Analysis with word: Application to food nutrients classification

Chris Tseng[1, 2] Chingchia Li[1] Kevin Hung[1]

[1] CS department, San Jose State University, One Washington Square, San Jose, CA 95192, USA
[2] tseng@cs.sjsu.edu (corresponding author), Tel: 1-408-924-7255, Fax: 1-717-754-6112

Abstract

Fuzzy c-means is applied to 207 raw foods to classify what nutrients do they have in an Analysis with Word Engine. 27 nutrients of varying minerals and vitamins contained in each of these 207 foods are analyzed with a Analysis with Word engine. Nutrient content for each food is mapped from numerical to verbal description like high, medium, and low. With this analyzer we verify that the Zhang-Fu relationship in Chinese food cure theory. Most nutrients beneficial to the Zhang organ is shown to be also beneficial to its affiliated Fu organ.

Keyword: Fuzzy clustering. Food nutrient. Computing with Words. Web application.

1. Introduction

For large data set with uncertainties, fuzzy C-means is an ideal setting for data clustering. The wide spread use of this intelligent classification technology in various fields indicates that the concepts are gaining acceptance. With a given number of groups required, fuzzy c-means clustering technique can generate the needed number of data clusters as well as the representative centers of each these clusters. With fuzzy logic, each data point will have a corresponding membership index that show how much the data belongs to each of the cluster centers. Unlike the crispy logic methodologies that typically designate yes or no membership, fuzzy logic technique provides membership values that range from 0 to 1. In so doing, it can compensate for uncertainties and allow the system to rank the result in the analysis process. Note in the food and ingredients data, it is highly possible that some of the data may be imprecise as they collected from many other sources. This technique should prove to be more reliable and effective than the traditional statistical technique. The success of this FCM technique in data mining and pattern recognition can be found in numerous literatures [1, 2].

We will show in this paper how an Analysis with Word engine can make use of Fuzzy C-Means to obtain a qualitative relation between two related sets of data. The main idea of Analysis with Word using fuzzy c-mean will be described in section 2. In section 3, we apply the engine to validate a Zhang-Fu food cure theory in Chinese medicine. Future research is proposed in the conclusion.

2. Fuzzy C-mean clustering of food ingredients

We propose an Analysis with Word engine as shown in Figure 2.1. A typical fuzzy c-mean analysis can be described by the following.

$$\text{fcm}(\text{data}, n) = (\mu_1, \mu_2, \ldots, \mu_n) \qquad (2.1)$$

fcm outputs the locations of n cluster centers as well as the membership values of each datum to these centers. We specify the data set as 206 foods containment on a vitamin or mineral, and the number of cluster as 3 in our case for low, medium, high separately. For the 206 raw foods' nutrition facts – what kinds of nutrient a food contains and how much of each nutrient a food contains, we focus on the amount of the 27 vitamins and minerals that a serving of these raw foods contains.

As a result, we have a matrix as below:

Nutrition facts matrix

$$D = \begin{bmatrix} d_{1,1} & d_{1,2} & \ldots & d_{1,27} \\ d_{2,1} & d_{2,2} & \ldots & d_{2,27} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ d_{206,1} & d_{206,2} & \ldots & d_{206,27} \end{bmatrix} \qquad (2.2)$$

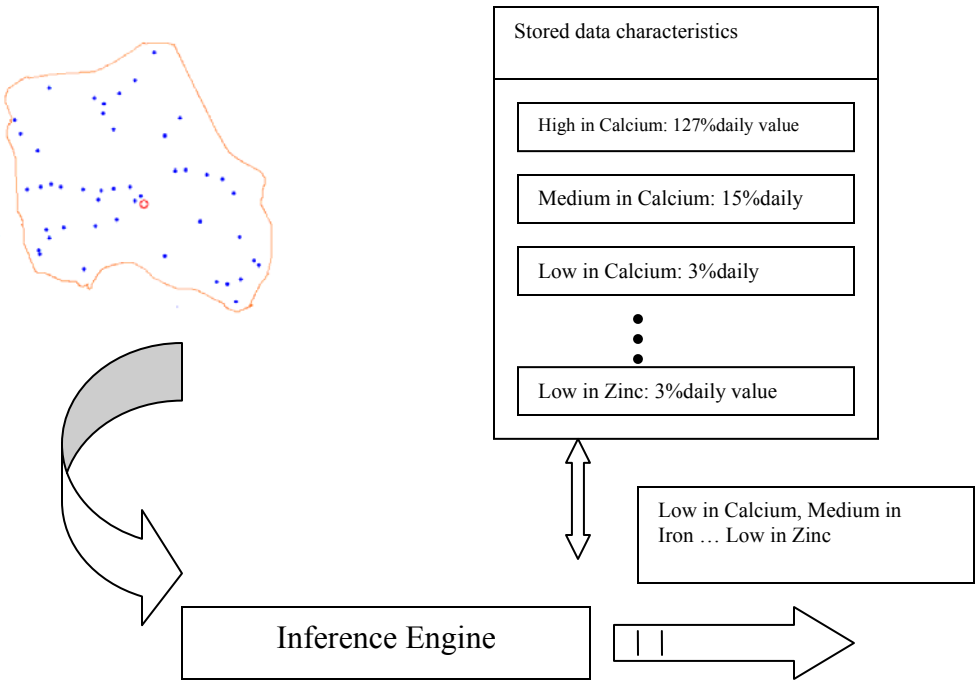

Fig. 2.1: Data flow of Inference Engine

Each row of the matrix records a raw food's nutrition facts on the 27 vitamins and minerals; each column of the matrix records the 206 raw foods' nutrition facts on one nutrition or mineral. Thus, for a vitamin or mineral N_j (j ϵ [1,27]), we have a data column with 206 pieces of data D[:,j] (":" means all data in a column of matrix and "j" specify the column number), each of which describes how much nutrient N_j a raw food contains. We substitute D[:,j] for data in (2.1) to get three clustering centers $\mu_{1,i}$, $\mu_{2,i}$, and $\mu_{3,i}$ of the data set ranging from low center, medium center, to high center respectively.

C = [fcm(D[:,1], 3), fcm(D[:,2], 3), … , fcm(D[:,27], 3)] (2.3)

Consequently, we attain the Low-medium-high clustering matrix below:

$$C = \begin{bmatrix} \mu_{1,1} & \mu_{1,2} & \cdots & \mu_{1,27} \\ \mu_{2,1} & \mu_{2,2} & \cdots & \mu_{2,27} \\ \mu_{3,1} & \mu_{3,2} & \cdots & \mu_{3,27} \end{bmatrix}$$

We have two options to classify how a data subset belongs to the data set with indicators.

Individual data approach: Each data of the data subset to be compared with the overall dataset is evaluated to find out its index of belonging (membership) to each of the indicators of the overall dataset. Note the concept of fuzzy logic will be employed here and the membership values will be anywhere from 0 to 1, not just 0 or 1. An aggregation scheme will be researched to combine all the membership function values of each data to conclude if the data subset as a whole is classified as closer to one indicator than the other.

After fuzzy C-mean clustering analysis on global data, we attain the high, medium, and low values of global data as indicators for sub dataset; then we calculate the attributability of each datum in the subset; finally we average the attributabilities for low, medium, and high cluster and conclude the ultimate subset membership by finding the largest average attributability. Specifically, individual data approach means the following steps to decide the nutrient memberships for a group of foods (FO_p,

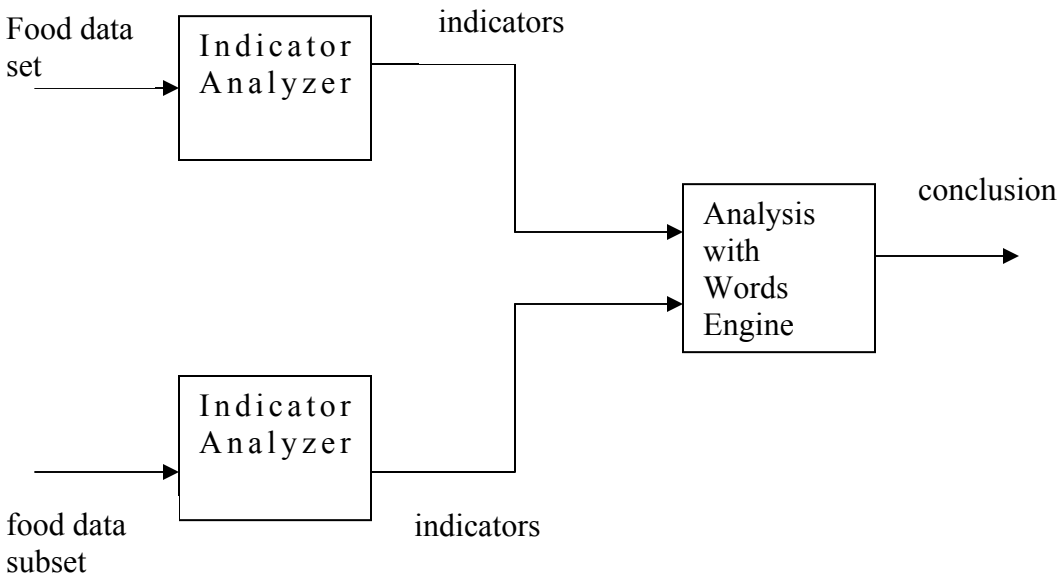

Fig. 2.2: Classifying a data subset by comparing the data subset indicators with the overall indicators

$FO_q \ldots FO_n$ where p, q… n ϵ [1, 206]) that are beneficial to an organ:

1) For each nutrient, calculates the low, medium, and high membership values of every food in the food group. For any $d_{i,j}$ (i ϵ [p, q,…, n] and j ϵ [1, 27]) recording the amount of nutrient N_j in food FO_i the in matrix D, calculate the low, medium, and high membership values ($m_{i,j,1}$, $m_{i,j,2}$, $m_{i,j,3}$) by substituting $\mu_{1,j}$,$\mu_{2,j}$, $\mu_{3,j}$ in matrix C respectively for μ, and $d_{i,j}$ for y in membership function expression, resulting in the following three dimensional all nutrients membership values array:

$V_{i,j,k} = m(D_{i,j}, C_{j,k}, 0.01)$ (2.4)

For a nutrient N_j in 3D Array V, there is a matrix: Single nutrient membership value matrix

$M_j = V[:,j,:]$ (2.5)

$$M_j = \begin{bmatrix} m_{p,1} & m_{p,2} & m_{p,3} \\ m_{q,1} & m_{q,2} & m_{q,3} \\ \cdots & \cdots & \cdots \\ m_{n,1} & m_{n,2} & m_{n,3} \end{bmatrix}$$

2) Calculate the average of each column in matrix M and find out the max average value among the three columns.

$$V = \begin{bmatrix} v_{p,1} & v_{p,2} & \cdots & v_{p,27} \\ v_{q,1} & v_{q,2} & \cdots & v_{q,27} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ v_{n,1} & v_{n,2} & \cdots & v_{n,27} \end{bmatrix}$$

Where $v_{i,j}$ i ϵ [p, q,…, n] and j ϵ [1, 27] is a one dimensional array with length 3 containing the low, medium, and high membership value ($m_{i,j,1}$, $m_{i,j,2}$, $m_{i,j,3}$).

3) Decide the whole food group's representative membership for nutrient N_j.

If the column with the max average is column one, the analyzer concludes that the whole food group is low in providing nutrient N_i; if the column with the max average is column two, the analyzer concludes that the whole food group is medium in providing nutrient N_i; if the column with the max average is column three, the analyzer concludes that the whole food group is high in providing nutrient N_i.

4) Repeat steps 2) and 3) to decide the food group's representative memberships for all 27 nutrients.

Lump-sum approach: The analysis with words engine is applied to both the overall data and the

selected data subset. The problem of where a given data subset stand in view of the overall data can be solved by comparing their indicators, representatives of these data set. This illustrated in Figure 2.2.

After fuzzy C-mean clustering analysis on global data, we attain the high, medium, and low values of global data as indicators; on the other hand, we calculate a representative datum for the whole subset; then we calculate the attributability of that representative datum to the high, medium, low indicators; finally we conclude the subset membership by finding the largest attributability. Specifically, lump-sum approach means the following steps to decide the nutrient memberships for a group of foods (FO_p, FO_q… FO_n where p, q… n ϵ [1, 206]) that are beneficial to an organ:

1) The input data composes a food group nutrition facts sub matrix SD by withdrawing data associated with the food group from matrix D: SD ϵ D

$$SD = \begin{bmatrix} d_{p,1} & d_{p,2} & \cdots & d_{p,27} \\ d_{q,1} & d_{q,2} & \cdots & d_{q,27} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ d_{n,1} & d_{n,2} & \cdots & d_{n,27} \end{bmatrix}$$

(2.6)

2) Calculate the average amount of each nutrient the food group contains by calculating the average of each column in matrix SD.

Food group average nutrition facts
$A = [a_1, a_2, \ldots, a_{27}] = [\text{avg}(SD[:,1]), \text{avg}(SD[:,2]), \ldots, \text{avg}(SD[:,27])]$
(2.7)

3) Evaluate the membership function values by substituting a_j for y, and $\mu_{1,j}$, $\mu_{2,j}$, $\mu_{3,j}$ in matrix C for μ to calculate low, medium, and high membership values respectively for nutrient N_j.

4) Decide the representative group membership for N_j as the one with the largest value among low membership values, medium membership values, and high membership values.

5) Decide the representative group membership for all 27 nutrients.

3. Application to Chinese food cure theory

According to Chinese medicine's Zhang-Fu organ theory, each Zhang organ – heart, liver, spleen, lung, and kidney – is closely related to a Fu organ – small intestine, gallbladder, stomach, large intestine, and urinary bladder [3, 4, 5]. Consequently, the nutrient membership pattern of a Zhang organ's beneficial food group should be similar to the pattern of a corresponding Fu organ's beneficial food group. In other words, two food groups beneficial to two related organs respectively should have the same membership in most of the nutrients. Compare the food nutrient patterns for two organs shows that they are closely related to each other.

4. Conclusion

Analysis with Word enables quick and easy analysis of data in natural language. The conclusive result from this proposed engine allow people to comment on the relationship between two data set in terms of their simple guidelines. Although the number of food and nutrients applied in this paper may no be the most challenging problem, it is nevertheless representative of numerous similar analyses of data in real life application. Future research calls for extension of our proposed Analysis of Word engine to problem with mixed linguistic and numeric characteristic.

5. References

[1] Besdek, J. C., "Some recent applications of Fuzzy C-Means in pattern recognition and image processing," IEEE workshop on Lang. Autom., P247-252, 1983.

[2] Trivedi, M. M., "Analysis of aerial images using fuzzy clustering," Analysis of Fuzzy Information, Vol. III (Bezdek, ed.) pp.133-151, CDC Press.

[3] Gao Duo, and Bernie Barbara. (1997). *Chinese Medicine* p.60-121.

[4] Lu, C. Henry. (1986). *Chinese System of Food Cures* p. 46-136.

[5] Junying, Geng. (1996). *Practical Traditional Chinese Medicine & Pharmacology – Herbal Formulas*