



International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

Analysis of Brain Tumor Using Neural Network Based PCA and Clustering

¹ Prof. Shanthi Mahesh, ² Dr. Neha Mangla, ³ Suhash A. Bhyratae

¹ Assistant Professor, ² Associate Professor, ³ Lecturer, Dept. Information Science & Engg.,
Atria Institute of Technology, Bangalore, India

Abstract: This paper discusses the analysis of Brain tumor in detail by using principal component (PCA) and clustering approach. Principal component analysis (PCA) is a widely used statistical technique for unsupervised dimension reduction. K-means clustering is a commonly used data clustering for unsupervised learning tasks. Here we prove that principal components are the continuous solutions to the discrete cluster membership. These results indicate that unsupervised dimension reduction is closely related to unsupervised learning. On dimension reduction, the result provides new insights to the observed effectiveness of PCA-based data reductions; SOM clusters, SOM topology, SOM neighbor connection etc. are obtained by using MATLAB. The results are interesting.

Keywords: Principal Component Analysis (PCA), Self Organization Map(SOM), Neural Network, SOM topology, brain tumor.

I. Introduction

Data analysis methods are essential for analyzing the ever-growing massive quantity of high dimensional data. On one end, cluster analysis (Duda et al., 2000; Hastie et al., 2001; Jain & Dubes, 1988) attempts to pass through data quickly to gain first order knowledge by partitioning data points into disjoint groups such that data points belonging to same cluster are similar while data points belonging to different clusters are dissimilar. One of the most popular and efficient clustering methods is the K-means method (Hartigan & Wang, 1979; Lloyd, 1957; MacQueen, 1967) which uses prototypes (centroids) to represent clusters by optimizing the squared error function. (A detail account of K-means and related ISODATA methods are given in (Jain & Dubes, 1988)) On the other end, high dimensional data are often transformed into lower dimensional data via the principal component analysis (PCA) (Jolliffe, 2002) (or singular value decomposition) where coherent patterns can be detected more clearly. Such unsupervised dimension reduction is used in very broad areas such as meteorology, image processing, genomic analysis, and information retrieval. It is also common that PCA is used to project data to a lower dimensional subspace and K-means is then applied in the subspace (Zha et al., 2002). In other cases, data are embedded in a low-dimensional space such as the eigenspace of the graph Laplacian, and K-means is then applied (Ng et al., 2001). The main basis of PCA-based dimension reduction is that PCA picks up the dimensions with the largest variances. Mathematically, this is equivalent to finding the best low rank approximation (in L2 norm) of the data via the singular value decomposition (SVD) (Eckart & Young, 1936). However, this noise reduction property alone is inadequate to explain the effectiveness of PCA.

The main objectives of PCA are: (i) To discover or to reduce the dimensionality of the data set and (ii). to identify new meaningful underlying variables.

PCA is often incorporated into genome-wide expression studies. Several measurement techniques used in the life sciences gather data for many more variables per sample than the typical number of samples. For instance, DNA microarrays and mass spectrometers can measure levels of thousands of mRNAs or proteins in hundreds of samples.

Mathematical background on Principal Component Analysis is constructed on a mathematical technique used in PCA is called eigen analysis: It can be solved for the eigen values and eigenvectors of a square symmetric matrix with sums of squares and cross products. The eigen vector associated with the largest eigen value has the same direction as the first principal component. The eigen vector associated with the second largest eigenvalue determines the direction of the second principal component. The sum of eigenvalues equals the trace of the square matrix and the maximum number of eigenvectors equals the number of rows (or columns) of the matrix. The paper is organized as follows: section:2 deals with the literature survey, section:3 deals with the data with the data set description, section:4 deals with methodology, section:5 deals with experiments and results. Finally, the conclusions are drawn.

II. Related work

Adaptive hypermedia learning system aim to improve the usability of hypermedia by personalizing domain knowledge to the student need. This study [2] investigates student modeling via machine-learning techniques. Two techniques are applied and compared to provide meaning analysis and class labels of the student clusters. It was found that implementing the SOM as a preprocessor to PCA improves the quality of cluster analysis. The author[3] presents a new approach for Data set compiler from spreading hot spot, responsibility for fire risk in many regions of Indonesian forests, are complex, primarily induced by the large size of the observed regions and high variation of hot spot distribution. This study presents the use of PCA and SOM to reduce the dimensionality of the input variables and subsequently to visualize the dataset into 2D space. The result indicates that the first two principal component of the PCA provide a large percentage of cumulative variance to explain the data patterns. A new technique to analyses microarray data which uses a combination of PCA consensus ensemble k-clustering find robust clusters and gene markers in the data, author[4] identify the clusters and their pathways with distinct subtypes of breast cancer. It is conform that the cancer phenotype develops early and find from our analysis that each subtype progresses from ADH to DCIS to IDC along its own specific pathway, as if each was a distinct disease. The author[5] presents Robust cancer molecular pattern identification from microarray data not only plays an essential role in modern clinic oncology, but also presents a challenge for statistical learning. Although PCA is a widely used feature selection algorithm in microarray analysis, its holistic mechanism prevents it from capturing the latent local data structure in the following cancer molecular pattern identification. As a more robust high-performance classifier, NPCA-SVM can be used to replace the general SVM and k-NN classifier in cancer biomarker discovery to capture more meaningful oncogenes. The author [6] present paper attempts to generate visual clustering and data extraction of cell formation problem using both PCA and SOM from input of sequence based machine-part incidence matrix. Most importantly, in the visual clustering of original data, the use of U-matrix alone cannot be efficient to get the cluster structure but with color extraction, hit map, labeling via SOM node it became powerful clustering visualization methodology. The author[7] proposed Independent Principal Component (IPCA) that combines the advantages of both PCA and ICA. It uses ICA as denoising process of the loading vectors produced by PCA to better highlight the important biological samples on graphical representations. IPCA and SIPCA are both implemented in the R package mixomics dedicated to the analysis and exploration of high dimensional biological data sets, and on mixomics web-interface. A new method for clustering analysis of QRS complexes was proposed by the author[8]. The method integrates PCA with SOM neural network. The QRS feature is extracted based on PCA and the unsupervised SOM is employed to cluster the data. It demonstrated that QRS complexes feature can be presented by four largest principal components and the PCA results can be used to cluster analysis efficiently. The relationship between cluster results and clinical categories are also investigated.

III. Data Set Description

BrainTumor Data[9]

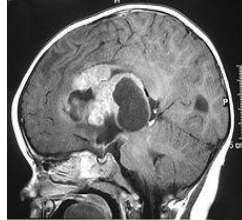
The dataset comes from a study of 5 human brain tumor types and includes 90 samples. Each sample has 5920 genes:

- A. **Medulloblastoma:** Medulloblastoma is a highly malignant primary brain tumor that originates in the cerebellum or posterior fossa. Previously, medulloblastomas were thought to represent a subset of primitive neuroectodermal tumor (PNET) of the posterior fossa. However, gene expression profiling has shown that medulloblastomas have distinct molecular profile and are distinct from other PNET tumors.
- B. **Malignant glioma:** A glioma is a type of malignant brain tumor. A malignant tumor is a mass of abnormal cells that is cancerous. Tumors can develop in any part of the brain or its nerves and covering tissues. The two major types of brain tumor are primary and secondary. Primary brain tumors start in the brain. Secondary brain tumors start in another part of the body, then spread to the brain. A glioma is a primary brain tumor, accounting for 45% of cancers that begin in brain cells. The three main types of glioma include: astrocytoma, ependymoma, and oligodendroglioma. Each of these types can be assigned a grade, either low grade or high grade, with high grade being more malignant and aggressive.
 - i. **Astrocytomas** are named for the cells where they originate, the astrocytes. These tumors can either show clear borders between normal brain tissue and the tumor (called focal) or no clear border (called diffuse). Focal astrocytomas are most common in children and are not often found in adults.
 - ii. **Ependymomas** begin in cells called ependymal cells that are found lining certain areas of the brain and spinal cord. These cells help repair damaged nerve tissue. They usually occur in children and young adults.
 - iii. **Oligodendrogliomas** form in oligodendrocyte cells, which produce a fatty substance called myelin that protects the nerve. More common in adults, these tumors may move to other parts of the brain or spinal cord.

C. AT/RT (atypical teratoid/rhabdoidtumours):

Atypical teratoidrhabdoid tumor (AT/RT) is a rare tumor usually diagnosed in childhood. Although usually a brain tumor, AT/RT can occur anywhere in the central nervous system (CNS) including the spinal cord. About 60% will be in the posterior rianal fossa (particularly the cerebellum). One review estimated 52% posterior fossa, 39% sPNET (supratentorialprimitive neuroectodermal tumors), 5% pineal, 2% spinal, and 2% multi-focal.[1]

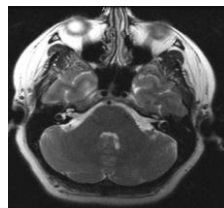
Figure 1: MRI of AT/RT



D. Normal cerebellum:

The cerebellum is part of the brain. It lies under the cerebrum, towards the back, behind the brainstem and above the brainstem. The cerebellum is largely involved in "coordination". Persons whose cerebellum doesn't work well are generally clumsy and unsteady. They may look like they are drunk even when they are not.

Figure 2: The cerebellum with surrounding skull and spinal fluid occupies the bottom 1/3 of this axis MRI image



The main clinical features of cerebellar disorders include incoordination, imbalance, and troubles with stabilizing eye movements. There are two distinguishable cerebellar syndromes -- midline and hemispheric. Midline cerebellar syndromes are characterized by imbalance. Persons are unsteady, they are unable to stand in Romberg with eyes open or closed, and are unable to well perform tandem gait. Severe midline disturbance causes "trunkal ataxia" a syndrome where a person is unable to sit on their bed without steadying themselves. Some persons have "titubation" or a bobbing motion of the head or trunk. Midline cerebellar disturbances also often affect eye movements. There may be nystagmus, ocular dysmetria and poor pursuit.

Hemispheric cerebellar syndromes are characterized by incoordination of the limbs. There may be decomposition of movement, dysmetria, and rebound. Dysdiadochokinesis is the irregular performance of rapid alternating movements. Intention tremors may be present on an attempt to touch an object. A kinetic tremor may be present in motion. The finger-to-nose and heel-to-knee tests are classic tests of hemispheric cerebellar dysfunction. While reflexes may be depressed initially with hemispheric cerebellar syndromes, this cannot be counted on. Speech may be dysarthric, scanning, or have irregular emphasis on syllables.

E. PNET (primitive neuroectodermaltumours)

PNET (pronounced pee-net) stands for a group of tumours known as Primitive Neuro Ectodermal Tumours. PNETS develop from cells that are left over from the earliest stages of a body's development in the womb. Normally these cells are harmless. But occasionally they turn into a cancer. These cancers are more common in children than adults.

Doctors use the term PNET to classify the tumour. They are divided into two main groups:

- a. PNETs of the brain and central nervous system
- b. Peripheral PNET (outside the brain and nervous system)

PNETs of the brain or spinal cord

Primitive neuroectodermal tumours that occur in the brain and spinal cord (the central nervous system or CNS) include

- Medulloblastoma (develops in the back part of the brain – the hindbrain)
- Pineoblastoma (develops in the pineal region of the brain)
- Non pineal supratentorial

IV. Methodology

Functional features of PCA

- `Princomp(X)` performs principal components analysis (PCA) on the n-by-p data matrix X, and returns the principal component coefficients, also known as loadings. Rows of X correspond to observations, columns to variables. For example `coeff=princomp(X)`, `coeff` is a p-by-p matrix, each column containing coefficients for one principal component. The columns are in order of decreasing component variance.
- `Princomp`, centers X by subtracting off column means, but does not rescale the columns of X.
- To perform principal components analysis with standardized variables, that is, bases on correlations, use `princomp(zscore(X))`.
- To perform principal components analysis directly on a covariance or correlation matrix, we use `pcacov`.
- `[coeff,score]=princomp(X)` return score, the principal component scores; that is, the representation of X in the principal component space. Rows of score correspond to observations, columns to components.
- `[coeff,score,latent]=princomp(X)` returns latent, a vector containing the eigenvalues of the covariance matrix of X.
- `[coeff,score,latent,tsquare]=princomp(X)` returns `tsquare`, which contains Hotelling's T2 statistic for each data point.
- The scores are the data formed by transforming the original data into the space of the principal components. The values of the vector latent are the variance of the columns of score. Hotelling's T2 is a measure of the multivariate distance of each observation from the center of the data set.
- When $n \leq p$, `score(:,n:p)` and `latent(n:p)` are necessarily zero, and the columns of `coeff(:,n:p)` define directions that are orthogonal to X.
- `[...]=princomp(X,'econ')` returns only the elements of latent that are not necessary zero, and corresponding columns of `coeff` and `score`, that is, when $n \leq p$, only the first n-1. This can be significantly faster when p is much larger than n.
- o Principal-component analysis (PCA) is a useful technique which can use be reduce the dimensionality of large data sets, such as those from microarray analysis, we can also use PCA to find signals in noisy data. The function `princomp` is statistics Toolbox software to calculate the principal components of a data set.
- o For example:
- o `[pc,zscores,pcvars] = princomp(braintumor)`
- o The cumulative sum of the variances is found using the function `cumsum`
`cumsum(pcvars./sum(pcvars) * 100)`

Which showed 100% of the variance is accounted for the first two principal components.

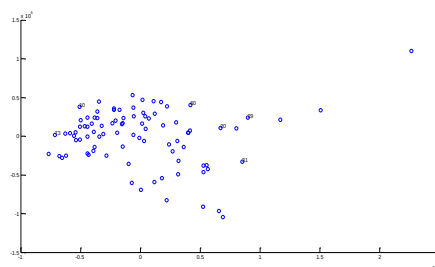
- ☐ The principal components can be visualized using the scatter function
- ☐ Scatter plot of the scores of the first two principal components shows that there are two distinct regions. This is not unexpected, since the filtering process removed many of the genes with low variance or low information.
- ☐ Another way to create scatter plot is with the `gscatter` function. This function creates a grouped scatter plot where points from each group have a different color or marker.
- ☐ Cluster Analysis using Self-Organizing Maps(SOM)
- ☐ The PCA can now be clustered using SOM clustering algorithm in Neural Network. The function `newsom` is used to create a SOM network which can then be trained with the `train` function. The function `plotsom` is used to display the network over a scatter plot.

V. Experiment and Results

The experiments are conducted on the microarray data of Brain tumor using the MATLAB tool.

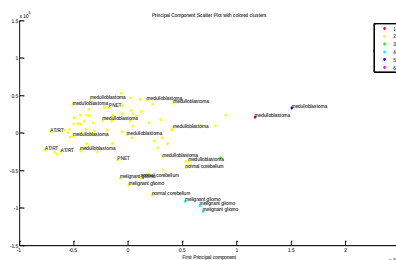
- a. A Scatter plot of the scores of the first two principal components shows that there are distinct regions

Figure 3: Scatter plot



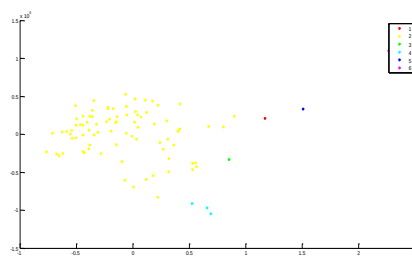
- b. The function `gname` can be used to identify genes on a scatter plot

Figure 4: Principal Component plot with colored clusters



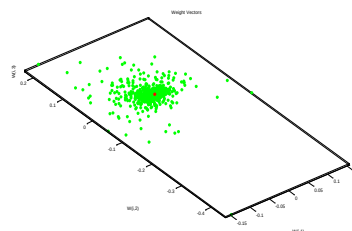
- c. Alternate way to create scatter plot using gscatter function. This function creates a group scatter plot where points from each group have a different color.

Figure 5: Scatter plot using gscatter



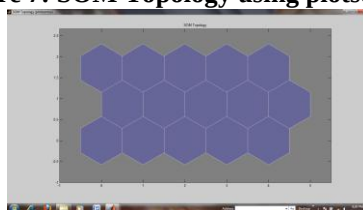
- d. Gene Expression Analysis: Cluster Analysis-Self Organizing Maps(SOM)- The principal components can be clustered using SOM clustering using Neural Network. The newsom function is used to create SOM map network which can then be trained with the train function.

Figure 6: SOM map network



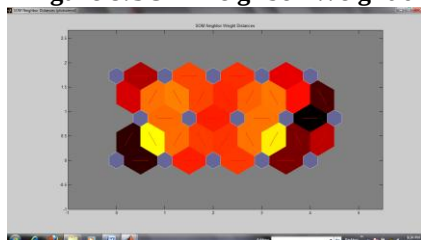
- e. For SOM training, the weight vector associated with each neuron moves to become the center of a cluster of input vectors. In addition, neurons that are adjacent to each other in the topology should also move close to each other in the input space. The figure 7 shows SOM topology using plotsom. The function plotsom is used to display the network over a scatter plot of the data.

Figure 7: SOM Topology using plotsomtop



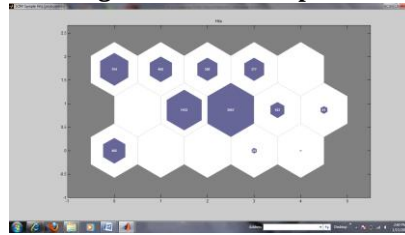
- f. Because the above SOM has a 2D topology, we can visualize in two dimensions the relationships among the 4D cluster centers. Figure 8 shows weight distance matrix also known as U-matrix). Clusters are assigned using SOM by finding the nearest node to each point in the data set

Figure 8:SOM Neighbor Weight distances



- g. Plotsomhits function plots a SOM layer, with each neuron showing the number of input vectors that it classifies. The figure 9 shows SOM sample hit.

Figure 9: SOM sample Hit



VI. Conclusions

In the present work the analysis based on PCA and clustering on Brain tumor data is presented and discussed. Here we prove that principal components are the continues solution to the discrete cluster membership. The result provides new insight to the observed effectiveness of PCA-based data reduction; SOM cluster, SOM topology, SOM neighbor connections etc. are obtained using SOM based neural network. The results obtained are interesting.

VII. References

- [1] Chris Ding and Xiaofeng He, Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, "K-means clustering via Principal Component Analysis. 21st Internal Conference on Machine Learning, Banff, Canada, 2004.
- [2] Chien-sing Lee, Fac.ofinf.Technol, Multimedia, Selangor, Malaysia Singh, Y.P, "Student modeling using PCA of SOM clusters", Advanced Learning Technologies, 2004, proceeding,IEEE
- [3] Suwardi annas*, Takenori Kanai and Shuhei Koyama, Osaka Prefecture University, Gakuen cho i-1, Nakaku Sakai, Osaka 599-8531, Japan , "Principal component analysis and SOM for visualizing and classifying Fire Risks in Forest Regions" Agricultural Information Research 16(2), 2007.44-51
- [4] Alexe G, Dalgin GS, Ganesan S, Delisi C, Bhanot G, The Broad Institute of MIT and Harvard, 7 Cambridge, MA 02142, USA." Analysis of breast cancer progression using PCA and clustering", The Broad Institute of MIT and Harvard, 7 Cambridge, MA 02142, USA.
- [5] XIAOXU HAN, Department of Mathematics and Bioinformatics Program, Eastern Michigan University," IMPROVING GENE EXPRESSION CANCER MOLECULAR PATTERN DISCOVERY USING NONNEGATIVE PRINCIPAL COMPONENT ANALYSIS" , Genome Informatics 21: 200-211 (2008)
- [6] Manojit Chattopadhyay, Pranab K. Dan, Sitanath Majumdar, "Principal component analysis and self organizing map for visual clustering of machine-part cell formation in cellular manufacturing system", Cornell University Library on 26 Jan 2012.
- [7] Shanghai University of Finance and Economics, Shanghai, P.R. China and Queensland Facility for Advanced Bioinformatics, University of Queensland, St Lucia, QLD 4072, Australia , "Independent Principal Component Analysis for biologically meaningful dimension reduction of large biological data sets", Sup'Biotech, Villejuif, F-94800, France .
biomedcentral.com/bmcbioinformatics/article/10.1186/1471/2105/13/24
- [8] Ye Wenyu, Coll.of precision Instrum & Opto-Electron.Eng. Tianjin Univ and Li Gang; Lin Ling; Yu Qilian. "ECG analysis based on PCA and SOM", Neural Networks and Signal Processing, 2003. Proceeding of 2003 Internation conference:IEEE Xplore.
- [9] Pomeroy SL, Tamayo P, Gaasenbeek M, et al. Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature. 2002;415(6870):436–442. [[PubMed](#)]