Fuzzy Reasoning In K-Means Classification Method

Irene DIAZ, Manuel VELASCO and Jose M. MOLINA

Departamento de Informática
Universidad Carlos III of Madrid, Butarque, 15
28911 Leganés, Madrid, España

Summary

Domain analysis tries to reuse software in an effective way. New methodologies are starting to be able to automate the process, in different degrees, with the construction of a domain model for each problem. The general process is divided into several phases. One of the most difficult tasks is the generation of the relationships which have to be defined between the components in the domain. In this paper the use of fuzzy logic and a statistical classification method in order to get the semantic relationships for each pair of components is presented.
Keywords: Domain analysis, Classificaton, Fuzzy System.

1 INTRODUCTION

Information classifying techniques usually use clustering algorithms to order the information. Each of those methods must measure distances between two vectors (that represent the information [8]). These distances are used to decide if a given vector belongs to a cluster or not. As the nature of the problem is often inexact and subjective, the use of a traditional distance does not provide successful results because such distances do not take account of all semantic information known about the data. Then, the use of a knowledge-based system to solve the problem allows us to take advantage of the semantic information that a domain provides us with. So, this distance should improve results obtained from traditional distances. These kind of fuzzy distances have been used, for instance, in the processing of stereoscope images to calculate distances among points of different images [3].
In this paper we use fuzzy reasoning in the subject of construction and validation of a methodology for automatic construction of domains [5]. Domain analysis was introduced in [6] as "An activity of identifying objects and operations of a class of similar systems in a domain of the specific problem". A process of global domain construction is shown in [8]. That method has five phases which are *Plan of the project and domain characterisation, domain components acquisition, data analysis, classification of components, and evaluation of the domain model.*
Classification phase tries to cluster the knowledge considered as similar. Once the knowledge (represented as terms or 'descriptors') has been clustered at a first level, common characteristics more relevant of each cluster are abstracted again and each cluster is divided into other clusters. The result of that process organised the knowledge in hierarchies.
In order to classify the information, statistical, neural and bibliometric methods are used to get these relationships. In this work, a statistical method is presented, the K-means. To sum up, when the process is beginning, we have a set of vectors $\{x_1, ..., x_m\}$ which represent information about each term. Each component x_{ij} of a vector x_i means the number of occurrences of term i in document j. Our purpose is to obtain a set of classes $\{a_1, ..., a_n\}$ in which these vectors are grouped. The problem is how these vectors are divided along the classes and how many classes will be.
Most of these methods are based on minimisation on some index. This process implies the use of a distance. In this work, it is proposed the use of a fuzzy distance [4], result of applying a fuzzy rule based system, which improves results obtained with an usual distance.

2 CLASSIFICATION PROBLEM

The problem is the classification of information over a particular domain [1],[5].To classify the information clustering analysis is used in this work. Clustering can be defined as an unsupervised process to classify objects. The distribution of the objects into the different

classes is not well known; but a vectorial representation of the objects is available $\{x_1, ..., x_p\}$. From this set of vectors, the classification process try to achieve the set of classes that group them $\{a_1, , a_n\}$. The problem is that neither the distribution nor the number of classes is known before the process. However the structure of the features vector is well defined (built up with the number of occurrences of the current candidate in each document).
Therefore, the problem consists in getting groups from the set of vectors in function of the similarities found between them.
Clustering is carried out according to the pattern similarity, by means of the used distance. The introduction of a fuzzy distance is based on the need of gathering as semantic information as possible.
The classification algorithm used in this work is K-Means. This algorithm is one of the most popular for clustering analysis. There are several algorithm versions, and one of the most efficient is the Anderberg's convergence algorithm [2]. It is an iterative and more formalised algorithm, in front of other excessive heuristic, for example the max-min algorithm. It is a quick and efficient algorithm which tries to minimise an index, based in adding all the quadratic Euclidean distances (fuzzy distances in this case) of all the members in a cluster to the cluster centre. It needs to know the k number that defines the number of cluster in the collection. If the number of classes is unknown then it is a trouble. One efficient solution consists in defining the number of classes using some user-defined parameters. If a collection is available with N training patterns, represented by x_i vectors, the centre (centroid) in this set of vectors is defined by

$$\frac{1}{N}\sum x_i \tag{1}$$

The way the algorithm works consists in moving each vector to the cluster which centre is the closest to the selected vector, and later the algorithm calculates again the centre in each cluster. Algorithm convergence depends on the number of classes. This algorithm works quite well when the number of classes is known or estimated.

3 FUZZY TERM MEASURE

Let us consider multi-input, single-output fuzzy systems mapping $X \longrightarrow Y$, where $X \subset \Re^n$ and $Y \subset \Re$. A basic configuration of fuzzy system is composed of a fuzzifier, a rule system and a defuzzifier [6]. The fuzzifier performs a mapping from the observed crisp input space to the fuzzy systems defined.
A rule base of a fuzzy system consists on a number of fuzzy rules , expressed as IF-THEN form. For the fuzzy system we consider suppose you have a set of N rules

$$R^k : IF\ x_i^k\ is\ X_i^k\ AND\ ...\ AND\ x_n^k\ is\ X_n^k$$

$$THEN\ y\ is\ Y^k \tag{2}$$

where X_i^k and B^k, $i = 1, ..., n$ and $k = 1, ..., N$ are linguistic terms characterized by fuzzy membership functions. Now we define a fuzzy distance which allow us to classify the set of vectors $\{x_1, ..., x_m\}$ into n classes a_i, $i = 1, ..., n$. Each vector means the number of occurrences of a term in each document, i.e., there are vectors $x_i = \{x_{ij}\} j = 1, ..., n$, where x_{ij} means number of times that the term i appears on document j. Using an expert system to build the fuzzy distance adequate to the problem, it is possible to use some information that should be overlook if an usual distance is used. This information is essential to solve the problem successfully. The objective is to know if two terms are related in some sense, that is, if two terms are synonyms, one is generic to other one, or if they treat about the same theme. The proposed multi-input single-output fuzzy system has four input variables and one output variable [1].These are briefly explained in the next subsections.

3.1 INPUT VARIABLES

3.1.1 NOT ZERO COINCIDENCES

It seems that terms that belong to same documents treat about more or less the same subject. Then, it is significant that a vector's component would be zero or not because this fact indicates that a term is, or is not, in a document. We define "Not zero coincidence" (nzc) input variable as follows. Let x_i and x_j be two vectors, It is said that one of their component k, coincide iff

$$x_{ik} \neq 0 \neq x_{jk} \tag{3}$$

The number of these coincidences will be an input to the system. This input treats to group those terms that are in the same documents. That fact is reasonable because two terms in different documents will have less likelihood to be related in some sense. Note that it is not very important the number of documents in which a term should be. The really important things are two, the first one is that two terms are, or not, in the same document and the second one how many documents are simultaneously two terms in.

3.1.2 NOT COINCIDENCE

It is also important to know if two terms are not simultaneously in a document, because if two terms are in different documents, the relationship between

them will probably be small. We define "Not coincidence"(nc) input variable as follows. Let x_i and x_j be two vectors, It is said that one of their component k, coincide iff

$$x_{ik} \neq 0\, AND\, x_{jk} = 0\, OR\, x_{ik} = 0\, AND\, x_{jk} \neq 0. \quad (4)$$

Summing over k we obtain the input value for this input variable,that is, we obtain the number of documents in which the two terms under study are not simultaneously.

3.1.3 MAXIMUM DIFERENCE IN THE NUMBER OF OCCURRENCES

This input variable (no) tries to weight the difference between the number of times that two vectors appear in the corpus. So for each corpus the maximum number of occurrences for a term is estimated as follows. *Given two vectors(representing terms) x_i and x_j, the maximum difference in the number of occurrences for those terms are*

$$max_{k=1,\ldots,n}|x_{i_k} - x_{j_k}| \quad (5)$$

3.1.4 NUMBER OF DOCUMENTS

The number of documents (nd) is a variable used to weight the relevance of the size of the corpus used. In fact, the number of documents is very important for the classification, because the information does not grow linearly with the growing of the number of documents.

3.2 OUTPUT VARIABLE: DISTANCE

The output is a value that indicates whether two vectors are near or not. The relationship between two descriptors is inversely proportional to the obtained value. To calculate output fuzzy sets, the bound of the value range must be known.

3.3 SYSTEM RULES

We have used three linguistic labels

Small, Medium, High

for each input variable and five for the output variable

Very Small, Small, Medium, High, Very$h$igh .

Now we present some of the rules we use:

IF nzc is S AND nc is S AND no is S AND nd is A

THEN distance is S

IF nzc is S AND nc is S AND no is H AND nd is S

THEN distance is Vh

IF nzc is S AND nc is M AND no is M AND nd is S

THEN distance is H

IFnzc is M AND nc is SAND no is S AND nd is S

THEN distance is S

IFnzc is M AND nc is S AND no is S AND nd is M

THEN distance is S

IFnzc is M AND nc is S AND no is SAND nd is H

THEN distance is Vs

IFnzc is M AND nc is S AND no is M AND nd is S

THEN distance is M

IFnzc is H AND nc is S AND no is H AND nd is H

THEN distance is Vs

IFnzc is H AND nc is M AND no is S AND nd is S

THEN distance is Vs

IFnzc is H AND nc is M AND no is S AND nd is M

THEN distance is S

IFnzc is H AND nc is M AND no is S AND nd is H

THEN distance is S

4 RESULTS

In the experiments fulfilled to contrast the results using the euclidean distance in k-means or the fuzzy distance in K-means algorithm presented here, we have made several experiments varying the number of documents (that is, the dimension of the vectors)in the corpus. Here we present one of those experiments.
We have selected only two documents in this example so that the number of terms is low (212).
The number of terms that is represented by the same vector is shown in table 1. In table 2 we can see in the

Table 1: Number of terms represented by the same vector

VECTOR	NUMBER OF TERMS
(1, 0)	41
(2, 0)	16
(3, 0)	5
(4, 0)	1
(5, 0)	3
(6, 0)	2
(8, 0)	1
(10, 0)	2
(15, 0)	2
(0, 1)	24
(0, 2)	3
(0, 6)	1
(0, 7)	1
(1, 1)	2
(2, 2)	1
(3, 3)	1
(2, 1)	1
(3, 2)	1
(4, 1)	1

Table 2: K-Means Classification versus K-Means with Fuzzy Reasoning Classification

VECTOR	K-Means Class	FK-Means Class
(1, 0)	1	1
(2, 0)	3	1
(3, 0)	3	1
(4, 0)	3	3
(5, 0)	3	3
(6, 0)	3	3
(8, 0)	5	3
(10, 0)	5	3
(15, 0)	5	5
(0, 1)	1	2
(0, 2)	1	2
(0, 6)	2	2
(0, 7)	2	2
(1, 1)	1	4
(2, 2)	4	4
(3, 3)	4	4
(2, 1)	4	4
(3, 2)	4	4
(4, 1)	3	4

second column the final result obtained by k-means and in the third one, the final result using k-means with fuzzy reasoning.

As result of this research we can conclude the classification that provide us K-Means with Fuzzy Reasoning is better for the purposes of this research because that classification is according with the semantic of the terms.

References

[1] Díaz,Velasco and Molina,*An application of fuzzy logic to domain analysis*,Proc. EUROFUSE-SIC'99/Pag 434-439. Budapest, May 1999.

[2] Gallant, *Neural Network Learning and Expert Systems* The MIT Press. 1995.

[3] Jiménez,Molina and Casar,*A Fuzzy Reasoning System for Binocular Scene Matching*, Proc. IEEE Workshop on NonLinear Signal/Image Processing, Pag 309-312. Greece, June 1995.

[4] Klir,Folger,*Fuzzy Sets, Uncertainty and Information*, Prentice-Hall, 1992.

[5] Lloréns,Velasco,Moreiro,Martínez and Amescua, *Automatic Domain Analysis using Thesaurus Structures*,Accepted in Journal of the American Society of Information Science, Marzo 1998.

[6] Neighbors, *Software Construction using Components*,Ph. D. Thesis. Department of Information and Computer Science. University of California. 1981.

[7] Rutkowska and Nowicki, *Constructive and Destructive Approach to Neuro-Fuzzy Systems*.Proc. EUROFUSE-SIC'99/Pag 100-105. Budapest, May 1999.

[8] Velasco,Martínez,Lloréns and Amescua, *Automatic Domain Analysis: Generation of Domain Representations*. IT-KNOWS 98. Viena. September 1998.