High Frequency Patterns in Data Mining

Tsau Young Lin
San Jose State University, USA

INTRODUCTION

The principal focus is to examine the foundation of association (rule) mining (AM) via granular computing (GrC). The main results is: The set of all high frequency patterns can be found by sloving linear inequalities within a polynomial time.

BACKGROUND

Some Foundation Issues in Data Mining

What is data mining? The following informal paraphrase of Fayad et al. (1996)'s definition seems quite universal:

> Deriving useful patterns from data.

The keys are data, patterns, derivation system, and useful-ness. We will examine critically the current practices of AM.

Some Basic Terms in Association Mining (AM)

In AM, two measures, support and confidence, are the main criteria. It is well known among researchers the support is the main hurdle, in other words, high frequency patterns are the main focus. AM is originated from the market basket data (Agrawal, 1993). However, we will be interested in AM for relational tables. For definitive, we assert:

1. A relational table is a bag relation, that is, repetitions of tuples are permissible (Garcia-Monila et al. 2002)
2. An item is an attribute value,
3. A q-itemset is a subtuple of length q,
4. A high frequency pattern of length q is a q-subtuple if its number of occurrences is greater than or equal to a given threshold.

Emerging Data Mining Method - Granular Computing

Bitmap index is a common notion in database theory. The advantage of bitmap representation is computationally efficient (Louis & Lin, 2000), and the drawback is the order of the table has to be fixed (Garcia-Molina, 2002). Based on granular computing, we propose a new method, called granular representations, that avoids this drawback. We will illustrate the idea by examples. The following example is modified from the text cited above (p. 702). A relational table K is viewed as a knowledge representation of a set V, called the universe, of real world entities by tuples of data; see Table 1.

A bitmap index for an attribute is a collection of bit-vectors, one for each possible value that may appear in the attribute. For the first attribute, BusinesSize (the amount of business in millions), the bitmap index would have nine bit-vectors. The first bit-vector, for value TWENTY, is 100011100, because the first, fifth, sixth, and seventh tuple have BusinesSize = TWENTY. The other two, for values TEN and THIRTY, are 011100000 and 000000011 respectively; Table 1 shows both the original

Table 1. K and B are isomorphic

V		BusinesSize	Bmonth	City		BusinesSize	Bmonth	City
v_1		TWENTY	MAR	NY		100011100	110011000	101000000
v_2		TEN	MAR	SJ		011100000	110011000	010011100
v_3		TEN	FEB	NY		011100000	001100000	101000000
v_4	K	TEN	FEB	LA		011100000	001100000	000100011
v_5	$\rightarrow$	TWENTY	MAR	SJ		100011100	110011000	010011100
v_6		TWENTY	MAR	SJ		100011100	110011000	010011100
v_7		TWENTY	APR	SJ		100011100	000000100	010011100
v_8		THIRTY	JAN	LA		000000011	000000011	000100011
v_9		THIRTY	JAN	LA		000000011	000000011	000100011
Relational Table K						Bitmap Table B		

Table 2a. Granular data model (GDM) for BusinesSize attribute

BusinesSize	Granular Representation	Bitmap Representation
TWENTY	$=\{v_1, v_5, v_6, v_7\}$	=100011100
TEN	$=\{v_2, v_3, v_4\}$	=011100000
THIRTY	$=\{v_8, v_9\}$	=000000011
	GDM in Granules	GDM in Bitmaps

Table 2b. Granular data model (GDM) for Bmonth attribute

Bmonth	Granular Representation	Bitmap Representation
Jan	$=\{v_8, v_9\}$	=000000011
Feb	$=\{v_3, v_4\}$	=001100000
Mar	$=\{v_1, v_2, v_5, v_6\}$	=110011000
APR	$=\{v_7\}$	=000000100
	GDM in Granules	GDM in Bitmaps

Table 2c. Granular data model (GDM) for CITY attribute

City	Granular Representation	Bitmap Representation
LA	$=\{v_4, v_8, v_9\}$	=000100011
NY	$=\{v_1, v_3\}$	={v1, v3}
SJ	$=\{v_2, v_5, v_6, v_7\}$	=010011100
	GDM in Granules	GDM in Bitmaps

Table 2. K and G are isomorphic

V		BusinesSize	Bmonth	City	BusinesSize	Bmonth	City
v_1		TWENTY	MAR	NY	{v1,v5,v6,v7}	{v1,v2,v5,v6}	{v1,v3}
v_2		TEN	MAR	SJ	{v2,v3,v4}	{v1,v2,v5,v6}	{v2,v5,v6,v7}
v_3		TEN	FEB	NY	{v2,v3,v4}	{v3,v4}	{v1,v3}
v_4	K	TEN	FEB	LA	{v2,v3,v4}	{v3,v4}	{v4,v8,v9}
v_5	$\rightarrow$	TWENTY	MAR	SJ	{v1,v5,v6,v7}	**{v1,v2,v5,v6}**	**{v2,v5,v6,v7}**
v_6		TWENTY	MAR	SJ	**{v1,v5,v6,v7}**	**{v1,v2,v5,v6}**	**{v2,v5,v6,v7}**
v_7		TWENTY	APR	SJ	**{v1,v5,v6,v7}**	$\{v_7\}$	**{v2,v5,v6,v7}**
v_8		THIRTY	JAN	LA	**{v8, v9}**	$\{v_8,v_9\}$	**{v4,v8,v9}**
v_9		THIRTY	JAN	LA	**{v8, v9}**	$\{v_8,v_9\}$	**{v4,v8,v9}**
Bag Relation K					GRANULR TABLE G		

table and bitmap table. Bmonth means Birth month; City means the location of the entities.

Next, we will interpret the bit-vectors in terms of set theory. A bit-vector can be viewed as a representation of a subset of V. For example, the bit-vector, 100011100, of BusinesSize = TWENTY says that the first, fifth, sixth, and seventh entities have been selected, in other words, the bit-vector represents the subset $\{v_1, v_5, v_6, v_7\}$. The other two bi-vectors, for values TEN and THIRTY, represent the subsets $\{v_2, v_3, v_4\}$ and $\{v_8, v_9\}$ respectively. We summarize such translations in Table 2a,b,c. and refer to these subsets as elementary granules.

Some easy observations:

1. The collection of elementary granules of an attribute (column) forms a partition, that is, all granules of this attribute are pairwise disjoint. This fact was observed by Pawlak (1982) and Tony Lee (1983).
2. From Tables 1 and 2, one can easily conclude that the relational table K, the bitmap table B and granular table G are isomorphic. Two tables are isomorphic if one can transform a table to the other by renaming all attribute values in a one-to-one fashion.

Related Content

Algebraic Reconstruction Technique in Image Reconstruction Based on Data Mining
Zhong Qu (2008). *Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications (pp. 3493-3508).*
www.irma-international.org/chapter/algebraic-reconstruction-technique-image-reconstruction/7845

Combining Induction Methods with the Multimethod Approach
Mitja Lenic, Peter Kokol, Petra Povalejand Milan Zorman (2005). *Encyclopedia of Data Warehousing and Mining (pp. 184-189).*
www.irma-international.org/chapter/combining-induction-methods-multimethod-approach/10590

Predicting Resource Usage for Capital Efficient Marketing
D. R. Mani, Andrew L. Betzand James H. Drew (2005). *Encyclopedia of Data Warehousing and Mining (pp. 912-920).*
www.irma-international.org/chapter/predicting-resource-usage-capital-efficient/10726

Data Mining In the Federal Government
Les Pang (2005). *Encyclopedia of Data Warehousing and Mining (pp. 268-271).*
www.irma-international.org/chapter/data-mining-federal-government/10605

A Framework for Data Warehousing and Mining in Sensor Stream Application Domains
Nan Jiang (2010). *Evolving Application Domains of Data Warehousing and Mining: Trends and Solutions (pp. 113-128).*
www.irma-international.org/chapter/framework-data-warehousing-mining-sensor/38221