Performing Dimensionality Reduction of AMEO 2015 Dataset and Wine Dataset

Raj Choudhary*
**School of Computer Science and Engineering, VIT Chennai, Tamil Nadu, India 600127*
*Email: raj.choudhary2016@vitstudent.ac.in **

***Abstract*: In statistics, machine learning, and information theory, dimensionality reduction or dimension reduction is the process of reducing the number of random variables under consideration by obtaining a set of principal variables. It can be divided into feature selection and feature extraction. Feature selection approaches try to find a subset of the original variables (also called features or attributes). There are three strategies: the filter strategy (e.g. information gain), the wrapper strategy (e.g. search guided by accuracy), and the embedded strategy (features are selected to add or be removed while building the model based on the prediction errors). See also combinatorial optimization problems.**
In some cases, data analysis such as regression or classification can be done in the reduced space more accurately than in the original space.
Feature projection transforms the data in the high-dimensional space to a space of fewer dimensions. The data transformation may be linear, as in principal component analysis (PCA), but many nonlinear dimensionality reduction techniques also exist. For multidimensional data, tensor representation can be used in dimensionality reduction through multilinear subspace learning.
Feature extraction and dimension reduction can be combined in one step using principal component analysis (PCA), linear discriminant analysis (LDA), canonical correlation analysis (CCA), or non-negative matrix factorization (NMF) techniques as a pre-processing step followed by clustering by K-NN on feature vectors in reduced-dimension space. In machine learning this process is also called low-dimensional embedding.

***Index Terms*: Dimensionality Reduction, PCA, Principal Component Analysis, KPCA, Kernel PCA**

1. Introduction

In this paper, the performance of Support Vector Machines (SVM) and Logistic Regression are evaluated on AMEO dataset and wine dataset respectively. These learning models are evaluated based in the accuracy score as well as execution time. Different dimension of the dataset will be used and their performance will be compared with the performance of the complete dataset. For dimensionality reduction – **Principal Component Analysis (PCA)** and **Kernel Principal Component Analysis (KernelPCA)** will be used. The scikit-learn implementation of these algorithms will be used for this experiment.

2. Methodology

Principal component analysis (PCA) is an unsupervised linear transformation technique that is widely used across different fields, most prominently for dimensionality reduction. Other popular applications of PCA include exploratory data analyses and de-noising of signals in stock market trading, and the analysis genome data and gene expression levels in the field of bioinformatics. PCA helps us to identify patterns in data based on the correlation between features. In a nutshell, PCA aims to find the directions of maximum variance in high-dimensional data and projects it onto a new subspace with equal or fewer dimensions that the original one. The orthogonal axes (principal components) of the new subspace can be interpreted as the directions of maximum variance given the constraint that the new feature axes are orthogonal to each other as illustrated in the following figure.
If we use PCA for dimensionality reduction, we construct a $d \times k$ -dimensional transformation matrix W that allows us to map a sample vector x onto a new K - dimensional feature subspace that has fewer dimensions than the original d -dimensional feature space:

$$x=[x1,x2,...,xd],\ x \in Rd$$

$$\downarrow xW,\ W \in Rd \times k$$

$$z=[z1,z2,...,zk],\ z \in Rk$$

As a result of transforming the original d -dimensional data onto this new k -dimensional subspace (typically $k << d$), the first principal component will have the largest possible variance, and all consequent principal components will have the largest possible variance given that they are uncorrelated (orthogonal) to the other principal components.

3. Dataset – AMEO 2015 and Wine

For every engineer, AMEO [1] dataset provides anonymised bio data information along with their respective skill scores and employment outcome information. Specifically, the following information is available for every engineer:

1. Scores on Aspiring Minds' AMCAT - a standardized test of job skills. The test includes cognitive, domain and personality assessments.

2. Personal information like gender and date of birth.
3. Pre-university information like 10th and 12th grade marks, board of education and 12th grade graduation year.
4. University information like GPA, college major, college reputation proxy, graduation year and college location.
5. The following employment outcome information is available for every engineer: First job, annual salary, First job title, First job location, Date of joining and leaving of first job.

AMEO 2015 has gained traction since its public release. Aspiring Minds annually publishes the National Employability Report, a data-driven commentary on graduates and their employability. A recent NER was based on an extension of this dataset.

In Wine[2], the data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines.

4. Analysis and Result

Number of attributes	Time taken for execution	Train Accuracy	Test Accuracy
70	9.1 minutes	57.598%	43.375%
PCA			
2	27.2 seconds	39.744%	19.00%
3	13.3 seconds	38.524%	19.875%
9	1.2 minutes	41.901%	23.125%
KernelPCA kernel = 'rbf			
3	7.3 minutes	28.205%	30.250%
4	5.2 minutes	28.924%	30.500%
9	6.6 minutes	29.143%	30.125%

The above results were found for the GridSearchCV of the SVM classifier with 2 cross validation folds. A total of 144 fits we made. In it the kernel, C and gamma parameters of the SVM were tuned and different hyperparameters were obtained for different dimensions of data. From the above results of the analysis of the AMEO dataset, we can confirm that for the dimension reduction performed, even though the accuracy is not as high as that of the complete dataset, but the execution time is much smaller in case of PCA but almost same in case of KernelPCA. The best accuracy from the reduced dataset was obtained by PCA with n_components = 9 and accuracy score of 41.901% and 23.125% for training set and test set respectively.

Number of attributes	Train Accuracy	Test Accuracy
13	99.194%	100%
PCA		
2	97.581%	92.593%
1	82.285%	87.037%
3	96.774%	92.593%
4	97.581%	92.593%
KernelPCA kernel = 'rbf'		
2	96.774%	98.148%
3	95.968%	98.148%
4	95.968%	98.148%

The above results were found by training Logistic Regression with default parameters on the Wine dataset. We see something unusual as we expect dimensionality reduction to perform better than complete dataset. But here, it is the other way around. Here the complete dataset performs better than the reduced dataset. This shows that dimensionality reduction works best where the number of attributes is large as well as the number of samples in the dataset is large.

5. Conclusion

From the above experiment, we conclude that dimensionality reduction performs best for dataset with high dimensionality as well as high number of samples. Due the presence of large number of features and samples, the reduced attributes are better able to recognize the patterns present in the dataset. Moreover, with more hyperparameter tuning the performance of the model could be further improved.

6. Reference

[1]. V. Aggarwal, S. Srikant, and H. Nisar, "Ameo 2015: A dataset com- prising amcat test scores, biodata details and employment outcomes of job seekers," in AMEO 2015: A Dataset Comprising AMCAT Test Scores, Biodata Details and Employment Outcomes of Job Seekers. ACM, 2016.

[2]. C. Blake: Wine recognition data. https://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.names