2003 International Symposium on Advanced Intelligent Systems
September 25-28, 2003, Jeju, Korea

Fuzzy *c*-Logistic Regression Model in the Presence of Noise Cluster

Arnold C. Alanzado
Graduate School of Systems and Information Engineering
University of Tsukuba, Ibaraki 305-8573, Japan
arnold@odin.esys.tsukuba.ac.jp

Sadaaki Miyamoto
Institute of Engineering Mechanics and Systems
University of Tsukuba, Ibaraki 305-8573, Japan
miyamoto@esys.tsukuba.ac.jp

Abstract – **In this paper we introduce a modified objective function for fuzzy *c*-means clustering with logistic regression model in the presence of noise cluster. The logistic regression model is commonly used to describe the effect of one or several explanatory variables on a binary response variable. In real application there is very often no sharp boundary between clusters so that fuzzy clustering is often better suited for the data.**

1 Introduction

The increasing availability and accessibility of data require advanced tools to extract information from these data. Automated data acquisition and the rise of the Internet are contributing to a strong increase in the amount of electronically available data. In today's world of data analysis, one of the significant problems is the presence of noise within data sets. The concept of "noise" plays a crucial role in the statistical analysis of data.

A strange value that stands out because it is not like the rest of the data in some sense is commonly called noisy point or outlier. Outliers are vectors, or called data point, in the data domain which are so distant from the rest of the other vectors in the data set [1]. The handling of outlying or anomalous observations in a data set is one of the most important tasks in data analysis, because outlying observations can have a considerable influence on the cluster results.

In real applications there is very often no sharp boundary between clusters so that fuzzy clustering is often better suited for the data. In fuzzy clustering each object belongs to multiple clusters to a different degree. Fuzzy clustering give belongedness to groups at each point of data set by a membership function [2]. Its advantage can adapt to noisy data and classes that are not well separated. The logistic regression model is commonly used to describe the effect of one or several explanatory variables on a dichotomous response variable. It has been used successfully in many areas of application that use statistical modeling

In this paper we introduce a modified objective function for fuzzy *c*-means clustering with logistic regression model in the presence of noise cluster. The aim of our approach is to develop a model utilizing both the concept of fuzzy *c*-means clustering and logistic regression. The modified fuzzy *c*-means objective function is an extension of the standard fuzzy *c*-means with the entropy term added to maximize the entropy of the centroids with respect to the data points. Additional variable α is added for controlling the cluster sizes. Noise term is added to handle the noise cluster.

The noise clustering approach is based on the concept of first defining a noise cluster and then defining a similarity or dissimilarity measure for the noise cluster. The noise is considered to be a separate class represented by an extra centroid that has a constant distance δ from all feature vectors. Most fuzzy cluster analysis algorithms are based on either Euclidean distances or density. However, most of today's data often consists of a lot of features which form a high dimension space. We introduce a new similarity measure based on the logistic regression for our modified objective function and applied it to a medical data. The error sum of squares contains the logit term.

2 Fuzzy c-means clustering

Fuzzy clustering allows an object to belong to multiple clusters. Each object is bound to each cluster to a certain degree, $u \in [0,1]$, also known as membership [3]. The basic idea of fuzzy *c*-means is very similar to k-means algorithm. It assumes the number of cluster *c*, is

known a priori, and tries to minimize the objective function.

Let $x_k=(x_k^1,\cdots,x_k^p)$, $k=1,\cdots,n$, be individuals to be clustered. They are points in p-dimensional Euclidean space. Notice that we use row vectors and hence x_k^t transpose of x_k, is a column vector.

We consider two types of objective functions:

$$J_1(U,V)=\sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m D_{ik}$$

$$J_2(U,V)=\sum_{i=1}^{c}\sum_{k=1}^{n}u_{ik}D_{ik}+\lambda^{-1}\sum_{i=1}^{c}\sum_{k=1}^{n}u_{ik}\log u_{ik}$$

where u_{ik} is an element of cluster membership matrix $U = (u_{ik})$. The constraint of the fuzzy partition

$$M=\{U:\sum_{j=1}^{c}u_{jk}=1,\ 0\le u_{ik}\le 1,\ i=1,\cdots,c,\ k=1,\cdots,n\}$$

is assumed as usual.

The term D_{ik} is assumed to be the square of the Euclidean distance between the individual x_k and the center $v_i = (v_i^1,\cdots,v_i^p)$ of the cluster i:

$$D_{ik}=\|x_k-v_i\|^2,$$

unless otherwise assumed. V = $(v_1,\cdots,v_c)$: the vector collecting all cluster centers.

The former J_1 is well-known [3,4], while the latter J_2 is for the method of entropy proposed by the authors [5] (see also [6]).

We next note that the following alternative optimization algorithm is used for finding optimal U and V, in which either $J = J_1$ or $J = J_2$.

Algorithm of fuzzy c-means

FCM1 Set initial value for $\bar{V}$.

FCM2 Find optimal solution $\bar{U}$:

$$J(\bar{U},\bar{V})=\min_{U\in M}J(U,\bar{V})$$

FCM3 Find optimal solution $\bar{V}$:

$$J(\bar{U},\bar{V})=\min_{V\in R^{pc}}J(\bar{U},V)$$

FCM4 Check stopping criterion and if convergent, stop. Otherwise go to FCM2.

Assume $J = J_1$ (the method of the standard fuzzy c-means is used). It is then easy to see that the solutions in FCM2 and FCM3 are

$$\bar{u}_{ik}=\left[\sum_{j=1}^{c}\left(\frac{D_{ik}}{D_{jk}}\right)^{\frac{1}{m-1}}\right]^{-1};$$

$$\bar{v}_i=\frac{\sum_{k=1}^{n}(u_{ik})^m x_k}{\sum_{k=1}^{n}(u_{ik})^m}.$$

We omit the over bars, e.g., we write u_{ik} instead of $\bar{u}_{ik}$ in the case of no ambiguity.

If we use $J = J_2$, we have

$$u_{ik}=\frac{e^{-\lambda D_{ik}}}{\sum_{j=1}^{c}e^{-\lambda D_{jk}}},$$

$$v_i=\frac{\sum_{k=1}^{n}u_{ik}x_k}{\sum_{k=1}^{n}u_{ik}}$$

in FCM2 and FCM3, respectively.

3 Noise clustering

The idea of noise clustering approach was proposed by Davé [7] to deal with noisy data for fuzzy clustering methods. A noise cluster, cluster number 0 with membership $u_{0k}, k=1,\cdots,n$, is introduced, with the hope that all noisy points can be dumped into this cluster. The objective function is given by

$$J_3(U,V)=\sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m D_{ik}+\sum_{k=1}^{n}(u_{0k})^m\delta^2$$

in which the constraint is

$$M=\{U:\sum_{j=0}^{c}u_{jk}=1,\ 0\le u_{ik}\le 1,\ i=0,\cdots,c,\ k=1,\cdots,n\}$$

Ichihashi *et al.* [8] propose a generalized objective function using K-L information with additional variables. A variation of which has been proposed by Miyamoto and Alanzado [9] to handle noise clusters.

$$J_4(U,V,\alpha)=\sum_{i=1}^{c}\sum_{k=1}^{n}u_{ik}D_{ik}+\sum_{k=1}^{n}u_{0k}\delta^2+\lambda^{-1}\sum_{i=0}^{c}\sum_{k=1}^{n}u_{ik}\log\frac{u_{ik}}{\alpha_i}$$

with the constraint (M) and

$$A=\{\alpha:\sum_{i=0}^{c}\alpha_i=1,\ \alpha_i\ge 0,\ i=1,\cdots,c\}.$$

The conditions for local minimum for the objective function J_4 are derived using Lagrangian multipliers and the result are:

$$u_{ik}=\frac{\alpha_i e^{-\lambda D_{ik}}}{\sum_{j=1}^{c}\alpha_j e^{-\lambda D_{jk}}+\alpha_0 e^{-\lambda\delta^2}}$$

for $i=1,\cdots,c$, and

$$u_{0k} = \frac{\alpha_0 e^{-\lambda\delta^2}}{\sum_{j=1}^{c} \alpha_j e^{-\lambda D_{jk}} + \alpha_0 e^{-\lambda\delta^2}}$$

$$v_i = \frac{\sum_{k=1}^{n} u_{ik} x_k}{\sum_{k=1}^{n} u_{ik}}$$

$$\alpha_i = \frac{1}{n} \sum_{k=1}^{n} u_{ik}$$

for $i=0,\cdots,c$.

4 Fuzzy c- logistic regression

Fuzzy c-logistic regression model yield simultaneous estimates of parameters of c regression models together with a fuzzy c-partitioning of the data. Fuzzy c-logistic regression model takes the following form.

$$f(x_k;\beta_i) = \sum_{j=1}^{p} \beta_i^j x_k^j + \beta_i^{p+1}$$

where $x_k = [x_k^1,\ldots,x_k^p]^T$ denotes the kth data sample. The c-regression constant β_i is expressed as $(\beta_i^1,\ldots,\beta_i^{p+1})^T$. For simplification $z_k = (x_k^1,\ldots,x_k^p,1)^T$ is used and the scalar product is written as

$$\langle \beta_i, z_k \rangle = \sum_{l=1}^{p+1} \beta_i^l z_k^l$$

The membership degree $u_{ik} \in U$ is interpreted as a weight representing the extent to which the value predicted by the model $f(x_k;\beta_i)$ matches y_i. The dissimilarity is defined by

$$D_{ik} = \left(y_i - \langle \beta_i, z_k \rangle\right)^2$$

where $y_i = \log\left(\frac{P_i}{1-P_i}\right)$ is a logit term.

The term $P_i/1\text{-}P_i$ is the odds ratio for each level of x. It is calculated as observed probability over one minus the observed probability. P_i describes risk in logistic model for individual x [10].

In logistic regression, the dependent variable is a logit, which is the natural log of the odds, that is,

$$\log(odds) = \operatorname{logit}(P) = \ln\left(\frac{P}{1-P}\right)$$

A logit is a log of odds and odds are a function of P, the probability of a 1. In logistic regression we find

$$\operatorname{logit}(P) = a + bx$$

Converting the odds to a simple probability yields:

$$\ln\left(\frac{P}{1-P}\right) = a + bx$$

$$\frac{P}{1-P} = e^{a+bx}; \quad P = \frac{e^{a+bx}}{1+e^{a+bx}}$$

The objective function is given as

$$J_5(U,\beta,\alpha) = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik} \left(y_i - \langle \beta_i, z_k \rangle\right)^2 + \sum_{k=1}^{n} u_{0k}\, \delta^2 + \lambda^{-1} \sum_{i=0}^{c}\sum_{k=1}^{n} u_{ik} \log \frac{u_{ik}}{\alpha_i}$$

The method of fuzzy c-means uses an alternative minimization of the objective function $J(U,\beta,\alpha)$ to find the optimum value of U and V. An additional variable $\alpha = (\alpha_i,\ldots,\alpha_c)$ with the constraint A is used for controlling the sizes of the clusters.

The following alternative optimization algorithm is used for clustering

Algorithm of Fuzzy c-logistic regression

FCLRA 1 Set the initial values

FCLRA 2 β and α are treated as constant. Minimizing the objective function $J(U,\beta,\alpha)$ the membership value U can be found.

FCLRA 3 U and α are treated as constant. Minimizing the objective function $J(U,\beta,\alpha)$ the prototype β can be found.

FCLRA 4 U and β are treated as constant. Minimizing the objective function $J(U,\beta,\alpha)$ the value of α can be found.

The optimal solutions of U, β and α are as follows.

$$u_{ik} = \frac{\alpha_i e^{-\lambda\left(y_i - \langle \beta_i, z_k \rangle\right)^2}}{\sum_{j=1}^{c} \alpha_j e^{-\lambda\left(y_i - \langle \beta_i, z_k \rangle\right)^2} + \alpha_0 e^{-\lambda\delta^2}}$$

for $i=1,\ldots,c$ and

$$u_{0k} = \frac{\alpha_0 e^{-\lambda\delta^2}}{\sum_{j=1}^{c} \alpha_j e^{-\lambda\left(y_i - \langle \beta_i, z_k \rangle\right)^2} + \alpha_0 e^{-\lambda\delta^2}}$$

$$\beta_i = \left(\sum_{k=1}^{n} u_{ik} z_k z_k^T\right)^{-1} \sum_{k=1}^{n} u_{ik} y_i z_k$$

$$\alpha_i = \frac{1}{n}\sum_{k=1}^{n} u_{ik}$$

The parameter δ is an arbitrary value.

5 Numerical Example

Figure 1 is an example of coronary heart disease data consisting of eight age groups [11]. In this experiment we modify the data for illustration purposes. Two points are added to serve as noisy points. The initial values of the memberships are assumed to be random. The objective function J_5 and the algorithm FCLRA is used to cluster this data. From the resulting memberships, the two added points are clustered into the noise cluster.

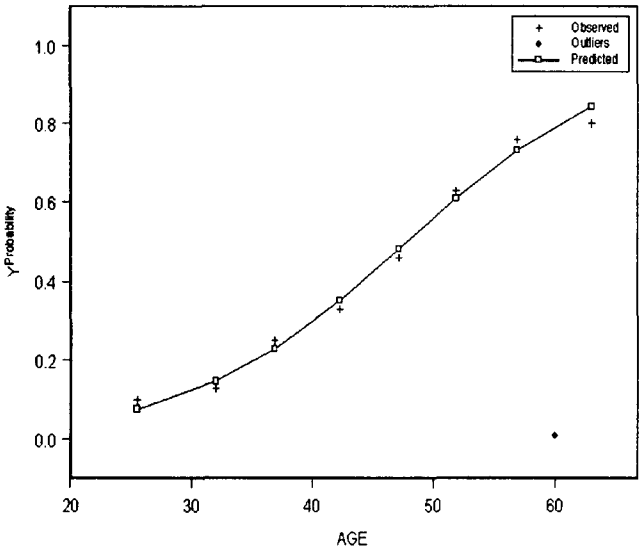

Figure 1

6 Conclusion and Future works

An alternative fuzzy clustering technique has been proposed in this paper. We have developed a modification of the well-known fuzzy *c*-regression objective function to handle the noise cluster. In our experiment the proposed method worked effectively. The noise points have been well separated from the other cluster.

Future studies include application to larger data sets with real noise present. The effectiveness of the method should moreover be investigated.

REFERENCES:

[1] V. Barnet and T. Lewis, *Outliers in statistical data, 3rd ed.* John Wiley, Chichester, 1994.

[2] J. Hartigan, *Clustering Algorithms*, Wiley New York, 1975.

[3] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, New York, 1981.

[4] F.Höppner, F. Klawonn, R. Kruse, T. Runkler, *Fuzzy Cluster Analysis*, Wiley, Chichester, 1999.

[5] S. Miyamoto and M. Mukaidono, Fuzzy C-means as a regularization and maximum entropy approach, *Proc. of the 7th International Fuzzy Systems Association World Congress* (IFSA'97), June 25-30, 1997, Prague, Chech, Vol. II, pp. 86-92, 1997.

[6] K. Miyagishi, Y.Yasutomi, H. Ichihashi, K. Honda, Fuzzy Clustering with regularization by K-L information. *16th Fuzzy System Symposium*, Akita, Sept. 6-8, 2000, pp.549-550 (in Japanese).

[7] R. N. Davé, "Characterization and detection of noise in clustering", *Pattern Recognition Letters.*, Vol. 12, pp. 657-664, 1991.

[8] K. Miyagishi, Y.Yasutomi, H. Ichihashi, K. Honda, Fuzzy Clustering with regularization by K-L information. *16th Fuzzy System Symposium*, Akita, Sept. 6-8, 2000, pp.549-550 (in Japanese).

[9] . S. Miyamoto and A.C. Alanzado, Fuzzy *C*-Means and Mixture Distribution Models in the Presence of Noise Clusters. *International Journal of Image and Graphics*, Vol. 2, No. 4, pp 573-586, 2002.

[10] D. G. Kleibaum and M. Klein, Logistic Regression: A self learning text, 2nd edition. Springer-Verlag New York, Inc. 2002

[11] D. W. Hosmer and S. Lemeshow, Applied Logistic Regression, 2nd edition, John Wiley & Sons, Inc. 2000.