JAIST Repository
https://dspace.jaist.ac.jp/

| | |
|---|---|
| Title | 解釈可能な機械学習とアプリケーションに関する研究 |
| Author(s) | NGUYEN, Thanh Phu |
| Citation | |
| Issue Date | 2021-03 |
| Type | Thesis or Dissertation |
| Text version | ETD |
| URL | http://hdl.handle.net/10119/17464 |
| Rights | |
| Description | Supervisor:Huynh Nam Van, 先端科学技術研究科, 博士 |

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

| | | | | |
|---|---|---|---|---|
| 氏　　名 | NGUYEN Thanh Phu | | | |
| 学　位　の　種　類 | 博士(知識科学) | | | |
| 学　位　記　番　号 | 博知第 279 号 | | | |
| 学位授与年月日 | 令和 3 年 3 月 24 日 | | | |
| 論　文　題　目 | A Study on Interpretable Machine Learning and Applications | | | |
| 論文審査委員 | 主査 | HUYNH Van Nam | 北陸先端科学技術大学院大学 | 教授 |
| | | 藤波　努 | 同 | 教授 |
| | | DAM Hieu Chi | 同 | 教授 |
| | | 由井薗　隆也 | 同 | 准教授 |
| | | 領家 美奈 | 筑波大学 | 准教授 |

**論文の内容の要旨**

Machine learning and data mining techniques have been developed rapidly in recent times. In tasks such as classification, machine learning techniques have been shown to equal to and even surpass human performance. However, high-performance models are usually complex, opaque and have low interpretability thus making it difficult to explain the underlying behaviors of those models that lead to the final outcomes. In many domains such as healthcare, medicine or finance, interpretability is one of the most important factors when considering the adoption of those models. In our research, we aim to develop transparent machine learning models that are not only able to provide users with knowledge about the underlying data but also still can achieve competitive performance compared with other commonly used techniques.

Firstly, in the field of unsupervised learning, clustering is a fundamental task that has been utilized in many scientific fields. Clustering groups data into clusters. For each cluster, objects in the same cluster are similar between themselves and dissimilar to objects in other clusters. The *k*-means is a popular interpretable method for the clustering task. However, it suffers from the problem of underfitting data with simple dissimilarity measures - a key part in formulating clusters of the *k*-means. In our work, we proposed a new *k*-means-based clustering method with a novel dissimilarity measure that can better fit with the underlying data. The effectiveness of the proposed clustering algorithm is proven by a comparative study conducted on popular clustering methods for categorical data.

Specifically, one inherent limitation of the *k*-means technique is its data type constraint, as the *k*-means technically can only work with the numerical data type. Several attempts have been made in order to remove the numeric data only limitation of *k*-means to make it applicable to clustering for categorical data. However, the measures to quantify the dissimilarity/similarity for categorical values

are still not well-understood because there is no coherent metric available between categorical values thus far. Especially, most previous works have unfortunately neglected the semantic information potentially inferred from relationships among categories. In this work, we proposed a new clustering algorithm that is able to integrate those kinds of information into the clustering process for categorical data. Specifically, the new categorical clustering algorithm takes account of the semantic relationships between categories into the dissimilarity measure.

Secondly, for the field of supervised learning, applying deep learning techniques could bring higher accuracy when dealing with big and heterogeneous data. However, such high accuracy comes with high complexity and opaqueness in the models. This situation leads to the difficulty of interpretability of those models - one of the important and required properties when implementing them within a decision support system, especially in healthcare, medicine and domains which require transparency for high-stake decisions. Due to the need for high-performance interpretable machine learning models, especially for healthcare applications, we proposed a binary classifying system named GSIC (GSOM-based Interpretable Classifying System) that is based on a systematic combination of unsupervised and supervised machine learning techniques.

GSIC is a two-stage binary classification system that is applicable for healthcare (or general) data. It benefits from a high level of interpretability and can at the same time achieve results comparable to commonly used classification techniques. The motivation behind the proposed system is the lack of effective classification methods for handling data generated by various distributions (such as healthcare or banking data) that can harmonize both performance and interpretability perspectives. In the proposed system, GSOM (The Growing Self-Organizing Map) plays a key role to help overcome the curse of dimensionality problem as well as improve efficiency and interpretability by analyzing its generated mapping results. GSOM is selected as a popular dimensional reduction and visualization method that has the advantages of dynamically learning new data representation and the capability of revealing salient relations between objects from underlying data contexts. In other to evaluate the performance of our proposed system, an experiment on the classification task is conducted. Furthermore, a use case on the specific data of sepsis patients in the Intensive Care Unit (ICU) is demonstrated in order to prove the merit of our proposed system.

On the other hand, we also realized the limitation of our proposed classification system when dealing with uncertainty. Specifically, real data with high uncertainty and ambiguity is challenging for the classification task. In related literature, E-KNN is a popular evidence theory-based classification method developed for handling uncertainty data. However, as a distance-based technique, it also suffers from the problem of high dimensionality as well as mixed distributed data where closed data

points originated from different classes. Based on that motivation, we enhanced our proposed classification system GSIC with the capability of handling the uncertainty existing in the underlying data. The classification experiment conducted on various real data and popular classifiers has shown that the proposed technique has competitive results compared with state-of-the-art methods.

Specifically, more efforts have been put in to remedy the above-mentioned problems by proposing a new classification method that can induce evidence about the classes of new data objects based on groups of data that belong to various distributions. By assuming that a data set contains instances that are generated by several different distributions, data generated by each distribution can be represented in the form of heterogeneous clusters. Each distribution has different rules for characterizing the classes of its generated data. For each cluster, we gauge the behaviors of the data distribution by using decision trees on the whole set of data belonging to that cluster. Results from those decision trees could be considered as evidence for determining the classes of new data points. For making a final decision about the predicted class, Dempster's combination rule is used to fuse the evidence collected from previous steps.

**Keywords:** Interpretable Machine Learning, Explainable Artificial Intelligence, *K*-means-based Clustering, Classification, Dempster–Shafer Theory.

## 論文審査の結果の要旨

Recently, machine learning (ML) models have shown to be competitive with or even surpass human performance in some tasks such as classification. However, such high-performance ML models like deep neural networks are usually complex and work as black-box models which don't provide clear interpretations of their decisions. In such application domains as medicine and healthcare, interpretability of ML models is one of the most important issues as humans need to trust the decisions made by these models when relying on them. In this research, the main objective is therefore to develop transparent ML models that are not only able to provide users with interpretations of their decisions but also achieve competitive performance compared with other commonly used techniques. The main results of this research are summarized as follows.

In the area of unsupervised learning, this research proposes a new *k*-means-based clustering method for categorical data with a novel similarity measure that incorporates not only the distributions of categories but also their relationship information into the quantification of similarity between data objects. The effectiveness of the proposed clustering algorithm is proven by a comparative study conducted on popular clustering methods for categorical data. In the area of supervised learning, it proposes a two-stage binary classification system named GSOM-based Interpretable Classifying System (GSIC, for short) that is applicable for healthcare (or general) data. GSIC has a high level of interpretability and at the same time can achieve the results comparable to commonly used classification techniques. The motivation behind the proposed system is the lack of effective classification methods for handling data generated by various

distributions (such as in healthcare or banking data) that can harmonize both performance and interpretability perspectives. GSIC was experimentally tested with a use case in sepsis patients staying in ICU to show its efficiency and effectiveness. In addition, the proposed classification system GSIC was also enhanced with the capability of handling the uncertainty existing in the underlying data. The experimental study conducted on various real data sets and popular classifiers has shown that the developed system provided competitive results compared to the state-of-the-art methods.

This dissertation has made good contributions to methodological and experimental developments within the area of interpretable machine learning. The research work presented in this dissertation has resulted in one journal paper and three refereed conference papers.

In summary, Mr. NGUYEN Thanh Phu has completed all the requirements in the doctoral program of the School of Knowledge Science, JAIST and finished the examination on February 3, 2021, all committee members approved awarding him a doctoral degree in Knowledge Science.