Data Mining with Independent Component Analysis *

Fasong Wang
China University of Geosciences
Wuhan, China
China Electronics Technology Group Corporation, the 27th Institute Zhengzhou, China
E-mail: fasongwang@126.com

Hongwei Li
School of Mathematics and Physics,
China University of Geosciences
Wuhan, China
E-mail: hwli@cug.edu.cn

Rui Li
School of Mathematics and Physics,
Henan University of Technology
Zhengzhou, China
E-mail: lyric@126.com

Abstract – Independent Component Analysis(ICA)/Blind Source Separation(BSS) has received many attentions in neural network and signal processing area recent years. In this paper, we consider the data mining problem with ICA. The data model of under-complete ICA in data mining is given and then gives the most popular ICA algorithm-Natural Gradient Algorithm (NGA). Several applications of data mining with ICA is considered, such as latent variable decompositions, multivariate time series analysis and prediction, text document data analysis, extracting hidden signals in satellite images, weather data mining and so on. All these discussions suggest the huge potential outlook of data mining using ICA. The other contribution of this paper is it contains several literature surveys on various aspects of data mining using ICA.

Index Terms - Independent Component Analysis(ICA); Blind Source Separation(BSS); Data Mining; Dimensionality Reduction; Neural Network.

I. INTRODUCTION

Independent Component Analysis(ICA) was originally developed for signal processing applications including communications, medical signal processing, speech signal processing and so on [1-3]. Recently it has been found out that ICA is a powerful tool for the problem of finding latent structure in high dimensional data which is considered in data mining area usually. It is assumed that the observed data are generated by unknown latent variables and their interactions. The task is to find these latent variables and the way they interact, given the observed data only. It is assumed that the latent variables do not depend on each other but act independently [4].

Data mining is a topic used for variety of computational methods and techniques for analyzing large data sets [5]. The aim in data mining is to describe the data either in a global or a local level. Global descriptions include clustering, joint probability density estimation, or visualization of the data; local descriptions might be repeating or exceptional patterns in the data, or statistical dependencies between the variables. Although data mining is closely related to traditional statistical data analysis, it has distinguishing characteristics: the data are not originally aimed for a particular study and so the analyst cannot affect the process of data collection; the data set is often so large that its storage and retrieval must be carefully
designed [6].

As popular methods for finding latent structure in high dimensional data, ICA is a powerful tool for data mining, which expresses a set of multidimensional observations as a combination of unknown latent variables that are statistically independent of each other. Here the term latent means hidden, unknown or unobserved; the term structure refers to some regularities in the data; high dimensional may be tens or tens of thousands of dimensions, depending on the situation; and data is any information that can be transformed into numerical values, most often represented as a matrix of observations where each dimension corresponds to a variable whose value we can somehow measure. The aims in this paper are to find What is there in the data and how they interact?, to form a simple representation of a large data set that is understandable to observer with ICA.

For using ICA, we will assume that the observed data are generated by interactions between latent variables. The objective is to find out what these latent variables are and how they interact—this is the key to understanding what the data are about. The latent variables will be called components, sources or topics. Throughout this paper, we will assume that there are no inherent dependencies among the latent variables.

Several applications of data mining by ICA are proposed. First, the ICA based data mining model—under-complete ICA model is given in section 2; In section 3, we propose that ICA can be used to the latent variable decompositions which is a conventional problem in various applications. Compared with other latent variable decompositions method, such as principle component analysis(PCA), ICA can give a more meaningful representation structure; In section 4, we see that ICA can be used in multivariate time series analysis and prediction especially in financial time series; Next, section 5 presents that ICA can be used in text document data as well after giving a suitable numerical form; In section 6, some other applications of ICA are given, such as extracting hidden signals in satellite images produced with many sensors and weather data mining. Finally, we give a brief summary and concluding remarks in section 7.

II. BASIC UNDER-COMPLETE ICA MODEL AND PROBLEM

* This work is partially supported by National Natural Science Foundation of China (Grant No. 60472062) and Natural Science Foundation of Hubei Province, China (Grant No.2004ABA038).
In this section, we formulate the basic model of ICA used in data mining[1-3]. $N$ source signals are represented as a whole by a vector $s(t) = (s_1(t), \cdots, s_N(t))^T$. It is assumed that the source signals are zero means and they are independent each other. Observed signals are represented by $x(t) = (x_1(t), \cdots, x_M(t))^T$ and they correspond to the recorded signals as sensors or microphones. In instantaneous mixtures, we assume that the unknown source signals and the observed noisy mixtures are related by

$$x(t) = As(t) + n(t)$$  \hfill (1)

where $A \in \mathbb{R}^{M \times N}$, called mixing matrix, is an $M \times N$ full rank real number matrix. We speak of complete, over-complete or under-complete ICA if $M = N$ , $M < N$ or $M > N$ respectively [7]. Because in data mining applications, we can get a big dataset, so we consider the complete or under-complete ICA model that is $M \geq N$. $n(t)$ is the additive noise vector and $t$ is the sample index, in this paper we consider $n(t) = 0$, then (1) becomes $x(t) = As(t)$.

The propose of ICA of the under-complete random observed signals $x(t)$ is to obtain the estimated signal $y(t)$ by estimating a $N \times M$ full rank separating matrix $W$, such that the estimated sources are the components of the output signal vector that are as independent as possible, as measured by an information-theoretic cost function such as K-L divergence or other criteria like sparseness, smoothness or linear predictability. In other words, it is to adapt the weights $w_{ij}$ of the matrix $W$ of the system $y(t) = Wx(t)$, which often referred to as a signal-layer feed-forward neural network, to combine the observations $x_i(t)$ to generate estimates of the source signals, defined as

$$\hat{s}(t) = y(t) = Wx(t) = WAs(t)$$  \hfill (2)

The optimal weights correspond to the statistical independence of the output signals $y_i(t) = w_i^T x(t)$. After a possible permutation and scale change, we get the performance matrix $P$ and $P = WA$.

The model to this overall mixing and separating process is shown by figure 1.

ICA algorithm is mainly performed using the information on signal statistics. Since Comon [8] gave a good insight to ICA problem from the statistical point of view, there has been emerged a set of efficient ICA algorithms, such as the Infomax algorithm [9], the FastICA algorithm [10], the Natural Gradient algorithm[7,11], the Flexible Score Function algorithm [12] and so on. Many algorithms are based on the NGA framework, such as Infomax algorithm. The NGA update equation is given by [11]

$$W(k+1) = W(k) + \eta(k) \left[ I - \phi(y(k)) y'(k) \right] W(k)$$  \hfill (3)

where $\phi(y(k))$ is an odd non-linear function of the output called the score function, whose choice dependents on the statistics of the source signals and $\eta(k)$ is a positive adaptive parameter.

Another efficient deflation ICA algorithm is the fixed point algorithm. The procedure of it can be obtained in [10].

III. ICA BASED DIMENSIONALITY REDUCTION VS PCA

ICA is a method for finding underlying factors or components from multidimensional statistical data. There are many latent variable decompositions method, such as PCA, singular value decomposition(SVD), factor analysis, projection pursuit and so on. What distinguishes between ICA from this methods is that it looks for components that are both statistically independent and non-Gaussian [13,14].

In PCA or factor analysis, an observed vector $x(t)$ is first centered by removing its mean(in practice, the mean is estimated as the average value of the vector in a sample). Then the vector is transformed by a linear transformation into a new vector, possibly of lower dimension, whose elements are uncorrelated with each other. The linear transformation is found by computing the eigenvalue decomposition of the covariance matrix, which for zero-mean vectors is the correlation matrix $E[x(t)(x(t))^T]$ and the eigenvectors of it form a new coordinate system in which the data are presented. As a result, the number of components $y_i(t)$ will be quite small, maybe only 1 or 2, but these components contain most information which may provide an insight into the structure of the data in the meaning of second order statistics. The basic PCA network can be described by

$$\begin{align*}
y_i(t) &= \sum_{j=1}^{N} w_{ji} x_j(t) \\
x_i'(t) &= x_i(t) - \sum_{j=1}^{N} w_{ji} y_j(t) \\
\Delta w_{ij} &= \eta x_i'(t) y_j(t) \\
\forall i,j &= 1,2,\cdots,N
\end{align*}$$  \hfill (4)

But in many applications, uncorrelatedness is not enough, we must find the independent components(ICS). Here, the independence is not correspond to the independence in factor analysis. Factor analysis originally developed in social sciences, which is often claimed that the factors are independent, but this is only partly true, because factor analysis assumes that the data has a Gaussian distribution. If the data has a Gaussian distribution, it is easy to find the ICS, for Gaussian data, uncorrelated components are equivalent to independent.

On the other hand, ICA tries to find statistical
independent sources by additionally minimizing higher order statistics between various components. Most popular independence measure is the contrast functions [8], there are many statistics can be used as contrast function, such as Kullback-Leibler(K-L) divergence and mutual information.

Another insight can be gained by interpreting the directions defined by the transformation. The first principle component(PC) finds the direction that captures the maximum variance possible, the second PC finds the direction of maximum variance in the remaining orthogonal subspace, and so forth. ICs, in contrast, do not find the directions of maximum variances but rather the most interesting direction. This can be operationalized by maximizing higher order statistics along these directions, such as maximizing the fourth order self-cumulants \((i = j = l = m)\). This is equivalent to minimizing the cross- cumulants(where the indices are not identical).

So in the applications, when we intend to find the independent factors among the huge data set, ICA is a ideal method more than PCA or factor analysis(in applications, most components are not have Gaussian distributions).

IV. ICA IN MULTIVARIATE TIME SERIES

Finding the latent functional factors is the main task in many time series analysis problem. Traditional method such as PCA is a powerful tool. But in some applications, such as financial time series analysis, there may be some underlying factors like seasonal variations or economic events that affect a number of simultaneous time series but can be assumed to be quit independent [14-19,30].

ICA provides a mechanism of decomposing a given signal into statistically ICs. The goal of this section is to explore whether ICA can give some indication of the underlying structure of the financial time series. Such ICs could include news(government intervention, natural or manmade disasters, political upheaval), response to very large trades and of course, unexplained noise. Ultimately, we hope that this might yield new ways of analyzing and forecasting financial time series, contributing to a better understanding of financial markets.

First we show that multivariate time series forecasting can be done by using ICA. The problem in time series prediction is to learn the mapping from the observations. In recent years this has been proposed done using various neural network models [15-17,19,30].

Here we consider a specific problem that may occur in some applications of multivariate time series analysis. Given a set of correlated signals, e.g. time series of geophysical observations or economical indexes, we find that in one series (or a few) some parts are missing. We assume that the missing sections extend over some tens of samples. The problem of recovering the missing parts may be defined as a prediction and interpolation task, where we want to use past samples to predict future values, but we also want to take advantage of having observations of the other signals in the set during the time of the missing section [17]. This problem can be done by combining the properties of ICA with those of the dynamical functional artificial neural network(D-FANN). The approach assumes that the observed signals are generated from independent sources through a linear mixing. Let \(x(t)\) and \(s(t)\) be the observed signals and the original source signals, respectively, and let \(A\) denote the mixing matrix. Then the model assumes that \(x(t)\) is related to \(s(t)\) as

\[
x(t) = As(t)
\]  

Let \(y(t) = Wx(t)\) be an estimate of the sources \(s(t)\) obtained by first estimating de-mixing matrix \(W\), and then calculating \(y(t)\) as \(\hat{s}(t) = y(t) = Wx(t) = WAs(t)\). The aim is to find the signals \(y(t)\) that, generally have a much simpler structure than the observed signals \(x(t)\), makes them easier to predict.

As shown in [17], We use the Discrete Cosine Transform(DCT)-based D-FANN predictor to perform nonlinear one-step predictions on each of the individual component signals. By doing the predictions in the ICA domain, i.e. on \(y(t)\), while calculating prediction errors for updating the model in the observation domain, i.e. using \(x(t)\), we are effectively utilizing the dependencies between the observed signals in the prediction process.

Then we describe how to combine the D-FANN predictor and the ICA transform to recover a missing section in one of a set of multivariate time series. Up to time \(t \leq T\), the complete set \(x(t) = (x_1(t), \ldots, x_N(t))\) is given , and for \(t > T\) time series \(x_i(t)\) is the missing samples. By using ICA algorithm, we estimate \(A\) and \(s(t)\). The estimated source signals are now given by \(\hat{s}(t) = y(t) = Wx(t) = A\hat{s}(t)\), \# represents the Moore-Penrose pseudoinverse. Then, we will refer to the observation space as the \(x\)-space and the ICA space (space of separated sources) as the \(y\)-space. The procedure is to use \(W\) to transform \(x(t)\) to \(y\)-space and predict in \(y\)-space. Then the prediction results are transformed back to the \(x\)-space. The prediction error is computed in \(x\)-space, and re-transformed to the \(y\)-space. Based on these errors in \(x\)-space, we update the predictor parameters. Specifically, let

\[
\hat{y}(t+1) = \{\hat{y}_1(t+1), \ldots, \hat{y}_N(t+1)\}\n\]  

denote the one-step predictions \(\hat{y}(t+1)\). Then we obtain the predictions of \(x(t+1)\) as

\[
\hat{s}(t+1) = A\hat{y}(t+1)
\]

Let \(e_x(t+1)\) denote the prediction error in \(x\)-space, and
data. Interestingness is measured as a low Kolmogorov coding to find interesting structure in multidimensional time series algorithm proposed in [26]. The idea in complexity pursuit is topographic visualization of time-varying data. Hidden Markov Model (HMM)-type algorithm for the dynamically evolving topics. [24] have recently presented a [22,23] give the ICA algorithm for the identification of the underlying characteristics–here the topics–of the discussion methods of time series processing may be used to extract the interest changes according to participants’ contributions. The chat rooms daily news topics are discussed and the topic of discussion is found in the internet relay chat rooms. In these textual data has arisen. An example of a dynamically evolving time.

A common practice in text mining is to compute the SVD of the data matrix and project the data into the subspace spanned by the left singular vectors corresponding to the largest singular values. Thus the observed documents are represented as linear combinations of some orthogonal features, called latent semantic factors(LSA). LSA is said to tackle the problem of synonymy and partially also polysemy, and take advantage of the implicit structure in how terms are associated with documents. LSA uses only second-order statistics of the data, so a natural step forward is to apply more powerful methods such as ICA.

First approaches of using ICA in the context of text data were presented [20]. Then in [21] some other ICA based text data mining are given. In these approaches, the textual data is not a dynamic time series but rather an instantaneous mixture of independent topics. The underlying assumption which they also adopt is that the textual data consists of some more or less independent topics. In the text retrieval parlance, a topic is a probability distribution on the universe of terms; it is typically concentrated on terms that might be used when discussing a particular subject. Here the word ‘topic’ also refers to a hidden, more or less independent random variable with time structure. Thus we can analyze the ‘independent components’ of text both by the terms they concentrate on, and by their activity in time.

Recently the issue of analyzing dynamically evolving textual data has arisen. An example of a dynamically evolving discussion is found in the internet relay chat rooms. In these chat rooms daily news topics are discussed and the topic of interest changes according to participants’ contributions. The online text stream can thus be seen as a time series, and methods of time series processing may be used to extract the underlying characteristics–here the topics–of the discussion [22,23] give the ICA algorithm for the identification of the dynamically evolving topics. [24] have recently presented a Hidden Markov Model (HMM)-type algorithm for the topographic visualization of time-varying data.

In [25,31], The algorithm is based on complexity pursuit algorithm proposed in [26]. The idea in complexity pursuit is to find interesting structure in multidimensional time series data. Interestingness is measured as a low Kolmogorov coding complexity of a projection of the data. Intuitively stated, projections with a short coding length are typically structured in some way, that is, they are far from random noise and/or Gaussianity. The complexity pursuit method is quite similar to ICA except that it exploits the temporal dependencies in the data in addition to higher order statistics. No special emphasis is put on the statistical independence of the estimated latent sources, though.

Let us briefly describe its main characteristics of the algorithm. The data are first whitened by PCA. At each iteration step, a first order autoregressive(AR) model is estimated for each latent variable \( s_j \). Then an approximation of the Kolmogorov complexity of the residual of the AR model is minimized by gradient descent. The estimated projection directions \( w_j \) are decorrelated after every iteration step.

VI. OTHER DATA MINING USING ICA

Recently there has been increased interest in the use of the ICA for image analysis. ICA can be considered as one approach to component analysis. Among other approaches, the traditional PCA is most popular. The component analysis that extracts the most important components of the data is useful for data mining in remote sensing which normally involves a very large amount of data. While PCA method attempts to decorrelate the components in a vector, ICA methods are to make the components as statistically independent as possible. [27,28,32] have developed several ICA algorithms which can be implemented efficiently by a neural network. As such it is a very useful tool for data mining in remote sensing. The use of the algorithm especially in hyperspectral image analysis is presented in the paper [27].

Statistical approaches to weather and climate prediction have a long and distinguished history that predates modelling based on physics and dynamics. This trend continues today with new approaches based on machine learning algorithms. The central problem in weather and climate modelling is to predict the future states of the atmospheric system. It is therefore possible to view the weather variables as sources of spatio-temporal signals. The information from these spatio-temporal signals can be extracted using data mining techniques. The variation in the weather variables can be viewed as a mixture of several independently occurring spatio-temporal signals with different strengths. Therefore, if the assumption of independent stable activity in the weather variables holds true then it is also possible to extract them using the same technique of ICA.

One basic assumption of the approach in [29] is that we view the weather phenomenon as a mixture of a certain number of signals with independent stable activity. The weather changes due to the changes in the mixing patterns of these stable activities over time. For linear mixtures, the change in the mixing coefficients gives rise to the changing nature of the global weather. The purpose of [29] is to investigate if there exist any such set of spatio-temporal stable patterns such that the variation of the mixture gives rise to the observed weather or climate phenomena. The conjecture is
that there exist independent stable spatio-temporal activities, the mixture of which give rise to the weather variables; and these stable activities can be extracted by ICA of the data arising from the weather and climate patterns, viewing them as spatio-temporal signals. If the conjecture about the existence of stable spatio-temporal activity in the weather is true, then the mixing coefficients will vary in accordance with the changes in the weather variables.

VII. CONCLUSIONS

In this paper, we introduce the basic instantaneous under-complete ICA model and the most popular NGA ICA algorithm briefly. Several data mining applications are given by using ICA. First we propose that ICA can be used to the latent variable decompositions which is a conventional problem in various applications. Compared with other latent variable decompositions method, such as PCA, ICA can give a more meaningful representation structure; we also show that ICA can used in multivariate time series analysis and prediction especially in financial time series; Next, we presents that ICA can be used in text document data as well after giving a suitable numerical form; At last, some other data mining applications of ICA are given, such as extracting hidden signals in satellite images produced with many sensors and weather data mining. All these aspect can give us a wide view of data mining by using ICA.

REFERENCES