AN ANALYSIS OF DENSITY BASED CLUSTERING TECHNIQUE WITH DIMENSIONALITY REDUCTION FOR DIABETIC PATIENT

Rekha Awasthi, Anil Kumar Tiwari and Seema Pathak
Disha College, Raipur, India

ABSTRACT:

Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. In a case of Medical Database, Data mining plays a vital role in helping physician to identify effective treatment and patients to receive better and more affordable medical services. There are large quantities of information about patients and their medical conditions. The discovery of trends and patterns hidden within the data could significantly enhance understanding of disease and medicine progression and management by evaluating stored medical documents. Methods are needed to facilitate discovering the trends and patterns within such large quantities of medical documents.

Normalization is an important preprocessing step in Data Mining to standardize the values of all variables from dynamic range into specific range. And PCA[6] and multidimensional scaling are widely used in Medical Database analysis and are often considered for machine learning techniques. To obtain an efficient processing time and increase performance of cluster use DBSCAN[5] and OPTICS [4] (density based clustering algorithm, apply normalization and PCA) on dataset and obtain a reduced dataset containing possibly uncorrelated variables.

Diabetic patient have continuously increase due to genetic Susceptibility, Autoimmune Destruction of Beta Cells, Autoimmune Destruction of Beta Cells, Obesity and Physical Inactivity, Abnormal Glucose Production by the Liver etc. Therefore, here Pima Indian diabetic database (PIDD) is considered and evaluated by univariate analysis and feature selection method are used.

Keywords: Density Based cluster, Diabetic Data, Normalization, PCA.

[1] INTRODUCTION

As we know large amount of data is stored in Health database, so in order to get required data & to find the hidden relationship, different data mining techniques are developed & used. There are varieties of popular data mining task within the Medical data mining e.g. classification, clustering, outlier detection, association rule, prediction etc. We can use the data mining in Health system as: predicting human characteristics whether they are prone to decease or not, relationship between the human disease test
reports & their impact on human health, predicting patient's risks to be infected with disease, discovery of strongly related factors causing disease, knowledge discovery on medical data.

In this paper, explore the connection between following used methods. We prove that principal components are actually the continuous solution of the cluster membership indicators in the DBSCAN [5] and OPTICS [4] clustering method, i.e., the PCA[6] dimension reduction automatically performs data clustering according to the DBSCAN and OPTICS objective function. This provides an important justification of PCA-based on Diabetic database reduction.

**Data Set**

In this study, Pima Indian diabetic database (PIDD) was used.

- **Number of Instances: 768**
- **Number of Attributes: 8 plus class**

Attributes :
- 1. Number of times pregnant
- 2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
- 3. Diastolic blood pressure (mm Hg)
- 4. Triceps skin fold thickness (mm)
- 5. 2-Hour serum insulin (mu U/ml)
- 6. Body mass index (weight in kg/(height in m)^2)
- 7. Diabetes pedigree function
- 8. Age (years)
- 9. Class variable (0 or 1)

These features are detailed in Table 1. All patients in this dataset are Pima Indian women at least 21 years old and living near Phoenix, Arizona, USA. The binary target variable takes the values ‘0’ or ‘1’. While ‘1’ means a positive test for Diabetes, ‘0’ is a negative test. There are 268 cases in class ‘1’ and 500 cases in class ‘0’.

**Table 2 (Brief statistical analysis)**

<table>
<thead>
<tr>
<th>Features number</th>
<th>Mean</th>
<th>Standard deviation</th>
<th>Min/max</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.8</td>
<td>3.4</td>
<td>0/17</td>
</tr>
<tr>
<td>2</td>
<td>120.9</td>
<td>32.0</td>
<td>0/199</td>
</tr>
<tr>
<td></td>
<td>69.1</td>
<td>19.4</td>
<td>0/122</td>
</tr>
<tr>
<td>4</td>
<td>20.5</td>
<td>16.0</td>
<td>0/99</td>
</tr>
<tr>
<td>5</td>
<td>79.8</td>
<td>115.2</td>
<td>0/846</td>
</tr>
<tr>
<td>6</td>
<td>32.0</td>
<td>7.9</td>
<td>0/67.1</td>
</tr>
<tr>
<td>7</td>
<td>0.5</td>
<td>0.3</td>
<td>0.078/2.42</td>
</tr>
<tr>
<td>8</td>
<td>33.2</td>
<td>11.8</td>
<td>21/81</td>
</tr>
</tbody>
</table>

[2] METHODS

2.1 **OPTICS and DBSCAN**

Density based clustering is to discover clusters of arbitrary shape in spatial databases with noise. It forms clusters based on maximal set of density connected points. The core part in Density-Based clustering is density-reach ability and density connectivity. Also it requires two input parameters i.e. Eps which is known as radius and the MinPts i.e. the minimum number of points required to form a cluster. It starts with an arbitrary starting point that has not visited once. Then the e- neighborhood is retrieved, and if it contains sufficiently many points than a cluster is started. Otherwise, the point is labeled as noise. This
section describes two density based clustering algorithms briefly i.e. DBSCAN (Density Based Spatial Clustering of Application with Noise)[5] and OPTICS (Ordering Points to Identify the Clustering Structure)[4]. Here, Density=number of points within a specified radius. Density-Based clustering Algorithms mainly include two techniques:
- DBSCAN which grows clusters according to a density-based connectivity analysis.
- OPTICS extends DBSCAN to produce a cluster ordering obtained from a wide range of parameter setting

2.2 NORMALIZATION

Normalization Approach[7]:
Data Mining can generate effective result if proper and effective data mining technique can apply to the dataset.
Data Normalization standardize the raw data by converting them into specific range using linear transformation which can generate good quality clusters and improve the accuracy of clustering algorithms. According, normalization is used to standardize all the features of the dataset into a specified predefined criterion so that redundant or noisy objects can be eliminated and use made of valid and reliable data which can effect and improve accuracy of the result. The importance of normalization is that it enhances the accuracy of the results that are obtained during clustering. Better results are generated when data preparation with data preprocessing and normalization is carried out with different techniques. Data normalizations techniques include min-max normalization, Z-Score normalization.

2.3 PCA

Principal component analysis[6]

Probabilistic and Statistical Model for Medical Database - Principal component analysis perform covariance analysis between factor decrease dimensionality of the data. By itself, it is fit for data sets in multiple dimensions on Medical Database For our proposed work, we are performing PCA using the covariance method. Following is a comprehensive explanation of PCA using the covariance method:

a. Organize the original dataset in a matrix form.

b. Subtract the mean from each of the data dimensions.
c. Find the data covariance matrix.
d. Rearrange the eigenvectors and Eigen values of the covariance matrix.
e. Remove weaker components from PCs and form transformation matrix consisting of significant PC’s.
f. Find the reduced data set using the reduced PCs.

i. Eigenvector matrix \( V \) and Eigen value matrix \( D \) in the decreasing order of Eigen values. This gives us the components in order of significance.

f. Remove weaker components from PCs and form transformation matrix consisting of significant PC’s: Suppose we have a sample data \( X \) comprising of ‘A’ records and each having ‘B’ attributes, and we want to reduce the data such that each record will have only ‘C’ attributes in such a way that \( C < B \).

a. Organize the high dimensional dataset \( X \) in a matrix

\( S \): Arrange data as a set of ‘N’ data vectors \( S_1, S_2, \ldots, S_N \) where each \( Sn \) represents a single grouped record of the ‘B’ attributes. Write \( S_1, S_2, \ldots, S_N \) as column vectors, each of which has \( B \) rows. Place the column vectors into a single matrix \( S \) of dimensions \( B \times N \)

b. Subtract the mean from each of the data dimensions: For PCA to work properly, we have to subtract the mean from each of the data dimensions. Find the mean along each dimension \( b = 1, 2, \ldots, B \). Then,
place the calculated mean values into mean vector of dimensions $B \times I$. Later, subtract the mean vector from each $S_n$ values of the data matrix $S$. Now, store the mean-subtracted data in the $B \times N$ matrix $U$. This produces a data set whose mean is zero.

c. Find the data covariance matrix $C$: Find the $B \times B$ covariance matrix $C$ from the outer product of matrix $U$ with itself.

$$C = \frac{1}{N} \sum UU^* \quad (1)$$

d. Find the eigenvectors and Eigen values of the covariance matrix $C$: Compute the matrix $V$ of eigenvectors which diagonalizes the covariance matrix $C$

$$V^{-1}CV = D \quad (2)$$

- $D$ is the diagonal matrix containing Eigenvalues of $C$ which will take the form of $B \times B$ diagonal matrix.
- Matrix $V$ is also of dimension $B \times B$ contains $B$ column vectors corresponding to the $B$ eigenvectors of the covariance matrix $C$.
- For a covariance matrix, the eigenvectors correspond to principal components and the eigen values to the variance explained by the principal components.

e. Rearrange the eigenvectors and Eigen values in decreasing order: The eigenvector with the highest Eigen value is the principle component of the data set. Thus, assemble the columns of the eigenvector matrix $V$ and Eigen value matrix $D$ in the decreasing order of Eigen values. This gives us the components in order of significance.

f. Remove weaker components from PCs and form transformation matrix consisting of significant PC’s: Selecting the number of PC’s is a significant question. The largest eigen values correspond to the principal components that are related with a large amount of the covariability amongst a number of observed data. Hence, we will remove weaker principal components from the set of components obtained. For the removal, perform any one of the three suitable methods explained. And generate the transformation matrix $P$ with reduced PCs is formed.

g. Find the reduced data set using the reduced PC’s The transformation matrix $P$ is applied to the original data set $X$ to produce the new reduced projected dataset $H$ which we can make use for data clustering.

[3] Experiment Investigation and Result
3.1 Data Pre-processing:

This module calculates tuples with missing values using different options like maximum, minimum, constant, average and standard deviation for the treatment of missing values tuples before we apply normalization approach on the dataset. This process gives the treatment of missing value data and then it applies to the second part (data normalization) of data preparation.

Apply normalization approach:
The importance of normalization is that it enhances the accuracy of the results that are obtained during clustering

- The Min-Max normalization technique involves the linear transformation on raw data.
- Z-Score normalization technique is useful when the actual minimum and maximum value of attribute

Apply PCA

It is a powerful tool for analyzing data by finding these patterns in the data. Then compress them by dimensions reduction without much loss of information. This showed that using four selected features by PCA is more promising than using all the eight features. The experiments was conducted in Weka 3.7

Weka Output Screen:
principal components might look like
PC1 : 1*Insulin + 0.01*Glucose + ..
PC2 : 1*Glucose + 0.12*Age + 0.12*DiastolicBP + ..
PC3 : 0.92 * DiastolicBP + 0.31*Triceps

Apply Density Based Clustering Algorithm

It forms clusters based on maximal set of density connected points. The core part in Density-Based clustering is density-reach ability and density connectivity. Also it requires two input parameters i.e. Eps which is known as radius and the MinPts i.e. the minimum number of points required to form a cluster.

Table 3 : (Result )

<table>
<thead>
<tr>
<th>Diabetic data set</th>
<th>Algorithm</th>
<th>Run time</th>
<th>cluster</th>
<th>Unassigned</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DBSCAN</td>
<td>1.31</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>DBSCAN with Normalization</td>
<td>1.25</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>DBSCAN with PCA</td>
<td>1.15</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>DBSCAN Normalization with PCA</td>
<td>1.09</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>OPTICS</td>
<td>1.78</td>
<td>0</td>
<td>768</td>
</tr>
<tr>
<td></td>
<td>OPTICS with Normalization</td>
<td>1.75</td>
<td>0</td>
<td>768</td>
</tr>
</tbody>
</table>
**OPTICS with PCA** | 1.75 | 0 | 768  
**OPTICS Normalization with PCA** | 1.53 | 0 | 768

[4] CONCLUSION
Data mining is becoming an important instrument for researchers and clinical practitioners in medicine. Before clustering, normalization is used to standardize all the features of the dataset into a specified predefined criterion so that redundant or noisy objects can be eliminated and use made of valid and reliable data which can effect and improve accuracy of the result. Normalization is that it enhances the accuracy of the results that are obtained during clustering [8]. Better results are generated when data preparation with data preprocessing and then dimension reduction before cluster and increase accuracy of cluster performance. Based on result dimension reduction can shorten processing time has lowest processing time. The advantages of PCA are identifying patterns in data, and expressing the data in such a way as to highlight their similarities and differences. It is a powerful tool for analyzing data by finding these patterns in the data. Then compress them by dimensions reduction without much loss of information. DBSCAN and OPTICS Normalization with PCA has lowest processing time for Diabetic Data.

REFERENCES
[1]. J. Han and M Kamber “data mining concept and technique “
[3]. Martin Ester, Hans-Peter Kriegel, Jörg Sander and Xiaowei Xu, A Density- Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, The Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, Oregon, USA, 1996
[5]. K. Muntaz, Dr. K. Duraiswamy, An Analysis on Density Based Clustering of Multidimensional Spatial Data in Indian Journal of Computer Science and Engineering Vol 1 No 1 8-12
[6]. Lindsay I Smith, A tutorial on Principal Components Analysis, 2002
[7]. K.A.J. Doherty a R.G. Adams a N. Davey a Unsupervised Learning with Normalised Data and Non-Euclidean Norms
[8]. Qi Luo “Knowledge Discovery and Data Mining” Work shop on Knowledge Discovery and Data Mining , 2008 pp.3-5.
[17]. Rene´ Vidal, Member, IEEE, Yi Ma, Member, IEEE, and Shankar Sastry, Fellow, IEEE “Generalized Principal Component Analysis (GPCA)” VOL. 27, NO. 12, DECEMBER 2005.