Multivariate time series classification analysis: State-of-the-art and future challenges

To cite this article: T Handhika et al 2019 IOP Conf. Ser. Mater. Sci. Eng. 536 012003

View the article online for updates and enhancements.
Multivariate time series classification analysis: State-of-the-art and future challenges

T Handhika*, Murni, D P Lestari, and I Sari

Computational Mathematics Study Center, Gunadarma University, Depok, 16424, Indonesia

Abstract. In the era of big data and internet of things (IoT) revolution, it is possible to observe some variables for an object in real time, e.g. sensor data obtained from one or more sensor devices. In addition to extrapolate these data simultaneously, the other hot issue is how to use these data for classifying some variables into some groups, respectively. Multivariate time series classification (MTSC) analysis provides various models to represent this problem according to its characteristics. In this paper, we tried to summarize state-of-the-art methods for MTSC analysis complete with their strengths and weaknesses. Furthermore, we also focused on some limitations from previous research for developing MTSC analysis in the future.

1. Introduction

In the era of industrial revolution 4.0, big data implementation supported by the internet of things (IoT) technology plays an important role in assisting data-driven decision making. They produce output in the form of predictions as well as insights from the patterns of the dynamics of existing data as a basis for managerial decision making. It was not limited to a certain field. As an example, in the health sector, a lung or breast cancer pathology diagnosis study was massively conducted through chest x-ray images [1, 2, 3, 4, 5, 6]. Another example in the field of risk management, big data technology is implemented in increasing efficiency of the credit scoring process for a new borrower [7, 8, 9, 10, 11, 12]. Both speech and gesture recognition are some of the innovations that result from the implementation of big data in the consumer electronics industry [13, 14, 15, 16, 17, 18]. In fact, big data technology is also implemented to improve driving safety procedures [19, 20, 21].

To simplify the complex problems faced in the real world, a problem can be represented in a mathematical model based on two approaches, that is statistics and machine learning. Statistics focused on generalizing results through a level of confidence, while machine learning focuses on the results of predictions measured by some performance indicators of accuracy. Perspectives on the data also differ between statistics and machine learning where the data processed by statistical approach is considered as sample data and the results are inferred to generalize the population. As for machine learning, the data processed is divided into two sets, that is training and testing data, where the magnitude of the accuracy of the models generated in the testing set becomes a reference for predicting new dataset. As for the problems that have been mentioned before, if seen in statistical perspective they will be
categorized into classification analysis. Equivalently, if they are seen in the context of machine learning they will be categorized into supervised learning [22]. Both are used to model classification problems, but with different emphases. Statisticians are interested in knowing what factors are significant in classifying objects. Conversely, machine learning focused on the classification results from the generated functions.

There are many methods that can be used to solve classification problems, such as discriminant analysis, Bayesian model, Markov model, tree-based classifiers, support vector machines, neural networks, linear vector quantization model, $k$-nearest neighbor, periodogram-based metric, dynamic time warping, functional echo state network and shapelet selection algorithm [11, 22, 23, 24, 25, 26, 27, 28, 29]. These methods can be used as an individual classifier or as a combination of several methods (ensemble), either homogeneous or heterogeneous, in each iteration [12]. Each method has its own characteristics. It is necessary to analyze the suitability of the classification problem and the characteristics of the method. Furthermore, the performance of selected methods then compared by some indicators of accuracy such as the correctness of categorical prediction, the discriminatory ability and the accuracy of probability prediction [12]. Based on both statistical and machine learning perspectives, these methods can be generally grouped into econometric and artificial intelligence methods, respectively. Some researchers even tried to combine the two methods into a hybrid method, named statistical learning [30].

Classification problems are often not represented by the right model. It has an impact in determining the appropriate method and also in measuring the performance indicators of accuracy. Some of the researchers use various types of classification methods arbitrarily without regard to the characteristics of the method used. One example occurs in gesture recognition problems that work with dynamic data. They treated it per frame as an image recognition problem. In fact, classification in the gesture recognition problem is based on the sequence of frames [27] which belongs to the time series classification category. Measurement of accuracy also becomes less precise because it is calculated based on the classification accuracy per frame. This can be misleading interpretation. In this paper, we focused on the Multivariate Time Series Classification (MTSC) analysis because big data technology is currently able to integrate data from various sources, e.g. sensor devices, in real-time and the calculation process runs efficiently using parallel computing like the Graphics Processing Unit (GPU) computing.

This paper presents a structured and systematic guidance to solve any classification problems. Starting with the introduction of the types of classification problems that depends on the types of data, i.e. static data and dynamic data. Various classification methods are then presented based on each approach, statistics and machine learning, along with their respective characteristics. We summarize the state-of-the-art methods for MTSC analysis, complete with their strengths and weaknesses. Finally, we provide some limitations that need to be completed in the future works and some problems that may arise from the MTSC analysis.

2. Types of problems classification

The complexity of the problems that occur in the real world, requires scientists to classify these problems into several categories in order to facilitate researchers in solving the problems encountered, not least in the classification problem. In this paper, we tried to classify the classification problem into two categories based on the type of data, i.e. static data and dynamic data. We also provide several examples of problems related to each type of data. As for the model in each type of data is classified into a univariate and multivariate model with each of numerical and/or categorical attributes. We propose a hybrid method to complete these models consisting of feature selection and classification stages. The econometric method is always used in the feature selection stage. While at the classification stage, both econometric or artificial intelligence can be used as a method. Thus, the expected results are in accordance with both statistical and machine learning perspectives where the prediction results have a high level of accuracy based on the significant variables in classifying each group. The MTSC which is a multivariate model on the dynamic data is discussed separately in the next chapter.
2.1. Static Data

Static data or commonly known as cross-section data contains a number of information consisting of several variables observed from some objects in a certain period of time. In the classification problem, these variables can be interpreted as numerical and/or categorical attributes. In the univariate model, the characteristics of each group in a categorical variable can be known, or vice versa, we can predict an object with some characteristics will be in which group on that categorical variable. Whereas in the multivariate model, the same results can be obtained for more than one categorical variable simultaneously. Credit rating based on the static personal data is one of the classification problems that can be represented by a univariate model. However, if at the same time that data is used to guess what types of credit products are suitable to be offered to prospective borrowers, then this problem should be represented in the multivariate model. Table 1 shows a sample of methods that can be used to solve the classification model, either the univariate or multivariate model, for static data.

<table>
<thead>
<tr>
<th>Economic Method</th>
<th>Artificial Intelligence Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear discriminant analysis</td>
<td>Genetic algorithm</td>
</tr>
<tr>
<td>(Multinomial) logistic regression model</td>
<td>Artificial neural network</td>
</tr>
<tr>
<td>Hidden-Markov model</td>
<td>Support vector machine</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>Generalized linear vector Quantization</td>
</tr>
<tr>
<td>Random forest</td>
<td>$k$-Nearest Neighbor</td>
</tr>
</tbody>
</table>

2.1. Dynamic Data

In contrast to static data obtained from a certain period of time, several variables from one or more objects, in dynamic data, are observed in a series of time. If observations are made on a particular object, dynamic data is usually called as time series data, while observations on some objects are called as panel data. Similar stochastic data, the observed variables can also be numerical and/or categorical attributes in dynamic data. In addition, model for dynamic data is also still divided into univariate and multivariate models. The classification model with dynamic data is divided into two types, i.e. the classification model with balanced or unbalanced data. Dynamic data is called balanced when observations can be carried out thoroughly on all objects in a series of times. Otherwise, when several variables of each object cannot be full observed in the same timeframe then it is called unbalanced.

The univariate classification model with time series data can model problems such as how to detect level of economic conditions based on its macro and micro economics historical data, e.g. inflation and stock market index. When that problem is developed to detect the level of economic conditions in several countries where observations can be carried out thoroughly in the same timeframe, it can be represented by a univariate classification model with panel (balanced) data. If some countries cannot be observed in a certain series of time (e.g. due to differences in their historical stock exchanges activity) then the appropriate model is the univariate model with panel (unbalanced) data. When detection is carried out not only for the level of economic conditions, but also for the level of Foreign Direct Investment (FDI) simultaneously, the appropriate model is a multivariate classification model that is adjusted to the observed dynamic data. There are several methods that are often used to solve the univariate classification model for dynamic (time series) data, such as hidden-Markov model, periodogram-based metric, dynamic time warping, functional echo state network and shapelet selection algorithm.
3. Multivariate time series classification analysis

Multivariate time series classification (MTSC) has attracted great interest in recent years, in many areas such as economic, finance, gesture recognition and health. Therefore, many researchers in the last decade have tried to examine MTSC, some of which are summarized on Table 2.

<table>
<thead>
<tr>
<th>References</th>
<th>Year</th>
<th>Problem</th>
<th>Type</th>
<th>Method</th>
<th>Strength</th>
<th>Weakness</th>
</tr>
</thead>
<tbody>
<tr>
<td>[32]</td>
<td>2008</td>
<td>A new method for MTSC analysis</td>
<td>Theory</td>
<td>Two-dimensional singular value decomposition</td>
<td>Outperforms one-dimensional singular value decomposition</td>
<td>Less efficient and multivariate time series samples have the equal length</td>
</tr>
<tr>
<td>[33]</td>
<td>2008</td>
<td>A new method for MTSC analysis</td>
<td>Theory</td>
<td>Locality preserving projections</td>
<td>Can be used for multivariate time series samples of different lengths</td>
<td>The dimensionality of the locality preserving projections subspace and the number of nearest neighbors are chosen manually</td>
</tr>
<tr>
<td>[34]</td>
<td>2010</td>
<td>A new method for MTSC analysis</td>
<td>Theory</td>
<td>Combination of discrete support vector machine and fixed cardinality warping distances</td>
<td>Outperforms traditional support vector machines and 1-nearest neighbor</td>
<td>The kernel function in the temporal discrete support vector machine is still linear</td>
</tr>
<tr>
<td>[20]</td>
<td>2015</td>
<td>Drunk driving detection</td>
<td>Applied</td>
<td>Support vector machines</td>
<td>Feasible and effective</td>
<td>Only tested on urban curves, only differentiated between normal state and drunk state, and relatively small sample size</td>
</tr>
<tr>
<td>[25]</td>
<td>2015</td>
<td>Human activity recognition</td>
<td>Applied</td>
<td>Dynamic time warping template selection</td>
<td>High accuracy (more than 80%)</td>
<td>Only recognizing simple activity and there are redundant templates</td>
</tr>
<tr>
<td>[35]</td>
<td>2015</td>
<td>A new method for MTSC analysis</td>
<td>Theory</td>
<td>Parametric derivative dynamic time warping distance</td>
<td>Adaptive to different data sets without showing signs of overfitting</td>
<td>Increase computation time</td>
</tr>
<tr>
<td>[36]</td>
<td>2015</td>
<td>A new method for MTSC analysis</td>
<td>Theory</td>
<td>Imprecise hidden-Markov model</td>
<td>Robust and outperforms other precise and imprecise methods</td>
<td>Only use an interval-valued dissimilarity measure</td>
</tr>
<tr>
<td>[37]</td>
<td>2015</td>
<td>A new method for MTSC analysis</td>
<td>Theory</td>
<td>Reliable early classification</td>
<td>Outperform the state-of-the-art methods in terms of accuracy and earliness</td>
<td>There is no significance checking for feature selection</td>
</tr>
</tbody>
</table>
A new method for MTSC analysis

Features extracted are interpretable and effective early classification

Cannot be performed on the unbalanced multivariate time series data

Combination of recurrent neural network and adaptive differential evolution

More promising, quite robust and adapts well to different data sets without showing signs of overfitting

Parameter values for all data sets are not optimal

Dynamic time warping based on hesitant fuzzy sets (generalized dynamic time warping)

Higher accuracy and lower time-consuming

There is no suggestion about the type of data sets which has an impact on the time consuming and parameter values are set empirically

4. Conclusion and future challenges

This paper produces a structured and systematic guidance containing models, approaches, and methods, especially for MTSC analysis. This paper can be used as a guide to help researchers to solve any classification problems, either static or dynamic data, which is adapted to each characteristic such that the analysis is not carried out by trial-and-error procedures. Some limitations still need to be completed in the guidance related to the interval of parameter values allowed for each method that are adjusted to the classification problems encountered. In this paper, we also ignored a stationary problem in the classification model for dynamic data. Furthermore, we need to explore the appropriate methods for the multivariate classification model with dynamic (panel) data. In addition to time series classification, one of the problems that may arise is how to cluster the time series data, such as the identification of words in a sentence of sign language.

Acknowledgments

This research was supported by PUP 2018 Research Grant, Ministry of Research, Technology and Higher Education, Indonesia to the Gunadarma University No. 010.14/LP/UG/III/2018.

References


[8] Chen M C and Huang S H 2003 Credit scoring and rejected instances reassigning through evolutionary computation techniques Expert Systems with Applications24(4) pp 433-441


[20] Li Z, Jin X and Zhao X 2015 Drunk driving detection based on classification of multivariate time series Journal of Safety Research54 pp 61-67


[34] Orsenigo C and Vercellis C 2010 Combining discrete SVM and fixed cardinality warping distances for multivariate time series classification *Pattern Recognition* **43** pp 3787-3794


[37] Lin Y F, Chen H H, Tseng V S and Pei J 2015 Reliable early classification on multivariate time series with numerical and categorical attributes *Advances in Knowledge Discovery and Data Mining* **9077** pp 199-211


[40] Liu S and Liu C 2018 Scale-varying dynamic time warping based on hesitant fuzzy sets for multivariate time series classification *Measurement* **130** pp 290-297