Volume 6, Issue 3, March 2016 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

PCA based Lattice Clustering Methodology for Heterogeneous Dataset Modeling and Analysis

B. Gowri, R. Rajmohan, D. Dinagaran, M. Pajany, A. Divya
Department of Computer Science and Engineering, IFET College of Engineering,
Villupuram, Tamilnadu, India

Abstract- Data mining is a powerful tool for extraction of hidden predictive information from large database. Sorting through large data sets to identify patterns and establishing relationship is a critical task. Clustering is a parallel processing technology which aims at finding uniformity in data characteristic for mining large datasets. The dimensionality of data patterns depends on quantitative and qualitative numerical measures. In this paper, propose a novel Lattice based Clustering Methodology (LCM) which uses lattice clustering technique and principal component analysis (PCA) technologies for modeling and analyzing heterogeneous dataset using binary coded factorial analysis. The Proposed methodology combines the advantages of lattice based factorial data analysis technique to achieve high accuracy in heterogeneous data mining.EM-based Gaussian mixture modeling finally performs as a pattern recognition tool in order to establish a classification of the population into an optimal number of homoscedastic subgroup.The clustering approach has great potential for scalable knowledge discovery from heterogeneous database and for application in open distributed environment such as semantic web.

Keyword: PCA, LCM, Lattice, Heterogeneous, Clustering

I. INTRODUCTION

Clustering [1] is a multivariate data analysis and modeling practice of great important. Principal Component Analysis (PCA) is the strategy that uses orthogonal projections to transform a set of observations of possible correlated variables into a set of values of linear uncorrelated variables called Principal Components. However, PCA has been developed to handle quantitativedata sets, whereas medical data are rather mixed-type, i.e., quantitative and qualitative [5]. In comparison with other data analysis problem, techniques for treating this type of data have not evolved enough in spite of their great importance.Many non-responses in the quantitative data set are often unsuitably treated, which lead to misinterpretation and erroneous assumptions are made.A common problem in multivariate data analysis arises when using a large number of variables.In this system, proposed a simple data mining strategy for analyzing and modeling datasets that contain both quantitative and categorical characters. The strategy exposed below is essentially developed to resolve such a problem and allows furthermost to treat the missing data problem that often thwarts analysts. Explicitly, the strategy interface consists in its first stage in prediction quantitative missing data, and this by projection them on some subspaces generated by the other quantitative factors. This task is performed with almost the same principal in order to replace qualitative observations by their quantitative predictions, where it is rather a logistic regression that guarantees the projection.

In the penultimate stage of the data mining interface, a lattice based PCA is applied to the restored data sets, which is only composed of quantitative variables. An EM-based Gaussian mixture modeling finally performs as a pattern recognition toolin order to establish a classification of the population into an optimal number of homoscedastic subgroups. To overcome this problem we introduced a lattice based clustering method to analyze the missing data problem and predicted the datasets. The algorithm evaluates clusters locally, it also maximizes the overall clustering quality and the local evaluation goes through all levels of the hierarchy in a bottom-up function. The objective of this method is to reduce the complexity of the algorithm and useful in control strategy of the algorithm.

II. RELATED WORK

The MIXMOD (MIXtureMODeling) [3] software fits mixture models to a given data set with a density estimation, a clustering or a discriminate analysis purpose. A large variety of algorithm to estimate the mixture parameter are proposed (EM, Classification EM, Stochastic EM) and it is possible to combine them to lead to different strategies in order to get a sensible maximum of the likelihood (or complete-data likelihood) function. MIXMOD is currently focused on multivariate Gaussian mixtures and fourteen Gaussian models can be considered according to different assumptions on the component variance matrix eigenvalue decomposition. Moreover, different information criteria for choosing a parsimonious model (the number of mixture components, for instance), some of them favoring either a cluster analysis or a discriminate analysis view point, are included.

Kiers[4] considered the orthogonal rotation in PCAMIX, a principal component method for a mixture of qualitative and quantitative variables. PCAMIX includes the ordinary principal component analysis (PCA) and multiple correspondence analysis (MCA) as special cases. In this paper, we give a new presentation of PCAMIX where the principal components and the squared loadings are obtained from a Singular Value Decomposition. The loadings of the quantitative variables and the principal coordinates of the categories of the qualitative variables are also obtained directly. In this context, we propose a computationally efficient procedure for varimax rotation in PCAMIX and a direct solution for the optimal angle of rotation.

In many applications the objects to cluster are described by quantitative as well as qualitative features. A variety of algorithms has been proposed for unsupervised classification if fuzzy partitions and descriptive cluster prototypes are desired. However, most of these methods are designed for data sets with variables measured in the same scale type (only categorical, or only metric). A new fuzzy clustering [6]approach is proposed based on a probabilistic distance measure. Thus a major drawback of present methods can be avoided which lies in the vulnerability to favor one type of attributes

The k-means algorithm [2]is well known for its efficiency in clustering large data sets. However, working only on numeric values prohibits it from being used to cluster real world data containing categorical values. In this paper we present two algorithms which extend the k-means algorithm to categorical domains and domains with mixed numeric and categorical values. The k-modes algorithm uses a simple matching dissimilarity measure to deal with categorical objects, replaces the means of clusters with modes, and uses a frequency-based method to update modes in the clustering process to minimize the clustering cost function. With these extensions the k-modes algorithm enables the clustering of categorical data in a fashion similar to k-means. The k-prototypes algorithm, through the definition of a combined dissimilarity measure, further integrates the k-means and k-modes algorithms to allow for clustering objects described by mixed numeric and categorical attributes

III. PROPOSEDSYSTEM

The system is in Fig.1 is implemented by using the following modules A.Preprocessing, B.Data Measurement, C.Clustering and D.Clustering Interface.

A. Preprocessing

This step is to preprocess extracted words. There are four aspects: splitting, elimination stop words, stemming and removing specific tags. Splitting: combined words are split into verbs and nouns. Because of naming convention, verb and noun are combined in operation names, service names and message of web services. Eliminating stop words: words are stop list. The stop list contains a few common words (including preposition, article, etc.), which are recognized that they are not useful in information retrieval. And in web services classification, stop words could not represent any category either. Stemming, verbs are taken back to infinitive; nouns are taken from plurals to singles; English words are taken to American words.

B. Data Measurement

The first and second tasks are mainly preparative and concern deciding which data will be used as input for data mining methods in the subsequent step. The third task corrects the data set by replacing hollow cells by their suitable predictions. Predictions are carried out as a result of a partial regression in the basis of the observable part of the data of the data set. The fourth task is considered exceptionally interesting because it homogenizes numerical measures by ensuring adequate conversion of quantitative variables into qualitative ones. The principal of conversion is to project qualitative variables on subspaces spanned by some quantitative variables.

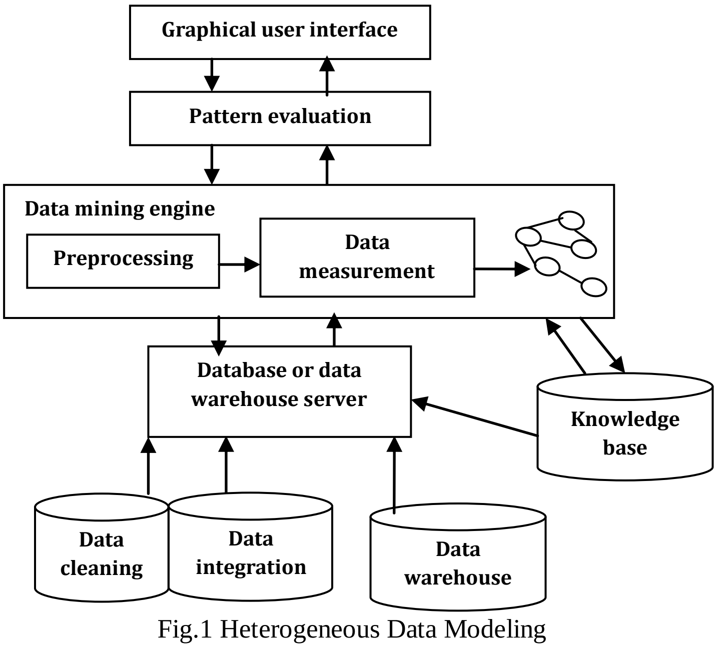


Fig.1 Heterogeneous Data Modeling

C. Clustering

In Fig.2 the finite mixture clustering is a popular unsupervised data analysis principal very useful for automatically classifying quantitative or qualitative objects in a database.

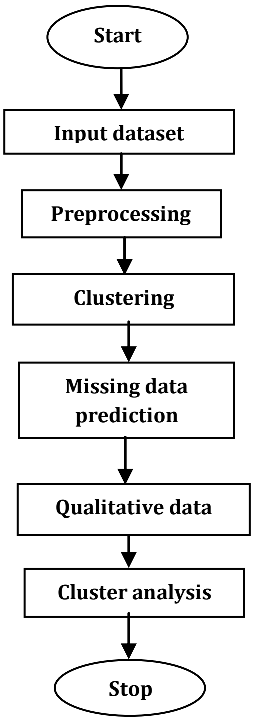


Fig.2 Data Clustering

It is pillar of exploratorydata mining used in many fields, including machine learning, pattern recognition, image processing, information discovery, and bioinformatics.One aspects of originality of such an approach is to consider aparameterization of the variance matrix of a cluster through its eigenvalue decomposition leading to many meaning meaningful models for clustering and classification. In clustering, observations are moved iteratively from one cluster to another, starting from an initial position. The number of groups has to be specified in advance and typically does not chance duration the course of the iteration

D. Clustering Interface

Fig.3shows the proposed approach that combines several techniques in order to analyze and model heterogeneous-type data. We consider that we observe another number of qualitative binary variables for all individuals.

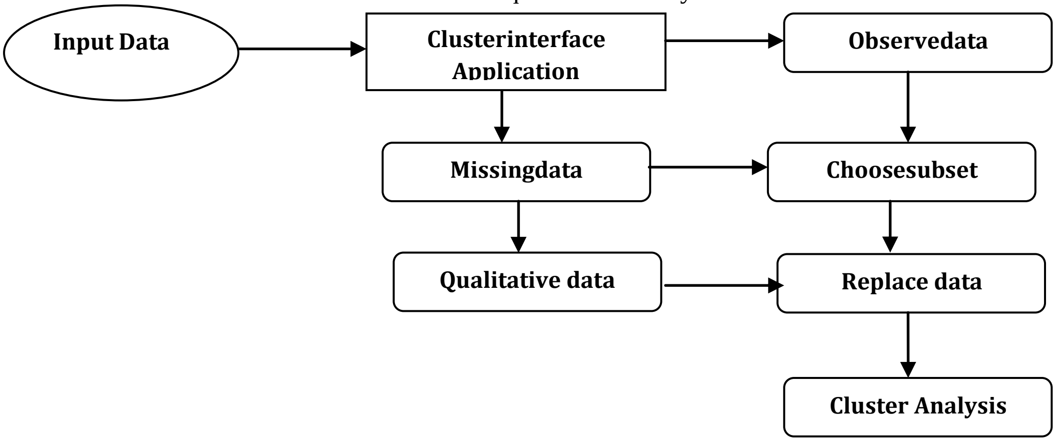


Fig.3 Clustering interface

The qualitative data set is essentially composed of binary responses that, either replaced (coded) quantitative unobservable ones, or simple has been used to code some categorical variables. The proposed algorithm can be summarized into three sub stages. The first one consists in interpolating the quantitative missing data, by firstly modeling the observed part of, and then, computing the predictions of the missing values. The fig.4 describes the qualitative and quantitative data conversionsThe first sub stage is principally the step-1 of table 1, which is exposed below. The step 2- consists in converting qualitative data to their quantitative predictions denoted and resulting from a logistic regression on the basis of the quantitative subsets of the dataset.The following Data mining algorithm is used for treating heterogeneous data sets.

Begin

1. From the standard data Set X, choose factors that contain missing values.
2. Predict the missing values using least square method and insert them.
3. For each column of the data set Y, choose from X, a set of variables make the prediction of the qualitative variables by a binomial logistic regression.

4. Compute the coefficient of the model and perform significance tests. Calculate the quantitative prediction to replace the qualitative ones.
5. Compose the new data set D′ = (W,P) containing only quantitative observations, where W is the normalized and repaired quantitative variables and P is the quantitative predictions replacing the qualitative data sets.
6. Perform a LPCA to D′ to reduce the variables dimension.
7. Clustering individuals on the reduced subspace. Information criteria allow choosing an adequate GM model, which essentially consists of on optimal number of clusters and appropriate covariance decomposition.

End

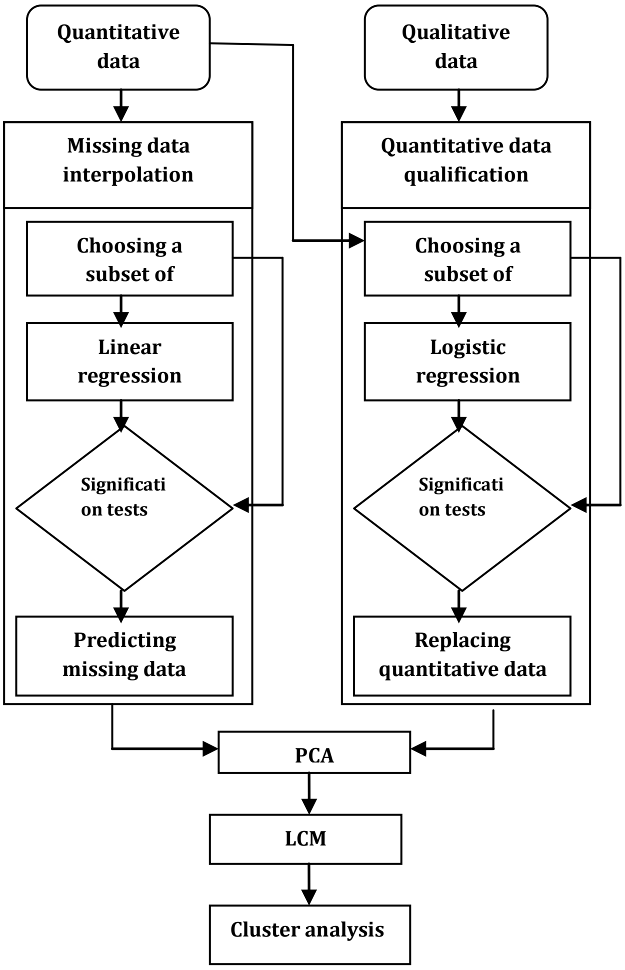
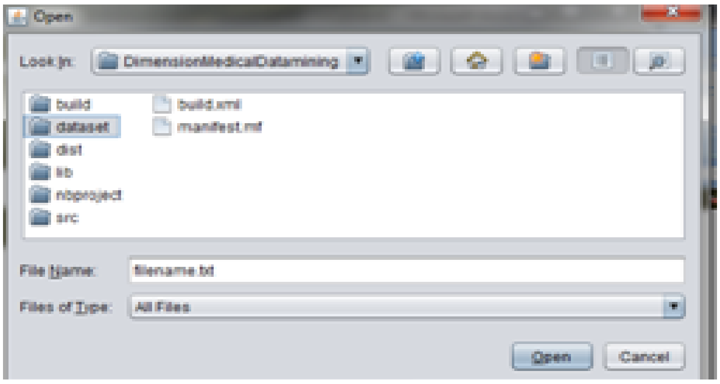


Fig.4 Qualitative and Quantitative attributes

IV. EXPERIMENTAL STUDY

Our study is based on a sample of Iris database .The choice of focusing on a iris structure is essentially grounded on the fact that it is often associated with traumas like stress, overwork physical and mental tiredness, etc.The input data set is 813*16 dimensional and consists of both qualitative and quantitative factors. Hence the data set will be divided into two sub sets X and Y .the first subset contains 12 quantitative variables X. the second subset is qualitative and contains 4 variables, Y. The general frame work of our proposed system is shown in fig.5.

Fig.5 LPCA clustering



Open

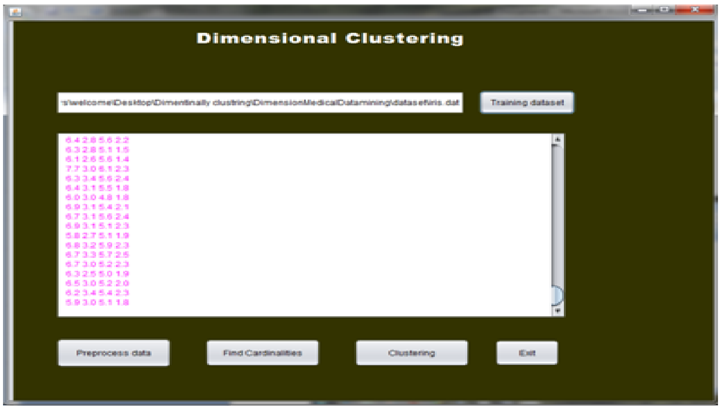
Look In: DimensionMedicalDatamining

build
dataset
dist
lib
nbproject
src
build.xml
manifest.mf

File Name: filename.txt

Files of Type: All Files

Fig.6 Dataset



Dimensional Clustering

s\welcome\Desktop\Dimentinally clustring\DimensionMedicalDatamining\dataset\iris.dat

```
6.4 2.8 5.6 2.2
6.3 2.8 5.1 1.5
6.1 2.6 5.6 1.4
7.7 3.0 6.1 2.3
6.3 3.4 5.6 2.4
6.4 3.1 5.5 1.8
6.0 3.0 4.8 1.8
6.9 3.1 5.4 2.1
6.7 3.1 5.6 2.4
6.9 3.1 5.1 2.3
5.8 2.7 5.1 1.9
6.8 3.2 5.9 2.3
6.7 3.3 5.7 2.5
6.7 3.0 5.2 2.3
6.3 2.5 5.0 1.9
6.5 3.0 5.2 2.0
6.2 3.4 5.4 2.3
5.9 3.0 5.1 1.8
```

Fig. 7 Matrix Data

In the execution part, first it starts dimensional clustering. Then, choose the training data sets as shown in Fig.6. It displays the height, width and staple of the iris. It shows column and rows of the data as shown in Fig.7. The values will be preprocessed. Here it finds the eigenvalues and eigenvector which is shown in Table 1.

Table 1 Eigen values and Eigen vectors

```
PREPROCESSING
 Input file name: C:\Users\welcome\Desktop\Dimentinally
clustring\DimensionMedicalDatamining\dataset\iris.dat
 No. of rows, n = 150, No. of cols, m = 4
 Input data sample follows as a check, first 4 values.
Value = 5.1, value = 3.5,value = 4.9, value = 3.0
 Matrix to be diagonalized
 0.21   0.04   0.13   0.12   0.16   0.09   0.15   0.09
 0.04   0.21   0.11   0.14   0.12   0.13   0.11   0.14
Eigenvectors (leftmost col <--> largest eval, first always = 1):
 1.0000   -0.9512   0.6850   1.5301   -0.5258   -0.2030   0.2776  1.6976
 1.0000    0.9532  -0.6864  -1.5333    0.5268   -0.2030   0.2776  1.6976
```

Innext step, it finds a cardinality it process the standard deviationof the values. The sample cardinality value is given below in Table 2.

Table 2 Cardinality values

```
Finding cardinality
Input file name: C:\Users\welcome\Desktop\Dimentinally
clustring\DimensionMedicalDatamining\dataset\iris.dat
 No. of rows, n    = 150, No. of columns, m  = 4,No. of clusters, clus = 5
 Variable means:  3.0573     3.7580     1.1993     5.8213
 Variable standard deviations: 0.4344     1.7594     0.7597     0.8865
 Cluster number, cardinality = 0 31.0
 Cluster number, cardinality = 1 30.0
Initial cluster means: 3.6000       1.4581       0.2355       4.9129
3.0200       2.4200       0.6400       5.1633
Compactness = 147.36772710419004
EPOCH NUMBER IS = 1
New compactness = 146.5101063218393      Final cluster cardinalities:
50.0000      12.0000      37.0000      29.0000      22.0000
```

In the clustering process, the data values are grouped together under a single cluster. The sample cluster value is given below in Table 3.

Table 3 Cluster values

```
Clustering
Input file name: C:\Users\welcome\Desktop\Dimentinally
clustring\DimensionMedicalDatamining\dataset\iris.dat
No. of rows, n row = 150, No. of cols, n col = 4
Input data sample follows as a check, first 4 values.
Value = 5.1, value = 3.5,value = 4.9,value = 3.0
Clus#1: 102; clus#2: 143; new card: 2.0; # clus left: 150; mindiss: 0.0
Clus#1: 8; Clus#2: 40; new card: 2.0; # clus left: 149; mindiss:
0.0049999999999999645
Clus#1: 1; Clus#2: 18; new card: 2.0; # clus left: 148; mindiss:
0.0049999999999999975
```

V. CONCLUSION

In this paper, we have addressed the issue of analyzing and modeling a multivariate data set composed of both quantitative and qualitative observations in a single research study. In this paper, propose a novel Lattice based Clustering Methodology (LCM) which uses lattice clustering technique and principal component analysis (PCA) technologies for modeling and analyzing heterogeneous dataset using binary coded factorial analysis. The Proposed methodology combines the advantages of lattice based factorial data analysis technique to achieve high accuracy in heterogeneous data mining. The simple interconnection has created a system that can fill a significant number of unobserved data, before allowing analyzing and modeling on the same dimensionally reduced subspace and heterogeneous set of quantitative and qualitative factors. The classification of the variability on the principal spaces has shown a compound structure of the data possibly due to the heterogeneous nature of the studied dataset.

REFERENCES

[1] A.Ahmadand, L. Dye, “A feature selection techniquefor classificatory analysis,”Pattern Recog.Lett.,pp.43-56, 2005. 1

[2] A. Ahmadand and L.Dye,“Ak-mean clustering algorithm formixednumeric and categorical data”, Data Knowl.Eng.,vol. 63,pp. 503–527,2007. 2

[3] C.Biernacki,G.Celeux,G.Govaert,andF.Langrognet,“Model-basedclusteranddiscriminantanalysiswiththemixmodsoftware,”Comput.Stat. Data Anal.,vol. 51,pp. 587–600, 2006. 5

[4] M.Chavent,V.Kuentz-Simonet,andJ.Saracco,“OrthogonalrotationinPCAMIX,”Adv. Data Anal. Classification,vol.6,pp.131–146,2012. 6

[5] A. DiCiacco, “Simultaneous clustering of qualitative and quantitative data with missing observations,” Statistical Applicata, vol. 4, pp.599–609, 1992. 7

[6] C.Doring,C.Borgelt,andR.Kruse,“Fuzzy clustering of quantitativeand qualitative data,” inProc. IEEEAnnul.Meet.Fuzzy Inf. (NAFIPS2004), 2004, pp. 84–89. 8