NEW DATA MINING MODELS BASED ON FORMAL CONCEPT ANALYSIS AND PROBABILITY LOGIC

by

Liying Jiang

A DISSERTATION

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Doctor of Philosophy

Major: Computer Science

Under the Supervision of Professor Jitender S. Deogun

Lincoln, Nebraska

July, 2006

UMI Number: 3216105

UMI Microform 3216105

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

NEW DATA MINING MODELS BASED ON FORMAL CONCEPT ANALYSIS AND PROBABILITY LOGIC

Liying Jiang, PhD

University of Nebraska, 2006

Advisor: Jitender S. Deogun

This dissertation enhances data mining processes by formalizing them in a logic framework, with the focuses on improving the efficiency of association rule mining and extending the use of association rules to make predictions based on the proposed framework.

Although extensive studies have been done on data mining, most of them concentrate on specific application domains. A logic framework to formally represent important notions and processes in data mining has attracted little attention. SPICE — *Symbolic integration of Probability Inference and Concept Extraction*, is therefore proposed, in which the logic representations of *concepts*, *patterns*, *previously unknown* and *potentially interesting patterns* are formalized. Two primary data mining tasks, association rule mining and classification, are formally represented as pattern discovery processes in SPICE.

Based on the SPICE framework, a new special type of patterns, *Maximal Potentially Useful* (MaxPUF) patterns, is formalized. The MaxPUF patterns lead to a new class of association rules, called MaxPUF rules. These rules are characterized by the minimum antecedents among all the high-confidence rules for the same consequent. At the same time, this minimum antecedent includes the most important factors to imply a consequent with high confidence. Thus, the MaxPUF rules are very interesting and potentially useful to the user. The mining of MaxPUF rules provides a

solution to the rule redundancy problem in association rule mining, which occurs when a large number of rules are generated and many of them are uninteresting or unimportant.

A new mining approach called *Succinct Worthy Association Rule Mining* (SWARM) is proposed to improve mining efficiency. Different from previous mining approaches that only prune the infrequent itemsets, SWARM adopts a new pruning strategy that deletes less important items in the mining process. Because a much smaller number of itemset candidates are generated after the items have been deleted, SWARM is more efficient than previous approaches. In SWARM the MaxPUF rules are used to help identify less important items.

In addition, the possible use of association rules for prediction is studied and a new prediction rule model is proposed. The experimental results show that the discovered prediction rules can be used for prediction with good results.

Overall, this dissertation introduces a logic framework for data mining and develops methodologies based on the proposed framework to enhance data mining.

ACKNOWLEDGEMENT

Many people have greatly contributed to the successful completion of this dissertation. First of all, I would like to thank my advisor Professor Jitender S. Deogun for his encouragement, inspiration, guidance and above all his advice in the past four years. He has helped me to expand the breadth and depth of my knowledge and research by providing many insights and directions to my research. I am greatly indebted to him.

I am very thankful to my committee members, Dr. Steve Goddard, Dr. Ashok Samal and Dr. Jamie Radcliffe for their great help during my PhD studies. Especially, I want to thank Dr. Goddard for his valuable comments and help to improve quality of this dissertation.

Thanks to my fellow colleagues Dan Li who has been very helpful to my start on data mining research. I also would like to extend my gratitude to Chenguang Wang and Qin Li for many insightful and thought provoking conversations for solving problems using statistical methods in the dissertation. Thanks to Professor Robert Haller and Elissa Dahlberg to help me review the presentation of this dissertation.

I want to thank my friends Wei Li, Yuancui Huang, Yun Feng, Jun Gao, Xuming Wu, Christina Cassel, Lixin Li, Xin Liu, Qing Yang, Wang Yao, Jingyi Yang, Xuli Liu and Jianzhi Wang. Their support and friendship are valuable and precious during my time at UNL.

I am deeply grateful to my parents and my sister for their unlimited support and love throughout all these year. They have been extremely helpful and a constant source of strength.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

With the popularity of the Internet and advanced data collection technologies, data across a wide variety of application domains are accumulating at a dramatic pace. Data are valuable assets, only if we know how to reveal the useful knowledge hidden in the raw data. With large volumes of data at hand, there is an urgent need for a new generation of computational theories and tools to assist humans to extract useful information and knowledge from the large datasets. Knowledge Discovery in Databases (KDD) is therefore developed to accommodate such needs. Data mining is usually seen as one of the most important steps in a KDD system, which extracts the "diamonds" of knowledge from historical data and predicts outcomes for the future. Data mining is defined as a "*nontrivial extraction of hidden, previously unknown and potentially useful information from large datasets*" [25].

In this dissertation, we enhances data mining processes by formalizing them in a logic framework. Specifically, we focus on improving the efficiency of association rule mining and extending the use of association rules to make predictions based on the proposed framework.

1.1 Motivation

Although extensive studies have been done on data mining, most of them concentrate on application domains. A logic framework to formally represent important notions and processes in data mining has attracted little attention. Building a logic framework for data mining is a nontrivial task, since data mining processes are usually complicated and application-dependent. In this dissertation, we make an initial attempt at a logic framework for data mining, in which the essential notions and processes in data mining can be represented logically and formally. Based on such a framework, we can build new models (e.g., new association rule models) for data mining tasks and improve data mining processes (e.g., association rule mining) for knowledge discovery.

Association rule mining is one of the most important data mining techniques, which has gained popularity because it presents the discovered knowledge in a way that people can easily understand. However, association rule mining has two main flaws for large database applications, namely rule redundancy and mining inefficiency. The rule redundancy problem is that usually a large number of rules are generated, many of which are not interesting or important. Previous work, such as [43, 86], proposed to reduce the number of rules by selecting a subset so that the other rules can be derived from them. However, they do not make attempts to develop a rule model excluding those rules that may not be important or interesting to the user. In this dissertation, we intend to define a new set of rules that are potentially most useful and interesting to the user.

In association rule mining, *support threshold* specifies the minimum frequency requirement for the rules to be discovered. When the rules involving the itemsets with relatively low frequency are also of interest, a smaller support threshold value is needed. Current rule mining algorithms have good performance when the support

threshold used is large. However, association rule mining is still confronted with the efficiency problem for large database applications when the support threshold used is small [38, 88]. On the other hand, lowering the support threshold will generate a large number of redundant rules. Therefore, we intend to develop an enhanced mining approach based on the idea of dealing with these two related problems together.

Traditionally, association rules are claimed to represent the co-occurrence relationships among data. When considered in the context of temporal databases, association rules may also reflect the purported cause-effect relationships over time. Applying association rules for prediction is therefore a promising research topic for extending the use of association rule mining.

In this dissertation, we investigate these problems in data mining and propose solutions to them. The goals of this thesis include:

1 Development of a logic framework for data mining, in which the important notions and processes in data mining can be logically and formally represented.

2 Development of a new type of association rules that are potentially most useful and interesting to the user based on the logic framework.

3 Development of a new mining approach to enhance the efficiency of association rule mining by making use of the new type of association rules.

4 Extension of the use of association rule mining for prediction.

1.2 Background Knowledge

1.2.1 Data Mining Basics

Data mining at its core is useful knowledge discovery, which focuses on finding patterns, associations, changes, anomalies, and statistically significant structures and

events in vast amounts of data. The main data mining techniques include: classification, clustering, association rule mining, and regression [15, 25, 61].

- Classification is to discover a predictive model or function that classifies the objects based on their attributes.

- Clustering is to identify a finite set of categories or clusters to describe the data based on different similarity measures.

- Association rule mining is to produce dependency rules which describe the occurrence of some events based on the occurrences of other events.

- Regression is to find a linear or nonlinear model of dependency to predict a value of a given continuous valued variable based on the values of other variables.

Data mining can be viewed as an automatic or semi-automatic learning process. In general, all data mining methods use induction-based learning, which is the process of forming general concepts by observing specific examples of the concepts to be learned [61]. Each concept can be seen as a set of objects, symbols, or events grouped together based on the characteristics that they share. Thus, the modeling of data mining is based on concept generalization and concept extraction.

Note concepts can be constructed based on two different views [61]. A classical view requires that all concepts have defining properties and these properties determine that if an individual item is an example of a particular concept. In contrast, a probabilistic view allows representation of concepts with properties that the concept may possibly possess. The assumption is that people store and recall concepts as the generalizations created from individual observations. While sometimes the observations are not complete or accurate, we may need to estimate the concepts with a probability. The probabilistic view may not be directly applied to achieve an answer

about whether a specific instance belongs to a group. However, it is useful for decision making by associating a probability indicating the membership of an instance to a specific group. As both classic and probabilistic views are useful, we believe that a logic with probability inferences is needed for modeling data mining.

Association Rule Mining Basics

Association rule mining was described in [1, 2]. In the following, we briefly introduce the terminologies in association rule mining based on the descriptions in [1, 2].

Let $I = \{i_1, i_2, ..., i_m\}$ be a set of distinct literals. Each literal is called an item. A *transaction* t is a set of items such that $t \subseteq I$. $D = \{t_1, t_2, ...t_n\}$ denotes a database of n transactions. If X is a set of k items, then X is called a *k-itemset*. The *support* of an itemset X, denoted by *sup(X)*, is defined as the percentage of transactions in D that contain X. An itemset is said to be *frequent* if its support is greater than or equal to a user pre-specified support threshold.

An association rule is in the form of $X \Rightarrow Y$, where X, Y are itemsets and $X \cap Y = \emptyset$. X is called the *antecedent* of the rule and Y is called the *consequent* of the rule. *Support* and *confidence* are two frequently used metrics for evaluating a rule. The *support* of a rule $X \Rightarrow Y$ is denoted by $sup(X \Rightarrow Y)$, indicating the percentage of transactions in a dataset that contain both X and Y, that is, $sup(X \Rightarrow Y)$=$sup(X \cup Y)$. The *confidence* of a rule $X \Rightarrow Y$ indicates the probability that a transaction containing Y given that it contains X, $conf(X \Rightarrow Y) = sup(X \Rightarrow Y)/sup(X)$.

Generally, a high support value implies that a rule is statistically significant. Similarly, a high confidence value is characteristic of a *strong* association between the antecedent and the consequent. Strong and statistically significant association rules are attractive. Therefore, the association rule mining problem is formulated as to discover all the association rules that have support and confidence values at

least equal to the user-specified *minimum support* and *confidence thresholds* [1]. The minimum support threshold is denoted as $\mathcal{S}_{min}$ and the minimum confidence threshold is denoted as $\mathcal{C}_{min}$ in this dissertation.

The *Apriori* algorithm is a well-known and standard algorithm for mining association rules, which includes two steps: first it discovers frequent itemsets and then generates association rules using the frequent itemsets. It is shown that the first step is critical and determines the overall efficiency of the mining process [2]. Efficiently discovering the frequent itemsets thus has been a focus in improving association rule mining [2, 33, 65].

To discover all frequent itemsets, the methodology is to literately enumerate the frequent itemsets from 1-itemsets to n-itemsets. The search space for all itemsets is $2^{|I|}$, which means that it is exponential in the number of items in the database. To prune the itemset candidates, the Apriori algorithm addresses an important property — *downward closure property* in frequent itemset mining. The downward closure property states that if an itemset X is not a frequent itemset, then none of its supersets are frequent itemsets. Alternatively, if an itemset A is a frequent itemset, all of its subsets are frequent itemsets. Therefore, the Apriori algorithm generates a candidate only if all of its subsets are frequent. The Apriori algorithm is efficient when $\mathcal{S}_{min}$ used is large, in which case the number of itemset candidates generated shrinks quickly in the iteration process. However, Apriori and its variations are not efficient when a smaller support threshold is used, due to the large number of frequent itemsets generated in each of the iterations [88].

1.2.2 Formal Concept Analysis and Probability Logic

Formal concept analysis (FCA) [80] is a method for deriving conceptual structures out of data, allowing the analysis of complex structures and discovering dependencies

within the data [79, 80]. As the modeling of data mining is based on the concepts and concept extraction, FCA offers a suitable theoretical foundation to this end.

The basic notions in formal concept analysis are that of *formal context* and *formal concept* [63]. A formal context is a logic representation of the dataset. Formal concepts are extracted from a given formal context. A formal concept is a pair of *intent* and *extent*. The intent of a formal concept is a conjunct of features that each object of the extent must possess. The extent of a formal concept is the maximal set of objects sharing the features in the intent [79]. Formal concepts are useful in that given either a set of features or objects, we can directly know the objects that share the set of features or the common features the set of objects possess. Formal concept analysis formulates the concept learning as a process that groups indiscernible objects.

Bacchus probability logic [6] is based on first-order logic and extends it with the probability expression capability. It defines a logic language on *statistical probability* and *propositional probability*. Statistical probability is defined over a set of elements, which relates to the properties of the set while not necessarily to the individual elements of the set. Propositional probability is a probability assigned to particular propositions or assertions about certain elements.

Probability logic sets up a general connection between propositional probability and statistical probability based on *direct inference*, the principle of which is: *the probability of an individual to possess a property is equal to the relative frequency of that property among the smallest class, which the individual belongs to and which one has "reliable statistics" about* [6]. That is, the propositional probability can be assigned based on one's ability to choose an appropriate reference class. Once the reference class is chosen, the statistical information of the class can be used for indicating the propositional probability [6].

Probability logic is an expressive logic, which makes it suitable for modeling com-

plicated notions and processes in data mining.

1.3 Research Contributions

The contributions of this dissertation include the following aspects. First, we propose a logic framework for data mining called SPICE, in which we formulate the logic representations of the important notions and processes in data mining.

Second, based on the SPICE framework, we develop MaxPUF patterns, which lead to a new class of association rules — MaxPUF rules. We show that MaxPUF rules are potentially most useful and very interesting to the user. The mining of MaxPUF rules provides a solution to the rule redundancy problem, as MaxPUF rules make a much smaller set than traditional rules.

Third, we propose a novel association mining strategy, based on which we develop a new mining approach called SWARM. SWARM greatly reduces the number of itemset candidates generated, and thus becomes more efficient than previous approaches.

Fourth, we propose a new prediction rule model, which defines a set of rules that are suitable for prediction.

Figure 1.1 shows an overview of this dissertation.

1.4 Dissertation Outline

The rest of this dissertation is organized as follows. In Chapter 2, we present a review on the formal concept analysis, rough set, probability logic and data mining. In Chapter 3, we develop SPICE, a new logic framework for data mining, in which we formulate the logic representations of important notions and processes in data mining. In Chapter 4, we propose a new type of association rules — MaxPUF rules. MaxPUF rules are a small set of potentially most useful and interesting rules to the user. In

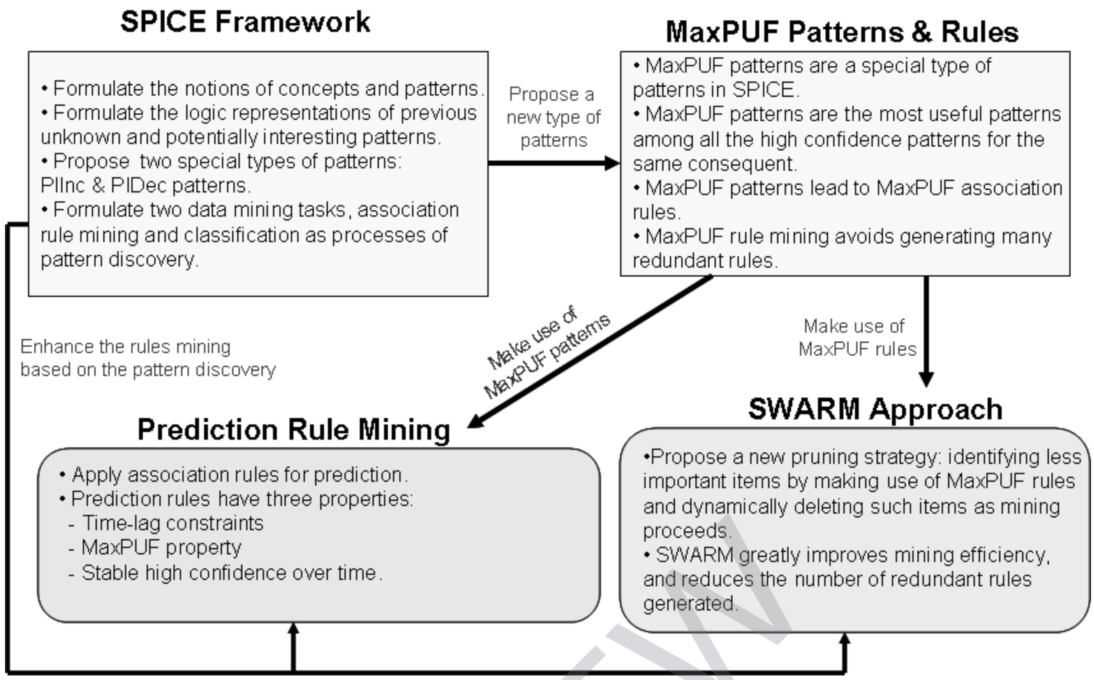

Figure 1.1: An Overview of this Dissertation.

Chapter 5, we present an enhanced association rule mining approach — SWARM, which adopts a new pruning strategy that dynamically deletes less important items in the mining process. In Chapter 6, we propose the prediction rule mining approach, which extends the application domains of association rules. Concluding remarks are given in Chapter 7.

Chapter 2

Formal Concept Analysis, Rough Sets, Probability Analysis and Data Mining — A Survey

2.1 Chapter Summary

Formal Concept Analysis (FCA), Rough Sets and Probability Analysis are three important and useful theoretical foundations for data mining. In this chapter we briefly review these theories and their applications to data mining. We also discuss the possible integration of these techniques into a framework useful for modeling the data mining processes.

2.2 Introduction

Formal concept analysis, rough set and statistical probability analysis are suitable for supporting and modeling the data mining processes from different perspectives.

However, they can play complementary roles and thus their integration is expected to lead to better data mining models. In this chapter, our objective is to review these theories and try to reveal the differences and commonness among these theories in facilitating data mining.

2.3 Formal Concept Analysis and Data Mining

FCA was developed by Wille and his colleagues at Darmstadt University in early 1980's as a method for deriving conceptual structures from data [80]. It can be used for the analysis of complex structures or to discover the dependencies within data [79, 80]. In general FCA can help data mining in two aspects. First FCA provides tools for formal representation of knowledge that facilitates efficient computer processing of the knowledge. Second FCA helps to formalize the conceptual knowledge discovery for different data mining tasks.

2.3.1 FCA Foundations

FCA is a theory for the formalization of *concepts* within a *formal context* [26, 63]. A *formal context* is a binary relation between a set of objects and a set of attributes [63].

Definition 1 Formal context. *Formal context is a triplet* (G, M, R)*, where* G *and* M *are two finite and nonempty sets, namely object set and attribute set. The relationships between objects and attributes are described by a binary relation* R *between* G *and* M*, which is a subset of the Cartesian product* $G \times M$*. If object* x *possesses attribute* g*, it is denoted as* $(x, y) \in R$*, or* xRy *[63].*

Based on the definition of formal context, we know that an object $x \in G$ has the set of attributes $xR = \{y \in M | xRy\} \subseteq M$. And an attributes y is possessed by the

set of objects $Ry = \{x \in G | xRy\} \subseteq G$.

FCA is an unsupervised learning technique for discovering conceptual structures in data. To perform FCA, we define a set-theoretic operator "$*$" to associate a subset of *objects* and to a subset of *attributes* in a formal context (G, M, R).

$$\begin{aligned} X^* &= \{y \in M | \forall x \in G (x \in X \Rightarrow xRy)\} \\ &= \{y \in M | X \subseteq Ry\} \\ &= \bigcap_{x \in X} xR. \end{aligned}$$

The above shows that the "*" operation associates a subset of attributes X^* to the subset of objects X. Similarly, for any subset of attributes $Y \subseteq M$, we can associate a subset of objects $Y^* \subseteq G$ as follows:

$$\begin{aligned} Y^* &= \{x \in G | \forall y \in M (y \in Y \Rightarrow xRy)\} \\ &= \{x \in G | Y \subseteq xR\} \\ &= \bigcap_{y \in Y} Ry. \end{aligned}$$

The "*" operation induces the following properties. For $X, X_1, X_2 \subseteq G$ and $Y, Y_1, Y_2 \subseteq M$,

$$\begin{array}{lcc} (1) & X_1 \subseteq X_2 \Longrightarrow {X_1}^* \supseteq {X_2}^*, & Y_1 \subseteq Y_2 \Longrightarrow {Y_1}^* \supseteq {Y_2}^*. \\ (2) & X \subseteq X^{**}, & Y \subseteq Y^{**}. \\ (3) & X^{***} = X^*, & Y^{***} = Y^*. \\ (4) & (X_1 \cup X_2)^* = X_1^* \cap X_2^*, & (Y_1 \cup Y_2)^* = Y_1^* \cap Y_2^*. \end{array}$$

A mapping (*) is called a Galois connection if it satisfies (1) and (2), and hence (3). By definition, $x^* = xR$ is the set of attributes possessed by x, and $y^* = Ry$ is the set of objects having attribute y. For a set of objects X, X^* is the maximal set

of attributes shared by all objects in X. Similarly, for a set of attributes Y, Y^* is the maximal set of objects that have all attributes in Y [83].

Definition 2 Formal concept. *A pair* (X, Y) *where* $X \subseteq G$ *and* $Y \subseteq M$ *is a formal concept if* $X = Y^*$ *and* $Y = X^*$. X *is called the extent of the concept and* Y *is called the intent of the concept.*

By property (3), for any subset $X \subseteq G$, we have a formal concept (X^{**}, X^*), and for any subset $Y \subseteq M$, we have a formal concept (Y^*, Y^{**}). With formal concepts, given either a set of attributes (or objects), we can directly obtain all the objects that share the set of attributes (or the common attributes that the set of objects possess).

The formal concepts defined over a formal context form a complete lattice, called *concept lattice* [26]. The *meet (infima)* and *join (suprema)* of the lattice are respectively given by:

$$(X_1, Y_1) \wedge (X_2, Y_2) = (X_1 \cap X_2, (Y_1 \cup Y_2)^{**}), \text{ and}$$
$$(X_1, Y_1) \vee (X_2, Y_2) = ((X_1 \cup X_2)^{**}, Y_1 \cap Y_2).$$

2.3.2 Knowledge Representation with FCA

FCA formalism is very suitable for knowledge representation. In [55] Davis et al. address five principles that a knowledge representation should confirm, which are 1) a medium of human expression, 2) a set of ontological commitments, 3) a surrogate, 4) a fragmentary theory of intelligent reasoning, and 5) a medium for pragmatically efficient computation. But each knowledge representation formalism may be in some way a trade-off among these principles.

The first three principles have been the driving forces for the development of FCA. Moreover, as the orientation of FCA shifted from mathematics towards computer science, it also gradually embodies the last two principles. Initially Wille developed

FCA theory in order to reconstruct the lattice theory, and to promote better communications between lattice theorists and potential users of the lattice theory. Wille referred to the roots of the lattice idea, namely hierarchies of concepts, to formalize the logic of human expressions [79]. Therefore, FCA is a theory of medium of human expression.

The ontological commitment of FCA lies in that it formalizes the notions of *concept* and *conceptual hierarchy* [55]. The FCA theory understands the concepts as the most basic units of thought, based on which more complex entities of thought, e.g. judgments and conclusions, can be built. Formal concepts explicitly formalize the extension and intension of a concept and the fact that an increasing intent implies a decreasing extent and vice versa [55].

The basic surrogates in FCA are formal contexts and concepts. The notion of formal context follows the understanding that one can analyze and argue only in the restricted contexts. In real-life applications, the transitions from reality to formal models (and back) are made explicitly by the use of formal contexts. Formal concepts, being surrogates, only consider the selected aspects of concepts, excluding the fuzziness, modifications over time and so forth [55, 63].

Though different from first order logic, FCA is also a theory of intelligent reasoning, which emphasizes on inter-subjective communication and argumentation. Wille discussed some real-world applications of FCA and showed how the intelligent reasoning is supported by FCA in application domains [81].

FCA provides an efficient information organization mechanism, the concept lattice. Lattices represent the knowledge with a clear structure, and comprise a partial order that allows of multiple inheritances. Lattices also facilitate retrieval of information from the stored knowledge.