International Journal of Computer Applications (0975 - 8887)
Volume 167 - No.14, June 2017

Type-2 Projected Gustafson-Kessel Clustering Algorithm

Charu Puri
Department of Computer Science
University of Delhi

Naveen Kumar
Department of Computer Science
University of Delhi

ABSTRACT

We propose a type-2 based clustering algorithm to capture data points and attributes relationship embedded in fuzzy subspaces. It is a modification of Gustafson Kessel clustering algorithm through deployment of type-2 fuzzy sets for high dimensional data. The experimental results have shown that type-2 projected GK algorithm perform considerably better than the comparative techniques.

General Terms: Data Mining, Fuzzy sets.

Keywords: Type-2, Subspace Clustering, Gustafson Kessel.

1. INTRODUCTION

Clustering aims at grouping data objects into classes so that the objects within a class are similar while the objects in different classes are dissimilar. Conventional clustering algorithms compute the distances between objects in the entire space of dimensions. However, as the number of dimensions increases, the data objects become sparse. Indeed any two points may become nearly equidistant. In such scenarios, clusters are often hidden in specific subspaces of the original feature space rather than in the original feature space. To cope with the problem of high dimensional feature spaces, feature reduction and feature selection techniques have hitherto been used in literature. Feature reduction techniques such as principal component analysis (PCA) suffer from usability problem as it becomes hard to interpret the results intuitively. Feature selection techniques project the whole feature space to a lower dimensional subspace so that cluster structures become apparent. However, these techniques do not deal effectively with clusters in varying subspaces. Hence, there is a need for more generalized techniques that can be used to obtain meaningful clusters in varying subspaces. Subspace clustering finds clusters on the subsets of dimensions of a data set. However, different dimensions may be relevant to different clusters to varying degree. A refinement of subspace clustering called soft subspace clustering attempts to cluster data objects in the entire data space with continuous feature weighting. Even though, traditional fuzzy logic has numerous applications it lacks in modeling high levels of uncertainty because the memberships are imprecise in nature [17][27]. In order to enhance the modeling capability of high level of imprecision, Zadeh [22]extended fuzzy sets to type-2 fuzzy sets by fuzzifying membership in type 1 fuzzy sets. The characteristic feature of type-2 fuzzy sets is that it uses the notion that type-1 fuzzy sets can be thought of as first order approximation to uncertainty and, therefore type-2 fuzzy sets provide a second order approximation. The membership function of type-2 fuzzy sets model the imprecise nature of fuzzy membership grades. The two main categories of type-2 fuzzy sets are interval and generalized type-2 fuzzy sets. Type-2 interval fuzzy sets have a secondary membership function as a crisp interval in [0, 1] and generalized type-2 fuzzy sets have secondary membership function as a fuzzy number between zero and one. Due to the addition of third dimension in the concept of type-2 fuzzy sets, literature is largely focused on interval type-2 fuzzy logic as the various techniques have been proposed to reduce the computational complexity of the logical operators. Theoretical work describing the terminology, a new representation theorem and new derivation of union, intersection and complement of type-2 fuzzy sets is introduced in [19]. Recently, Mendel and John[17] proposed a new representation theorem that could be used in union, intersection and complement for type-2 fuzzy sets without using the extension principle. Type-2 fuzzy sets have found applications in those areas where it is difficult to determine an exact membership function for a fuzzy set like linguistic uncertainties. Type-2 fuzzy sets have been applied in transport scheduling, forecasting of time series, signal processing, pattern recognition, decision making, speech recognition[7],[13] etc. Mizumoto and Tanaka[24],[25] and Duboid and Prade[6] explored the logical operation on type-2 fuzzy sets. John[27] stated that " Type-2 fuzzy sets allow for linguistic grades of membership, thus assisting in knowledge representation, they also offer improvement on inferencing with type-1 fuzzy sets". Mendel[17] stated that " type-2 fuzzy sets provide more degrees of freedom, so using type-2 fuzzy sets has the potential to outperform type-1 fuzzy sets. In [11] Klir and Floger, explained that representation of fuzziness by using membership grades that are themselves precise real numbers is problematic hence, the concept of Type-1 fuzzy sets should be extended to type-2 fuzzy sets and its higher degrees. We propose a new objective function for type-2 projected clustering which automatically detects the relevant cluster dimensions. Experimental results indicate that it enhances the efficiency of clustering solution by simultaneously pruning away the irrelevant subspaces.

2. PROBLEM FORMULATION

In this section, we introduce all necessary notations. Given a data set $X = \{x_1, x_2, ..., x_n\}$ in the d-dimensional space the objective is to partition the data set X into k cluster prototypes $Z = \{z_1, z_2, ..., z_k\}$ based on identification of subspaces. The main idea is to impose weights on the dimensions corresponding to each cluster. Fuzzy partition of the data set X can be represented

by a $k \times n$ matrix $U = [\mu_{ij}]$, where μ_{ij} denotes the degree of membership with which j^{th} pattern belongs to the i^{th} cluster, for $1 \leq j \leq n$, $1 \leq i \leq k$. The matrix U is called the fuzzy partition matrix which satisfies the following conditions:

$$\mu_{ij} \in [0,1],\ 1 \leq j \leq n,\ 1 \leq i \leq k, \tag{1}$$

$$\sum_{i=1}^{k} \mu_{ij} = 1,\ 1 \leq j \leq n, \tag{2}$$

The above constraint express the fact that the sum of memberships of pattern over the set of clusters must be equal to 1. The fact that there are at least two number of clusters is expressed by the following constraint:

$$0 < \sum_{j=1}^{n} \mu_{ij} < n,\ 1 \leq i \leq k. \tag{3}$$

The fuzzy partition space for (X, k), is the set:

$M^{fk} = \{U \in \Re^{k \times n} | \mu_{ij} \in [0,1]\,,\ \forall i,j;\ \sum_{i=1}^{k} \mu_{ij} = 1, \forall j; 0 < \sum_{j=1}^{n} \mu_{ij} < n, \forall i\}$

The fuzzy c-means objective function is formulated as:

$J_m = \sum_{j=1}^{n} \sum_{i=1}^{k} \mu_{ij}^m d^2\,(x_j, z_i)$

The coefficient $m \in (1, \infty)$ is a fuzzification parameter. However this objective function is constrained to find the clusters in the the entire feature space and therefore cannot determine the respective natural subspaces of each cluster in high dimensional data set.
Now, we associate with each cluster a weight vector in order to capture the subspace information of each cluster. Let $W = [\omega_{ir}]$ be a $k \times d$ matrix expressing the memberships of each protoype along different dimensions. In this matrix ω_{ir} denotes the contribution of i^{th} cluster to the r^{th} dimension. The sum of contributions from all dimensions adds to 1 for any cluster. This expressed by the constraint, where

$$\sum_{r=1}^{d} \omega_{ir} = 1,\ 1 \leq i \leq k, \tag{4}$$

$$\omega_{ir} \in [0,1]\,,\ 1 \leq i \leq k,\ 1 \leq r \leq d, \tag{5}$$

Also as there are at least two dimensions, we get the constraint: $0 < \sum_{i=1}^{k} \omega_{ir} < k,\ \forall r$
Thus, fuzzy partitioning subspace for (X,d) is the set

$M^{fd} = \{W \in \Re^{k \times d} | \omega_{ir} \in [0,1] \forall i,r;\ \sum_{r=1}^{d} \omega_{ir} = 1, \forall i; 0 < \sum_{i=1}^{k} \omega_{ir} < k, \forall r\}$

The objective function J_m of GK algorithm is defined as follows[14]:

$$J_m = \sum_{j=1}^{n} \sum_{i=1}^{k} \mu_{ij}^m d_{ij}^2, \tag{6}$$

where

$$d_{ij}^2 = (x_j - z_i) A_i (x_j - z_i)^T \tag{7}$$

where, fuzzy partitioning subspace for (X, k) and (X, d) together forms the new partition space for high dimensional data set X.

Parameters $\alpha \in (1, \infty)\,,\ \beta \in (1, \infty)$ are weighting components. These parameters control the fuzzification of $\mu_{ij}(\omega_{ir})$. Larger the value of $\alpha(\beta)$ the more equal the distribution of μ_{ij} and ω_{ir} giving each pattern an equal chance to impact all clusters and dimensions. Value of $\alpha(\beta)$ closer to 1 indicates good clustering behaviour as $\mu_{ij}(\omega_{ir})$ assigns higher values to clusters(subspaces. A_i is a symmetric, positive definite matrix that induces for each cluster a norm of its own[9][14]. In order to avoid singularity problem, A_i is constrained in such a way that $det(A_i) = \rho_i > 0$, ρ_i being fixed for each i permitting different sizes of cluster. The exponent $m \in (1, \infty)$ is a fuzzification parameter, that controls the extent by which clusters may overlap. The objective function J_m is minimized using an alternating optimization (AO) technique. The AO optimization technique leads to the local optimum as it proceeds by fixing a set of parameters and optimizing the rest of parameters in an alternating manner. Iteratively updating in such a fashion yields the optimum value of J_m. However, such techniques do not ensure the global optimum and the algorithm may get stuck in the local optimum. In order to counter the possibility of getting stuck at local optimum, one often performs several runs of the algorithm.
The minimization of this objective function is carried out with respect to $\mu_{ij},\ \omega_{ir},\ z_{ir}$.
The final update equations are given below:

$$\mu_{ij} = 1 / \sum_{l=1}^{k} \left[\frac{\sum_{r=1}^{d} \omega_{ir}^{\beta} d_{ijr}^2}{\sum_{r=1}^{d} \omega_{lr}^{\beta} d_{ljr}^2} \right]^{1/(\alpha-1)} \tag{8}$$

$$\omega_{ir} = 1 / \sum_{l'=1}^{d} \left[\frac{\sum_{j=1}^{n} \mu_{ij}^{\alpha} d_{ijr}^2}{\sum_{j=1}^{n} \mu_{ij}^{\alpha} d_{ijl'}^2} \right]^{1/(\beta-1)} \tag{9}$$

$$z_{ir} = \sum_{j=1}^{n} \omega_{ir}^{\beta} \mu_{ij}^{\alpha} x_{jr} / \sum_{j=1}^{n} \omega_{ir}^{\beta} \mu_{ij}^{\alpha} \tag{10}$$

$$A_i = ((det(F_i)\rho_i))^{1/d} F_i^{-1} \tag{11}$$

2.1 Type-2 Projected GK

In this section we present the proposed a new objective function based type-2 projected clustering which automatically detects the relevant cluster dimensions.
Membership functions of type-1 fuzzy sets are two dimensional. Type-2 fuzzy sets can be considered as fuzzification of membership in type-1 fuzzy set. We give below the formal definition of type-2 fuzzy set. A type-2 fuzzy set is characterized by a type-2 membership function $\mu_{\tilde{A}}\,(x, u)$ where $x \in X$ and $u \in J_x \subseteq [0,1]$:
$\tilde{A} = \{\ (\,(x,u),\ \mu_{\tilde{A}}(x,u)\,) \mid \forall x \in X, \forall u \in J_x \subseteq [0,1]\ \}$

where, $0 \leq \mu_{\tilde{A}}(x,u) \leq 1$. J_x is known as the primary memberships. So, a type-2 fuzzy set has membership grades that are type-1 fuzzy sets, which are referred to as *secondary membership functions*. At each value x , $\mu_{\tilde{A}}(x,u)$ is a secondary membership function of $\tilde{A}$. Let us consider taking a picture of an object with a camera. For simplicity, we consider grey scale picture. The gray scale image may be thought of type-1 fuzzy sets. However, whenever we take picture using a camera, there is an element of uncertainty introduced by environmental conditions apart from skills of photography. This introduces another level of uncertainty, in the imaging process. Thus, for a given pixel, gray scale value may be thought of as distributed (possibly normally) in a range about the measured gray scale value.
Rhee and Hwang[8] extended the conventional FCM to type-2 FCM. They have argued that when the prototype is computed, it may be uncertain whether each pattern properly contributes in updating the location of prototype. Patterns with high membership contribute more in prototype determination as they are considered to have less uncertainty as opposed to patterns with low membership. Also, memberships generated are based on relative distance, as they are relative numbers which lacked typicality. Since, type-2 fuzzy sets have more degrees of freedom therefore, they have the potential to outperform type-1 fuzzy sets.
They have extended FCM by designing type-2 membership functions. Type-2 membership functions are assigned as the base length of each triangular function as 1 minus the corresponding type-1 membership value and by taking the difference of each type-2 membership function triangular area with the corresponding type-1 membership grade. The type-2 membership grades of i^{th} pattern corresponding to j^{th} cluster can be obtained by the equation:

$$a_{ij} = \mu_{ij} - (1 - \mu_{ij})/2 \tag{12}$$

We have extended it to determine the membership grades of i^{th} cluster corresponding to r^{th} dimension which can be obtained by the equation:

$$b_{ir} = \omega_{ir} - (1 - \omega_{ir})/2 \tag{13}$$

where $a_{ij}(b_{ir})$ are type-2 membership grades for type-1 membership grades $\mu_{ij}(\omega_{ir})$.
Substituting $-a_{ij}(-b_{ir})$ for $\mu_{ij}(\omega_{ir})$ in the objective function, updated equations have been computed.

Since, $\sum_{i=1}^{k} \mu_{ij} = 1$,

$a_{ij} = (1 - 3\mu_{ij})/2$,

$\Rightarrow \sum_{i=1}^{k} a_{ij} = (k-3)/2$.

Also, $\sum_{r=1}^{d} \omega_{ir} = 1$,

$b_{ir} = (1 - 3\omega_{ir})/2$,

$\Rightarrow \sum_{r=1}^{d} b_{ir} = (d-3)/2$.

2.2 Type-2 Projected GK Clustering Algorithm

We present type-2 projected GK algorithm. Given the high dimensional data set X, choose the number of clusters $1 < k < n$, the weighting exponent $\alpha > 1$, $\beta > 1$, the termination tolerance $\in > 0$. Initialize the partition matrix U, W randomly.
Based on the update equations obtained above, we describe below the Type-2 Projected GK algorithm.

[H] Type-2 Projected GK Clustering Algorithm **Inputs**
n: size of data set X
k: number of clusters, $1 < k < n$
α: the weight exponent of matrix U, $\alpha > 1$
β: the weight exponent of matrix W, $\beta > 1$
ϵ: the termination tolerance, $\epsilon > 0$
A: the norm-inducing matrix
Outputs
U: membership matrix of objects in clusters
W: matrix indicating relevance of dimensions for clusters
Z: cluster centers
Initialize the partition matrices U, W randomly
t=0
$\|U^t - U^{t-1}\| > \in$

Step 1. Compute the cluster prototypes

$$z_{ir}^t = \frac{\sum_{j=1}^{n} (\omega_{ir}^{t-1})^{\beta} (\mu_{ij}^{t-1})^{\alpha} x_{jr}}{\sum_{j=1}^{n} (\omega_{ir}^{t-1})^{\beta} \mu_{ij}^{\alpha}}$$

Step 2. Compute the cluster covariance matrices

$F_i^t = \sum_{j=1}^{n} (\omega_{ir}^{t-1})^{\beta} (\mu_{ij}^{t-1})^{\alpha} (x_{jr} - z_{ir}^t)(x_{js} - z_{is}^t) 1 \leq r \leq d, 1 \leq s \leq d.$

Step 3. Compute the distances

$$d_{ijr}^2 = [(x_{j1} - z_{i1}^t) \ldots (x_{jd} - z_{id}^t)] A_i [(x_{j1} - z_{i1}^t) \ldots (x_{jd} - z_{id}^t)]^T$$

Step 4. Update the partition matrices

$$\mu_{ij}^h = 1/3 + (k/3 - 1) / \sum_{l=1}^{k} \left[\frac{\sum_{r=1}^{d} [(1 - 3\omega_{ir}^{h-1})/2]^{\beta} d_{ijr}^2}{\sum_{r=1}^{d} [(1 - 3\omega_{lr}^{h-1})/2]^{\beta} d_{ljr}^2} \right]^{1/(\alpha-1)}$$

$$\omega_{ir}^h = 1/3 + (d/3 - 1) / \sum_{t=1}^{d} \left[\frac{\sum_{j=1}^{n} [(1 - 3\mu_{ij}^{h-1})/2]^{\alpha} d_{ijr}^2}{\sum_{j=1}^{n} [(1 - 3\mu_{ij}^{h-1})/2]^{\alpha} d_{ijt}^2} \right]^{1/(\beta-1)}$$

Step 5. t = t+1

3. THEORETICAL ANALYSIS

In this section, we discuss the necessary condition for minimization of the proposed algorithm along with its proof. The necessary condition for minimization of the type-2 projected GK objective function yields the following update equations:

$$\mu_{ij} = 1/3 + (k/3 - 1) / \sum_{l=1}^{k} \left[\frac{\sum_{r=1}^{d} [(1 - 3\omega_{ir})/2]^{\beta} d_{ijr}^2}{\sum_{r=1}^{d} [(1 - 3\omega_{lr})/2]^{\beta} d_{ljr}^2} \right]^{1/(\alpha-1)} \tag{14}$$

$$\omega_{ir} = 1/3+(d/3-1)/\sum_{t=1}^{d}\left[\frac{\sum_{j=1}^{n}[(1-3\mu_{ij})/2]^{\alpha}d_{ijr}^2}{\sum_{j=1}^{n}[(1-3\mu_{ij})/2]^{\alpha}d_{ijt}^2}\right]^{1/(\beta-1)} \tag{15}$$

$$z_{ir} = \frac{\sum_{j=1}^{n}[(1-3\omega_{ir})/2]^{\beta}[(1-3\mu_{ij})/2]^{\alpha}x_{jr}}{\sum_{j=1}^{n}[(1-3\omega_{ir})/2]^{\beta}[(1-3\mu_{ij})/2]^{\alpha}} \tag{16}$$

PROOF. We have to minimize $J_{\alpha,\beta}$ with respect to U, W, subject to the respective constraints $\alpha \in (1,\infty)$, and $\beta \in (1,\infty)$. Then the constraints have been adjoined to $J_{\alpha,\beta}$ with a set of Lagrange multipliers $\{\lambda_j\}$ $1 \leq j \leq n$ and $\{\phi_i\}$ $1 \leq i \leq k$ to formulate:

$$J_{\alpha,\beta} = \sum_{j=1}^{n}\sum_{i=1}^{k}\sum_{r=1}^{d}[(1-3\mu_{ij})/2]^{\alpha}[(1-3\omega_{ir})/2]^{\beta}d_{ijr}^2$$
$$+\sum_{j=1}^{n}\lambda_j\left(\sum_{i=1}^{k}[(1-3\mu_{ij})/2]-(k-3)/2\right)$$
$$+\sum_{i=1}^{k}\phi_i\left(\sum_{r=1}^{d}[1-3\omega_{ir})/2]-(d-3)/2\right)$$

Now, we compute the first order derivative of $J_{\alpha,\beta}$ with respect to μ_{ij}, which is a necessary condition for optimality.

$$\frac{\partial J_{\alpha,\beta}}{\partial \mu_{ij}} = \alpha\sum_{r=1}^{d}[(1-3\mu_{ij})/2]^{\alpha-1}[(1-3\omega_{or})/2]^{\beta}d_{ijr}^2+\lambda_j = 0 \tag{17}$$

$$[(1-3\mu_{ij})/2]^{\alpha-1} = \frac{-\lambda_j}{\alpha\sum_{r=1}^{d}[(1-3\omega_{ir})/2]^{\beta}d_{ijr}^2+\lambda_j} = 0 \tag{18}$$

$$\mu_{ij} = 1/3+2/3\left[\frac{\lambda_j}{\alpha\sum_{r=1}^{d}[(1-3\omega_{ir})/2]^{\beta}d_{ijr}^2}\right]^{1/(\alpha-1)} \tag{19}$$

$$\sum_{i=1}^{k}\mu_{ij} = k/3+2/3\sum_{i=1}^{k}\left[\frac{\lambda_j}{\alpha\sum_{r=1}^{d}[(1-3\omega_{ir})/2]^{\beta}d_{ijr}^2}\right]^{1/(\alpha-1)} \tag{20}$$

Substituting the value of λ_j in 19 we obtain:

$$\mu_{ij} = 1/3+(k/3-1)/\sum_{l=1}^{k}\left[\frac{\sum_{r=1}^{d}[(1-3\omega_{ir})/2]^{\beta}d_{ijr}^2}{\sum_{r=1}^{d}[(1-3\omega_{lr})/2]^{\beta}d_{ijr}^2}\right]^{1/(\alpha-1)} \tag{21}$$

Now, we compute the first order derivative of J with respect to ω_{ir}, which is again a necessary condition for optimality.

Table 1. Data Sets

Data Sets	Instances	Attributes	Classes
Forest Fire	517	13	3
Alzheimr	45	8	3
Parkinson	197	23	2
Breast Cancer	569	32	2

Table 2. Accuracy

Data Sets	GKS	PROCLUS	GK
Forest Fire	0.8588	0.8696	0.8337
Alzheimr	0.7556	0.6667	0.6000
Parkinson	0.7538	0.7641	0.7538
Breast Cancer	0.8875	0.7907	0.8401

$$\frac{\partial J_{\alpha,\beta}}{\partial \omega_{ir}} = 2\sum_{j=1}^{n}[(1-3\mu_{ij})/2]^{\alpha}[(1-3\omega_{ir})/2]^{\beta-1}d_{ijr}^2+\phi_i = 0 \tag{22}$$

Computing in the similar fashion as above we obtain:

$$[(1-3\omega_{ir})/2]^{\beta-1} = \left[\frac{-\phi_i}{\beta\sum_{j=1}^{n}[(1-\mu_{ij})/2]^{\alpha}d_{ijr}^2}\right]^{\frac{1}{(\beta-1)}} \tag{23}$$

$$\sum_{r=1}^{d}\omega_{ir} = d/3+2/3\sum_{r=1}^{d}\left[\frac{-\phi_i}{\beta\sum_{j=1}^{n}[(1-\mu_{ij})/2]^{\alpha}d_{ijr}^2}\right]^{\frac{1}{(\beta-1)}} \tag{24}$$

Substituting the value of ϕ_i in 23 we obtain:

$$\omega_{ir} = 1/3+(d/3-1)/\sum_{t=1}^{d}\left[\frac{\sum_{j=1}^{n}[(1-3\mu_{ij})/2]^{\alpha}d_{ijr}^2}{\sum_{j=1}^{n}[(1-3\mu_{ij})/2]^{\alpha}d_{ijt}}\right]^{1/(\beta-1)} \tag{25}$$

To minimize $J_{\alpha,\beta}$ with respect to prototypes, we fix U and V. We obtain

$$\frac{\partial J_{\alpha,\beta}}{\partial z_{ir}} = 2\sum_{j=1}^{n}[(1-3\mu_{ij})/2]^{\alpha}[(1-3\omega_{ir})/2]^{\beta}(x_{jr}-z_{ir}) = 0 \tag{26}$$

Solving it for z_{ir} we obtain:

$$z_{ir} = \frac{\sum_{j=1}^{n}[(1-3\omega_{ir})/2]^{\beta}[(1-3\mu_{ij})/2]^{\alpha}x_{jr}}{\sum_{j=1}^{n}[(1-3\omega_{ir})/2]^{\beta}[(1-3\mu_{ij})/2]^{\alpha}} \quad \square$$

4. EXPERIMENTS

In this section we discuss the experiments.
For evaluating the efficiency of the Type-2 Projected GK clustering algorithm, we have compared its performance with PROCLUS and GK algorithms using real and synthetic data sets. The parameters of each algorithm were fine tuned and multiple runs of the experiments were conducted to minimize the effect of initialization.

Table 3. F1-Measure

Data Sets	GKS	PROCLUS	GK
Forest Fire	0.2695	0.1004	0.3443
Alzheimr	0.1301	0.0851	0.5497
Parkinson	0.2958	0.5631	0.6489
Breast Cancer	0.5921	0.8330	0.8332

Table 4. Recall

Data Sets	GKS	PROCLUS	GK
Forest Fire	0.2992	0.2731	0.3230
Alzheimr	0.1667	0.0953	0.5375
Parkinson	0.2757	0.6914	0.6703
Breast Cancer	0.5710	0.8906	0.7911

Table 5. Precision

Data Sets	GKS	PROCLUS	GK
Forest Fire	0.2451	0.0615	0.3686
Alzheimr	0.1067	0.0769	0.5587
Parkinson	0.3191	0.7101	0.6289
Breast Cancer	0.6289	0.7825	0.8800

4.1 Data Sets

Forest Fire, Breast Cancer, Parkinson and Alzheimer data sets from the UCI data repository were used for experimentation [29]. These data sets have no missing values. Table 1 describes these data sets.

4.2 Cluster Validity

Cluster validity measures have been used to asses the quality of the output produced by clustering algorithms [14] [9]. We have used the validity measures accuracy, recall, precision, specificity, F1-measure in our experiments. These are described below:

(1) Accuracy: We use clustering accuracy measure defined in [12]. Accuracy may not be an effective measure of evaluation in several situations such as fraud detection in banking transactions or intrusion detection as in such situations it is important to label the exception cases correctly. Alternative measures of evaluation such as recall, precision, specificity, F1-measure are used in such situations.

(2) Recall: ratio between the number of correct positive predictions and the number of positive examples.

(3) Precision: ratio between the number of correct positive predictions and the number of positive predictions.

(4) Specificity: ratio of number of true negatives and sum of number of true negatives and false positives.

(5) F1-measure: is the harmonic mean of precision and recall. [1]

Table 2 shows, while the Type-2 Projected GK algorithm achieves highest accuracy for Alzheimer and Breast Cancer data sets, the performance of the Type-2 Projected GK algorithm is comparable to PROCLUS and GK algorithm for the other two data sets. In Table 4, 6, 5, and 3, we present the results of applying recall, specificity, precision and F1-measure to the outcomes of clustering schemes produced by different algorithms. GK algorithm achieves highest F1-measure and Precision for Breast Cancer,

[1] In order to compare the cluster accuracy results of PROCLUS, GK, with Type-2 Projected GK algorithm, we defuzzified the fuzzy assignments.

Table 6. Specificity

Data Sets	GKS	PROCLUS	GK
Forest Fire	0.6321	0.6481	0.8325
Alzheimr	0.5849	0.5155	0.7679
Parkinson	0.2757	0.6943	0.6703
Breast Cancer	0.5170	0.8906	0.7911

Alzheimer, Parkinson and Forest data sets. PROCLUS algorithm achieves highest recall and specificity Alzheimer for Breast Cancer and Parkisnon data sets, GK algorithm achieves highest recall and specificity for and Forest data sets and GK algorithm achieves comparable F1-measure, recall, precision and specificity for Alzheimer and Forest data sets as compared to PROCLUS.

5. CONCLUSION

The contribution in this paper is adaptation of GK algorithm for type-2 projected clustering algorithm. It captures higher degree of uncertainties along with the typicality in the data set. We have done the theoretical analysis of the type-2 projected GK algorithm. Experiments show improvement in results over other comparative techniques.

6. REFERENCES

[1] . Hinneburg, C. Aggarwal, and D.A. Keim, What is the nearest neighbor in high dimensional spaces? In Proceedings 26th International Conference on Very Large Data Bases (VLDB-2000), Cairo, Egypt, September 2000, pp. 506-515, Morgan Kaufmann (2000).

[2] . K. Jain, M.N. Murthy, P.J. Flynn, Data clustering: a review, ACM Comput. Survey, vol. 31(3) pp. 264-323, 1999.

[3] .Wiswedel and M.R.Berthold. Fuzzy clustering in parallel universies In Proc.Conf. North American Fuzzy Information Processing Society(NAFIPS 2005), pp. 567-572, 2005.

[4] . Aggarwal, J. Wolf, P. Yu, C. Procopiuc, and J. Park. Fast algorithms for projected clustering In Proceedings of the 1999 ACM SIGMOD international conference on Management of data, pp. 61-72. ACM Press, 1999.

[5] . Bohm, K. Railing, H.-P. Kriegel, P. Kroger, Density Connected Clustering with Local Subspace Preferences, ICDM, Fourth IEEE International Conference, pp. 27 - 34, Nov. 2004.

[6] . Dubois and H. Prade, Fuzzy Sets and Systems: Theory and Applications. New York: Academic, 1980.

[7] . Lin, M.S. Yang, A Similarity Measure between Type-2 Fuzzy Sets with Its Application to Clustering, Proc. of Int. Conf. on Fuzzy Systems and Knowledge Discovery, pp. 726-731, 2007.

[8] .C.H. Rhee and C. Hwang, A Type-2 Fuzzy-c-Means clustering algorithm, In Proceedings of IEEE FUZZ Conference,

Melbourne, Australia, pp.1926- 1929, December 2001.

[9] . Hoppner, F. Klawonn, R. Kruse, and T. Runkler, Fuzzy Cluster Analysis: Methods for Classification, Data Analysis, and Image Recognition, John Wiley Sons (1999).

[10] .J. Klir, B.Yuan, Fuzzy sets and Fuzzy Logic:Theory and Applications, Prentice Hall, Upper Saddle River, NJ, 1995.

[11] . J. Klir and T. A. Folger, Fuzzy Sets, Uncertainty and Information. Englewood Clifs, NJ: Prentice Hall, 1988.

[12] . Gan, J. Wu, A Fuzzy Subspace Algorithm for Clustering High Dimensional Data, ADMA, 2006.

[13] .B. Mitchell, Pattern recognition using type-II fuzzy Sets,Information Sciences pp. 409-418, 2005.

[14] . Abonyi and Balazas Feil, Cluster Analysis for Data Mining and System Identification, Birkhauser.

(1) J.C. Bezdek, Pattern recognition with Fuzzy Objective Function Algorithm, Plenum Press, New York, 1981.

[15] . Han and M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, 2006.

[16] .M. Mendel, Advances in type-2 fuzzy sets and systems, Information Sciences, pp. 84110, 2007.

[17] .M. Mendel, R.I. John, Type-2 fuzzy sets made simple, IEEE Transactions on Fuzzy Systems, pp. 117127, 2002.

[18] .N Karnik and J.M. Mendel, Introduction to Type-2 Fuzzy Logic Systems, In Proc. 7th Intl. Conf. on Fuzzy Systems FUZZ-IEEE'98, pp. 915-920, 1998.

[19] . Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is "nearest neighbor" meaningful? In C. Beeri and P. Buneman, editors, Database Theory - ICDT '99,7th International Conference, Jerusalem, Israel, January 10-12, 1999, Proceedings, volume 1540 of Lecture Notes in Computer Science, pp. 217-235, 1999.