FUZZY CLUSTER ANALYSIS USING UNSUPERVISED ALGORITHM FOR THE DIAGNOSIS OF TYPES OF DIABETES

Dr.R.Jamuna

**Professor, Department of Computer Science S.R.College,
Bharathidasan university, Trichy.**

Abstract

Technology can be defined as an instrument which allows improved understanding medical data and better management of their health records. Rapidly changing medical technology and changing practice pattern of physicians have revolutionized health care monitoring. Today's medical research could be more advanced, more effective for the society by the application of computer algorithms in large medical data analysis. There is an ever increasing demand for technology based diagnostic predictions to anticipate and prevent complications of major diseases like diabetes, cancer, hypertension, and heart and liver disorders. Clustering is one of the data mining techniques for analyzing such medical datasets. It is a technique for finding similarity groups in data, called clusters. It groups data instances that are similar to each other in one cluster from the data instances that are very different from each other into a different cluster. Clustering is classified as an unsupervised learning task**.** The paper identifies different symptoms in different types of diabetic patients which are the clusters to be grouped. Similarity measure technique isolates the probable disease group for a particular patient. Fuzzy clustering extends this notion to associate each pattern with every cluster using a membership function .The paper finally predicts the most probable type of diabetes pertaining to a patient using the Minkowski metric of the unsupervised algorithm from the cluster of assumed symptom levels over a range. The same technique can be applied to predict any type of disease given the symptoms.

Keywords: Cluster, diabetes, membership function, Minkowski metric, symptoms, data, gestational, matrices.

FUZZY CLUSTER ANALYSIS USING UNSUPERVISED ALGORITHM FOR THE DIAGNOSIS OF TYPES OF DIABETES

Introduction

Clustering is one of the frequently used data mining technique. It is a technique for finding similarity groups in data, called clusters. It groups data instances that are similar to each other in one cluster from the data instances that are very different from each other into a different cluster. Clustering is classified as an unsupervised learning task as no class values are given before grouping them. In our problem under study different symptoms for different diseases of patients are the clusters to be grouped. After applying the similarity measures it should isolate the probable disease group for a particular patient. The symptoms with upper and lower limits of the symptom levels with weight matrix of diseases are given. Cluster analysis groups the objects based on the medical information found in the data describing the objects or their relationships with the prediction of diseases. In this paper we identify the four types of symptom clusters into which the disease data can be grouped. The similarity criterion is distance: two or more objects belong to the same cluster if they are "close" according to a given distance. This is called *distance-based clustering*. Another kind of clustering is *conceptual clustering where* two or more objects belong to the same cluster if it forms a concept suited to all the objects. In other words, objects are grouped according to their fit to descriptive concepts, not according to simple similarity measures. Traditional clustering approaches generate partitions; in a partition, each pattern belongs to a single cluster. Hence, the clusters in a hard clustering are disjoint. Fuzzy clustering [2] extends this notion to associate each pattern with every cluster using a membership function .The output of such algorithms is a clustering, but not a partition matrix, depending on the particular field.

The Data Matrix

Objects are usually represented as vector points in a multi-dimensional space, where each dimension represents a distinct attribute describing the object. For simplicity, it is normally assumed that values are present for all attributes. Here a set of objects is represented as an m by n matrix, where there are m rows, one for each object, and n columns, one for each attribute. This matrix has different names, e.g., diabetes symptom matrix or data matrix.

Components of Clustering Task

Typical pattern clustering activity involves the following steps

(1) Pattern representation for feature extraction of symptom inputs.

(2) Definition of a pattern proximity measure appropriate to the data domain. (upper and lower bounds of symptom levels with weights)

(3) Clustering or grouping by finding Minkowski distance.

(4) Data abstraction for privacy of real time data set of patients.

Mathematical Model for the problem

Models of medical diagnosis that use cluster analysis usually perform a clustering algorithm on the set of patients by examining the similarity of the presence and severity of symptom patterns exhibited by each. The severity of the symptoms present can be designated with degrees of membership in fuzzy sets [3] representing each symptom category. The similarity measure is computed between the symptoms of the diabetic patient in question and the symptoms of a patient possessing the classical symptom pattern for each possible disease. The patient to be diagnosed is then clustered to varying degrees with

the prototypical patients whose diabetic symptoms are most similar. The most probable diagnostic candidates are those disease clusters in which the patient's degree of membership is the highest.

Many different methods of fuzzy clustering exist. One group of common methods use some form of distance measure [6] to determine the similarity between observed attributes (symptoms) and those present in the existing diagnostic clusters. We use a simplified adaptation of the method employed by Esogbue and Elder to illustrate this technique.

Esogbue and Elder Technique for classifying types of diabetes [1]

Let us assume that we are given a patient x who displays the symptoms s_i, s_2, s_3, and s_4 at the levels of severity given by the following sample of diabetes symptom vector:

$$X = \begin{matrix} s_1 & s_2 & s_3 & s_4 \end{matrix} \\ \begin{bmatrix} .1 & 0 & 0 & .5 \end{bmatrix}$$

Let $\mu_x(s_i) \in [0, 1]$ denote the grade of membership in the fuzzy set characterizing the diabetic patient X and defined on the set $S = \{s_1, s_2, s_3, s_4\}$, which indicates the severity level of the symptom s_i for the patient. We must determine a diagnosis for this patient among three possible types of diabetes categories Table[1.1] $d_{1(\text{type-1 diabetes})}$, $d_{2(\text{type-2 diabetes})}$, and $d_{3(\text{gestational diabetes})}$. Each of these disorder levels described by a matrix giving the upper and lower bounds of the normal range of severity of each of the four symptoms that can be expected in a patient with the disease. In fact diabetes is a pancreatic disorder than a disease. The types of diseases d_1, d_2, and d_3 are described by the matrices

$$d_1 = \begin{array}{c} \\ \text{Lower} \\ \text{Upper} \end{array} \begin{array}{c} \begin{array}{cccc} s_1 & s_2 & s_3 & s_4 \end{array} \\ \begin{pmatrix} 0 & .6 & .5 & 0 \\ .2 & 1 & .7 & 0 \end{pmatrix} \end{array}$$

$$d_2 = \begin{array}{c} \\ \text{Lower} \\ \text{Upper} \end{array} \begin{array}{c} \begin{array}{cccc} s_1 & s_2 & s_3 & s_4 \end{array} \\ \begin{pmatrix} 0 & .8 & .3 & .2 \\ .3 & 1 & 1 & .2 \end{pmatrix} \end{array}$$

$$d_3 = \begin{array}{c} \\ \text{Lower} \\ \text{Upper} \end{array} \begin{array}{c} \begin{array}{cccc} s_1 & s_2 & s_3 & s_4 \end{array} \\ \begin{pmatrix} .2 & .2 & 0 & .1 \\ .5 & .3 & .4 & .4 \end{pmatrix} \end{array}$$

Let $\mu_{jl}(s_i) \in [0, 1]$ denote the lower bound of the symptom i for disease j, and let $\mu_{ju}(s_i) \in [0,1]$ denote the upper bound of the fuzzy symptom for disease j. We further define a fuzzy relation W on the set of symptoms and diseases that specifies the significance of symptom s in the diagnosis of diabetic disorder d.

Types of diabetic disorders taken for the problem

Disease-1:**Type- 1 diabetes [4]**

Autoimmune destruction of insulin producing pancreatic beta islet cells .It could be controlled by insulin , diet control with exercises.

Disease-2: Type -2 diabetes

Insulin resistant condition with inadequate insulin secretion controlled by medicines and diet with exercises.

Disease-3: Gestational diabetes

Disorders in insulin secretion and glucose metabolism during the period of pregnancy. It could also be kept under controlled by insulin or medicines.[7]

Table [1.1]: Classic Symptoms in Diabetes Mellitus

Classic symptoms:-	
S_1	Poly urination
s_2	Sudden weight loss
s_3	False hunger and thirst.
d_1:	Type-1 diabetes
d_2:	Type-2 diabetes
d_3:	Gestational diabetes.

The relation W of these weights of relevance is given by

Table [1.2] Symptom level versus types of diabetes weight matrix.

$$W = \begin{array}{c} \begin{array}{ccc} d_1 & d_2 & d_3 \end{array} \\ \begin{pmatrix} .5 & .8 & 1 \\ .5 & .5 & .3 \\ .6 & .1 & .9 \\ .6 & .2 & .9 \end{pmatrix} \end{array}$$

Let μ_w (s_i,d_j) denote the weight of symptom s_i for disease d_j. In order to diagnose for the patient x, we use a clustering technique to determine to which diagnostic cluster the patient's characteristics is most similar. This clustering is performed by computing a similarity measure between the patient's symptoms and those typical of each category d_j.

To compute this similarity, we use a distance measure based on the **Minkowski distance** [5] that is appropriately given by the formula

$$Dp(d_j, x) = \left[\sum_{i \varepsilon A_j} \left| \mu\ w(s_i, d_j) (\mu\ d_{jl}(s_i) - \mu_x(s_i)) \right|^p + \sum_{i \varepsilon B_j} \left| \mu w(s_i, d_j) (\mu d_{ju}(s_i) - \mu_x(s_i)) \right|^p \right]^{1/p}$$

here

$A_j = \{i | \mu_x (s_i) < \mu\ d_{jl} (s_i), 1 \leq i \leq m\}$

$B_j = \{i | \mu_x (s_i) > \mu\ d_{ju} (s_i), 1 \leq i \leq m\}$

Where m equals the total number of symptoms.

Results and Conclusions

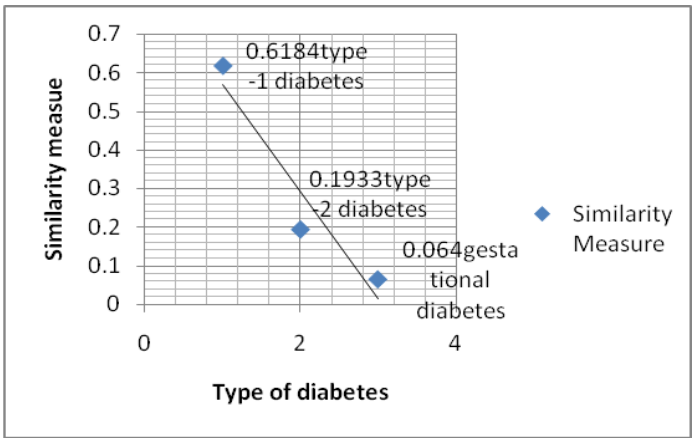
For this example, we give a value of 2 to the parameter p of the distance measure, thus creating a modified Euclidean metric. We use the above equation, with p = 2 to calculate the similarity between patient x and diseases $d_{1(\text{type-1 diabetes})}$, $d_{2(\text{type-2 diabetes})}$, and $d_{3(\text{other-type diabetes})}$ as follows:

$D_1 = (d_1, x) = [\ |(.5)(.6 - 0)|^2 + |(0.6)(.5-0)|^2 + |(0.9)(0-0.5)|^2]^{1/2} = .6184$

$D_2 = (d_2, x) = [\ |(.5)(.8 - 0)|^2 + |(0.1)(.3-0)|^2 + |(0.6(.2-.5)|^2]^{1/2} = .1933$

$D_3 = (d_3, x) = [\ |(.1)(.2 - .1)|^2 + |(0.3(0.2-0)|^2 + |(0.2(.4-.5)|^2]^{1/2} = .0640$

Fig [1] Similarity measure graph for types of diabetes



Conclusion

The most likely a candidate with disease is the one for which the similarity measure attains the minimum value. Fig [1] shows the same**.** In this case, the patient's symptoms are most similar to those typical of disease $d_{3(\text{type-3 gestational diabetes etc.})}$.The findings can be extended to and any number of symptoms which can be stored in a database and can be used in the diagnosis of any type of disease from the clusters of symptoms**.** Technology has brought immense benefit to both the physician and patient alike. When one has symptoms, they can simply type them into a medical expert system on internet and get multidimensional view of the disease and get total awareness about the same. Not only that, but prescribed drugs can even be researched by both physicians and patients alike. Technology brings with it its own clutter that must be sorted through, and sorting through it can become more complicated for the physician. Despite the fact that technology may streamline healthcare and diagnosis easier, it can be a bottleneck as records of medical databases can still be trapped by hackers. So with a very careful security measure such clustering techniques can serve as a best disease diagnostic system.

References

[1] Esogbue, A.O. and Elder, R.C., "Fuzzy sets and the modeling of physician decision processes: part I: the initial interview - information gathering process", Fuzzy sets and systems, 1979, 2, pp. 279-291.

[2] Jumarie, G., "A Minkowskian theory of observation: application to uncertainty and fuzziness", Fuzzy sets and systems, 24, pp. 231-254.

[3] Trillas, E., Alsina, C. and Valverde, L., "Do we need max, min I-J in fuzzy set theory?", 1982, pp. 275-297.

[4] http://www.diabetes.org. Home page for American Diabetes Association.

[5] Jiawei Han, Micheline Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, 2001, p.296 to 299.

[6] Trillas, E., Alsina, C. and Valverde, L. [1982] "Do we need max, min I-J in fuzzy set theory?" In: Yager [1982a], pp 275-297.

[7] Data base source: http:///www.niddk.nih.gov/.From National Institute of digestive and Kidney diseases.