An Application of PCA on Uncertainty of Prediction

Santi Phithakkitnukoon

Department of Computer Engineering, Faculty of Engineering
Chiang Mai University, Chiang Mai, Thailand
santi@eng.cmu.ac.th

Abstract. Principal component analysis (PCA) has been widely used in many applications. In this paper, we present the problem of computational complexity in prediction, which increases as more input of predicting event's information is provided. We use the information theory to show that the PCA method can be applied to reduce the computational complexity while maintaining the uncertainty level of the prediction. We show that the percentage increment of uncertainty is upper bounded by the percentage increment of complexity. We believe that the result of this study will be useful for constructing predictive models for various applications, which operate with high dimensionality of data.

Keywords: PCA, uncertainty, prediction.

1 Introduction

Prediction plays an important role in many applications. It is widely applied in various areas such weather, economic, stock, disaster (e.g. earthquake and flooding), network traffic, and call center forecasting. Several techniques and predictive models have been utilized to generate prediction such as regression analysis, Bayesian networks, Markov model, and neural network. These techniques have different computational costs associated with them. As more predicting event's information is provided, the computational cost (and complexity in some predictive models) increases. Reducing the dimensionality of the input to the predictive model in order to lower the computational cost may also increase the uncertainty of the prediction. To avoid degrading the uncertainty of the predictive model while reducing its computational cost, in this paper we present an application of the PCA method as a solution to our problem.

The rest of this paper is structured as follows. Section II carries out the main contribution of this paper. The paper is concluded in section III with a summary and an outlook on applying our finding to the future work.

2 Applying PCA to Complexity in Prediction

Prediction is a statement about the future observation. The actual future event does not always occur as its prediction. Hence there is an uncertainty associated with a prediction. The uncertainty of prediction can be measured using information entropy or Shannon's entropy, which is a measure of uncertainty of a random variable [1] defined by (1).

$$H(X) = -\sum_{x} p(x) \log_2 p(x), \tag{1}$$

where X is a discrete random variable, $x \in X$, and the probability mass function $p(x) = Pr\{X=x\}$.

Cover, T. M., and J. A. Thomas [2] shows that the uncertainty of prediction of event X given the information Y is less than uncertainty of prediction of event X without given information about event X,

$$H(X \mid Y) < H(X). \tag{2}$$

We can extend (2) to a scenario where there is more one known information about predicting event X. Since conditioning reduces entropy, (2) still holds for multiple given information,

$$H(X \mid Y_1, Y_2, Y_3, ...) < H(X). \tag{3}$$

Inequation (3) tells us that theoretically the more information about event X, the less uncertainty about predicting event X. This implies that one may collect infinite information about event X to make the optimal prediction in the sense of having the least uncertainty.

In practical prediction problem, an infinite dimensional dataset is not available. However, there is a high dimensional dataset in which taking entire data will result a high computational complexity in prediction. It is desired to lower the dimensions of dataset while retaining as much as possible of the characteristics of the dataset in order to reduce the computational complexity in prediction. This can be achieved by applying the Principal Component Analysis (PCA) method.

The PCA is a technique of multivariate analysis. It was first introduced by Pearson [3] in 1901 and developed by Hotelling [4] in 1933. The idea of PCA is to reduce the dimensionality of a dataset while retaining as much as possible of the variation present in the dataset. The PCA method composes transformation matrix from the set of input vectors containing correlated components to another set of vectors containing orthogonal and uncorrelated components. PCA reduces dimension of the dataset by keeping the most relevant information and discarding the statistically less relevant information from the multidimensional dataset. The PCA transformation is based on the autocorrelation matrix which is given by (4).

$$R_{xx} = \frac{1}{n} \sum_{j=1}^{n} \left(X_j X_j^T \right), \tag{4}$$

where n is the number of input vectors, X_j is the j^{th} vector.

Principal components are obtained by arranging eigenvalues corresponding to eigenvectors of matrix R_{xx} in order. The first principal component contains the largest percentage of the variation in the original dataset. The second principal component contains the second largest percentage of the variation in the original dataset and so on. The main idea of PCA is to reduce dimensionality of dataset to first m principal components. I.T. Jollife [5] suggests three types of rule for choosing m; cumulative percentage of total variation, size of variances of principal components, and the scree graph and the log-eigenvalue diagram. Choosing the number of principal components is not in the scope of this paper. Interested readers may find more information from [5].

By choosing the number of principal components m, one can construct a matrix P which is given by (5).

$$P = \begin{bmatrix} P_1 & P_2 & \dots & P_m \end{bmatrix}, \tag{5}$$

where P_j is the j^{th} principal component.

The PCA can also be visualized as a simple transformation from one domain to another by projecting original data points onto the new principal component axis where the first principal component contains the largest percentage of the variation in the original data points and so on. Figure 1 shows the original data points from X-Y axis to be projecting onto the principal component axis (PC_1 and PC_2).

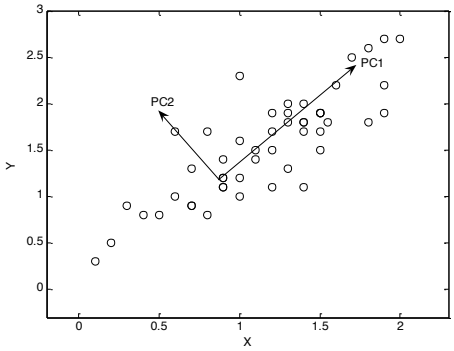

Fig. 1. Example of projection of the data points onto the principal component axis.

Since the first principal component contains the largest percentage of the variation in the original dataset, therefore range of the percentage of the variation in the original dataset that the first principal component contains ($\%Var(P_1)$) lies between 100/n% and 100% where n is the dimension of the original dataset.

$$\frac{1}{n} \le \%Var(P_1) \le 1. \tag{6}$$

The percentage of the variation in the original dataset contained in the first principal component has the maximum value of 100% when the original data points form a straight line. On the other hand, the percentage of the variation in the original

dataset contained in the first principal component has the minimum value of $100/n\%$ when the original data points form a perfect sphere cloud where the variation of the original data contained in the first principal component is equal to the variation of the original data contained other principal components. In general, having all features (dimension) of the data equally represent the overall data is very rare.

Given a high n-dimensional dataset to predict event X with the computational complexity of $O(n)$ and uncertainty of $H(X|Y_1,Y_2,...,Y_n)$, using the PCA method reduces the dimension of the data to n-k while increases the uncertainty to $H(X|P_1,P_2,...,P_{n-k})$, where P_j is the j^{th} principal component. Thus, the complexity is reduced by $(n-k)/n$ % while the uncertainty is increased at most by $(n-k)/n$ % since $\%Var(P_1)$ has lower bound of $1/n$ and its equality holds if and only of all principal components contain the same variation of the original data (Eq. 6). Since the dimension of the data is reduced to n-k which means n-k principal components are retained, thus

$$\sum_{i=1}^{n-k} \%Var(P_i) \geq \frac{n-k}{n}. \quad \textbf{(7)}$$

This means that by applying PCA, percentage decrement of computational complexity $(\%\downarrow C)$ is greater than and equal to the percentage increment of uncertainty $(\%\uparrow H)$,

$$\%\downarrow C \geq \%\uparrow H, \quad \textbf{(8)}$$

which implies that using PCA can reduce the computational complexity of the predictive model while maintaining the uncertainty level of the predicting event as much as possible with its percentage incretion never be more than percentage decrement of complexity.

3 Conclusion

In this paper, we present an application of the PCA in the prediction. We show that the computational complexity of the prediction increases as the dimensionality of the dataset grows and in order to reduce computational complexity in prediction while maintaining the level of the uncertainty of prediction, the PCA method can be applied.

References

1. Shannon, C. E.: A mathematical theory of communication. Bell System Technical Journal, vol. 27, pp. 379-423 and 623-656, July and October (1948)
2. Cover T. M., Thomas, J. A.: Elements of Information Theory, Wiley (1991)
3. Pearson, K.: On lines and planes of closest fit to systems of points in space. Phil. Mag. (6), 2, 559-572 (1901)
4. Hotelling, H.: Analysis of a complex of statistical variables into principal components. J. Educ. Psychol., 24, 417-441, 498-520 (1933)
5. Jolliffe, I. T.: Principal Component Analysis, Springer Series in Statistics (2002)