NONPARAMETRIC CLUSTERING ALGORITHM WITH AN APPLICATION IN FINANCIAL TIME-SERIES DATA

SHU-MAN A. CHANG
Center of General Education, Chang Gung University, Tao-yuan, Taiwan

TSAIR-CHUAN LIN
Department of Statistics, National Taipei University, Taipei, Taiwan

A new methodology for clustering nonstationary time series with nonparametric regression model is proposed. The new methodology first use the projection pursuit regression models for formulating each time series and then use the measure of cross-validation to calculating the similarity of fitted models and carry out the cluster analysis. Application of the proposed method is adopted in the collection of average personal income of 25 states in the US. Comparison with existing clustering method show several advantages of the proposed.

1. Introduction

The study on time series clustering analysis is an important subject paid to growing attention in data mining, and its theory and application have been greatly put into study on fields such as biology, medicine, economy, finance, machine learning, signal analysis, gene recognition, and others. Cluster analysis is a general designation of data classification, the researchers use it to simplify and group the data. In general, it aims at identifying the similar series based on certain characteristics to make the elements within a cluster bear the high similarity to one another but is very dissimilar to the elements in the other clusters. The time series clustering (Xiong, 2004) can be divided into two major classes of distance based methods and model based methods. The non-model based methods assume that each series can be represented as a point in certain multi-dimensional space of fixed dimensionality, and then base on the similarity or distance measurement to group the data set. For example, the well-known K-mean method assumes the data point with its each dimension being independent and then uses the Euclidean distance for clustering. Unfortunately, there is no natural distance function among time series data. Moreover, the problems of

unequal length, time delay, overall slower rate, premature cutoff, and correlation of each dimension in a time series will be overly emphasized by the distance measure.

Generally the model based time series clustering takes a probability model or statistical model as assumption for describing the mechanism to generate the data. Under assumptions about the joint probability density function of series data and prior information, the probability model clustering can derive the maximization of posterior probability of cluster model as the foundation of clustering. For examples, the Markov methods and hidden Markov methods (HMM) are two sets of well known probability model based clustering methods. The study on statistical model based clustering of time series model is usually to assume the realizations as the linear parameterized stationary autoregressive (AR) or autoregressive and moving-average (ARMA) model, and then to cluster the time courses based on similarity measure of the feature function based on fitted models. The definition of feature function between stochastic models is an important study subject, such as weighted autoregressive coefficient distance (WAR), autocorrelation functions (ACF), principle component vector, discrete Fourier transform (DFT), discrete wavelet transforms (DWT) and series spectral transform. Since the statistical model based clustering methods take into the stochastic properties of time series, the interpretation of their clustering result is better and more reasonable than that of probability models based methods.

In recent years, nonparametric time-series analysis has become a powerful statistical tool for exploring the underlying structure in a dataset and gained attention due to the limitations of autoregressive and moving-average models in describing various natural phenomena such as asymmetrical limit cycles, time irreversibility, amplitude-dependent frequency, and chaos. Let $\{Y_t\}$ be a time series, the most general nonparametric p-th order autoregressive model can be defined as

$$Y_t = f(Y_{t-1}, Y_{t-2}, ..., Y_{t-p}) + \varepsilon_t, \tag{1}$$

where the random error ε_t is i.i.d. with a zero mean and variance σ^2. In a linear regression setup, one assumes the response surface f can be expressed as a linear combination of predictor variables. If inadequate, model scope can be extended by adding terms, such product or high-order predictors, to the model. Parametric regression models have advantages such as ease in computations, interpretations and forecasting. However, guessing which terms should be included in the function when many predictors exist in the model, and which is the most appropriate functional form is difficult when just looking at the data.

Multivariate nonparametric smoothers only require a few assumptions; however, they frequently encounter the problem "curse of dimensionality," which is a neighborhood with a fixed number of points that become less local as the dimensions increase (Hastie and Tibshirani, 1990). Several nonparametric regression approaches have been developed in response to this dimensionality problem. Generalized additive models (GAMs),

$$Y_t = f_1(Y_{t-1}) + \cdots + f_p(Y_{t-p}) + \varepsilon_t, \tag{2}$$

or projection pursuit regression (PPR) (Friedman and Switzer, 1981) overcomes this problem by using analogs of the Taylor explanation to approach a complex response surface. The mean function in GAMs is the summation of several univariate smooth and unknown functions $f_j(\cdot)'s$. Once the additive model is fitted to data, plots of smooth functions can be examined to assess the contributions of predictors in predicting a response. Although it can be applied to non-linear, non-Gaussian distributed or nonstationary cases, GAMs cannot deal with interactions between predictors. In such models, the projections are done onto individual predictors rather than onto a projection vector, which is the linear sum of the predictors, as in PPR. These projection vectors, instead of individual predictors, allow PPR to deal with interactions, which is the main property of PPR. Thus, this study applies the PPR model to overcome these problems, and estimate nonparametric models and clustering.

The remainder of this paper is organized as follows. The existing time-series clustering methods are introduced in Section 2. Section 3 describes the PPR clustering method and its basic concepts. This clustering method is applied to Average personal income in 1929–1999 in 25 US states, and compared with other clustering analysis methods in Section 4.

2. Literature Review

2.1. *Statistical model based clustering*

The statistical-model-based clustering methods assume the series in the same cluster are from the same time-series model, such as common AR models, ARMA models or autoregressive integrated moving average (ARIMA) models. Ramoni *et al*. (2002) considered gene expression as an AR model, and applied the Bayesian method to perform clustering with an agglomerative algorithm. Xiong (2004) assumed data were mixtures of ARMA, and applied the Expectation Maximization (EM) algorithm (Dempster *et al*, 1976) to estimate

model parameters and calculate maximum posterior probability for a clustering model. Kalpakis *et al*. (2001) adopted the ARIMA time-series model, and applied the liner predictive coding cepstrum (LPC) (Furui, 1989) coefficient to extract coefficients of the time-series model. Trend and seasonality components were first removed from data, since stationary data can be well fitted with an autoregressive model of a certain order; thus, the LPC can be acquired through estimated autoregressive coefficients. Sequentially, the cluster result can be obtained by applying the partitioning around method (PAM) to the LPC.

2.2. *Nonparametric model clustering*

Nonparametric regression allows one to estimate nonlinear fits between continuous variables with few assumptions about the functional space. This feature results in wide-ranging techniques that can be employed to numerous practical situations in diverse fields. In some current nonparametric clustering studies, the "time" factor is used as a predictive variable of a model; that is,

$$Y_t = f(t) + \varepsilon_t \tag{3}$$

is the generation mechanism producing the series in each cluster. Luan and Li (2003) applied this model to cluster time-series data, gene expression was analyzed by a nonparametric mixed-effects model, and a parameter model was obtained based on the B-splines (De Boor, 1978) transform. One crucial step in this clustering method is to apply the EM algorithm to obtain maximum likelihood estimates; optimum clusters can then be determined by the Bayesian information criteria (BIC) evaluation. Ma *et al*. (2006) also demonstrated that gene expression changes over time; thus, different gene sequences have different characterization functions, and each gene can have the same stochastic effect in the same clusters. Notably, nonparametric regression was applied to estimate the mean curve of each series during clustering. James *et al*. (2003) proposed a clustering procedure that is applicable to various curve data but is especially useful when individuals are observed at a sparse set of time points.

3. Methodology

3.1. *PPR Models*

The primary concept underlying projection pursuit regression (PPR) is as follows. Let Y and $\underline{X} = (X_1, X_2, ..., X_p)'$ be the response and explanatory

vectors, respectively. Suppose one has observations y_i and corresponding predictors $x_i = (x_{i1}, x_{i2}, ..., x_{ip})'$, $i = 1, \cdots, n$. Let $\alpha_1, \alpha_2, \cdots, \alpha_{M_0}$ be p-dimensional unit "directional" vectors, and let

$$\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i .$$

The PPR function locates the $M = M_0$ directional vectors $\alpha_1, \alpha_2, \cdots, \alpha_{M_0}$ and a good nonlinear transformation $f_1, f_2, \cdots, f_{M_0}$, such that

$$y \approx \overline{y} + \sum_{m=1}^{M_0} \beta_m f_m(\alpha_m^T x)$$

provides a good model for data. Formally, y and x are assumed to satisfy the conditional expectation model,

$$E(Y \mid X_1, \cdots, X_p) = \mu_y + \sum_{m=1}^{M_0} \beta_m f_m(\alpha_m^T \underline{X}) \tag{4}$$

where f_m has been standardized to have mean zero and unity variance:

$$E f_m(\alpha_m^T x) = 0, \quad E^2 f_m(\alpha_m^T x) = 1, \quad m = 1, \cdots, M_0.$$

Model parameters $\beta_m, f_m, a_m, m = 1, \cdots, M_0$ in Eq. (4) minimize mean squared error

$$E^2[Y - \mu_y - \sum_{m=1}^{M_0} \beta_m f_m(\alpha_m^T \underline{X})]^2. \tag{5}$$

For instance, suppose $E(Y \mid X_1, X_2) = X_1 X_2$. This is described by (4) with $\mu_y = 0$, $M_0 = 2$, $\beta_1 = \beta_2 = 1/4$, $a_1 = (1,1)^T$, $a_2 = (1,-1)^T$, $f_1(x) = x^2$, $f_2(x) = -x^2$. Thus, this PPR model is a simple interaction model. Due to the length limitation of this work, researchers interested in the PPR estimation method can refer to Friedman and Switzer (1981).

3.2. *PPR Clustering*

To apply the PPR model for time-series clustering, we assume the series in the same cluster are generated by the same model. Hence, if two sequences, S_i and S_j, are similar, the fitted models m_i and m_j can also reflect the similar structural relationship. Furthermore, if such a similar fitted model relationship exists, these fitted models should have similar predictive results for any sequence in two sequences. Thus, if predictive vector $\hat{y}_j$ is acquired by predictive series S_j through fitted model m_j, a similar predictive vector, $\hat{\hat{y}}_j$,

can be deduced through the m_i fitted model of this predictive vector; and *vice versa*. Therefore, clustering can be processed through recursively searching two similar series or cluster according their fitted models. Cross validation (CV) is a model-selection method that takes the predictive ability of a model as the basis for model selection. The basic purpose of CV is to divide a dataset into two parts—a training set and testing set. For each series, this work fits a PPR model using training data, and validates the model using testing data. Thus, the concept of CV is applied to determine the similarity between two fitted models during the clustering process. Let $\{S_1, \cdots, S_n\}$ be the set of time-series data that is divided into k clusters $\{C_1, \cdots, C_k\}$; the PPR clustering method is described simply as follows:

1. **Initiate:** Set the initial clusters $C_i = \{S_i\}$, $\forall 1 \le i \le n$.
2. For any two clusters, $\{C_i, C_j\}$, $\forall 1 \le i < j \le n$, and calculate the CV value to determine the similarity between any two series:
   (a) (a) Thus, the PPR method is individually applied to fit $\{C_i, C_j\}$ for obtaining the fitted model $\{m_i, m_j\}$ and fit vectors $\{\hat{y}_i, \hat{y}_j\}$.
   (b) Model m_i is applied to predict $\hat{y}_j$ for obtaining $\hat{\hat{y}}_j$; at the same time, model m_j is applied to predict $\hat{y}_i$ for obtaining $\hat{\hat{y}}_i$.
   (c) Define $CV(i, j) = \dfrac{\sum_{i=1}^{n_i - p} (\hat{y}_i - \hat{\hat{y}}_i)^2 + \sum_{j=1}^{n_j - p} (\hat{y}_j - \hat{\hat{y}}_j)^2}{n_i + n_j - 2p}$, where n_i means the summation of all sequence lengths in cluster C_i.
3. Define the upper triangular matrix $CV = [CV(i, j)]$ of size $n \times n$.
4. Select two clusters, (C_{i^*}, C_{j^*}), for meeting $(i^*, j^*) = aug(\min_{i<j} CV_{i,j})$ merged to the same clusters. Additionally, take CV_{i^*,j^*} as the CV* of this iteration.
5. Cycle. Repeat steps 2–4 until the data merges as a cluster.

The CV* value is calculated during clustering, as CV is a measurement of predictive ability of a fitted model. If the optimum number of a cluster is obtained, variance between clusters increases and variance within clusters decreases; thus, the CV* value increase markedly in the next recursion, and the optimum number of clusters can be determined based on these phenomena**.**

To evaluate clustering quality or compare clustering results with other methods proposed for similar time series clustering tasks, this work uses the cluster similarity measure developed by Gavrilov (2000) to assess the performance of clustering methods. Given two clustering sets, $G = (G_1, \cdots, G_C)$ and $A = (A_1, \cdots, A_C)$, the cluster similarity measure is defined by

$$sim(G, A) = \frac{1}{C} \sum_{i=1}^{C} \max_{1 \le j \le C} Sim(G_i, A_j) \tag{6}$$

where $sim(G_i, A_j) = 2|G_i \cap A_j| / (|G_i| + |A_j|)$, G is the clustering for the "ground truth" and A is obtained by a cluster method under evaluation.

4. Real Data Analysis

The data for demonstration analysis are personal annual average income for 25 states in the US in 1929–1999. An economist divided high and low growth rates in personal income in 25 states into two clusters. The first cluster includes 17 states (CT, DC, DE, FL, MA, ME, MD, NC, NJ, NY, PA, RI, VA, VT, WV, CA, IL) on the eastern seacoast; California and Illinois are areas with high growth rates in personal income. The second cluster includes 8 inland states (ID, TA, IN, KS, ND, NE, OK, SD); these states are areas with low growth rates in personal income. In cluster analysis, this work applied the PPR method to group these 25 series into two clusters. In Table 5, p and M adopt the number of different candidate values used in the PPR clustering to assess the impact to performance. Table 1 shows analytical results. Clustering similarity was 0.762–0.802; thus, analytical results did not change obviously.

The number of clusters is examined further. The arithmetic recursion is combined 24 times until all data are in one cluster, and the $_{CV^*}$value of each time is calculated for each recursion.

The CV* value increases obviously between CV*(23)=374150.37 and CV*(24)=1744352.57 (Fig. 1); hence the optimum number of clusters is 2. Notably, this number of clusters agrees with the statement of 2 average personal income groups by the economists.

Kalpakis (2001) first assumed all personal income data are ARIMA, and applied the different methods—LPC, DFT, DWT, PCA and MSE—to extract the coefficients, and then took the Euclidean distance between coefficients as the

basis for clustering. Table 2 compares clustering results. The PPR clustering similarity is 0.81, which approaches the current optimum clustering result.

Table 1. Clustering quality of personal income for 25 US states with various orders.

	M=1	M=2	M=3	M=4	M=5
P=2	0.79	0.76	0.76	0.76	0.76
P=3	0.79	0.79	0.79	0.79	0.79
P=4	0.79	0.79	0.79	0.80	0.79

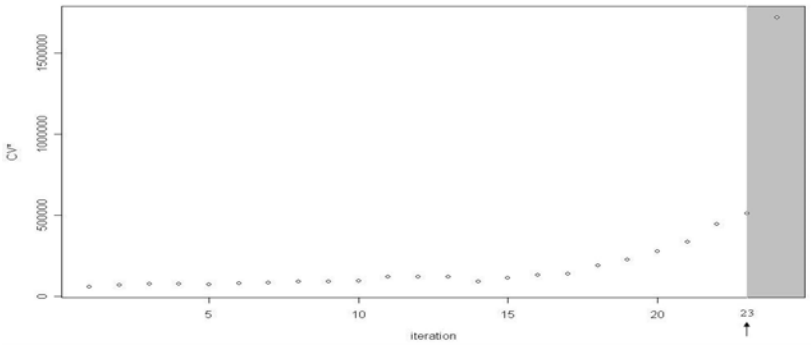

FIGURE 1 The CV*value for personal income in 25 US states

Table 2 Similarity of personal incomes in 25 US states clustered by various methods.

Method	PPR	LPC	DFT	DWT	PCA	MSE
Sim	0.81	0.84	0.68	0.60	0.68	0.78

References

[1] De Boor, C., “A Practical Guides to Splines.” Springer, 1978.
[2] A.P. Dempster, N. Laird, and D.B. Rubin, “Maximum likelihood estimation from incomplete data using the EM algorithm (with discussion).” J. Roy. Statist. Soc. Series B, 39, pp. 1–38, 1976.

[3] J. H. Friedman, and W., Stuetzle, “Projection Pursuit Regression,” Journal of the American Statistical Association, Vol. 76, pp. 817-823, 1981.
[4] S. Furui., 1989. “Digital Speech Processing, Synthesis, and Recognition.” Marcel Deckker, Inc., New York. 1989.
[5] M. D. Gavrilov, P. I. Anguelov, and R. Motwani, “Mining the stock market: Which measure is best?” In Proc. of the KDD, pp. 487-496, 2000.
[6] T. J. Hastie and R. J. Tibshirani, “Generalized Additive Models.” Chapman & Hall, London. 1990.
[7] K. Kalpakis, D. Gada and V. Puttagunta, “Distance measure for effective clustering of ARIMA time series.” In proc. of the IEEE ICDM, pp. 273-280, 2001.
[8] Y. Luan, and H. Li, “Clustering of time-course gene expression data using a mixed-effects model with B-splines.” Bioinformatics. Vol 19, No. 4, pp. 474-482, 2003.
[9] Y. Luan, and H. Li, “Model-based methods for identifying periodically regulated genes based on the time course microarray gene expression data.” Bioinformatics, 20, pp. 332–339, 2004.
[10] P. Ma, C. I. Castillo-Davis, W. Zhong, and J.S. Liu, “A data-driven clustering method for time course gene expression data. Nucleic Acids Res. Vol.34, No. 4, pp. 1261–1269, 2006.
[11] M. Medvedovic, and S. Sivaganesan, “Bayesian infinite mixture model based clustering of gene expression profiles.” Bioinformatics, Vol.18, No.9, pp 1194–1206, 2002.
[12] Peter J. Huber., “Projection Pursuit.” The annals of statistics, Vol. 13, No. 2., pp. 435-475, 1985
[13] L. R. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition.” Proc. of the IEEE, Vol. 77, No.2, pp. 257-286, 1989.
[14] L. R. Rabiner, C. H. Lee, B. H Juang, and J. G. Wilpon, “HMM clustering for connected word recognition”. IEEE. Int. conf. Acoust., Speech, Signal Processing, 1989, pp. 405-408, 1989.
[15] M. F. Ramoni, P. Sebastiani, and I. S. Kohane, “Cluster analysis of gene expression dynamics.” Proc Nat Acad Sci USA , Vol. 99, No. 14, pp. 9121-9126, 2002.
[16] Y. Xiong, and D. Y. Yeung, “Time series clustering with ARMA mixtures.” Pattern Recognition, Vol. 37, No. 8, pp. 1675-1689, 2004.