Data mining Application of Data Reduction and Clustering Domain of Textile Database

M. Salomi, R. Lakshmi Priya, Manimannan G, N. Manjula Devi

Abstract: ***This research paper attempts to identify the textile data structure and hidden pattern of original database with certain important parameters. The main objectives of this study are to identify the first n number of factors that explained over the study period. Initially factor analysis is performed to extract factor scores. Principal extraction is performed through Data mining package with sixteen textile fabrics parameters. Factor extraction is aimed to uncover the intrinsic pattern among the textile parameters considered and an important point of factor analysis is to extract factor scores for further investigation. Thus, factor analysis consistently resulted in three factors for the whole datasets. The amount of total variation explained is over 75 percent in factor analysis with varimax rotation. The factor loadings or factor structure matrix with unassociated rotation methods are not always easy to interpret. The nonhierarchical k-mean clustering is also used to identify meaningful cluster based on their parameter means of original database.***

Keywords: Data Mining, Principal Component Analysis, k-mean Clustering, Sillohoutte plot and Scatter plot,

I. INTRODUCTION

Normally a textile industry uses fibres, yarn and fabrics for testing their properties based on the quality. In general, for every end-use, the mechanical properties of fabrics are of special interest, this study how these properties are related to fabric. Yarn properties could help in make a decision how to produce a suitable fabric at a minimum cost when its likely range of use is known. The textile industry depends on raw materials such as cotton, jute and silkfibres. The fibres were acquired from natural resources. After mining, materials are then sent to spinning mills for processing and yarns are produced. Yarn structure and properties are primarily influenced by fibre properties based on their length, fineness and cross-sectional shape.The spinning method consists of ring, rotor and airjet and process variables based on their twist insertion rate, rotor speed, nozzle pressure, etc. Classifying of yarns is another important need in assessing textile quality control for business purposes, since it provides a useful means of expressing yarn standards in the market.

II. REVIEW OF LITERATURE

Many scholars quoted their research, the traditional methods; it is done by a similarity of measurements of various quality related variables with their values as suggested in a standard, which is diverse for different countries. Yarns are used for producing fabrics.

Revised Manuscript Received on October 20, 2020.

M. Salomi, Assistant Professor, Department of Statistics, Madras Christian College, Chennai (Tamil Nadu), India.

R. Lakshmi Priya, Assistant Professor, Department of Statistics, Dr. Ambedkar Govt. Arts College, Chennai (Tamil Nadu), India.

Manimannan G, Assistant Professor, Department of Mathematics, TMG College of Arts and Science, Chennai (Tamil Nadu), India.

N. Manjula Devi, Bio Statistician, Department of Community Medicine, Karpaka Vinayakar Institute of Medical Sciences, Chengalpet, (Tamil Nadu), India.

These fabrics can be classified according to weave structure, knitting method, base material used such as cotton, jute and silk. Syntheticmaterial is based on polyester, softness or roughness of the material. The classification is very essential for textile industry. Textile technologists have introduced various methods for solving the above mentioned problems. Most of the measurement properties are obtained using Kawabata instrument. Kawabata (1980) and his associates established the *Kawabata Evaluation System for Fabrics (KES-F),* which is used to measure mechanical properties of fabrics. It's shown to offer benefit over other instruments in the routine dimensions of the properties. Fabric is tested for their tensile, bending, shearing, compression, surface properties and also for thickness and weight. To categorize a fabric based on measurements, some scientific methodology is to be implemented. The relationship between yarn structure and tensile properties has been studied extensively. A number of statistical models have been developed for textile parameters. More number of publications of research papers shows that fabric handle received wide attention from scientist and technologists from very early days. Fabric database handle has become one of the most likely areas of textile research during the past twenty years. Fabric 'hand' or 'handle' is defined as the quality of a fabric or yarn assessed by the reaction obtained from the sense of touch. KES-F system of assessment of hand value has been used by fibre, yarn and fabric producers all over the world. Many researchers have used multivariate statistical methods of Cluster Analysis, Principal Component Analysis in the study of fabric hand. Recently, Rong, Slater and Fei (1994) have used cluster analysis method for grading of yarns in textile industry. Pan, Zeronian and Ryu(1993) suggest statistical methods such as PCA, Variable Cluster Analysis, D-optimal method and multi collinearity tests for identifying the most important mechanical properties. Pan, Yen, Zhao and Yang (1988) have conversed classification of fabrics by hierarchical clustering methods. Recently Artificial Neural Network technique (ANN) has been applied for textile database for the classification purpose by Ramesh, Rajamanickam and Jayaraman (1995)

III. DATABASE

The database consists of three data sets, combined together which has 105 fabric samples with 16 KES-F parameters. In the present study, the combined data sets as well as individual data sets are analyzed. All data setwas subjected to normality test and equality of variance test and the same established. The three data sets are, Data Set 1 comprises different types of Polyester Fabrics (Regular/Micro fibres). It includes 27 fabric samples with 16 parameters (Table 1).

Data Set 2 consists of 40 Lyocell/Viscose fabric samples with 16 KES-F parameters and Chemically Treated Polyester fabric samples constitute Data set 3.

Table 1. Sixteen Textile Parameters and their abbreviation

Parameters and their Abbreviation
1. Linearity of the Stress Strain Curve (LSSC)
2. Tensile Energy (TE):
3. Tensile Resilience (TR)
4. Bending Rigidity (BR):
5. Hysteresis of Bending (HB):
6. Shear Rigidity (SR):
7. Hysteresis of Shear Force (HSF):
8. Hysteresis of Shear Force 2 (2HSF) :
9. Linearity of the Compression Curve (LCC):
10. Work of Compression Curve (WCC):
11. Compressional Resilience (CR):
12. Coefficient of Surface Friction (CSF):
13. Mean Deviation of Coefficient of Friction (MDCF):
14. Surface Roughness (SR):
15. Fabric Thickness (FT):
16. Fabric Weight (FW):

IV. METHODOLOGY

4.1 MacQueen k-means Clustering Algorithm

In general context, one of the most popular clustering algorithms suggested by MacQueen (1967) known as ***k-means*** is used to identify q classes in the data set. This technique uses Euclidean distance measure computed on textile paparameters to partition or group the data set into mutually exclusive groups such that the members of each group are as close as possible to one another and different groups are as far as possible. Thus, ***k***-means clustering assigns group labels to data sets which are unknown initially based on the nuclei of clusters or groups as seed points exhibited in factor analysis. The number of cluster is determined as a part of clustering procedure. In its simplest report, the process is composed of three steps.

Step 1: Partition the data sets in to ***k*** initial clusters.

Step 2:The initial cluster starts from *k=2…n* based cluster centroid. The distance is computed using Euclidean distance with standardized or unstandardized observations. Recalculate the centroids for the cluster receiving the new item and the cluster losing the item.

Step 3: Repeat Step 2 until no more reassignment takes place.

Rather than starting with partition of all items in to *k* preliminary groups in Step 1, Specify ***k*** initial centroids and proceed Step 2.The final assignment of items to clusters will be, to some extent depends upon initial partition or initial selection of seed points. The present study deals with textile database with sixteen parameters as input data matrix. Centroids are calculated and assigned the samples based on Euclidean distance measure. The cemtroids are calculated by using following equations.

$$d(x,y)=\sum_{i=1}^{n}\left|y_i - x_i\right|$$

4,2 Data Mining k- Mean Clustering Methods

The data mining widget applies the k-Means clustering algorithm to data and outputs of textile dataset in which cluster index is used as a class attribute. The original class attribute, if it exists, is moved to Meta attributes. The scores of clustering results for various k are also shown in the widget. The following algorithm execute the widget

Step 1: Initially, select the number of clusters.

Fixed: *The* algorithm clusters data in a specified number of clusters.

Optimized: widget shows clustering scores for the selected cluster range:

Silhouette: This method contrasts average distance to elements in the same cluster with the average distance to elements in other clusters

Inter-cluster distance: This measures distances between clusters, normally between centroids values

Distance to centroidsmeasures distances to the arithmetic means of clusters.

Step 2: Select the initialization method:

k-Means++ to evaluate first centre is selected randomly and subsequent are chosen from the remaining points with probability proportioned to squared distance from the closest center values

Random initialization clusters are assigned randomly at first and then updated with further iterations. *Re-runs* to evaluate how many times the algorithm is run from random initial positions; the result with the lowest within-cluster sum of squares will be used and **maximal** *iteration***s** is the maximum number of iterations within each algorithm run can be set manually.

Step 3: The output widget create a new dataset with appended cluster information. Select how to append cluster information based on class, feature or Meta attribute and name the column.

Step 4: If Apply automatically is ticked, the widget will assign changes automatically. Alternatively, click *Apply* button.

Step 5: Produce a report for a given database.

Step 6:Scores of clustering results for various k in Table 2

4.3 Principal Component Analysis

Principal Component Analysis (PCA) is generally preferred for purposes of data reduction and variable reduction but not when the goal is to detect the latent construct or factors. Factor analysis is similar to principal component analysis, in that factor analysis also involves linear combinations of variables withPCA linear transformation of input data.

4.3.1 Inputs of Data:

Textile database as a input dataset with sixteen parameters

4.3.2 Outputs

The outputs based on PCA transformed data Eigen Values Components.

Principal Component Analysis (PCA) computes the PCA linear transformation of the input data. It outputs either a transformed dataset with weights of individual instances or weights of Principal Components Analysis.

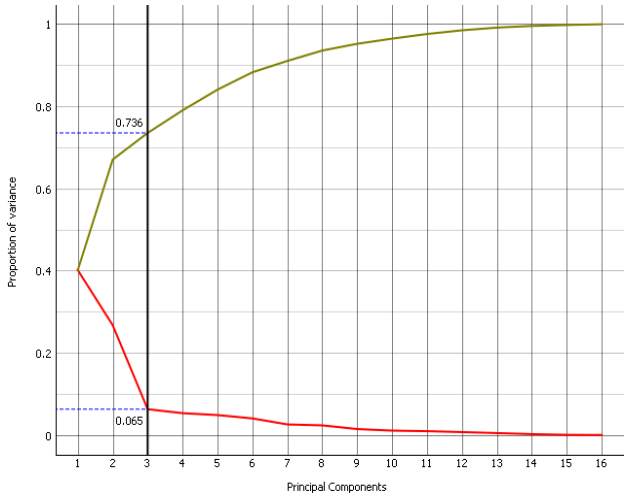

Figure 1. PCA Screw Plot for Extracted Components

4.3.3 Data mining Algorithm for Principal Component Analysis

Step 1: The file Widget to import the database from various formats of data file like, .dat. txt. CSV, .tab, etc. (Three sets of database with sixteen parameters)
Step 2: Select the number of principal components to be included in output. It is the best to choose as few as possible with variance covered as high as possible.
Step 3: Set how much variance to be covered with principal components.
Step 4: Normalize data to adjust the values to common scale.
Step 5: When Apply Automatically is ticked, the widget will automatically communicate all changes. Alternatively, click *Apply*.
Step 6: Press *Save Image* to save the created image to your computer. Produce a report.
Step 7: Principal components graph, where the red (lower) line is the variance covered per component and the green (upper) line is cumulative variance covered by components.
The number of components of transformation can be selected either in Components Selection input box or by dragging the vertical cutoff line in the graph(Figure 1).

4.3.4 Proposed Algorithm for k++ (k-means)

In this section, the researcher explains the widget with the following schema.
Step 1: The file Widget to import the database from various formats of data file like, .dat. txt. CSV, .tab, etc. (Three sets of database with sixteen parameters)
Step 2: The k- mean widget analyze for any given data and group them for requirements. The k-mean widget by default i=2 clusters, then increase up to i=n meaningful clusters, then stop the iteration process.
Step 3: Data widget shows that grouped data of k- mean clustering.
Step 4: Select Row Widget to highlight the grouped data.
Step 5: Finally, the scatter plot widget, to visualize the grouped data using k-means clustering method (Figure 2)

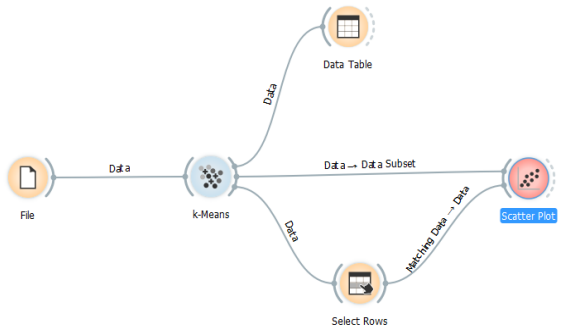

Figure 2. k- Means Work Flow

V. RESULT AND DISCUSSION

The input textile database load the file widget, the k-mean widget divides it into three clusters and presentedin Data Table. The interesting parts are the Scatter Plotachieved from Select Rows widget (Figure 3). Since k-means added the cluster index as a class attribute, the scatter plot will color the points according to the clusters they are shown inFigure 3.

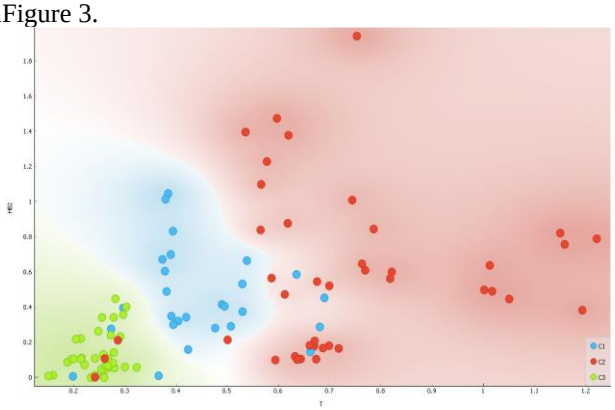

Figure 3. Scatter Plot for k- Means

It may be noticed that removal of noisy data or attributes to be unusedclasses in Select Rows widget is unchecked. This is important: if the widget modifies the attributes, it outputs a list of modified instances and the scatter plot cannot compare them with original data.Conceivably a simpler way to test the match between clusters and the original classes is to use Distributions widget (Figure 4).

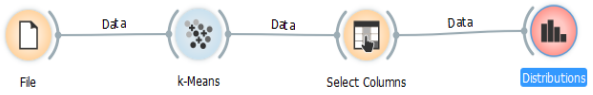

Figure 4. k- Means Distribution Work Flow

The widget visualizes normal attributes bar chart for three datasetonly. This is achieved by Select Columns widget: Reinstate the original class *textile* Data set as the class and put the cluster index among the attributes. The match is perfect for 27 Polyester Fabrics: all parameters of Polyester Fabrics are in first cluster (blue). 40 Lyocell/Viscose fabrics are in second cluster (red), while two ended up in first. Chemically Treated Polyester fabric samples are in third cluster and 38 in the second.

The following figure represents the cluster information of three data sets (Figure 5).

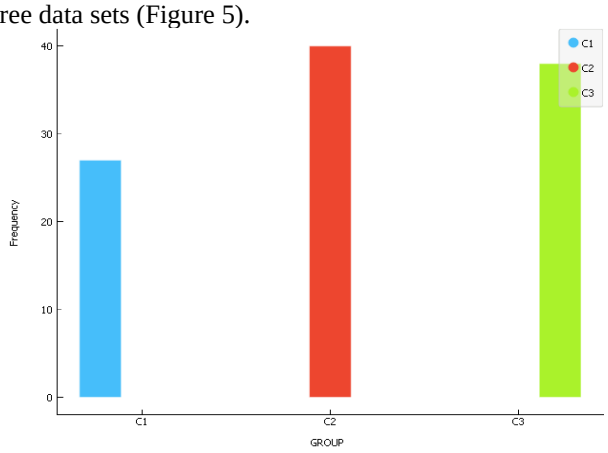

Figure 5. k- Means Classification Work Flow

Principal Component Analysis can be used to simplify visualizations of large datasets. The researcher used the Textile Fabricsdataset to show how visualization of dataset with PCA can be improved. The transformed data in Scatter Plot and Silhouette plot shows a cleardistinction between classes than the default locations (Figure 6-8)

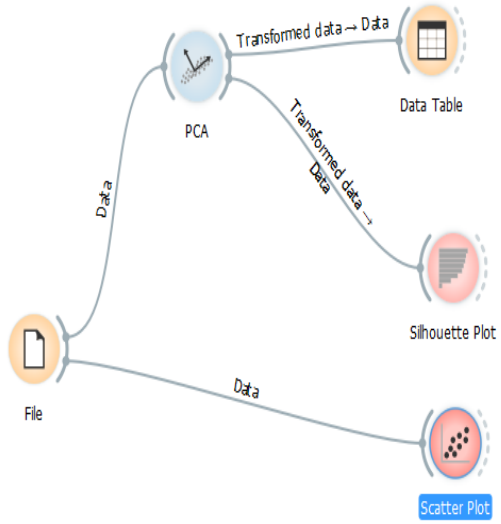

Figure 6. Principal Component Analysis Work Flow

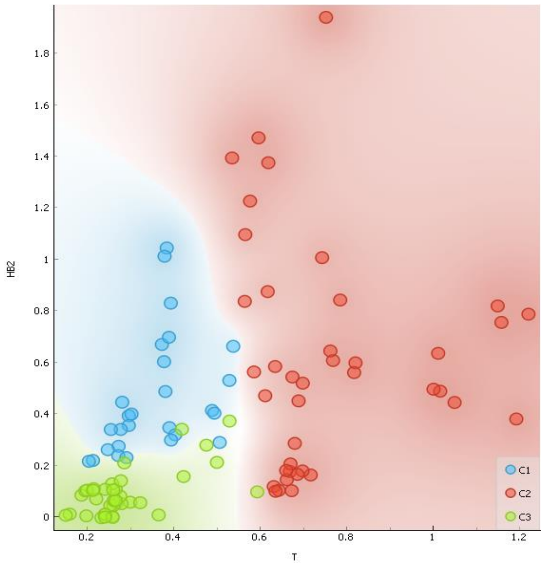

Figure 7. Scatter Plot for Principal Component Analysis

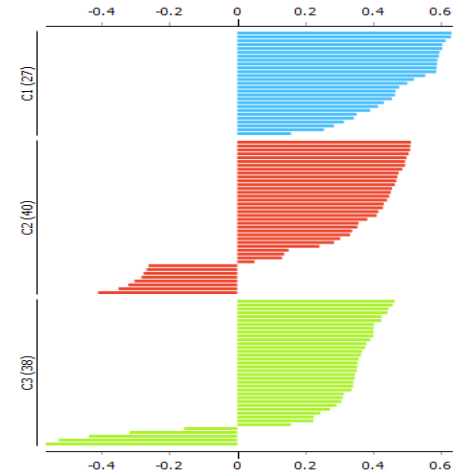

Figure 8. Silhouette for Principal Component Analysis

The below flow diagram widget provides two outputs: transformed data and principal components. Transformed data are weights for individual instances in new coordinate system, while components are system descriptors based on their weights for principal components. When fed into the Data Table widget, both outputs in numerical form can be achieved. In order to provide a more clean visualization of workflow, two data tables were used, but one can also choose to edit the links in such a way that it displays data in just one data table widget. It is necessary to create two links and connect the *Transformed data* and *Components* inputs to *Data* to extract PCA output. The following workflow shows the performance of PCA (Figure 9).

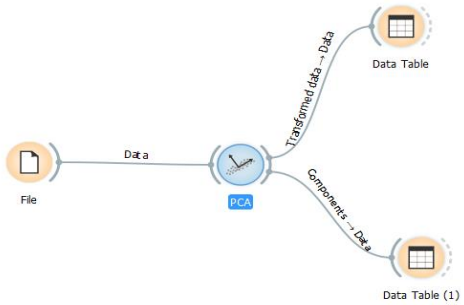

Figure 9. Principal Component Analysis Workflow for Components and Transformed Data

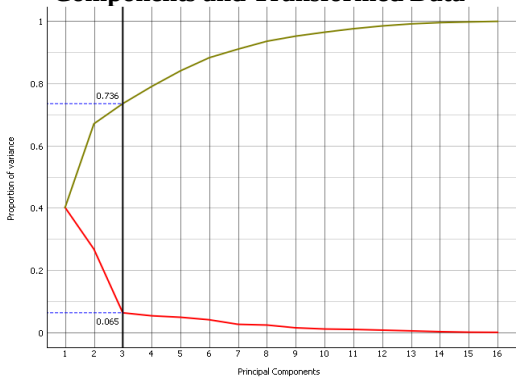

Figure 10. Screw Plot for Components and Transformed Data

The Screw plot shows that variance and number factor in textile database (Figure 10). The total variance explained 75 percent with three factors and they are labeled as, Polyester Fabrics, Lyocell/Viscose fabrics and Chemically Treated Polyester fabrics. The three extracted factor scores are shown in the following table. Finally, the removed three factor component score are listed in next table of factor scores (Table 1 to 2).

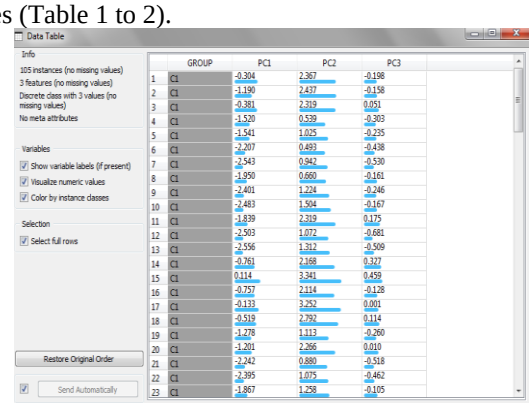

Table 1. PCA Extracted Factor Score

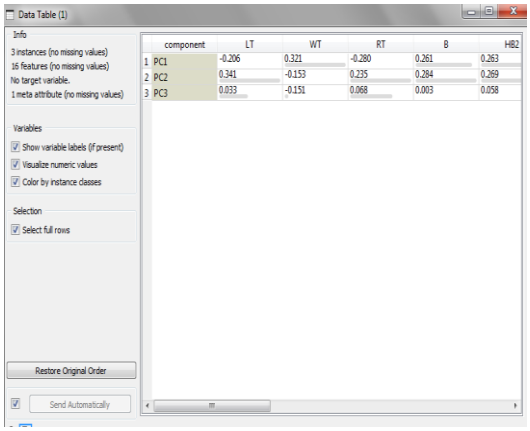

Table 2. PCA Extracted Factor Component Score

VI. CONCLUSION

The main objective of this study is to identify the first **n** number of factors that explained over 75 percent of variation in the data set. Initially factor analysis is performed to extract factor scores. Principal extraction is performed through data mining package on sixteen textile fabrics database. Factor extraction is aimed to uncover the intrinsic pattern among the ratios considered and an important point of factor analysis is to extract factor scores for further investigation. Thus, factor analysis consistently resulted in three factors for the three datasets. The amount of total variation explained is over 75 percent in factor analysis. The factor loadings or factor structure matrix with unassociated rotation methods are not always easy to interpret. Both Varimax and Quatrimax orthogonal criterion are employed to improve the interpretability of the set of variables on factors. Though both rotations are almost similar in structure, the grouping of variables is more meaningful under Varimax rotation. Variables have been ordered and grouped by sizes of loadings to facilitate interpretation. Factor analysis results show that the set of variables are highly correlated with their respective factors. After identifying all significant loadings, an attempt is made to assign meaning to the factors and hence factors are labeled as ***Polyester Fabrics, Lyocell/Viscose fabrics and Chemically Treated Polyester fabrics*** respectively. The naming of factors is based on previous study and also with clustered variables on factors. In addition, the mechanical properties stability of all the samples was measured to certain extent without any uncertainty though the number of samples kept varying during the study period.

REFERENCES

1. Chen Y and Collier B J(1997), Characterizing Fabric End-Use by Fabric Physical Properties, Textile Research Journal, Vol. 67, No. 4, pp.247-252.
2. G.K. Gupta (2012), Introduction to Data Mining with Case Studies, PHI Learning Private Limited, New Delhi, India.
3. Kawabata S. (1980), The Standardisation and Analysis of Hand Evaluation, 2nd Edition, The Textile Machinery Society of Japan, Osaka, Japan.
4. MacQueen, J. Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, 281--297, University of California Press, Berkeley, Calif., 1967
5. Pan N, Zeroniam S H and Ryu H S(1993), An Alternative Approach to the Objective Measurements of Fabrics, Textile Research Journal, Vol.63, pp.33-43.
6. Pan N, Yen K C, Zhao S J and Yang S R(1988), A New Approach to the Objective Evaluation of Fabric Handle from Mechanical Properties(Fuzzy Cluster Analysis for Fabric Handle Sorting), Textile Research Institute, pp.565-571.
7. Ramesh M.C., Rajamanickam R. & Jayaraman S. (1995). The Prediction of Yarn Tensile Properties by Using Artificial Neural Networks, J. Text. Inst., 86: 3, pp. 459 – 469.
8. Rong G H, Slater K and Fei R C(1994), The Use of Cluster Analysis for Grading Textile Yarns, Journal Textile Institute, Vol. 85, No.3, pp.389-396.
9. R. Ruhin Kouser and R. Gunasundri (2015), Data Warehousing and Data Mining, Lakshmi Publications, Chennai, India

AUTHORS PROFILE

Dr. Manimannan G. M. Sc. M. Phil. Ph. D, PGDCA,, MCA. He has good research experience by working for many Project Guidance and consultation work in application of Statistics, Neural Networks and Data Mining. He has published more than seventy research papers in various National and International journals.

Dr. Salomi M. M. Sc. M. Phil. Ph. D, She has good research experience by working for many project guidance and consultation work in application of Statistics and Bio – Statistics. She has published more than twenty research papers in various national and International journals.

Dr. R. Lakshmi Priya M. Sc. M. Phil. Ph. D, She has good research experience by working for many project Guidance and consultation work in application of Statistics and Data Mining. She has published more than forty five research papers in various National and International journals.

Ms. N. Manjula Devi M. Sc (Bio Statistics), She has good research experience by working for few project Guidance and consultation work in application of Bio Statistics and Data Mining. She has published more than Ten research papers in various National and International journals.
.