The Design of Pre-Processing Multidimensional Data Based on Component Analysis

Rahmat Widia Sembiring & Jasni Mohamad Zain
Faculty of Computer System and Software Engineering, Universiti Malaysia Pahang
Lebuhraya Tun Razak, 26300, Kuantan, Pahang Darul Makmur, Malaysia
E-mail: rahmatws@yahoo.com, jasni@ump.edu.my

Received: February 17, 2011 Accepted: March 27, 2011 doi:10.5539/cis.v4n3p106

Abstract

Increased implementation of new databases related to multidimensional data involving techniques to support efficient query process, create opportunities for more extensive research. Pre-processing is required because of lack of data attribute values, noisy data, errors, inconsistencies or outliers and differences in coding. Several types of pre-processing based on component analysis will be carried out for cleaning, data integration and transformation, as well as to reduce the dimensions. Component analysis can be done by statistical methods, with the aim to separate the various sources of data into a statistical pattern independent. This paper aims to improve the quality of pre-processed data based on component analysis. RapidMiner is used for data pre-processing using FastICA algorithm. Kernel K-mean is used to cluster the pre-processed data and Expectation Maximization (EM) is used to model. The model was tested using wisconsin breast cancer datasets, lung cancer datasets and prostate cancer datasets. The result shows that the performance of the cluster vector value is higher and the processing time is shorter.

Keywords: Pre-processing data, Data cleansing, Data noisy, FastICA

1. Introduction

Applications related to multidimensional data continue to grow. Techniques to support more efficient query becomes an important research issue at this time. This technique is needed to open the multimedia content, data exploration in areas of health, population issues, decision-making in education, as well as to analyse the time-series. Processes such as data pre- processing, data cleaning, data integration and transformation, and reduction of dimension can be applied to improve the quality of the results.

Real data are often incomplete (Magnani, et.al., 2004), lack of attributes, noisy, contains outlier, and also inconsistent, thus requiring the data pre-processing. Pre-processing of data is to improve algorithm (Orfanidis, et.al., 2008), accuracy, completeness, consistency, timeliness, value added, interpretation, and better accessibility.

Pre-processing is the process of transforming data into simpler, more effective, and in accordance with user needs. More accurate results and shorter computation time can be used as indicators. The data also becomes smaller without changing the information in it. Some pre-processing method is done by selecting a subset of a large population sample of data, referred to as denoising. This will be followed by normalization and feature extraction.

Dimensional reduction becomes a fundamental problem in most of the data mining process. It benefits not only for computational efficiency, but also can improve the accuracy of the analysis (Cunningham, et.al.,2007). Dimension reduction techniques are often used to overcome "the curse of dimensionality", as part of pre-processing in addition to simplify the data model.

Dimension reduction techniques can be grouped into feature selection and feature extraction. Feature selection is the process of finding a subset of the original variables, with the aim to reduce and eliminate the noise dimension. It can improve the performance of data mining, including improving the speed and accuracy. In some cases, regression or classification analysis can be done to reduce the dimension, which produces more accurate dimensions. Several algorithms have been proposed such as ReliefF (Sikonja, et.al., 2003), Focus, Support Vector Machine Recursive Feature Elimination (SVM RFE) and Feature Subset Selection using Expectation Maximization (FSSEM).

Feature extraction is a technique to transform high-dimensional data into lower dimensions. Several supervised learning algorithms have been proposed, namely Linear Discriminant Analysis (LDA), Canonical Correlation Analysis (CCA), Partial Least Square (PLS), Latent Semantic Indexing (LSI), Singular Value Decomposition

(SVD). While for unsupervised learning, algorithms such as Principal Component Analysis (PCA), Independent Component Analysis (ICA), FastICA (extension of ICA) can be used as a basic component analysis.

This paper organized into a few sections. Section 2 will present related work. Section 3 presents material and method, followed by result and discussion in Section 4, and followed by concluding remarks in Section 5.

2. Related Work

2.1 Pre-processing Data

Research in pre-processing has been done, and produced commercial products. However, given the number of attributes and multidimensional data continues to grow, this research has the potential to grow. *DB-H* algorithm is one of the researches relating to the pre-processing of data, namely with discretizes technique to eliminate the numerical attributes and generalizes by eliminating the symbolic attributes (Hu, et.al, 2003). Pre-processing and data transformation is often required before applying the data mining of clinical data, namely the compilation of data using information from the data itself (Lin, et.al, 2009). This process is used to ensure reliability of data used in data mining (Wahab, et.al., 2008).

2.2 Dimension Reduction

Dimension reduction methods associated with regression, additive models, neural network models, and methods of Hessian (Fodor, et.al. 2003). Local Dimension Reduction (LDR) looks for relationships in the dataset and reduces the dimensions of each individual using a multidimensional index structure (Chakrabarti, et.al., 2000). Nonlinear algorithm gives better performance than PCA for sound and image data (Kambhatla, et.al. 1994). Principal Component Analysis (PCA) which is based on dimension reduction and texture classification scheme can be applied to manifold statistical framework (Sang, et.al., 2007). The semantics of linear algebra is significantly simpler than Probabilistic Latent Semantic Analysis (PLSA) and LDA, while PLSA is much simpler than the LDA (Chua, et.al., 2009).

In most applications, dimension reduction performed as pre-processing step (Ding, et.al. 2007), performed with traditional statistical methods that will parse an increasing number of observations (Fodor, et.al., 2003). Dimension reduction creates a more effective domain characterization (Bi). Sufficient Dimension Reduction (SDR) is a generalization of nonlinear regression problems, where the extraction of features is as important as the matrix factorization (Globerson, et.al. 2003), while SSDR (Semi-Supervised Dimension Reduction) is used to maintain the original structure of high dimensional data (Zhang, et.al., 2008).

2.3 Normalization

Important aspect of pre-processing of data is the normalization (Shalabi, et.al, 2007). There are many methods of doing data normalization, i.e. normalization min-max, z-score, and normalization with the decimal scale (Han, et.al., 1998). Normalization *min-max* performed with a linear transformation on the original data. Suppose min_A and max_A is the minimum and maximum values of an attribute A. Normalization Min-max will map the value v, A into v' in the range [new_min_A, new_max_A] by using the formula $v' = \frac{v-min_A}{max_A-min_A}(new\_max_A - new\_min_A) + new\_min_A$

In the normalization *z-score*, value for attribute A, normalized by the mean and standard deviation values A. Value v for A normalized to v' by counting $v' = \frac{v-\bar{A}}{\sigma A}$, where $A^{-}$and σA are *mean* and standard deviation of attribute A. Normalization method is useful if the minimum and maximum value of attribute A is unknown, or when there are outliers that dominate normalization *min-max*.

Normalization by decimal scale by moving the decimal point normalized attribute values A. The number of decimal points moved depends on the absolute value of A. The value v from A normalized to v' count by $v' = \frac{v}{10^{j'}}$, where j is minimum value of $Max(|v'|) < 1$.

To achieve a good dataset we need to pay attention to data noise, which may contain errors or outliers. Binning method (Davis, et.al. 2007) will be used in this research to deal with data noise. This can be further improved by using clustering, semi-automatic method, and regression. There are several other factors that need to be addressed, such as measurement information (Jebara, et.al., 2000, Chen, et.al., 2004, Shutin, et.al., 2010, Harrie, et.al., 2009), measuring the distance (Sikonja, et.al., 2003, Wong, et.al., 2010, Zheng,et.al, 2010, Xu, et.al, 2010, Li, et.al, 2008, Lee, et.al, 2010), the measurement of dependence (Claesken, et.al, 2002, Kao, et.al, 2009, Bukor, et.al, 2009, Warrens, et.al, 2009), consistency, and accuracy.

2.4 Outlier

Outlier which often also be interpreted as an anomaly, is a set of data that is considered to have different properties compared with other data. Outlier analysis is also known as anomaly analysis or anomaly detection, or deviation detection (object attribute values are significantly different from others). Outlier based on density-based approach, where the outlier is a point which is located in an area with low density. To find outliers, we can use the formula:

$$density(x,k) = \left(\frac{\sum_{y \in N(x,k)} dist(x,y)}{|N(x,k)|} \right)^{-1}$$

where *N(x, k)* is the set containing the k nearest neighbours *x*, *y* is the nearest neighbour of *x* and |*N(x,k)*| is the number of members of the set *N(x, k)*. Meanwhile, to calculate the LOF (local outlier factor) can be done with the approach of:

$$average\_relative\_density(x,k) = \frac{density(x,k)}{\sum_{y \in N(x,k)} density(y,k) / |N(x,k)|}$$

3. Material and Methods

The proposed design consist of five stages, i.e. data cleansing, data denoising, data extraction, data clustering & cluster modelling and data visualization (*Figure 1*).

3.1 Data Cleansing

Most of real data cannot be used directly because of the lack of the value attribute, or containing only aggregate data. Data can also be noisy because it contains errors, have outliers, or is not consistent due to differences in coding or naming conventions. This can be solved by cleaning the data. The cleaning of data starts with the process of centering, to reduce the data by finding the average of each attribute, using the formula: $\hat{X} = X - \bar{X}$, where $\hat{X}$ is the result after *centring*, *X* is column vector and $\hat{X}$ is the average of the corresponding column. The process of centering done for all in order, if null value is found, the value will be replaced by an average value to that column, the result of the centering process can be used to find the spread by using the formula *Scatter* = $\hat{X}\,'\,\widehat{X}$

The results of the scatter can be used to find the value of covariance using the formula, $Kovarian = \frac{\hat{X}\,'\hat{X}}{m-1}$. After the process of centering followed by a Gaussian function, hereinafter referred to as normalization, by the formula $\hat{X} = \frac{X - \bar{X}}{\sigma x}$

3.2 Data Denoising

After cleaning, the data will be denoise using Binning method. This study will use the outlier detection method with D_n^k (Ramaswamy et.al., 2000) where D=distance, k=nearest neighbour, and n=top n point. Denoising outlier data can be done through a search by an equal size distance (Knorr, et.al, 1997). This method states that every object with the greatest distance from the k-nearest neighbor can be called outliers from data set.

3.3 Data Extraction

Independent Component Analysis (ICA) was introduced by Jeanny Hérault and Christian Jutten in 1986, later developed by Pierre Comon in 1994. FastICA is one of the extensions of ICA, which is based on point iteration scheme to find nongaussianity (Hyvaerinen, et.al., 2000). It can also be derived as approximate Newton iteration, using the following formula $W^{+} = W + diag(\alpha_i)[diag(\beta_i) + E\{g(y)y^T\}]W$, where $y = W_x, \beta_i = -E\{y_i g(y_i)\}$ and $\propto_i = -1/(\beta_i - E\{g'(y_i)\})$, matrices W need to *orthogonalized* after each phase have been processed.

3.4 Data Clustering

This research implements the Kernel K-Mean clustering, with k = 2 and 100 sets maximum optimization. Radial kernel used by the kernel gamma = 1.0. For the cluster model used Expectation Maximization (EM), with maximum runs = 5, maximum optimization steps = 100, with quality 1.0E-10, and with k-mean initial distribution runs.

3.5 Data Visualization

Most of the data mining research is a predictive modelling (Kohavi, et.al., 2000), with the primary tasks are defining goals and to measure the predictive test-set independency. Data visualization is another important factor so that users understand the result of applying the model. The visualization is represented in 3-dimensional image.

4. Result and Discussion

This paper proposed a pre-processing model using component analysis, as shown in Figure 1. The model is then tested against three medical datasets (cancer datasets):

a. Wisconsin breast cancer dataset (Mangasarian, et.al., 1990) with 698 example sets, and 22 attributes.

b. Lung cancer dataset (Hong, et.al., 1991) with 31 example sets and 6 attributes.

c. Prostate cancer dataset (Byar, et.al., 1980) with 502 example datasets and 36 attributes.

To view the comparison of model results, we have conducted two tests of classification, namely the implementation of pre-processing, FastICA and clustering, as shown in Figure 2, and compared the results with no pre-processing, as shown in Figure 3. The overall results of experiments on the two models above can be seen in Table 3. Overall, these experiments have been carried out except for prostate cancer datasets with the test datasets without pre-processing has yet to find results despite testing more than seven hours.

Analysis of the performance vector for the number of clusters generated showed better value in all three datasets, i.e. from 0.990 to 0.993 for wisconsin breast cancer dataset, 0.867 to 0.882 for lung cancer datasets and 0.991 for prostate cancer datasets. Meanwhile, if viewed from the time of processing, the implementation of the model with the application of pre-processing also showed positive results, for the three datasets, i.e. 63 to 61 seconds for testing wisconsin breast cancer datasets, 7 to 5 seconds lung cancer datasets and 46 seconds for prostate cancer datasets.

Cluster modelling results with the implementation of EM have also been carried out, for the application of pre-processing of data as shown in Figure 4 for wisconsin breast cancer dataset, Figure 6 for the lung cancer dataset and Figure 8 for prostate cancer datasets. Implementation of the model without pre-processing for wisconsin breast cancer datasets results as shown in Figure 5, to lung cancer datasets in Figure 7, while for prostate cancer datasets found no results and shows no significant difference.

5. Conclusion

This paper aims to improve the quality of pre-processed data. We proposed a model for the design of pre-processing multidimensional data based on component analysis. RapidMiner is used for data pre-processing using FastICA algorithm. Kernel K-mean is used to cluster the pre-processed data and Expectation Maximization (EM) is used to model the cluster. The model was tested using wisconsin breast cancer datasets, lung cancer datasets and prostate cancer datasets. The result shows that the performance of the cluster vector value is higher and the processing time is shorter.

Acknowledgment

The authors wish to thank Universiti Malaysia Pahang. This work supported in part by a grant from GRS090116.

References

Bi, Jinbo, Kristin Bennett, Mark Embrechts, Curt Breneman and Minghu Song. (2003). Dimensionality Reduction via Sparse Support Vector Machine, *Journal of Machine Learning Research* 3 (2003) p.1229-1243, [doi>10.1.1.10.203].

Bukor, Jozsef, Ladislaf Misik, Janos T. Toth. (2009). *Dependence of densities on a parameter,* [doi>10.1016/j.ins.2009.04.014].

Byar, DP, Green SB. (1980). *Bulletin Cancer,* Paris 67:477-488, http://lib.stat.cmu.edu/S/Harrell/data/xls/prostate.xls.

Chakrabarti, Kaushik, Shrada Mehrotra. (2000). Local Dimensionality Reduction: A New Approach to Indexing High Dimensional Spaces, Proceedings of the 26th VLDB Conference, Cairo, Egypt, [doi> 10.1.1.36.4091].

Chen, Pei, David Suter. (2004). Recovering the Missing Components in a Large Noisy Low-Rank, *IEEE Transaction on Pattern Analysis and Machine Intelligence,* pp.1051-1063, [doi>10.1109/tpami.2004.52].

Chua, Freddy Chong Tat. (2009). Dimensionality Reduction and Clustering of Text Document, [Online] Available: www.mysmu.edu/phdis2009/freddy.chua.2009/papers/probabilistic.pdf.

Claesken, Gerda, Peter Hall. (2002). Effect of Dependence on Stochastic Measures of Accuracy of Density Estimators, *The Annual of Statistics* 2002, Vol.50 No.2, pp.451-454.

Cunningham, Pádraig. (2007). Dimension Reduction, *Technical Report* UCD-CSI-2007-7, [doi> 10.1.1.98.1478].

Davis, Richard. A, Adrian J. Charlton, John Godward, Stephen A. Jones, Mark Harrison, Julie C. Wilson. (2007). Adaptive binning: An improved binning method for metabolomics data using the un-decimated wavelet transform, [doi> 10.1016/j.chemolab.2006.08.014].

Ding, Chris, Tao Li. (2007). Adaptive Dimension Reduction Using Discriminant Analysis and K-means Clustering, International Conference on Machine Learning, Corvallis, OR, 2007, [doi>10.1.1.118.2712].

Fodor, Imola K. (2003). *A Survey of Dimension Reduction Technique,* [Online] Available: http://citeseerx.ist.psu.edu, [doi>10.1.1.8.5098]

Globerson, Amir, Naftali Tishby. (2003). Sufficient Dimensionality Reduction, Journal of Machine Learning Research 3, pp. 1307-1331, [doi>10.1.1.2.6467]

Han, Jiawei, Shojiro Nishio, Hiroyuki Kawano, Wei Wang, (1998), Generalization-based data mining in object-oriented databases using an object cube model, *Data & Knowledge Engineering* 25, pp.55-97.

Harrie, Lars, Hanna Stigmar. (2009). *An evaluation of measures for quantifying map information*, [doi> 10.1016/j.isprsjprs.2009.05. 004]

Hong, Z.Q. and Yang, J.Y. (1991). Optimal Discriminant Plane for a Small Number of Samples and Design Method of Classifier on the Plane, *Pattern Recognition,* Vol. 24, No. 4, pp. 317-324

Hu, Xiaohua. (2003). *DB-H Reduction- A Data Preprocessing Algorithm for Data Mining Applications,* [doi>10.1016/S0893-9659(03)90013-9].

Hyvaerinen, Aapo, Erkki Oja. (2000). Independent Component Analysis: Algorithms and Applications, Neural Networks, 13(4-5), pp. 411-430.

Jebara, Tony, Tommi Jaakola. (2000). Feature Selection and Dualities in Maximum Entropy Discrimination, [doi>10.1.1.26.9953].

Kambhatla, Nanda , Todd K. Leen. (1994). *Fast "Non_Linear Dimension Reduction",* [doi> 10.1.1.41.1646].

Kao, Shih-Chieh, Auroop R. Ganguly, Karsten Steinhaeuser. (2009). Motivating Complex Dependence Structures in Data Mining: A Case Study with Anomaly Detection in Climate, [doi>10.1109/icdmw.2009.37].

Knorr, Edwin M., Raymond T. Ng. (1997). *A Unified Approach for Mining Outliers,* [doi>10.1.1.45.9715]

Kohavi, Ron. (2000). *Data Mining and Visualization, National Academy of Engineering (NAE),* [Online] Available: http://citeseerx.ist.psu.edu.

Lee, Paul H, Philip L.H, Yu. (2010). *Distance based tree models for ranking data,* [doi> 10.1162/neco.1991.3.1.79].

Li, Yuanhong, Ming Dong, Jing Hua. (2008). *Localized feature selection for clustering,* [doi>10.1016/j.patrec.2007.08.012].

Lin, Ping, Zhenming Lei, Luying Chen, Jie Yang, Fang Liu. (2009). Decision tree network traffic classifier via adaptive hierarchical clustering for imperfect training dataset, Wireless Communications, Networking and Mobile Computing, 2009. WiCom '09. 5th International Conference on, pp.1-6, [doi> 10.1109/wicom.2009.5302133].

Magnani, Matteo, Danilo Montesi. (2004). A New Reparation Method for Incomplete Data in the Context of Supervised Learning, ITCC, vol. 1, pp.471, International Conference on Information Technology: Coding and Computing (ITCC'04) Volume 1.

Mangasarian, O. L. and W. H. Wolberg. (1990). *Cancer diagnosis via linear programming,* SIAM News, Volume 23, Number 5, September 1990, pp 1 & 18.

Orfanidis, Paraskevas, David J. Russomanno. (2008). *International Journal of Business Intelligence and Data Mining (IJBIDM),* Vol. 3, No. 2.

Ramaswamy, Sridhar, Rajeev Rastogi, Kyuseok Shim. (2000). Efficient Algorithms for Mining Outliers from Large DataSets, [doi> 10.1145/342009.335437]

Sang, Mook Lee, A. Lynn Abbott, Philip A. Araman. (2007). *Dimensionality Reduction and Clustering on Statistical Manifolds,* [doi> 10.1.1.134.6544]

Shalabi, Luai Al, Zyad Shaaban. (2007). Normalization as a Preprocessing Engine for Data Mining and the Approach of Preference Matrix, International Conference on Dependability of Computer Systems, [doi>10.1109/depcos-elcomex.2006.38]

Shutin, Dmitriy, Olga Zlobinskaya. (2010). Application of Information-theoretic Measures to Quantitative Analysis of Immuno fluorescent microscope imaging, [doi>10.1016/j.cmpb. 2009.05.009]

Sikonja, Marko Robnik, I. Kononenko, (2003), Theoretical and Emphiritical Analysis of ReliefF and RReliefF, *Machine Learning,* 53, p.23–69, [doi> 10.1111/j.1467-8640.1990.tb00298.x]

Wahab, Mohd Helmy Abd, Mohd Norzali Haji Mohd, Hafizul Fahri Hanafi, Mohamad Farhan. (2008). Data Pre-processing on Web Server Logs for Generalized Association Rules Mining Algorithm [doi>10.1.1.140.5102].

Warrens, Matthijs J., Willem J. Heiser. (2009). Diagnostics for regression dependence in tables re-ordered by the dominant correspondence analysis solution, [doi> 10.1016/j.csda.2008. 07.035].

Wong, Tzu-TsungWong, Kuan-LiangLiu. (2010). A probabilistic mechanism based on clustering analysis and distance measure for subset gene selection, [doi> 10.1093/bioinformatics/btm207].

Xu, Zeshui. (2010). A method based on distance measure for interval-valued intuitionistic fuzzy group decision making, [doi> 10.1002/int. v20:8].

Zhang, Daoqiang, Hua Zhou Zhi, Songcan Chen. (2008). Semi-Supervised Dimensionality Reduction, 7th SIAM International Conference on Data Mining, [doi>10.1.1.102.7089]

Zheng, Jianping, Jiangjiao Duan, Chengrong Wu. (2010). *A new distance measure for hidden Markov models,* [doi> 10.1016/j.eswa.2007. 11.021]

Table 1. Origin Example Datasets

skc0101	skc0102	skc0103	skc0201	skc0202	skc0203	skc0204	skc0301	skc0302	skc0303	skc0304	skc0305	skc0401	skc0402	skc0403
8	6	5	10	6	1	10	5	4	6	2	9	4	3	5
7	3	1	6	2	7	10	7	4	7	1	3	3	8	1
4	10	10	4	9	6	1	7	1	🟥	2	4	8	9	2
3	9	2	10	6	8	2	10	1	8	9	3	5	3	4
9	2	🟥	8	3	8	7	8	4	4	8	🟥	1	8	1
6	10	7	5	2	6	7	8	5	2	8	10	1	10	4
10	2	5	2	2	🟥	2	1	9	4	5	5	1	3	🟥
6	4	1	1	7	1	8	3	1	5	7	5	6	7	6
9	9	5	3	9	6	2	10	5	6	3	8	8	10	3
10	8	7	10	6	6	1	5	6	7	4	3	3	5	3

Table 2. Example Datasets after Cleaning

skc0101	skc0102	skc0103	skc0201	skc0202	skc0203	skc0204	skc0301	skc0302	skc0303	skc0304	skc0305	skc0401	skc0402	skc0403
8	6	5	10	6	1	10	5	4	6	2	9	4	3	5
7	3	1	6	2	7	10	7	4	7	1	3	3	8	1
4	10	10	4	9	6	1	7	1	5	2	4	8	9	2
3	9	2	10	6	8	2	10	1	8	9	3	5	3	4
9	2	5	8	3	8	7	8	4	4	8	6	1	8	1
6	10	7	5	2	6	7	8	5	2	8	10	1	10	4
10	2	5	2	2	5	2	1	9	4	5	5	1	3	3
6	4	1	1	7	1	8	3	1	5	7	5	6	7	6
9	9	5	3	9	6	2	10	5	6	3	8	8	10	3
10	8	7	10	6	6	1	5	6	7	4	3	3	5	3

Table 3. Experiment Result

	Wisconsin breast cancer datasets		Lung cancer datasets		Prostate cancer datasets	
	No pre-processing	With pre-processing	No pre-processing	With pre-processing	No pre-processing	With pre-processing
Datasets:					More than 7 hours uncompleted the process	
Example Set	698	698	31	31		502
Special Attributes	0	0	4	4		0
Regular Attributes	22	22	2	2		36
Cluster Model (Kernel K-Mean)						
Cluster 0	261	261	17	17		250
Cluster 1	437	437	14	14		252
Cluster Model (EM)						
Cluster probabilities						
Cluster 0	0.52	0.52	0.46	0.42		0.57
Cluster 1	0.48	0.48	0.54	0.57		0.42
Cluster means						
Cluster 0	0.58 ; -0.48	0.58 ; -0.48	-0.77 ; -0.27	-0.87 ; -0.37		-0.02 ; -0.72
Cluster 1	-0.63 ; 0.52	-0.63 ; 0.52	-0.67 ; 0.23	0.64 ; 0.28		0.03 ; 0.95
Cluster covariance matrices						
Cluster 0	0.54 0.51 0.51 0.62	0.54 0.51 0.51 0.62	0.49 -0.51 -0.51 0.99	0.20 -0.23 -0.23 0.50		0.98 -0.13 -0.13 0.38
Cluster 1	0.78 0.08 0.08 0.89	0.78 0.08 0.08 0.89	0.48 0.12 0.11 0.90	0.63 -0.24 -0.25 1.18		1.01 0.11 -0.12 1.22
Performance Vector						
Number of clusters	0.990	0.993	0.867	0.882		0.991
FastICA						
Number of Component	2	2	2	2		
Time to process	63s	61s	7s	5s		46s

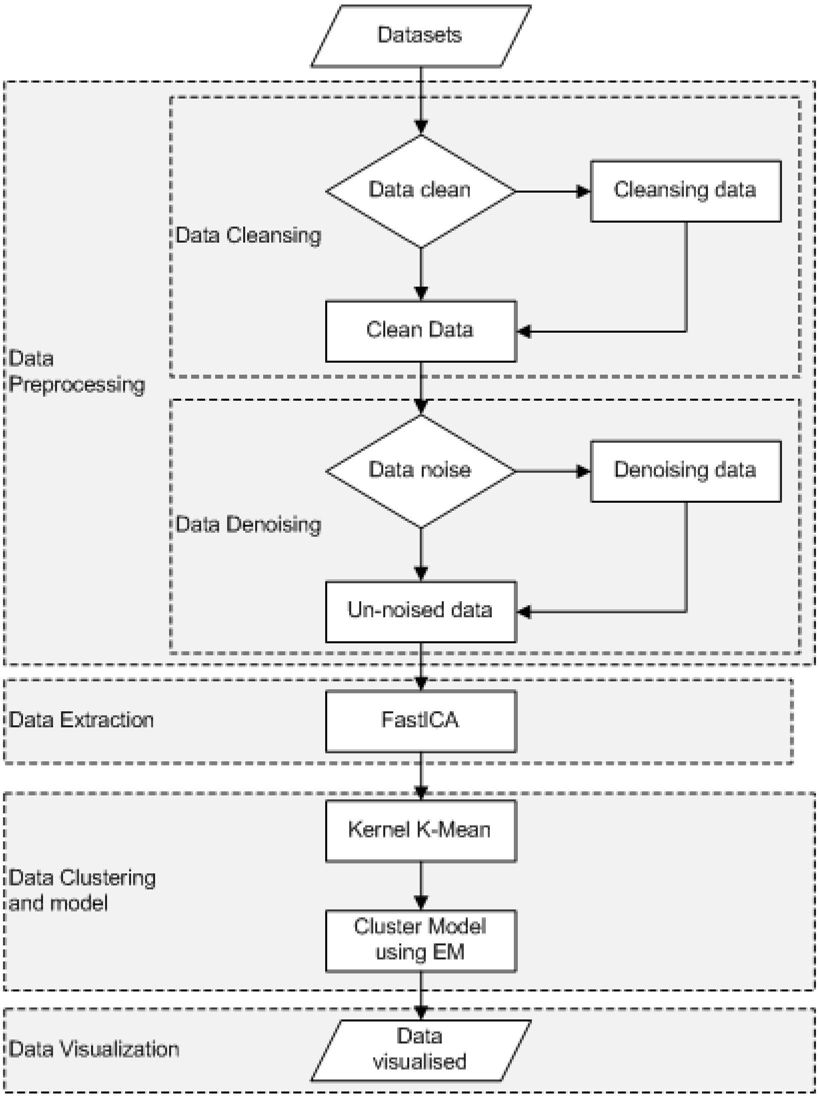


Figure 1. Pre-processing model based on component analysis

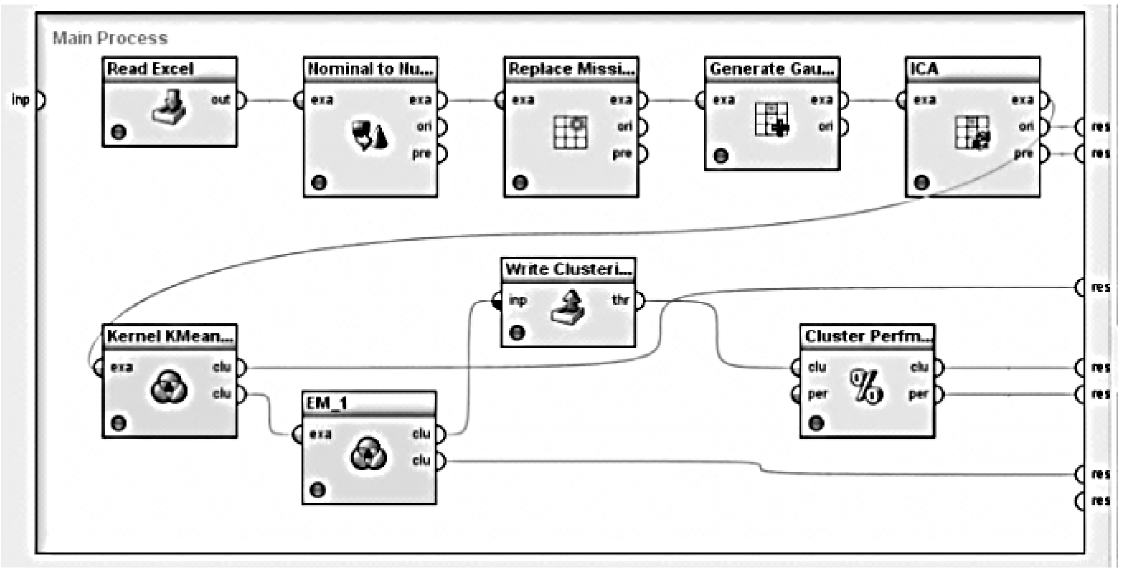


Figure 2. Phase Using Pre-processing data and Fast ICA

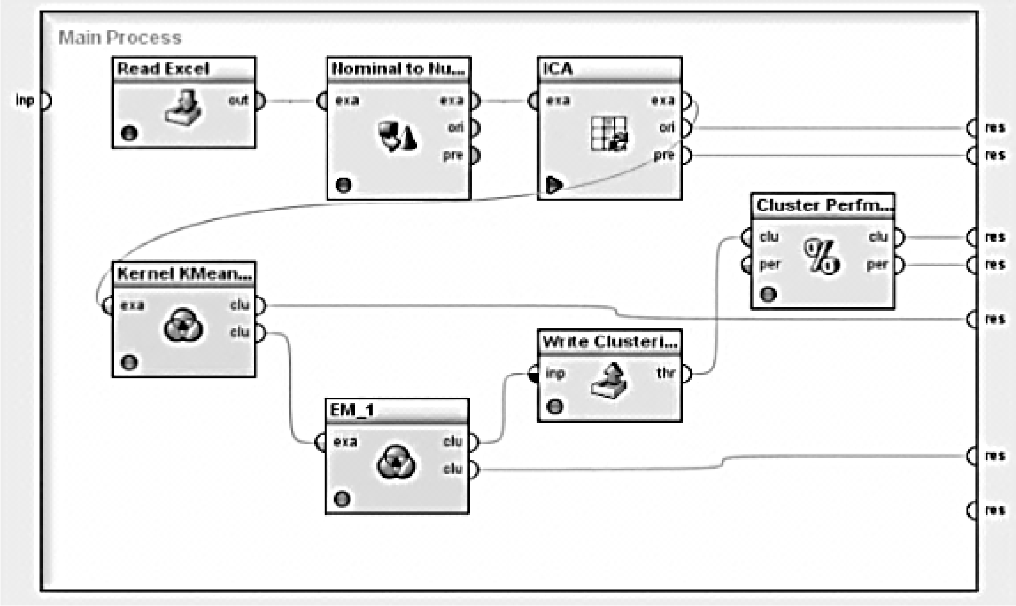
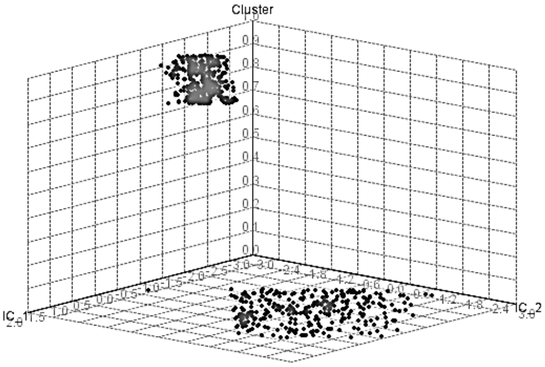


Figure 3. Phase Using Pre-processing data and Fast ICA



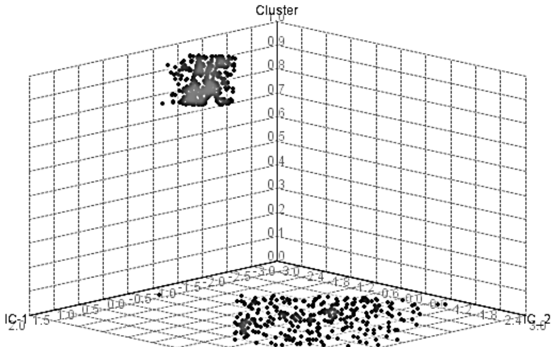
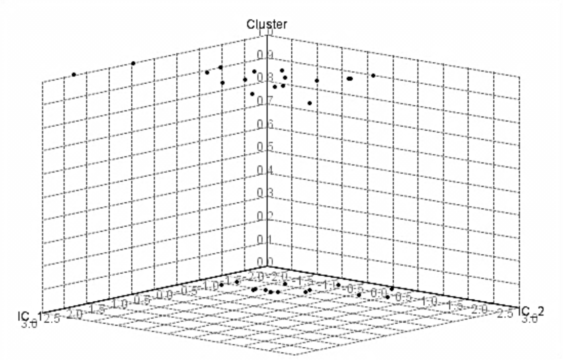


Figure 4. Wisconsin breast cancer datasets clustering using FastICA, Kernel K-mean and EM through pre-processing

Figure 5. Wisconsin breast cancer datasets clustering using FastICA, Kernel K-mean and EM without pre-processing



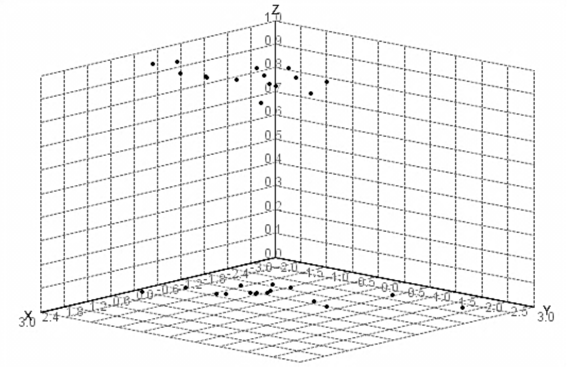


Figure 6. Lung cancer datasets clustering using FastICA, Kernel K-mean and EM through pre-processing

Figure 7. Lung cancer datasets using FastICA, Kernel K-mean and EM without pre-processing

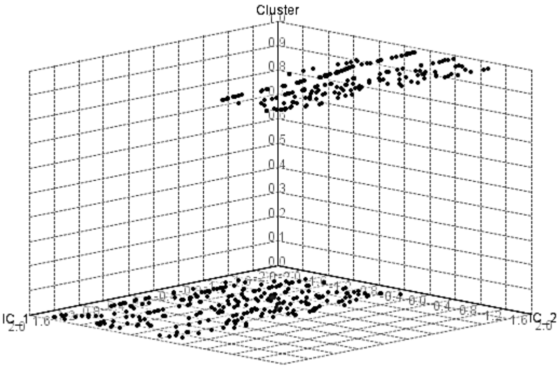


Figure 8. Prostate cancer datasets clustering using FastICA, Kernel K-mean and EM through pre-processing