Scientific Bulletin of the Petru Maior University of Tirgu Mures
Vol. 6 (XXIII), 2009
ISSN 1841-9267

Fuzzy Clustering in an Intelligent Agent for Diagnosis Establishment

Zdrenghea Vlad
Iuliu Hategan University of Medicine and Pharmacy
Cluj-Napoca, Romania
vladzdrenghea@yahoo.com

Man Diana Ofelia
Department of Computer Science, Babes-Bolyai University,
Cluj-Napoca, Romania
mandiana@cs.ubbcluj.ro

Tosa-Abrudan Maria
Department of Computer Science, Babes-Bolyai University,
Cluj-Napoca, Romania
maria@cs.ubbcluj.ro

Abstract

In this paper we present a way to use fuzzy clustering for generating fuzzy rule bases in the implementation of an intelligent agent that interacts with human for diagnosis establishment: The (Psychiatric) Medical Diagnostics System. The system is intended to be a software learning application mainly destined to orientate the resident doctors in the process of establishing a diagnostic for the patients they are examining.

Learning techniques have been used for clustering and to form the premise portion of the If-Then rules. The general idea is to generate a set of fuzzy rules that not only best describe the data at hand but also are robust enough to show good generalization capabilities.

1. Introduction

Intelligent Agents Systems and Multi-agent systems are the common words that largely supplant for Distributed Artificial Intelligence (DAI) Systems. Distributed Artificial Intelligence (DAI) systems can be defined as cooperative systems where a set of agents act together to solve a given problem. These agents are often heterogeneous, for example in Decision Support System the interaction takes place between a human and an artificial problem solver. In the Medical Decision System the human operator - a resident medical doctor - is interacting with the software in order to achieve the patient diagnostic.

In DAI, there is no universal definition of “agent”, but Ferber’s definition is quite appropriate for drawing a clear image of an agent: ”An agent is a real or virtual entity which is emerged in an environment where it can take some actions, which is able to perceive and represent partially this environment, which is able to communicate with the other agents and which possesses an autonomous behavior that is a consequence of its observations, its knowledge and its interactions with the other agents” [1].

The Medical Diagnosis System is an agent kind software program that is taking somehow the role of an experienced medical person, which benefits of a vast medical knowledge regarding symptoms and diseases and haze the role to orientate the young resident doctors in the process of diagnosis establishment. The action this agent system is taken is to generate at each iteration the next more appropriate question whose answer will bring the diagnosis process closer to its end: the diagnostic of the patient. The environment the agent is able to represent is given by a knowledge database that contains general symptoms like temperature, symptoms values (e.g. 38 degrees is a symptom value for the symptom temperature) , diseases hierarchical structure, the symptom values that are associated to a disease (38 degree temperature is associated to a flue) and the relation between this disease symptom values, some may be mandatory, some may be optional and each symptom value has a weight meaning its ”importance” for a specific disease. The medical diagnostic systems is not communicating with some other software intelligent agents at this time at least, but is interacting with a human operator, re-evaluates the situation and gives

Keywords: *fuzzy clustering, medical diagnostics systems*

an new suggestion question at each iteration/ after each answer.

Medical diagnosis is an excellent field where fuzzy sets theory can be applied with success, due to the high prominence of sources of uncertainty that should be taken into account when the diagnosis of a disease must be formulated.

The medical diagnosis problem is inherently a classification problem, where for each vector of symptoms measurements one or a set of possible diagnoses are associated [2].

Fuzzy system can be designed based on expert knowledge. Several approach have been proposed to built fuzzy system from numerical data, including fuzzy clustering-based algorithms, neoro-fuzzy systems and genetic fuzzy rules generation. First, in order to obtain a good initial fuzzy system, a fuzzy clustering algorithm is used, to identify the antecedents of fuzzy system, while the consequents are designed separately to reduce computational burden. Second, the precision performance, the number of fuzzy rules and the number of fuzzy sets are taken into account. Among the different fuzzy modeling techniques, the Takagi-Sugeno (TS) model has attracted most attention.

This paper is concerned with rule extraction from data by means of fuzzy clustering in the product space of inputs and outputs where each cluster coresponds to a fuzzy IF-THEN rule.

The rest of paper is organized as follows. In Section II, the TS fuzzy model is presented, next we describe a variety of fuzzy clustering methods and present some examples. The last subsection concludes the paper.

2. Fuzzy modeling and fuzzy clustering

2.1. Takagi-Sugeno (TS) fuzzy model

The Takagi-Sugeno fuzzy model is a fuzzy rule-based model suitable for the approximation of many systems and function. The construction of a TS fuzzy model is usually done in two steps. In the first step, the fuzzy sets (membership function) in the rule antecedents are determined. In the second step, the parameters of the consequent functions are estimated [3].

In the TS fuzzy model, the rule consequents are typically taken to be either crisp numbers or linear functions of the inputs: R_i: IF x is A_i THEN $y_i = a_i^T x + b_i, i = 1, 2...M$, where $x \in R^n$ is the input variable (antecedent) and $y \in R$ is the output (consequent) of the i^{th} rule R_i. The number of rules is denoted by M and A_i is the (multivariante) antecedent fuzzy set of the i^{th} rule:

$$A_i(x) : R^n \rightarrow [0,1], A_i(x) = \prod_{i=1}^{n} u_{ij}(x_j) \qquad (1)$$

where $u_{ij}^{(x_j)}$ is the univariate membership functions. For the k^{th} input x_k, the total output y(k) of the model is computed as follows:

$$y(k) = \sum_{i=1}^{n} u_{ki} y_i(k) \qquad (2)$$

where u_{ki} is the normalised degree of the fulfilment of the antecedent clause of rule R_i:

$$u_{ik} = \frac{A_i(x_k)}{\sum_{j=1}^{M} A_j(x_k)} \qquad (3)$$

2.2. Fuzzy clustering algorithm

2.2.1 Fuzzy C-Means algorithm

The most popular fuzzy clustering algorithm is *Fuzzy c-Means* (Bezdek, 1981). It is a data clustering technique wherein each data point belongs to a cluster to some degree that is specified by a membership grade. It tooks several names before FCM such as: Fuzzy ISODATA, Fuzzy K-Means.

Given a set $X = \{x_1, x_2, ..., x_n\} \subset \Re^p$ of sample data, the aim of the algorithm is to determine the prototypes in such a way that the objective function is minimized.

The objective function is:

$$J(M, p_1, p_2, ..., p_c) = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c} (\sum_{k=1}^{n} u_{ik}^q d_{ik}^2)) \qquad (4)$$

subject to:

$$\sum_{k=1}^{n} u_{ik} > 0, \forall i \in \{1, 2...c\}, \sum_{i=1}^{c} u_{ik} = 1, \forall k \qquad (5)$$

where u_{ik} stands for the membership degree of datum x_k to cluster i, d_{ik} is the distance of datum x_k to cluster i, represented by the prototype p_i and c is the number of clusters. The parameter q ($q \in [1, \infty)$) is a weighting exponent (fuzziness exponent). Usually q=2 is chosen. At q=1, FCM collapses to HCM algorithm.

The first constraint guarantees that no cluster is empty and the second condition ensures that the sum of the membership degrees for each datum equals 1.

The output of FCM algorithm is not a partition, thus: $C_i \bigcap C_j \neq \emptyset, i \neq j$.

There are two necessary conditions for J to reach a minimum:

$$p_i = \frac{\sum_{k=1}^{n} u_{ik}^q x_k}{\sum_{k=1}^{n} u_{ik}^q} \qquad (6)$$

$$u_{ik} = \frac{(\frac{1}{d_{ik}})^{1/(q-1)}}{\sum_{j=1}^{c} (\frac{1}{d_{jk}})^{1/(q-1)}} \qquad (7)$$

where d_{ik} is the distance between object x_k and the center of cluster C_i [4].

2.2.2 Clustering with fuzzy covariance matrix

The Gustafson-Kessel algorithm (Gustafson and Kessel, 1979) is an extension of FCM algorithm and it uses the Mahalanobis distance, in order to adapt to various sizes and forms of the clusters.

It uses for each cluster a separate matrix:

$$A_i = (detC_i)^{1/n} C_i^{-1} \quad (8)$$

where C_i is the covariance matrix, which defines the shape of the cluster.

The objective function is:

$$J(P, U, A; X, c, q) = \sum_{i=1}^{c} u_{ik}^q ((x_k - p_i)^T A_i (x_k - p_i)) \quad (9)$$

To minimize the objective function with respect to prototypes, they are updated according to the following equations:

$$p_i = \frac{\sum_{k=1}^{n} u_{ik}^q x_k}{\sum_{k=1}^{n} u_{ik}^q} \quad (10)$$

$$C_i = \frac{\sum_{k=1}^{n} u_{ik}^q (x_k - p_i)(x_k - p_i)^T}{\sum_{k=1}^{n} u_{ik}^q} \quad (11)$$

Gath-Geva algorithm (Gath and Geva, 1989) is an extension of Gustafson-Kessel algorithm and it can be used to detect ellipsoidal clusters with varying size. It can be used to detect lines. This algorithm, in contrast with FCM algorithm, is not based on an objective function, but is a fuzzification of statistical estimators.

This algorithm takes the data as a comprehension of p-dimensional normally distributed random variables. Statistics provides for expected value of the i^{th} normal distribution, the estimators:

$$m_i = \frac{\sum_{j=1}^{n} u_{ij} x_j}{\sum_{j=1}^{n} u_{ij}} \quad (12)$$

and for covariance matrix:

$$A_i = \frac{\sum_{j=1}^{n} u_{ij} (x_j - p_i)(x_j - p_i)^T}{\sum_{j=1}^{n} u_{ij}} \quad (13)$$

The normal distribution N_i with value v_i and covariance matrix A_i is chosen to generate a datum with a priori probability p_i. The distance function is chosen to be indirectly proportional to this unnormalized apriori probability [5].

2.2.3 Clustering with linear prototypes

The Fuzzy c-Varieties algorithm (Bezdek, Coray, Gunderson and Watson,1981) is one of those algorithms where the prototype of each cluster is a multi-dimensional linear variety.

Thus, the prototypes are tuples of the form $(v_i, e_i^{(1)}, ..., e_i^{(r)}) \in (\Re^p)^{r+1}$ and induce r-dimensional linear varieties:

$$\{y \in \Re^p | y = v_i + \sum_{j=1}^{r} t_j e_i^{(j)}, t_1, ..., t_r \in \Re\} \quad (14)$$

i.e. lines for r=1 and planes for r=2. The distance function

$$d^2(v_i, x_k) = \|x_k - v_i\|^2 - \sum_{j=1}^{r} ((x_k - v_i)^T e_i^{(j)})^2 \quad (15)$$

assigns the distance 0 exactly to those data that are lying in the linear variety determined by the prototype.

The Fuzzy c-Elliptotypes algorithm (Bezdek, 1981) detects clusters that have the shape of lines or planes.

It uses distance measures which is a convex combination of the distance (12) and the Euclidian distance between the cluster center v_i and the datum x_k in order to avoid that, for instance for r=1, collinear, separated, short lines are lumped together in one cluster.

The eigenvectors are arranged in descending order of the corresponding eigenvalues. The first eigenvector describes the direction of the longest axis of the cluster. When r=1, the algorithm detects lines, when r=2 it detects planes [6].

2.3. Fuzzy Diagnosis

The problem of medical diagnosis can be formalized as a classification problem, where a set of c diagnoses are defined for a certain medical problem and formalized as class labels:

$$C = \{C_1, C_2, ..., C_c\} \quad (16)$$

In order to assign a diagnosis to a patient, a set of symptoms are measured and formalized as a n-dimensional real vector $x = (x_1, x_2, , x_n)$. To perform diagnosis, a classifier is needed to perform a mapping:

$$D : X \subseteq \Re^n \rightarrow C \quad (17)$$

The domain X defines the range of possible values that each component of x can hold [2].

Our dataset consists of some cases with several attributes. Each symptoms are associated the symptoms values. For instance: Daily Disposition (Sad, Happy, Normal), Weight Change (Growth, Drop, No), Insomnia (Yes, No), Attempted Suicide (Yes, No), Social Life (Isolation, Normal), Personal outfit (Damaged, Good), Language (Vague, Normal), Delusion (Yes, No), Hallucinations (Yes, No), Thinking (Magical, Normal). We can consider four classes of diagnoses: schizophrenia, mood disorders, personality disorder, disorders due to substance use psychoactive.

2.4. Identification by Fuzzy Clustering and Cluster Reduction

A clustering method that has proven suitable for the identification of TS fuzzy model is the Gustafson-Kessel algorithm. Compared with Fuzzy C-Means algorithm, it employs an adaptive distance norm in order to detect clusters of different geometric shapes in the data set.

Each cluster in the product space of the input/output data, represents a rule in the rule base. The goal is to establish the fuzzy antecedents A_i in the rule (1) and these are defined by the fuzzy clusters found in the data. Univariate membership function u_{ij} can be obtained by projections onto the various input variables x_j spanning the cluster space.

The Fuzzy clustering algorithm produces a fuzzy partition of the product space of all data, whereas fuzzy i-then rules are usually defined on the basis of fuzzy partitions of the single domains, which means that the projection of the fuzzy cluster can lead to unusual fuzzy partitions on the single domains and enforces again a loss of information. To reduce this loss of information it is recommended to restrict to diagonal matrices C_i when using the Gustafson-Kessel algorithm. In this way, the fuzzy clusters are forced to be in the form of axis-parallel hyperellipsoids [7].

An important issue in clustering is the determination of the relevant number of clusters in the data. Cluster validity techniques attempt to assess the "correctness" of a particular set of clusters in a given data set. Among them ar: Partition Index(SC), Separation Index (S) and Xie and Beni's Index (XB).

2.5 Experiments and Results

Iris Dataset. The Iris classification problem are considered. The data set contains 50 measurements of four features (sepal length, sepal width, pental length, pental width) from each of three species (setosa, versicolor, virinica). The first class is linearly separable from the others ; the second and third class are overlap slightly.

Figure 1-2 shows the two-dimension (sepal length, sepal width) measurement, for three classes.

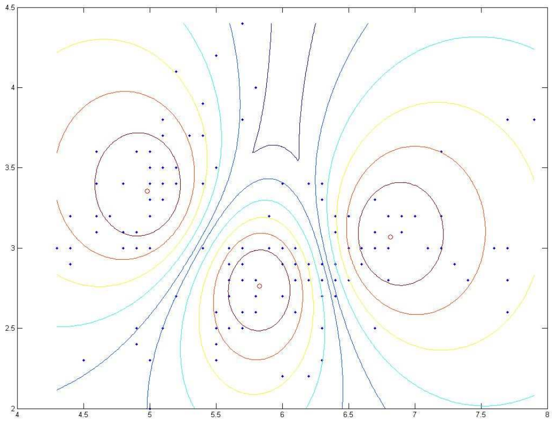

Figure 1: Fuzzy C-Means clustering for Iris data: c=3

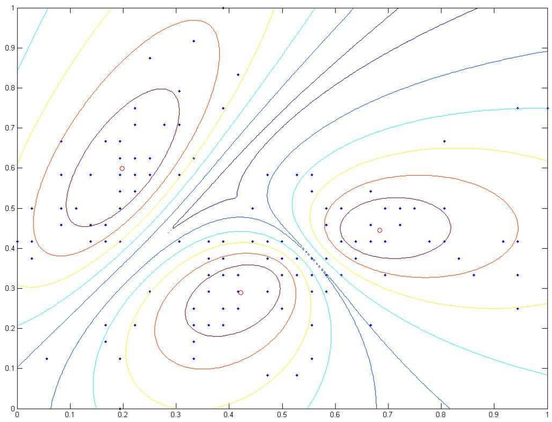

Figure 2: Gustafson-Kessel clustering for Iris data: c=3

Comparing with FCM, the Gustafson-Kessel algorithm can detect the elongated clusters better.

In Tabel 1, the validity measures indicate that three cluster is the optimal number.

Methods	Validity index	2 Cl	3 Cl	4 Cl
FCM	SC	0.6348	0.4872	0.4281
FCM	S	0.0042	0.0050	0.0043
FCM	XB	7.4704	4.9451	5.3422
GK	SC	2.2197	1.5931	1.7250
GK	S	0.0148	0.0140	0.0188
GK	XB	13.0666	8.5185	8.9598

Table 1: The validity measures for 2, 3 and 4 clases

The initial fuzzy system is obtained by the Gustafson-Kessel algorithm. The number of fuzzy rules is 9 and the number of fuzzy sets is 36.

The multi-objective genetic algorithm can be used to optimize the initial fuzzy system.

Finally, the fuzzy system of the Iris problem will be:
R1: If sepal length is big, sepal width is big then output is virinica
R2: If sepal length is small, sepal width is small then output is setosa
R3: If sepal length is medium, sepal width is big then output is virinica
R4: If sepal length is medium, sepal width is small then output is versicolor

Parameters of antecedents of the fuzzy system:
sepal length: small=[0.048457, 0.040494]
medium=[0.53457, 0.020748]
big=[0.88115, 0.030973]
sepal width:
small=[0.39215, 0.084652]
big=[0.82529, 0.030372]

All simulation programs are realized under Matlab 7.0 environment.

•

2.6 Conclusion

The application of fuzzy clustering to the identification of Takagi-Sugeno (TS) fuzzy models has been addressed. Methods to extract a fuzzy model from fuzzy clusters obtained by Fuzzy C-Means and Gustafson-Kessel algorithm is presented.

As further work we want to use these techniques for contructing an intelligent agent that interacts with human for diagnosis establishment.

References

[1] J. Ferber, *Multi-Agent System: An Introduction to Distributed Artificial Intelligence*, Addison-Wesley, 1999

[2] G. Castellano, A.M. Fanelli, C. Mencar *A Fuzzy Clustering Approach for Mining Diagnostic Rules*,Conference on Systems, Man and Cybernetics, 2003. IEEE International, 2, 2003, 2007-2012

[3] M. Setnes, *Supervised Fuzzy Clustering for Rule Extraction*,IEEE Transactions on Fuzzy Systems, 8, 4, 2000, 416-424

[4] A.I. Shihab, *Fuzzy clustering algorithms and their application to Medical Image Analysis*,PhD Thesis, 2001, http://en.wikipedia.org/wiki/clusteranalysis

[5] R.Kruse, C.Doring, MJ.Lesot, *Advances in Fuzzy Clustering and its Applications*, Hardcover, 2007

[6] R.J. Hathaway, J.C. Bezdek, *Switching Regression Models and Fuzzy Clustering*, IEEE Transations on Fuzzy Systems, 1, 3, 1993, 195-204

[7] F. Klawonn, A. Keller *Fuzzy clustering and fuzzy rules*, IFSA 97 Prague Proceedings Volume I, 6, 1997, 193-198