DOI: 10.21917/ijsc.2016.0158

AN EFFECTIVE MULTI-CLUSTERING ANONYMIZATION APPROACH USING DISCRETE COMPONENT TASK FOR NON-BINARY HIGH DIMENSIONAL DATA SPACES

L.V. Arun Shalin[1] and K. Prasadh[2]
[1]*Department of Computer Science and Engineering, Manonmanium Sundaranar University, India*
E-mail: arunshalin7@gmail.com
[2]*Mookambika Technical Campus, India*
E-mail: ksprasadh@gmail.com

Abstract

Clustering is a process of grouping elements together, designed in such a way that the elements assigned to similar data points in a cluster are more comparable to each other than the remaining data points in a cluster. During clustering certain difficulties related when dealing with high dimensional data are ubiquitous and abundant. Works concentrated using anonymization method for high dimensional data spaces failed to address the problem related to dimensionality reduction during the inclusion of non-binary databases. In this work we study methods for dimensionality reduction for non-binary database. By analyzing the behavior of dimensionality reduction for non-binary database, results in performance improvement with the help of tag based feature. An effective multi-clustering anonymization approach called Discrete Component Task Specific Multi-Clustering (DCTSM) is presented for dimensionality reduction on non-binary database. To start with we present the analysis of attribute in the non-binary database and cluster projection identifies the sparseness degree of dimensions. Additionally with the quantum distribution on multi-cluster dimension, the solution for relevancy of attribute and redundancy on non-binary data spaces is provided resulting in performance improvement on the basis of tag based feature. Multi-clustering tag based feature reduction extracts individual features and are correspondingly replaced by the equivalent feature clusters (i.e.) tag clusters. During training, the DCTSM approach uses multi-clusters instead of individual tag features and then during decoding individual features is replaced by corresponding multi-clusters. To measure the effectiveness of the method, experiments are conducted on existing anonymization method for high dimensional data spaces and compared with the DCTSM approach using Statlog German Credit Data Set. Improved tag feature extraction and minimum error rate compared to conventional anonymization methods are demonstrated with experiments.

Keywords:

High-Dimensional Data Space, Data points, Non-Binary Database, Quantum Distribution, Dimensionality Reduction

1. INTRODUCTION

With the recent improvements in clustering algorithms four types are approximately separated namely projection clustering, hierarchical clustering, density-based clustering and subspace clustering algorithms. Different types of algorithm as mentioned investigate the clusters in lower-dimensional projection of the original data. It is normally favored when dealing with information that is high dimensional. Motivated by the fact of high dimension which has more dimensions leads to the so called curse of dimensionality with which the performance of usual machine learning algorithms becomes impaired. This is frequently due to two types of pervasive effects such as empty space incident and deliberation of distances.

The term ‘curse of dimensionality’ refers to the fact that all high dimensional data sets tend to be sparse, because the number of points necessary to symbolize any distribution grows exponentially with the number of dimensions. This results in bad estimates for high-dimensional data resulting in complexity over density based approaches on non- binary database. The latter is a rather counterintuitive property of high dimensional data point representations, where all distances between data points tend to turn out to be harder to differentiate with the increase in dimensionality which is omnipresent and copious.

Novel anonymization methods for sparse high-dimensional data [1] were based on estimated Nearest Neighbor (NN) search in high-dimensional spaces, which was evaluated using Locality Sensitive Hashing (LSH). The data transformation involved extracts the establishment using the underlying reduction into a band matrix and gray encoding-based sorting. These band matrixes and gray encoding made the establishment of anonymization in groups resulting in lesser information loss with the help of an efficient linear-time heuristic but problem related to non-binary databases was not solved. Anonymization methods for sparse high-dimensional data do not use dimensionality reduction techniques for more effectual anonymization.

The idea of selecting subset of good features with high variance and feature subset selection are proved to be certain efficient methods for dimensionality reduction. With the selection of good feature, the irrelevant data are removed that increases the accuracy related to learning and maximizing the comprehensibility for non-binary database. The feature subset selection methods for non-binary database are divided into four types namely, extensive category explicitly embedded, wrapper, filter, and hybrid techniques. The embedded methods integrate feature selection as a part of the training process and are usually precise, and therefore proved to be more efficient than the other three types.

On the basis of the aforementioned techniques and methods applied, the proposed work uses Discrete Component Task Specific Multi-Clustering (DCTSM) approach for non-binary high dimensional data spaces to improve accuracy. DCTSM first clusters different types of pertinent attributes to recognize the constituent records of it. The problem of projected clustering is addressed by identifying the clusters and their attributes are extracted using Statlog German credit dataset. Subsequently, discrete dimensional projection clusters using the quantum distribution model are applied that considers the problem related to attribute relativity and redundancy.

DCTSM approach offers a multi-cluster formation based on objective function and evolve a discrete dimensional projection clusters. Finally, dimensionality of the data is reduced by ignoring the lower Eigen value components. On the other hand, DCTSM approach uses the class information to perform a projection of the features which best separate two or more classes.

Experiments using Statlog German Credit Data Set extracted from UCI repository confirm that, the DCTSM approach facilitates higher level of accuracy and also multi-clustering process is considerably efficient. Empirical studies show that the application of DCTSM approach improves the level of accuracy and minimizes the error rate when compared to significantly more efficient state of art technique. The contribution of Discrete Component Task Specific Multi-Clustering (DCTSM) approach on non-binary database for dimensionality reduction mining includes the following:

1. To identify sparseness degree of dimensions using Discrete Component Task Specific Multi-Clustering (DCTSM) approach
2. To provide solution for relevancy of attribute and redundancy on non-binary data spaces
3. To reduce dimensionality on the basis of tag based feature
4. To apply multi-clusters instead of individual tag features and then during decoding the individual features are replaced by the corresponding multi-clusters to reduce the error rate.

The paper is structured as follows: Section 2 provide the related work and motivation behind our proposal. Section 3 describes the Discrete Component Task Specific Multi-Clustering approach. Section 4 demonstrates the experimental parameter with data extracted from Statlog German Credit Data Set. Section 5 details the performance with resultant table and graph. Finally the concluding section is explained in section 6.

2. RELATED WORKS

Clustering has been considered as one of the significant and valuable method in the field of data mining. Fuzzy similarity based self-constructing algorithm [2] exactly extracted single feature for every cluster. The feature being extracted was equivalent to a cluster that represented weighted mixture of words restricted in a specific cluster. With the introduction of fuzzy similarity based self-constructing algorithm, the derived membership functions were equivalent in a direct way and explained suitably for the real allocation of the training data.

DENsity Conscious Subspace clustering [8] followed a divide-and-conquer scheme to competently determine clusters that referred to the phenomenon that the region density varied in different subspace cardinalities. Some of the existing clustering algorithms were considered to be incapable if the required similarity measure calculated was present between the data points in high-dimensional space. To deal with the difficulty related to high-dimensional space, a number of projected clustering algorithms were developed in [3]. The clustering of high-dimensional data was not reviewed by recent researchers which also included kernel mappings and shared-neighbor clustering. The main disadvantage was that it only detected hyperspherical clusters, just as similar to that of the K-means. The work in [11] used hubs to repeatedly determine the number of clusters in the data.

The paper [16] analyzed and compared four data clustering algorithms, namely k-means algorithm, hierarchical clustering algorithm, self-organizing maps algorithm, and expectation maximization clustering algorithm. Concluding remarks stated that partitioning algorithms like k-means and EM were recommended for huge size dataset whereas the hierarchical clustering algorithms were recommended for small dataset.

Vector approximation employed scalar quantization, while the hyper plane bounds were better than MBR and MBS bounds explained in [5] but still movable compared to the true query-cluster distance. At the same time, the cluster-distance bounds were further optimized by introducing and optimizing the clustering algorithm. But many of the optimization methods faced certain level of difficulties whenever the clusters hide in subspaces with maximum level of low dimensionality. The projected clustering in [3] evaluated the space between the values of the attribute to obtain the relevance result, size of the selected attribute in the cluster and correlated the reports in the cluster. A novel approach of 2-D fractal dimension estimation based on contour wavelet transform was presented in [14] that used the properties of the autocorrelation function to improve the accuracy but at the cost of time.

An extensive analysis on enhancing the FCM clustering algorithm was developed in [17] to solve the difficulties related to K-means and FCM algorithms over high dimensional data. The algorithm presented called as Projected Clustering based on FCM Algorithm (PCFA) used the standard FCM clustering algorithm for sub-clustering high dimensional data into reference centroids and the reference values were fed as an input to the modified FCM algorithm.

Motivated from the fact of the manifold learning and L1-regularized methods for selecting the subset, Multi-Cluster Feature Selection (MCFS) [4] was used. Specifically, it selected those features that were in a way arranged according to the characteristics of multi-clustering. The equivalent optimization problem was capably solved as it only involved a sparse Eigen-problem and a L1-regularized least squares problem.

Data stream classification technique that integrates a narrative class detection mechanism into traditional classifiers [6], compliant very high false negative rate. With the introduction of concept-drift, feature detection problem occurred whenever the problem of fundamental data distribution occurred in the streams. But it failed in addressing the data stream classification problem beneath active feature sets.

The main task of sparse coding is to discover embedding of data by transferring the feature values based on subspace cluster membership [10]. The design of sparse coding by focusing on the identification of shared clusters between data when source and target data have dissimilar distribution did not offered a good starting point for addressing complex task multiple heterogeneous data sources. Probabilistic reverse nearest neighbor query that get back the objects from uncertain data had superior probability than a given threshold to be the RNN of an uncertain query object but not extensive to the probabilistic RkNN queries [13].

The work in [7] called as the feature relation network used text feature selection method based on the rule that considered semantic information were more appropriate for other classification problems when related to text. The drawback of the method was that the semantic information was obtainable but it could not be applied to additional potential feature relations. Other measurements, such as event frequency and various positional distributional features were added resulting in non-multidimensional FRN.

A multi-objective approach referred to as the elitist non-dominated sorting in genetic algorithm (NSGA-II) was presented in [15] and applied to identify high-level knowledge representation using IF-THEN rules with eight-dimensional search space. Three probable outcomes were labeled called as *very low, low* and *high* were chosen for determining the rule. Multi viewpoint-based similarity measure does not describe alternative forms as the relative resemblance did not used standard relative similarities according to the different viewpoints. It failed in applying criterion functions for hierarchical clustering algorithms [12]. Fast clustering-based feature selection algorithm (FAST) was divided into clusters using graph-theoretic clustering methods and related to target classes from each cluster to form a subset of features but it did not calculated different correlations [9]. To investigate dimensionality reduction on non-binary high dimensional data space, a technique named Discrete Component Task Specific Multi-Clustering approach is designed.

3. MATERIALS AND METHODS

In this section, an approach for effective reduction of dimensionality based on tags in the non-binary database called as Discrete Component Task Specific Multi-Clustering is presented. The design considerations of Discrete Component Task Specific Multi-Clustering starts with the attribute relevance analysis followed by the model of quantum distribution and tag specific feature extraction for precise dimensional projection. In addition an algorithm for DCTSM approach is presented. The architecture diagram of the DCTSM approach for reducing the dimensionality on non-binary high dimensional data spaces is illustrated in Fig.1.

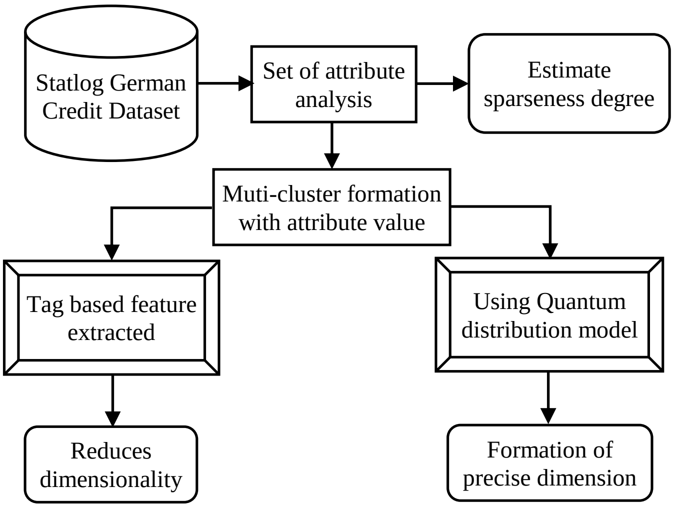
Statlog German Credit Dataset
Set of attribute analysis
Estimate sparseness degree
Muti-cluster formation with attribute value
Tag based feature extracted
Using Quantum distribution model
Reduces dimensionality
Formation of precise dimension

Fig.1. Diagrammatic representation of the DCTSM approach

The Fig.1 illustrates the diagrammatical representation of the DCTSM approach which is split into three phases. During the initial phase, analysis of attribute in the non-binary database takes place and the process of projecting clusters is deployed to identify sparseness degree even with small subset of dimensions. Once the analysis of non-binary database is accomplished, the second phase develops a multi-cluster dimension on quantum distribution for identifying the relevancy of attribute and redundancy on non-binary data spaces. Finally, the third phase extracts the feature based on the tag to reduce the dimensionality on non-binary data using the attributes extracted from the UCI repository, Statlog German Credit Dataset.

3.1 DCTSM ATTRIBUTE RELEVANCE ANALYSIS

To facilitate attribute relevance analysis, all proportions in a Statlog German credit dataset are identified which display certain level of cluster composition by determining dense regions and identify their position with the help of correct measurement. The fundamental hypothesis for attribute relevance analysis phase is that in scenarios including estimated clustering, a cluster contain appropriate proportions in which the projection of every point of the cluster is in an adequate number of further expected points, and this notion of "proximity" is qualified with all the proportions. The dimensions that are further symbolized are then used as the probable aspirants for appropriate proportions of the clusters.

In DCTSM approach, let us consider DB a Statlog German credit dataset of *n*-dimensional points, where the attributes is denoted by $A = \{A_1, A_2,\dots, A_n\}$ and a set of N non-binary data points, where $x_i = (x_{i1}, x_{i2},\dots x_{in})$. Each x_{ij} communicates with the assessment of data point x_i on attribute A_j where, x_i is termed as one dimensional point. Each data point x_i fits in either one estimated cluster or to the position of outliers OUT.

For a given number of clusters, 'n_c', acts as an input parameter, with an estimated cluster C_s, $s = 1, 2, \dots, n_c$ is termed as a pair (SP_s, SD_s) where SP_s is a division of non-binary data points of database and SD_s is a division of dimensions of set of attributes A, such that the projections of the data points in SP_s beside every dimension in SD_s are directly grouped. The proportions in SP_s are termed as significant dimensions for the cluster C_s. The last proportions, (i.e.,) A-SD_s are termed as inappropriate dimensions for the cluster C_s. The cardinality of the position SD_s is indicated by d_s where d_s, d and n_s specifies the cardinality of the set SP_s where $n_s < N$.

With the help of attribute significance analysis, the sparseness level y_{ij} are determined for diverse proportions. The sparseness level y_{ij} are specified by,

$$y_{ij} = \Sigma(r-c_{ij})4/k \tag{1}$$

where, $r \in p_{ij}\left(x_{ij}\right)$. The least assessment of y_{ij} signifies solid region whereas the highest assessment signifies thin region of non-binary data. Likewise different y_{ij} values are determined for different spatial images for different dimensions which facilitate assessment of y_{ij} for every image, i.e., it simply perceive the dense regions. The images with better values of y_{ij} indicate thin regions whereas the attribute with less values of y_{ij} signify the opaque regions.

3.2 QUANTUM DISTRIBUTION MODEL IN DCTSM APPROACH

Once the attribute relevance analysis is performed, the quantum distribution model is applied in DCTSM approach. Each cluster is precise in selecting the attribute value using DCTSM approach. Followed by this multiple clusters are formed for different types of attribute value and each attribute value of a cluster is a restricted mean chosen precisely from the domain. Each report in the multi-cluster then follow the precise attribute values according to the data error rate. Quantum distribution is optimized with the given constraints and with variables that need to be minimized or maximized using programming techniques.

The DCTSM approach necessitates quantum distribution model by simpler form so that available computational objective function approach is used on non-binary database. This leads to a natural criterion for selecting the most excellent precise attribute value. Statlog German credit dataset consists of a set of incoming reports which are denoted by $Y_1',\ldots Y_i'$.

Let us assume that the data point 'Y_i'' is received during the time stamp 'S_i' followed by the assumption that the discrete dimensionality of the non-binary database is 'h'. The 'h' dimensions of the report Y_i' are denoted by $\left(y_i^1,\ldots y_i^d\right)$. In addition, each non-binary data point has an error associated with dissimilar dimensions. The error associated with the k^{th} dimension for non-binary data point Y_i is denoted by point by $\Psi_k(Y_i')$.

Since different dimensions of data replicate diverse quantities, they correspond to different scales in DCTSM approach. In order to take the precise behavior of the different discrete dimensions into account, quantum distribution across different discrete dimensions is performed. DCTSM approach maintains the global statistics to compute global variances. These variances are used to scale the data over time with values.

In order to include the greater importance of recent data points in a developing stream, concept of an objective function $f(s)$ quantifies the relative importance of the different data points over time. The objective function is drawn from the range (0, 1), and serves as a quantize factor for the relative importance of a given data point with a decreasing function that represents the objective of importance of a data point over time.

The objective function of quantum distribution model for DCTSM approach is the exponential objective function on non-binary database. The exponential objective function $f(s)$ with parameter λ is defined as follows as a function of,

$$f(s) = 2 - \lambda.s(1) \tag{2}$$

The value of $f(s)$ reduces by a factor of 2 every $1/\lambda$ time units and corresponds to the half of the function $f(s)$.

3.3 TAG SPECIFIC FEATURE EXTRACTION

Finally in order to reduce the dimensionality, DCTSM approach extracts the feature extraction based on the tag, where several similar features are grouped together in order to reduce the feature dimension. As a result, feature reduction for individual feature is substituted by the equivalent feature tag clusters. During classification multi-clysters are used instead of the individual tag features and then through reverse process the individual tag features are replaced by the corresponding clusters.

To perform tag clustering on non-binary database, the tags are represented as vectors and the similarity between the vectors are computed. Two different vector representations are developed using DCTSM approach. A target tag is represented as a vector of other tag in its proximity. If all the tags in the lexicon are used during vector representation, the vector dimension becomes very high and the calculations are complicated. For efficient implementation, only the tags which occur in the context of the discrete component are considered.

Let us consider a tag with v_i. If v_i is the beginning tag of a discrete component then the most frequent tags are marked as Register_Prev that occur as v_{i-1} or v_{i-2} whereas if v_i is the last tag of a discrete component then the most frequent next tag in positions v_{i+1} or v_{i+2} are marked as Register _Next. Using the position specific tag lists, the items in Register_Prev and Register _Next are replaced using DCTSM approach. Let us further assume that a particular tag't' occurs 'n' times in the corpus. For each occurrence v_k of v, its previous tags v_{k-1} or v_{k-2} are checked to identify that if they match any element of Register_Prev. If there is a match, then the corresponding position of the vector is set to one and is set to zero to the other positions related of Register_Prev. Similarly the next tag v_{k+1} or v_{k+2} in Register _Next are checked and the values of the corresponding positions are evaluated. The final tag vector $\overset{\rightarrow}{v_k}$ is obtained by taking the average of the 'n' vectors corresponding to the 'n' occurrences of 'v'. It measures the similarity of the contexts of the occurrences of the tag't' in terms of the proximal words.

Discrete Component (DC) tasks the position of a context, performs tag selection and positions specific clusters using DCTSM approach. The two preceding and two following positions (i -1, I + 2, i + 1, i + 2) of a tag are considered, and corresponding to these positions, four different tags vectors are defined. Each vector is of dimension m + 1 corresponding to 'm' discrete component group (M_j). For a particular tag v_k, the fraction ($R_j(v_k)$) is measured for obtaining the total occurrences of the tag in a particular position belonging to a group M_j.

$$\text{DC} = \frac{\text{Occurrence of } v_i + pos \text{ position of DC}}{\text{Total occurrence of } v_i \text{ in quantity}} \tag{3}$$

where, pos denote a particular position like (+1), (+2). Four positions (-2, -1, +1 and +2) are considered and for each position different sets of tags are selected. The component of the tag vector $\overset{\rightarrow}{v_k}$ for the position corresponding to M_j is $R_j(v_k)$.

Once the vectors are obtained for non-binary database, these are multi-clustered and chosen using a training process on a validation set during non-binary database experiments. The multi-cluster seeds are selected arbitrarily and two types of vector similarity representation measures are used for clustering the vectors, namely cosine similarity and Euclidean distance. The DCTSM approach defined are applied on other high dimensional features space based on tags and also experimented with affix using the task specific approach.

3.4 ALGORITHM FOR DCTSM APPROACH

The algorithm given below describes the steps performed in Discrete Component Task Specific Multi-Clustering (DCTSM):

Input: Let A_j denote the attribute values of Statlog German Credit Data Set 'S'. $Q_{\min,}$ is minimum number of Quantum index of a selected attribute, 'MC' is multi-cluster k^{th} dimensions of dataset, t occurs 'n' times in the corpus.

Output: Attribute analysis with sparseness degree and Dimensionality reduction on non-binary database.

```
Begin
// Analysis of attribute on ‘$S$’
Compute the sparseness degree $y_{ij}$;
Normalize $y_{ij}$ in the interval [0, 1];
For $m$ = 1 to m_max do
    If ($m$ = = 1) then
        Estimate the parameters of the gamma distribution based on the probability
        Compute the value of sparseness using Eq.(1)
    End If
// recognize relevancy of attribute and redundancy
For each cluster $C$, SelectAttrisVal($C$, $Q_{\min}$)
BuildQDmodel($|B_{\min}|$, $Q_{\min}$)
    While (Quantizeresult)
        $MC1$ and $MC2$ are multiple clusters formed with objfunc
        Various attribute value forms the new cluster $C_n$
        $C_n$: = $MC1$, $MC2$,…, $MCn$
    End While
Update Quantize result
End for each
// Dimensionality reduction on multi-cluster
SelectAttrisVal($C_n$, $Q_{\min}$)
If (Tag‘t’)
Register_Prev vector representation on $v_{i-1}$ or $v_{i-2}$ tag set
Register _Next vector representation on $v_{i+1}$ or $v_{i+2}$ tag set
    If (position ‘pos’)
        Register_Prev vector representation on $v_{k-1}$ or $v_{k-2}$ tag set
        Register _Next vector representation on $v_{k+1}$ or $v_{k+2}$ tag set
        Tag vector $\overset{\rightarrow}{v_k}$ computed
    End If
Calculate DC using Eq.(3) to reduce feature tags
End If
End
```

The above algorithmic step is used to identify the sparseness degree and dimensionality reduction on non-binary database using sparseness degree y_{ij} using Eq.(1) and the parameters are evaluated. In DCTSM approach, relevancy of attribute and redundancy are identified using the quantum distribution model. The quantum model produces result using the objective function after selecting the appropriate attribute value. The quantized result is updated. Then, dimensionality reduction in the non-binary database based on the tag based feature is obtained. Register_Prev and Register _Next are two form of vector representation for the positions v_{i-1} or v_{i-2}, v_{i+1} or v_{i+2}, v_{k-1} or v_{k-2}, v_{k+1} or v_{k+2}. Finally, $\overset{\rightarrow}{v_k}$ is computed and DC is obtained using the above equation to reduce the dimensionality.

4. RESULTS ON DISCRETECOMPONENT TASK SPECIFIC MULTI-CLUSTERING

The main goal of the experiments presented here is to evaluate the capability of non-binary database to correctly identify dimensionality reduction in various simulation settings. Experimental evaluation is conducted to estimate the performance of the Discrete Component Task Specific Multi-Clustering Approach in high dimensional data spaces. The DCTSM approach is implemented in Java. The first phase analysis the factor involved with related to redundancy in non-binary database with the help of Statlog German Credit Data Set. With the outcomes of the first phase, the objective of the second process is to provide solution for relevancy of attribute and redundancy, whereas the third phase efficiently reduces the dimensionality on multi-cluster dimensions.

Statlog German Credit Data Set classifies the publicly available data by a set of attributes as good or bad acclaim risks. Also Statlog German Credit Data Set contents come up with a cost matrix. In cost matrix, rows represent the actual classification and the columns denote the predicted classification. Statlog German Credit Data Set holds 20 attributes with 1000 instances. Only discrete dimension of attributes is considered with each multi-cluster having different set of precise attribute value. An attribute is precise to a cluster if it helps to identify the member reports of it. This means the values at the precise attributes are distributed among certain specific values in the cluster, while the reports of other clusters are less likely to have such values. Determining multi-clusters and their precise attribute value from a Statlog German Credit dataset is known as the discrete dimensional projected clusters.

5. PERFORMANCE COMPARISON OF DISCRETE COMPONENT TASK SPECIFIC MULTI-CLUSTERING

Discrete Component Task Specific Multi-Clustering (DCTSM) approach is compared with the Anonymization methods for sparse high-dimensional data through locality-sensitive hashing (LSH) for measuring the accuracy, running time and error rate. Average Accuracy (AA) is needy on how data points (i.e.) information is collected, and is usually judged by comparing numerous capacity from the same or different multi-clustered data. The non-binary high dimensional database average accuracy is measured on dimensionality reduction.

$$\text{AA} = \left.\frac{\text{Number of correctly classified tags}}{\text{Number of tags on binary data}}\right\} \tag{4}$$

The running time of tag feature extraction for DCTSM is defined as the amount of time taken to extract the tags from the non-binary data points. It is measured in terms of milliseconds (ms). The Error Rate (ER) for DCTSM approach measures the number of errors separated by the whole number of transferred data points on non-binary database during specific time interval.

$$\text{ER} = \frac{\text{Number of Errors}}{\text{Total Number of Transferre d error points}} \tag{5}$$

The error rate is measured as an estimate of the error probability and precise for a long time interval and a high number of errors. ER is measured in terms of percentage (%).

The Table.1 given below shows the average accuracy measured with respect to novel instances in the range of 100 to 700. Comparisons are made with the existing LSH technique to measure the effectiveness of the proposed DCTSM approach.

Table.1. Novel instances vs. Average Accuracy

Novel Instances	**Average Accuracy (%)**	
	LSH Technique	**DCTSM Approach**
100	82	88
200	84	90
300	83	91
400	87	93
500	88	95
600	85	92
700	86	95

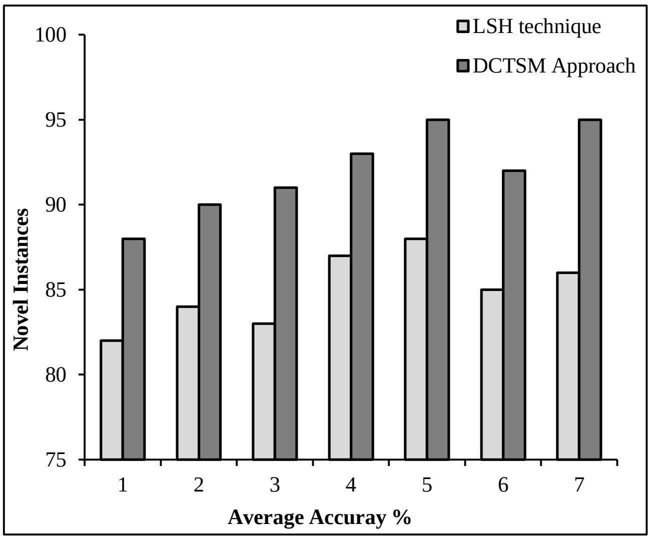

Fig.2. Measure of Average Accuracy

The Fig.2 describes the average accuracy based on the instances extracted from the Statlog German credit dataset. The accuracy is measured in terms of percentage (%) and instances ranges from 100 to 700. From the figure it is illustrative that the average accuracy is increased using the proposed DCTSM approach when compared to the existing LST technique. This is because the two vector representations, Register_Prev and Register_Next on the vectors v_{i-1} or v_{i-2}, v_{i+1} or v_{i+2}, v_{k-1} or v_{k-2}, v_{k+1} or v_{k+2} reduces the dimensionality reduction and improves the average accuracy. The average accuracy using DCTSM approach is improved from 5-10% when compared to the existing LSH technique anonymization method for high dimensional data spaces [1].

The Table.2 given below show the tag feature extraction running time with respect to the number of extracted tag features and elaborate comparisons are made with the existing LSH technique to measure the effect of running time using the proposed DCTSM approach.

Table.2. Number of extracted tag features vs. running time

No. of extracted tag features	**Tag feature Extraction Running time (ms)**	
	LSH Technique	**DCTSM Approach**
50	787	771
80	767	752
120	778	762
150	884	866
190	895	881
250	873	861
320	921	903

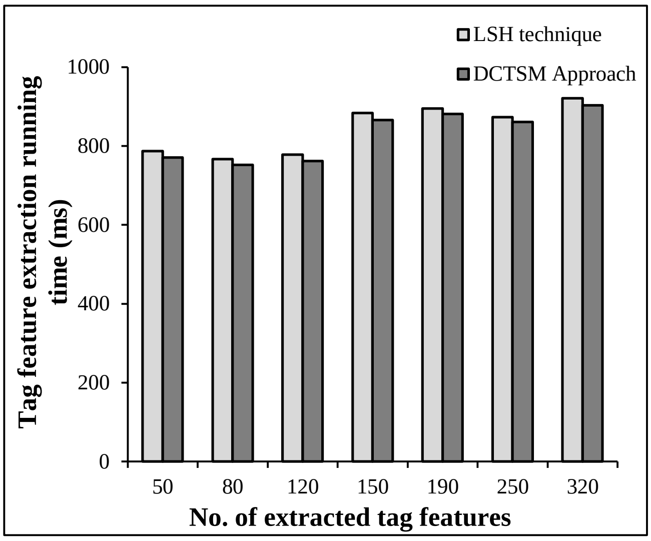

Fig.3. Measure of running time

The Fig.3 illustrates the tag feature extraction running time based on the extracted tag features number. The figure illustrates that the time taken to run using DCTSM approach is lesser than when compared to that using the LSH technique. This is because with the application of multi-clustered quantum distribution model and with the application of the exponential objective function *f*(*s*) in DCTSM approach the running time gets reduced in the proposed system when compared to that of the LSH technique which is measured in terms of milliseconds (ms). The running time is reduced from 2-5% in DCTSM approach when compared to that using the LSH technique.

Table.3. Non-binary data dimensions vs. Error Rate

Non-binary data dimensions	Error rate (%)	
	LSH Technique	DCTSM Approach
50	6.1	5.6
100	10.6	9.3
150	13.5	12.1
200	22.8	19.3
250	29.1	25.5
300	29.5	25.9
350	32.3	27.9

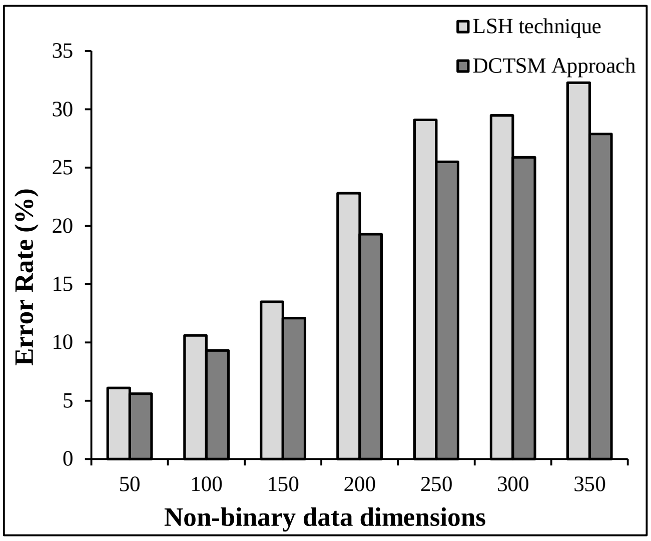

Fig.4. Measure of Error rate

The Table.3 and Fig.4 illustrates the error rate measured based on the non-binary data dimensions and measured in terms of the percentage (%). The error rate in DCTSM approach is reduced to 10-20% using the discrete component (DC). Each vector in DC is of dimension m + 1 corresponding to 'm' discrete component group (M_j). In addition with the introduction of indecisive DCTSM algorithm the possibility density function characterized the underlying behavior and decreases the error rate.

To conclude with, dimensionality is reduced in the non-binary high dimensional data space points. The ratio between the number of tag occurrences and its total number of occurrences in the quantity are used as a metric of DC task specific tag selection. Some tags occur only once in the training quantity and considered as a context tag. These tags with higher discrete component find place in the important tag list. But these tags which are not much frequent are removed for further reduction and improve the accuracy rate.

6. CONCLUSION

The performance of Discrete Component Task Specific Multi-Clustering approach initially analysis the attribute present in the non-binary Statlog German credit dataset and the process of projecting clusters identify the degree of sparseness related to dimensions of data. Then multi-cluster dimension provide relevancy of attribute and redundancy on non-binary data spaces using quantum distribution. As a final point, multi-clustering tag based feature reduction takes individual features and is replaced by the equivalent feature clusters for dimensionality reduction. Recognition of an appropriate position based on tag feature in DCTSM approach is very imperative that provide the best performance for dimensionality reduction. Generally the context and affix information efficiently identifies the discrete component from a tag. The DCTSM approach gives up results with 7.05% improved accuracy, minimal time taken for tag feature extraction, and lesser error rate. The effectiveness of the feature reduction based on tag approach is carried out in better way than the analogous attribute sets.

REFERENCES

[1] G. Ghinita, P. Kalnis and Yufei Tao, "Anonymous Publication of Sensitive Transactional Data", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 23, No. 2, pp. 161-174, 2011.

[2] Jung-Yi Jiang, Ren-Jia Liou and Shie-Jue Lee, "A Fuzzy Self-Constructing Feature Clustering Algorithm for Text Classification", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 23, No. 3, pp. 335-349, 2011.

[3] M. Bouguessa and Shengrui Wang, "Mining Projected Clusters in High-Dimensional Spaces", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 21, No. 4, pp. 507-522, 2009.

[4] Deng Cai, Chiyuan Zhang and Xiaofei He, "Unsupervised Feature Selection for Multi-Cluster Data", *Proceedings of 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 333-342, 2010.

[5] S. Ramaswamy and K. Rose, "Adaptive Cluster Distance Bounding for High-Dimensional Indexing", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 23, No. 6, pp. 815-830, 2011.

[6] M.M. Masud, Jing Gao, L. Khan, Jiawei Han and B. Thuraisingham, "Classification and Novel Class Detection in Concept-Drifting Data Streams under Time Constraints", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 23, No. 6, pp. 859-874, 2011.

[7] A. Abbasi, S. France, Zhu Zhang and Hsinchun Chen, "Selecting Attributes for Sentiment Classification Using Feature Relation Networks", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 23, No. 3, pp. 447-462, 2011.

[8] Yi-Hong Chu, Jen-Wei Huang, Kun-Ta Chuang, De-Nian Yang and Ming-Syan Chen, "Density Conscious Subspace Clustering for High-Dimensional Data", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 22, No. 1, pp. 16-30, 2010.

[9] Qinbao Song, Jingjie Ni and Guangtao Wang, "A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 25, No. 1, pp. 1-14, 2013.

[10] B. Quanz, Jun Huan and M. Mishra, "Knowledge transfer with Low-Quality Data: A Feature Extraction Issue", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 24, No. 10, pp. 769-779, 2012.

[11] N. Tomasev, M. Radovanovic, D. Mladenic and M. Ivanovic, "The Role of Hubness in Clustering High-Dimensional Data", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 26, No. 3, pp. 739-751, 2013.

[12] Duc Thang Nguyen, Lihui Chen and Chee Keong Chan, "Clustering with Multiviewpoint-Based Similarity Measure", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 24, No. 6, pp. 988-1001, 2012.

[13] M.A. Cheema, Xuemin Lin, Wei Wang, Wenjie Zhang and Jian Pei, "Probabilistic Reverse Nearest Neighbor Queries on Uncertain Data", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 22, No. 4, pp. 550-564, 2010.

[14] M. Yazdi and A.G. Mahyari, "A New 2-D Fractal Dimension Estimation based on Contourlet Transform for Texture Segmentation", *The Arabian Journal for Science and Engineering*, Vol. 35, No. 13, pp. 293-317, 2010.

[15] Z.M. Nopiah, M.H. Osman, S. Abdullah and M.N. Baharin, "Application of a Multi-Objective Approach and Sequential Covering Algorithm to the Fatigue Segment Classification Problem", *Arabian Journal for Science and Engineering*, Vol. 39, No. 3, pp. 2165-2177, 2014.

[16] Osama Abu Abbas, "Comparisons between Data Clustering Algorithms", *The International Arab Journal of Information Technology*, Vol. 5, No. 3, pp. 320-325, 2008.

[17] Ilango Murugappan and Mohan Vasudev, "PCFA: Mining of Projected Clusters in High Dimensional Data Using Modified FCM Algorithm", *The International Arab Journal of Information Technology*, Vol. 11, No. 2, pp. 168-177, 2012.