AN ENSEMBLE DATA SUMMARIZATION APPROACH BASED ON FEATURE TRANSFORMATION TO LEARNING RELATIONAL DATA

CHUNG SENG KHEAU

THESIS SUBMITTED IN FULFILLMENT FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

FAKULTI KOMPUTERAN DAN INFORMATIK
UNVERSITI MALAYSIA SABAH
2015

UNIVERSITI MALAYSIA SABAH

BORANG PENGESAHAN STATUS THESIS

JUDUL: AN ENSEMBLE DATA SUMMARIZATION APPROACH BASED ON FEATURE TRANSFORMATION TO LEARNING RELATIONAL DATA

IJAZAH: DOKTOR FALSAFAH

Saya, Chung Seng Kheau, Sesi Pengajian 2008-2015, mengaku membenarkan tesis Doktor Falsafah ini disimpan di Perpustakaan Universiti Malaysia Sabah dengan syarat-syarat kegunaan seperti berikut:-

1. Thesis ini adalah hak milik Universiti Malaysia Sabah.
2. Perpustakaan Universiti Malaysia Sabah dibenarkan membuat salinan untuk tujuan pengajian sahaja.
3. Perpustakaan dibenarkan membuat salinan tesis ini sebagai bahan pertukaran antara institusi pengajian tinggi.
4. Sila tandakan (/)

☐	SULIT (Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysia seperti yang termaktub di dalam AKTA RAHSIA RASMI 1972)
☐	TERHAD (Mengandungi maklumat TERHAD yang telah ditentukan oleh organisasi/badan di mana penyelidikan dijalankan)
☑	TIDAK TERHAD

Disahkan oleh,

NURULAIN BINTI ISMAIL
LIBRARIAN
UNIVERSITI MALAYSIA SABAH

(Tandatangan Penulis)

(Tandatangan Pustakawan)

Alamat Tetap:

Tarikh : 14 August 2015

(PROF. MADYA DR. RAYNER ALFRED)
Penyelia

DECLARATION

Here, I declared all the research work in this thesis was made by myself to meet the requirements to pass the degree of doctor of philosophy.

14 August 2015

CHUNG SENG KHEAU
PK20088406

CERTIFICATION

NAME : **CHUNG SENG KHEAU**

MATRIC NO. : **PK20088406**

TITLE : **AN ENSEMBLE DATA SUMMARIZATION APPROACH BASED ON FEATURE TRANSFORMATION TO LEARNING RELATIONAL DATA**

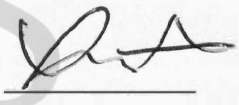
DEGREE : **DOCTOR OF PHILOSOPHY (COMPUTER SCIENCE)**

VIVA DATE : **14th AUGUST 2015**

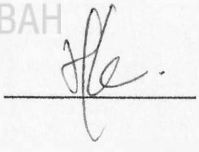
DECLARED BY

1. SUPERVISOR
Assoc. Prof. Dr. Rayner Alfred

Signature



2. CO-SUPERVISOR
Dr. Lau Hui Keng



ACKNOWLEDGEMENTS

To God for the gift of life.

To Assoc. Prof. Dr. Rayner Alfred and Dr. Lau Hui Keng for everything during this PhD.

To my family for the support, love, and understanding.

To my friends for the discussions, ideas, food, and companionship.

To the M.O.H.E of Malaysia, UMS, and SEIT for the financial support.

ABSTRACT

DARA is a framework that is designed particularly to summarize data stored in a multi-relational database having non-target tables associate with the target table. In the process of summarizing data, the data stored in multi-relational database need to be transformed into Term Frequency – Inverse Document Frequency (TF-IDF) vector space. Due to the fact that the size of TF-IDF is directly affected by the number of unique terms that are found in the data stored in target tables, increasing the number of unique terms also increase the clustering complexity and it could produce less accurate clustering results. In this thesis, a Feature Selection algorithm is investigated and proposed to optimize the TF-IDF vector space by selecting only relevant features from the initial TF-IDF vector space. In addition to that, a Feature Construction algorithm is also investigated and proposed to optimize the TF-IDF vector space by merging two or more features in the TF-IDF vector space that best represent the datasets. The Information Gain borrowed from Information Retrieval theory and Term-term Correlation algorithm are used to determine the relevancy of these features to be selected or merged in order to form a new generation of TF-IDF vector space. Consequently, the size of the TF-IDF vector space is reduced. This will indirectly minimize the complexity the TF-IDF vector space that makes the clustering work more efficient while trying to maintain or improve its clustering result accuracy. A genetic algorithm (GA) is also used to find the best centroids for all the clusters generated cluster centroids. A ensemble clustering is designed, used and evaluated to generate the final classification framework that will take all input generated from the GA based clustering with Feature Selection and Feature Construction algorithms and perform the classification task for the relational datasets. Several experiments have been conducted to evaluate the predictive performance of a classification task (C4.5 classifier) when using these clusters results on several relational datasets from mutagenesis, financial and hepatitis databases. The experimental results obtained show some improvements on the predictive accuracy tasks when using the clustering results obtained. Finally, there are further improvements shown when a GA is applied to the whole framework of the classification task by using the WEKA C4.5 classifier and taking the predictive accuracy as the fitness function. The experiment result shows that the ensemble clustering shows a good sign that indicates the consensus function works correctly. This study shows the task of optimizing the TF-IDF vector space by reducing the number of features in TF -IDF vector space increases the efficiency of the clustering task in order to produce cluster result with better accuracy. A better cluster result can also be produced by combining the cluster results generated from the GA based clustering with Feature Selection and Feature Construction algorithms.

ABSTRAK

PENDEKATAN ENSEMBLE DATA SUMMARIATION BERDASARKAN CIRI TRANSFORMASI UNTUK PEMBELAJARAN RELASIONAL DATA

DARA adalah satu rangka kerja terutamanya untuk meringkaskan data yang disimpan dalam pangkalan data multi-relational dimana table non-target bersekutu dengan table target. Dalam proses meringkaskan data, data yang disimpan dalam pangkalan data multi-relational perlu diubah menjadi ruang vektor Term Frequency - Inverse Document Frequency (TF-IDF). Peningkatan bilangan nilai unik akan meningkat saiz ruang vector TF-IDF dan merumitkan lagi clustering yang boleh menghasilkan hasil kelompok yang ketepatannya rendah. Dalam tesis ini, kaedah Feature Selection telah disiasat and bertujuan untuk mengoptimumkan ruang vektor TF-IDF dengan memilih hanya feature yang relaven dari ruang vektor TF-IDF asal. Tambahan, kaedah Feature Construction disiasat dan bertujuan untuk mengoptimumkan ruang vektor TF-IDF dengan menggabungkan dua atau lebih feature dalam ruang vektor TF-IDF bagi mewakili dataset yang terbaik. Ruang vektor TF-IDF adalah matriks kekerapan wajaran berubah dari pangkalan data multi-relational dengan satu-ke-banyak hubungan. Information Gain dipinjam dari teori Information Retrieval dan algoritma Term-term Correlation digunakan untuk menentukan keutamaan feature untuk dipilih atau digabungkan bagi membentuk satu generasi baru ruang vektor TF-IDF. GA juga digunakan untuk mencari sentroid yang terbaik untuk semua cluster yang dijana dengan sentroid cluster tersebut. Akhirnya, Clustering Ensemble direka, digunakan dan dinilai untuk menghasilkan rangka kerja klasifikasi terakhir yang akan mengambil semua input yang dijana daripada GA berasas clustering bersama kaedah Feature Selection dan Feature Construction dan melaksanakan tugas klasifikasi bagi relational dataset. Beberapa eksperimen telah dilaksanakan untuk menilai prestasi ramalan tentang tugas klasifikasi C4.5 apabila menggunakan kelompok ini hasil daripada beberapa dataset relational dari Mutagenesis, Financial dan Hepatitis domain. Keputusan eksperimen menunjukkan peningkatan pada ketepatan hasil kelompok. Akhirnya, terdapat penambahbaikan ketepatan hasil kompok apabila GA digunakan pada rangkah kerja kasifikasi disamping menggunakan WEKA C4.5 classifier sebagai penilaian ramalan ketetapan keatas hasil kelompok. Akhir sekali, hasil eksperimen bagi ensemble clustering menunjukkan satu petanda yang baik yang menunjukkan kerja-kerja fungsi konsensus betul seperti yang direka. Kajian ini menunjukkan tugas mengoptimumkan ruang vektor TF-IDF dengan mengurangkan bilangan feature di ruang vektor TF-IDF mampu meningkatkan kecekapan tugas kelompok dalam usaha untuk menghasilkan keputusan kelompok dengan ketepatan yang lebih baik. Hasil kelompok yang lebih baik juga boleh dihasilkan dengan menggabungkan hasil kelompok dari GA berasas clustering bersama kaedah Feature Selection dan Feature Construction.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

TF-IDF	Term Frequency–Inverse Document Frequency
DARA	Dynamic Aggregation of Relational Attributes
GA	Genetic Algorithms
KDD	Knowledge Discovery in Databases
ILP	Inductive Logic Programming
IG	Information Gain
eDARA	Ensemble DARA
SFP	Selected Features Percent
FS	Feature Selection
FC	Feature Construction
BM	Bit Merge
BMU	Bit Merge Unsorted
BMFC	Bit Merge Feature Construction
BMUFC	Bit Merge Unsorted Feature Construction
TCL	Term-term Correlation Level
TTCFC	Term-term Correlation Feature Construction
FFS	Forward Feature Selection
BFS	Backward Feature Selection
GFFS	GA Based Clustering with Forward Feature Selection
GBFS	GA Based Clustering with Backward Feature Selection
GBMFC	GA Based Clustering with Bit Merge Feature Construction
GTTCFC	GA Based Clustering with Term-term Correlation Feature Construction
SCR	Single Consolidated Clustering Result File
CCR	Consensus Clustering Result

CHAPTER 1

INTRODUCTION

1.1 Background

The demand for valuable information is growing rapidly as more raw digital data are collected and stored. The massive amount of raw digital data collected limits the capability of the current approaches to process this data into accessible and actionable knowledge. Most structured data are stored in relational databases. Several approaches have been introduced to learn data stored in relational databases and extract valuable information from it. For instance, a Dynamic Aggregation of Relational Attributes (DARA) (Alfred, 2007) is one of the approaches introduced to learn relational data. In DARA approach, all data stored in non-target tables that are associated with the target table are summarized before further data analysis can be performed. A table that is used for patterns extraction is considered as a target table in which each row is representing a single unique object. A table that is used as a reference information is considered a non-target table and more explanation can be found Section 2.3. In the process of summarizing relational data, a data transformation process needs to be performed before the data summarization can be performed completely.

The DARA algorithm is designed particularly to summarize the entire non-target tables that are associated with the target table by clustering records into several clusters in which each cluster is formed based on distinct characteristics. In the process of summarizing relational data, the data stored in a multi-relational setting need to be transformed into term frequency-inverse document frequency (TF-IDF) weighted frequency matrix (Salton and McGill, 1984) that represents data in a vector space representation. The process of transforming a relational data into a vector space representation in the form of TF-IDF vector space will be described and discussed in Chapter 3 (3.2.1 Configuration Phase). The number of features generated by this transformation depends on the number of unique values that exist

in the data stored in multi-relational databases. A single unique value that exists in the relational data is considered as a single feature in the TF-IDF vector space representation. When a large number of features exist in the TF-IDF vector space, the clustering task on the data become more complicated. The feature selection and construction algorithms are used to reduce the complexity of the TF-IDF vector space representation by selecting or combining only relevant features in the TF-IDF vector space representation so that a new generation of TF-IDF is populated with lesser features. Future selection (FS) algorithm is a process of selecting a set of features in a matrix dataset. Future construction (FC) algorithm is a process of combining two or more features in a matrix dataset. However the FS and FC algorithms that applied in previous work of DARA is not consistently producing high yield predictive accuracy clustering result for the mutagenesis and financial and hepatitis datasets. For example, FS that used in DARA can produce high yield predictive accuracy clustering result when compared to feature construction algorithm in the mutagenesis but not the hepatitis datasets. As a result, an ensemble data summarization approach (eDARA) is proposed to enhance the data summarization algorithm of DARA (Alfred, 2007) by combining several clustering results that are produced from FS and FC algorithms in order to generate final clustering result. Further explanation of eDARA can be found in Chapter 3.

In this study, a feature transformation process refers to the algorithms that select relevant features or construct a set of new features for learning purposes. This feature transformation is used in the process of transforming relational data into a vector space representation in order to reduce the number of features represented in the vector space so that the complexity of the vector space can be minimized. Then, the process of data summarization can be performed with efficiently while improving the performance of cluster accuracy. Due to the fact that several distinct sets of relevant features can be selected or generated from the large vector space produced in the data transformation process, an ensemble clustering technique can be applied to find the clustering consensus of several clustering runs that yields a more stable and robust final clustering result. An ensemble clustering method refers to the process of finding clustering consensus in which several clustering results are combined to produce the final clustering result. DARA approach can be implemented

with different methods of feature transaction process based on feature selection and feature construction algorithms to produce a better predictive accuracy of the clustering result. However there is no mechanism to choose the best from from the clustering results that are produced by DARA that using different methods of feature transaction process. Thus, ensemble technique is introduced to this study to combine the clustering results from different feature transaction process into a final clustering result.

1.2 Motivation

In DARA approach, a data summarization method is proposed to summarize data stored in relational databases in order to extract useful information. However, the process of data summarization used in this approach has a drawback due to the large size of relational databases. The large size of relational database coupled with the large number of distinct values that will increase the level of complexity of the vector space representation because the higher the number of unique values that exist in the relational database, the more features will be produced in the TF-IDF vector space representation. The study of reducing the complexity of the vector space model by reducing the number of features in the TF-IDF vector space has motivated this work to explore more algorithms such as FS and FC algorithms that can be used to reduce the features in TF-IDF vector space as well as minimizing the level of complexity of the TF-IDF vector space.

When the FS and FC algorithms are used to reduce and enrich the representation of features in the TF-IDF vector space, various clustering results can be obtained respectively. For instance, when a FS algorithm is applied to the dataset, one can get a good clustering result. However, when a feature construction algorithm is used to enrich the vector space representation of the data, one can obtain a different clustering result. Similarly, when a feature construction algorithm is coupled with the FS algorithm in order to optimize the representation of data in a vector space model, a different clustering result can be obtained. These clustering results can be used to yield an ensemble of clustering results. As a result, this research is motivated to improve the previously proposed DARA algorithm by

proposing an ensemble data summarization approach based on feature transformation in order to learn relational data.

1.3 Research Questions

As stated in section 1.2, the method proposed in DARA approach needs to be enhanced in order to increase its efficiency in summarizing massive data stored in relational databases. This research will adopt an optimization technique in summarizing relational data in which data are populated across multiple tables with records having a one-to-many relationship in the relational database. Therefore, in this thesis, three fundamental research questions will be investigated:

1. Is it feasible to select a set of relevant features from the vector space representation and construct a new set of relevant features in order to improve the classification task?
2. Is it feasible for the classification task to be improved by optimizing the clustering results using a GA?
3. It is feasible to propose an ensemble data summarization method based on the feature transformation process in order to effectively learn relational data?

1.4 Research Objectives

The aim of this work is mainly to enhance the data summarization algorithm proposed in the DARA approach by optimizing the data representation of the vector space modal. Thus, there are three main objectives of this work that include the tasks:

- To develop the methods that can optimize and reduce the size of the vector space used in DARA while improving the predictive accuracy of the clustering result.
- To develop a GA based clustering with FS and FC algorithms for the proposed data summarization approach.
- To develop a data summarization ensemble that applies a consensus clustering based on the feature transformation algorithms in order to produce a better clustering result.

1.5 Research Contributions

The main goal of this research is to enhance the DARA data summarization technique in order to learn relational data more effectively. Therefore, this research contributes towards the data mining field that focuses on the data summarization by demonstrating:-

1. The overall data summarization technique can be improved by optimizing the FS and FC processes.
2. The overall prediction task can be improved by optimizing the FS and FC algorithms using GA to select the best cluster centroids for the clustering process.
3. A consensus clustering can be applied to produce a better clustering result by using the ensemble clustering technique that combines several clustering results that are generated by the FS and FC algorithms.

1.6 Research Scope and Strategy

The scope of this research refers to three predefined stages that include data initialization, data preparation and transformation, data enrichment and data summarization. Figure 1.1 shows the three predefined stages.

1.6.1 The Data Preparation and Transformation Stage

There are three main relational databases that will be used in this research namely, Financial, Mutagenesis and Hepatitis datasets and they were chosen because they are also used in the work of the previous DARA. Section 2.6 will describe in detail about these datasets. In this data preparation and transformation stage, the data stored in multiple tables are transformed into a TF-IDF vector space representation via the TF-IDF transformation process. A TF-IDF vector space is a record-pattern matrix in which a row represents a primary record referred from the target table in a relational database and the column represents a pattern that exists in the non-target table associated to the primary record. Data represented in the TF-IDF vector space is stored in a form of numerical statistic that represents a weight of a term in the document which reflects how frequent each term (represented a column) exists in a

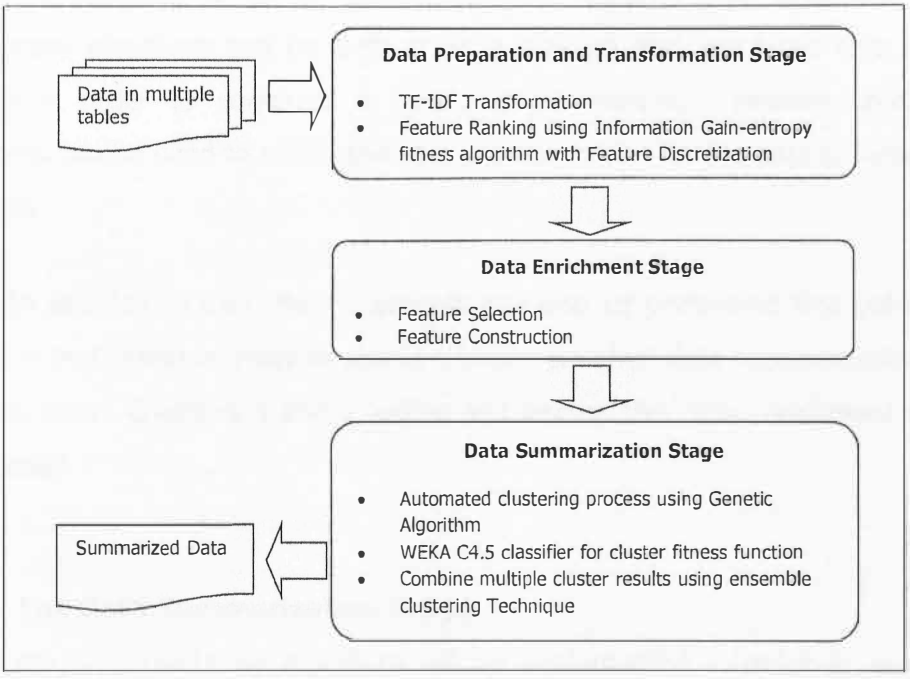


Figure 1.1: Three predefined stages in this research: Data Preparation and Transformation, Data Enrichment and Data Summarization Stages.

collection of the dataset.

All features stored in the TF-IDF vector space representation will then be ranked based on the value of feature scoring. In this work, the Information Gain (Jaynes, 1957) feature scoring is used to rank the features. In the process of ranking all the features, all numerical values in the TF-IDF vector space will be discretized in order to reduce the number of unique values and this will facilitate the computation of feature scoring by Information Gain fitness algorithm. After the Information Gain values for all features are computed, it will be used to rank these features in the TF-IDF vector space. Then, the features stored in TF-IDF vector space are sorted according to their ranking in ascending or descending order. This process will be explained in more details in Chapter 3 (Section (iii)).

1.6.2 The Data Enrichment Stage

In this stage, two main feature reduction techniques will be implemented. They are called FS and FC algorithms. In general, the FS algorithm can be defined as a

process of selecting a set of several relevant features only while the feature construction algorithm can be defined as a process that combines two or more features in order to construct a new relevant feature. Feature construction algorithms can be used to enrich the data representation for the data summarization purposes.

In addition to that, the FC process can also be performed first before a FS process is performed in order to obtain a better enriched data representation of the relational data. Chapters 4 and 5 outline and discuss this data enrichment stage in more details.

1.6.3 The Data Summarization Stage

In this stage, a clustering algorithm will be implemented in order to assess the quality of selected features and constructed features which are performed in the data enrichment stage. In this stage, this clustering technique is optimized by using a GA. Chapters 4 and 5 discuss about the experimental design and setup for the assessment of the FS and FC algorithms.

A GA based clustering with FS and FC algorithms are also implemented by embedding these processes into the classification task. A WEKA C4.5 classifier is used as the main classifier for searching the best clustering result when clustering process is taking place. WEKA (Witten and Frank, 1999) is a machine learning tools that purposely to evaluate the accuracy of a clustering result and C4.5 (Quinlan, 1993b) is a classification algorithm generates a decision tree from training data for classification. Chapter 6 discusses the implementation of the proposed GA based clustering with FS and FC algorithms in more details. A 10-folds cross validation setting will be used in the evaluation setup.

Finally, the implementation of a data summarization ensemble technique will be discussed in Chapter 7, where a consensus clustering technique is proposed and assessed.

1.7 Thesis Organization

This thesis consists of eight chapters. The organization of this thesis is as follows:

Chapter 2 provides a literature review of the approaches, theories and techniques related to the Knowledge Discovery and Data Mining. Recent researches conducted by other researchers who contributed ideas to this research are also reviewed here.

Chapter 3 outlines the overview description of the proposed solutions using the cluster ensemble in order to enhance the DARA approach. FS and FC algorithms are added in order to obtain multiple clustering results for the proposed ensemble clustering technique. Preliminary experiments are also conducted and discussed in this chapter in order to find the best set of values for the parameters used in the experimental studies conducted in Chapters 4, 5, 6 and 7.

Chapter 4 describes the implementation of the FS algorithm in order to reduce the number of features in the TF-IDF vector space representation. A comprehensive analysis of the results obtained is also discussed in this chapter.

Chapter 5 describes the implementation of the FC algorithm in order to enrich the number of features in the TF-IDF vector space representation. This chapter concludes by discussing the experimental results obtained.

Chapter 6 describes the implementation of the GA and WEKA C4.5 classifier in the data summarization process which involves FS and FC algorithms discussed previously in chapters 4 and 5 respectively. The experimental setup and results obtained are discussed in details.

Chapter 7 describes the implementation consensus phase in eDARA and discusses the experimental results obtained.

Chapter 8 concludes the entire research work done. In this chapter, the main contributions of the thesis are summarized in line with the proposed objectives of this thesis. Finally, future works and research directions are also discussed here.