. Kruglyak, et al. (1999). "Exploring Expression Data: Identification and Analysis of Coexpressed Genes." Genome Research 9(11): 1106-1115.

[3] Krishnapuram, R. and J. M. Keller (1993). "A possibilistic approach to clustering." Fuzzy Systems, IEEE Transactions on 1(2): 98-110.

[4] Lukman, S., Y. He, et al. (2007). "Computational methods for Traditional Chinese Medicine: A survey." Computer Methods and Programs in Biomedicine 88(3): 283-294.

[5] Pal, N. R., K. Pal, et al. (1997). A mixed c-means clustering model. Fuzzy Systems, 1997., Proceedings of the Sixth IEEE International Conference on.

[6] Pal, N. R., K. Pal, et al. (2005). "A Possibilistic Fuzzy c-Means Clustering Algorithm." Fuzzy Systems, IEEE Transactions on 13(4): 517-530.