Middle-East Journal of Scientific Research 23 (10): 2533-2545, 2015
ISSN 1990-9233

DOI: 10.5829/idosi.mejsr.2015.23.10.11

Automated Subspace Two Variable Weighted Clustering Method for Mixed Numeric and Categorical Values with Complete and Incomplete Dataset

M. Kalaiarasu and R. Radhakrishnan

Department of IT, Sri Ramakrishna Engineering College,
Principal, Vidhya Mandhir Institute of Technology

Abstract: Novel Subspace Two Variable Weighted Clustering (NSTVWC) is proposed that can function with apartial and complete mapping. The proposed approach is competent for clustering complete and incomplete view dataset samples and moreovervalid for measuring the values of categorical, numerical and mixed data attributes. Thismethodis executed impartingTerm Frequency (TF) -Inverse Document Frequency (IDF), Kullback -Leiber Divergence (KLD) and Shannon Entropy (SE) method. Application of Modified Independent component analysis (MICA) contrast functions based on canonical correlations in a replicating kernel Hilbert space to convert incomplete view dataset samples into complete view datasets. For view and variable weight estimation, centroid values of the subspace clustering methods are optimized based on Stepsizesteepest descent (SSD) and Fuzzy Artificial Fish Swarm (FAFS) Optimization algorithm. Experimental results on three real-life data sets have shown that the designed NSTVWC significantly executes all the challenging processes in terms of Precision, Recall, F Measure, Average Cluster Entropy (ACE) and Accuracy consistently in complete and incomplete view of the data in regard to the true clusters in the data. The NSTVWC framework is to bid a remedy for the mixed attributes multiview clustering problem, the potential of various mode of existing clustering algorithms and features of different types of complete and incomplete views of datasets could be totally exploited and integrated. Results shows that proposed NSTVWC framework achieves enhanced clustering.

Key words: Data mining · Clustering with Mulitiviewdata · Subspace clustering · Augmented Lagrangian Cauchy step computation (ALCS) · Fuzzy Artificial Fish Swarm (FAFS) Optimization · Singular value decomposition (SVD) · Incomplete view data

INTRODUCTION

Actually the datasets are represented in the form of multiple views[1, 2]. For example, Web Pages are generally in cooperation with the page-text and hyperlink features; images on the web have related designation linked; in multi-lingual information retrieval, the same document has multiple statements in various languages and so on. Albeit these individual views might be adequate on their own for a given data assignment, providing parallel information to each other can direct to flexible enhancement of the performance. Each source of information is fundamentally a view of the data and learning with such type of data is characteristically referred to as multi-view learning [1, 3].

There are various approaches to express the same set of data objects in several data analysis tasks. This directs to the accessibility of multiple distinct interpretation that encode patterns significant to the domain [1]. Ahead of learning from a single view, multiple different views frequently have a synergistic effect on learning, recuperating the performance of the resulting model. Multi-view learning is mainly valid to applications that sync collection of data from various modalities with each single modality presenting one or more views of the data. Each view incorporates unique complementary information about an object; only in combination do the views generate an absolute depiction of the original object. In another view, if possibility to improve learning is provided, concepts that are challenging to learn in one view may be simpler approach. Multi-view learning can distribute learning progression in a single view via the direct correspondences between views.

Corresponding Author: M. Kalaiarasu, Department of IT, Sri Ramakrishna Engineering College, Principal, Vidhya Mandhir Institute of Technology, India.

The means of learning from multiple views is to stimulate each view's own knowledge base in order to surpass the basically concatenating views. As unlabelled data are ample in real life and its rising quantities come in multiple views from different sources, the trouble of unacceptable learning from multiple views of unlabelled data has alerted attention [4], termed to as multi-view clustering. The objective of multi-view clustering is to partition objects into clusters based on multiple representations of the object. Current multi-view clustering algorithms can be generally classified into three categories. Algorithms in the first category [4] incorporate multi-view integration into the clustering process directly through optimizing particular loss functions whereas algorithms in the second category such as the ones based on Canonical Correlation Analysis [5] first predict multi-view data into a regular lower dimensional subspace and then operate any clustering algorithm such as k-means to learn the partition. The third category in which a clustering solution is derived from each individual view and then all the solutions are fused base on consensus is called late integration or late fusion [6-7]. Recently, number of algorithms, both supervised and unsupervised, has been proposed to develop multiple views of the data.

Research focuses on variable weighting clustering in cluster analysis [8-9]. It involuntarily calculates a weight for each variable and recognizes significant and insignificant variables through variable weights. The multiview data could be considered as have two levels of variables. The variance of views and the significance of individual variables in each view should be taken into account in a clustering the multiview data. The conventional variable weighting clustering approach simply computes weights for individual variables and ignores the differences in views in the multiview data which is not suitable for multiviewdata. Though, there are many circumstances in which complete datasets are not accessible in the actual world applications.

Existing multi-view algorithms usually indicate each object is represented in all views presumimg that there is a complete bipartite mapping between instances in the various views to characterize these correspondences. Prior works majorly focus on numerical data whose intrinsicgeometric properties can be manipulated naturally to classify distance function between data points. However the contemporary datas in the databases is predominantly categorical in which attribute values cannot be naturally ordered as numerical attribute values. Due to the variation in the characteristics of attributes, challenge to develop criteria function for mixed data.

The chief incentive of the proposed design is to crack the complexity of criteria function for mixed data, weight value calculation and centroid selection in multiview data with incomplete data point of view .In this work, the limitations of criteria function for mixed data is overcome by introducing a Term Frequency (TF) -Inverse Document Frequency (IDF), Kullback -Leiber Divergence (KLD) and Shannon Entropy (SE) method to compute the attribute value of numerical and categorical data. In the premeditated KICASDSTWC system is contemplated for clustering both complete and incomplete view data in mutliviewdata. Incomplete view of data is aided by Modified Kernel-Based Independent Component analysis (MKICA) that differentiates the complete and incomplete view of multi view data, the effects of different views and different variables in clustering and the weights of views and individual variables are consequentially calculated based on the FAFS. Accordingly, the view weights and variable weights indicate the importance of the views in the entire data and the importance of variables in the view respectively.

Selection or optimization of the fuzzy centroid values for Subspace Two Variable Weighted Clustering (STVWC) is recommended by Step size Steepest Descent (SSD) Algorithm. Augmented Lagrangian Cauchy step computation (ALCS) is to tally the objects in subspaces where they are consistent and have high correlated utilities. The proposed a Novel Subspace Two Variable Weighted Clustering (NSTVWC) is a comprehensive depiction that support both incomplete and complete view data which is competent in clustering large high dimensional multiview data, value of the numerical and categorical data values are assessed to differentiate numerical and categorical data using the metrics like TF-IDF, KLD and SE. Experimental results on real datasets validates the efficacy of NSTVWC methodology and it is evaluate against the existing algorithm for both complete and incomplete view data.

MATERIALS AND METHODS

In Aiming multi-view clustering with both complete and incomplete view of the data, a novel Modified Kernel based Independent kernel Component Analysis (MKICA) and StepsizeSteepest Descent (SSD) methods for Subspace Two Variable Weighted Clustering (STVWC) so a Novel Subspace Two Variable Weighted Clustering (NSTVWC) approach has been proposed in this paper. Computing thesignificance of the mixed attributes values still develops an inexplicable problem in existing multiview

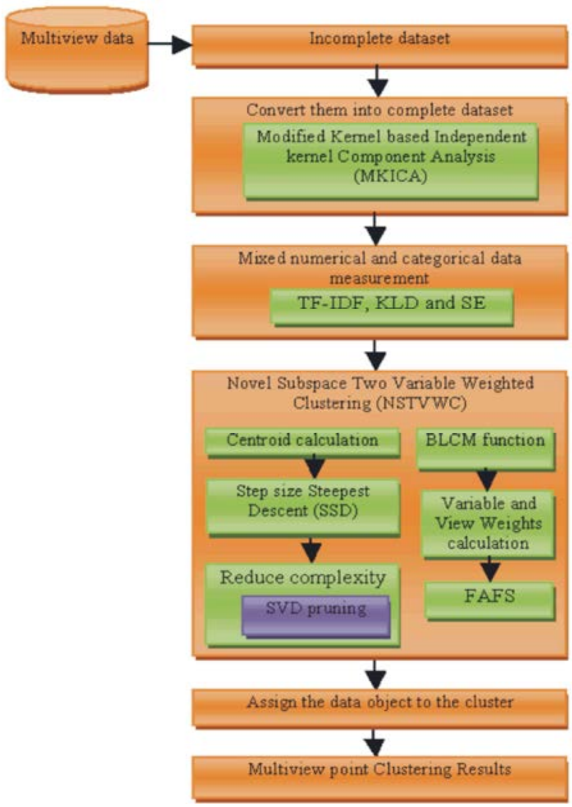

Fig. 1: Flowchart Representation Of Proposed Methodology

point clustering methods which can be overcome by conversion of dataset from SSD which is divided into categorical and numerical attributes and is reserved as same, if anyone of the attributes corresponds to both categories. Therefore the measurement of the attributes value becomes significant, in order to perform the categorical attributes, numerical attributes and mixed attributes values are measured based on the metrics like Term Frequency (TF) -Inverse Document Frequency (IDF), Kullback-Leiberuncertain (KLD) and ShannonEntropy (SE). The premeditated NSTVWC, the partial view data are transformed into entire data by proposing the MKICA in which the subspaces are created in accordance with a set of centroids by which calculation impart SSD methodalong with utility function. During commencement, the input data results from MKICA values are transformed which then calculate the attribute values and subsequently convert them into the fuzzy centroid values, optimized using the SSD method. The projectedmethodology distinguishes the impacts of several views and variables by setting up the weights of views and individual variables to the distance function. With the assistance of the algorithm, the view and variable weights values is the objective function which is optimized Fuzzy Artificial Fish Swarm (FAFS). The entirefunction of the proposed work is illustrated in Figure 1.

At first the incomplete dataset is transformed into complete dataset by proposing Modified Kernel based Independent kernel Component Analysis (MKICA) in order to carry out multiview clustering for both complete and incomplete dataset. For easiness, consider X and Y represent the two number complete and incomplete dataset respectively. Generalization to over two types of complete and incomplete dataset with one complete and remaining incomplete dataset can be performed in aanalogous way. Assume that complete mutliview data is indicated as X while incomplete multiview dataset is indicated as Y i.e., the variables values of the multiview data are accessible for only a subset of the entire examples. In the proposed MKICA method, Hilbert-Schmidt Independence Criterion (HSIC) is applied to compute values of complete and incomplete dataset values which achieved as the squared Hilbert-Schmidt (HS) norm of the covariance operator between mappings to replicating kernel Hilbert spaces (RKHSs) [10] and simplify the featureof complete and incomplete dataset samples based on Hilbert space F. The Hilbert space F for complete and incomplete view dataset is defined based on the point evaluation operator d: X, $Y \rightarrow R$, which maps attributes of the complete and incomplete dataset $a \in A$ to $a(u) \in$ R, is a continuous linear functional. To each complete and incomplete dataset point of view $x, y \in U$, there corresponds to a feature (attribute) elements value $\alpha_u \in A$, where ψ: $U \times U \rightarrow R$ is a unique positive definite kernel for complete and incomplete dataset samples. Also additionally define a second reproducing kernel Hilbert spaces (RKHSs) G with respect to complete and incomplete dataset U, with attribute map $B \in \beta \in A$ and corresponding kernel $<\beta_x,\beta_y>_g=\hat{\psi}(x,y)$. Let be a joint measure on ($U \times U$, $G \times \wedge$) (here G and $\wedge$ are Borel salgebras on complete and incomplete dataset U), with associated covariance C_{xy}: $G \rightarrow U$ and $f \rightarrow x$, $g \rightarrow y$,

$$< f, C_{xy}(g) >_U = E[f(x)g(y)] - E[f(x)]E[g(y)]$$

For all complete x $\in$ U and incomplete dataset x $\in$ U, the squared HS norm of the covariance operator C_{xy} is denoted as, HSIC, is then.

$$\begin{aligned}\left\|C_{xy}\right\|_{HS}^2 &= E_{x,x',y,y'}[\psi(x,x')\hat{\psi}(y,y')] \\ &+E_{x,x'}[\psi(x,x')]E_{y,y'}[\hat{\psi}(y,y')] \\ &-2E_{x,y}[E_{x'}[\psi(x,x')]E_{y'}\hat{\psi}(y,y')]]\end{aligned}$$

where $E[.]$ represent the expectation over the corresponding random variables for complete and incomplete view dataset. In this work define a Gaussian kernel and use the same kernel for both complete and incomplete view dataset samples

$$\psi(x,y) = \hat{\psi}(x,y) := \phi(x,y) = \exp\left(\frac{-(x,y)^2}{2\lambda^2}\right)$$

In the ICA model the Hilbert-Schmidt Independence Criterion (HSIC) based model is represented over the Gaussian kernel space for both complete and incomplete view dataset is represented as H

$$H(U) := \sum_{1\le i<j\le m}^{m} E_{x,y}[\phi(u_i^T \bar{W}_{xy})\phi(u_j^T \bar{W}_{xy})] + E_{x,y}[\phi(u_i^T \bar{W}_{xy})\phi(u_j^T \bar{W}_{xy}] - 2E_x[E_{x,y}[\phi(u_i^T \bar{W}_{xy})\phi(u_j^T \bar{W}_{xy})]]$$

where $U := [u_1,..,u_m], \bar{W}_{xy} = W_x - W_y \in \Re^m$ denotes the difference between the x^{th} and y^{th} samples for completedataset and $E_{x,y}[.]$ denotes the empirical evaluation of the complete x^{th} and incomplete y^{th} samples. The above empirical evaluation based system govern the similarity value between the incomplete and complete view for multi view data. Data mining often enclose both numeric and categorical values. The conventional way to treat categorical attributes as numeric does not constantlygenerate meaningful results as many categorical domains are not ordered. In order to overcome these problems, object similarity measure is derived from both numeric and categorical attributes. Three major important metrics such as TF-IDF, KLD and SE is proposed in this work to compute the resemblance value for categorical and numerical attributes.

Term Frequency –Inverse Document Frequency (TF-IDF): In In the TF-IDF [11], the multiplication of two performance statistics, term as frequency and inverse document frequency. Various ways for determining the exact values of both statistics exist for numerical and categorical data. In case of the term frequency for selected categorical data and numerical data is represented as t and the d, symbolized by the dataset $tf(a, d)$ the selected attributes occurs the number of times in a dataset d.

$$tf(a,d) = 0.5 + \frac{0.5 * f(a,d)}{\max\{f(a,d)\}}$$

The inverse document frequency is a measure of how much selected attribute is significant for complete dataset samples. It is the logarithmically scaled fraction of the dataset that contain the categorical and numerical attribute value, attained by dividing the total number of datasets by the number of the dataset samples containing the numerical and categorical data and then taking the logarithm of that quotient.

$$idf(a,D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

Then tf–idf is calculated as,

$$tfidf(a,d,D) = tf(a,d) \times idf(a,D)$$

Kullback-Leibler (KL) Divergence [12] based measure for estimation of posterior probability values.

$$KL(p(a \mid p(a \mid D_i), p(a \mid D_j)) = \int_{-\infty}^{+\infty} p(a \mid D_i) \log \frac{p(a \mid D_i)}{p(a \mid D_j)} da + \int_{-\infty}^{+\infty} p(a \mid Dj) \log \frac{p(a \mid D_j)}{p(a \mid D_i)} da$$

Let $cvd_1,...,cvd_k$ is a sequence of k independent attributes for complete view dataset samples and identically distributed drawings with uniformly distributed in the alphabet $\{1,...,n\}$. Let k_q be the number of times symbol that selected categorical and numerical attribute data occurs in the complete view. Dataset is measured as follows:

$$H = H\left(\frac{k_1}{k}, \frac{k_2}{k}, ..., \frac{k_n}{k}\right)$$

When multiview dataset incomplete and complete view of the data are discovered, the attributes values are measured to differentiate categorical and numerical data in the dataset subsequently attain centroid value computation with the assistance of Step size Steepest Descent (SSD) for that purpose the complete and incomplete view results from KICA is given as $MVD = \{z_1,...,z_n\}$, its dimensions are described by set of n objects mo represented by the set A of m variables and view weights vw. Consider the value of object mo on attribute a and in time weight values is indicated by muv_{moaw}. Also consider csd represent an object chosen as the centroid from SSD. In addition, h_{csd} ($muv_{moawvw} = Smu_{moawvw}$) is indicated as a homogeneous function to determine the homogeneity among object m_0 and centroid csd, on attribute a in a multiview weight value. The users are permitted to define the homogeneous function, however the homogeneous values must be normalized to [0, 1] in order that $Smu_{moawvw} = 1$ points out that the value muv_{oat} is "perfectly" homogeneous of the centroid $muv_{cqnoawvw}$, or else which indicated by value $Smu_{oawvw} = 0$.

The Distribution Centroid: The fuzzy centroid as upgradeby Yang *et al*. [13] simulated the idea of distribution centroid for anenhanced representation of categorical variables. The cluster centers for the categorical variable part will be better represented by a fuzzy scenario. For $Dom(V_j) = \{v_i^1, v_i^2, v_i^3, .., v_i^t\}$ the distribution centroid of a cluster *mvc* is specified as C'_{mvc} and given as below.

$$C'_{mvc} = \{c'_{mvc1}, c'_{mvc2}, ..c'_{mvcj}, ..c'_{mvcm}\}$$

where, $$c'_{mvcj} = \left\{ \begin{array}{l} \{b_j^1, w_{mvcj}^1\}, \{b_j^2, w_{mvcj}^2\}, ... \\ \{b_j^k, w_{mvcj}^k\}, ... \{b_j^t, w_{mvcj}^t\} \end{array} \right\}$$

In the above equation

$$w_{mvcj}^k = \sum_{i=1}^{n} \mu(x_{ij})$$

where $$\mu((z_{ij}) = \begin{cases} \dfrac{u_{imvc}}{\sum_{i-1}^{n} u_{imvc}} if z_{ij} = b_j^k \\ 0 if z_{ij} \neq b_j^k \end{cases}$$

At this point, the value of 1 is assigned to u_{imvc}, if the data object x_i belongs to cluster *mvc* otherwise which 0 is allocated, if the data object x_i do not belong to cluster *mvc*. Based on the above citedequations 10, 11, 12 and 13, it is apparent that the number of repetitions of each categorical value is been considered by the cluster computation of distribution centroid. Consequently, the distribution characteristics of categorical variables are considered to specify the center of a cluster. In the proposed approach, optimisation of fuzzy membership centroid values is prepared with the assistance of SSD [14]. SSDis an iterative and computations depending on the computation of the objective function *mvfcf*, at each iteration are commonly concerned. The SSD approach of choosing the best centroid values by maintaining least amount of cluster multiviewdatapoints for each cluster, subsequently recur the step until maximum number of points in the cluster is attained, or else go to step 3 and negative direction that is remaining points in the multiview data are elected to choose optimized centroid value. The optimized centroid values are reduced through estimation of Cauchy step size in step 3 of the algorithm then revise the chosen optimized fuzzy centroid value results in step 5 and move to step 2.

Algorithm 1 Step Size Steepest Descent Algorithm for Centroid Calculation:

- Compute distance matrix $Dist_{m\times m}$ in which $dist(z_i, z_j)$ signifiesdistance from z_i to z_j;
- Make an preliminary guess $\mu(z_{ij})$ at the minimum; keep $k = 0$. Choose convergence parameter $\varepsilon > 0$, is calculated from distance matrix.
- Cauchy step and that by the Newton's method is bounded below by,

$$\frac{4\mu}{(1+\mu)^2}$$

  where $\mu = \frac{\lambda_1(Dist_{m\times m})}{\lambda_N(Dist_{m\times m})}$ is the ratio between the largest Eigen value and the smallest eigenvalueof $Dist_{m\times m}$
- Calculate the gradient steepest descent of the centroid objective function $\nabla mvfcf(z)$ at a point z^k to all other points and centroid is denotedas $c^{(k)} = \nabla mvfcf(z^{(k)})$.
- Compute centroid value as $\|C\| = \sqrt{C^T C}$, when $\|C\| < \varepsilon$ and $utility(u_{mowvw}) > u_{mowvw}$ (min) > 0.5 then terminate the iteration process $z^* = z^{(k)}$ is minimum number of cluster multiview data cluster datapoints. Otherwise go to step 3.
- Consider the search direction at the current point $z^{(k)}$ as $d^{(k)} = -c(^k)$
- Compute a step size $a^{(k)}$ to reduce fuzzy centroid value $(z^{(k)} + \alpha^{(k)} d^{(k)})$.
- One dimensional search is exploited to determine $a^{(k)}$.
- Revise the chosen fuzzy centroid values as $z^{(k+1)} = z^{(k)} + \alpha^{(k)} d^{(k)}$

10. Keep $k = k + 1$and move to step 2

- Gaussian function which is employed above is the homogeneous function as similarity among data object *mo* and centroid *cqn* is been normalized on feature (a, w) to [0;1]. The homogeneous function is specified as below:

$$h_{csd} = \exp\left(-\frac{|v_{csdawvw} - v_{moawvw}|}{2\sigma_{csd}^2} \right) \nabla f(\mu(x_{ijk}))$$

where σ^2_{csd} indicates a parameter which sustainsthe width of the Gaussian function centered at centroid *csd*. At this point, the similarity function is not symmetric, i.e. $h_{csd}(v_{moawvw}) \neq h_{mo}(v_{csdawvw})$, as the calculation depends on the distribution of objects centered at the former object. The evaluation of width of the Gaussian function is completed with the help of k-nearest neighbor's heuristic [10] and is given as:

$$\sigma_{csd} = \frac{1}{k} \sum_{n \in Neigh_{csdawvw}} dist_a(csd, n)$$

where $Neigh_{csdawvw}$ represents the set of k-nearest neighbors of object mo on feature a,vww and $k = \rho|mo|$ with a supposition that ρ is the neighbourhood parameter expressedby users. In accordance with the distribution of the objects projected in the data space of attribute, the width of the Gaussian function is being implemented by the k-nearest neighbours heuristic, consequently showing that σ_{csd} is stronger than keeping a constant value. Calculating and pruning the homogeneous tensor using SVD [15] for optimized centroid σ_{csd}, a homogeneous tenso $S \in [0,]^{|mo|\times|\alpha|\times|vww|}$ is characterized comprising the homogeneity values MUS_{moawvw} with respect to centroid *csd*.

Algorithm 2: SVD Pruning Input Homogenous Tensor Output: Pruned Homogenous S:

- M = unfold(S)
- Add dummy row and column to M
- While true do
- $N \leftarrow$ zeromeannormalization(M)
- $U\Sigma V^{\wedge} \leftarrow N$ //SVD decomposition on
- $u \leftarrow$ principalcomponent
- $v \leftarrow$principalcomponent
- Calculate threshold τ_u τ_v
- Prune row i of M if $|u(i)| < \tau_u$, $1 \le I \le r$
- Prune column i of M if $|v(i)| < \tau_v$, $1 \le j \le r$
- If there is no pruning then break
- Remove dummy row and column from M
- S = fold(M)

Initially, zero mean normalization is carried out on matrix M to get hold of the zero mean normalized matrixes N (Line 4), whichlatterly exploited to compute the covariance matrices. Zero mean normalization is carried out by computing the mean αvg_j of the matrix M that $\forall_j \in \{1,..c\}$ of each column.

$$avg_c = \frac{1}{r} \sum_{i=1}^{r} M(i, j)$$

Subsequently, from each entry of M, its equivalent column mean $\forall_j \in \{1,..c\}$ is been subtracted.

$$N(i, j) = M(i, j) - avg_j$$

At some point in the performance of the clustering process for the above returned centroid values, the homogeneous tensor S together with the utilities of the objects were exploited to compute the probability of each value $muv_{cadoawvw}$ of the data to be clustered with the centroid *csd*. Subsequently, the covariance matrices of the homogeneous values in the object space and feature space called NN' and $N'N$ respectively were calculated (N' is the conjugate transpose of matrix N).

$$NN' = U \sum{}^2 U'$$

$$N'N = V \sum{}^2 V'$$

where U represents a $r\times r$ orthonormal matrix (its columns are the eigenvectors of NN'), Σ^2 represents a $r \times c$ diagonal matrix with the Eigen values on the diagonal and V is a $c \times c$ orthonormal matrix (its columns are the eigenvectors of $N'N$). If the magnitude of the pruned objects in their related elements of their principal components is little (Line 9 and 10), a heuristic however parameter-free approach can be proposed to find out the threshold τ_u for pruning objects. For pruned rows (objects) and columns (features) of matrix M, the homogeneous values are fixed to "0". The process of computing SVD and pruning the matrix M is replicated until there is no more pruning. The clustering process for computing the probability value is carried out in which $p_{moawvw} \in \Re$ represent the probability of object mo to be clustered with centroid *csd* on attribute α. The view weight v, variable weight w for multiview data is computed with the help of Artificial Fish Swarm Algorithm (AFSA). Consider $P \in \Re|mo|\times|a|\times|vww|$ be the probability tensor, such that p_{moawvw} is an element of it provided with the respective indices $|mo|\times|a|\times|vww|$. The following objective function is then maximised to calculate the probabilities: To perform the clustering process, the objective functions are defined as:

$$f(p) = \sum_{mo \in MO} \sum_{a \in A} \sum_{vw \in VW} \sum_{w \in W} h^{p_{moawvw}}_{(v_{moawvw})}$$

$$g(p) = \sum_{mo \in MO} \sum_{a \in A} \sum_{vw \in VW} \sum_{w \in W} p_{moawvw} - 1$$

The Optimization of $f(p)$ under constraint $g(p)$ is a linear programming problem, as $f(p)$ and $g(p)$ is linear functions of the design variable p. Augmented Lagrangian multiplier technique is then exploited to maximize the objective function $f(p)$ for clustering multiview data in the subspace clustering technique. As a result, the modified objective function is defined as,

$$F(p) = -f(P) - \lambda g(P) + \frac{\mu}{2} g(P)^2$$

$$utility(u_{mowvw})$$

The optimization of $F(p)$ (Algorithm 3) depends on Augmented Lagrangian Cauchy Step computation (ALCS) methods, $f(P)$ and $g(P)$ would be employed by ALCS with the intention that the constrained optimization problem are been replaced with iterations of unconstrained optimization sub problems, Hence, the iterations continue until the solution converges. For algorithm 3, ALCS necessitates three parameters such as μ_k, Θ_k, ϵ_k to calculate the optimized probability value for clustering process. In the majority of situations, the results are insensitive to these parameters and therefore can be fixed to their default values. The closeness cluster results for multiview data results is constantly indicated by parameter μ_k. Consequently, δ provides the standard tradeoff between accuracy and efficiency, i.e., smaller δ indicates longer computation time however better result. Parameter Θ_k maintains the level of clustering to the constraint $g(p)$. Parameter ϵ_k auxiliary nonnegative scalar quantities on $F(p)$ when the constraint is breached,

Algorithm 3: Augmented Lagrangian Cauchy step Computation (ALCS)
Input: Initial Probability Distribution P_a :
Output: The optimal probability distribution P_a^*

- Initialize $P_a^o, \mu_k > 0, \Theta_k > 0, \in_k > 0, \gamma \in (0,1)$
- While $P_a^i(z_k) < \mu_k$ true do
- Set $P_a^* \leftarrow P_a^i(z_k) - z_k$ then return P_a^*
- If not satisfied do
- $P_a^* \leftarrow \gamma P_a^i(z_k)$
- End while
- If $| g(P_a^i) | < | g(P_a^{i-1}) |$ then
- Return $g(P_a^i)$
- $\Theta_k \leftarrow \Theta_k . g(P_a^i)$
- Else $\in_k \leftarrow \in_k . g(P_a^i)$
- end if
- $\lambda^i \leftarrow \lambda^i - \in_k . g(P_a^i)$
- $i \leftarrow i+1$
- End procedure

From the results of the optimized probability values for multiview data both view and variable weights values are calculated using the fuzzy artificial fish swarm (FAFS) algorithm. Artificial Fish (AF) is a fictitious entity of true fish, which is exploited to carry on the analysis and explanation of the problem and can be recognized by exploiting an animal ecology conception. The functions multiview clustering data samples that comprise the behaviours of the AF: AF_Prey, AF_Swarm, AF_Follow, AF_Move. Every fish typically resides in the place with a best objective function (21). New fuzzy based artificial fish swarm algorithm [16] to control the visual and step parameters for variable (w) and view weights (vw) of global and local searching weight calculation adaptively.

Let (XW) is the current state of the variable (w) and view weights (vw) values and it is represented as $XW = \{xw_1,...,xw_n\}$ and $Xw_v = \{xw_{1v},...,xw_{nv}\}$ and then process can be expressed as follows.

$$xw_i^v = xw_i + visual.rand(), i \in (0,n]$$

$$XW_{next} = XW + \frac{XW_v - XW}{\|XW_v - XW\|} step.rand()$$

where Rand () produces random numbers between 0 and 1, Step is the step length to perform variable and view weight calculation for clustering complete and incomplete view dataset and xw_n is the variables parameter, n is the total number variables weights . Visual represents the visual distance of one weight value to another weight value and δ is the crowd factor $0 < \delta < 1$.

Generally AFSA the visual and step length kept same, it reduces the clustering results it is overcome by using fuzzy parameter is named as constriction weight. Weight must be greater than 0 and smaller than 1. Current iteration for variable and view weight calculation of visual and step values are determined according to the following formulas:

$$visual_{iter} = CW \times visual_{iter-1}$$

$$step_{iter} = CW \times step_{iter-1}$$

$visual_{iter}$ and $step_{iter}$ stand for the current view and variable weight calculation iteration for visual and step and $visual_{iter-1}$ and $step_{iter-1}$ is the previous view and

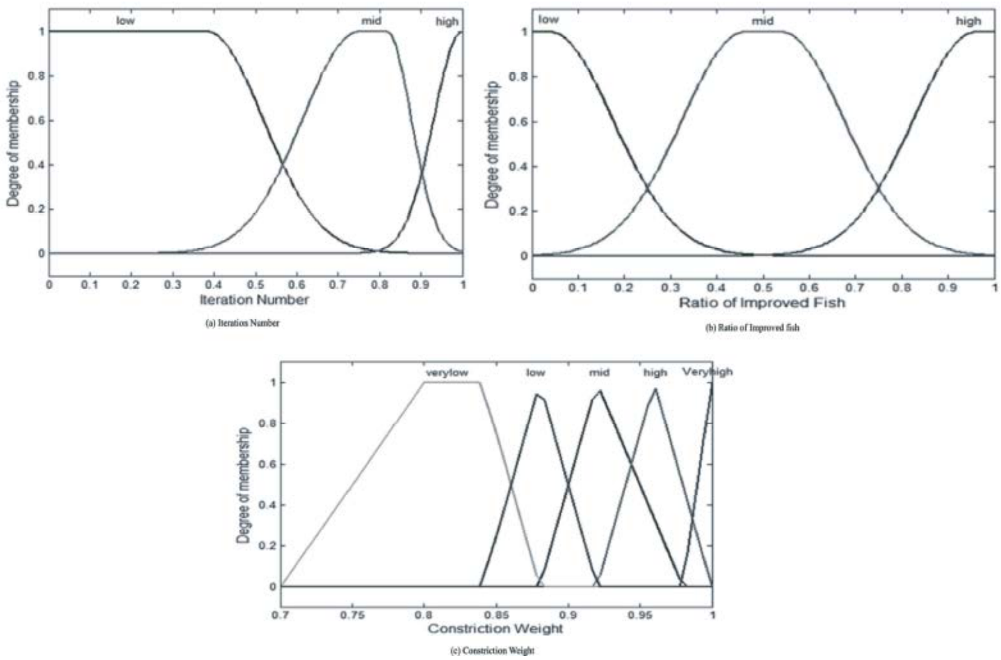

Fig. 2: Fuzzy Uniform Fish Membership functions

Table 1: Fuzzy associative memory rules

Iteration number	Ratio of improved fish	Construction weight
L	H	VH
L	M	H
L	L	M
M	H	H
M	M	M
M	L	L
H	H	M
H	M	L
H	L	VL

variable weight calculation iteration for visual and step respectively. In this work the construction weight of initial parameters such as $visual_{iter}$ and $step_{iter}$ of task (fishes) is updated using fuzzy membership function. Figures 2(a) and 2(b) show the membership functions for Inputs: Iteration Number and Ratio of Improved Fish. Constriction Weight is the output of the fuzzy engine which has the membership functions of Figure 1(c).The proposed fuzzy the rules shown in the fuzzy associative memory in Table 1, where, VL: very low, L: low, M: mid, H: high and VH: very high.

Now precede the variable and view weight calculation based on the basic behaviours of AF are defined [17] as follows with updated visual and step length results.

AF_Prey: This is a basic biological behaviour that tends to the each variable and view weights is allocated to clustering objective function, generally the fish (variable and view weights) perceives the concentration of best weight values food in water to determine the movement by vision,

$$XW_j = XW_i + visual_{iter}.rand()$$

If $Y_i < Y_j$ in the maximum optimized view and variable weight calculation results then it is forwarded tocluster; otherwise, select a state Y_i and judge whether it satisfies the forward condition. If it cannot satisfy clustering results for data pointsafter maximum number of iterations completed by fish, it moves a step arbitrarily to choose different variable and view weights.

$$XW_i^{t+1} = XW^t{}_i + visual_{iter}.rand()$$

AF_Swarm: The fish will assemble view and variable weights in groups that are naturally assign weight to data points in the moving process, which is a kind of living habits to satisfy clustering results for all datapointsand avoid dangers stage of the cluster. Behavior description:

Let Y_i and Y_c be the AF current state of variable and view weights and the center location for variable weightsrespectively. Let n_f be the number of its companions in the current neighborhood ($d_{ij} < visual_{iter}$), n is total fish number. If $Y_c > Y_i$ $Y_c > Y_i$ and $\frac{nf}{n} > \delta$, which means that the companion center has more best view and variable weights results for each cluster data points and is not very crowded, it goes forward a step to the companion center;

$$XW_i^{(t+1)} = XW_i^{(t)} + \frac{XW_j - XW_i^{(t)}}{\left\| XW_j - XW_i^{(t)} \right\|}$$

If not, executes the preying behaviour. The crowd factor limits the length of the searching space of variable and view weights.

AF_Follow: In the moving process of the variable and view weights from one place to many places then find best variable and view weights calculation by comparing the neighbourhood partners will trail and reach best clustering results rapidly.

$$XW_i^{(t+1)} = XW_i^{(t)} + visual_{iter}.rand()$$

AF_Move: Fish (tasks) swim randomly in the water; in fact, they are seeking best variable and view weights for each data points in the cluster food or companions in larger ranges

$$XW_i^{(t+1)} = XW_i^{(t)} + visual_{iter}.rand()$$

To increase likelihood value, results of view and variables weights by based on the fish behaviour attempt to add leaping behaviour to AF. The AF's leaping behaviour is defined as follows.

AF_Leap: If the objective functions of the view and variable weight values is less than or equal to objective function (21) for n_i iterations performed by fish, β is a parameter or a function that can make some fish have other abnormal actions (values), *eps* is a smaller constant for each view and variable weights,

$$IF(f(p) - f(p)(n_i)) < eps$$

$$XW_{some}^{(t+1)} = XW_{some}^{(t)} + \beta.visual_{iter}.rand()$$

This procedure is continued until all; the multiview clustering data samples is all completed. It can be basically validated whether the objective function (20) can get minimized with regard to *VW* and *W* if $\eta \geq 0$ and $\xi \geq 0$. In addition, it is carried out as given below: $\eta > 0$, based on (35), *vw* is inversely proportional to *E*. The smaller E_j and the larger v_i shows that the equivalent variable is more significant. $\eta > 0$ based on (35), $\eta = 0$ will generate a clustering result with only one significant variable in a view which possibly will not be desirable for high dimensional data. The attributes are presumed to be segmented into *T* views $\{G_t\}_{t=1}^{T}$

$$\delta(vw_j) = \frac{\exp\left\{-E_j/\eta\right\}}{\sum_{k \in G_t} \exp\left\{-E_j/\eta\right\}}$$

$$E_j = \sum_{l=1}^{k}\sum_{i=1}^{n}\sum_{j=1}^{m} \hat{N}(i,j,mvc) r\hat{v}w_t dist_a(nsd,n)$$

$\xi > 0$, based on (37), *w* is inversely proportional to *D*. The smaller D_t, the larger w_t, the more compact the corresponding view.

$\xi > 0$, based on (37), $\xi = 0$ will generate a clustering result with only one significant view. It is possibly not desirable for multiview data.

$$\delta(w_t) = \frac{\exp\left\{-D_t/\xi\right\}}{\sum_{k=1}^{T} \exp\left\{-D_t/\xi\right\}}$$

$$D_t = \sum_{l=1}^{k}\sum_{i=1}^{n}\sum_{j=1}^{m} \hat{N}(i,j,mvc) \hat{v}_j dist_a(nsd,n)$$

RESULTS AND DISCUSSION

To study the performance of the proposed NSTVWC for categorical and numerical data of incomplete view in classifying real-life data, three data sets from UCI Machine Learning Repository isselected [18]: the Multiple Features data set, the Internet Advertisement data set and the Image Segmentation data set. With these data, the performance of the proposed NSTVWC with existing Kernel-Based Independent Component Analysis and Steepest Descent Subspace Two Variable Weighted Clustering (KICASDSTWC), Quasi Newton's Subspace Two Variable Weighted Clustering (QNSTWC), TW-k-means with four individual variable weighting clustering algorithms(TW-k)[19], i.e., EWKM [20] compared.

Characteristics of Three Real-Life Data Sets: The Multiple Features (MF) data set contains 2,000 patterns of handwritten numerals that were extracted from a collection of Dutch utility maps. These patterns were classified into 10 classes ("0"-"9"), each having 200 patterns. Each pattern was described by 649 features that were divided into the following six views:

- Mfeat-fou view: contains 76 Fourier coefficients of the character shapes;
- Mfeat-fac view: contains 216 profile correlations;
- Mfeat-kar view: contains 64 Karhunen-Love coefficients;
- Mfeat-pix view: contains 240 pixel averages in 2 × 3 windows;
- Mfeat-zer view: contains 47 Zernike moments;
- Mfeat-mor view: contains 6 morphological variables. Here, use G_1,G_2,G_3,G_4,G_5 and G_6, to represent the six views.

The Internet Advertisement (IA) data set comprise a set of 3,279 images from various web pages that are categorized either as advertisements or non advertisements (i.e., two classes). The instances are described in six sets of 1,558 features, which are the geometry of the images (width, height and aspect ratio), the phrases in the url of the pages containing the images (base url), the phrases of the images url (image url), the phrases in the url of the pages the images are pointing at (target url), the anchor text and the text of the images alt (alternative) html tags (alt text). All views have binary features, apart from the geometry view whose features are continuous.

The Image Segmentation (IS) data set consists of 2,310 objects drawn randomly from a database of seven outdoor images. The data set contains 19 features which can be naturally divided into two views.

- Shape view: contains nine features about the shape information of the seven images;
- RGB view: contains 10 features about the RGB values of the seven images.

The graphical representations of the clustering results for variable and view weights with different variables and the methods results are shown in Figure 3. It shows the variation in variable weights for varying η = 8 values and view weights ξ = 1, ξ = 32, for TW-K means, QNSTWC, KICASDSTWC with incomplete view (ICV) ξ = 1, ξ = 32 and NSTVWC with incomplete view (ICV) ξ = 1, ξ = 32 results are shown in Figure 1. It shows that proposed NSTVWC with incomplete view (ICV) attainshigher clustering accuracy with less view weights values are automatically calculated using FAFS. The proposed method not only efficient for clustering complete and incomplete view dataset samples, it is also easily applicable for categorical, numerical and mixed data attributes by measuring the values of categorical,

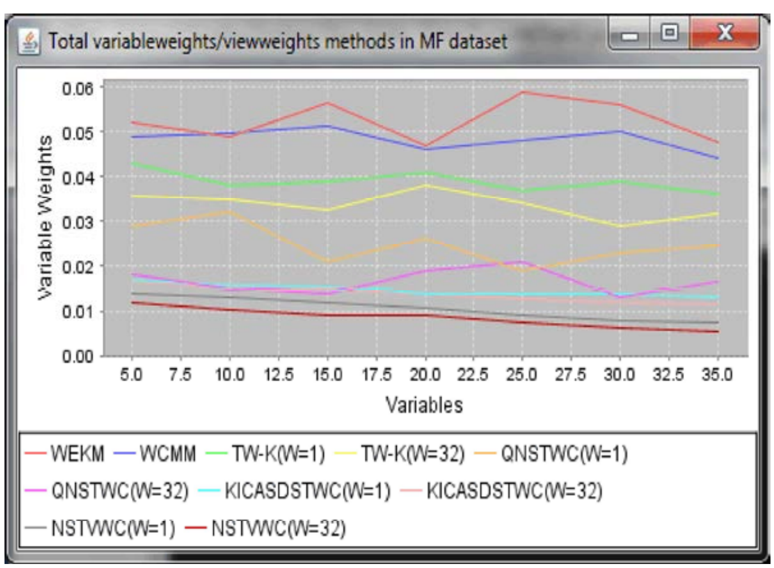

Fig. 3: Comparison of the total variable weights and view weights for methods in Multiple Features (MF) data set

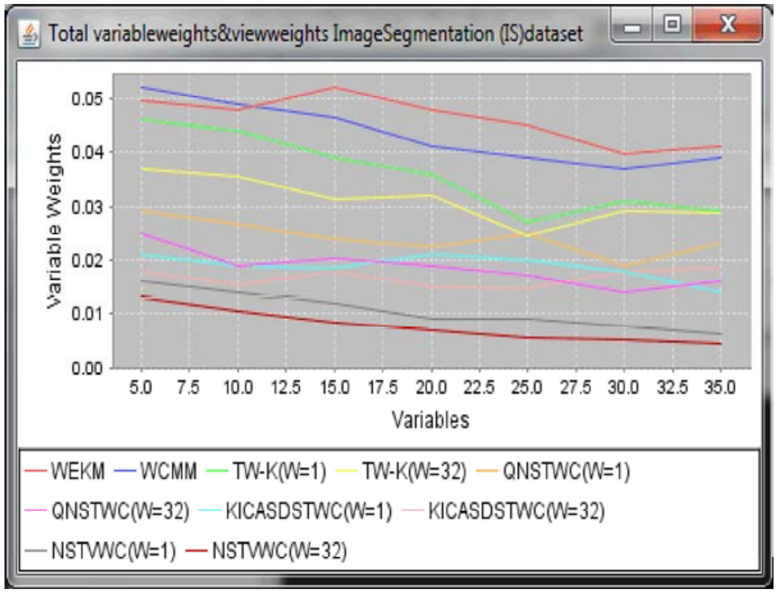

Fig. 4: Comparison of the total variable weights and view weights for methods in Internet Advertisement (IA) data set

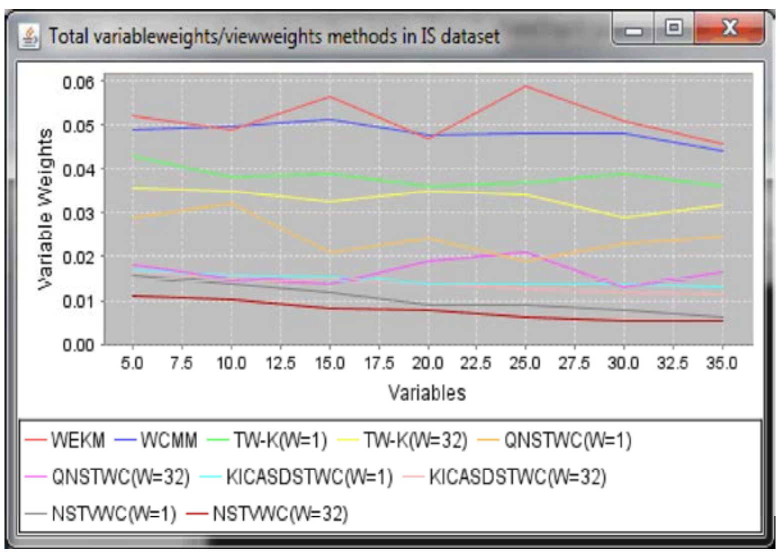

Fig. 5: Comparison of the total variable weights and view weights for methods in Image Segmentation (IS) dataset

numerical and mixed data. Because of this reason, it concludes that the proposed work have higher clustering accuracy when compare to existing clustering methods, distributed fuzzy centroid values are optimized are optimized using SSD.

The graphical representations of the clustering results for variable and view weights with different variables and the methods results are shown in Figure 4 for Internet Advertisement (IA) dataset. It shows the variation in variable weights for varying $\eta = 8$ values and view weights $\xi = 1$, $\xi = 32$, for TW-K means, QNSTWC, KICASDSTWC with incomplete view (ICV) $\xi = 1$, $\xi = 32$ and NSTVWC with incomplete view (ICV) $\xi = 1$, $\xi = 32$, results. It shows that proposed NSTVWC with incomplete view (ICV) achieves higher clustering accuracy, as proposed work additionally compute the values to differentiate the categorical and numerical mixed data in efficient manner. Automatic calculation of centroid values using SSD.

The graphical representations of the clustering results for variable and view weights with different variables and the methods results are shown in Figure 5 for Image Segmentation (IS) dataset. It representsthe variation in variable weights for varying $\eta = 8$ values and

Table 2: Summary of Clustering Results on Three Real-Life Data Sets by Six Clustering Algorithms

Dataset	Evaluation	WCMM	TW-K	QNSTWC	KICASDSTWC	NSTVWC
MF	Precision	0.59	0.81	0.83	0.88	0.91
	Recall	0.58	0.83	0.85	0.87	0.905
	F measure	0.64	0.84	0.86	0.89	0.924
	Accuracy	0.62	0.86	0.87	0.895	0.936
	ACE	2.15	1.89	1.56	1.05	0.95
	CE	0.21	0.18	0.12	0.09	0.05
IA	Precision	0.59	0.74	0.76	0.81	0.88
	Recall	0.36	0.75	0.76	0.815	0.85
	F measure	0.49	0.72	0.73	0.82	0.88
	Accuracy	0.39	0.74	0.76	0.83	0.915
	ACE	1.98	1.56	1.05	0.98	0.95
	CE	0.23	0.19	0.16	0.13	0.09
IS	Precision	0.36	0.64	0.72	0.79	0.82
	Recall	0.43	0.65	0.71	0.76	0.835
	F measure	0.45	0.62	0.73	0.78	0.846
	Accuracy	0.42	0.65	0.71	0.79	0.827
	ACE	1.85	1.36	1.25	0.85	0.849
	CE	0.25	0.235	0.21	0.19	0.16

view weights $\xi = 1$, $\xi = 32$, for TW-K means, QNSTWC, KICASDSTWC with incomplete view (ICV) $\xi = 1$, $\xi = 32$ and NSTVWC with incomplete view (ICV) $\xi = 1$, $\xi = 32$, results. It also shows that proposed NSTVWC with incomplete view (ICV) achieves higher clustering accuracy, as proposed work additionally measure the values to differentiate the categorical and numerical mixed data in efficient manner. In order to perform the measuring the clustering accuracy used Precision, Recall, F-measure, accuracy and average cluster entropy to evaluate the results.

Precision: Precision is calculated as the fraction of correct objects among those that the algorithm consideredto fit into the relevant cluster.

Recall: Recall is the fraction of actual objects that were identified.

F-measure: F-measure is the harmonic mean of precision and recall and accuracy is the proportion of correctly clustered objects.

The results of the different clustering methods with the above mentioned parameter results are shown in the Table 1. The performance comparison results of the proposed QNSTWC shows higher Precision, Recall, F measure and average accuracy, since the weight and centroid values are automatically calculated instead of using fixed values.

Average Cluster Entropy (ACE): Is based on the impurity of a cluster given the true classes in the data. If p_{ij} be the fraction of class j in obtained cluster i, N_i be the size of cluster i and N be the total number of examples, then the average cluster entropy is defined as:

$$E = \sum_{i=1}^{K} \frac{N_i(-\sum_j p_{ij} \log(p_{ij}))}{N}$$

where K is the number of clusters.

Table 2. summarizes the total clustering results. From theseresults, NSTVWC-means significantly out other four algorithms in almost all results,especially on the Multiple Features and Internet sets. Although NSTVWC -means is an extension to TW-kmeans,introduction of weights on views enhanced itsresults. WCMM produced the worst results on all threedata sets. One of the most important observation is the measurement of the clustering error by underlying the outputs into various clusters. In final proposed NSTVWC work produces less clustering error when compare to existing traditional methods, as proposed NSTVWC work is easily applicable to mixed types of attribute data.

CONCLUSION

This paper presents a novel multiview data clustering method for mixed numeric and categorical data attributes with complete and incomplete dataset. To support mixed numeric and categorical data attributes, it values are measured based on the Term Frequency (TF) -Inverse Document Frequency (IDF), Kullback -Leiber Divergence (KLD) and Shannon Entropy (SE) metrics. Before that initially the incomplete dataset samples are converted into the complete view dataset by using MICA. After this process the original dataset samples are divided into categorical and numerical dataset and measure values. The proposed NSTVWC framework then fuzzy centroid values are learned and optimized using SSD. Given multiple-view data, compute weights for views and individual variables simultaneously using FAFS. In order to reduce the complexity in the subspace clustering method SVD is proposed with ALCS methods for probability distribution optimization. Finally the clustering results of numerical and categorical dataset are combined as categorical dataset to get ultimatemultiview clustering results. The proposed system have been experimented to three dataset namely Multiple Features (MF),Internet Advertisement (IA) and Image Segmentation (IS), capabilities of different existing clustering methods and characteristics of three types of dataset could be fully measured based on the clustering parameters such as Precision, Recall, F measure and accuracy. It shows that proposed NSTVWC framework achieves enhanced clustering results than the existing conventionalclustering methods. In the future, combine the two-level variable weighting method with other techniques such as fuzzy techniques, semi-supervisedtechniques etc. so as to apply our method to more applications. Furthermore, approaches that can automatically group variables in the clustering process will examined.

REFERENCES

1. Kamalika Chaudhuri, Sham M. Kakade, Karen Livescu and Karthik Sridharan, 2009. Multi-view Clustering via Canonical Correlation Analysis. In International Conference on Machine Learning.
2. Zhao Haitao, 2006. Combining labeled and unlabeled data with graph embedding. Neurocomputing, 69(16): 2385-2389.
3. Nguyen Duc Thang, Lihui Chen and Chee Keong Chan, 2012. Clustering with Multiviewpoint-Based Similarity Measure. IEEE Transactions on Knowledge and Data Engineering, 24(6): 988-1001.
4. Kumar, A., P. Rai and H. Daume, 2011. III. Co-regularized multi-view spectral clustering. In NIPS, pp: 1413-1421.
5. Blaschko, M. and C. Lampert, 2008. Correlational spectral clustering. In CVPR, pp: 1-8.
6. Bruno, E. and S. Marchand-Maillet, 2009. Multiview clustering: a late fusion approach using latent models. In SIGIR, pp: 736-737.
7. Greene, D. and P. Cunningham, 2009. A matrix factorization approach for integrating multiple data views. In ECML PKDD, pp: 423-438.
8. Deng, Z., K. Choi, F. Chung and S. Wang, 2010. Enhanced Soft Subspace Clustering Integrating Within-Cluster and Between- Cluster Information, Pattern Recognition, 43(3): 767-781.
9. Cheng, H., K.A. Hua and K. Vu, 2008. Constrained Locally Weighted Clustering", Proceedings of VLDB Endowment, 1: 90-101.
10. Gretton, Arthur, *et al*., 2005. Measuring statistical dependence with Hilbert-Schmidt norms. Algorithmic learning theory. Springer Berlin Heidelberg.
11. Roul, R.K., O.R. Devanand and S.K. Sahay, 2014. Web document clustering and ranking using Tf-Idf based Apriori Approach.
12. Polani Daniel, 2013. Kullback-Leibler Divergence. Encyclopedia of Systems Biology, pp: 1087-1088.
13. Yang, M.S., Y.H. Chiang, C.C. Chen and C.Y. Lai, 2008. A fuzzy k-partitions model for categorical data and its comparison to the GoM model. Fuzzy Sets and Systems, 159(4): 390-405.
14. Wen, G.K., M. Mamat, I. Bin Mohd and Y. Dasril, 2012. A Novel of Step Size Selection Procedures for Steepest Descent Method. Applied Mathematical Sciences, 6(51): 2507-2518.
15. Kleibergen, F. and R. Paap, 2006. Generalized reduced rank tests using the singular value decomposition. Journal of Econometrics, 133(1): 97-126.
16. Peng, Y., 2011. An improved artificial fish swarm algorithm for optimal operation of cascade reservoirs. Journal of Computers, 6(4): 740-746.
17. Neshat, M., G. Sepidnam, M. Sargolzaei and A.N. Toosi, 2012. Artificial fish swarm algorithm: a survey of the state-of-the-art, hybridization, combinatorial and indicative applications. Artificial Intelligence Review, pp: 1-33.
18. Frank, A. and A. Asuncion, 2010. UCI Machine Learning Repository," http://archive.ics.uci.edu/ml, 2010.
19. Chen, X., X. Xu, J.Z. Huang and Y. Ye, 2013. TW-$(k) $-Means: Automated Two-Level Variable Weighting Clustering Algorithm for Multiview Data. Knowledge and Data Engineering, IEEE Transactions on, 25(4): 932-944.
20. Tzortzis, G. and C. Likas, 2010. Multiple View Clustering Using a Weighted Combination of Exemplar-Based Mixture Models, IEEE Transaction on Neural Networks, 21(12): 1925-1938.