Independent Component Analysis for Clustering Multivariate Time Series Data

Edmond H.C. Wu and Philip L.H. Yu

Department of Statistics & Actuarial Science, The University of Hong Kong,
Pokfulam Road, Hong Kong
hcwu@hkusua.hku.hk, plhyu@hku.hk

Abstract. Independent Component Analysis (ICA) is a useful statistical method for separating mixed data sources into statistically independent patterns. In this paper, we apply ICA to transform multivariate time series data into independent components (ICs), and then propose a clustering algorithm called ICACLUS to group underlying data series according to the ICs found. This clustering algorithm can be used to identify stocks with similar stock price movement. The experiments show that this method is effective and efficient, which also outperforms other comparable clustering methods, such as K-means.

Keywords: Clustering, Independent component analysis, Statistics, Time series.

1 Introduction

The goal of clustering is to find an intrinsic structure in a set of unlabeled data. As a common data mining technique, clustering is useful in finding suitable groupings and representatives for homogeneous groups, and in detecting unusual data objects. A cluster is therefore a collection of objects which are 'similar' among them and are 'dissimilar' to the data objects belonging to other clusters. For instance, we could be interested in finding groups of stocks with similar return performance from a large database of historical stock prices. Also, we need to consider the scalability and robustness of the clustering algorithms that can deal with high-dimensional data with noises and outliers.

Many clustering algorithms have been developed in the literature. Clustering algorithms can be classified as partitioning methods, hierarchical methods, density-based methods, grid-based methods etc. For partitioning methods, MacQueen [6] first proposed the well-known K-means clustering algorithm. The K-means algorithm first randomly selects k of the objects as cluster centers. The remaining objects will be assigned to a cluster to which it is the most similar by computing the distance between the object and the cluster mean. Then, it recomputes the new mean for each cluster and reassigns the objects to the new clusters. This process will iterate until certain criterion function converges. However, a disadvantage of K-means is that the clustering results will be influenced by noises or outliers in the data.

X. Li, S. Wang, and Z.Y. Dong (Eds.): ADMA 2005, LNAI 3584, pp. 474–482, 2005.

Many statistical techniques have been used in various data mining algorithms. A recently developed linear transformation method is the Independent Component Analysis (ICA) [3], in which the desired representation is the one that minimizes the statistical dependence of the components of the representation. Such a representation can capture the essential structure of the data in many potential applications. In this paper, we investigate ICA for clustering applications.

The rest of this paper is organized as follows: In Section 2, we will introduce the independent component analysis technique. Then, in Section 3, we will propose a clustering model for time series data by using independent component analysis. In Section 4, we give some experimental results on testing the effectiveness and scalability of this clustering model with artificial time series data and real financial datasets. Finally, we give some conclusions in Section 5.

2 Independent Component Analysis

ICA [1, 3, 5] is a statistical method aiming to express the observed data in terms of a linear combination of underlying latent variables. The latent variables are assumed to be non-Gaussian and mutually independent. The task is to identify both the latent variables and the mixing process. A typical ICA model is:

$$X = AS \tag{1}$$

where $X = (x_1, ..., x_m)$ is the vector of observed variables, $S = (s_1, ..., s_m)$ is the vector of statistically independent latent variables called the independent components, and A is an unknown constant mixing matrix. The independent components S in the ICA model (1) are found by searching for a matrix W such that $S = WX$ up to some indeterminacies.

The FastICA algorithm [2, 4] is a computationally efficient and robust fixed-point type algorithm for independent component analysis and blind source separation. The iterative fixed-point algorithm for finding one unit is:

$$\tilde{w}_{n+1} = E\{x(w_n x) * g(|w_n x|^2)\} - E\{g(|w_n x|^2) + |w_n x|^2 g'(|w_n x|^2)\}w_n \tag{2}$$

where $w_{n+1} = \frac{\tilde{w}_{n+1}}{\|\tilde{w}_{n+1}\|}$. Getting the estimate of w, we can obtain an IC by $s = wx$.

The above algorithm can be extended to the estimation of the whole ICA transformation $S = WX$. To prevent converging to the same ICs, the outputs $w_1 x, ..., w_n x$ are decorrelated after every iteration. When we have estimated n independent components, or n vectors $w_1, ..., w_n$, we run the one-unit fixed-point algorithm for w_{n+1}, and after every iteration step subtract from w_{n+1} the projections of the previously estimated n vectors, and then renormalize w_{n+1}:

$$\tilde{w}_{n+1} = \tilde{w}_{n+1} - \sum_{j=1}^{n} w_j w_j' \tilde{w}_{n+1}. \tag{3}$$

where $w_{n+1} = \frac{\tilde{w}_{n+1}}{\|\tilde{w}_{n+1}\|}$. The above decorrelation scheme is suitable for deflationary separation of the ICs. Using FastICA, we can estimate A and S from observations X, where $A = W^{-1}$. The vectors $w_1, ..., w_n$ compose W, i.e., $W = [w_1; ...; w_n]$.