STUDIA UNIV. BABEŞ–BOLYAI, INFORMATICA, Volume **LV**, Number 2, 2010

FUZZY CLUSTERING IN AN INTELLIGENT AGENT FOR DIAGNOSIS ESTABLISHMENT

VLAD ZDRENGHEA, DIANA OFELIA MAN, AND MARIA TOSA-ABRUDAN

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 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.

We used fuzzy c-means clustering to assign symptoms to the different types of aphasia categories. The results were compared with the results in some subtests of the Aachen Aphasia Test (AAT).

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

Received by the editors: December 7, 2009.

2010 *Mathematics Subject Classification.* 03B52, 93C42.

1998 *CR Categories and Descriptors.* I.2.1 [**Computing Methodologies**]: Artificial Intelligence – *Applications and Expert Systems*; I.2.3 [**Computing Methodologies**]: Artificial Intelligence – *Deduction and Theorem Proving*.

Key words and phrases. Fuzzy clustering, Fuzzy systems, Expert Systems, Uncertainty.

This paper has been presented at the International Conference Interdisciplinarity in Engineering (INTER-ENG 2009), Târgu-Mureş, Romania, November 12–13, 2009.

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 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:

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

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:

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

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

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

2.2. **Fuzzy clustering algorithm. 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:

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

subject to:

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

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:

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

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

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

FCM Algorithm

1. Initialize the membership matrix U with random values between 0 and 1 within the constraints of (2).
2. Calculate c cluster centres $p_i, i = 1..c$ using (3).
3. Compute the objective function according to (1). Stop if either it is below a certain threshold level or its improvement over the previous iteration is below a certain tolerance.
4. Compute a new U using (4).
5. Go to step 2.

Advantages:

1. Compared with the HCM, it is quite insensitive to its initialization.
2. FCM algorithm generalised the notion of membership to emulate the fuzzy clustering structures found in real-world.
3. It employs an intuitive objective function.
4. It performs robustly, thus it always converges to a solution.
5. It is simply in terms of programming implementation.
6. It minimizes intra cluster variance as well.

Disadvantages:

1. We must know the number of clusters a priori.

2. It is sensitive to initialisation.
3. It is sensitive to noise and outlier points.
4. It finds clusters of the same shape.
5. It does not use a good clustering criterion when clusters are close to one another but are not equal in size or population [5].

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:

$$C = \{C_1, C_2, ..., C_c\} \tag{8}$$

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:

$$D : X \subseteq \Re^n \to C \tag{9}$$

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.

2.5. **Experiments and Results. Aachen Aphasia Test (AAT).**

Aachen Aphasia Test (AAT) is publicly available at the following web address: http://fuzzy.iau.dtu.dk/aphasia.nsf/PatLight . Aphasia is the loss or impairment of the ability to use or comprehend words often a result of stroke

or head injury. Data of 256 aphasic patients, treated in the Department of Neurology at the RWTH Aachen, were collected in a database since 1986.

In aphasiology, there are many inconsistencies concerning the definition and interpretation of aphasic syndromes. In a clinical setting, the following aphasic syndromes are distinguished. These syndromes are strictly empirical and based on a statistically reliable co-occurrence of a set of symptoms: Broca's Aphasi, Wernicke's Aphasia, Global Aphasia, Anomic Aphasia, Conduction Aphasia.

Factor analysis was applied on a correlation matrix of 26 symptoms of language disorders and led to five factors (Keyserlingk et al., 2000). These factors displayed meaningful indication of the disease.

Factor-No.	Meaning
I	severity of disturbance
II	expressive vs. comprehensive
III	granularity of phonetic mistakes
IV	awareness of disease
V	deficits in communication

Table 1-Factors derived from the factor analysis

After the factors have been gained they are usually transformed into 'simple structure' to render easier interpretation of their significance. The principle of the 'simple structure' is to work out from all possible feature configurations - how scattered they may be - the ideal configuration, in which the variable possesses the simplest complexity, i.e., it can be described by only one single factor. We treated the factors with the so-called varimax method (Weber, 1980).

Fuzzy c-mean clustering (Bezdek, 1981) was then used to advise the symptoms to the different entities, because of polarization of the five factors results in at least 10 categories.

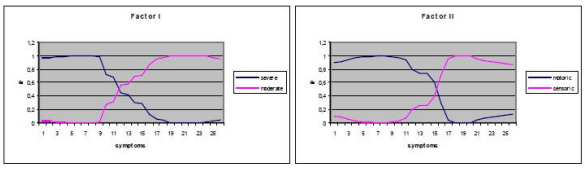

Figure 1: Graphical presentation of the results of the c-means clustering. Factor I (left) represents the overall severity of disturbance whereas factor II (right) indicates the more expressive or more comprehensive character of the language disorder

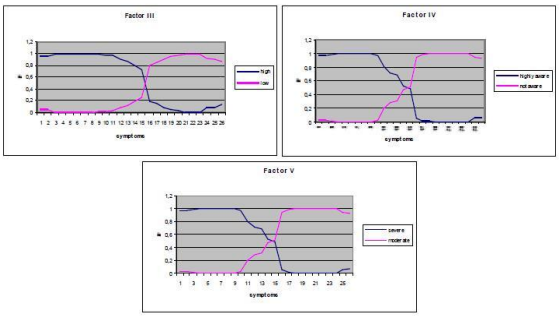

Figure 2: Graphical presentation of the results of the c-means clustering. Factor III and IV (upper row) represent the granularity of the phonetic language disorders and the patients awareness of the disease, factor V (below) exposes the deficits in communication.

The resulting classes of the clustering method are presented in the Figure . For graphical interpretation the different symptoms were put in order according to their membership to the respective feature, severe or moderate overall severity of disturbance. The clustering procedure leads to clearly distinguishable classes of symptoms. The clusters can be separated simply, as it is indicated in the small areas of overlap between the respective features. Furthermore, the description of language failures by Fuzzy C-Mean classification of analyzed factors correspond in many but not in all cases to the traditional diagnostic scheme [6].

2.6. **Conclusion.** The application of fuzzy clustering to the identification of Takagi-Sugeno (TS) fuzzy models has been addressed.

We used c-mean fuzzy clustering for classification after feature extraction from an aphasia database. The additional feature extraction allows to ensure the statistical validity of the factors. The clustering method seems to be insufficient to distinguish the granularity of the phonetic mistakes correctly. However, overall severity of the disease and the character of the language disorder can be distinguished much better.

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] F. Hoppner, F. Klowonn, R. Kruse, T. Runkler, *Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition*, John Wiley and Sons, 1999.
[5] R. Kruse, C. Doring, MJ. Lesot, *Advances in Fuzzy Clustering and its Applications*, Hardcover, 2007
[6] G. Berks, D.G. von Keyserlingk, J. Jantzen, M. Dotoli, H. Axer, *Fuzzy clustering - A versatile mean to explore medical databases*, ESIT 2000, 2000, 453–457.

Iuliu Haţieganu University of Medicine and Pharmacy, Cluj-Napoca, Romania

Department of Computer Science, Babeş-Bolyai University, Cluj-Napoca, Romania

Department of Computer Science, Babeş-Bolyai University, Cluj-Napoca, Romania
E-mail address: vladzdrenghea@yahoo.com, {mandiana,maria}@cs.ubbcluj.ro