IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 1, Ver. I (Jan. 2014), PP 37-41
www.iosrjournals.org

Unsupervised Clustering Classify theCancer Data withtheHelp of FCM Algorithm

Mr.S.P.shukla
Assist.Professor In BIT,Durg
and
Mrs. RituDwivedi
Research Scholer In Dr.C.V.Raman University KOTA Bilaspur(C.G)

Abstract:*There is structure in nature; also it is believed that there is an underlying structure in most of phenomena, to be understood. in image recognition ,mole biology applications such as protein folding and 3D molecular structure, cancer detection many others the underlying structure exists .by finding structure one classifies the data according to similar patterns, features and other characteristics .this general idea is known as classification. In classification, also termed clustering, the most important issue is deciding what criteria to classify against.This paper presents the fuzzy classification techniques to classify the data of cancer disease. Cluster analysis, cluster validity and fuzzy C-mean (FCM) technique are proposed to be discussed and applied in cancer data classification.*
Keywords:…*classification,clustering,cancerdata.*

I. Introduction

The most important issue in classification and clustering of cancer data is deciding what criteria to classify against .for example suppose it is desirous to classify cancer disease .in describing cancer one will look at its type, spot stage and duration and so on. Many of these features are fuzzy and qualitative in nature. For this classification some criterion is to be decided. One can classify cancer on the basis of its types, its human parts of manifestation i.e. mouth, thought ,tongue intestine, image, liver or similar other parts of the body .its fatal status duration of manifestation and life risk may also be classified. For this classification one might only need two features describing cancer i.e. spot and stage means at different part of body and stages like first, second and third stage .a criterion for classification must be prepared before one can segregate the data into definable classes .as often the case in classification studies ,the number and kind of feature and the type of classification criteria are choices that are continually changed as the data are manipulated ,and this iteration continues until one think that there have a grouping of the data that seems plausible from a structural and physical perspective.
The popular method of classification is very well-known as fuzzy –means (FCM) ,so named because of its close analog in the crisp word, hard C-means (Bezdek 1981).this method uses concepts in n-dimensional Euclidean space to determine the geometric closeness of data points by assigning them various clustering or classes and then determining the distance between the clusters.

1.1 Cluster analysis

Clustering refers to identifying the number of subclasses of C-clusters in a data universe X comprised of n data samples and partitioning X into c clusters (2<=c<=n) .note that c=1 denotes rejection of the hypothesis that there are clusters M the data, where c=n constitutes the trivial case where each sample is in a "cluster" by itself. Two important issues to consider in cluster analysis are how to measure the similarity between pairs of observation and how to evaluate the partition ones they are formed.

One of the simplest similarity measures is distance between pairs of feature vectors in the feature space. If one can determine a suitable distance measure and compute the distance between all pairs of observations, then one may expect that the distance between points in the same cluster will be considerable less than the distance between points in different clusters. the clustering method defines "optimum" partitions through a global criterion function that measures the extent to which candidate partitions optimize a weighted sum of squared errors between data points and cluster centers in feature space. it is emphasized here that the method of clustering must be closely matched with the particular data under study (cancer data) for successful interpretation of sub structure in the data.

1.2 Cluster Validity

In many cases, the number c of clusters in the data is known. In other cases, however if may be reasonable to except cluster substructure at more than one value of c. in this situation if is necessary identify the

value of c that gives the most plausible number of clusters in the data for the analysis at hand. This problem is a unique and absolute measure of cluster validity i.e. the c that is given.
For unlabeled data, no absolute measure of clustering validity exists. Although the importance of these difference is not known, if is clear that the features nominated should be sensitive to the phenomena of applications at hand.
The rest of the paper is organised as follows: The 2 section outlines the reason for using ear as biometric for newborn. This section is followed by details of database acquisition in section 3. Covariates of newborn ear is explained in section 4 followed by automated ear masking in section 5. The details of feature extraction and matching are explained in section 6 and this section also explains proposed methodology for ear recognition. Section 7 describes performance evaluation of different algorithms on newborn ear. Finally section 8 and 9 present future direction and key conclusion.

II. C-Means Clustering

Bezdek(1981) developed an extremely powerful classification method to accommodate fuzzy data. it is an extensive of a method known as C-means, or hard c-means ,when employed in a crisp classification sense .to introduce this method, sample of set of n data samples are defined as:
$X = \{x_1, x_2, x_3 \ \dots x_n\}(1)$
Each data sample x_i, is defined by m features,
$X_i = \{x_{i1}, x_{i2}, x_{i3} \ \dots x_{im}\}(2)$
Where each x_i in the universe X is an *m-dimensional* vector of *m* elements or m features. Since the m features all can have different units, in general, each of to do two things simultaneously, first minimize the Euclidean distance between each data point in a cluster and its cluster center, and second, maximize the Euclidean distance between cluster centers.

III. Fuzzy C-Means (FCM)

It is surprising that very little research on new born biometric identification is published, while that of adults receives much funding for research [25]. One of the main reasons of limited research for newborn identification is the non availability of reference database in public domain.

to developed these methods in classification ,one can define a family of fuzzy sets$\{Ai, i = 1,2,3 \dots \dots \dots \dots . . c\}$ as a fuzzy C-partition on a universe of data points *X*. one can assign membership function to the various data points in each fuzzy set (fuzzy class, fuzzy cluster) .hence a single point can have partial membership in more than one class. The membership function of k^{th} data point in the ith class is given by:
$\mu_{ik} = \mu_{Ai}(X_k) \epsilon [0, 1](3)$
With the restriction that the sum of all membership values for a single data point in all of the classes has be unity:
$\mu_{ik} = 1 \ for \ all \ k = 1,2,3 \dots n(4)$
Also there can be no empty classes the there can be no classes that contain all the date of points, as given by the expression:
$0 < \sum_{k=1}^{n} \mu_{ik} < n(5)$
Because each date of point can have partial membership in more than one class, the restriction for fuzzy classification is:
$\mu_{ik} \Lambda \mu_{jk} \neq 0(6)$
The following provision also holds for the fuzzy case,
$\bigvee_{i=1}^{c} \mu_{Ai}(X_k) = 1 \ for \ all \ k(7)$
$0 < \sum_{k=1}^{n} \mu_{Ai}(X_k) < n \ for \ all \ I(8)$
Before in the case of c=2, the classification problem reduced to that of the excluded middle lows for two classes $A_i$and $A_j$as below:
$Ai \cap Aj \neq \emptyset(9)$
$\emptyset \subset A_i \subset X(10)$
A family of fuzzy partition matrices, M_{fc}, for the classification involving *c* classes and *n* date of points can be defined as:
$M_{fc} = \{Ul \ \mu_{ik} \epsilon [0, 1], \sum_{i=1}^{c} \mu_{ik} = 1 , 0 < \sum_{i=1}^{n} \mu_{ik} < n\} (11)$
Where $i= 1, 2 \dots c \ and \ k = 1, 2 \dots n$
Any $U\epsilon$, M_{fc} is a fuzzy C-partition, and it follows from overlapping character of the classes and the infinite number of membership values possible for the describing class membership that the cardinality of $M_{fc}$is also infinite, i.e. η_{mfc} = *infinite*.

3.1 FCM algorithm

To describe the method to determine the fuzzy C- partition matrix U for grouping the collection of n data sets in to c classes, an objective functions J_m for a fuzzy C- partition is defined as:

$J_m(U,v) = \sum_{m=2}^{N}\sum_{m=2}^{c}(\mu_{ik})\ (d_{ik})$(12)

Where$d_{ik} = d\ (X_k - V_i) = [\sum(X_{ki} - V_{ii})^2]^{1/2}$(13)

and where $\mu_{ik}$is the membership of the kith data point in the i^{th}class.The distance measure d_{ik}in eqn 13 in again a Euclidean distance between the i^{th}cluster center and k^{th} data set. (Data point in m-space). *m*' is a weighting parameter *M eqn /2* and its value has a range $m'\epsilon\ [1, infinite]$. v_i is i^{th} cluster center and can be arranged as $v_i = \{v_{i1}, v_{i2}, \ldots v_{im}\}$.

Each of the cluster co-ordinates for each class can be calculated in a manner as given below:

$v_{ij} = \sum_{K=1}^{n} \mu_{ik} m' X_{kj} \Big/ \sum_{K=1}^{n} \mu_{ik} X_{kj} m'$(14)

Where j is a variable on the feature space i.e. *j=1, 2, 3…m*

The optimum fuzzy C-partition be the smallest of the partitions described as:

$J_m\ (U,V) = \min_{M_{fc}} J\ (U,v)$(15)

3.1.1 Algorithm

A recent effective algorithm for fuzzy classification, called iterative optimization, was proposed by Bezdek (1981), the step in this algorithm are as bellows:

1. Fix $c\ (2 <= c < n)$ and select a value for parameter *m'*. Initialize the partition matrix$U^{(0)}$.each step in this algorithm will be labeled r, where *r=0, 1, 2……*.
2. Calculate the c centers $\{v_i^{(r)}\}$ fir each step.
3. Update the partition matrix for the $rth\ step, U(r)$as follows:

$\mu_{ik}^{(r+1)} = [\sum(d_{ik}(r)\ /d_{ik}(r)^{2/(m'-1)}] - 1^{-1}\ \ for\ I_k = \emptyset$(16)

$\mu_{ik}(r+1) = 0\ for\ all\ classes\ i\ where\ i\ \epsilon\ I_k$(17)

Where $I_k = \{i/2 <= c < n, d_{ik}(r) = 0\}$(18)

And $I_k = \{1, 2 \ldots c\} - I_k$(19)

And $\sum \mu_{ik}(r+1) = 1$(20)

4. If $\|U^{(r+1)} - U(r)\| <= \epsilon L$, stop, other set $r = r + 1$ and return to step 2.

3.1.2 Example for cancer data classification

In cancer data classification many frustration process of cancer has a relation between penetration speed of medicine and medicine efficiency. Two classes of date are known from the treatment efficiency. Points of high medicine efficiency and high penetration speed are indicator of improving patient (Class1) and points of low medicine efficiency and low penetration speed are indicative of collapsing patient (Class2). Suppose one measures the medicine efficiency and penetration speed of medicine four patients and attempt to characterize them as improving and collapsing. The four data points (n = 4) are shown , where Y axis is medicine efficiency and X axis is the penetration speed of the medicine. The data are described by two features (*m=2*), and have following coordinate in 2 D space.

$$X1 = \{1, 3\}$$
$$X2 = \{1.5, 3.2\}$$
$$X3 = \{1.3, 2.8\}$$
$$X4 = \{3, 1\}$$

It is desired to classify these data points into two classes(*c=2*).

The fuzzy classification method generally converges quite rapidly, even when the initial guess for the fuzzy partition is quite poor, in classification sense. The fuzzy iterative optimization method for this case would process as follows:

Using U^*as the initial fuzzy partition $U^{(0)}$,and assuming a weighting factor of *m'=2* and a criterion for convergence of $\epsilon\ L = 0.01$,i.e.

$$\max_{l,k} |\mu_{ik}\ (r+1) - \mu_{ik}\ (r)| <= 0.01$$

Optimum fuzzy 2-partition U^* is be determined.

$$U^{(0)} = \begin{bmatrix} 1 & 1 & 10 \\ 0 & 0 & 01 \end{bmatrix}$$

next is the calculation of the initial cluster center using eqn 14 where $m' = 2$

$$V_{ij} = \sum_{K=1}^{N}(\mu_{ik})^2 X_{kj} \Big/ \sum_{K=1}^{N}(\mu_{ik})^2$$

Where for $c = 1$

$$V_{ij} = \frac{\mu_1^2 X_{1j} + \mu_2^2 X_{2j} + \mu_3^2 X_{3j} + \mu_4^2 X_{4j}}{\mu_1^2 + \mu_2^2 + \mu_3^2 + \mu_4^2}$$

$$V_{ij} = \frac{(1)^2 X_{1j} + (1)^2 X_{2j} + (1)^2 X_{3j} + (1)^2 X_{4j}}{(1)^2 + (1)^2 + (1)^2 + (1)^2}$$

$$V_{ij} = \frac{X_{1j} + X_{2j} + X_{3j} + X_{4j}}{3}$$

$$\left.\begin{aligned} V_{11} &= \frac{1 + 1.5 + 1.3}{3} = 1.26 \\ V_{12} &= \frac{3 + 3.2 + 2.8}{3} = 3.0 \end{aligned}\right\}$$

$$V_1 = \{1.26, 3.0\}$$

For $c = 2$

$$V_{2i} or\ V_i = \frac{X_{4j}}{0 + 0 + 0 + 1} = X_{4j}$$

$$\left.\begin{aligned} V_{21} &= {}^{3}/_{1} = 3 \\ V_{22} &= {}^{1}/_{1} = 1 \end{aligned}\right\}$$

$$V_2 = \{3, 1\}$$

Now the distance measures (distance of each data point from cluster center) are found using eq[n]13

$d_{11} = \sqrt{(1 - 1.26)^2 + (3 - 3)^2} = 0.26$
$d_{12} = \sqrt{(1.5 - 1.26)^2 + (3.2 - 3)^2} = 0.31$
$d_{13} = \sqrt{(1.3 - 1.26)^2 + (2.8 - 3)^2} = 0.20$
$d_{14} = \sqrt{(3 - 1.26)^2 + (1 - 3)^2} = 2.65$
$d_{21} = \sqrt{(1 - 3)^2 + (3 - 1)^2} = 2.82$
$d_{22} = \sqrt{(1.5 - 3)^2 + (3.2 - 1)^2} = 2.66$
$d_{23} = \sqrt{(1.3 - 3)^2 + (2.8 - 1)^2} = 2.47$
$d_{24} = \sqrt{(3 - 3)^2 + (1 - 1)^2} = 0.00$

With the distance measures, update can be found using eq[n] 18 to 20.

$$\mu_{ik}^{(r+1)} = [\sum_{n=1}^{c} i = 1 \, (d_{ik}(r) \, / d_{ik}(r)^2]^{-1}$$

And it can be found

$$\mu_{11} = \left[\sum_{n=1}^{c} i = 1 \left(\frac{d_{11}}{d_{i1}}\right)^2\right]^{-1} = \left[\left(\frac{d_{11}}{d_{11}}\right)^2 + \left(\frac{d_{11}}{d_{21}}\right)^2\right]^{-1}$$

$$\mu_{11} = \left[\left({}^{0.26}/_{0.26}\right)^2 + \left({}^{0.26}/_{0.28}\right)^2\right]^{-1} = 0.991$$

$$\mu_{12} = \left[\left({}^{d_{12}}/_{d_{12}}\right)^2 + \left({}^{d_{12}}/_{d_{22}}\right)^2\right]^{-1}$$

$$\mu_{12} = \left[1 + \left({}^{0.31}/_{2.66}\right)^2\right]^{-1} = 0.986$$

$$\mu_{13} = \left[\left({}^{d_{13}}/_{d_{13}}\right)^2 + \left({}^{d_{13}}/_{d_{23}}\right)^2\right]^{-1}$$

$$\mu_{13} = \left[1 + \left({}^{0.2}/_{2.47}\right)^2\right]^{-1} = 0.993$$

$$\mu_{14} = \left[\left({}^{d_{14}}/_{d_{14}}\right)^2 + \left({}^{d_{14}}/_{d_{24}}\right)^2\right]^{-1}$$

$$\mu_{14} = \left[1 + \left({}^{2.65}/_{0}\right)^2\right]^{-1} > 0.00 \; for \; l_4 \neq \emptyset$$

Using Eq[n] 4 for the other partition values, μ_{2j}, for *J=1, 2, 3, 4*, the new membership functions from an updated fuzzy partition given by:

$$U^{(1)} = \begin{bmatrix} 0.991 & 0.986 & 0.9930 \\ 0.009 & 0.014 & 0.0071 \end{bmatrix}$$

To determine whether convergence has been achieved, a matrix hormוווו is chosen such as the maximum absolute value of pair wise comparison of each of the values in $U^{(0)}$ and $U^{(1)}$ i.e.

$$\max_{l,k} l\, \mu_{ik}^{(1)} - \mu_{ik}^{(0)} = 0.0134 > 0.01$$

This result suggests that convergence criteria have not yet been satisfied, so one needs another iteration of the method. The cluster center are again calculated from the latest fuzzy partition $U^{(1)}$for $C = 1$,

$$V_{ij} = \frac{(0.991)^2 X_{1j} + (0.986)^2 X_{2j} + (0.993)^2 X_{3j} + (0)^2 X_{4j}}{(0.991)^2 + (0.986)^2 + (0.993)^2 + (0)^2}$$

$$\left.\begin{aligned} V_{11} &= \frac{0.98\,(1) + 0.97(1.5) + 0.99\,(1.3)}{2.94} = \frac{3.719}{2.94} = 1.26 \\ V_{12} &= \frac{0.98\,(3) + 0.97(1.5) + 0.99\,(0.28)}{2.94} = \frac{8.816}{2.94} = 3.0 \end{aligned}\right\}$$

$$V_1 = \{1.26, 3.0\}$$

For $c = 2$

$$V_{ij} = \frac{(0.009)^2 X_{1j} + (0.014)^2 X_{2j} + (0.007)^2 X_{3j} + (1)^2 X_{4j}}{(0.009)^2 + (0.014)^2 + (0.007)^2 + (1)^2}$$

$$V_{21} = \frac{(0.009)^2(1) + (0.014)^2(1.5) + (0.007)^2(1.3) + (1)^2(3)}{(0.009)^2 + (0.014)^2 + (0.007)^2 + (1)^2}$$

$$V_{22} = \frac{(0.009)^2(3) + (0.014)^2(1.5) + (0.007)^2(2.8) + (1)^2(1)}{(0.009)^2 + (0.014)^2 + (0.007)^2 + (1)^2}$$

$$\left.\begin{aligned} V_{21} &= \frac{3.0004387}{1.0004387} = 3.00 \\ V_{22} &= \frac{1.0006742}{1.0004387} = 1.00 \end{aligned}\right\}$$

The final partition $U^{(2)}$will result in a classification.

IV. Conclusions

Previously, the partition for cancer patient shown in fig1 explains that one patient is classified (c_1) as an improving patient and three patients are classified (c_2) in collapsing patient. After a fuzzy C- means classifications it can be seen that three patients have been classified (c_1) as an improving patients as one patient is classified (c_2) as a collapsing patient.

References

[1] George j.klir / boyuan "fuzzy set and fuzzy logic" theory and application, year 2003, pages 50-61
[2] Gurney, Kevin: An Introduction to Neural Networks. UCL Press, London, UK 1999.
[3] Fausett, Laurene: Fundamentals of Neural Networks: Architectures, algorithms, and Applications. Prentice Hall, NJ, USA 1994.
[4] Hertz, J.A., Krogh, A. & Palmer, R. Introduction to the Theory of Neural Computation(Addison-Wesley, Redwood City, 1991)
[5] Tanaka, Makoto: "Application of The Neural Network and Fuzzy Logic to The Rotating Machine Diagnosis" Fusion of Neural Networks, Fuzzy Sets, and Genetic Algorithms: Industrial Applications. CRC Press LLC, CRC Press LLC, Boca Raton, FL, USA 1999.
[6] Lee, S. and E. Lee: "Fuzzy Sets and Neural Networks" Journal of Cybernetics. Volume 4, No. 2, pp. 83-013, 1974.
[7] Zadeh, Lotfi: "The Role of Soft Computing and Fuzzy Logic in the Conception, Design, Development of Intelligent Systems" Plenary Speaker, Proceedings of the International Workshop on soft Computing Industry. Muroran, Japan, 1996.
[8] [172] Zadeh, Lotfi: "What is Soft Computing" Soft Computing. Springer-Verlag Germany/USA 1997.
[9] Kacpzyk, Janusz (Editor): Advances in Soft Computing. Springer-Verlag, Heidelberg, Germany, 2001.