Abstract

Principal component analysis (PCA) guided clustering approach is widely used in high dimensional data to improve the efficiency of K-means cluster solutions. Typically, Pearson correlation is used in PCA to provide an eigen-analysis to obtain the associated components that account for most of the variations in the data. However, PCA based Pearson correlation can be sensitive on non-Gaussian distributed data, which involve skewed observations such as outlying values. Thus, applying PCA based Pearson correlation on such data could affect cluster partitions and generate extremely imbalanced clusters in a high dimensional space. In this study, Tukey's biweight correlation based on M-estimate approach in PCA is used as an alternative to Pearson correlation. This approach is more resistant to outlying values as it examines each observation and down weight those that lie far from the center of the data. In particular two major features are highlighted: (1) fewer components are retained and imbalanced clusters at the recommended cumulative percentage of variation threshold is avoided; (2) the cluster quality with respect to external, internal and relative criteria as shown in Rand, Silhouette and Davies-Bouldin indices, outperform that of the clusters from PCA based Pearson correlation.
References

Keywords
Tukey's biweight  K-means  Principal Component Analysis.

Index Terms
Computer Science  Data Mining