Neha Jain, Seema Shukla / International Journal of Engineering Research and Applications (IJERA)
ISSN: 2248-9622 www.ijera.com
Vol. 2, Issue 3, May-Jun 2012, pp.1444-1451

Fuzzy Databases Using Extended Fuzzy C-Means Clustering

Neha Jain*, Seema Shukla**
*(Department of Computer Science, JSS Academy of Technical Education, Noida, India)
** (Department of Computer Science, JSS Academy of Technical Education, Noida, India)

ABSTRACT
In recent years, the Fuzzy Relational Database and its queries have gradually become a new research topic. Fuzzy Structured Query Language (FSQL) is used to retrieve the data from fuzzy database because traditional Structured Query Language (SQL) is inefficient to handling uncertain and vague queries. The proposed model provides the facility for naïve users for retrieving relevant results of non-crisp queries and improves the relevance of results provided by Fuzzy C-Means (FCM) through the use of extended Fuzzy C-Means (EFCM). An extended fuzzy clustering algorithm based on the Gustafson-Kessel (GK) algorithm. Fuzzy C-Means and Gustafson-Kessel algorithm both are well known fuzzy clustering algorithms. Gustafson-Kessel algorithm is needed because the clustering results of the traditional Fuzzy C-Means clustering algorithm are less stable and all the clusters are spherical shaped only. Gustafson-Kessel algorithm is useful for making clusters of different geometrical shapes. The result analysis of both the algorithms is on the basis of cluster validity measures which indicate that Gustafson-Kessel algorithm is better than Fuzzy C-Means fuzzy clustering algorithm.

Keywords - **Fuzzy C-Means, Fuzzy Databases, Fuzzy Systems, Gustafson-Kessel**

1.Introduction

Database is the most important part of every organization. It is used for storing the data and retrieving the data. Generally, Structured Query Language (SQL) is used for maintaining the data. Although Structured Query Language is a very powerful tool of Relational Data Base Management System (RDBMS) but there is also some limitations with the data. In traditional databases, data is stored in the numeric and alphanumeric format. So, finder should know his actual requirements in which boundary of data he wants. Only then output comes in precise form. But in the real world user is uncertain with his requirements. If user applies his thoughts in the form of query then lot of ambiguity, uncertainty and vagueness arise. For the uncertainty or approximation of the user another type of SQL is required. So, Fuzzy Structured Query Language (FSQL) is developed.

Fuzzy relational databases extend the conventional relational database model to allow for representation of imprecise data. In general, each value in a crisp relational database is taken from a specified domain and is strongly typed and thus, the data is essentially homogeneous across all rows in the relation. Fuzzy relational databases, however, may allow heterogeneous data for an attribute. To establish its theoretical validity, fuzzy relational database theory is based on fuzzy set theory, which is extended from classical set theory. Zadeh (1965) is credited with developing both fuzzy logic and fuzzy set theory as a way to model the imprecision and uncertainty that is inherent in both the world and language.

Clustering is a mathematical tool that attempts to discover structures or certain patterns in a data set, where the objects inside each cluster show a certain degree of similarity. Clustering is useful with database in Data Storage and Retrieval Process**.** When a query is made for the address of a Person the archived data is clustered according to the various criteria, e.g.- by similar street names, within the same zip code or by similar last name.

There have been many researches for cluster analysis. Fuzzy clustering is an extension of cluster analysis. For finding the similarity in the data and grouping the data many fuzzy clustering algorithms are defined in the literature. Fuzzy C-Means algorithm and Gustafson-Kessel algorithm are two of them. They are very useful with the database.

The proposed approach is the extension of Fuzzy C-Means (FCM) algorithm. The Gustafson-Kessel (GK) algorithm is an extension of the FCM, which can detect clusters of different orientation and shape in a data set by employing norm-inducing matrix for each cluster. Gustafson-Kessel (GK) algorithm is required because Fuzzy C-Means (FCM) algorithm has some limitations. The downside with using a single matrix A is that all clusters will have the same shape and orientation. When there are clusters with different shapes, FCM will be undesirable. Gustafson and Kessel extended the FCM by employing an adaptive distance norm for each cluster to detect different geometrical shapes in data sets. Each ith cluster has its own norm-inducing matrix A_i which affects the distance norm in the FCM. Euclidean norm in the FCM is now changed as Mahalanobis distance norm.

2. Fuzzy Systems, Fuzzy Databases and Clustering

This section introduces the basics of fuzzy systems and fuzzy databases and then the concepts of clustering are described. Fuzzy clustering algorithms Fuzzy C-Means and Gustafson-Kessel are described in detail. Then section explains the concept of cluster validity measurement indexes.

2.1 Fuzzy Systems

Fuzzy logic [1],[4],[5] is a form of many-valued logic

derived from fuzzy set theory to deal with reasoning that is fluid or approximate rather than fixed and exact. In contrast with "crisp logic", where binary sets have two-valued logic, fuzzy logic variables may have a truth value that ranges in degree between 0 and 1. In simple words it can be said fuzzy logic is a super set of conventional (Boolean) logic that has been extended to handle the concept of partial truth- the truth values between completely true and completely false. Furthermore, when linguistic variables are used, these degrees may be managed by specific functions.

When A is a fuzzy set [1],[4],[5] and x is a relevant object, the proposition "x is a member of A" is not necessarily either true or false, as required by the two-valued logic, but it may be true only to some degree, the degree to which x is actually a member of A, is a real number in the interval [0, 1].
Theoretically, if X is a collection of objects denoted generically by x, then a fuzzy set F in X is a set of ordered pairs,

$$F = \{(x, \mu_F(x)) \mid x \varepsilon X\},$$

$\mu_F(x)$ is called the membership function (or grade of membership) of x in F that maps X to the membership space M. The range of the membership function is a subset of the nonnegative real numbers whose supremum is finite.

2.1.1 Fuzzy Set Operators and Fuzzy Logic

For crisp sets, the basic operations are, namely,

- Union, OR
- Intersection, AND
- Complement, NOT

As an analogy, for fuzzy sets fuzzy operators are defined which allow to manipulate the fuzzy sets. Similarly it has fuzzy complements, intersection and union operators but they are not uniquely defined i.e. as membership functions, they are also context – dependent.

However an important dissimilarity exists there between traditional set/logic and fuzzy set theory. Traditionally there is a distinction between a union operation of sets and OR of logic as is the case with intersection and AND also. But in fuzzy theory there is no such distinction between the logical and set operators i.e.

Fuzzy union ≡ Fuzzy OR
Fuzzy intersection ≡ Fuzzy AND
Fuzzy complement ≡ Fuzzy NOT

Some standard fuzzy operations are:

- Fuzzy Complement,
  $\sim A(x) = 1 - A(x)$
- Fuzzy Union,
  $(A \cup B)(x) = \max[A(x), B(x)]$
- Fuzzy Intersection,
  $(A \cap B)(x) = \min[A(x), B(x)]$

2.2 Fuzzy Databases

The data stored in the database is normally crisp in nature. But the request for the required information may be of fuzzy in nature. The fuzziness may be classified into two categories viz. Impreciseness and Vagueness[1]. In the real time situation, people express their ideas using the natural languages. Normally natural language has a lot of vagueness and ambiguity. However, while applying one's thoughts as a query in terms of natural languages into the database, a lot of problems are experienced due to the inefficiency of RDBMS to handle such queries. Consider the query "Give me the names of the young age and high salary employees". This query cannot be processed directly by the SQL, since it contains a lot of vagueness like "Young" and "High". The best remedy for modeling the above situation is by the use of Fuzzy Sets.

2.2.1 Linguistic Variables and Hedges

Natural language consists of fundamental terms called "atomic terms"[1]. Examples of some atomic terms are "medium", "young" and "beautiful", etc. Collection of atomic terms are called composite terms. Examples of composite terms are "Very slow car", "Slightly Young student", "fairly beautiful lady", etc. The atomic terms are called linguistic variable[1] in Fuzzy set theory. A linguistic variable differs from a numerical variable in that; its values are not numbers but words or sentences in Natural languages. The purpose of using the linguistic Variable is to provide a means of approximate characterization of phenomena that is not defined properly. Linguistic variables can be characterized by the use of trapezoidal shaped possibility distribution. In linguistics, fundamental atomic terms are often modified with adjectives (noun) or adverbs (Verbs) like very, low, slightly, more-or-less, fairly, almost, roughly, etc. These modifiers are called linguistic hedges[1]. When a fuzzy set is used for interpretation, the linguistic hedges have the effect of modifying the membership function for a basic atomic term.

2.2.2 The FSQL Language

The FSQL language is an authentic extension of SQL language to model fuzzy queries [2]. It means that all the valid statements in SQL are also valid in FSQL. The SELECT command is extended to express flexible queries and due to its complex format, we only show an abstract with the main extensions.

Example: "*Give me the names of the young age and high salary employees*". This query is modelled in FSQL language as follows:
SELECT Emp_name, Emp_age, Emp_salary FROM Employee WHERE (Emp_age between 22 and 30) and (Emp_salary>8)

The FSQL server uses Fuzzy Meta information to model the different types of fuzzy attributes.

The attribute *age*, presented in Fig. 1, has the linguistic labels Young, Adult and Old, defined on the trapezoidal possibility distributions as following: YOUNG (18, 22, 30, 35), ADULT (25, 32, 45, 50), OLD (50, 55, 62, 70). An approximate value has a margin of 5.

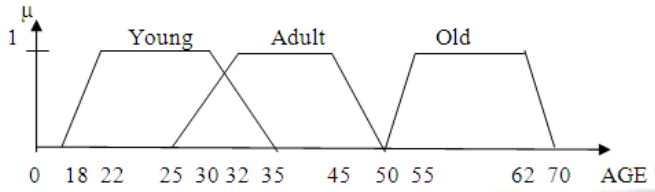


figure 1: Definition of Age attribute

The attribute SALARY, presented in Fig. 2, has the linguistic labels Low, Medium and High, defined on the trapezoidal possibility distributions as following: LOW (1, 1.5, 2.5, 4), MEDIUM (3, 4.5, 6.5, 8), HIGH (6, 8, 10, 12). An approximate value has a margin of 1.5 L.

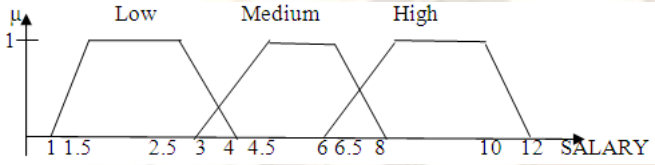


figure 2: Definition of Salary attribute

2.3 Clustering

Clustering [2],[4],[5],[8],[9] is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar to each other than to those in other clusters. Clustering is a main task of explorative data mining, and a common technique for statistical data analysis used in many fields, including pattern recognition, image analysis, information retrieval and bioinformatics.

2.3.1 Cluster Analysis

The objective of cluster analysis [5],[8],[9] is the classification of objects according to similarities among them and organizing of data into groups. Clustering techniques are among the *unsupervised* methods, they do not use prior class identifiers. The main potential of clustering is to detect the underlying structure in data, not only for classification and pattern recognition, but for model reduction and optimization.

2.3.2 The Data

Clustering techniques can be applied to data that is quantitative (numerical), qualitative (categorical), or a mixture of both. In this work, the clustering of quantitative data is considered. The data are typically observations of some physical process. Each observation consists of n measured variables, grouped into an n-dimensional row vector $x_k = [x_{k1}, x_{k2}, \ldots, x_{kn}]^T$, $x_k \in R^n$. A set of N observations is denoted by $X = \{x_k | k = 1, 2, \ldots, N\}$, and is represented as an $N \times n$ matrix:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Nn} \end{bmatrix}$$

In pattern recognition terminology, the rows of X are called *patterns* or objects, the columns are called the features or attributes, and X is called the *data matrix*.

The meaning of the columns and rows of X with respect to reality depends on the context. In medical diagnosis, for instance, the rows of X may represent patients, and the columns are then symptoms, or laboratory measurements for the patients. When clustering is applied to the modeling and identification of dynamic systems, the rows of X contain samples of time signals, and the columns are, for instance, physical variables observed in the system (position, velocity, temperature, etc.). In order to represent the system's dynamics, past values of the variables are typically included in X as well.

2.3.3 Cluster Partition

Since clusters can formally be seen as subsets of the data set, one possible classification of clustering methods can be according to whether the subsets are fuzzy or crisp (hard). Hard clustering methods are based on classical set theory, and require that an object either does or does not belong to a cluster. Hard clustering in a data set X means that partitioning the data into a specified number of mutually exclusive subsets of X. The number of subsets (clusters) is denoted by c. Fuzzy clustering methods allow objects to belong to several clusters simultaneously, with different degrees of membership. The data set X is thus partitioned into c fuzzy subsets. In many real situations, fuzzy clustering is more natural than hard clustering, as objects on the boundaries between several classes are not forced to fully belong to one of the classes, but rather are assigned membership degrees between 0 and 1 indicating their partial memberships. The discrete nature of hard partitioning also causes analytical and algorithmic intractability of algorithms based on analytic function, since these functions are not differentiable.
The structure of the partition matrix $U = [\mu_{ik}]$:

$$U = \begin{bmatrix} \mu_{1,1} & \mu_{1,2} & \cdots & \mu_{1,c} \\ \mu_{2,1} & \mu_{2,2} & \cdots & \mu_{2,c} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{N,1} & \mu_{N,2} & \cdots & \mu_{N,c} \end{bmatrix}.$$

2.3.4 Fuzzy C-Means Algorithm [11]

The Fuzzy C-Means clustering algorithm is based on the minimization of an objective function called *C-means functional*. It is defined by Dunn as:

$$J(\mathbf{X};\mathrm{U},\mathbf{V}) = \sum_{i=1}^{c}\sum_{k=1}^{N}(\mu_{ik})^m \|\mathbf{x}_k - \mathbf{v}_i\|_A^2 \qquad (1)$$

Where

$$\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_c],\ \mathbf{v}_i \in \mathbf{R}^n \qquad (2)$$

is a vector of cluster prototypes (centres), which have to be determined, and

$$D_{ikA}^2 = \|\mathbf{x}_k - \mathbf{v}_i\|_A^2 = (\mathbf{x}_k - \mathbf{v}_i)^T A(\mathbf{x}_k - \mathbf{v}_i) \qquad (3)$$

is a squared inner-product distance norm.

Statistically, equation (1) can be seen as a measure of the total variance of x_k from v_i. The minimization of the C-Means functional equation (1) represents a nonlinear optimization problem that can be solved by using a variety of available methods, ranging from grouped coordinate minimization, over simulated annealing to genetic algorithms.

$D^2_{ikA} > 0, \forall i, k$ and $m > 1$, then $(U, V) \in M_{fc} \times R$ may minimize equation (1) only if

$$\mu_{ik} = \frac{1}{\sum_{j=1}^{c}(D_{ikA}/D_{jkA})^{2/(m-1)}}, \quad 1 \le i \le c,\ 1 \le k \le N, \qquad (4)$$

$$\mathbf{v}_i = \frac{\sum_{k=1}^{N}\mu_{ik}^m \mathbf{x}_k}{\sum_{k=1}^{N}\mu_{i,k}^m}, \quad 1 \le i \le c. \qquad (5)$$

Note that equation (5) gives $\mathbf{v}_i$ as the weighted mean of the data items that belong to a cluster, where the weights are the membership degrees. That is why the algorithm is called "C-Means". One can see that the FCM algorithm is a simple iteration through equation (4) and (5).

The FCM algorithm computes with the standard Euclidean distance norm, which induces hyperspherical clusters. Hence it can only detect clusters with the same shape and orientation, because the common choice of norm inducing matrix is: A = I or it can be chosen as an $n \times n$ diagonal matrix that accounts for different variances in the directions in the directions of the coordinate axes of X:

$$\mathrm{A_D} = \begin{bmatrix} (1/\sigma_1)^2 & 0 & \cdots & 0 \\ 0 & (1/\sigma_2)^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & (1/\sigma_n)^2 \end{bmatrix},$$

or A can be defined as the inverse of the $n \times n$ covariance matrix: $A = \mathrm{F}^{-1}$, with

$$\mathrm{F} = \frac{1}{N}\sum_{k=1}^{N}(\mathbf{x}_k - \overline{\mathbf{x}})(\mathbf{x}_k - \overline{\mathbf{x}})^T. \qquad (6)$$

Here x denotes the sample mean of the data. In this case, A induces the Mahalanobis norm on R^n.

2.3.5 The Gustafson-Kessel Algorithm

Gustafson and Kessel [6],[8],[9] extended the standard Fuzzy C-Means algorithm by employing an adaptive distance norm, in order to detect clusters of different geometrical shapes in one data set. Each cluster has its own norm-inducing matrix A_i, which yields the following inner-product norm:

$$D_{ikA}^2 = (\mathbf{x}_k - \mathbf{v}_i)^T A_i(\mathbf{x}_k - \mathbf{v}_i), \quad 1 \le i \le c, \quad 1 \le k \le N. \qquad (7)$$

The matrices A_i are used as optimization variables in the C-Means functional, thus allowing each cluster to adapt the distance norm to the local topological structure of the data. Let A denote a c-tuple of the norm-inducing matrices: A = (A_1, A_2, ..., A_c). The objective functional [6],[8],[9] of the GK algorithm is defined by:

$$J(\mathbf{X};\mathrm{U},\mathbf{V},\mathbf{A}) = \sum_{i=1}^{c}\sum_{k=1}^{N}(\mu_{ik})^m D_{ikA_i}^2. \qquad (8)$$

The objective function equation (8) cannot be directly minimized with respect to A_i, since it is linear in A_i. This means that J can be made as small as desired by simply making A_i less positive definite. To obtain a feasible solution, A_i must be constrained in some way. The usual way of accomplishing this is to constrain the determinant of A_i. Allowing the matrix A_i to vary with its determinant fixed corresponds to optimizing the cluster's shape while its volume remains constant:

$$\|A_i\| = \rho_i, \quad \rho > 0, \qquad (9)$$

where ρ_i is fixed for each cluster. Using the Lagrange multiplier method, the following expression for A_i is obtained:

$$A_i = [\rho_i det(\boldsymbol{F}_i)]^{1/n}\boldsymbol{F}_i^{-1}, \qquad (10)$$

where F_i is the fuzzy *covariance matrix* of the ith cluster defined by:

$$\boldsymbol{F}_i = \frac{\sum_{k=1}^{N}(\mu_{ik})^m(\mathbf{x}_k - \mathbf{v}_i)(\mathbf{x}_k - \mathbf{v}_i)^T}{\sum_{k=1}^{N}(\mu_{ik})^m}. \qquad (11)$$

Note that the substitution of equation (10) and (11) into equation (7) gives a generalized squared Mahalanobis distance norm between x_k and the cluster mean v_i, where the covariance is weighted by the membership degrees in U.

2.4 Validation of Clusters

Different scalar validity measures have been proposed in the literature, none of them are perfect by themselves and therefore three indexes are used here, which are described below:

1. **Partition Coefficient (PC):** PC [11] measures the amount of "overlapping" between clusters. It is defined by Bezdek as follows:

$$PC(c) = \frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{N}(\mu_{ij})^2 \qquad (12)$$

Where μ_{ij} is the membership of data point j in cluster i. The disadvantage of PC is lack of direct connection to some property of the data themselves. The optimal number of cluster is at the maximum value.

2. **Classification Entropy (CE):** CE [11] measures the fuzziness of the cluster partition only, which is similar to the Partition Coefficient.

$$CE(c) = -\frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{N}\mu_{ij}log(\mu_{ij}), \qquad (13)$$

3. **Xie and Beni's Index (XB):** XB [11] aims to quantify the ratio of the total variation within clusters and the separation of clusters.

$$XB(c) = \frac{\sum_{i=1}^{c}\sum_{j=1}^{N}(\mu_{ij})^m||x_j - v_i||^2}{Nmin_{i,j}||x_j - v_i||^2} \qquad (14)$$

The optimal number of clusters should minimize the value of the index.

3. Proposed Methodology

An application layer is placed over the SQL and it will perform the necessary translation by acting as a middleware. It is assumed that the underlying database will be crisp. Therefore the fuzziness is incorporated in the front end only. At the front end, initially the Fuzzy sets / Linguistic Variables on the necessary domains are defined. For example, the fuzzy sets Young, Adult and Old are defined on the attribute AGE.

After understanding the user's query, it is converted into the SQL format and gets the relevant result and applies the clustering algorithms to find the interesting patterns and groupings in the given result set.

Fig. 3 shows the model used for incorporating Extended Fuzzy C-Means in Fuzzy Databases. It represents an integrated set of components that enables the transformation of fuzzy query and extraction of data from the database. This model is useful for the naive user for retrieving relevant results of non-crisp queries and further for analyzing the relevance of results provided by fuzzy clustering algorithms Fuzzy C-Means (FCM) and Gustafson-Kessel algorithm (GK). For analyzing the performance of both the clustering algorithms FCM and GK the partition coefficient (PC), Classification Entropy (CE) and Xie Beni's Index (XB) clustering validity measures are used.

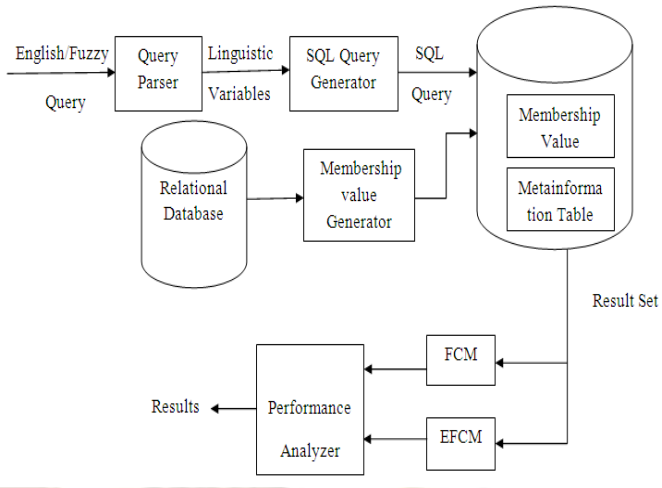


figure 3: Proposed Model Overview

4. Implementation

The work is developed in MATLAB version R2009b with the use of FUZZY tool. The experiment is run on Windows XP. In this section some of the screenshots are given from the software. The naïve user is uncertain with his requirements. So, he selects his requirements in the form of Age and Salary. Experiments are done on some particular english queries which are normally used by the user.

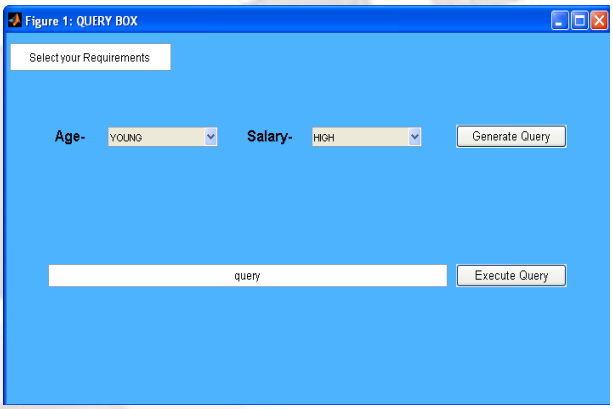


figure 4: User interface for choosing requirements

Firstly, user chooses his requirement like age is YOUNG and salary is HIGH then he clicks on the query generation button as shown in the Fig. 4.

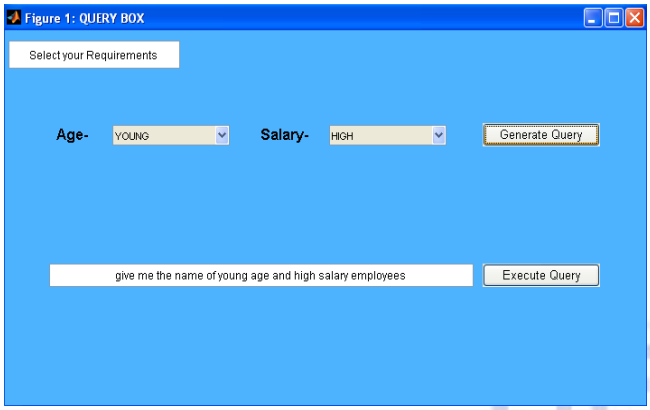


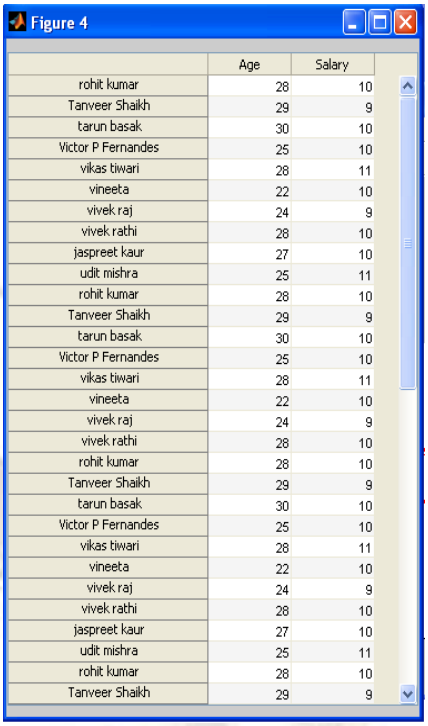
figure 5: Fuzzy query generation

As the user clicks on the Generate Query button, fuzzy query "give me the name of young age and high salary employees" (Q) generates as shown in the Fig. 5. After that, user clicks on the Execute Query button.

After clicking on Execute Query button some intermediate stages have to be processed before getting the result. Fuzzy Query or English query is checked that any fuzzy attributes and linguistic variables are present in this query or not. If yes, then the calculation of membership value is performed according to the linguistic variables. If linguistic Hedges are also present in the Fuzzy query then membership value is updated with the manipulated membership value.

With the help of Meta Information table fuzzification and defuzzification are done and all the fuzzy attributes are replaced with the particular range. After that Result Data Set is generated for the query "give me the name of young age and high salary employees" (Q) as shown in the Table 1 which consists of 60 records.

Table 1: Result Data set for query Q



	Age	Salary
rohit kumar	28	10
Tanveer Shaikh	29	9
tarun basak	30	10
Victor P Fernandes	25	10
vikas tiwari	28	11
vineeta	22	10
vivek raj	24	9
vivek rathi	28	10
jaspreet kaur	27	10
udit mishra	25	11
rohit kumar	28	10
Tanveer Shaikh	29	9
tarun basak	30	10
Victor P Fernandes	25	10
vikas tiwari	28	11
vineeta	22	10
vivek raj	24	9
vivek rathi	28	10
rohit kumar	28	10
Tanveer Shaikh	29	9
tarun basak	30	10
Victor P Fernandes	25	10
vikas tiwari	28	11
vineeta	22	10
vivek raj	24	9
vivek rathi	28	10
jaspreet kaur	27	10
udit mishra	25	11
rohit kumar	28	10
Tanveer Shaikh	29	9

After getting the result Data set, for finding the similarities in the data, the clusters are made of the result data set. Therefore, the well known Fuzzy C-Means (FCM) and Gustafson-Kessel (GK) fuzzy clustering algorithms are applied separately on the same result set and the clusters are made as shown in the Fig. 6 and Fig. 7 in the form of contour graph.

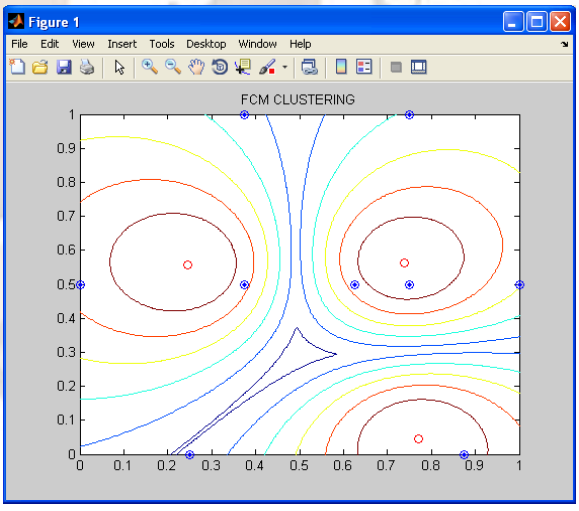


figure 6: Clusters after applying FCM for query Q

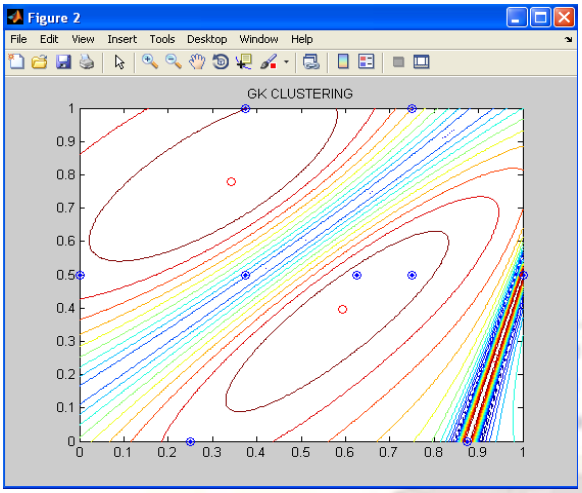


figure 7: Clusters after applying GK for query Q

In Fig. 6 and Fig. 7 the '.' remarks are the data points and the 'o' are the cluster centers, which are the weighted mean of the data. The FCM algorithm can only detect clusters with circle shape, that is why it cannot really discover the orientation and shape of the cluster. Gustafson-Kessel algorithm is an extension of the Fuzzy C-means algorithm (uses adaptive distance norm), it detects the elongated clusters. The orientation and shape can be "mined" from the eigen structure of the covariance matrix: the direction of the axes are given by the eigenvectors. In Fig. 7 the contour-map shows the superposition of the three ellipsoidal clusters for Gustafson-Kessel algorithm.

5. Performance Analysis

First of all it must be mentioned, that these algorithms use random initialization, so different running issue in different partition results, i.e. values of the validation measures. On the other hand the results hardly depend on the structure of the data, and no validity index is perfect by itself for a clustering problem. Several experiments and evaluations are needed that are not the proposition of this work.

The only difference between Fig. 6 and Fig. 7 stands in the shape of the clusters, while the Gustafson-Kessel algorithm can find the elongated clusters better because the partition coefficient (PC) which is membership based measure. The partition coefficient aims to measure the degree of fuzziness of the clusters. The rationale is that the fuzzier the clusters are, the worse the partition is. Subsequently, another membership based validity measures is classification entropy (CE). The entropy measure increases as the fuzziness of partition increases. Therefore, a cluster with higher partition coefficient (PC) and lower classification entropy (CE) is preferred. Xie and Beni introduced a validity measures that consider both the compactness of clusters as well as the separation between the clusters. Intuitively, the more compact the clusters are and the further the separation between clusters, the more desirable a partition. So, the lower the Xie and Beni's cluster index the better the soft partition is.

After running some fuzzy queries the experimental results of partition coefficient (PC), Classification Entropy (CE) and Xie and Beni's Index (XB) for Fuzzy C-means clustering algorithm and Gustafson-Kessel (GK) algorithm are shown in Table 2.

Table 2: The numerical values of validity measures

Query	Partition Coefficient (PC)		Classification Entropy (CE)		Xie and Beni Index (XB)	
	FCM	GK	FCM	GK	FCM	GK
give me the name of young age and high salary employees	0.67040	0.86760	0.59320	0.22730	1.28960	0.14740
give me the name of adult age and medium salary employees	0.65860	0.70570	0.61550	0.53680	1.24460	0.80000
give me the name of old age and low salary employees	0.84510	1.00000	0.31430	0.00000	1.35960	0.26800
give me the name of young age and medium salary employees	0.72580	0.80170	0.50540	0.36470	1.97270	0.35110
give me the name of adult age and high salary employees	0.94870	1.00000	0.11520	0.00000	1.28250	0.12500
give me the name of old age and high salary employees	0.74470	0.83910	0.46800	0.31230	1.24830	0.17290
give me the name of adult age and low salary employees	0.86940	0.91540	0.26770	0.19060	1.19830	0.19730
give me the name of old age and medium salary employees	0.65360	0.70400	0.62790	0.52610	1.64780	0.30630

On the score of the values of these "most popular and used" indexes for fuzzy clustering the GK clustering has the very best results for this data set.

The results of the FCM and GK algorithms have been evaluated by Partition Coefficient (PC), Classification Entropy (CE) and Xie and Beni's Index (XB).

6. Conclusion

The proposed model has been successfully implemented and the translation of fuzzy query into SQL in relational databases has been carried out. The Fuzzy C-Means and Gustafson-Kessel fuzzy clustering algorithms have been successfully implemented.

One of the requirements in clustering is the handling of arbitrary shaped clusters and there are some efforts in this context. However, there is no well-established method to describe the structure of arbitrary shaped clusters as defined by an algorithm. The main problem of Fuzzy C-Means clustering algorithm is that all the clusters should have spherical shape only solved by Gustafson-kessel algorithm. Result analysis represents that Gustafson-Kessel clustering algorithm is more accommodating for the employee data set when compared to Fuzzy C-Means.

Future scope of this work is to another two fuzzy clustering methods can be applied for comparison on the same fields or domains that have been provided in this work.

References

[1] V.Balamurugan and K.Senthamarai Kannan, A Framework for Computing Linguistic Hedges in Fuzzy Queries, *IEEE The International Journal of Database Management System (IJDMS), Vol. 2, No. 1,* 2010,1-7.

[2] Amel Grissa Touzi, An Alternative Extension of the FCM Algorithm for Clustering Fuzzy Databases, *IEEE*

Second international conference on Advances in Databases, Knowledge and Data Applications, 2010,135-142.

[3] Shawki A. Al-Dubaee and Nesar Ahmad, Search Result Clustering Using Fuzzy C-Mean and Gustafson Kessel Algorithms: A Comparative Study, *IEEE 2010 First International Conference on Integrated Intelligent Computing*, 2010.

[4] Jiawei Han and Micheline Kamber, *Data Mining: Concepts and Techniques* (Second edition, Morgan Kaufmann 2006).

[5] John Yen, Reza Langari, *Fuzzy Logic Intelligence, Control and Information* (Pearson Education 2003).

[6] Marie-Jeanne Lesot and Rudolf Kruse, Gustafson-Kessel-like clustering algorithm based on typicality degrees, *international conference on Information Processing and Management of Uncertainty, Paris, France*, 2006.

[7] Yauheri Veryha, Implementation of fuzzy classification in rational databases using conventional SQL querying, *Science direct, information and Software Technology Vol. 47*, Issue 5, 2005, 357-364.

[8] Young-Il Kim, Dae-Won Kim, Doheon Lee, Kwang H. Lee, 2004. A cluster validation index for GK cluster analysis based on relative degree of sharing, *Elsevier, Information sciences 168*, 2004, 225–242.

[9] Maria Halkidi, Yannis Batistakis and Michalis Vazirgiannis, 2001. On Clustering Validation Techniques, *Springer, Journal of intelligent information, VOL 17; PART 2/3*, 2001, 107-146.

[10] Miroslav Hudec, 2009. An Approach to Fuzzy Database Querying, Analysis and Realisation, *ComSIS Vol. 6*, No. 2, 2009, 127-140.

[11] Balazs Balasko, Janos Abonyi and Balazs Feil, Fuzzy Clustering and Data Analysis Toolbox, *Department of Process Engineering University of Veszprem, Hungary.* 2005 [Online] Available at: http://www.fmt.vein.hu/softcomp