SOFT COMPUTING REPORT

Water Quality Analysis

Submitted to : Dr M Venkatesan
Submitted on: 29th April, 2018

Title:

Water Quality Analysis

Cluster members:

Team 1 - Principal Component Analysis (PCA)
15co239 - RuhiTaj Reddypalli
15co253 - Y M Greeshma

Team 2 - Support Vector Machines (SVM)
15co211 - Vasudha Boddukuri
15co253 - Sai Priya D

Team 3 - Integration of PCA and SVM
15co225 - Ranjith M
15co152 - Suhag A

Aim:

To detect the water quality using PCA and SVM from the collected set of the data in the past.

Objective:

- Feature Extraction using Principal Component Analysis
- Prediction using Support Vector Machines
- Integration to get more optimised water quality analysis
- Prediction of Water Quality

Abstract :

Over the past few years, due to excessive pollution and poor water treatment and management, the quality of water has gone down drastically. This has affected drinking water and water for domestic purposes as well. In many places the water people get for domestic use is hard water full of dissolved minerals and salts. The effects of unclean water are far-reaching, impacting every aspect of life.

Therefore, management of water resources is very crucial in order to optimize the quality of water. The effects of water contamination can be tackled efficiently if data is analyzed and water quality is predicted beforehand. The proposed architecture evaluates the quality of water using Principal Component Analysis and Support Vector Machine. Principle Component Analysis(PCA) reduces the number components by choosing the ones that contribute most to the data's variance. Support Vector Machine(SVM) is a supervised learning model that predicts the category to which a given example belongs.

Introduction:

Principal Component Analysis (PCA) is a dimension-reduction tool that can be used to reduce a large set of variables to a smaller set that still contains most of the information in the large set. PCA does this by calculating the variance contributed by each attribute. The attributes/variables are selected based on their variance starting from the highest. This is done so that a subset of all variables can account for most of the variance in the data. Data was collected from multiple sources about the water quality in the rivers and water bodies in many places. 8 variables were considered including the amount of minerals like Magnesium, Phosphorus, Potassium and Carbonates and Bicarbonates present in the water sample. PCA was used to reduce this number to 5.

Support Vector Machines (SVMs) are used to classify the given dataset into classes based on a particular distribution. The distribution may be linear or nonlinear based on the applications. The data collected for PCA was used for this step as well. We used 80% of the data as training data and the rest 20% as the testing data. The SVM predicts the value of water quality given the other parameters.

When we have a lot of attributes and we want to include only the important ones, we use PCA. In this project, we combine the PCA step with SVM by first reducing the number of attributes and then learning based on the remaining data. The performance of the above three methods are analysed.

Methodology:

1.Principal Components Analysis:

PCA is a way of identifying patterns in data, and expressing the data in such a way as to highlight their similarities and differences. Since patterns in data can be hard to find in data of high dimension, where the luxury of graphical representation is not available, PCA is a powerful tool for analysing data. The other main advantage of PCA is that once you have found these patterns in the data, and you compress the data, ie. by reducing the number of dimensions, without much loss of information.Using a PCA we can now identify what are the most important dimensions and just keep a few of them to explain most of the variance we see in out data. Hence we can drastically reduce the dimensionality of the data and make EDA feasible again. Moreover, it will also enable us to identify what the most important variables in the *original* feature space are, that contribute most to the most important PCs. Intuitively, one can imagine, that a dimension that has not much variability cannot explain much of the happenings and thus is not as important as more variable dimensions.

Eigenvalue: Column sum of squared loadings for a factor, i.e., the latent root. It conceptually represents that amount of variance accounted for by a factor.

Specific or unique variance: Variance of each variable unique to that variable and not explained or associated with other variables in the factor analysis.

Eigenvectors:
Principal components (from PCA - principal components analysis) reflect both common and unique variance of the variables and may be seen as a variance-focused approach seeking to reproduce both the total variable variance with all components and to reproduce the correlations.

The ratio of eigenvalues is the ratio of explanatory importance of the factors with respect to the variables. If a factor has a low eigenvalue, then it is contributing little to the explanation of variances in the variables and may be ignored as redundant with more important factors. Eigenvalues measure the amount of variation in the total sample accounted for by each factor.

INPUT:
Data set containing a list of water quality at different places. The number of original dimensions in the data set are 8 - ['MG','PH','K','NITRATE','SULPHATE','CARBONATE','CHLORIDE','FLUORIDE', 'Water Quality']

PROCESS:

The above given 8-dimensional data set is reduced into n dimensional data set required.

OUTPUT:

- The first obtained component has high variance and higher importance in the output and hence the rate of other components.
- 1st comp > 2nd comp > ..... > nth component
- The eigenvectors of the original data are also obtained.
- From this, we obtain n-reduced dimensions of the original dataset.

- From all these datasets, we obtain the dimension which contributes the most to output, with chlorine having the highest contribution(with n=5):

    1 CHLORIDE
    2 SULPHATE
    3 MG
    4 FLUORIDE
    5 NITRATE
    6 K
    7 PH
    8 CARBONATE

2.Support Vector Machine (SVM):

The given 8 dimensional input is reduced to 'n' dimensions(n<=8) using PCA.Here,SVM maps the original sample space to a high-dimensional linear space through a nonlinear transformation, then in this new space to seek the optimal linear classification surface.
The following description explains about SVM

Support Vector Machine (SVM) is a pattern recognition technique proposed by Cortes and Vapnik. SVM has a good theoretical basis in solving classification, regression and density function estimation problems. It is mainly used to solve small sample size problems and it is a nonlinear pattern recognition technique. SVM is mapping the original sample space to a high-dimensional linear space through a nonlinear

transformation, then in this new space to seek the optimal linear classification surface. This nonlinear transformation is realized by defining the appropriate kernel function.
In general, SVM has some outstanding characteristics compared with traditional methods:
a. High generalization ability. SVM uses maximal margin to reduce the dimension , hereby the upper bound of SVM can be reduced.
The maximal margin method is useful to solve small sample size problems .
b. Kernel trick. SVM uses kernel function to define the similarity between two datasets. It converts the problem from lower dimension to higher dimension, while the computation complexity remains the same.
c. Sparseness. The less number of SVs (Support vectors, the data taking effect) means better generalization ability.
d. Unique solution. The optimal solution of SVM is solved by a quadratic optimization method. The convex property of the formulation makes the solution unique.
Therefore SVM is used in this study to identify if there are significant differences in element concentrations in drinking water from both counties.

Trace elements in drinking water can influence the longevity of human beings. Support vector machine (SVM),
as a pattern recognition technique for small sample size classification problem, was applied to analyze the difference of element concentrations in the drinking water.

3. Integration of PCA and SVM:

For the third part of the project, we used the dataset with 23 entries and 8 attributes. 18 entries were used as training data while the rest 5 were used as testing data. We first perform Principal Component Analysis of the 8 attributes and reduce it to 5. The next step is reducing the number components in the training data. This is done by transforming the training and testing data with respect to the PCA. We used Python to implement the above steps. While performing PCA before SVM, we may lose some data due to the reduction in the number of attributes. While this may look like a con, it is not.

Experiment and Result Analysis

Phase 1:

Data set : Two- dimensional array of m, n size

- m - no. of places
- n - no. of dimensions

Result :

- New_Dimesions * Old Dimensions

```
[[ 1.17999644e-01 -8.71133036e-06  1.59020640e-02  3.25898958e-02
   1.62202335e-01 -0.00000000e+00  9.79005361e-01  1.09592838e-04]
 [ 4.44560518e-01 -6.17014671e-04 -6.58607365e-02  6.68929699e-01
   5.67507303e-01  0.00000000e+00 -1.68806155e-01  9.64280636e-04]
 [ 1.16383279e-01 -5.71121746e-04 -4.18585981e-01  5.22618913e-01
  -7.27241444e-01 -0.00000000e+00  9.58646898e-02 -6.30068448e-03]
 [ 5.18044165e-02  1.98661593e-03  9.05030565e-01  3.05637380e-01
  -2.90734619e-01  0.00000000e+00  1.70504488e-02 -6.97953696e-04]
 [ 8.78750274e-01 -6.61953928e-04  3.32769342e-02 -4.30028077e-01
  -1.95404671e-01 -0.00000000e+00 -5.97661372e-02 -3.32260448e-03]]
```

- Eigenvectors of the each old dimension in terms of new-

Example: Similarly, the eigenvectors of other dimensions are also calculated.

```
Eigen Vector of the original component - MG :

0.11799964369
0.444560517759
0.116383279179
0.0518044165054
0.878750274093



Eigen Vector of the original component - PH :

-8.71133036007e-06
-0.000617014671151
-0.000571121745916
0.00198661593257
-0.000661953928407



Eigen Vector of the original component - K :

0.0159020639866
-0.065860736534
-0.418585980644
0.905030565064
0.0332769341594
```

- Variance of each new dimension:

```
Explained Variance of the reduced dimensions :

 [0.9487465  0.03810494 0.00749253 0.00303955 0.00261548]
```

- The dataset in reduced dimensions:

```
  1.100e+00]
 [7.330e+01 7.300e+00 2.000e+00 4.000e+00 4.800e+01 0.000e+00 1.140e+02
  3.000e-01]
 [4.020e+01 7.800e+00 1.500e+01 2.000e+00 3.000e+01 0.000e+00 1.910e+02
  1.500e+00]
 [9.490e+01 7.600e+00 2.500e+01 0.000e+00 2.000e+02 0.000e+00 5.320e+02
  2.300e+00]]



Out Matrix after dimension reduction:

[[ 7.09112812e+02  1.69759344e+02  9.34232315e+01  2.84763856e+01
  -5.29281074e+01]
 [-4.84762103e+02 -6.31243309e+01  3.59846435e+01 -5.13085760e+00
   3.78073641e+01]
 [ 1.16863982e+01  4.07885990e+00  4.33362686e+00 -2.89112486e+01
   9.20435776e+01]
 [ 8.05584681e+01  1.36675841e+02 -1.28545665e+01 -1.95473851e+01
   9.37735133e-01]
 [-2.86250512e+02  2.21168511e+02 -9.21458852e+00 -9.90376511e+00
   3.67833231e+01]
```

- Importance of the old data set in terms of new:

```
1 CHLORIDE
2 SULPHATE
3 MG
4 FLUORIDE
5 NITRATE
6 K
7 PH
8 CARBONATE
```

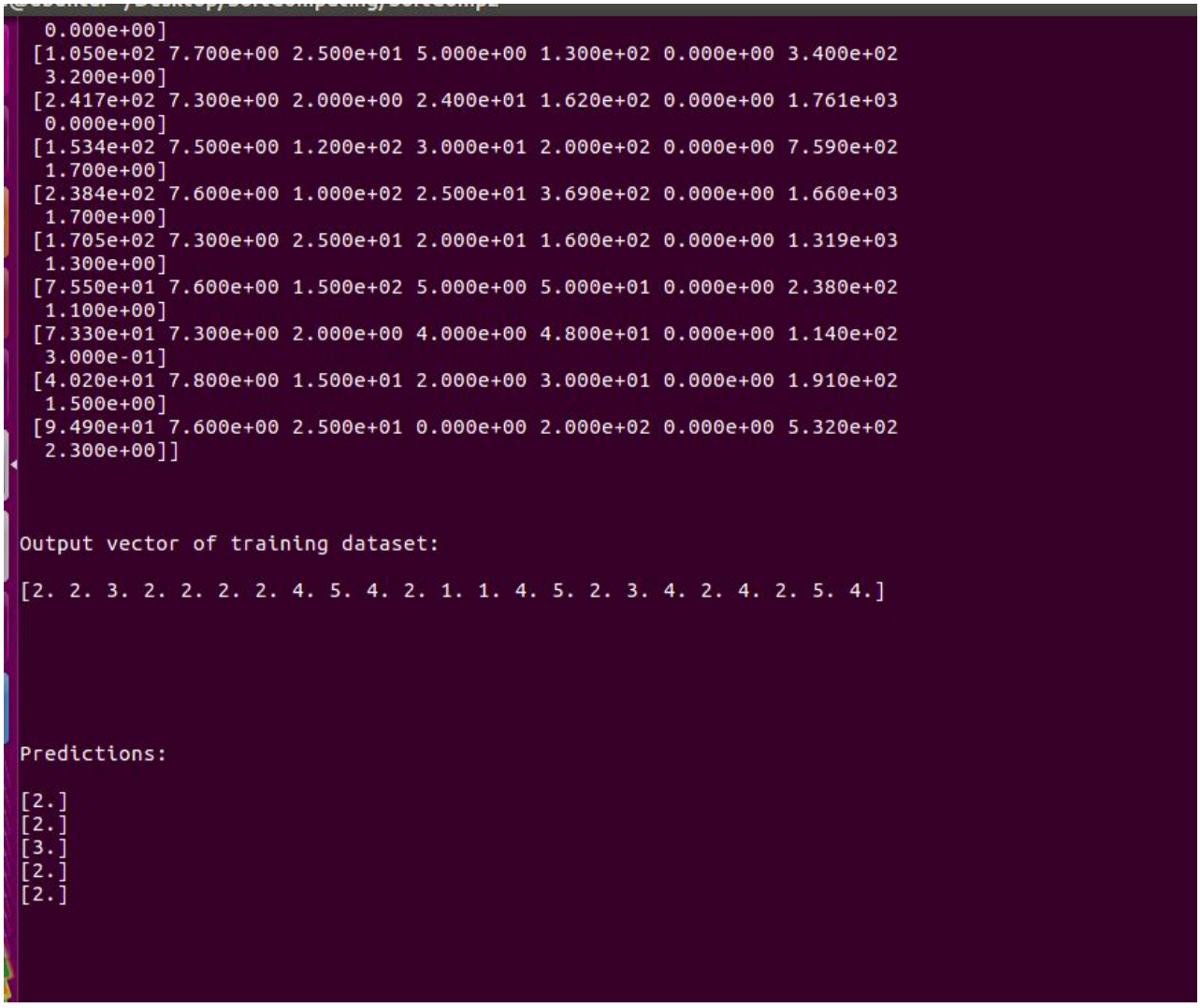
Phase 2:

Data set : Training and Testing data set

- Input and output of training dataset
- Output of training dataset

The input data set is taken from the .csv file.

Result:



Phase 3:

Data set : Training and Testing data set

- Input and output of training dataset
- Output of training dataset

In this case, the training set is taken from PCA component analysis

Output 1: Training dataset with PCA

```
[[ 1.01177457e+03  1.87228272e+01  4.84390451e+02  1.16223297e+02
  -4.64643331e+02 -5.46932122e+02 -3.92147142e+02 -3.98344730e-14]
 [ 1.34571861e+02 -6.18111448e+00  7.69710718e+01  1.88047326e+01
  -7.73256747e+01 -3.85079242e+00 -7.32495431e+01 -6.05713153e-15]
 [ 5.02746187e+02 -2.56454632e+01  2.85525180e+02  3.49864898e+01
  -2.90848191e+02 -1.45947998e+02 -1.73325909e+02 -2.20453989e-14]
 [ 6.22310757e+02  2.27566829e+01  2.54004149e+02  6.14526837e+01
  -2.52830852e+02 -2.43364688e+02 -1.31601078e+02 -1.90148567e-14]
 [ 4.57362663e+02  9.02535412e+01  7.34466087e+01  7.41811128e+01
  -7.77135755e+01 -5.60426843e+01 -1.84780547e+01 -3.65005037e-15]
 [ 8.24042944e+02  1.46841693e+02  1.88758286e+02  1.40980412e+02
  -1.84905139e+02 -2.39223657e+02 -1.55497772e+02 -1.35237124e-14]
 [ 7.33707956e+01  1.45145533e-01  2.44940319e+01  1.11915965e+01
  -2.47668327e+01 -9.62139965e+00 -7.95163599e+00 -1.50141727e-15]
 [ 2.99260903e+02  6.33953478e+00  1.22587076e+02  3.23139970e+01
  -1.22627794e+02 -1.76525078e+02 -5.79933396e+01 -9.44961888e-15]
 [ 3.13936525e+02 -6.00347016e+01  2.54254233e+02  2.49698484e+01
  -2.34641810e+02 -2.29890961e+02 -2.42034484e+02 -2.19902454e-14]
 [ 2.07082978e+02  2.51843373e+01  3.27610861e+01  2.18513246e+01
  -3.62016632e+01 -3.91099488e+01  4.14460074e+01 -8.61549782e-16]
 [ 1.48622050e+03 -1.98142899e+02  9.71042936e+02  3.17318106e+01
  -9.27179766e+02 -8.73550874e+02 -5.38958853e+02 -7.70487974e-14]
 [ 1.67118835e+02 -5.90679009e+00  8.13598180e+01  1.63219587e+01
  -8.18994503e+01 -4.67057984e+01 -4.45897626e+01 -6.02606001e-15]
 [ 8.77736685e+01 -2.42098890e+00  4.04779142e+01  1.30798685e+01
  -4.01843108e+01 -2.35158900e+01 -2.70764014e+01 -2.98370467e-15]
 [ 1.19885885e+02 -1.76980358e+00  3.86232822e+01  1.04281494e+01
  -3.15146378e+01 -1.85389176e+01  1.06034987e+00 -2.02948527e-15]
 [ 2.94810204e+02 -3.43249161e+01  1.54496911e+02  5.41455724e+00
  -1.41579251e+02 -9.69058673e+01 -5.30964178e+01 -1.09825365e-14]
 [ 1.05043828e+03 -1.56341153e+02  7.94020316e+02  5.32624053e+01
  -7.79407198e+02 -6.66724661e+02 -6.10943165e+02 -6.66696526e-14]
 [ 5.87323687e+02 -1.15054712e+02  3.50852341e+02 -1.02676792e+01
  -2.52564777e+02 -2.70491801e+02 -1.91382936e+02 -2.56659712e-14]
 [ 1.14096439e+03 -1.94272230e+02  7.55724948e+02  1.65113269e+00
  -6.71610004e+02 -6.49878171e+02 -4.18587170e+02 -5.88988708e-14]]
```

Training dataset

```
[[ 8.09416229e+02 -1.31777765e+02  5.94756881e+02  2.94207788e+01
  -5.66954372e+02 -5.13267906e+02 -4.28379139e+02 -4.91307639e-14]
 [ 2.18865041e+02 -9.92922873e+01  1.31322020e+02 -3.04535157e+01
  -6.73667892e+00 -5.45679342e+01 -9.20403373e+01 -7.96808329e-15]
 [ 1.20284976e+02 -1.60211792e+00  5.22364364e+01  1.32421479e+01
  -5.40818326e+01 -4.22415650e+00 -2.07012077e+01 -3.45740588e-15]
 [ 1.31963045e+02 -2.67770156e+01  8.75985087e+01  7.08881395e+00
  -7.68042396e+01 -6.21798711e+01 -6.05714336e+01 -6.82647318e-15]
 [ 4.16342014e+02 -5.95032525e+01  2.40990948e+02 -1.39005046e+00
  -2.27501265e+02 -2.03507642e+02 -6.77199766e+01 -1.75890074e-14]]
```

Explained Variance with and without scaling and Accuracy:

```
[2. 2. 3. 2. 2. 2. 2. 4. 5. 4. 2. 1. 1. 4. 5. 2. 3. 4.]
[2. 4. 2. 5. 4.]
Explained Variance without Standard Scaling: [9.49074765e-01 3.93131645e-02 7.29717487e-03 2.94217551e-03
 1.37177660e-03 8.07312640e-07 1.36574850e-07 3.25086417e-40]



[ 94.90747647  98.83879291  99.5685104   99.86272795  99.99990561
  99.99998634 100.         100.        ]



40.00%

Explained Variance after Standard Scaling: [4.39958626e-01 2.39783990e-01 1.35330434e-01 1.06108607e-01
 5.22997864e-02 1.49255412e-02 1.15930149e-02 9.22241885e-34]



[ 43.99586261  67.97426166  81.50730502  92.11816575  97.34814439
  98.84069851 100.         100.        ]
```

Discussion with method and Comparison:

This method uses Principal Components Analysis and creates its own dimensions, with supportable variances. The first component obtained has higher variance and is more weighted in the output.

Support Vector Machines, on the other convert a lower-dimensional into higher dimension using different approaches for classifying input vectors to obtain desirable outputs.

The PCA technique was first used to reduce and orthogonalize the original input variables (data). Then these treated data were used as new input variables in SVM model.

The distribution of Eigenvalues for the covariance matrix of the data, and they get very small. In general SVM's are robust in the cases where data spans a subspace of the full feature dimension. The reason for this is that the SVM operates at the sample level (the kernel is computed between samples) and not at the feature level. A logistic regression without regularization can be thrown off by using PCA.
Hence, when these are combined , good output can be obtained.

Existing models include:

1. A water quality prediction model with the help of water quality factors using Artificial Neural Network (ANN) and time-series analysis. The data includes the measurements of 4 parameters which affect and influence water quality. For the purpose of evaluating the performance of model, the performance evaluation measures used are Mean-Squared Error (MSE), Root Mean-Squared Error (RMSE) and Regression Analysis

2. Predictive Analysis of Water Quality Parameters using Deep Learning- Deep Belief Network, Linear Regression and Multi Layer Perceptron.This system can be implemented on system to continuously monitor the quality of the water. It can be helpful to monitor the quality of water in any uncertain condition.

3. Prediction of with combination of support vector machine and principal component analysis, which is the procedure we have followed.

4. Water Quality Index- It is explained as follows

WQI: It may be defined as a rating, reflecting the composite influence of different water quality parameters on the overall quality of water. The main objective of computing of water quality

index (WQI) is to turn the complex water quality data into information which is easily understandable and usable. This is used for calculation generally.

Computation of WQI: The WQI is computed following the three steps.

First step – Assigning of weight (wi) to the selected water parameters (e.g., pH, TDS, TH, HCO3, Cl, SO4, NO3, Fe, ...)

Second step – Computation of a relative weight (Wi) of the chemical parameter using the following equation:

$Wi = wi / \sum wi \ (i = 1 \text{ to } n)$

Third step - Assigning of a quality rating scale (qi) for each parameter, as below:

$qi = (Ci / Si) \times 100$

$SIi = Wi \times qi$

$WQI = \sum SIi \ 1 - n$

Classification of water :The water may be classified into five types based on computed WQI as given below:

WQI range and water type:

< 50

Excellent water;

50 - 100

Good water;

100- 200

Poor water;

200 – 300

Very poor water;

> 300

The 4th method is the very obvious method for Water Quality Analysis without any machine learning techniques.

After comparing the results, we found that **Deep Learning Approach -** 2nd one of the existing models is proven to be the most efficient one, with the order **2 > 3 > 1** of the above order.

Related Work:

- http://jotterbach.github.io/2016/03/24/Principal_Component_Analysis/
- https://www.utdallas.edu/~herve/abdi-awPCA2010.pdf
- https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7005973
- https://users.ics.aalto.fi/praiko/papers/ilin10a.pdf
- http://pubs.rsc.org/en/content/articlehtml/2014/ay/c3ay41907j
- https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1245090
- https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7019692
- https://www.hindawi.com/journals/mpe/2018/6486345/
- https://pdfs.semanticscholar.org/06a7/5994c62465802ac1b454bce6d0e643ed233c.pdf
- https://pdfs.semanticscholar.org/4eab/310142d1a6971d27b13c621b6adf0cda8750.pdf
- www.iptek.its.ac.id/index.php/jps/article/viewFile/306/471
- https://ac.els-cdn.com/S1878029611002696/1-s2.0-S1878029611002696-main.pdf?_tid=db4d1e7b-f410-4005-807b-69f3915aab84&acdnat=1525027656_f73521f9a3731f93212225b2a90e3e92
- http://psrcentre.org/images/extraimages/37.%200112217.pdf
- https://pdfs.semanticscholar.org/47cd/a5f7694cb542465b34925a5ddb6b2b0fd91a.pdf
- https://www.researchgate.net/topic/Water-Quality-Analysis