이학석사학위논문

A survey on model based time series classification and clustering

시계열 분류, 군집분석에 대한 조사

2023 년 2 월

서울대학교 대학원

통계학과

김 세 호

A survey on model based time series classification and clustering

시계열 분류, 군집분석에 대한 조사

지도교수 이 상 열

이 논문을 이학석사 학위논문으로 제출함

2022 년 10 월

서울대학교 대학원

통계학과

김 세 호

김세호의 이학석사 학위논문을 인준함

2023 년 1 월

위 원 장	김용대	(인)
부위원장	이상열	(인)
위 원	원중호	(인)

Abstract

Time series classification and clustering have become a significant challenge in data mining with the availability of storing vast amounts of time series data. Due to its tricky property, traditional methods, such as K-means, K-nn, and SVM, do not directly apply to time series analysis. However, despite its challenging aspects, time series classification and clustering are helpful in understanding data structure and finding new patterns in unstructured time series. For this reason, it has emerged as a popular topic in data mining, and there are many relevant articles. This review holistically discusses the essential parts of some existing research, focusing on a model-based approach to time series classification and clustering. Although there are several comprehensive reviews on this topic, they are too broad to get specific knowledge or insight quickly. Thus, we give brief instructions about the overall process for those interested in statistical applications.

Keywords: Time series clustering, classification, model-based approach

Student Number: 2021-25928

Contents

Chapter 1

Introduction

Clustering and classification have become a popular and significant challenge in the machine learning field throughout the years, owing to advanced data storage systems. Thus, many researchers have attempted to develop various methods or algorithms to enhance the overall process and to apply them to diverse fields. However, due to the characteristics of time series data, practitioners must conflict with some problems attributed to temporality. Despite these problems, time series classification and clustering have rich applications in diverse areas like engineering, finance, etc. (Keogh and Kasetty (2002); Geurts (2001)).

The clustering and classification process are somewhat different when one deals with time series data other than classic (or static) data. This is attributed to the tricky property of the time series data when one tries to solve problems with a traditional statistical approach. Generally, the time series dataset has large dimensions and heavy sizes because of ordinality. Also, each data point in one series may have high autocorrelations. Thus in clustering or classification, it is reasonable to consider the whole time series as one object. In literature,

there are ways to transform the entire time series. But in this process, another challenge arises, such as determining whether one object (the whole time series) is close to the other. Consequently, determining the distance (similarity) between two time series in conjunction with specific representation methods has been a central topic in recent years. So, researchers must consider both the representation methods and similarity distances simultaneously. To cope with this, the three ways, "shape-based," "feature-based," and "model-based" approaches, are adopted (see Abanda et al. (2019); Aghabozorgi et al. (2015)), which are addressed below.

1.1 Time series clustering and classification (overview)

We can define time series as real-valued series which has its domain (typically) in $\mathbb{R}^+$. Time series data is dynamic since its output values are function or random elements depending on the varying time. As a result, dimension, size, autocorrelation, and the unequal length of time series can be a problem. Thus, in time series clustering and classification (abbreviated as from now on TSCL and TSC, respectively), one usually regards the whole series as one object. TSCL and TSC processes have similarities in approach though they are basically different. More formally, we can summarize as follows:

Time series clustering : Given n time-series data set $\{\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_n\}$, where each $\mathbf{S}_i$ is whole time series, partition this to K-classes using similarity measures.

Time series classification : Given n time-series data set $\{\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_n\}$, and assumption is given as each $\mathbf{S}_i$ has specific label from 1 to K, assign each $\mathbf{S}_i$ to one of the labels using also similarity measures.

The main difference is whether known labels exist, which is also a significant challenge in some situations. Indeed, clustering is a pre-processing in many cases before classification. Thus it is justifiable to view these processes in a comprehensive view.

As mentioned above, in many literatures, TSCL and TSC methods are typically classified into three main categories. These are so-called "shape," "feature," and "model"-based approaches. Though their names and taxonomies can be somewhat different through literature, we use these three terms. In the shaped-based method, time series is used in raw form or transformed by non-linear transformation. This approach is mainly functional when dealing with a relatively short time series. A main interest is then distinguishing just shape profiles (Maharaj et al. (2019)). After this, an appropriate similarity measure can be chosen for raw/transformed time series.

In the feature-based method, observed raw time series must be sent to (or extracted to) some new vector space (usually has a lower dimension) to use Euclidean distances. The feature-based approach can remedy many problems of shaped-based approaches like high dimensionality and autocorrelation issues.

In the model-based method, specific stochastic models are assumed, and then time series are generated from one of the underlying models. This approach can be considered parametric since the observed time series is first converted to model parameter vectors. Then suitable metrics (similarity measures) are given using these parameters. Herein, we will mainly focus on the model-based approach.

1.2 Agenda of the review

This review presents the overall literature on time series clustering and classification, emphasizing more detail in the model-based approach. This review would be beneficial for those aiming to develop theoretical approaches since most review papers focus on algorithms.

The next chapter will give a conventional framework and literature review for the shape and feature-based approach. In Chapter 3, before going to the model-based part, we briefly review some of the well-known stochastic models and how they can be related to TSCL and TSC. Subsequently, the model-based approach is presented in detail with explanations of time series models. Finally, in Chapter 4, the conclusion and further discussion are provided.

Chapter 2

Framework of shape and feature based approach

2.1 Shape-based approach of TSCL and TSC

The shape-based approach is often called the observation-based approach as it uses raw time series data. Hence, the standard Euclidean metric should be modified to measure similarity. This section will briefly review the one-to-one Euclidean distance-based approach called "Lockstep measures" and the concept of dynamic time warping.

Euclidean distance-based method is proper when dealing with local geometric shapes. For example, point-wise Euclidean distance-based measures are given in D'Urso (2000). See also D'Urso (2000) who proposed these measures to cluster multivariate time series. For example, a straightforward form of this measure is given as follows:

D'Urso (2000) : Put two multivariate time series data point as $\mathbf{x}_t^{(i)} = (x_{1t}^{(i)}, \ldots, x_{pt}^{(i)})$, $\mathbf{x}_t^{(j)} = (x_{1t}^{(j)}, \ldots, x_{pt}^{(j)})$, where each component $\mathbf{x}_{kt}^{(i)}$ represents k-th feature

of i-th observed time series at time t. And further assumes that time domain is common set $\{1, \ldots, T\}$. Then similarity measure between two time series $\mathbf{x}^{(i)}$, $\mathbf{x}^{(j)}$ is given as :

$$d(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \sqrt{\sum_{t=1}^{T} (\|\mathbf{x}_t^{(i)} - \mathbf{x}_t^{(j)}\| w_t)^2}$$

where w_t is weight parameter at time t and $\|\cdot\|$ is standard euclidean norm.

He also devised a similar metric using in place of vector at time t, namely $\mathbf{x}_t^{(i)}$, linearly transformed observed series to measure the deviation on slope and convexity. Finally, he also devised the so-called polygonal coefficient to measure the geometrical oscillation in terms of the time domain, which could determine weight parameters in time intervals. Note that, as seen in the distance equation, one needs an equal time point to measure the distance, which is impossible when dealing with an unequal time series length. Furthermore, this method will be computationally expensive and only locally applicable when analyzing longer time series lengths. However, despite these drawbacks, this distance-based method is proper when distinguishing local patterns. Also, one can try the conventional clustering/classification method directly. For more information, see also Coppi and D'Urso (2001).

Next, many distances are based on dynamic time warping (DTW). As we can see in the above example, euclidean distance cannot capture the similarity between the unequal length of two time series. Dynamic time warping emerged to solve problems in lock step measures which finds optimal passage of time points (Sakoe and Chiba (1978); Berndt and Clifford (1994)). Its main use in conjunction with classic machine learning algorithms like k-NN, k-medoids, k-

means, etc., shows substantial accuracy. For example, in Wang et al. (2013), DTW presents significant accuracy in time series classification more than euclidean distance.

DTW algorithm seeks minimal cost over all possible warping paths. So many use dynamic programming to get DTW scores iteratively. However, despite these algorithms, its algorithmic complexity amounts to $\mathbf{O}(nm)$ where n and m represent each time series' length, respectively. So its computational cost is somewhat expensive and has limitations when dealing with long time series (Berndt and Clifford (1994)). Also, in Wang et al. (2013), he concluded that the DTW method's accuracy converges with euclidean's accuracy. Furthermore, DTW is not a standard distance since it does not obey triangle inequality, making it hard to use algorithms like the K-dimensional tree or the ball tree (Faouzi (2022)). Although there are many drawbacks, DTW is used substantially in many areas. For more examples, see Aach and Church (2001).

2.2 Feature based approach of TSCL and TSC

As discussed in the previous section, using lock step measures requires the same time domain and comparing each data point independently. Thus it cannot capture the structure of autocorrelation, which is very common in time series data. To escape from this, the concept of DTW-based metric appeared, which nonlinearly transforms time domains and uses a time path to compare two time series. But it also has drawbacks like algorithmic complexity, semi-metric property, and etc. Thus variants of the DTW metric have appeared. Although there have been many improvements to overcome such problems, shape based approach has some intrinsic limitations. These include non-robustness attributed to noise in data, which can classify or cluster series wrongly (Ratanamahatana

and Keogh (2005); Ratanamahatana et al. (2005)). Furthermore, most shape-based (or observed) approaches require high costs and make it expensive to implement clustering/classifying analysis. In this perspective, the feature-based method has arisen to overcome these problems. Because the feature-based approach aims at distinguishing generating process, it is superior to shape based in some aspects since the shape-based method focuses mainly on geometric profiles. Furthermore, dimension reduction usually occurs when extracting certain features in the original series. As a result, the computational cost reduces, which is also an important goal in contemporary data science.

Typically, feature-based methods are based on the notion that one carries time series to another transformed (vector) space. According to the many advantages listed above, the feature extraction method is considered a base solution for time series classification/clustering. In general, feature-based methods can be classified into three types: time domain feature, frequency domain feature, and wavelet-based feature approach. The remainder of this section will be devoted to a brief introduction to some well-known methods in this domain and some of the literature using this.

The autocorrelation function (ACF) is used for time domain features to measure the distance between two time series. This method is somewhat similar to the model-based approach, which will be discussed in the next chapter. Other autocorrelation types, like partial ACF(PACF) or inverse ACF(IACF), are also used to define the metrics between two time series. Some examples of these are illustrated in Alonso and Maharaj (2006); Caiado et al. (2009); D'Urso and Maharaj (2009).

Next, in frequency domain features, periodogram-based distance measures are used. This method can also be applied to unequal time series lengths based on spectral analysis. Other approaches include discrete Fourier transform (DFT,

Agrawal et al. (1993)) and discrete cosine transform (DCT). Also, see Caiado et al. (2009); Maharaj and D'Urso (2011) for more detailed explanations. Finally, in the wavelet-based feature approach, discrete wavelet transform (DWT) parameters are used to cluster/classify time series(Chan and Fu (1999); Kawagoe and Ueda (2002)). Other wavelet-based methods include Chebyshev Polynomials (Cai and Ng (2004)).

Chapter 3

Stochastic time series model approach

3.1 Introduction

In the model-based approach, assuming each time series originates from a specific probabilistic model, the first step is to measure the distance between a pair of models. Then, using these measures one can apply standard classification or clustering methods to all of given time series datasets. This chapter presents a framework for this approach and investigates the literature on this topic.

In the model-based approach, many literatures assume the ground stochastic model to be linear and Gaussian. With this assumptions, Shumway and Unger (1974) used Kullback-Liebler divergence to discriminate between the underlying two models. Also, Kailath (1967) used Bhattacharyya distance to measure the distance between two probabilistic models. These assume that underlying stochastic processes are gaussian or stationary. See also Korzhik et al. (2008); Sharif et al. (2010); Georgiou and Lindquist (2003); Grivel et al. (2021). Another

approach is transforming the time series into well-known model parameter vectors. Piccolo (1990) introduced the AR distance between ARIMA models, which is calculated as the Euclidean distance between the coefficients of AR(∞) expansion. This research was based on the assumption that the underlying model is an invertible and causal ARIMA process. Related works are presented in Maharaj (1996, 2000) aiming to remedy the problem of controlling the number of model parameters. These papers all provided hypothesis testing methods to discriminate or cluster a given time series dataset and test statistics' properties. Also, similarly to Corduas and Piccolo (2008), asymptotic distributional properties of AR distance were provided.

For models dealing with heteroskedastic time series, GARCH models are often used. In financial time series, estimating the volatility of financial data, such as stock prices, market indices, etc., has become one of the main parts of econometrics. Clustering models based on GARCH appeared in recent years, see Otranto (2008), Caiado and Crato (2010), D'Urso et al. (2016), Khan et al. (2019).

3.2 AR expansion based method

We will first present some general notation and definitions discussed in this section. Put zero mean stationary stochastic process $(X_t)_{t\in\mathbb{Z}}$ following ARMA(p,q) process as

$$\phi_p(B)X_t = \theta_q(B)Z_t$$

with each term $\phi_p(z) = 1 + \phi_1 z + \ldots + \phi_p z^p$, $\theta_q(z) = 1 + \theta_1 z + \ldots + \theta_q z^q$, B is back-shift operator, and Z_t is generally white noise with constant variance σ^2. It is well known that (see Brockwell and Davis (2002) or Montgomery et al. (2015)) above X_t has a unique stationary solution and also causal and invertible

if all roots of $\phi_p(z)$, $\theta_q(z)$ lie outside of the unit disk and are not common. That is, $\phi_p(z)\theta_q(z) \neq 0$, if $|z| \leq 1$. Then integrated process autoregressive integrated moving average (ARIMA(p,d,q)) is defined as $Y_t = (1-B)^d X_t$.

We define invertibility using the above notation as the existence of sequence $\{\pi_i\}$ that the series absolutely converges and also satisfies the relationship:

$$Z_t = \sum_{i=0}^{\infty} \pi_i X_{t-i}, \qquad t \in \mathbb{Z}.$$

And similarly, define causality as the existence of sequence $\{\psi_i\}$ which the series also absolutely converges and :

$$X_t = \sum_{i=0}^{\infty} \psi_i Z_{t-i}, \qquad t \in \mathbb{Z}.$$

As introduced above, one can cluster or classify time series data based on a model-based approach with a minor assumption, not using specific models. The next step is to use classic or general divergence measures such as kullback-leiber distance, bhattacharyya distance, etc., rather than model-dependent specific similarity measures. When a particular model holds(i.e., the model assumption holds), the latter approach generally gives better results. Many time series datasets are well suited to econometrics, finance, and biostatistics models. Thus, we will explain from now on a model-based approach. The ARMA, ARIMA, and related AR metric is our first topic.

If underlying model is causal and invertible ARMA(p,q), then by aforementioned definition, sequence $\{\pi_i\}$ is determined by :

$$\pi(z) = \sum_{i=0}^{\infty} \pi_i z^i = \phi(z)/\theta(z).$$

And each coefficient can be calculated by recursion algorithms. In Piccolo (1990), he defined a dissimilarity measure between two ARIMA class processes

using this AR(∞) expansion. As a remark, although we illustrate ARMA(p,q) process (i.e., ARIMA(p,0,q) process), basically ARIMA(p,d,q) process can be dealt with using the same method. Now the metric between two time series X_t, and Y_t from the ARIMA class is given as :

$$d = \sqrt{\sum_{j=1}^{\infty} (\pi_{xj} - \pi_{yj})^2}$$

where each π_{ij}, $i = x, y$ represents AR(∞) coefficients from each time series. This measurement satisfies all axioms of metrics: non-negativity, symmetry, triangle inequality, and also convergent since each series are absolutely convergent. Since π coefficients carry the structure of the underlying stochastic models, comparing the euclidean distance to measure the dissimilarity between two series seems reasonable. Furthermore, since π coefficients and observations fully determine prediction value at some specific time until the given time, a smaller AR distance would imply similar prediction values. Thus in terms of forecasting perspective, it also seems reasonable to use AR metrics. Subsequently, conventional clustering, like hierarchical methods, can be applied based on this AR distance. However, some limitations exist, like controlling the number of parameters of underlying models. Since in Piccolo (1990), he calculated distance after fitting each ARMA model under consideration. Consequently, there needed to consider the case of a different number of parameters of underlying models.

In Maharaj (1996, 2000), he developed Piccolo's idea by proposing a statistical test and using p-value to cluster the given time series datasets. Also, he tried to solve the problem above by directly fitting the π coefficients through truncated AR(∞) models. In fitting the truncated AR(∞) model, selection criteria such as Akaike's information criteria (AIC) can be used. Also, in Corduas and Piccolo (2008), he solved the problem of fitting the original ARIMA model

by setting the number of π coefficients first and using ML estimators. Some of the details will be followed from now on.

Using the above notation, an invertible ARMA model can be expressed as AR(∞) type as :

$$X_t = \sum_{i=1}^{\infty} \pi_i X_{t-i} + Z_t$$

.

For two time series (X_t) and (Y_t), the number of coefficients of AR(∞) expression is chosen first respectively for (X_t) and (Y_t) as m_1, m_2 by model selection method such as AIC or BIC. Then we put $m = m_1 \vee m_2$, and corresponding parameters and estimated vectors as :

$$\pi_i = (\pi_{1i}, \ldots, \pi_{mi})^t, \quad \hat{\pi}_i = (\hat{\pi}_{1i}, \ldots, \hat{\pi}_{mi})^t \qquad i = x, y$$

Then, assuming without loss of generality $m_1 < m_2$, component of vector π_x ($\hat{\pi}_x$ also) would be considered as $\pi_{jx} = 0$ if $j > m_1$.

Now hypothesis is given as $H_0 : \pi_x = \pi_y \quad H_1 : \pi_x \neq \pi_y$, and test statistics can be obtained by generalized least squares method implemented at combined models. Then finally, obtained estimator $\hat{\pi}$ follows asymptotical normal where $\pi = \begin{pmatrix} \pi_x \\ \pi_y \end{pmatrix}$ and $\hat{\pi} = \begin{pmatrix} \hat{\pi}_x \\ \hat{\pi}_y \end{pmatrix}$. Also, converting null hypothesis H_0 as equivalent form using an augmented matrix :

$$\pi_x = \pi_y \quad \Longleftrightarrow \quad D\pi := [I_m \ \ -I_m]\pi = 0$$

, leads to $D\hat{\pi}$ being asymptotically normal under the null hypothesis also. Furthermore, a quadratic form based on these statistics can be obtained asymptotically as χ^2(m). For more detailed explanations of the deriving procedure, see Maharaj (2000) or chapter 7 on Maharaj et al. (2019).

A similar but somewhat different hypothesis testing is also given in Corduas and Piccolo (2008). As mentioned, he suggested truncating π coefficients after fitting each ARIMA model, leaving the possibility of a different number of original parameters. Then with the same hypothesis $H_0 : \pi_x = \pi_y$ or alternatively $H_0 : d = 0$, distribution of $\hat{\pi}_x - \hat{\pi}_y$ is derived as following asympotically normal. Finally, since $\hat{d} := (\hat{\pi}_x - \hat{\pi}_y)^t(\hat{\pi}_x - \hat{\pi}_y)$ represents the estimated distance between X_t and Y_t, some well known quadratic theorem can be applied to use $\hat{d}$ as a test statistics. It is represented asymptotically as a linear combination of χ-squared random variables.

After obtaining the p-value from this hypothesis framework, one can implement clustering algorithms like hierarchical clustering, k-means clustering, or k-medoids based on the obtained p-values. Although AR metrics can carry conventional clustering algorithms, like agglomerative or divisive hierarchical clustering, its interpretability is not better than p-value-based clustering. Two time series objects are considered members of the same clusters at a given significance level if the corresponding p-value is greater than the given level. Subsequently, one cluster should have a property that all objects in that cluster have pairwisely bigger p-values than the given level. As a result, the closeness of each object can be measured or interpreted statistically.

In summary, the essence of using AR metric to measure the nearness between two items is converting the underlying probabilistic model to well known Euclidean vector. Then by the AR metric, one can implement clustering algorithms followed by elementary statistical procedures. Furthermore, classification or discriminating analysis can be done in this framework. The asymptotic normality of test statistics and estimated distance makes it possible to view these procedures as equivalent to Fisher's lda or qda type analysis (for more information, see also Corduas and Piccolo (2008)).

3.3 GARCH model approach

Approaches considered until now have mainly focused on dealing with the mean values. But as mentioned earlier, the modeling approach has a significant drawback: the accuracy of the classification or clustering is highly dependent on assumed models. Thus, the AR metric-based method would also fail if the homoskedastic variance assumption fails. Furthermore, many financial time series like stock market indices, risk indexes, and portfolio investments exhibit heteroskedastic variance. So, it is natural to focus on the variance part of the time series instead of the mean part. Indeed, clustering or classification algorithms is helpful to investors since clustering or classifying volatile financial items correctly will prevent investors' failure.

Below, we consider the GARCH model (Bollerslev (1986)):

$$Y_t = \mu_t + \epsilon_t$$

$$\epsilon_t = \sqrt{h_t} u_t,$$

where u_t is usually i.i.d normal with mean 0, variance one (or more generally just white noise) and ϵ_t is disturbance term. Also we call h_t as conditional variance following GARCH(p,q) if:

$$h_t = \sum_{i=0}^{p} \alpha_i \epsilon_{t-i}^2 + \sum_{j=1}^{q} \beta_j h_{t-j},$$

with the restriction $\alpha_0 > 0$ and all $\alpha_k, \beta_k \geq 0$.

Assuming GARCH(p,q) model, Otranto (2008) introduced AR metric-based clustering, who represented model volatility as unconditional, time-varying, and structural volatility. Putting $e_t := \epsilon_t^2 - h_t$ leads to ARMA(p*,q) model and assuming suitable coefficients condition, AR(∞) expression can be derived:

$$\epsilon_t^2 = \frac{\alpha_0}{1 - \sum_{j=1}^{q} \beta_j} + \sum_{k=1}^{\infty} \pi_k \epsilon_{t-k}^2 + e_t.$$

Now, the conditional expectation of squared disturbance term at time t+1 given the information until t is:

$$\mathbb{E}[\epsilon_{t+1}^2|H_t] = \frac{\alpha_0}{1-\sum_{j=1}^{q}\beta_j} + \sum_{k=1}^{\infty}\pi_k\epsilon_{t-k}^2,$$

where H_t represents information till time t. The first term represents constant volatility or risk, and the second is time-varying volatility. Subsequently, taking expectation gives unconditional volatility, which is represented as:

$$\mathbb{E}[\epsilon_{t+1}^2] = \frac{\alpha_0}{(1-\sum_{j=1}^{q}\beta_j)(1-\sum_{k=1}^{\infty}\pi_k)}.$$

He measured the similarity of time-varying volatility by π coefficients, somewhat similar to AR metric approach. Since the same AR distance between two AR expressions yields the same time-varying volatility term, these can capture similar volatility structures. Furthermore, divergence from the null model (all $\pi_k = 0$) was used as an amount of volatility.

Using these unconditional and time-varying volatilities, he clustered following the three steps using hypothesis testing and p-value as before: first by unconditional volatility, second by time-varying volatility, and finally by similar parametric structure. This is because similar unconditional and time-varying volatility does not guarantee the same underlying model due to the nonlinear combination of the parameters. On the other hand, similar parameter estimates (i.e., similar structure estimates) will guarantee other ones. To test the last step, one should check that whole GARCH parameter $(\alpha_0, \alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_q)$ are equivalent in the whole time series dataset under consideration. This test of volatility structure was introduced in Otrano and Triacca (2007). Major differences with the previous section's AR metric-based clustering approach are that hypothesis testing is conducted holistically, and at steps 1 and 2, each cluster

has a numerical hierarchy. Thus, one can readily see that the top-down clustering method is justifiable for these reasons. Also, benchmark series are chosen automatically to cluster at each step, and Wald statistical test is implemented.

In D'Urso et al. (2016), weighted distance using the above unconditional, time-varying volatility term is introduced. For given two time series X_t and Y_t, and weight parameter w_1, w_2, metric is given accordingly :

$$d_{xy} = \sqrt{[w_1^2(uv_x - uv_y)^2 + w_2^2(tvv_x - tvv_y)^2]},$$

where uv and tvv stand for each series' unconditional volatility and time-varying volatility terms. Furthermore, restrictions on weight w_1, w_2 are imposed as $w_1 + w_2 = 1$ and $w_1, w_2 \geq 0$. Based on this distance, he introduced some robust clustering models by a fuzzy-clustering method which assigns a certain probability of belonging to a cluster to each object. These models are constructed by partitioning around the medoids (PAM) procedure and have the robustness to the anomalies.

For other GARCH-based or volatility approaches, in Caiado and Crato (2010), they introduced Mahalanobis-type distance between the dynamic features using a threshold GARCH model. Another recent GARCH-based fuzzy clustering work is illustrated in Cerqueti et al. (2021). This paper extended the original GARCH-based fuzzy clustering methods using higher conditional moments. Also, some applications of GARCH model clustering can be found in Niyitegeka and Tewar (2013); Caiado and Crato (2007).

3.4 Remark

The essence of the model-based approach is that it assumes a stochastic model as a generating process, which is a parametric approach unlike the shape or feature-based approach. Although some intrinsic limitations could cause low

accuracy in a specific situation, its careful use can yield a substantial accuracy in financial, econometric, and other applied fields. This section presented model-based distance and how it can be applied to clustering/classifying procedures. At the clustering/classifying stage, the p-value of the hypothesis testing plays a significant role as a base instrument for many clustering algorithms. Thus, it is crucial to understand that constructing hypothesis testing is essential in a model-based approach.

Chapter 4

Conclusions

In this review, we have introduced various time series clustering/classification methods. In particular, to ease the difficulty of understanding the vast amount of research work, we first classified the types of approaches: shape-based, feature-based, and model-based. Also, this review did not try to emphasize each specific methodology but focused on understanding the framework of whole procedures.

As may be noticed, shape-based, feature-based, and model-based approaches are classified according to how dissimilarity is measured. In the shape-based approach, one mainly focuses on the dissimilarity in a geometric shape. On the other hand, characterizing the underlying process is a major issue in a feature or model-based approach. The feature-based method is nonparametric, while the model-based approach is parametric. Considering the characteristics and limitations of each type of approach, one should choose proper methods when dealing with time series data.

Since the model-based method is essentially parametric, well-known statistical tools can be applied easily. Many famous literatures use the distance ob-

tained by coefficients of the underlying stochastic model. In this paper, an AR-based metric and its variant have been introduced. Also, for the heteroscedastic variance model, a GARCH-based metric was also introduced. In many situations, p-value and hypothesis testing is used to cluster or discriminate time series data.

In recent studies, many dissimilarity measures have been combined with post hoc clustering/classifying algorithms to find the highest accuracy procedure. As there is no typical standard for selecting specific metrics in conjunction with post hoc clustering or classifying method, it has become a significant challenge to combine proper metrics and algorithms. Also, for statisticians who mainly deal with model-based approaches, choosing the best models in each case and modifying the metrics is a quite challenging task. Furthermore, since time series datasets are generally large, reducing the complexity remains a significant problem.

Bibliography

Aach, J. and Church, G. M. (2001). Aligning gene expression time series with time warping algorithms. *Bioinformatics*, 17(6):495–508.

Abanda, A., Mori, U., and Lozano, J. A. (2019). A review on distance based time series classification. *Data Mining and Knowledge Discovery*, 33(2):378–412.

Aghabozorgi, S., Shirkhorshidi, A. S., and Wah, T. Y. (2015). Time-series clustering–a decade review. *Information systems*, 53:16–38.

Agrawal, R., Faloutsos, C., and Swami, A. (1993). Efficient similarity search in sequence databases. In *Foundations of Data Organization and Algorithms: 4th International Conference, FODO'93 Chicago, Illinois, USA, October 13–15, 1993 Proceedings 4*, pages 69–84. Springer.

Alonso, A. M. and Maharaj, E. A. (2006). Comparison of time series using subsampling. *Computational statistics & data analysis*, 50(10):2589–2599.

Berndt, D. J. and Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA, USA:.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics*, 31(3):307–327.

Brockwell, P. J. and Davis, R. A. (2002). *Introduction to time series and forecasting*. Springer.

Cai, Y. and Ng, R. (2004). Indexing spatio-temporal trajectories with chebyshev polynomials. In *Proceedings of the 2004 ACM SIGMOD international conference on Management of data*, pages 599–610.

Caiado, J. and Crato, N. (2007). A garch-based method for clustering of financial time series: International stock markets evidence. In *Recent advances in stochastic modeling and data analysis*, pages 542–551. World Scientific.

Caiado, J. and Crato, N. (2010). Identifying common dynamic features in stock returns. *Quantitative Finance*, 10(7):797–807.

Caiado, J., Crato, N., and Peña, D. (2009). Comparison of times series with unequal length in the frequency domain. *Communications in Statistics—Simulation and Computation®*, 38(3):527–540.

Cerqueti, R., Giacalone, M., and Mattera, R. (2021). Model-based fuzzy time series clustering of conditional higher moments. *International Journal of Approximate Reasoning*, 134:34–52.

Chan, K.-P. and Fu, A. W.-C. (1999). Efficient time series matching by wavelets. In *Proceedings 15th International Conference on Data Engineering (Cat. No. 99CB36337)*, pages 126–133. IEEE.

Coppi, R. and D'Urso, P. (2001). The geometric approach to the comparison of multivariate time trajectories. In *Advances in Classification and Data Analysis*, pages 93–100. Springer.

Corduas, M. and Piccolo, D. (2008). Time series clustering and classification by the autoregressive metric. *Computational statistics & data analysis*, 52(4):1860–1872.

D'Urso, P., De Giovanni, L., and Massari, R. (2016). Garch-based robust clustering of time series. *Fuzzy Sets and Systems*, 305:1–28.

D'Urso, P. (2000). Dissimilarity measures for time trajectories. *Journal of the Italian Statistical Society*, 9(1):53–83.

D'Urso, P. and Maharaj, E. A. (2009). Autocorrelation-based fuzzy clustering of time series. *Fuzzy Sets and Systems*, 160(24):3565–3589.

Faouzi, J. (2022). Time series classification: A review of algorithms and implementations. *Machine Learning (Emerging Trends and Applications)*.

Georgiou, T. T. and Lindquist, A. (2003). Kullback-leibler approximation of spectral density functions. *IEEE Transactions on Information Theory*, 49(11):2910–2917.

Geurts, P. (2001). Pattern extraction for time series classification. In *European conference on principles of data mining and knowledge discovery*, pages 115–127. Springer.

Grivel, E., Diversi, R., and Merchan, F. (2021). Kullback-leibler and rényi divergence rate for gaussian stationary arma processes comparison. *Digital Signal Processing*, 116:103089.

Kailath, T. (1967). The divergence and bhattacharyya distance measures in signal selection. *IEEE transactions on communication technology*, 15(1):52–60.

Kawagoe, K. and Ueda, T. (2002). A similarity search method of time series data with combination of fourier and wavelet transforms. In *Proceedings Ninth International Symposium on Temporal Representation and Reasoning*, pages 86–92. IEEE.

Keogh, E. J. and Kasetty, S. (2002). On the need for time series data mining benchmarks: A survey and empirical demonstration. *Data Mining and Knowledge Discovery*, 7:349–371.

Khan, M. S., Khan, K. I., Mahmood, S., and Sheeraz, M. (2019). Symmetric and asymmetric volatility clustering via garch family models: An evidence from religion dominant countries. *Khan, MS, Khan, KI, Mahmood, S., & Sheeraz, M.(2019). Symmetric and asymmetric volatility clustering via GARCH family models: An evidence from religion dominant countries. Paradigms*, 13(1):20–25.

Korzhik, V., Imai, H., Shikata, J., Morales-Luna, G., and Gerling, E. (2008). On the use of bhattacharyya distance as a measure of the detectability of steganographic systems. In *Transactions on Data Hiding and Multimedia Security III*, pages 23–32. Springer.

Maharaj, E. A. (1996). A significance test for classifying arma models. *Journal of Statistical Computation and Simulation*, 54(4):305–331.

Maharaj, E. A. (2000). Cluster of time series. *Journal of Classification*, 17(2).

Maharaj, E. A., D'Urso, P., and Caiado, J. (2019). *Time series clustering and classification*. Chapman and Hall/CRC.

Maharaj, E. A. and D'Urso, P. (2011). Fuzzy clustering of time series in the frequency domain. *Information Sciences*, 181(7):1187–1211.

Montgomery, D. C., Jennings, C. L., and Kulahci, M. (2015). *Introduction to time series analysis and forecasting*. John Wiley & Sons.

Niyitegeka, O. and Tewar, D. (2013). Volatility clustering at the johannesburg stock exchange: Investigation and analysis. *Mediterranean Journal of Social Sciences*, 4(14):621.

Otrano, E. and Triacca, U. (2007). Testing for equal predictability of stationary arma processes. *Journal of Applied Statistics*, 34(9):1091–1108.

Otranto, E. (2008). Clustering heteroskedastic time series by model-based procedures. *Computational Statistics & Data Analysis*, 52(10):4685–4698.

Piccolo, D. (1990). A distance measure for classifying arima models. *Journal of time series analysis*, 11(2):153–164.

Ratanamahatana, C., Keogh, E., Bagnall, A. J., and Lonardi, S. (2005). A novel bit level time series representation with implication of similarity search and clustering. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 771–777. Springer.

Ratanamahatana, C. A. and Keogh, E. (2005). Multimedia retrieval using time series representation and relevance feedback. In *International Conference on Asian Digital Libraries*, pages 400–405. Springer.

Sakoe, H. and Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, 26(1):43–49.

Sharif, M., Uyaver, S., Djeraba, C., et al. (2010). Crowd behavior surveillance using bhattacharyya distance metric. In *International Symposium Computational Modeling of Objects Represented in Images*, pages 311–323. Springer.

Shumway, R. and Unger, A. (1974). Linear discriminant functions for stationary time series. *Journal of the American Statistical Association*, 69(348):948–956.

Wang, X., Mueen, A., Ding, H., Trajcevski, G., Scheuermann, P., and Keogh, E. (2013). Experimental comparison of representation methods and distance measures for time series data. *Data Mining and Knowledge Discovery*, 26(2):275–309.

국문초록

시계열 분류, 군집분석은 시간에 따른 방대한 데이터를 저장하는 능력과 함께 데이터 마이닝 분야에서 주요한 과제로 떠오르고 있다. 시계열 자료의 까다로운 성질에 의해 전통적 기법인 K-평균, K-근접이웃, SVM 등등은 직접적으로 적용이 쉽지 않다. 그러나 이러한 어려움에도 불구하고 시계열 분류, 군집분석은 데이터의 구조를 이해하는데 도움을 주고 구조화 되지않은 데이터에서 새로운 패턴을 발견할 수 있도록 도움을 준다. 이러한 이유로 데이터 마이닝 분야에서 인기있는 주제로 여겨지고 있고 수 많은 해당 연구들이 존재하고 있다. 이번 재검토 연구에서는 전체적으로 여러 연구들을 검토한 다음 통계적 응용에 목적이 있는 모델 기반의 접근법을 알아본다. 비록 이 분야에서 벌써 몇 가지의 재검토 연구가 존재하지만 대부분의 재검토 연구는 특정한 지식이나 통찰을 얻기에는 너무 방대하게 설명되고 있다. 따라서 이 연구에서는 처음 주제를 접하는 연구자를 위해 전반적인 과정들에 대한 간단한 설명을 하고 특히 통계적 응용과 연구에 관심이 있는 이들을 위해 모델 기반의 접근법을 소개한다.

주요어: 시계열 군집 분석, 시계열 분류, 모델 기반의 접근.

학번: 2021-25928

Acknowledgements

부족한 논문이지만 석사 과정 2년을 무사히 마칠 수 있게 물신양면으로 도와주신 저희 지도교수님 이상열 교수님께 먼저 감사의 마음을 전합니다. 또한, 아들이 원하는 바를 이루기 위해 대학원 생활을 지원해 주고 계신 부모님께 깊은 감사의 말씀을 전하고 싶고, 옆에서 항상 저를 응원해주고 어떤 일을 하더라도 지지해주는 금도에게도 깊은 고마움과 미안함을 전합니다. 마지막으로 2년동안 학업과 연구 자세에 대해 보고 배울 수 있게 도움을 준 연구실의 많은 선배들과 동기들에게도 감사의 말씀을 올립니다. 모두 훗날 학계, 기업 등 어느 곳에서라도 일이 잘 풀리길 기원합니다.

마지막으로, 앞으로 더 나아가는 박사과정이란 항해에서 석사 과정 2년이 든든한 배가 될 수 있도록 도와주신 여러 교수님들과 행정직원분들 다른 동기, 후배들에게도 감사함을 전합니다. 잘한 것들은 계속 살리고 부족했던 점은 깊이 성찰하여 더 나은 연구자가 되고자 노력하겠습니다. 감사합니다.