World Journal of Engineering Research and Technology

WJERT

www.wjert.org

SJIF Impact Factor: 5.924

STUDY ON DIMENSIONALITY REDUCTION USING COPULA APPROACH IN DATA MINING

Sumaiya Maryam*[1] and Sriram Yadav[2]

[1]M. Tech. Scholar, MITS, Bhopal.

[2]A. P., CSE, MITS Bhopal.

Article Received on 17/12/2019 **Article Revised on 07/01/2020** **Article Accepted on 28/01/2020**

***Corresponding Author**
Sumaiya Maryam
M. Tech. Scholar, MITS, Bhopal.

ABSTRACT

Copula Approach in Data Mining is Sampling-based dimensionality reduction technique eliminating linearly redundant combined dimensions, providing a convenient way to generate correlated multivariate random variables, maintaining the integrity of the original information, reducing the dimension of data space without losing important information. The recent trends in collecting huge and diverse datasets have created a great challenge in data analysis. One of the characteristics of these gigantic datasets is that they often have significant amounts of redundancies. The use of very large multi-dimensional data will result in more noise, redundant data, and the possibility of unconnected data entities. To efficiently manipulate data represented in a high-dimensional space and to address the impact of redundant dimensions on the final results, a new technique for the dimensionality reduction using Copulas and the LU-decomposition (Forward Substitution) method has been proposed. The proposed method is compared favorably with existing approaches on real-world datasets: Diabetes, Waveform, two versions of Human Activity Recognition based on Smartphone, and Thyroid Datasets taken from machine learning repository in terms of dimensionality reduction and efficiency of the method, which are performed on statistical and classification measures.

KEYWORDS: Data mining, Data pre-processing, Multi-dimensional Sampling, Copulas, Dimensionality reduction, Decomposition.

1. INTRODUCTION

High-dimensionality reduction has emerged as one of the significant tasks in data mining applications and has been effective in removing duplicates, increasing learning accuracy, and improving decision making processes. High-dimensional data are inherently difficult to analyze, and computationally intensive for many learning algorithms and multi-dimensional data processing tasks. In this paper, a new approach has been proposed which reduces the size of the data by eliminating redundant attributes based on sampling methods. The proposed technique is based on the theory of Copulas and the LU-decomposition method (Forward Substitution). A Copula provides a suitable model of dependencies to compare well-known multivariate data distributions to better distinguish the relationship between the data. The detection of dependencies is thereafter used to determine and to eliminate the irrelevant and/or redundant attributes. The critical issues for the majority of dimensionality reduction studies based on sampling and probabilistic representation are how to provide a convenient way to generate correlated multivariate random variables without imposing constraints to specific types of marginal different assumptions about the data distribution; to specify the dependencies between the random variables; to reduce the redundant data and remove the variables which are linear combinations of others; and to maintain the integrity of the original information. For these reasons, the main goal of this paper is to propose a new method for dimensionality reduction based on sampling methods addressing the challenges mentioned before. The paper uses both statistical and classification methods to improve the efficiency of the method. In the statistical part, a standard deviation of the final dimensionality reduction results will be computed for all databases with each dimensionality reduction method studied (PCA, SVD, SPCA, and proposed approach). However, the effectiveness of dimensionality reduction in classification methods will be improved using from one side the full set of dimensions and from the other the reduced set of provided data in terms of precision and recall, for the three classifiers: Artificial Neural Network (ANN), k-nearest neighbors (k-NN), naive Bayesian.

Dimensionality reduction technique is based on probabilistic and sampling models; therefore, one needs to.

The following table gives basic notations used throughout this paper

Table 1: Basic Notations.

Primitive	Definition
X	$n \times m$ data matrix (random variable).
Xi	i^{th} row of the matrix X.
Xj	j^{th} column of the matrix X.
$Fj(.)$	CDF of the j^{th} column.
$fj(.)$	PDF of the j^{th} column.

Let f be the Probability Density Function (PDF) of a random variable X. The probability distribution of X consists in calculating the probability P(X1 _ x1;X2 _ x2; :::;Xm _ xm), for all(X1; :::;Xm) belongs to Rm. It is completely specified by the CDF F which is defined in (Rubinstein &Kroese, 2011) as follows:

F(x1; x2; :::; xm) = P(X1 _ x1;X2 _ x2; :::;Xm _

xm) ----------(1)

1.1 Random Variable Generation

The problem of generating a sample from a one-dimensional cumulative distribution function *CDF* by calculating the inverse transform sampling. To illustrate the problem, let X be a continuous random variable with a *CDF* $F(x) = P[X \leq x]$, and U be a continuous uniform distribution over the interval [0, 1]. The transform $X = F^{-1}(U)$ denotes the inverse transform sampling function of a given continuous uniform variable $U = F(X)$ in [0, 1], where $F^{-1}(u) = \{min\ x, F(x) \geq u\}$ (Rubinstein & Kroese, 2011). So the simple steps used for generating a sample $X \sim F$ are given as follows (Rubinstein & Kroese, 2011):

1. Generate $U \sim U[0,1]$;
2. Return $X = F^{-1}(U)$.

The usual problem is how to combine one- dimensional distribution functions to form multivariate distributions and how to estimate and simulate their density $f(x1, x2, xm)$ to obtain the required number of random samples of $Xi, i=1,...,m$, especially in high- dimensional spaces. This problem will be explained in the following section.

1.2 Modeling with Copulas

The first usage of Copulas is to provide a convenient way to generate correlated multivariate random variable distributions and to present a solution for the difficulties of transformation of the density estimation problem.

C	Gaussian Copula of the matrix X.To illustrate
c	Density associated with C. transformations
Cij	Empirical Copula of the matrix X.random variabl
Σ	Correlation matrix of C. their CDF, in
Xt	Transposed matrix of X. Th distributed va
vij	Value of the ith row and j column.

The problem of invertible of *m*-dimensional continuous es $X1$ ..., Xm according to to *m* independently uniformly- riables $U1 = F1(X1)$, $U2 =$

$F2(X2)$, ..., $Um = Fm(Xm)$, let $f(x1, x2, ..., xm)$ be the probability density function of $X1$, ..., Xm, and let $c(u1, u2, ..., um)$ be the joint probability density function of $U1$, $U2$, ..., Um. In general, the estimation of the probability density function f $(x1, x2,..., xm)$ can provide a nonpara- metric form (unknown families of distributions). In this case, we estimate the probability density function c $(u1, u2,..., um)$ of $U1$, $U2$, ..., Um instead of that $X1$, ..., Xm to simplify the density estimation problem, and then simulate it to achieve the ran- dom samples $X1$, ..., Xm by using the inverse transformations $Xi = F^{-1}(Ui)$.

Sklar's Theorem showed that there exists a unique m-dimensional Copula C in $[0; 1]^m$ with standard uniform marginal distributions U1,…..Um. (Nelsen, 2007) states that every distribution func- tion F with margins $F1$, ..., Fm can be written $\forall(X1,..., Xm) \in IR^m$ as:

$$F(X1,..., Xm) = C(F1(X1),..., Fm(Xm)). \quad (2)$$

To evaluate the suitability of a selected Copula with estimated parameter and to avoid the introduction of any assumptions on the distribution $Fi(Xi)$, one can utilize an empirical *CDF* of a marginal $Fi(Xi)$, to transform *m* samples of *X* into *m* samples of *U* . An empirical Copula is useful for examining the dependence structure of multivariate random vectors. Formally, the empirical Copula is given by the following equation:

$$C = 1/m\,(\sum^{n} \quad I(vkj \leq vij) \quad (6)$$

Where the function $I(arg)$ is the indicator function, which equals 1 if *arg* is true and 0 otherwise. Here, *m* is used to keep the empirical *CDF* less than 1, where *m* is the number of observations. In the following, we will focus on the Copula that results from a standard multivariate Gaussian Copula.

1.3 Gaussian Copula

The difference between the Gaussian Copula and the joint normal *CDF* is that the Gaussian Cop- ula allows to have different marginal *CDF* types from the joint distribution (Nelsen, 2007). How- ever, in probability theory and statistics, the multi- variate normal distribution is a generalization of the one- dimensional normal distribution. The Gaus- sian Copula is defined as follows:

$C(\Phi(x1),\ldots.. \Phi(xm)) =$

$$1/\sqrt{|\sum|}\exp(-1/2\, X^t(\Sigma^{-1}-I)X) \quad (4)$$

Where $\Phi(xi)$ is the CDF standard Gaussian distribution of $fi(x\boldsymbol{i})$, i.e., $Xi \sim N(0, 1)$, and Σ is the correlation matrix. The resulting Copula $C(u1,..., um)$ is called Gaussian Copula. The density associated with $C(u1, ..., um)$ is obtained with the following equation:

$$c(u1,\ldots., um) =1/\sqrt{|\sum|}\exp[-1/2\xi^t(\sum^{-1}-I)\xi] \quad (5)$$

where ui $=\Phi(xi)$, and $\xi = (\Phi^{-1}(u1),\ldots\ldots\ldots\ldots\ldots,$

$\Phi^{-1}(um))^T$.

1.4 Dependence and Rank Correlation

Since the Copula of a multivariate distribution describes its dependence structure, it might be ap- propriate to use measures of dependence which are Copula-based. The Pearson correlation measures the relationship $\Sigma = cov(X_i, X_j)/(\sigma_X\boldsymbol{i}\ \sigma_X\boldsymbol{j}\ )$ where $cov(Xi, Xj)$ is the covariance of $X\boldsymbol{i}$ and $X\boldsymbol{j}$ while $\sigma_X\boldsymbol{i}, \sigma_X\boldsymbol{j}$ are the standard deviations of X_i and $X\boldsymbol{j}$. Kendall rank correlation (also known as Kendall's coefficient of concordance) is a non- parametric test that measures the strength of dependence between two random samples Xip ;Xip of n observations. The notion of concordance can be defined by the following equation:

$$\tau = P\,[(X^i - X^j)(X^i\mathrm{t} - X^j) > 0] - P$$

$$[(X^i - X^j)\,(X^i\mathrm{t} - X^j) < 0] \quad (6)$$

For the Gaussian Copula, Kendall's τ can be calculated as follows: by linear combinations of the original variables and are uncorrelated. Several models and techniques for data reduction based on PCA have been proposed (Sasikala & Balamuru- gan, 2013). (Zhai et al., 2014) proposed a max- imum likelihood approach to the multi-size PCA problem. The covariance based approach was extended to estimate errors within the resulting PCA decomposition. Instead of making all the vectors of fixed size and then computing a covariance matrix, they

directly estimate the covariance matrix from the multi-sized data using nonlinear optimization. (Kerdprasop et al., 2014) studied the recognition accuracy and the execution times of two different statistical dimensionality reduction methods ap- plied to the biometric image data, which are: PCA and linear discriminant analysis (LDA). The learning algorithm that has been used to train and recognize the images is a support vector machine with linear and polynomial kernel functions. The main drawback of reducing dimensionality with PCA is that it can only be used if the original variables are correlated, and homogeneous, if each component is guaranteed to be independent and if the dataset is normally distributed. If the original variables are not normalized, PCA is not effective.

$\tau = 2/\pi \arcsin \Sigma XiXj(7)$

2 LITERATURE REVIEW

2.1 Linear dimensionality reduction

Principal Component Analysis (PCA) is a well established method for dimensionality reduction. It derives new variables (in decreasing order of im- portance) that are linked.

The Sparse Principal Component Analysis (SPCA) (Zou et al., 2006) is an improvement of the classical method of PCA to overcome the problem of correlated variables using the LASSO technique. LASSO is a promising variable selection technique, producing accurate and sparse models. SPCA is based on the fact that PCA can be written as a regression problem where the response is predicted by a linear combination of the predictors. There- fore, a large number of coefficients of principal components become zero, leading to a modified PCA with sparse loading. Many studies on data reduc tion based on SPCA have been presented. (Shen & Huang, 2008) proposed an iterative algorithm named sparse PCA via regularized SVD (sPCA- rSVD) that uses the close connection between PCA and singular value decomposition (SVD) of the data matrix and extracts the PCs through solving a low rank matrix approximation problem. (Bai et al., 2015) proposed a method based on sparse principal component analysis for finding an effective sparse feature principal component (PC) of multiple physiological signals. This method identifies an active index set corresponding to the non-zero entries of the PC, and uses the power iteration method to find the best direction.

Singular Value Decomposition (SVD) is a powerful technique for dimensionality reduction. It is a particular case of the matrix factorization approach and it is therefore also related to PCA. The key is- sue of an SVD decomposition is to find a lower dimensional feature space by

using the matrix product $U\ S\ V$, where U and V are two orthogonal matrices and S is a diagonal matrix with $m\ \ m$, $m\ n$, and $n\ n$ dimensions, respectively. SVD retains ×nly×$r\ \ n$ positive singular values of low effect to reduce the data, and thus S becomes a diagonal matrix with only r non-zero positive en- tries, which reduces the dimensions of × these three matrices to $m\ \ r$, $r\ \ r$, and $r\ \ n$, respectively. Many studies on data reduction have been presented which are built upon SVD, such as the ones used in (Zhang et al., 2010) and (Watcharapinchai et al., 2009). (Lin et al., 2014) developed a dimensionality reduction approach by applying the sparsified singular value decomposition (SSVD). Their paper demonstrates how SSVD can be used to identify and remove nonessential features in order to facilitate the feature selection phase, to analyze the application limitations and the computational complexity. However, the application of SSVD on large datasets showed a loss of accuracy and makes it difficult to compute the eigenvalue decomposition of a matrix product $A^T A$, where A is the matrix of the original data.

2.2 Nonlinear dimensionality reduction

A vast literature devoted to nonlinear techniques has been proposed to resolve the problem of dimensionality reduction, such as manifold learning methods, e.g., Locally Linear Embedding (LLE), Isometric mapping (Isomap), Kernel PCA (KPCA), Laplacian Eigenmaps (LE), and a review of these methods is summarized in (Gisbrecht & Hammer, 2015; Wan et al., 2016). KPCA (Kuang et al., 2015) is a nonlinear generalization of PCA in a high-dimensional kernel space constructed by using kernel functions. By comparing with PCA, KPCA computes the principal eigenvectors using the kernel matrix, rather than the covariance matrix. A kernel matrix is done by computing the in- ner product of the data points. LLE (Hettiarachchi & Peters, 2015) is a nonlinear dimensionality reduction technique based on simple geometric in- tuitions. This algebraic approach computes the low-dimensional neighborhood preserving embed- dings. The neighborhood is preserved in the embedding based on a minimizing cost function in in- put space and output space, respectively. Isomap (Zhang et al., 2016) explores an underlying manifold structure of a dataset based on the computation of geodesic manifold distances between all pairs of data points. The geodesic distance is determined as the length of the shortest path along the surface of the manifold between two data points. It first constructs a neighborhood graph between all data points based on the connection of each point to all its neighbors in the input space. Then, it estimates geodesic distances of all pairs of points by calculating the shortest

path distances in the neighborhood graph. Finally, multidimensional scaling (MDS) is applied to the arising geodesic distance matrix to find a set of low-dimensional points that greatly match such distances.

2.3 Sampling dimensionality reduction

Other widely used techniques are based on sampling. They are used for selecting a representative subset of relevant data from a large dataset. In many cases, sampling is very useful because processing the entire dataset is computationally too expensive. In general, the critical issue of these strategies is the selection of a limited but representative sample from the entire dataset. Various random, deterministic, density biased sampling, pseudo-random number generator and sampling from non-uniform distribution strategies exist in the literature (Rubinstein & Kroese, 2011). How- ever, very little work has been done on the Pseudo-random number generator and sampling from non- uniform distribution strategies, especially in the multi-dimensional case with heterogeneous data. Naive sampling methods are not suitable for noisy data which are part of real-world applications, since the performance of the algorithms may vary un- predictably and significantly. The random sampling approach effectively ignores all the information present in the samples which are not part of the reduced subset (Whelan et al., 2010). An advanced data reduction algorithm should be developed in multi-dimensional real-world datasets, taking into account the heterogeneous aspect of the data. Both approaches (Colom´e et al., 2014)(Fakoor & Hu- ber, 2012) are based on sampling and a probabilistic representation from uniform distribution strategies. The authors of (Fakoor & Huber, 2012) pro- posed a method to reduce the complexity of solv- ing Partially Observable Markov Decision Processes (POMDP) in continuous state spaces. The paper uses sampling techniques to reduce the complex- ity of the POMDPs by reducing the number of state variables on the basis of samples drawn from these distributions by means of a Monte Carlo approach and conditional distributions. The authors in (Colom´e et al., 2014) applied dimensionality reduction to a recent movement representation used in robotics, called Probabilistic Movement Primitives (ProMP), and they addressed the problem of fitting a low-dimensional, probabilistic representation to a set of demonstrations of a task. The authors fitted the trajectory distributions and estimated the parameters with a model- based stochastic using the maximum likelihood method. This method assumes that the data follow a multivari-ate normal distribution which is different from the typical assumptions about the relationship between the empirical data. The best we can do is to exam- ine the sensitivity of results for different assumptions about the data distribution and estimate the optimal space dimension of

the data.

2.4 Similarity measure dimensionality reduction

There are other widely used methods for data reduction based on similarity measures (Wencheng, 2010) (Pirolla et al., 2012)(Zhang et al., 2010). Ac- cording to (Dash et al., 2015), the presence of redundant or noisy features degrades the classification performance, requires huge memory, and consumes more computational time. (Dash et al., 2015) proposes a three- stage dimensionality reduction technique for microarray data classification using a comparative study of four different classifiers, multiple linear regression (MLR), artificial neural network (ANN), *k*-nearest neighbor (*k*-NN), and naive Bayesian classifier to observe the improvement in performance. In their experiments, the authors reduce the dimension without compromising the performance of such models. (Deegalla et al., 2012) proposed a dimensionality reduction method that employ s classification approaches based on the *k*-nearest neighbor rule. The effectiveness of the reduced set is measured in terms of the classification accuracy. This method attempts to derive a minimal consistent set, i.e., a minimal set which correctly classifies all the original samples (Whelan et al., 2010). (Venugopalan et al., 2014) discussed the ongoing work in the field of pattern analysis for bio-medical signals (cardio-synchronous waveform) using a Radio Frequency Impedance Interrogation (RFII) device for the purpose of user identification. They discussed the feasibility of reducing the dimensions of these signals by projecting them into various sub-spaces while still preserving inter-user discriminating information, and they compared the classification performance using traditional dimensionality reduction methods such as PCA, independent component analysis (ICA), random projections, or *k*-SVD-based dictionary learning. In the majority of cases, the authors see that the space obtained based on classification carries merit due the dual advantages of reduced dimension and high classification.

Developing effective clustering methods for high- dimensional datasets is a challenging task (Whe- lan et al., 2010). (Boutsidis et al., 2015) studied the topic of dimensionality reduction for *k*-means clustering that encompasses the union of two approaches: 1) A feature selection-based algorithm selects a small subset of the input features and then the *k*-means is applied on the selected features. 2) A feature extraction-based algorithm constructs a small set of new artificial features and then the *k*- means is applied on the constructed features. The first feature extraction method is based on random projections and the second is based on fast

approximate SVD factorization. (Sun et al., 2014) developed a tensor factorization based on a clustering algorithm (k-mean), referred to as Dimensionality Reduction Assisted Tensor Clustering (DRATC). In this algorithm, the tensor decomposition is used as a wayto learn low-dimensional representation of the given tensors and, simultaneously, clustering is con- ducted by coupling the approximation and learning constraints, leading to the PCA Tensor Clustering and Non-negative Tensor Clustering models.

3. METHODOLOGY

The approach presented in this paper for dimensionality reduction in very large datasets is based on the theory of Copulas and the LU-decomposition method (Forward Substitution). The main goal of the method is to reduce the dimensional.

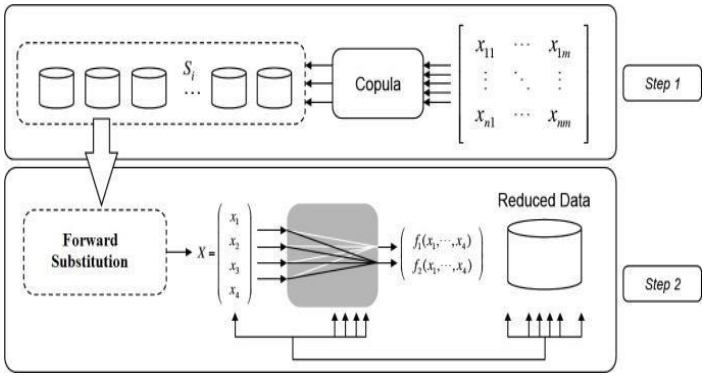

Figure 1: Overview of the proposed reduction method.

Spaces of data without losing important/interesting information. On the other hand, the goal is to estimate the multivariate joint probability distribution without imposing constraints on specific types of marginal distributions of dimensions. Figure 1 shows an overview of the proposed reduction method which operates in two main steps.

In the first step, large raw datasets are decomposed into smaller subsets when calculating the data dependencies using a Copula by taking into account heterogeneous data and removing the data which are strongly dependent. In the second step, we want to reduce the space dimensions by eliminating dimensions that are linear combinations of others. Then we will find the coefficients of the linear combination of dimensions by applying the LU-decomposition method (Forward Substitution) to each subset to obtain an independent set of variables in order to improve the efficiency of data mining algorithms. The two different steps of the pro- posed method are as follows (See also Figure 1):

Step 1: Construction of dependent sample sub- sets $S_{i,(i=1,...,k}$t)

In order to decompose the real-world dataset into smaller dependent sample subsets, vectors are considered which are linearly dependent in the original data.

Empirical Copula will be calculated first to better observe the dependencies between variables. According to the marginal distributions from the observed and approved empirical Copula, we can determine the theoretical

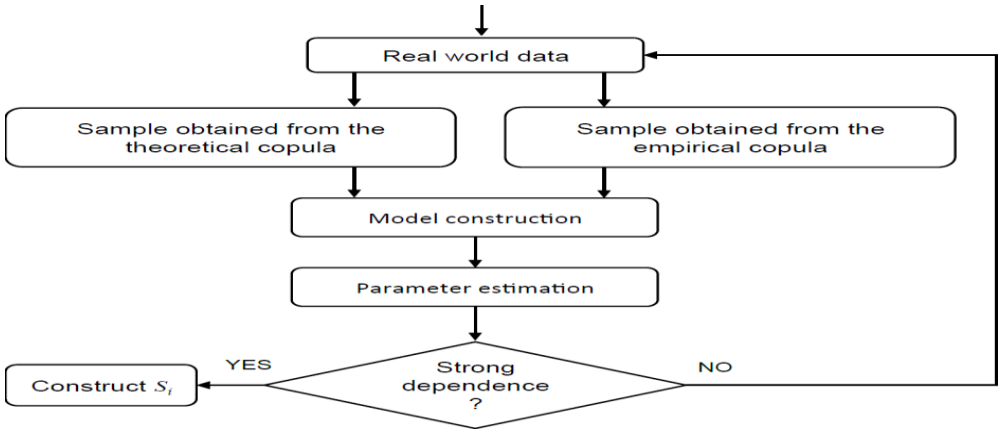

Figure 2: Construction of the subsets Si,(i=1,...,kt).

Copula, that links univariate marginal distributions to their joint multivariate distribution function, and then we will regroup dimensions having the strong correlation relationship in each sample subset $S_{i,(i=1,...,k}$t) by estimating the parameters of the Copula. In this paper, we have presented the Gaussian Copula that corresponds to our experimental results. An illustration of the Copula method is given in Figure 2. The dependence between two continuous random variables $X1$ and $X2$ is defined as follows: If the correlation parameter ρ is greater than 0.5, then $X1$ and $X2$ are positively correlated, meaning that the values of $X1$ increase as the values of $X2$ increase (i.e., the more each at- tribute implies the other). Hence, a higher value may indicate that $X1$ and $X2$ are positively dependent, and probably have a highly redundant attribute, then these two samples will be made as in the same subset $S_{i,(i=1,...,k}$t). When the parameter of the Copula ρ of the two continuous random variables $X1$ and $X2$ is greater than 0.7, then $X1$ and $X2$ have a strong dependence. If the resulting value is equal or less than 0, then $X1$ and $X2$ are independent and there is no correlation between them.

The output of the sample subset $S_{i,(i=1,...,k}$t) represents a matrix that retains only dependent samples of the original matrix in order to detect, and remove a maximum of the redundant

dimensions, which are linear combinations of others, in the second step.

Step 2: LU-decomposition method

The key idea behind the use of the Forward Substitution method is to solve the lin- ear system equations as given by the samples $S_{i,(i=1,\ldots,k}t)$ with an upper-triangular coefficient matrix in order to find the coefficients of linear sample combinations and to provide a low linear space (Xi;i=1,……k) of the original matrix as shown in Figure 3.

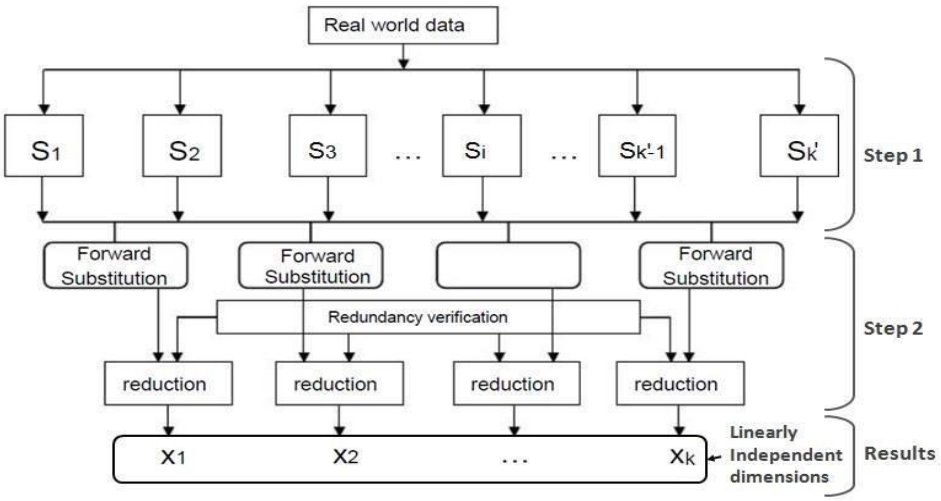

Figure 3: LU-decomposition method.

The LU decomposition method is an efficient procedure for solving a system of linear equations $\alpha \chi \acute{S} = C$, and it can help accelerate the computation. When C is a column vector in the dependent sample subsets $S_{i,(i=1,\ldots,k}t)$, and α_j is an output vector representing the relationship between dimensions or the coefficients of the linear combination of dimensions, **SSi;(i=1,….k'-1)** induces a lower triangular matrix without column C. We conclude that each matrix $\acute{S}i_{,(i=1,\ldots,k}t-1)$ induces a lower triangular matrix of the following form:

$$
\begin{aligned}
&(S)\ \{\alpha 1x11 = c1\\
&\{\alpha 1x21 + \alpha 2x22 = c2 \qquad (8)\\
&\{\alpha 1xn1 + \alpha 2xn2 + \ldots + \alpha nxnn = cn
\end{aligned}
$$

From the above equations, we see that $\alpha 1 = c1/x11$. Thus, we compute $\alpha 1$ from the first equation and substitute it into the second to compute $\alpha 2$,..., etc. Repeating this process, we reach equation i, $2 \leq i \leq n$, using the following formula:

$$\alpha i = 1/xi[ci-\sum_{\square=1}^{i-1} \alpha ixij],\ i=2,,n \qquad (9)$$

From the above equations, it has been observed that α1 =c1/x11. Thus, α1 is computed from the fist equation and substituted it into the second to compute α2,..., etc. Repeating this process, we reach equation i, 2 ≤ i ≤ n, using the following formula:

$\alpha i = 1/xii[ci - \sum_{\square=1}^{i-1} \alpha j xij], i = 2, \ldots n$ (10)

Algorithm 1: Dimensionality linear combination reduction method

Input: Vector C and a lower triangular matrix S; Output: Vector α. Begin α 1 = c1=x11 for i := 2 to n do α i = ci; for j := 1 to i - 1 do α i = α i - xij α j end α i = α i/xii end

The algorithm 1 used for this resolution makes (*n X* (*n* - 1))/2 additions and subtractions, (*n X* (*n* 1))/2 multiplications and *n* divisions to calculate the solution, a global number of operations in the order of n^2.

Statistical precision: The goal of this part is to test the statistical efficiency, after the final dimensionality reduction of all the databases using PCA, SVD,

SPCA, and the proposed approach. The most common precision measure is the standard deviation (sd) measured by the formula (11).

$Sd = \sqrt{1/N-1} \sum\sum^{n} (xi - x\ bar)$ (11)

Classification accuracy: The goal of this part is to improve the effectiveness of dimensionality reduction before and after the final reduction of the dimensionality of Pima Diabetes and Waveform databases, by using the classification methods for PCA, SVD, Sparse PCA, and our proposed approach.

In general, the performance of a classification process can be evaluated by the following quantities:

True Positives (TP), False Positives (FP), and False Negatives (FN), and the use ofdifferent metrics such as precision and recall. The precision P and the recall R are measured by the following formulas:

P =TP/TP + FP (12)

R =TP/TP + FN (13)

Where N is the number of values of the sample, and xi bar is the mean of the values xi.

4. RESULTS AND ANALYSIS

Table 2: Dimensionality reduction results (number of columns reduced.)

Methods	SVD	PCA	SPCA	PA
Pima Diabetes	5	9	7	10
Waveform	10	20	20	22
Human Activity 1	28	107	98	112
Human Activity 2	117	342	343	344
Thyroid Disease	185	505	524	525

Dimensionality reduction results

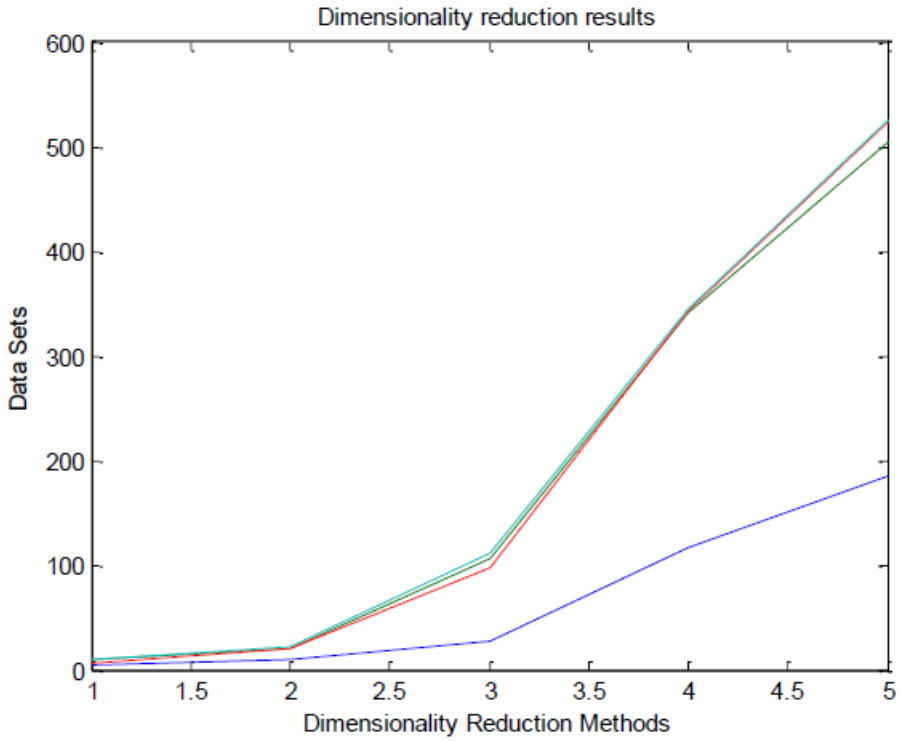

Figure 4: Dimensionality reduction (number of columns reduced.)

5. CONCLUSION AND FUTURE WORK

A new method for dimensionality reduction in the data pre-processing phase of mining high-dimensional data has been introduced. This approach is based on the theory of Copulas (sampling techniques) to estimate the multivariate joint probability distribution without constraints of specific types of marginal distributions of random variables that represent the dimensions of proposed datasets. A Copula based model provides a complete and scale free description of dependency that is thereafter used to detect the redundant values. A more extensive evaluation is made by eliminating dimensions that are linear combinations of others after having decomposed the data, and using the LU decomposition method. The problem of data reduction has been reformulated as a constrained optimization problem. Proposed

approach is compared with well-known data mining methods using five real-world datasets taken from the machine learning repository in terms of the dimensionality reduction and the efficiency of the methods. The efficiency of the proposed method was improved by using the both statistical and classification methods. The different results obtained show the effectiveness of our approach which outperforms significantly the performance of dimensionality reduction comparing to other methods, i.e., it provided a smaller bias with more better standard deviations, a highest precision, and a lowest recall with all classifiers for all databases. Further work can be carried out in several directions. Researchers have made great efforts to improve the performance of the dimensionality reduction approach in very large datasets. However, the most serious problem is the presence of missing values in datasets. Missing value scan result in loss of efficiency of the dimensionality reduction approach, lead to complications in handling and analyzing the data, or distort the relationship between the data distribution. Also, it would be interesting to investigate the possibility of using meta-heuristics or hybrid approaches to determine a solution of the proposed optimization problem in the Big Data setting.

REFERENCES

1. Agatonovic-Kustrin, S., & Beresford, R. Basic concepts of arti_cial neural network (ann) modeling and its application in pharmaceutical research. Journal of pharmaceutical and biomedical analysis, 2000; 22: 717-727.
2. Bai, D., Liming, W., Chan, W., Wu, Q., Huang, D., & Fu, S. Sparse principal component analysis for feature selection of multiple physiological signals from ight task.In Control, Automation and Systems (ICCAS), 2015 15th International Conference on (pp. 627{631). IEEE. Boutsidis, C., Zouzias, A., Mahoney, M. W., & Drineas, P. Randomized dimensionality reduction for-means clustering. Information Theory, IEEE Transactions on, 2015; 61: 1045-1062.
3. Colom_e, A., Neumann, G., Peters, J., & Torras, C. Dimensionality reduction for probabilistic movement primitives. In Humanoid Robots (Humanoids), 201414th IEEE-RAS International Conference on, 2014; 794-800.
4. Dash, R., Misra, B., Dehuri, S., & Cho, S.-B. Ef- _cient microarray data classi_cation with three-stage dimensionality reduction. In Intelligent Computing, Communication and Devices, 2015; 805-812.
5. Deegalla, S., Bostr• om, H., & Walgama, K. Choice of dimensionality reduction methods for feature and classifier fusion with nearest neighbor classifiers. In Information Fusion

(FUSION), 15th International Conference on, 2012; 875-881.

6. Derrac, J., Chiclana, F., Garcia, S., & Herrera, F. Evolutionary fuzzy k-nearest neighbors algorithm using interval-valued fuzzy sets. Information Sciences, 2016; 329: 144-163.
7. Fakoor, R., & Huber, M. A sampling-based approach to reduce the complexity of continuous state space POMDPs by decomposition into coupled perceptual and decision processes. In Machine Learning and Applications (ICMLA), 2012 11th International Conference on, 2012; 687-692.
8. Gisbrecht, A., & Hammer, B. Data visualization by nonlinear dimensionality reduction. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2015; 5: 51-73.
9. Han, J., & Kamber, M. (San Francisco). Data Mining: Concepts and Techniques. 13: 978-1-55860-901-3 (2nd ed.). Diane Cerra. Hettiarachchi, R., & Peters, J. (2015). Multi-manifold LLE learning in pattern recognition. Pattern Recognition, 2006; 48: 2947-2960.
10. Houari, R., Bounceur, A., & Kechadi, M.-T. A new method for dimensionality reduction of multidimensional data using copulas. In Programming and Systems (ISPS), 2013 11th International Symposium on, 2013; 40-46.
11. Houari, R., Bounceur, A., & Kechadi, T. A New Approach for Dimensionality Reduction of Large MultiDimensional Data Based on Sampling Methods for DataMining, 2013.
12. Houari, R., Bounceur, A., Kechadi, T., Tari, A., & Euler, R. A new method for estimation of missing data based on sampling methods for data mining. In Advancesin Computational Science, Engineering and Information Technology, 2013; 89-100.
13. Kerdprasop, N., Chanklan, R., Hirunyawanakul, A., & Kerdprasop, K. An empirical study of dimensionality reduction methods for biometric recognition. In Security Technology (SecTech), 2014 7th International Conference on, 2014; 26-29.
14. Kuang, F., Zhang, S., Jin, Z., & Xu, W. A novel svm by combining kernel principal component analysis and improved chaotic particle swarm optimization for intrusion detection. Soft Computing, (pp. 1{13). Lichman, M. (2013). Uci machine learning repository, university of california, irvine, school of information and computer sciences, http://archive.ics.uci.edu/ml., 2015.
15. Lin, P., Zhang, J., & An, R. Data dimensionality reduction approach to improve feature selection performance using sparsi_ed svd. In Neural Networks (IJCNN), 2014 International Joint Conference on (pp. 1393{1400). IEEE. Nelsen, R. B. (2007). An introduction to copulas. (2nd ed.). Springer Science & Business Media, 2014.
16. Pirolla, F. R., Felipe, J., Santos, M. T., & Ribeiro, M. X. Dimensionality reduction to improve content based image retrieval: A clustering approach. In Bioinformatics and

Biomedicine Workshops (BIBMW), IEEE International Conference on, 2012; 752{753).

17. Rubinstein, R. Y., & Kroese, D. P. Simulation and the Monte Carlo method volume 707. John Wiley & Sons, 2011.
18. Saoudi, M., Bounceur, A., Euler, R., & Kechadi, T. Data mining techniques applied to wireless sensor networks for early forest _re detection. In Proceedings of the International Conference on Internet of things and Cloud Computing, 2016; 71.
19. Sasikala, S., & Balamurugan, S. A. A. Data classification using pca based on effective variance coverage (evc). In Emerging Trends in Computing, Communication and Nanotechnology (ICE-CCN), International Conference on, 2013; 727-732.
20. Shen, H., & Huang, J. Z. Sparse principal component analysis via regularized low rank matrix approximation. Journal of multivariate analysis, 2008; 99: 1015-1034.
21. Sun, Y., Gao, J., Hong, X., Guo, Y., & Harris, C. J. Dimensionality reduction assisted tensor clustering. In Neural Networks (IJCNN), 2014 International Joint Conference on (pp. 1565{1572). IEEE. Venugopalan, S., Savvides, M., Griofa, M. O., & Cohen, 2014.
22. K. Analysis of low-dimensional radio-frequency impedance-based cardio- synchronous waveforms for biometric authentication. Biomedical Engineering, IEEE Transactions on, 61, 2324{2335. Wan, X., Wang, D., Peter, W. T., Xu, G., & Zhang, Q. (2016). A critical study of di_erent dimensionality reduction methods for gear crack degradation assessment under di_erent operating conditions. Measurement, 2014; 78:138-150.
23. Watcharapinchai, N., Aramvith, S., Siddhichai, S., & Marukatat, S. Dimensionality reduction of sift using pca for for object categorization. In Intelligent Signal, 2009.
24. Processing and Communications Systems, ISPACS 2008. International Symposium on, 2008; 1-4. IEEE.
25. Wencheng, P. Theory and application of parameter self-optimizing intelligent sampling method. In Signal Processing (ICSP), 2010 IEEE 10th International Conference on (pp. 66{69). IEEE. Whelan, M., Khac, N. A. L., Kechadi, M. et al. (2010). Data reduction in very large spatio-temporal datasets. In Enabling Technologies: Infrastructures for Collaborative Enterprises (WETICE), 2010 19th IEEE International Workshop on, 2010; 104-109. IEEE.
26. Zhai, M., Shi, F., Duncan, D., & Jacobs, N. Covariance-based pca for multi-size data. In Pattern Recognition (ICPR), 22nd International Conference on, 2014; 1603-1608. IEEE.
27. Zhang, J., Nie, X., Hu, Y., Liu, S., Tian, Y., & Wu, L. A method for land surveying sampling optimization strategy. In Geoinformatics, 2010 18th International Conference on (pp. 1{5). IEEE. Zhang, T., Du, Y., Huang, T., & Li, X. (2016). Stochastic simulation

of geological data using isometric mapping and multiple-point geostatistics with data incorporation. Journal of Applied Geophysics, 2010; 125: 14-25.

28. Zou, H., Hastie, T., & Tibshirani, R. Sparse principal component analysis. Journal of computational and graphical statistics, 2006; 15: 265-286. 20.