The Application of Fuzzy Clustering in Life Quality Assessment of Lung Cancer Patients

Peiyu Chen
College of Electronic Information and Control Engineering
Beijing University of Technology
Beijing, China
530795037@qq.com

Liying Fang, Pu Wang, Shuang Li
College of Electronic Information and Control Engineering
Beijing University of Technology
Beijing, China
{fangliying, wangpu}@bjut.edu.cn,
690736804@qq.com

***Abstract*—In order to research the relationship between the FACT score and lung cancer patients' tumor progression with the patients' death situation, we manage to identify the variation tendency rules of the FACT score. This paper proposes a new, simple and efficient representation for the one-dimensional numerical time series. The representation uses the improved fuzzy C-means clustering algorithm to cluster all the sequence segments, realizing the symbolization of the numerical data sequence. At last,we find the frequent patterns in the new symbolic FACT score sequence. The relation between the FACT score and the patients' death situation can be found through the experiments.**

Keywords-fuzzy C-means clustering; time series; symbolization; GSP.

I. INTRODUCTION

This part contains a review of the data object, i.e., FACT score, and expounds the research necessity and significance.

A. FACT score

Life quality assessment of lung cancer patients has been widely focused, as the traditional disease evaluation and prognosis are not adapting the patients' needs, the health concepts and nowadays medical mode any more.

FACT is a Functional Assessment of Cancer Therapy which was developed by Cella etc. of the Chicago Rush-Presbyterian-St. Luke Medical Center at United States. It contains a basic module, FACT-G, measuring the life quality of cancer patients and some specific cancer subscale. The FACT-G has 27 indexes, divided into four parts: the somatic condition (7 indexes), the social/family situation (7 indexes), the emotional status (6 indexes) and the functional status (7 indexes). Specific cancer assessment contains the basic module and a specific module [1].

The FACT-L, a specific cancer assessment, that is a self-assessment of the lung cancer patients which containing the FACT-G and the lung cancer specific module (9 indexes). The lung cancer specific module is mainly used to the patients who receive chemotherapy, radiation therapy and have been well-educated. Each index is the grade entry, the forward entry direct count from zero to four points and reverse entries reverse count when computing the score, i.e., four points to fill the first one grade within three points on two levels, etc.

B. Overview

The basic task of data mining is to find frequent patterns from the data set. This kind of problem is to decide which mode is frequent in a class of data set mode possible existence. Frequent pattern is found based on the symbolic sequence that including several basic item sets, however, the large number of the item sets will inevitably reduce the support and typicality of frequent pattern. In this way, finding frequent pattern from the numerical data sequence will be a great challenge.

There are certain kinds of data types in the medical data source, such as the symbol type, the Boolean type, the numeric type, etc. Due to the limited number of basic item sets, the frequent patterns using the frequent pattern algorithm can be directly mined from the first two data types. The number of the numerical data in the basic set is hundreds of thousands that is not conducive to frequent pattern mining, so mining frequent patterns need to transforming dataset from numerical to symbol.

This paper proposed a new representation of time series, and then mined data pattern from FACT score in lung cancer cases for above problems. The experiments show that the representation can re-paint the original sequence in a certain accuracy conditions and convenient for their pattern mining.

In the second part, introducing the whole algorithm and detailed explain the each step of the algorithm. In the third part, the experiment is accomplished based on the algorithm and analysis. At last, the conclusion and future work is presented.

II. PATTERN FIND BASED ON THE FUZZY CLUSTERIN

During the process of finding the pattern, the first step is classifying the FACT score with tumor progression and death situation. Then symbolizes the original sequence to find the pattern. At last verifies the result, the sequential pattern, which is compliant with reservation requirements through the experimental results. If the results do not reach reservation requirements, it needs to re-select the dimensions

 ISBN: 978-1-61208-202-8

and other parameters, until the condition is met. Figure 1 shows the flowchart:

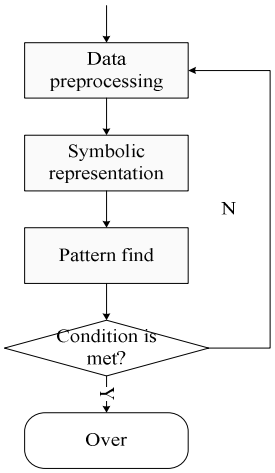

Figure 1. The flowchart of Pattern find

A. Data preprocessing

Each patient has at most 4 treatment periods and 14 follow-up periods, and the collected data content of each treatment and follow-up are the same. Preprocess the patients' FACT score. Take out the ID number of patients and the corresponding FACT score while ignoring the missing point directly from the database which stores the each treatment and follow-up data of patients.

B. Symbolic representation

First handle the patient's data thus we could segment an N-dimensionality into an N-1 one by making every treatment and follow-up data a break-point. The slope of the segment represent the trend of the curve, which has a better result compared with the Euclidean distance, Pearson correlation coefficient etc. Extract the slope of the segment as the standard of clustering, and symbolize the segment that each line in the corresponding class.

In this paper, we proposed an improved fuzzy C-means clustering algorithm used to cluster FACT score. Fuzzy clustering algorithm is based on the best practices function and it use calculus computing technology to get the optimal cost function. The fuzzy clustering algorithm defines the neighbor function between the vector and the cluster, and the membership function set provides the membership degree of the clustering vector. In the fuzzy method, the vector membership degree in the different cluster are interrelated. Simultaneously using the distance of slope of the segment replacing the traditional Euclidean distance could pay more attention to the trend of the FACT score in the clustering algorithm.

The fuzzy C-means clustering algorithm's input parameter is C and then divide N data objects into C clustering. Clustering results to meet high similarity in one cluster, and meet smaller similarity in the different clustering results [2].

Fuzzy C-means clustering algorithm:

Input: the number of clusters C, and the data sets with N data objects.

Output: the matrix of cluster centers and partition matrix.

Algorithm process:

Step 1: Initialize the partition matrix Uik with values in a random number between 0 and 1.

Step 2: Calculate the cluster center.

Step 3: Calculate the value function J. If it is less than a certain threshold or the value change comparing with last value is less than a threshold, the algorithm stop.

Step 4: Recalculate the partition matrix Uik, then return to step 2.

Figure 2 shows the flowchart:

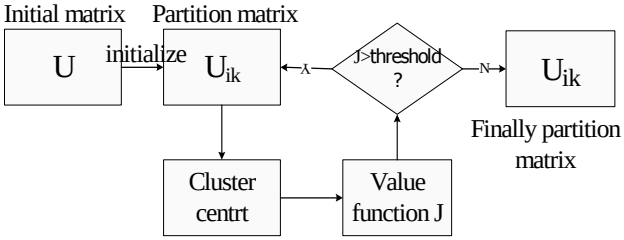

Figure 2. The flowchart of Fuzzy Cluster

The pseudo code of the algorithm:

```
Initialize the number of clusters = 3;
Initialize the partition matrix Uik;
While J> threshold
    For each cluster:
        Calculate the cluster center;
        Calculate the value function J;
Return Uik;
```

In the algorithm, the data object is a sub-sequence after segment; the distance between the segments is measured by the change value of the two points of the segment [3]. The distance between segment l_i and l_j is:

$$D_{ij} = \sqrt{(s_i - s_j)^2} \qquad (1)$$

C. Pattern find algorithm

Frequent pattern find algorithm uses pruning strategies of redundant candidate pattern and special data structure-a hash tree realizing the fast memory access of candidate patterns that using the GSP algorithm, which is similar to Apriori algorithm [4][5].

Algorithm process:

Step 1: Scan sequence database, getting sequential pattern L1 whose length is 1 as the initial seed set.

Step 2: According to the seed set Li, candidate sequential patterns Ci+1 whose length is i+1 by connecting operations and pruning operations; then scan the sequence database, calculating the support of each candidate sequence patterns and generate the sequence pattern Li+1 whose length is i+1 which is used to be a new seed set;

Step 3: Repeat the second step, until no new sequence mode or a new candidate sequence patterns.

 ISBN: 978-1-61208-202-8

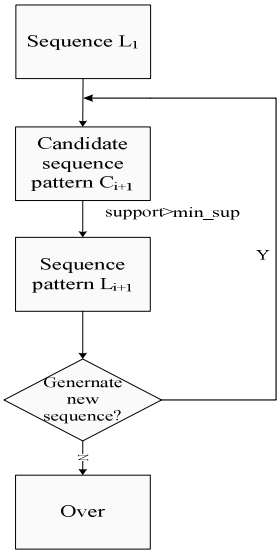

Figure 3. The flowchart of GSP

The pseudo code of the algorithm:

```
Initialize the sequence L1;
While generate new sequence
   Generate the candidate pattern Ci+1
   For each Ci+1
     If the support of Ck>min_sup
           Ck is a FP;
     Get the new sequence pattern Li+1;
   Return Li+1;
```

In above flowchart, the initial sequence is the result of the fuzzy clustering after restored base on the original numeric sequence by the patients.

III. EXPERIMENTAL RESULTS AND ANALYSIS

Experimental data is FACT score which is in the treatment and follow-up cases of non-small cell lung cancer provided by the Beijing Chinese Medicine Hospital. All patients' data were right-align (reversal the sequence), and only intercepting the first 4 data of the sequence. All cases have been deposited in the Oracle 10g database.

There are 4 class patients which classified by tumor progression and death situation: first class is no tumor progression but death; second class is tumor progressed and death; third class is no tumor progression and live; forth class is tumor progressed but live. Every patient have a sequence whose length is between 4 and 17; then cluster the FACT segment by the fuzzy C-means clustering algorithm, while C is 3 that is increment, decrement and constant which are represented by a,b,c, respectively; at last, use GSP algorithm to finding the frequent pattern from the sub-sequence with four data.

We use above algorithm to find the frequent pattern from FACT score. TABLE I shows the result:

TABLE I. THE RESULT OF THE PATTERN FIND

class (patients' number)	Support (percentage)	Frequent sub-sequence
I(22)	15(68.2%)	<a c c>
II(74)	52(70.3%)	<a c c>
III (59)	55(93.2%)	<c c>
IV (75)	66(88%)	<c c>

From the table, it can be seen that a clear upward trend in the last follow-up data in Class I and II patient, and lead to death; however, Classes III and IV patient population in the last several follow-up FACT score value is always maintained a flat trend, and survived.

IV. CONCLUSION AND FUTURE WORK

Transformed the numeric sequence to symbolic sequence by the improved fuzzy C-means clustering algorithm, and find the frequent pattern from the new sequence. The method shows the relationship between the FACT score and the patients' death situation. But there are some disadvantages, such as the support of the frequent sequence <a c c> in class I and II is not enough. Further improvements of the algorithm we expect better results.

ACKNOWLEDGMENT

This paper is supported by 2010 Program for Excellent Talents in Beijing Municipal Organization Department (2010D005015000001), the New Centaury National Hundred, Thousand and Ten Thousand Talent Project, and got the cooperation with Beijing Hospital of Traditional Chinese Medicine Affiliated to CPUMS. Special thanks have been given there.

REFERENCES

[1] C. Wan and C. Zhang, Development and evaluation of the Chinese version of the FACT-L (V1.0) for patients with lung cancer, Quality of Life Newsletter, pp. 19.

[2] H. Chu and B. Chao, Novel Optimization Method for Fuzzy C-Means Algorithms, Journal of Information Engineering University, vol.12, No.3, Jun.2011. (in Chinese)

[3] C. S. Mller-Levet, F. Klawonn, K. H. Cho, and O. Wolkenhauer. Fuzzy Clustering of Short Time-Series and Unevenly Distributed Sampling Points. in Advances in Intelligent Data Analysis V. 5th International Symposium on Intelligent Data Analysis, IDA 2003, 28-30 Aug. 2003. 2003. Berlin, Germany: Springer Verlag.

[4] M. Xia and X. Wang, Research on sequential pattern mining algorithms, Computer Technology and Development, 16th ed., vol. 4. (in Chinese)

[5] L. Liu and J. Cui, Comparison and Analysis between Algorithm of GSP and PrefixSpan, Journal of Liaoning Institute of Technology, vol. 26, No.5, Oct. 2006. (in Chinese)

 ISBN: 978-1-61208-202-8