DOI: http://dx.doi.org/10.26483/ijarcs.v9i1.5191

Volume 9, No. 1, January-February 2018

ISSN No. 0976-5697

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

# SELF-ORGANIZING MAP BASED CLUSTERING MODEL BY ANALYZING EIGEN SYSTEM OF PCA

Parthajit Roy
Department of Computer Science
The University of Burdwan
West Bengal, India-713104

Swati Adhikari
Department of Computer Science
The University of Burdwan
West Bengal, India-713104

***Abstract:*** Two novel clustering techniques, based on Principal Component Analysis (PCA), have been proposed in this paper that use Self Organizing Map as clustering model. The proposed models are differed by the number of principal components selection techniques in PCAand are applicable on clustering of non-categorical data. The present paper proposes, either to cluster the eigenvalues or to cluster the eigenvectors of the covariance matrix of the associated dataset in order to determine the number of principal components to be selected in PCA. It is also proposed that it is possible to further improve the performance of the SOM based clustering model by using either of the proposed techniques to select number of principal components.The benchmark *wine* dataset is used for testing purpose. Two existing principal components selection methods are used to evaluate the proposed clustering models.

***Keywords:*** Cluster Analysis; Self Organizing Map; Principal Component Analysis.

## 1. INTRODUCTION

In present day's scenario, the term *cluster analysis* has become very much familiar in the domain of pattern recognition. The notion of cluster analysis is to search for similar patterns in available data and group them to solve specific problems in an unsupervised manner i.e. when no labeled data are available at all. This is also afundamental task in pattern recognition.Although, the grouping of similar data in pattern recognition can also be done in a supervised waybut this time the task becomes trivial. When it is not possible to have labeled data, the task of finding similarity among data becomes absolutely complicated.This makes the task of clustering very challenging and attracts many researchers in this field as well.

Cluster analysis has many applications in different disciplines of pattern recognition [1] likedata mining [2], document classification [3], image processing [4], drug discovery [5] and many more.That means, cluster analysis can be applicable in any field where data need to be classifiedor retrieved based on some similarity value.

There exist a large number of standard clustering algorithmslike Hierarchical clustering algorithm, K-means algorithm, Self-Organizing Map (SOM)based algorithm etc. [6]. Along withthe merits and demerits of these algorithms,search for new one will always continue in an expectation to have more improved one.

Higher dimensions of data and scaling are two major clustering problems on which the performance of any clustering algorithm is highly dependent. Actually, data in real world exist with varying dimensions. When the dimension of data becomes higher, the task of clustering is also become complex and needs lot of computation time and storage space. Due to the other clustering problem, one attribute may influence other attributes to a great extent. So, these two factors have a great impact on accuracy of the results of clustering. Some pre-processing techniques,like normalization or some dimensionality reduction techniques prior to application of the actual clustering algorithm may help to overcome these problems.

To reduce dimensions in data, SOM and PCA are used. SOM is an artificial neural network based algorithm that groups data in an unsupervised way and projects data in higher dimensional space to a lower dimensional one in such a way that the topological order among data is also preserved [7]. On the other hand, PCA is a technique that selects only most informative attributes from the original dataset.

In the process of PCA,covariance matrix of a dataset and eigenvalues and eigenvectors of the same play a significant role. Covariance matrix is used to measure the spread of data in a higher dimensional dataset. The relevant information of data is captured by the eigenvalues and eigenvectors of the covariance matrix. So, these two things are considered in PCA algorithm while selecting principal components.

A vital point in PCA is to decide about the number of principal components to be selected.This considerably affects the performance of any PCA based model. Numbers of methods have been adopted to select principal components. Among these Kaiser Criterion and Cumulative Percentage of Total Variations are two standard methods of selecting principal components [8].

In this paper, two techniques have been proposed to fix the number of components to be selected in PCA. For this, one of the proposed method clusters the eigenvalues of the covariance matrix and the other proposed method clusters the eigenvectors of the same.The number of principal components to be selected is the number of members in the cluster with highest number of members in it.The selected principal components are clustered by using SOM algorithm in both of the proposed methods.It is also proposed in this paper that the performance of the PCA based SOM model can also be made more robust by the use of these two newly proposed principal component selection methods in comparison with the two existing methods.

The organization of the rest of the paper is as follows: Literature review appears in Section II. Section IIdiscusses the existing methods. Section IV is about the proposed methods. Section Vdescribesthe experimental setup followed by the discussion of Result and analysis which appears in Section VI and conclusion is given in Section VII. References come thereafter.

## 2. LITERATURE REVIEW

Dongkuan Xu and Yingjie Tianhavesurveyed on different clustering algorithms upto 2015 [9].
Among the available clustering algorithms, two standard algorithms have been used in each of the proposed clustering model. They are SOM and K-means.
The topological structure of SOM networkis constructed by T. Kohonenin 2001 [10].The researchers are influenced by the dimensionality reduction power of SOM very much. JuhaVesanto and EsaAlhoniemi have observed clustering of SOM model from different angles [11].
K-means algorithm has two versions — one is with known cluster size [12] and the other is with unknown cluster size [11]. Both of these versions have been used in the proposed clustering models.
Number of research works can be found in the field of SOM based clustering. One of them is the fusion of SOM and PCA.When these two dimensionality reduction tools are combined, the performance of the resulting model is far improved than their individual applications.Some of the research works on PCA based SOM model can be found in [13], [14].
General references on PCA are given by Jolliffe I. T., 2002 [8] and Narayan C. Giri, 2003 [15].
To determine the number of components to be selected in PCA is also an innovative area in recent research works [16] [17].

## 3. EXISTING METHODS

A short description of all the prerequisites for the proposed models appears in this section.
Two standard clustering algorithms, namely, SOM and K-means have been used in the proposed clustering models. In both of the proposed models, the input data are pre-processed by normalization followed by PCA prior to application of SOM. All these techniques are described below.

### *A.Self-Organizing Map*

The inherent structure used by SOM network is a two (or more) dimensional rectangular (or hexagonal) lattice.This lattice is used as a placeholder that contains number of output nodes; this number and the number of classes in data are same. All these output nodes have their respective weight vectors with the same dimension as that of input nodes.These weights are only used for making connections between output nodes and input nodes. There may or may not be connections present in between the output nodes. If present, those connections are weightless.
The algorithm of SOM is basically a two pass algorithm. The first pass is the actual SOM algorithm and the second pass is the merging of SOM outputs.
The first pass is again consisting of two phases – training phase and testing phase. In the training phase, input nodes are presented to the SOM network one after anotherrepeatedly.Anyinput node is assigned to the output node which is nearest to it and weights are updated accordingly.The initial learning rate and neighborhood size are very important in this phase. These two parameters are decreased over time. The training is done in such a way that all parts of the network respond similarly to a certain input node. While testing, the trained data are used and the whole process of the training phase is repeated. The outcomes of this pass are numbers of SOM prototypes.
In the second pass, the SOM prototypes are further goes under clustering by using *K*-means.

### *B. The K-means Algorithm*

Generally, in K-means clustering algorithm the number of clusters is well known in advance. In the proposed models, this algorithm is used to merge SOM prototypes.In this algorithm, an error function is minimized iterativelyand updating of the cluster centersis done accordingly. The error function is defined by equation 1.

$$E = \sum_{i=1}^{C} \sum_{j=1}^{N} \| \mathrm{X_j} - K_i \|^2, \qquad \ldots\ldots (1)$$

Where$X_j$ is the *j*-th input vector, *K* is the number of clusters, and $K_i$ is the center of *i*-thcluster.

When the number of groups in a dataset is not known, the algorithmic steps for K-means clustering are repeated from an initial cluster size 2 to$\sqrt{n}$where *N* is the sample size. In each step, the error function defined in equation 1 is minimized. This version of K-means algorithm is used in each of the proposed component selection method.

### *C. Normalization*

When the attributes in a dataset appear with different scales, they need to be made of the same scale. For this, normalization is used. Due to normalization, all the attribute values existbetween 0 and 1.
In the present paper, all the attributes are normalized before application of PCA.

### *D. Principal Component Analysis*

PCA is a technique that helps to decrease the dimension of data by selecting those attributes, called principal components, whichare comparatively more informativethan others. In the proposed clustering models, PCA is used as a pre-processing technique in addition to normalization. The selected principal components are used in subsequent cluster analysis.
Among the existing principal component extraction methods, Kaiser Criterion and Cumulative Percentage of Total Variation are used to evaluate the proposed principal component selection methods.
According to Kaiser Criterion, the component with an eigenvalue less than one is considered to be of lesser amount of varianceto the total variance in the dataset and it is rejected; rests are kept as principal components.
In Cumulative Percentage of Total Variation, the principal components are selected from a dataset on the basis of a pre-assumed cut-off percentage.

To calculate PCA either the covariance matrix or the correlation matrix of the dataset is used. In the proposed model, the covariance matrix is considered.

## 4. PROPOSED METHODOLOGY

This section describes the reason behind the construction of the proposed clustering model. Detailed discussions about the proposed principal component selection methods also appear in this section.

It has already been discussed in previous section that PCA and SOM both are capable of retaining the essential information of a dataset in lesser number of selected components individually in a sophisticated way.Number of principal components to be extracted by PCA has a great impact on the performance of any model. If these components are chosen properly followed by the application of SOM algorithm, it is possible to have more robust output.

Eigenvalues and eigenvectors of data capture essential information and pattern of data.The eigenvector corresponding to the highest eigenvalue is the principal component of the dataset, in general. The whole set of eigenvalues or eigenvectors can be decomposed into number of groups based on similarity in patterns in a dataset. As the covariance matrix of a dataset also resembles the significant information, it is proposed to cluster the eigenvalues or the eigenvectors of a covariance matrixto consider the cluster with highest number of members in it.As this cluster holds the most similar patterns, that highest number can be used to select number of principal components while discarding the components in other.

It is also proposed that if the principal components are selected by using those above mentioned proposed principal component selection methods and those components are fed to theSOM network for clustering,itcan improve the performance of the standard SOM model.

To implement the proposed algorithms,at first covariance matrix is generated from normalized data. The eigenvalues and eigenvectors of this covariance matrix are also computed which are thenindividually clustered by using the *K*-means algorithm.Next, the cluster with highest number of members is identified in each case and that number is used to select principal components. Then, SOM algorithm is applied on these selected components.The proposed clustering methods are commonly named as EIGENPCASOM algorithm. Steps of the EIGENPCASOM algorithm are given as follows.

***EIGENPCASOM Algorithm*:**

Input: *N* number of eigenvalues or eigenvectors *EV* of the covariance matrix of normalized data, number of clusters *K*

Output: *K* number of clusters

1) Apply K-means algorithmon *N* eigenvalues or eigenvectors. Suppose we get *C* number of clusters.
2) Findthe cluster *C*1 out ofthese *C* clusters that has highest number of members. Suppose the highest number is *n*.
3) Select *n*numbers of principal components by using PCA.
4) Cluster*n*transformed components by using standard SOM algorithm.
5) Apply K-means algorithm on SOM prototypes to produce *K* number of clusters.

## 5. EXPERIMENTAL SETUP

The dataset that is to be used for testing any PCA based model must have larger set of attributes. The benchmark *wine* dataset of UCI machine learning repository[18]has this feature. So this dataset is chosenfor testing purposeand hence it helps in analyzing the performances of the proposed PCA based SOM models.

The *wine*data are the results of a chemical analysis of wines grown in the same region in Italy but extracted from 3 different species. This dataare composed of 13 attributes representing the quantities of 13 componentsthat form each of the 3 types of wines.The whole dataset is divided into 3 classes with 59, 71 and 48 instances respectively and 178 instances in total. The dataset contains no missing attributes. First attribute in this dataset is the class label 1-3. The attributes arenamely Alcohol, Malic acid, Ash, Alcalinity of ash, Magnesium, Total phenols, Flavanoids, Nonflavanoid phenols, Proanthocyanins, Color intensity, Hue, OD280/OD315 of diluted wines, Proline.

## 6. RESULTS AND ANALYSIS

The performance of the two variants of the proposed PCA based SOM model have been tested and analyzed for different lattice sizes from 3×3 to 10×10.The following two subsections discuss and analyze the results obtained by the application of proposed principal components selection methods in clustering of the *wine* data through PCA based SOM model.

### *A. When eigenvalues clustering method is chosen as component selection method*

In this case, 11 principal components have been extracted by the proposed component selection method and these components are clustered by using SOM algorithm.

Here, Table I shows the results which are considered to be the best output.

It is clear from Table I thatthe percentage of accuracy varies from 89% to 100% for class I in all the cases upto lattice sizes 3×3 to 10×10 and this 100% accuracy has been achieved with very small lattice size 3×3. The percentage of accuracy varies from 88% to 95% for class II. For class III, it varies from 95% to 100% and this 100% of accuracy has been achieved with small lattice size 4×4.

*Comparison with standard SOM*:

Table II shows the comparative study between the proposed principal component selection method using eigenvalues clustering and the standard SOM model in clustering of *wine* dataset.

It is clear from Table II that the proposed method is able to classify the test dataset with almost same accuracy as that of standard SOM modelupto lattice sizes 3×3 to 6×6.For class II and class III, these accuracies have been much betterthan the standard SOM model for some lattice sizes. So, it has been possible to achieve 100% accuracy for class I and more than 95% accuracy for classes II and III with very small lattice size 3×3 and with reduced components than the standard SOM model that saves lots of computation time and space, in turn.

It can also be seen from Table II that although the standard SOM is unable to classify the data from lattice sizes 7×7 to

10×10, theproposed method is able to classify with more than 94% accuracy for class I, 88% to 95% accuracy for class II and 95% to 100% accuracy for class III. Although, for these higher percentages of accuracies,*K*-means algorithm needs to consider more number of SOM prototypes while merging them than the previous set of lattice sizes for a small dataset, it may become effective for a dataset with larger set of attributeswhere standard SOM fails to classify it properly.

*Comparison with the model when Kaiser Criterion used as component selection method*:
Table V shows the comparative study between the proposed principal component selection method using eigenvalues clustering for classification of *wine*datasetand the model where Kaiser Criterion is used as component selection method.
It is seen from Table V that the proposed approach is classifying the test dataset with improved accuracy than the existing one in all the classes with all lattice sizes. This is obvious because, the proposed method is classifying the dataset with 11 components whereas the existing one is classifying with only 8 components. As a result,the accuracyhas also been improved, because comparatively very few components are lost in the proposed case. Although, it seems that the proposed approach is taking more space and computation time than the existing one but this may not be always true for all datasets.How many componentswill be extracted from a dataset depends on the pattern of data, size of the datasetand original number of components in a dataset for which more or less number of clusters of eigenvalues can be obtained.

*Comparison with the model when Cumulative Percentage of Total Variation used as component selection method*:
Table V shows the comparative study between the proposed principal component selection method using eigenvalues clustering for classification of *wine*dataset and the model where Cumulative Percentage of Total Variation is used as component selection method.
From Table V, it can be seen that for almost all lattice sizes, theproposedmodelis classifying all the classes with improved accuracy than this existing method. In this case, the existing method is using 5 components and the proposed one is using 11 components. Obviously, the result is also muchbetter with the proposed model. Same explanation,as given for the other existing component selection method, is also applicablein the present case.
A graphical representation of the performance of the proposed principal component selection method based clustering model has been given in Figure 1where the top line represents the overall accuracy for different lattice sizesof the proposed clustering model and the bottom two lines represent that of the other two existing component selection methods based clustering models. From Figure 1 it is clear that the proposed method has the superiority over other two existing component selection methods.

### *B. When eigenvectors clustering method is chosen as component selection method*

In this case, 7 principal components have been extracted by the proposed component selection method and these components are presented to the SOM network.
Table III shows the results which are considered to be the best output.
It is clear from Table III that the percentage of accuracy varies from 79% to 93% for class I upto lattice sizes 3×3 to 10×10.For class II, the same varies from 88% to 95% and for class III, it varies from 91% to 97%.
*Comparison with standard SOM*:
Table IV shows the comparative study between the proposed principal component selection method using eigenvectors clustering and the standard SOM model in clustering of *wine* dataset.
It is seen from Table IV that with this proposed approach,in almost all the cases it is possible to have 88% to 92% accuracy for class II and 91% to 97%accuracy for class IIIupto lattice sizes 3×3 to 6×6. For these two classes, the accuracy is improving than the standard SOM for some of the lattice sizes. For class I, 84% to 89% accuracy is achieved through this set of lattice sizes. It is clear from Table IV that around 84% accuracy for class I and more than 90% accuracy for other two classes can be achieved with very small lattice sizes 3×3 or 4×4 with reduced components than the standard SOM which saves lots of computation time and space.
Table IV also shows that from lattice sizes 7×7 to 10×10, the standard SOM fails whereas theproposed approach is able to classify with 79% to 93% accuracy for class I, 90% to 95% accuracy for class II and 93% to 97% accuracy for class III. Although,due to thesehigher percentages of accuracy,*K*-means algorithm needs to consider more number of SOM prototypeswhile merging them than the previous set of lattice sizes, its significance can be well explained in the same way as it has been explained in the last subsection.
*Comparison with the model when Kaiser Criterion used as component selection method*:
Table VI shows the comparative study between the proposed principal component selection method using eigenvectors clustering for classification of *wine*datasetand the model where Kaiser Criterion is used as component selection method.
It can be seen from Table VIthat with theproposed approach, all the classes are classifying with improved accuracy than the existing methodin almost all the cases.Although in some cases, one class is classifying with decreased accuracy, the overall percentage of accuracy is improving because this proposedmethod is using only 7 components for classification of the dataset whereas the existing method is using 8 components. So, it can be said that proposedmethod is computationally efficient in terms space and time than the existing method.Again, this must be kept in mind that number of extraction of components depends on several factors.
*Comparison with the model when Cumulative Percentage of Total Variation used as component selection method*:
Table VI shows the comparative study between the proposed principal component selection method using eigenvectors clustering for classification of *wine*dataset and the model where Cumulative Percentage of Total Variation is used as component selection method.
It can be easily seen from Table VI that theproposedmethodis capable of classifying almost all the classes with improved accuracy than this existing method for almost all lattice sizes. Again, as this existing method of component selection is using only 5 components and the

proposed one is using 7, the result obtained with the proposed method is also better. But theproposed approach is showing improved accuracy for all the classes with higher order lattice sizes like 8×8 to 10×10.This leads to the requirement of more computation time.

Figure 2 represents a graphical view of the performance of the proposed principal component selection method based clustering model.Here, the top line represents the overall accuracy for different lattice sizes of the proposed clustering model and the bottom two lines represent that of the other two existing component selection methods based clustering models. It is clear from Figure 2 that the proposed method has the superiority over other two existing component selection methods.

*Comparison between the two proposed component selection methods*:

Table VII shows a comparative study of the two proposed principal component selection methods in PCA for classification of *wine*dataset using SOM model.

It is clear from Table VII that the accuracy level is same for all the classes of both of the proposed methods with almost all the cases. Although, sometimes it is deviating from one another in some of the cases, it is negligible in terms of number of components they are using; sometimes it also remains same. To maintain this accuracy level, the eigenvalues clustering method is using 11 components whereas eigenvectors clustering method is using only 7components. So, the second method is consuming less space and computation time than the first. And hence, it can be said that the second method is computationally more efficient than the first for clustering of the *wine* dataset.

Figure 3 shows a graphical representation of the performance of the clustering models based on two proposed principal component selection methods. Although from Figure 3 it seems that the eigenvalues clustering based component selection method gains the superiority over the other proposed component selection method, if individual accuracy level is compared for each class and the number of components the proposed methods are using is also taken into account, it can be said that the eigenvectors based clustering method is superior.

## 7. CONCLUSION

In this paper, two different techniques for selection of principal components in PCA have been proposed.Both of the methods are able toproduce satisfactory results with improved accuracy in clustering of the SOM model.So,it can be concluded that the proposed models are better than the existing models. Proposed models also have the abilityto cluster data at small lattice size and have the power of reducing the dimension of data.So, it can also be concluded that the proposed methods are computationally efficient.In spite of these satisfactory results, there is also scope for further improvement in the proposed models. Instead of *K*-means clustering algorithm,other clustering techniques can be used to cluster eigenvalues and eigenvectors of the covariance matrix and also to produce final clusters from the SOM model. Other pre-processing techniques can also become effectivein clustering SOM model with better output.

TABLEI. Classification of Wine Data [18] using PCA Based SOM Showing Best Outputs with Different Lattice Sizes by Clustering Eigenvalues and Choosing Highest Number of Eigenvalues Classified as the Principal Component Selection Method

| ***Lattice Size*** | ***Classification using Eigenvalues Clustering as the Principal Component Selection Method (No. of components = 11)*** | | | | | | ***Accuracy (%)*** | | |
|---|---|---|---|---|---|---|---|---|---|
| | ***Class I*** | | ***Class II*** | | ***Class III*** | | ***Class I*** | ***Class II*** | ***Class III*** |
| | ***Correct*** | ***Wrong*** | ***Correct*** | ***Wrong*** | ***Correct*** | ***Wrong*** | | | |
| **3×3** | 59 | 2 | 68 | 1 | 47 | 1 | **100** | **95.77** | **97.92** |
| **4×4** | 53 | 0 | 63 | 6 | 48 | 8 | **89.83** | **88.73** | **100** |
| **5×5** | 58 | 2 | 68 | 4 | 44 | 1 | **98.31** | **95.77** | **91.66** |
| **6×6** | 59 | 7 | 64 | 1 | 47 | 0 | **100** | **90.14** | **97.92** |
| **7×7** | 57 | 1 | 68 | 4 | 46 | 2 | **96.61** | **95.77** | **95.83** |
| **8×8** | 57 | 6 | 63 | 3 | 47 | 2 | **96.61** | **88.73** | **97.92** |
| **9×9** | 56 | 4 | 65 | 4 | 47 | 2 | **94.92** | **91.55** | **97.92** |
| **10×10** | 58 | 4 | 65 | 1 | 48 | 2 | **98.31** | **91.55** | **100** |

TABLE II. Comparative study of the Proposed Principal Component Selection Method for Classification of Wine Data [18] using Eigenvalues Clustering with the Standard SOM Model.

| ***Lattice Size*** | *Comparative study of the Proposed Principal Component Selection Method for Classification of Wine Dataset using Eigenvalues Clustering with the Standard SOM Model* ***[The values are percentage of Correct Classification, No. of Components (for Standard SOM)=13, No. of Components (for Eigenvalues Clustering)=11]*** | | | | | |
|---|---|---|---|---|---|---|
| | ***Class I*** | | ***Class II*** | | ***Class III*** | |
| | ***Standard SOM*** | ***Eigenvalues Clustering*** | ***Standard SOM*** | ***Eigenvalues Clustering*** | ***Standard SOM*** | ***Eigenvalues Clustering*** |
| **3×3** | **100** | **100** | **95.77** | **95.77** | **100** | **97.92** |
| **4×4** | **100** | **89.83** | **91.55** | **88.73** | **91.67** | **100** |

| | | | | | | |
|---|---|---|---|---|---|---|
| **5×5** | **100** | **98.31** | **88.73** | **95.77** | **100** | **91.66** |
| **6×6** | **100** | **100** | **85.92** | **90.14** | **100** | **97.92** |
| **7×7** | × | **96.61** | × | **95.77** | × | **95.83** |
| **8×8** | × | **96.61** | × | **88.73** | × | **97.92** |
| **9×9** | × | **94.92** | × | **91.55** | × | **97.92** |
| **10×10** | × | **98.31** | × | **91.55** | × | **100** |

TABLE III. Classification of Wine Data [18] using PCA Based SOM Showing Best Outputs with Different Lattice Sizes by Clustering Eigenvectors and Choosing Highest Number of Eigenvectors Classified as the PrincipalComponent Selection Method.

| ***Lattice Size*** | ***Classification using Eigenvectors Clustering as the Principal Component Selection Method (No. of components = 7)*** | | | | | | ***Accuracy (%)*** | | |
|---|---|---|---|---|---|---|---|---|---|
| | ***Class I*** | | ***Class II*** | | ***Class III*** | | ***Class I*** | ***Class II*** | ***Class III*** |
| | ***Correct*** | ***Wrong*** | ***Correct*** | ***Wrong*** | ***Correct*** | ***Wrong*** | | | |
| **3×3** | 50 | 7 | 65 | 4 | 47 | 5 | **84.75** | **91.55** | **97.92** |
| **4×4** | 50 | 6 | 66 | 1 | 46 | 9 | **84.75** | **92.96** | **95.83** |
| **5×5** | 51 | 7 | 65 | 6 | 44 | 5 | **86.44** | **91.55** | **91.66** |
| **6×6** | 53 | 11 | 63 | 1 | 45 | 7 | **89.83** | **88.73** | **93.75** |
| **7×7** | 47 | 3 | 68 | 8 | 47 | 5 | **79.66** | **95.77** | **97.92** |
| **8×8** | 54 | 9 | 64 | 1 | 46 | 4 | **91.53** | **90.14** | **95.83** |
| **9×9** | 55 | 9 | 64 | 1 | 45 | 4 | **93.22** | **90.14** | **93.75** |
| **10×10** | 50 | 3 | 68 | 7 | 46 | 4 | **84.75** | **95.77** | **95.83** |

TABLE IV. Comparative study of the Proposed Principal Component Selection Method for Classification of Wine Data [18] using Eigenvectors Clustering with the Standard SOM Model.

| ***Lattice Size*** | *Comparative study of the Proposed Principal Component Selection Method for Classification of Wine Dataset using Eigenvectors Clustering with the Standard SOM Model* ***[The values are percentage of Correct Classification, No. of Components (for Standard SOM)=13, No. of Components (for Eigenvalues Clustering)=7]*** | | | | | |
|---|---|---|---|---|---|---|
| | ***Class I*** | | ***Class II*** | | ***Class III*** | |
| | ***Standard SOM*** | ***Eigenvectors Clustering*** | ***Standard SOM*** | ***Eigenvectors Clustering*** | ***Standard SOM*** | ***Eigenvectors Clustering*** |
| **3×3** | **100** | **84.75** | **95.77** | **91.55** | **100** | **97.92** |
| **4×4** | **100** | **84.75** | **91.55** | **92.96** | **91.67** | **95.83** |
| **5×5** | **100** | **86.44** | **88.73** | **91.55** | **100** | **91.66** |
| **6×6** | **100** | **89.83** | **85.92** | **88.73** | **100** | **93.75** |
| **7×7** | × | **79.66** | × | **95.77** | × | **97.92** |
| **8×8** | × | **91.53** | × | **90.14** | × | **95.83** |
| **9×9** | × | **93.22** | × | **90.14** | × | **93.75** |
| **10×10** | × | **84.75** | × | **95.77** | × | **95.83** |

TABLE V. Comparative study of the Two Existing and the Proposed Eigenvalues Clustering Principal Component Selection Methods for Classification of Wine Data [18] using PCA Based SOM.

| ***Lattice Size*** | ***Comparative Study of the two existing and the proposed Eigenvalues Clustering Principal Component Selection Methods for classification of Wine Dataset using PCA based SOM [The values are percentage of Correct Classification, No. of Components (for Kaiser Criterion)=8, No. of Components (for C.P. of Total Variation)=5, No. of Components (for Eigenvalues Clustering)=11]*** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | ***Class I*** | | | ***Class II*** | | | ***Class III*** | | |
| | ***Kaiser Criterion*** | ***C.P. of Total Variation*** | ***Eval. Clust.*** | ***Kaiser Criterion*** | ***C.P. of Total Variation*** | ***Eval. Clust.*** | ***Kaiser Criterion*** | ***C.P. of Total Variation*** | ***Eval. Clust.*** |
| **3×3** | **86.44** | **93.22** | **100** | **88.73** | **81.69** | **95.77** | **85.42** | **87.50** | **97.92** |
| **4×4** | **86.44** | **91.52** | **89.83** | **88.73** | **80.28** | **88.73** | **93.75** | **93.75** | **100** |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **5×5** | **86.44** | **94.92** | **98.31** | **87.32** | **84.51** | **95.77** | **89.58** | **87.50** | **91.66** |
| **6×6** | **86.44** | **83.05** | **100** | **84.51** | **84.51** | **90.14** | **97.92** | **87.50** | **97.92** |
| **7×7** | **96.61** | **88.14** | **96.61** | **84.51** | **81.69** | **95.77** | **89.58** | **87.50** | **95.83** |
| **8×8** | **86.44** | **81.36** | **96.61** | **84.51** | **81.69** | **88.73** | **89.58** | **91.66** | **97.92** |
| **9×9** | **86.44** | **84.75** | **94.92** | **94.36** | **81.69** | **91.55** | **79.16** | **91.66** | **97.92** |
| **10×10** | **81.36** | **88.14** | **98.31** | **91.55** | **84.51** | **91.55** | **89.58** | **87.50** | **100** |

TABLE VI. Comparative study of the Two Existing and the Proposed Eigenvectors Clustering Principal Component Selection Methods for Classification of Wine Data [18]using PCA Based SOM

| ***Lattice Size*** | ***Comparative Study of the two existing and theproposed Eigenvectors Clustering Principal Component Selection Methods for classification of Wine Dataset using PCA based SOM [The values are percentage of Correct Classification, No. of Components (for Kaiser Criterion)=8, No. of Components (for C.P. of Total Variation)=5, No. of Components (for Eigenvectors Clustering)=7]*** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | ***Class I*** | | | ***Class II*** | | | ***Class III*** | | |
| | ***Kaiser Criterion*** | ***C.P. of Total Variation*** | ***Evect. Clust.*** | ***Kaiser Criterion*** | ***C.P. of Total Variation*** | ***Evect. Clust.*** | ***Kaiser Criterion*** | ***C.P. of Total Variation*** | ***Evect. Clust.*** |
| **3×3** | **86.44** | **93.22** | **84.75** | **88.73** | **81.69** | **91.55** | **85.42** | **87.50** | **97.92** |
| **4×4** | **86.44** | **91.52** | **84.75** | **88.73** | **80.28** | **92.96** | **93.75** | **93.75** | **95.83** |
| **5×5** | **86.44** | **94.92** | **86.44** | **87.32** | **84.51** | **91.55** | **89.58** | **87.50** | **91.66** |
| **6×6** | **86.44** | **83.05** | **89.83** | **84.51** | **84.51** | **88.73** | **97.92** | **87.50** | **93.75** |
| **7×7** | **96.61** | **88.14** | **79.66** | **84.51** | **81.69** | **95.77** | **89.58** | **87.50** | **97.92** |
| **8×8** | **86.44** | **81.36** | **91.53** | **84.51** | **81.69** | **90.14** | **89.58** | **91.66** | **95.83** |
| **9×9** | **86.44** | **84.75** | **93.22** | **94.36** | **81.69** | **90.14** | **79.16** | **91.66** | **93.75** |
| **10×10** | **81.36** | **88.14** | **84.75** | **91.55** | **84.51** | **95.77** | **89.58** | **87.50** | **95.83** |

[Note: C.P. – Cumulative Percentage, Eval. Clust. – Eigenvalues Clustering, Evect. Clust. – Eigenvectors Clustering]

TABLE VII**.** Comparative study of the Two Proposed Principal Component Selection Methods for Classification of Wine Data[18] using PCA Based SOM

| ***Lattice Size*** | ***Comparative Study of the two proposed Principal Component Selection Method for classification of Wine Dataset using PCA based SOM [The values are percentage of Correct Classification, No. of Components (for Eigenvalues Clustering)=11, No. of Components (for Eigenvectors Clustering)=7]*** | | | | | |
|---|---|---|---|---|---|---|
| | ***Class I*** | | ***Class II*** | | ***Class III*** | |
| | ***Eigenvalues Clustering*** | ***Eigenvectors Clustering*** | ***Eigenvalues Clustering*** | ***Eigenvectors Clustering*** | ***Eigenvalues Clustering*** | ***Eigenvectors Clustering*** |
| **3×3** | **100** | **84.75** | **95.77** | **91.55** | **97.92** | **97.92** |
| **4×4** | **89.83** | **84.75** | **88.73** | **92.96** | **100** | **95.83** |
| **5×5** | **98.31** | **86.44** | **95.77** | **91.55** | **91.66** | **91.66** |
| **6×6** | **100** | **89.83** | **90.14** | **88.73** | **97.92** | **93.75** |
| **7×7** | **96.61** | **79.66** | **95.77** | **95.77** | **95.83** | **97.92** |
| **8×8** | **96.61** | **91.53** | **88.73** | **90.14** | **97.92** | **95.83** |
| **9×9** | **94.92** | **93.22** | **91.55** | **90.14** | **97.92** | **93.75** |
| **10×10** | **98.31** | **84.75** | **91.55** | **95.77** | **100** | **95.83** |

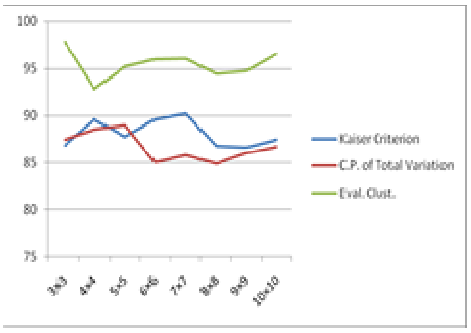


Fig. 1: Comparison of the performances of the proposed Eigenvalues Clustering model against the Two Existing Component Selection Methods

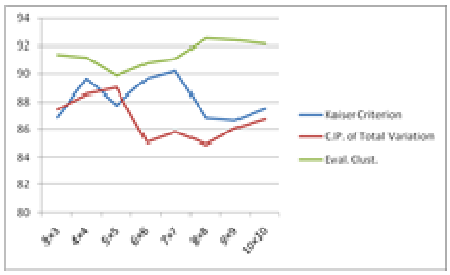


Fig. 2: Comparison of the performances of the proposed Eigenvectors Clustering model against the Two Existing Component Selection Methods

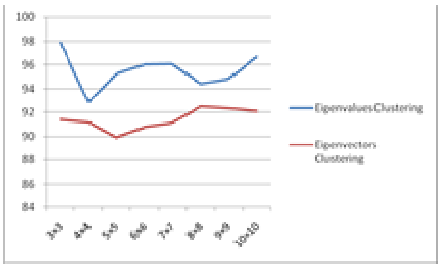


Fig. 3: Comparison of the performances of the PCA based model for two proposed principal component selection methods with different lattice sizes for clustering *wine*dataset

## 8. ACKNOWLEDGEMENT

Authors are grateful to the Department of Computer Science, The University of Burdwan.The authors are also thankful to UCI Machine Learning Centre for their online standard datasets.

## REFERENCES

[1] Yujing Zeng and J. Starzyk, "Statistical approach to clustering in pattern recognition," in *Proceedings of the 33rd Southeastern Symposium on System Theory (Cat. No.01EX460)*, Athens, OH, 2001, pp. 177-181. doi: 10.1109/SSST.2001.918513

[2] Neha D. and B. M. Vidyavathi, "A Survey on Applications of Data Mining using Clustering Techniques," International Journal of Computer Applications (0975 – 8887), vol. 126, no. 2,, pp. 7 – 12, September 2015.

[3] Tomasz Tarczynski, "Document Clustering – Concepts, Metrics and Algorithms," INTL Journal of Electronics and Telecommunications, vol. 57, no. 3, pp. 271–277, 2011.

[4] Md. Khalid Imam Rahmani, Naina Pal and Kamiya Arora, "Clustering of Image Data Using K-Means and Fuzzy K-Means" International Journal of Advanced Computer Science and Applications(ijacsa), vol. 5, no. 7, pp. 160 – 163, 2014, http://dx.doi.org/10.14569/IJACSA.2014.050724

[5] M. G. Malhat, H. M. Mousa and A. B. El-Sisi, "Clustering of chemical data sets for drug discovery," in Proceedings of *2014 9th International Conference on Informatics and Systems*, Cairo, pp. DEKM-11-DEKM-18, 2014, doi: 10.1109/INFOS.2014.7036702

[6] Rui Xu and Donald Wunsch II, "Survey of Clustering Algorithms," IEEE Transactions on Neural Network, vol. 16, no. 3, pp. 645-678, May 2005.

[7] KadimTasdemir, Pavel Milenov, and Brooke Tapsall, "Topology-Based Hierarchical Clustering of Self-Organizing Map," IEEE Transactions On Neural Networks, vol. 22, no. 3, pp. 474-485, March 2011.

[8] JolliffeI. T., "Principal Component Analysis," 2nd edition, Springer, 2002.

[9] Dongkuan Xu and Yingjie Tian, "A Comprehensive Survey of Clustering Algorithms," Annalysis of Data Science, vol. 2, no. 2, pp. 165 – 193, June 2015.

[10] Kohonen T., "Self-Organizing Maps," 3rd edn., New York: Springer-Verlag, 2001.

[11] JuhaVesanto and EsaAlhoniemi, "Clustering Of The Self-Organizing Map," IEEE Transactions On Neural Networks, vol. 11, no. 3, pp. 586-600, May 2000.

[12] Jian Yu, "General C-Means Clustering Model," IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 27, no. 8, pp. 1197-1211, Aug 2005.

[13] Ye Wenyu, Li Gang, Lin Ling and Yu Qilian, "ECG analysis based on PCA and SOM," in Proceedings of the 2003 International Conference on Neural Networks and Processing, Nanjing, pp. 37 – 40, 2003, vol. 1. doi: 10.1109/ICNNSP.2003.1279207

[14] SuwardiAnnas, Takenori Kanai and Shuhei Koyama, "Principal Component Analysis and Self-Organizing Map for Visualizing and Classifying Fire Risks in Forest Regions," Agricultural Information Research, vol. 16, no. 2, pp. 44 – 51, 2007.

[15] Narayan C. Giri, "Multivariate Statistical Analysis," CRC Press, 2nd edn., 2003.

[16] Gibbs Y. Kanyongo, "Determining The Correct Number of Components to Extract from A Principal Component Analysis: A Monte Carlo Study of the Accuracy of the Scree Plot," Journal of Modern Applied Statistical Methods, vol. 4, no. 1, pp. 120 – 133, May 2005.

[17] Y. Fu, H. Tao and H. Yang, "Simultaneous estimation of the number of principal components and kernel parameter in KPCA," *2017 6th International Symposium on Advanced Control of Industrial Processes (AdCONIP)*, Taipei, 2017, pp. 149-154. doi: 10.1109/ADCONIP.2017.7983771

[18]R. A. Fisher, "UCI machine learning repository," 1936. [Online]. Available: http://archive.ics.uci.edu/ml