AFRL-AFOSR-JP-TR-2019-0062

Integration of Clustering with Semantics Learning for Massive Categorical and Mixed Data

Van Nam Huynh
JAPAN ADVANCED INSTITUTE OF SCIENCE AND TECHNOLOGY
1-1, ASAHIDAI
NOMI, ISHIKAWA, 923-1211
JP

11/18/2019
Final Report

Air Force Research Laboratory
Air Force Office of Scientific Research
Asian Office of Aerospace Research and Development
Unit 45002, APO AP 96338-5002

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
18-11-2019	Final	16 Aug 2017 to 15 Aug 2019

4. TITLE AND SUBTITLE
Integration of Clustering with Semantics Learning for Massive Categorical and Mixed Data

5a. CONTRACT NUMBER

5b. GRANT NUMBER
FA2386-17-1-4046

5c. PROGRAM ELEMENT NUMBER
61102F

6. AUTHOR(S)
Van Nam Huynh

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
JAPAN ADVANCED INSTITUTE OF SCIENCE AND TECHNOLOGY
1-1, ASAHIDAI
NOMI, ISHIKAWA, 923-1211 JP

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AOARD
UNIT 45002
APO AP 96338-5002

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/AFOSR IOA

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-AFOSR-JP-TR-2019-0062

12. DISTRIBUTION/AVAILABILITY STATEMENT
A DISTRIBUTION UNLIMITED: PB Public Release

13. SUPPLEMENTARY NOTES

14. ABSTRACT
The PI has had very good performance with this grant. The main objective of this research was to conduct a systematic study of data-driven similarity measures based on information theory and kernel-based methods for representation of cluster centers for categorical objects so as to ultimately develop a k-means like clustering methodology capable of handling missing data for categorical and mixed datasets. Firstly, the PI has proposed a new unsupervised similarity measure for categorical data based on the information theoretic approach. Secondly, based on the newly developed similarity measure for categorical data, they have proposed a novel k-means like clustering framework making use of kernel-based methods for representation of cluster centers. Thirdly, they also developed the so-called k-CCM algorithm for clustering categorical data with missing values. Finally, they have further extended the proposed k -means like clustering framework so as to make it applicable for clustering mixed numeric and categorical datasets with missing data. The PI has had 3 journal papers and 7 conference/workshops as a direct result of this research grant. There was one graduate student supported by this research grant.

15. SUBJECT TERMS
clustering, categorical and mixed data sets

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			CHEN, JERMONT
Unclassified	Unclassified	Unclassified	SAR		19b. TELEPHONE NUMBER *(Include area code)* 315-227-7007

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18

Final Report for AOARD Grant FA2386-17-1-4046

"Integration of Clustering with Semantics Learning for Massive Categorical and Mixed Data"

12 November, 2019

Name of Principal Investigator: Van-Nam HUYNH

- e-mail address: huynh@jaist.ac.jp
- Institution: Japan Advanced Institute of Science and Technology
- Mailing Address: 1-1 Asahidai, Nomi, Ishikawa, JAPAN
- Phone: +81-761 51 1791
- Fax: +81-761 51 1149

Name of Co-Investigator: Canh-Hao NGUYEN

- e-mail address: canhhao@kuicr.kyoto-u.ac.jp
- Institution: Kyoto University
- Mailing Address: Gokasho, Uji, Kyoto, 611-0011, JAPAN
- Phone: +81-774-38-3024

Name of Co-Investigator: Sadaaki Miyamoto

- e-mail address: miyamoto.sadaaki.fu@u.tsukuba.ac.jp
- Institution: Tsukuba University
- Mailing Address: 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8577, JAPAN

Period of Performance: 15/08/2017 – 15/08/2019

Abstract:

This research project aims to develop a novel methodology for integration of clustering with semantics learning that enable us to discover and exploit hidden semantics of data while effectively learning clusters from massive categorical/mixed data sets. Specifically, we focused on investigating information theoretic-based measures of similarity and kernel-based methods for representing cluster means to reflect semantic information in the clustering process. We have proposed a new similarity measure based on information theoretic approach that could be able to integrate semantic information into the quantification of the similarity between categorical data. We have also developed, based on kernel smoothing techniques, kernel-based methods for representation of cluster means for categorical objects. Such kernel-based representation methods could provide an interpretation of cluster means being consistent with the statistical interpretation of the cluster means for numerical data. Eventually, the new similarity measure and kernel-based methods for representation of cluster means have been integrated into a novel *k*-means like clustering framework for categorical and mixed data. Within the proposed clustering framework, we have further developed a so-called *k*-CMM algorithm for clustering mixed numeric and categorical datasets with missing values, which incorporates the imputation procedure into the clustering framework so as to improve clustering results with missing data. A series of experimental studies have been

comprehensively conducted on real datasets from UCI Machine Learning Repository to evaluate the proposed clustering framework as well as illustrate its applicability.

Introduction

Clustering is one of fundamental operations in data mining and machine learning. Clustering has been widely applied in a variety of fields (Tellaroli et al., 2016): ranging from medical sciences, economics, computer sciences, engineering to social sciences and earth sciences. There is a vast body of knowledge in the area of clustering and there has been attempts to analyze and categorize them for a large number of applications (Berkhin, 2002; Xu & Wunsch, 2005; Fahad et al., 2014; Xu & Tian, 2015). However, clustering of massive data sets with categorical and mixed-type data is still a challenge in many applications of big data mining.

According to (LeCun, Bengio & Hinton, 2015), unsupervised learning will become far more important in the longer term, because labeling data is both costly and time consuming, and sometimes impossible especially in the context of big data. As an important branch in unsupervised learning, clustering has recently emerged as an active research topic in big data mining. Particularly, it can be considered as an important tool for analyzing the massive volume of data generated by modern applications: having big data clustered/grouped in a compact format that is still an informative version of the entire data (Fahad et al., 2014). It is also argued that human and animal learning is largely unsupervised.

There are difficulties for applying clustering techniques to big data due to new challenges that are raised with big data (Shirkhorshidi et al., 2014). Especially, big data mining adds to clustering the complications of very large datasets with many attributes of different types (Berkhin, 2002).

The classical k-means algorithm (MacQueen, 1967) is probably the most popular and widely used clustering technique. Numerous *k*-means like algorithms have been also proposed in the literature; each of which typically uses different similarity measure (Kogan et al., 2005). The *k*-means like algorithms are easy to implement and also efficient for handling large data sets. However, working only on numerical data prohibits k-means like algorithms from being used for clustering categorical data.

During the last decade or so, several attempts have been made in order to remove the numeric data only limitation of *k*-means algorithm so as to make it applicable to clustering for categorical data, such as *k*-modes algorithm (Huang, 98), *k*-representative algorithm (San, Huynh, Nakamori, 2004), modified *k*-means algorithm (Ahmad and Dey, 2007), k-centers algorithm (Chen and Wang, 2013) and *k*-means like algorithm (Nguyen and Huynh, 2016). While these *k*-means based algorithms use a similar clustering procedure to the *k*-means algorithm, they are different in the way of defining "cluster center" or "similarity measure" for categorical data.

In this research project we focused on the problem of clustering massive categorical and mixed data sets. The main objective of this research is to conduct a systematic study of data-driven similarity measures based on information theory and kernel-based methods for representation of cluster centers for categorical objects so as to ultimately develop a *k*-means like clustering methodology capable of handling missing data for categorical and mixed datasets. We have also conducted a series of experiments tested on real datasets from UCI Machine

Learning Repository to evaluate the performance of the proposed clustering framework against other previously developed clustering methods.

Results and Discussion:

Firstly, we have proposed a new unsupervised similarity measure for categorical data based on the information theoretic approach. The new proposed measures could be able to integrate the information of relations between attributes through co-occurrence content. In order to evaluate the new measures, experiments were conducted to compare its effectiveness with other categorical similarity measures. The results have shown that the new proposed measures have a competitive performance comparing to the others especially when handles with highly correlated data sets.

Secondly, based on the newly developed similarity measure for categorical data, we proposed a novel *k*-means like clustering framework making use of kernel-based methods for representation of cluster centers. Essentially in the proposed clustering framework it allows us to formulate the problem of clustering categorical data in the fashion similar to *k*-means clustering, while kernel-based representation of centers also provides an interpretation of cluster means being consistent with the statistical interpretation of the cluster means for numerical data. Also, the new similarity measure based on the information theoretic approach allows us to integrate not only the distributions of categories but also their relationship information into the quantification of similarity between data objects. Within this framework, a class of *k*-means like algorithms for clustering categorical data have been experimentally implemented and tested. The experiments have shown that the proposed clustering algorithm has a competitive performance when compared to other popular used clustering methods for categorical data.

Thirdly, we also developed the so-called *k*-CCM algorithm for clustering categorical data with missing values. The proposed algorithm *k*-CCM integrates imputation step and clustering into a common process. By this way, all incomplete objects are first imputed and then assigned into appropriate clusters. In particular, we have extended a decision tree-based imputation method to fill in missing values and then, for clustering, we used a kernel density estimation approach to define cluster centers and an information theoretic-based dissimilarity measure to quantify the differences between objects. An extensive experimental evaluation is conducted on benchmark categorical datasets to evaluate the performance of the *k*-CCM algorithm. According to the experimental results, the designed algorithm has a comparative result in terms of clustering quality when compared to other five algorithms showing that the imputation step has improved the quality of the clustering.

Finally, we have further extended the proposed *k*-means like clustering framework so as to make it applicable for clustering mixed numeric and categorical datasets with missing data. In particular, we developed a so-called *k*-CMM algorithm for clustering mixed numeric and categorical datasets with missing values by integrating the imputation step into the proposed clustering framework. In the imputation step, it first uses the decision-tree based method to find the set of correlated objects and then uses the IS and MCS measures to search for possible imputed values from the correlated set to impute for missing values in categorical attributes, while the missing values in numeric attributes are imputed using the mean of corresponding attributes from the correlated set. As for the clustering step, *k*-CMM uses the kernel density estimation approach and the mean to define cluster centers at categorical and numeric attributes, respectively. In addition, to quantify the proximity between

data objects, it uses the squared Euclidean and the information-theoretic based dissimilarity measures for numeric and categorical attributes, respectively. Experimental results have shown that *k*-CMM is more efficient than Huang's *k*-prototypes algorithm in terms of clustering quality in most cases. The particular case of *k*-CMM for clustering purely categorical datasets was also evaluated with the other five state-of-the-art clustering algorithms in terms of clustering quality. Experimental results indicated that the decision-tree based method and measures used for categorical attributes can enhance clustering results. Generally, *k*-CMM has a comparative performance in terms of Purity, Normalized Mutual Information (NMI) and Adjusted Rand Index (ARI). Moreover, we also evaluated the runtime, memory consumption, and scalability of *k*-CMM. The results obtained show that *k*-CMM is scalable with respect to the number of instances. It would be also extendable by designing a parallel clustering algorithm for efficiently handling large-scale and high-dimensional mixed datasets with missing values.

As for the future work, we will focus on a hybrid approach that is to uncover the semantic information and to cluster data simultaneously. This approach would allow to learn the semantic information that is consistent with clustering solutions. We will also apply the developed clustering methods and tools to practical applications including chemical compounds analysis, DNA sequences analysis, social networks analysis, and patient similarity for treatment regimen discovery.

The results obtained from this research project have been published, or submitted for publication, in the following papers. One PhD thesis has been completed within this research project. In addition, two journal papers are being written based on the work supported by this research project and it is anticipated that they will be submitted for publication shortly.

List of Publications (that resulted from this AOARD supported project):

a) Journal Papers:

[1] Thu-Hien Thi Nguyen, Duy-Tai Dinh, Songsak Sriboonchitta, Van-Nam Huynh. A Method for *k*-Means-Like Clustering of Categorical Data, *Journal of Ambient Intelligence and Humanized Computing* (accepted). DOI: 10.1007/s12652-019-01445-5 **(attachment)**

[2] Duy-Tai Dinh, Songsak Sriboonchitta, Van-Nam Huynh. Clustering Mixed Numeric and Categorical Data with Missing Values, *Information Sciences* (submitted for publication on February 12, 2019). **(attachment)**

[3] Duy-Tai Dinh, Van-Nam Huynh. *k*-PbC: An Improved Cluster Center Initialization for Categorical Data Clustering, *Applied Intelligence* (submitted for publication on September 25, 2019). **(attachment)**

b) Conference Papers:

[4] Thanh-Phu Nguyen, Duy-Tai Dinh, Van-Nam Huynh. A New Context-Based Clustering Framework for Categorical Data. **PRICAI 2018**: Trends in Artificial Intelligence - *15th Pacific Rim International Conference on Artificial Intelligence*, Nanjing, China, August 28-31, 2018, Springer-Verlag, pp. 697-709. **(attachment)**

[5] Thanh-Phu Nguyen, Mina Ryoke, Van-Nam Huynh. A New Context-Based

Similarity Measure for Categorical Data Using Information Theory. **IUKM 2018:** *Integrated Uncertainty in Knowledge Modelling and Decision Making - 6th International Symposium*, Hanoi, Vietnam, March 15-17, 2018, Proceedings, Springer-Verlag, pp. 114-125. **(attachment)**

[6] Sadaaki Miyamoto, Van-Nam Huynh, Shuhei Fujiwara. Methods for Clustering Categorical and Mixed Data: An Overview and New Algorithms. **IUKM 2018:** *Integrated Uncertainty in Knowledge Modelling and Decision Making - 6th International Symposium*, Hanoi, Vietnam, March 15-17, 2018, Proceedings, Springer-Verlag, pp. 75-86.

[7] Sadaaki Miyamoto, Jong Moon Choi, Yasunori Endo, Van-Nam Huynh. Optimal Clustering with Twofold Memberships. **MDAI 2018:** *Modeling Decisions for Artificial Intelligence - 15th International Conference*, Mallorca, Spain October 15-18, 2018, Springer-Verlag, pp. 221-231.

[8] Duy-Tai Dinh, Van-Nam Huynh. *k*-CCM: A Center-Based Algorithm for Clustering Categorical Data with Missing Values. **MDAI 2018:** *Modeling Decisions for Artificial Intelligence - 15th International Conference*, Mallorca, Spain October 15-18, 2018, Springer-Verlag, pp. 267-279. **(attachment)**

[9] Duy-Hung Nguyen, Van-Nam Huynh. Learning Individual and Group Preferences in Abstract Argumentation. **PRICAI 2019:** *Trends in Artificial Intelligence - 16th Pacific Rim International Conference on Artificial Intelligence*, Cuvu, Yanuca Island, Fiji, August 26-30, 2019, Proceedings, Part I., Springer-Verlag, pp. 704-717.

[10] Duy-Tai Dinh, Tsutomu Fujinami, Van-Nam Huynh. Estimating the Optimal Number of Clusters in Categorical Data Clustering by Silhouette Coefficient. **KSS 2019:** Knowledge and Systems Sciences - 20th International Symposium, Da Nang, Vietnam, November 29 – December 1, 2019, Springer-Verlag, pp. 1-17.

c) PhD Thesis

[11] Duy-Tai Dinh. *Effective Clustering Algorithms for Categorical and Mixed Data*. PhD Dissertation, Oct. 2019. (Supervisor: Prof. Van-Nam Huynh)