氏　　名　斎藤 侑輝

学位(専攻分野)　博士(統計科学)

学 位 記 番 号　総研大甲第 2184 号

学位授与の日付　2020 年 9 月 28日

学位授与の要件　複合科学研究科　統計科学専攻
学位規則第6条第1項該当

学 位 論 文 題 目　Multiple Data Matching by Modeling Data Structures

論 文 審 査 委 員　主　査　教授　日野　英逸

准教授　持橋　大地

教授　福水　健次

特別研究員　岩田　具治
NTT コミュニケーション科学基礎研究所

（様式３）

博士論文の要旨

氏　　名　斎藤　侑輝

論文題目　Multiple Data Matching by Modeling Data Structures

With the development of information technology in recent years, there is a need to apply machine learning to a wide range of industrial and academic fields, and emerging services and applications have started requiring matching multiple groups of data, namely, multiple data matching. Through the multiple data matching, we can investigate or infer the relationship between groups of data, such as common cluster structures or links. Furthermore, modeling the structure of the data is known as required to construct methods in some scenarios; however, several emerging use-cases are not well-studied, in which both modeling the structure of data and combining with new powerful functions (e.g., kernel functions and deep neural networks) is essential to match the multiple data.

This thesis considers matching up heterogeneous groups of objects by modeling the structures of data, involving various real-world applications. The problem scenarios divide into two cases: building methods to match (i) clusters, where the cluster structure commonly lies in heterogeneous domains, and (ii) two heterogeneous sets, where correct pairs are given as supervised information. This thesis focuses on extending the problems of multiple data matching onto the two different directions above.

(i) In the first case, we study a so-called supervised clustering to match common clusters that exist across two different domains, using given cluster assignments on one side of the domains as supervised information. The proposed method maximizes the similarity between the cluster structures within two domains in which kernel mean embeddings represent each cluster as probability distribution uniquely and nonparametrically. In the experiments, we use the datasets from meteoritics and planetary science and investigate taxonomical matching between the meteorites and asteroids. Here, the datasets consist of reflectance spectra of asteroids and meteorites, and also major chemical compositions of meteorites, where cluster assignments of the meteorites are known as the abovementioned supervised information, and the problem is to solve supervised clustering on the asteroidal domain. By comparing the clustering accuracy of the asteroid between with and without the guidance of the meteorite, we observe that the guidance of meteorite taxonomy improves the accuracy, either with the reflectance spectra or major chemical compositions of meteorites. This fact serves as a piece of evidence that there is a common taxonomic structure and links between meteorites and asteroids, implying a long-standing hypothesis of the taxonomy

matching.

(ii) Second, we investigate heterogeneous set-to-set matching problem building novel deep neural networks. In this case, we are only given paired group data for training and inference, and the learned neural network models must classify whether an unknown paired data matches or not in the inference. The difficulties of the heterogeneous set-to-set matching are to extract features to match a correct pair of different sets and also preserve two types of exchangeability required for set-to-set matching: the pair of sets, as well as the items in each set, should be exchangeable. In this study, we propose a deep learning architecture for heterogeneous set-to-set matching to address the abovementioned difficulties. The proposed framework includes two novel modules: (1) a cross-set feature transformation (CSeFT) module and (2) cross-similarity (CS) function. The former provides the exchangeable set feature based on the interactions between two sets in intermediate layers, and the latter performs the exchangeable set matching by calculating the cross-feature similarity of items between two sets. Furthermore, we propose a novel loss function, K-pair-set loss, to train our model effectively. The effectiveness of our approach is demonstrated in two real-world applications. First, we consider fashion set recommendations via matching fashion outfits, where provided examples of the outfits are used as correct combinations of items. Since the paired sets include images of different fashion items, we regard this case as heterogeneous set matching. Next, we evaluate our methods through group re-identification experiments using two datasets, a new extension of the Market-1501 dataset (Market-1501 Group) and the Road Group dataset. Considering group membership change, we regard group re-identification as a heterogeneous set matching problem. In the experiment, we further introduce the novel data augmentation method that augments paired data (set-data augmentation). In these experiments, we show that the proposed method provides significant improvements and results compared with the state-of-the-art methods, thereby validating our architecture for the heterogeneous set matching problem.

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　斎藤　侑輝

Title
論文題目
Multiple Data Matching by Modeling Data Structures

[論文の概要]
博士出願論文は，提出された論文は，データのグループ間のマッチングに対する教師付き学習の方法を論じたもので，英文で書かれており全４章と引用文献の計 81 頁からなる．

１章は本論文の序章である．データのグループ間のマッチング問題が説明され，関連する過去の研究が述べられた後，本論文で論じる２つの課題として，関連する異種ドメインのクラスタ構造が補助情報として与えられた場合のクラスタマッチングの問題と，多種のデータを含んだ集合をマッチングする問題が説明されている．

２章では，２つの異なるドメインのクラスタマッチングの方法を提案し，小惑星と隕石の分類体系のマッチングへ応用している．個々のデータのマッチングが存在しない状況で，一方のデータのクラスタ構造がわかっているときに，他方のドメインのクラスタリングと，２ドメイン間のクラスタのマッチングを同時に行うアルゴリズムを提案している．提案手法を，隕石の反射スペクトル／元素組成データを教師とした小惑星の反射スペクトルデータのクラスタリング・クラスタマッチング問題に適用し，専門家の分類と整合するクラスタリングが得られること，および小惑星-隕石の対応関係に関する仮説に一致したクラスタマッチングがなされることを確認している．

３章では，任意の個数のアイテムからなる集合が２つ与えられたときに，集合間の類似性を求める深層学習の方法を提案している．２つの集合が適合するかを判定することを目的としており，本論文では，与えられたファッションアイテム群とマッチする他のアイテム群を推薦する問題と，多数の人物が撮影された２つの画像データベースの中から，同じ人物からなるグループを発見するグループ再同定問題に応用している．アイテム間の置換および集合の置換に関する不変性を満足しながら高い性能を持たせるために，内積に基づく注意機構または類似度計算を集合間に跨るように発展させた集合間特徴変換層（Cross-set feature transformer）を多層に用い，最終層で２つの集合に関して対称な集合間類似度を計算する深層ネットワーク構造とその学習法を提案している．提案法を，上述のファッションアイテム推薦問題とグループ再同定問題に適用し，Set Transformer，BERT，グラフニューラルネットなどの既存手法と比較したところ，ファッションアイテム推薦問題に対して提案手法は既存手法を大きく上回る性能を，グループ再同定問題では既存手法と同等以上の性能を持つことが示されている．

４章は論文のまとめである．

[論文の評価]

データのグループ間マッチング問題に対し，本論文は，従来十分な研究がなされていなかった補助的情報を有効に用いる方法を，クラスタ情報が与えられる場合と，マッチングの正例が与えられる場合に対してそれぞれ提案し，実問題に対してその有効性を示しており，統計科学の博士論文として十分な意義を持つと考える．なお，2章の内容をまとめた論文が査読付き国際学術雑誌Meteoritics & Planetary Science (Wiley刊)に採択されている．