PROCEEDINGS OF SPIE

SPIE—The International Society for Optical Engineering

Data Mining and Knowledge Discovery: Theory, Tools, and Technology IV

Belur V. Dasarathy
Chair/Editor

1–4 April 2002
Orlando, USA

SPIE
P
PROCEEDINGS SERIES

Volume 4730

PROCEEDINGS OF SPIE
SPIE—The International Society for Optical Engineering

Data Mining and Knowledge Discovery: Theory, Tools, and Technology IV

Belur V. Dasarathy
Chair/Editor

1–4 April 2002
Orlando, USA

Sponsored and Published by
SPIE—The International Society for Optical Engineering

SPIE
P
PROCEEDINGS SERIES

Volume 4730

SPIE is an international technical society dedicated to advancing engineering and scientific applications of optical, photonic, imaging, electronic, and optoelectronic technologies.

The papers appearing in this book compose the proceedings of the technical conference cited on the cover and title page of this volume. They reflect the authors' opinions and are published as presented, in the interests of timely dissemination. Their inclusion in this publication does not necessarily constitute endorsement by the editors or by SPIE. Papers were selected by the conference program committee to be presented in oral or poster format, and were subject to review by volume editors or program committees.

Please use the following format to cite material from this book:

Author(s), "Title of paper," in *Data Mining and Knowledge Discovery: Theory, Tools, and Technology IV*, Belur V. Dasarathy, Editor, Proceedings of SPIE Vol. 4730, page numbers (2002).

ISSN 0277-786X
ISBN 0-8194-4480-4

Published by
SPIE—The International Society for Optical Engineering
P.O. Box 10, Bellingham, Washington 98227-0010 USA
Telephone 1 360/676-3290 (Pacific Time) • Fax 1 360/647-1445
http://www.spie.org/

Printed in the United States of America.

Conference Committee

Conference Chair

Belur V. Dasarathy, Dynetics, Inc. (USA)

Program Committee

Jafar Adibi, University of Southern California (USA)
Sarabjot S. Anand, MINEit Software Ltd. (UK)
Joydeep Ghosh, University of Texas/Austin (USA)
Wynne Hsu, National University of Singapore
Andrew Kusiak, University of Iowa (USA)
Tsau Young Lin, San Jose State University (USA)
Howard E. Michel, University of Massachusetts/Dartmouth (USA)
Valery A. Petrushin, Accenture (USA)
Stephan Rudolph, Universität Stuttgart (Germany)
Ishwar K. Sethi, Oakland University (USA)
James G. Shanahan, Xerox Research Center Europe (France)
Hannu T. T. Toivonen, Nokia Research Center (Finland)
Shusaku Tsumoto, Shimane Medical University (Japan)
Ari J. Visa, Tampere University of Technology (Finland)
Richard Weber, Universidad de Chile
Lian Yan, Athene Software, Inc. (USA)
Suk-Chung Yoon, Widener University (USA)
Yanqing Zhang, Georgia State University (USA)

Special Student Volunteer

Gurushyam Hariharan, Netaji Subhash Institute of Technology (India) and University of Delhi (India)

Session Chairs

1 Attributes Extraction, Assessment, and Selection
Tsau Young Lin, San Jose State University (USA)
Sarabjot S. Anand, MINEit Software Ltd. (UK)

2 Pattern Analysis and Visualization
Jafar Adibi, University of Southern California (USA)
Andrew Kusiak, University of Iowa (USA)

3 Neural Nets and Genetic Algorithms
Sarabjot S. Anand, MINEit Software Ltd. (UK)
Howard E. Michel, University of Massachusetts/Dartmouth (USA)

4 Association Rules
Richard Weber, Universidad de Chile
Tsau Young Lin, San Jose State University (USA)

5 Knowledge Discovery Processes
Stephan Rudolph, Universität Stuttgart (Germany)
Valery A. Petrushin, Accenture (USA)

6 Data Mining Tools
Andrew Kusiak, University of Iowa (USA)
Ari J. Visa, Tampere University of Technology (Finland)

7 Miscellaneous Tools I
Shusaku Tsumoto, Shimane Medical University (Japan)
Jafar Adibi, University of Southern California (USA)

8 Web Applications
Richard Weber, Universidad de Chile
Ari J. Visa, Tampere University of Technology (Finland)

9 Text Data and Electronic Commerce Applications
Howard E. Michel, University of Massachusetts/Dartmouth (USA)
Andrew Kusiak, University of Iowa (USA)

10 Medical, Laboratory, and Natural Resources Applications
Valery A. Petrushin, Accenture (USA)
Richard Weber, Universidad de Chile

11 Miscellaneous Applications
Valery A. Petrushin, Accenture (USA)
Jafar Adibi, University of Southern California (USA)

12 Aproximate Reasoning: Fuzzy Logic and Rough Sets
Howard E. Michel, University of Massachusetts/Dartmouth (USA)
Shusaku Tsumoto, Shimane Medical University (Japan)

13 Miscellaneous Tools II
Ari J. Visa, Tampere University of Technology (Finland)
Stephan Rudolph, Universität Stuttgart (Germany)

Introduction

In spite of the expected impact of the events of 11 September 2001 on the economy in general and travel in particular, we have had an impressive response for this conference as evidenced by the increase in both the number of initial abstract submissions and final contributed papers included here.

The two conferences on data mining and sensor fusion are again presented under a single track. A tradition started last year, this helps in recognizing, exploiting, and nurturing the synergy between the two fields. The resultant intermingling of the two groups should help the overall advancement of the state of the art. The fact that the printed volumes are being made available on-site aids the goal of rapid dissemination of the most recent developments in these areas. Thus, it effectively complements the peer-reviewed archival publications that, by their very nature, represent a much slower process.

The conference and papers have been grouped into 13 sessions dealing with attributes extraction, assessment and selection, approximate reasoning, pattern analysis and visualization, association rules, knowledge discovery processes, data mining tools, web applications, text data and electronic commerce applications, medical, laboratory, and natural resource applications, and miscellaneous tools and applications. As in previous years, the conference, reflecting the diversity of the program committee, has global representation from over 15 different countries (Australia, Brazil, Chile, China, Denmark, Finland, France, Germany, Hong Kong, India, Japan, Malaysia, Poland, Taiwan, Turkey, Ukraine, and USA). This increase from past years demonstrates a widening of the appeal of the conference in a geographical sense.

Given the continued growth in the response to this conference, we plan to offer it again in the coming year. Accordingly, we hope and expect to see an even larger and wider participation. Anyone interested in active participation in planning and conference program development process should contact me at belur@ieee.org as early as possible. Further details regarding the call for papers and schedule for the next year will be made available in due course on the Internet at SPIE (http://www.spie.org) as well as my home page (http://belur.tripod.com). Finally, it is indeed my pleasure to acknowledge the authors for choosing this avenue for the presentation of their work and thereby contributing to its success. I would also like to take this opportunity to thank the members of my program committee and the session chairs for their cooperation and support.

कायेन वाचा मनसेन्द्रियैर्वा
बुध्यात्मनावा प्रकृते स्वभावात
करोमि यद्यत सकलं परस्मै
श्रीमन्नारायणायेति समर्पयामि

"kaayena vaachaa manasendriyairvaa
budhyaatmanaavaa prakR^ite svabhaavaat
karomi yadyat sakalaM parasmai
shriiman naaraayaNaayeti samarpayaami"

Be it with my body, or with my mind
With words, or organs of any kind,
With my intellect, or with my soul,
Or by force of Nature pushing me to my goal,
Whatever it is, with all these I do,
Oh! Supreme Lord! I surrender to you.

Belur V. Dasarathy

Contents

SESSION 1 ATTRIBUTES EXTRACTION, ASSESSMENT, AND SELECTION

SESSION 2 PATTERN ANALYSIS AND VISUALIZATION

SESSION 3 NEURAL NETS AND GENETIC ALGORITHMS

SESSION 4 ASSOCIATION RULES

SESSION 11 MISCELLANEOUS APPLICATIONS

SESSION 12 APPROXIMATE REASONING: FUZZY LOGIC AND ROUGH SETS

SESSION 13 MISCELLANEOUS TOOLS II

Feature Transformations and Structure of Attributes

Tsau Young (T.Y.) Lin
Department of Mathematics and Computer Science
San Jose State University
San Jose, CA 95192

ABSTRACT

Let V be a set real world entities, that admits a relational model. An attribute induces an equivalence relation on V and will be regarded as such. Let A* be the set of such equivalence relations and their refinements. Then the smallest sublattice L(A*) generated by A* in Π(V) is called generalized relation lattice. The set of the granules in L(A*) whose cardinality is above certain threshhold is the set of all possible patterns derivable from the given attributes (features) of the given relation.

1. INTRODUCTION- WHAT IS DATA MINING?

What is data mining? There are many popular citations. To be specific, [4] defines data mining as the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns from data. Clearly it serves more as a guideline than a scientific definition. "Novel," "useful," and "understandable," involve subjective judgments; they cannot be used for scientific criteria. Nevertheless, it does imply two very important points:

patterns, and
real world implications.

For the first point, it is difficult to capture a perfect definition. In practice, however, working definitions that cover sufficient large class of examples are adequate. In *database* mining, *repeated occurrence is* a well accepted criterion. A sub-tuple is an association rule (a pattern), if the frequency of *occurrences* is above certain threshold.

For second point, we note that though subjective judgmental terms are meaningless scientifically, some terms, such as "useful," do have significant implication, namely, patterns should reflect real world phenomena; *this is critically important*. For example, the following assertions,

all data are ended with 5 characters "z"

all tuples are ended with 111 (a numerical constant)

are obviously patterns of data. The frequency is indeed high, however, we are not interested in such *pure patter of data*. In the current state of arts, the exclusion of such patterns is relied on human judgments. Naturally one would ask

Could we exclude them scientifically?

Using data model, it is rather difficult; it involve the notion of *invariance of attribute transformations;* see appendix We settle this question rather easily by taking an alternative approach; first, we propose

Data mining is to find *real world* patterns in the portion of the real world that the data is represented about.

This paper was presented at 25th IEEE computer software and application conference, Chicago, Oct 8-12, 2001 under the title "The Lattice Structure of Database and Mining Multiple Level Rules." Due to some error, the text was not in the proceedings. The verbatim copy will appear in the Bulletin of International Rough Set Society (no copy right). Here is the first publication with revised title.

Data Mining and Knowledge Discovery: Theory, Tools, and Technology IV, Belur V. Dasarathy, Editor,
Proceedings of SPIE Vol. 4730 (2002) © 2002 SPIE · 0277-786X/02/$15.00

To use such a notion of data mining, a mathematical model of real world is needed [7, 8, 9, 11]; this is the main topic of next section.

2. MODOLING THE REAL WORLD

How should we model the real world? One of the oldest models of real world is the model theory of first order logic. Briefly, a model is, among others, a Cantor set (of real world entities) together with a relational structure [3]. We *will* consider the simplest one, namely, the relational structure consists of a finite set of equivalence relations. Interestingly, such a structure turns out to be the world model of relational database theory.

2.1. The Basic World Model of a Relation

A *relation is a knowledge representation* of a set of real world entities by tuples of attribute values. Note that such a representation is merely a reflection of certain intrinsic mathematical structure of the real world. In this section, we will examine such a structure.

We will illustrate the idea by example. Table 1 illustrates the details of a knowledge representation,

$$K: V \rightarrow (S\#, STATUS, CITY),$$

which maps entities to tuples: The right hand side (of the arrow) is a classical relation, and the left hand side is the set of real world entities. Mathematically, Table 1 is the graph, (x, K(x)) of the function K (the representation). Such a graph is called an *information table* (also known as information system or knowledge representation system) in rough set theory.

V	K	(S#	STATUS	CITY)
v_1	→	$(S_1$	TWENTY	$C_1)$
v_2	→	$(S_2$	TEN	$C_2)$
v_3	→	$(S_3$	TEN	$C_2)$
v_4	→	$(S_4$	TWENTY	$C_1)$
v_5	→	$(S_5$	THIRTY	$C_3)$

Table 1. An Information Table; arrows and parentheses will be suppressed

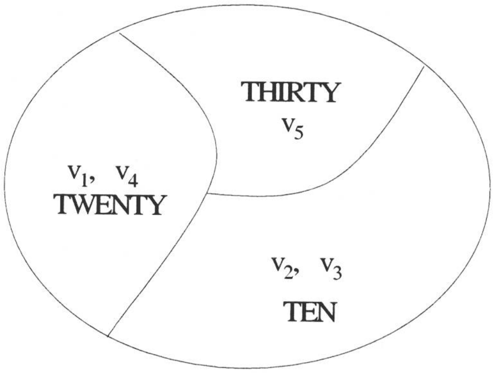

Figure 1. The Partition induced by STSTUS

Let us look at the STATUS in Table 1. The three distinct attribute values classify the real world entities into three distinct classes; see Figure 1.

Equivalence. Class	Attribute value (meaningful name)	Attribute
*	identity equiv. relation	S#
v_1, v_4	TWENTY	STATUS
v_2, v_3	TEN	
v_5	THIRTY	
v_1,v_4	C_1	CITY
v_2,v_3	C_2	
v_5	C_3	

Table 2. The three partitions induced from Table 1

Such a classification, mathematically defines a partition or an equivalence relation; note that a partition induces an equivalence relation and vice versa. By apply similar arguments to each attribute, we have three equivalence relations; see Table 2.

This illustration can easily be extended to general cases: Let K be a relation that represents the set V of real world entities. Let $A =\{A^1, A^2, \ldots, A^n\}$ be the set of attributes.

Proposition 2.1.1. Any subset of A, in particular, each attribute A^j induces a partition (an equivalence relation) on V, in particular the equivalence relation A_E^j.

Note that different A^j‘s may give the same A_E^i’s. Let $A_E =\{A_E^1, A_E^2, \ldots, A_E^n\}$ be the bag of (not necessary distinct) equivalence relations. We will regard attributes as the names of equivalence relations, that is, A^j = NAME(A_E^j), and attribute values the names of equivalence classes. When such names are given, we say it is a named equivalence relation. A_E is the set of *named* equivalence relations.

Definition 2.1.1. The pair (V, A_E) is the *basic world model (BWM)* of the given relation K.

BWM is the model in the first order logic, in which the relational structure is a finite set of equivalence relations. We believe the "correct" place to do the data mining is the (V, A_E), not the relation K which is merely a knowledge representation of BWM. From different consideration, we also reached BWM-ing and called it machine oriented modeling [9, 11].

Such phenomena were observed independently by Z. Pawlak (1982) and T. T. Lee (1983) [5], [18]. Pawlak called the pair (V, A_E) a knowledge base or an approximation space, if A_E has only one equivalence relation. The theory is called rough set theory; it is essentially a computing of granules [12, 20].

T.T. Lee took global view. Instead of granules, he considered the partitions. He expressed many relational database concepts by some notions of partitions, e.g., functional dependency by refinement We combine both approaches.

2.2. Relation Lattice

The power set 2^A of forms a lattice (Boolean algebra), where meet and join operations are the set theoretical union and intersection respectively; please notice the *twist* from the common usage. Let Π(V) be the lattice of all possible equivalence relations (partitions) on V, where meet is the intersection of equivalence relations and join is the "union;" the "union" $\cup A_E^j$ is the smallest refinement of all A_E^j, j= 1, 2, . . . Π(V) is called a partition lattice by Lee Proposition 2.1.1 implies

Proposition 2.2.1. There is a map

$$\theta: 2^A \rightarrow \Pi(V),$$

that respects the meet, but not the join, operations.

Definition 2.2.1. The image Imθ was called the *relation lattice* by Lee; we will denote it by $L(A_E)$ or simply $L(A)$; see Convention 2.3.3.

Lee noted that

1. The join in $L(A)$ is different from that of $\Pi(V)$.
2. So $L(A)$ is a subset, but not a sublattice, of $\Pi(V)$.

The latter one is the point of our departure: The fact that $L(E)$ is not a sublattice implies that the structure of $L(E)$ is incomplete; we will explore its full structure.

2.3. Multi-level World Models and Generalized Relation Lattice

In the Appendix, we show that a concept hierarchy of a given attribute A_E^j is a conceptual naming scheme of a finite chain of refinements of A_E^j, where $\theta(A^j) = A_E^j$. A refinement of A_E^j will be referred to as a *derived attribute*, denoted by $D(A_E^j)$; it is a partition of the base concepts (attribute values). A base concept is a name of an equivalence class of A_E^j, and a high level concept is a name of an equivalence class (granule) of base concepts and hence is an equivalence class of V, whose equivalence relation is coarser than A_E^j.

Definition 2.3.1. The smallest sublattice of $\Pi(V)$, that contains Imθ, is called the *generalized relation lattice* (it is a relation lattice of generalized concepts) and will be denoted by $L(A_E^*)$ or simply $L(A^*)$; see Convention 2.3.3.

The intent of the notation will be clear soon. Let X be a set of (or an) equivalence relation(s) on V. We will use $D(X)$, X^* and X^+ to denote a refinement, the set of all refinements (including X itself) and strict refinements (not including X itself) of X respectively. Each (strict) refinement will be called a *(strictly) derived equivalence relation.* A_E^{i*} and A_E^{i+} are the set of all derived and strictly derived equivalence relations of A_E^i, $i = 1, 2, \ldots, m$. A_E^* and A_E^+ consists of all derived and strictly derived equivalence relations of A_E; note that $A_E^* \supseteq \cup_i A_E^{i*}$. And that $A_E^+ \supseteq \cup_i A_E^{i+}$.

Proposition 2.3.1. The smallest sublattice of $\Pi(V)$, which contains E^*, is $L(A_E^*)$.

Proof: Let H be the smallest lattice that contains A_E^{i*}, $i = 1, 2, \ldots, m$, hence it contains A_E^*. Further, note that each $\theta(A^i) = A_E^i$ is contained in H and Imθ consists of all possible meets of $\theta(A^i)$, $i = 1, 2, \ldots, n$, so Imθ is contained in H. Obviously H is a sublattice of $L(A_E^*)$, by definition, $H= L(A_E^*)$. Q.E.D.

If each equivalence class of a refinement is given a conceptual name, then we have

Corollary 2.3.2. $L(A_E^*)$ contains all possible non-equivalent concept hierarchies of every attribute in A.

Convention. 2.3.3. We will suppress the subscript E from now on, so A^i means an attribute as well as the corresponding equivalence relation and A^* is the set of all refinements of A or derived attributes; see section 4.

Definitions

2.3.2. The pair (V, A^*) is called the multi-level world *model (MWM)* of the given relation K.
2.3.3. The pair (V, A^+) is called the high level world *model (HWM)* of the given relation K.

2.3.4. An equivalence class of any partition in L(A*) will be referred to as *a multi-level granule*; that of L(A) *a primary level granule*; that of E *a base granule*; and that of A^{i*} , i=1,2, . . . *a high level base granule*. A general term, granule, refers to one of the above.

A database can be viewed as a logic system [18, 19]. If we replace all multi-level base granules by their names, then lattice expressions become logic formulas.

Proposition 2.3.4. A concept in L(A*) can be expressed by a logic formula of multi-level base concepts.

3. DATA MINING VIA WORLD MODEL

Main Theorem All possible patterns in the given relation K are the granules in L(A*) whose cardinality is above certain threshhold.

From the prospect of this theorem, the goal to find all possible association rules in L(A) (the classical apriori algorithms) may not be a reasonable goal, since L(A), evwn not a sublattice of L(A*), is not a natural scope to search. We would rather serach in L(A*) or any of a sublattice. For a similar reason, the classical AOG try to find all possible rules in a given refinement chain of A may not be nature either? These will be the future research issues.

3.1. Association rules - high frequency patterns

Association rules are the primary examples of high frequency patterns. We will see in basic world model, the frequency of occurrences is translated into the cardinal number of some granules (equivalence classes) in L(A*).

Let us set up some notations: We will use the lower case b to denote an attribute value, and the upper case B the corresponding granule, that is, b = NAME(B). Observed that frequency of b_i is Card(B_i), the cardinal number of B_i, and frequency of b =(b_1, b_2 . . . b_n) is Card(B), where $B = \cap_i B_i$.

Theorem

3.1.1. b_i, is an association rule (of length 1), if Card(B_i) $\geq$ threshhold.

3.1.2. A subtuple b is an association rules (of length n), if Card(B) $\geq$ threshhold,

3.1.3. A high level concept p is a multi-level association rules, if P is a multi-level granule in L(A*) and Card(P) $\geq$ threshhold.

To search for high/multilevel rules [2, 5] reduce to search the granules in L(A*). For more general form [15] (using binary relations instead of equivalence relations) the problem is harder. Note that if we replace the h-level concept q by the equivalence class of (h-1)-level concept x, that is, the equivalence class q = NAME([x]), then any (p, y) for $y \in q$ is said to be a soft rule [16].

Theorem 3.1.4 A logical formula of attribute values is a pattern (*generalized association rules*) iff the cardinal number (of the corresponding algebraic expression of granules) $\geq$ threshhold

This general notion of association rules is a rephrase of [9, Theorem 2.2.1]. It is a comprehensive formulation, all primary and high level rules are included. The algebraic *expression of granules* is referred to as the *meaning* of the corresponding logic formula.

3.2. Decision Rules - high confidence pattern

The primary examples of high confidence pattern are Pawlak's rough set and Quinlan's decision trees. Basically, both divide the attributes into two categories, condition and decision attributes. Such an information table is called a decision table. Both theories offer various efficient ways of defining the condition and decision partitions so that the former is a refinement to the latter one with minimal errors.