ISSN No. 0976-5697

Volume 4, No. 7, June 2013 (Special Issue)

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Spatial Data Mining using PCA

Dr. Ritu Bhargava
Govt. Girls Engineering College,
Ajmer (Raj) India

Pooran Singh
Dept. Computer Science,
MJRP University, Jaipur (Raj) India

Prakash Singh Tanwar
Dept. Computer Science,
MJRP University, Jaipur (Raj) India

H.Ram
Rajasthan University,
Jaipur (Raj) India

Abstract: Spatial Data Mining is location based Data Mining. This paper shows the basic concept of PCA with its practical implementation in ArcGIS. This paper shows the process to calculate principal components from the GIS Image. This paper explores some of the basics of Principal Component Analysis. This paper also demonstrates its advantages.

Keywords: Geographic information systems, Principal Component Analysis, Data Mining, Spatial Data Mining.

I. INTRODUCTION

Data mining is a step in the KDD process that consists of applying data analysis and discovery algorithms that produce a particular enumeration of patterns (or models) over the data.[10]

Analysis is an important part of GIS which allows spatial operations with data (e. g. network analysis or filtering of raster data), measuring functions (e.g. distance, direction between objects), statistic analyses or terrain model analysis (e. g. visibility analysis).

Spatial data mining is a special kind of data mining . The main difference between data mining and spatial data mining is that in spatial data mining tasks we use not only non-spatial attributes (as it is usual in data mining in non-spatial data), but also spatial attributes [8].

The central idea of principal component analysis is to reduce the dimensionality of a data set in which there are a large number of interrelated variables, while retaining as much as possible of the variation present in the data set. This reduction is achieved by transforming to a new set of variables, the principal components, which are uncorrelated, and which are ordered so that the first few retain most of the variation present in all of the original variables. Computation of the principal components reduces to the solution of an eigenvalue-eigenvector problem for a positive-semidefinite symmetric matrix.[4]

Previous Work

PCA was first formulated in statistics by Pear- son, who formulated the analysis as finding “lines and planes of closest fit to systems of points in space”. This geometric interpretation will be further discussed in Section 4. PCA was briefly mentioned by Fisher and MacKenzie as more suitable than analysis of variance for the modelling of response data. Fisher and MacKen- zie also outlined the NIPALS algorithm, later rediscovered by Wold. Hotelling further developed PCA to its present stage[9].

PCA now goes under many names. Apart from those already mentioned, singular value decomposition (SVD) is used in numerical analysis and Karhunen-LoCve ex- pansion in electrical engineering. Eigenvector analysis and characteristic vector analysis are often used in the physical sciences. In image analysis, the term Hotelling transformation is often used for a principal component projection. Correspon- dence analysis is a special double-scaled[9]

PCA can be used in Simplification, Data Reduction, Modeling, Outlier detection, Variable Selection, Classification and Prediction

II. METHODOLOGY

Mathematical Definition of PCA

Principal component analysis (PCA) is a mathematical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components.

If n variables are there in the input then the total number of principal components are less than or equal to the number of input variables. This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it be orthogonal to (i.e., uncorrelated with) the preceding components.

Image Matrix

Image Matrix=(Img Vec1,
Img Vec2,
Img Vec3,…)

Variance and Covariance

$$\mathrm{var}(X) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(X_i - \overline{X})}{n-1}$$

$$C(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{n-1}$$

Covariance Matrix

$$C = \begin{pmatrix} \operatorname{cov}(x,x) & \operatorname{cov}(x,y) & \operatorname{cov}(x,z) \\ \operatorname{cov}(y,x) & \operatorname{cov}(y,y) & \operatorname{cov}(y,z) \\ \operatorname{cov}(z,x) & \operatorname{cov}(z,y) & \operatorname{cov}(z,z) \end{pmatrix}$$

Eigenvector and eigen value

The eigenvector's matrix V can be computed by the following equation, Matrix V diagonalizes the covariance matrix C:

$$V^{-1}CV=D$$

where D is the diagonal matrix of eigen values of C.

Matrix D is in the form of an M × M diagonal matrix, where

$D[p,q] = L_m$ for p=q=m

is the mth eigen value of the covariance matrix C

and D[p,q]=0 for p != q

Feature Vector

Feature Vector=(ev1,ev2,ev3,…)

PCA Image

Final_data=Row_feature_Vector * row_data_adj

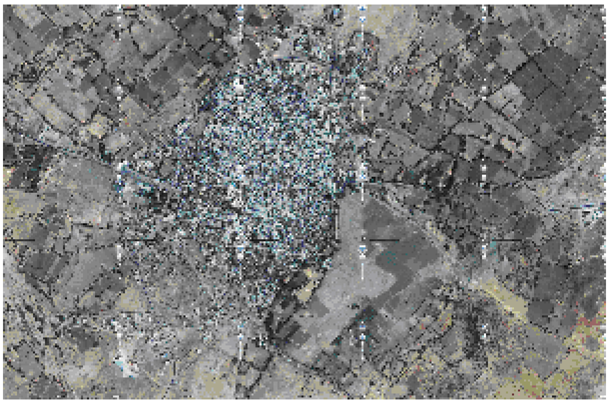


Figure 1 Image for PCA

III. ALGORITHM

Step 1: Input the Image

Now this image have three bands Red, Green and blue

Image Matrix=(Img Vec1,
Img Vec2,
Img Vec3,…)

Step 2: subtract the mean

Step 3:Calculate the covariance matrix from the formula given in methodology

$$C = \begin{pmatrix} \operatorname{cov}(x,x) & \operatorname{cov}(x,y) & \operatorname{cov}(x,z) \\ \operatorname{cov}(y,x) & \operatorname{cov}(y,y) & \operatorname{cov}(y,z) \\ \operatorname{cov}(z,x) & \operatorname{cov}(z,y) & \operatorname{cov}(z,z) \end{pmatrix}$$

Step 4: Calculate the eigenvectors and eigenvalues of the covariance matrix

V-1CV=D

Step 5: Choosing components and forming a feature vector

Feature Vector=(ev1,ev2,ev3,…)

Step 6: Deriving the new data set

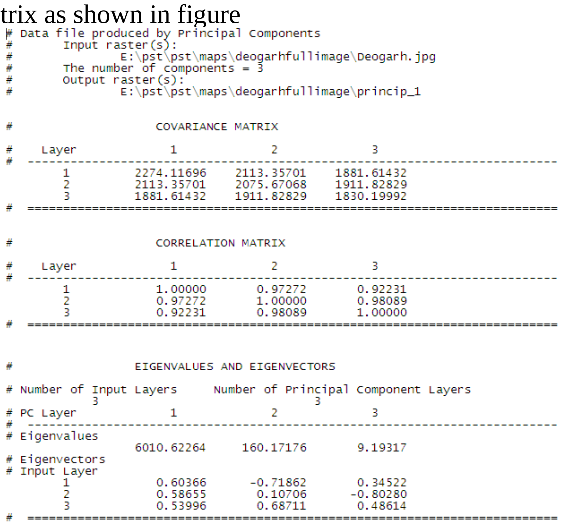
Final_data=Row_feature_Vector * row_data_adj

where

Row_feature_Vector is the matrix with the eigenvectors in the columns transposed so that the eigenvectors are now in the rows, with the most significant eigenvector at the top,and

row_data_adj is the mean-adjusted data transposed, ie. The data items are in each column, with each row holding a separate dimension.

Once we have performed PCA, we have our original data in terms of the eigenvectors we found from the covariance matrix as shown in figure



```
# Data file produced by Principal Components
#       Input raster(s):
#               E:\pst\pst\maps\deogarhfullimage\Deogarh.jpg
#       The number of components = 3
#       Output raster(s):
#               E:\pst\pst\maps\deogarhfullimage\princip_1


#                        COVARIANCE MATRIX

#    Layer              1              2              3
#   -------------------------------------------------------------------------
        1          2274.11696     2113.35701     1881.61432
        2          2113.35701     2075.67068     1911.82829
        3          1881.61432     1911.82829     1830.19992
#   =========================================================================


#                        CORRELATION MATRIX

#    Layer              1              2              3
#   -------------------------------------------------------------------------
        1             1.00000        0.97272        0.92231
        2             0.97272        1.00000        0.98089
        3             0.92231        0.98089        1.00000
#   =========================================================================


#                     EIGENVALUES AND EIGENVECTORS

# Number of Input Layers      Number of Principal Component Layers
           3                                  3
# PC Layer              1              2              3
#   -------------------------------------------------------------------------
# Eigenvalues
                     6010.62264      160.17176        9.19317

# Eigenvectors
# Input Layer
        1             0.60366       -0.71862        0.34522
        2             0.58655        0.10706       -0.80280
        3             0.53996        0.68711        0.48614
#   =========================================================================
```

Figure 2 Cavariance matrix, Correlation Matrix, Eigen Values and Eigenvectors for the image of figure 1

Now for given image we have three principal components.

Step 7 Use first Principal component as Red Band(now we have most of the details in this band)

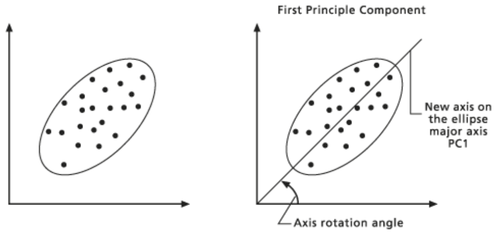

First Principle Component
New axis on the ellipse major axis PC1
Axis rotation angle

Figure 3 Graph between two bands of an image with first Principal component.

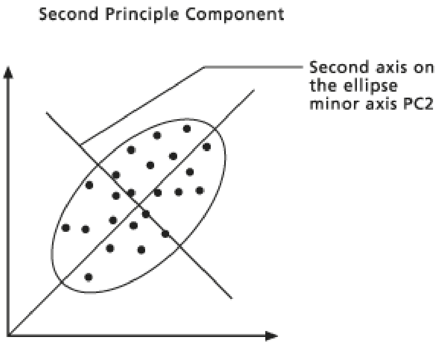

Second Principle Component
Second axis on the ellipse minor axis PC2

Figure 4 Second Principal Component

First principal component as shown in figure 3 describes the direction of maximum variance we use this as the first band of the image. Then an orthogonal perpendicular line to PC1 is the second principal component (PC2) as shown in figure 4 The new axis for the original y-axis, describes the second most variance not described by PC1. we use this as our second band of the image.

Discovering important structure from data

Data Mining is a data-driven hypothesis generation process. To describe the data we use various data description methods like summarizing data using mean, median, mode, and variance

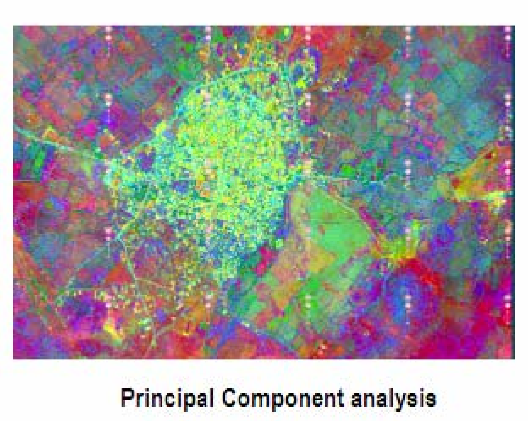


Figure 5. Image after rotating principle component

IV. Experiments and Results

After rotating the principle component we can see from the image that now we have more details on first band of the image and now it is easy to classify the image from unsupervised classification

V. Conclusion

In this study, we presented the basic concept of principal component analysis. We focused on the conversion of an Image using PCA method. The advantage of PCA is that we have more details on first or second band of an image, which is useful in classifying the image.

VI. References

[1] Ester M., Kriegel H. P., Sander J.: Spatial Data Mining: A Database Ap- proach. Proc.of the Fifth Int. Symposium on Large Spatial Databases (SSD ’97), Berlin, Germany, Lecture Notes in Computer Science, Springer, 1997.

[2] Fayyad U.M.,Piatetski-Shapiro G., Smyth P., Uthurusamy R. (eds.): Advances in Knowledge Discovery and Data Mining. AAAI/ MIT Press 1996.

[3] H. Tang and S. McDonald, “Integrating GIS and spatial data mining techniques for target marketing of university courses,” Symposium on Geospatial Theory, Processing and Applications, Ottawa, 2002.

[4] I. Jolliffe, Principal component analysis. Wiley Online Library, 2005.

[5] J. Mennis and D. Guo, “Spatial data mining and geographic knowledge discovery—An introduction,” Computers, Environment and Urban Systems, vol. 33, no. 6, pp. 403–408, 2009.

[6] K. Koperski, J. Han, and J. Adhikary, “Mining knowledge in geographical data,” Communications of the ACM, vol. 26, no. 1, pp. 65–74, 1998.

[7] Koperski K., Han J., Adhikary J.: Mining Knowledge in Geographical Data. To appear in Comm. of ACM 1998. http:/ / db.cs.sfu.ca/ sections/ publication/ kdd/ kdd.html

[8] P. Kuba, “Data structures for spatial data mining,” Masaryk University Brno, Czech Republic, September 2001, 2001.

[9] S. Wold, K. Esbensen, and P. Geladi, “Principal component analysis,” Chemometrics and intelligent laboratory systems, vol. 2, no. 1, pp. 37–52, 1987.

[10] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “From data mining to knowledge discovery in databases,” AI magazine, vol. 17, no. 3, p. 37, 1996.

[11] V. Bogorny, A. T. Palma, P. Engel, and L. O. Alvares, “Weka-gdpm: Integrating classical data mining toolkit to geographic information systems,” SBBD Workshop on Data Mining Algorithms and Aplications (WAAMD 2006), Florianopolis, Brasil, October, pp. 16–20, 2006.

[12] W. Wu, Modeling Spatial Dependencies for Data Mining. University of Minnesota, 2002.