氏　　名　松江 清高

学位(専攻分野)　博士(情報学)

学 位 記 番 号　総研大甲第 2330 号

学位授与の日付　2022 年 3 月 24 日

学位授与の要件　複合科学研究科 情報学専攻
学位規則第6条第1項該当

学 位 論 文 題 目　A Study on Unsupervised Feature Extraction for Multivariate Time Series

論 文 審 査 委 員　主　査　杉山 麿人
情報学専攻 准教授
宇野 毅明
情報学専攻 教授
水野 貴之
情報学専攻 准教授
速水 謙
国立情報学研究所／総合研究大学院大学
名誉教授
小林 亮太
東京大学 大学院新領域創成科学研究科
准教授

（様式３）

博士論文の要旨

氏名　　松江　清高

論文題目　A Study on Unsupervised Feature Extraction for Multivariate Time Series

It has been a while since Internet of things (IoT) devices that measure various types of events such as temperature, voltage and pressure are used in many systems, and the amount of data collected by such devices is increasing year by year. Accordingly, a variety of services using those collected data have been produced in a lot of fields and utilization of the data has become much more important than ever. Considering an office building as one of the examples, collected data by sensors installed on walls or ceilings, which measure temperature, humidity, or carbon dioxide concentration, can be used in an air conditioning control system or a lighting system. In the case of sensing at multiple locations in a building, the sensors located in the same local area are expected to record similar values. In this situation, if a sensor is broken, different patterns of values from others might be recorded, which indicates that the sensor should be replaced with new one immediately. However, finding such broken sensors is difficult because anomalous behavior of broken sensors may emerge combinatorially together with other healthy sensors, and the combinatorial relationship between sensors must be taken into account. Since those data collected by multiple sensors are in the form of multivariate time series, it is essential to extract features encoding association between multiple time series, and there are heavy demands particularly for industrial fields.

Once data taken by sensors are collected, features extracted from the data that properly takes relationships between multivariate time series into account can be used in various data science applications such as outlier detection and clustering. Since finding useful feature vector representation from time series is one of crucial tasks in those fields, a lot of methods to extract association between time stamps have been developed so far such as Discrete Fourier Transform (DFT), Discrete Wavelet Transform (DWT), and Discrete Cosine Transformation (DCT). These methods are widely used in signal processing fields and the methods are commonly targeted to univariate time series data, that is, they cannot be directly applied to multivariate time series even though they are widely seen in the real-world. Therefore, extraction of features with considering association between multivariate time series remains a challenging task because both time-wise and variable-wise associations should be taken into account. Although some algorithms using machine learning technology like deep learning are becoming popular among outlier detection tasks nowadays, which can implicitly take such time-wise and variable-wise associations into account, they

commonly need ground truth inlier (normal) time series that do not include any outliers (anomalous patterns) to train a model. As one of the examples, there is an autoencoder method that can detect outliers by calculating reconstruction errors. A model made by the autoencoder is expected to correctly decode inliers and wrongly decode outliers if only inliers are used in its training, and the difference of reconstruction errors makes it possible to discriminate outliers from inliers. However, if outliers exist in a training dataset, a model trained by the autoencoder may overlook outliers because they can be also correctly decoded, resulting in the suboptimal performance. To date, only few unsupervised algorithms have been proposed that do not require any labeled time series data, although such unsupervised algorithms are of high importance in practice. To address this issue, we focus on unsupervised feature extraction that can be used for various downstream tasks including outlier detection and clustering.

In this dissertation, we present unsupervised feature extraction algorithms for multivariate time series, called UFEKS (Unsupervised Feature Extraction using Kernel and Stacking) and UFEKT (Unsupervised Feature Extraction using Kernel Method and Tucker Decomposition). UFEKS (1) constructs a kernel matrix for the set of subsequences from each time series and (2) concatenates all matrices horizontally. Feature representation is obtained as row vectors in the concatenated matrix in a fully unsupervised manner, which can be used in subsequent machine learning problems. Likewise, UFEKT (1) constructs a kernel matrix from subsequences of each time series to account for time-wise association and (2) constructs a single tensor by stacking the kernel matrices and performs Tucker decomposition to account for variable-wise association. Tucker decomposition is one of the well-known tensor decomposition techniques and it decomposes the constructed tensor of a kernel into one core tensor and three factor matrices. Feature representation is obtained as a row vector in one of the decomposed factor matrices. In the decomposition process, ranks of a tensor must be given as one of the hyper-parameters. Although finding the best values of hyper-parameters in an unsupervised learning is known as a difficult task, we present an algorithm to find appropriate values of the ranks heuristically. The whole process of UFEKT is also fully unsupervised and can be used for subsequent machine learning tasks.

After we describe our new algorithms in detail, we empirically evaluate our algorithms and show experimental results in two tasks of outlier detection and clustering. Nine synthetic and six real-world datasets are used for outlier detection and 102 real-world datasets are used for clustering. Our methods are compared with two well-established existing feature extraction methods, the subsequence-based method (SS) and the page rank kernel-based method (PRK). Furthermore, we discuss reasons why our algorithms are superior to the existing methods using the principal component analysis (PCA). Finally, we summarize main findings of this dissertation and discuss future work.

Form8 : Separate Sheet
（様式８・別紙１）

Results of the doctoral thesis screening

博士論文審査結果

Name in Full
氏 名　　松江　清高

Title
論文題目　　A Study on Unsupervised Feature Extraction for Multivariate Time Series

本学位論文は、「A Study on Unsupervised Feature Extraction for Multivariate Time Series」と題し、多変量の時系列データに対する汎用的な特徴抽出手法に関する成果を述べている。複数の時系列からなる多変量時系列に対するデータ解析は、変数間の関連と時間軸方向の関連という 2 つの関連を同時に捉えてモデリングする必要があり、本質的に困難な課題である。そこで本論文では、これら 2 つの関連を同時に捉える特徴抽出を実施することで、多変量時系列を特徴ベクトル集合へ埋め込む手法を考案している。提案手法を用いて特徴ベクトルを得ることで、実数値ベクトルを対象とする既存の多様な機械学習手法がそのまま適用可能となるため、多変量時系列のデータ解析を簡便に実施することが可能となる。本論文では、2 つの特徴ベクトル抽出手法を提案するとともに、代表的な教師なし学習タスクである外れ値検出とクラスタリングにおいて提案手法の性能を検証し、その有用性を示している。本学位論文は英語で執筆されており、全 6 章から構成されている。

第 1 章「Introduction」では、研究の背景として多変量時系列データ解析問題を導入し、例題として複数センサーからの外れ値検出を紹介することで、多変量時系列データ解析が抱える課題である組合せ的に現れる変数間の関連と時間軸方向の関連を同時に取り込むことの困難さについてまとめている。さらに、この課題解決のための提案手法について、その概要を説明している。

第 2 章「Related Work」では、関連研究について議論している。まず時系列データに対する特徴抽出の既存手法について議論した後に、本論文で実施している教師なし学習タスクである、多変量時系列データに対する外れ値検出とクラスタリングについての関連技術について議論している。特に、教師なし外れ値検出の技術発展が進んでいないことを指摘している。教師なし学習では、一切のクラスラベルが使用できないことを前提とするが、時系列データに対する多くの外れ値検出手法において、外れ値が混入していない正常データが訓練データとして入手可能であることを仮定しており、これは厳密には教師なし学習ではない。これに対して、提案手法は完全な教師なし学習手法であり、訓練データを蓄積することなくすぐに適用できるという利点がある。

第 3 章「Algorithms」では、2 つの新規提案手法である UFEKS（Unsupervised Feature Extraction using Kernel and Stacking）と UFEKT（Unsupervised Feature Extraction using Kernel Method and Tucker Decomposition）を提案している。UFEKS は、各時系列に対する部分時系列への分割、部分時系列間の類似度に基づくカーネル行列の作成、そして時系列ごとに作成したカーネル行列の列方向連結、という手順によって、変数間及び時間軸方向の関連を取り込んだ特徴ベクトルを獲得する手法である。また、UFEKS にお

ける列方向連結によって、得られる特徴ベクトルが高次元になってしまうという問題を解決するために、UFEKT では、列方向の連結をテンソル作成へと置き換え、テンソル分解を実行し、因子行列から特徴ベクトルを獲得する。

第 4 章「Outlier Detection」では、外れ値検出タスクを用いて提案手法 UFEKS と UFEKT の性能を検証している。外れ値検出を定式化し、提案手法で構築した特徴ベクトルに対して標準的な外れ値検出手法である *k*NN（*k*th Nearest Neighbor）や LOF（Local Outlier Factor）、one-class SVM などを適用することで、外れ値検出を実施している。人工データ及び実データを用いた数値実験によって、部分時系列から直接特徴ベクトルを作る手法や、カーネル行列を作成したあとに時系列間で総和を取る既存手法と比較して、提案手法が優れた性質を持つことを示している。さらに、PCA を使って提案手法がもつ性質を実験的に考察するとともに、特徴ベクトル間の距離に着目した理論的な考察を実施している。

第 5 章「Clustering」では、クラスタリングを用いて提案手法 UFEKT の性能を検証している。クラスタリングを定式化し、提案手法で構築した特徴ベクトルに対して標準的なクラスタリング手法である *K*-means や DBSCAN などを適用することで、クラスタリングを実施している。実データを用いた実験において、特に DBSCAN との相性が良く、既存手法と比較して提案手法が優れた性質を持つことを示している。

最後に、第 6 章「Conclusion」で本論文の貢献をまとめ、提案手法の限界や欠点、そして今後の課題や展望について述べている。

公開発表会では、博士論文の章立てに従って発表がおこなわれ、その後におこなわれた論文審査会及び口述試験では、審査員からの質疑に対して適切に回答がなされた。質疑応答後に審査委員会を開催し、審査委員で議論をおこなった。審査委員会では、出願者が情報学分野の十分な知識と研究能力を持つと認められるとともに、博士研究が多変量時系列データ解析において十分な新規性を有しており、かつ学術的にも優れた貢献であることが評価された。

以上を要するに本学位論文は、多変量時系列データを解析するためのシンプルかつ効果的な手法を提案しており、そのシンプルさ故に時系列の慎重なモデリングが不要となるため、汎用性が高く、実応用における時系列データ解析の適用範囲を大幅に広げることが期待される。また、本学位論文の成果は、学術雑誌論文 1 件、フルペーパー査読付き国際会議論文 1 件として発表され、学術的な貢献も認められる。以上の理由により、審査委員会は、本論文が学位の授与に値すると判断した。