Data Mining of Medical Data: Opportunities and Challenges in Mining Association Rules

Dan A. Simovici
University of Massachusetts Boston

> The beginning clinical clerk, the house officer and the practicing physician are all confronted with conditions that are frustrating in every phase of medical action. ... To deal effectively with these frustrations it will be necessary to develop a more organized approach to the medical record, a more rational acceptance and use of the paramedical personnel and a more positive attitude about the computer in medicine.
> *L. L. Weed: Medical records that guide and teach, New England Journal of Medicine, 1968*

Abstract

Association rules represent knowledge embedded in data sets as probabilistic implications and are intimately related to computation of frequent item sets. We survey applications of frequent item sets and association rules in medical practice in such areas as nosocomial infections, adverse drug reactions, and the interplay between co-morbidities and the lack of transitivity of association rules.

To make this survey as self-content as possible we present in an appendix the Fisher exact test and the χ^2-test, enumeration of subsets, frequent item sets and the Apriori algorithm, and combinatorial properties of association rules.

Keywords: item sets, transactions, support, confidence, Apriori algorithm, nosocomial infections, adverse drug reactions

1 Introduction

Data Mining (DM) is the process that discovers new patterns embedded in large data sets. DM makes use of this information to build predictive models. DM is grounded in artificial intelligence, databases, and statistics.

The health care industry requires the use of DM because of it generates huge and complex volumes of data. Thus, un-automated analysis has become both expensive and impractical. The existence of insurance fraud and abuse

impels insurers to use DM. DM can generate information that can be useful to all stakeholders in health care, including patients by identifying effective treatments and best practices.

DM came into prominence in mid 90s because computers made possible the fast construction of huge data warehouses, containing potentially large amounts of information. The modern day statistical techniques and the advances in probability theory offered the necessary analytical tools.

The history of data and its contents is much older. Huge collections of data were built over hundreds and thousands of years by various forms of government and scientists. A famous case is the vast collection of very accurate planetary observations of the Danish astronomer Tycho Brahe (Dec. 14, 1546, Knutstorp Castle - Oct. 24, 1601, Prague). The knowledge embedded in this data - the laws of the movements of the planets were discovered by his successor Johannes Kepler (Dec. 27, 1571, Weil der Stadt -Nov. 15, 1630, Regensburg) and were confirmed by the work of Newton.

The main DM activities consist of description and visualization, seeking associations between data elements, grouping data into sets of similar records (a process known as clustering), data classification, prediction based on trends that can be extracted from data, etc.

DM applications in health care are numerous and already well established: evaluating treatment effectiveness, health care management, the analysis of relationships between patients and providers of care, pharmacovigilance, fraud and abuse detection. Despite the obvious benefits, there exist many limitations and difficulties in adapting DM analysis techniques.

DM can be limited by the accessibility to data that often is distributed in different settings (clinical, administrative, insurers, labs, etc.). Data may be incomplete, corrupted, noisy, or inconsistent. There exist ethical, legal and social issues (data ownership, privacy concerns).

Many patterns find in DM may be the result of random fluctuations, so many such patterns may be useless, which requires a serious statistical analysis.

DM of medical data requires specific medical knowledge as well as knowledge of DM technology and, last but not least, DM requires institutional commitment and funding.

In this survey we will focus on exploring data in the pursuit of association rules, a concept that formalizes probabilistic dependencies between parts of data. We begin by introducing the notion of *table*, the main data structure involved in this process. Then, we introduce formally association rules and their main parameters, support and confidence.

We discuss the use of AR in the study of nosocomial infections, adverse drug reactions and issues related to the lack of transitivity of association rules which are relevant for medical applications. To avoid interruptions in the flow of ideas of the paper we relegated the technicalities in an appendix.

2 Tables and relational databases

A table is an aggregate that consists of

- a *table name*;
- a *heading* that contains a set $A_1, \ldots, A_n$ of symbols called *attributes*;
- a *content* that is a multiset of rows: we could have multiple copies of the same row;
- each attribute A has a *domain* $\mathrm{Dom}(A_i)$, a set that contains at least two elements;
- a row is a sequence of values $(a_1, \ldots, a_n)$, such that a_i is a member of $\mathrm{Dom}(A_i)$ for $1 \leq i \leq n$.

Graphically, a table looks exactly as we would expect (see Figure 1).

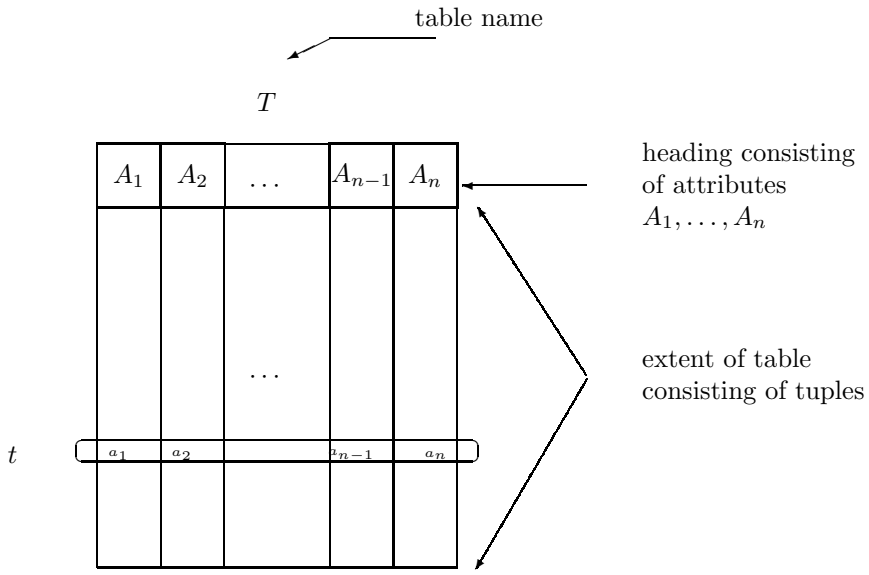

Figure 1: A Table and Its Constituents

Example 2.1 The heading of the table shown in Figure 2 consists of 5 attributes. The domains of attributes *shape* and *color* are

$$\begin{aligned} \mathrm{Dom}(\text{shape}) &= \{\text{cube, sphere, pyramid}\}, \\ \mathrm{Dom}(\text{color} &= \{\text{red, blue}\} \end{aligned}$$

have no natural ordering. Such domains are said to be *categorial*.

OBJECTS

	shape	length	width	height	color
1	cube	5	5	5	red
2	sphere	3	3	3	blue
3	pyramid	5	6	4	blue
4	cube	2	2	2	red
5	sphere	3	3	3	blue

Figure 2: Table containing categorial and numerical attributes

The domains of the remaining attributes, *length, width, height* are numerical. □

A special role is played by *binary tables* due to their capabilities for representing collections of sets. In such tables the domain of every attribute is the set $\{0, 1\}$ and every tuple is a sequence of 0s and 1s.

Let $S = \{s_1, \ldots, s_n\}$ be a set and let T be a subset of S. This subset can be represented by a sequence $(t_1, \ldots, t_n)$, where

$$t_i = \begin{cases} 1 & \text{if } t_i\text{is a member of } T, \\ 0 & \text{otherwise.} \end{cases}$$

Binary tables were used in analyzing purchase patterns of supermarket customers, documents containing words, etc. In the initial literature dealing with frequent item sets and association rules [4, 21, 5] the goal of this analysis was to determine what items people buy together, regardless of who they are.

Example 2.2 A fictional convenience store sells *milk*,*bread*,*butter*,*beer*, and *diapers*. The purchase records of seven customers are listed below

Customer basket	Content
1	{milk, bread, butter, diapers}
2	{bread, beer, diapers}
3	{milk, bread, butter, beer}
4	{bread, butter, diapers}
5	{milk, butter, beer, diapers}
6	{milk, butter}
7	{butter, beer}

and represented in the following binary table:

	milk	bread	butter	beer	diapers
t_1	1	1	1	0	1
t_2	0	1	0	1	1
t_3	1	1	1	1	0
t_4	0	1	1	0	1
t_5	1	0	1	1	1
t_6	1	0	1	0	0
t_7	0	0	1	1	0

This table indicates, for example, that the 4^{th} customer bought milk, bread, beer, and diapers. □

The tabular representation of collections of sets facilitates the introduction of the notion of support of an item set. If X is an item set, the *support* of X is the number of tuples that have 1s in all positions that correspond to the attributes of X. Equivalently, this is the number of baskets that contain X.

Example 2.3 For the transaction set defined in Example 2.2 we have

$$\begin{array}{ll} \mathsf{supp}(\text{milk}) = 4 & \mathsf{supp}(\text{bread}) = 4 \\ \mathsf{supp}(\text{milk bread}) = 2 & \mathsf{supp}(\text{milk bread butter}) = 2. \end{array}$$

□

Note that the larger the attribute set, the smaller the support: $X \subseteq Y$ implies $\mathsf{supp}(Y) \leq \mathsf{supp}(X)$. The number $\frac{\mathsf{supp}(X)}{N}$ estimates the probability that a randomly chosen transaction t contains all elements of X, where N is the total number of transactions.

Frequently, the support of an item set is expressed fractionally (as $\frac{\mathsf{supp}(X)}{N}$ or in percentages. For example, the relative value of the support of the set *milk bread* is $\frac{2}{7}$ or 28.57%.

3 Association Rules as Knowledge Embedded in Data

An *association rule* (AR) is a pair (X, Y) of sets of attributes, denoted by $X \to Y$. X is the *antecedent* and Y is the *consequent* of the rule $X \to Y$.

The simplest parameters associated to an AR are its support and confidence.

The *support of a rule* $X \to Y$ is the number of records that contain all items of X.

$$\mathsf{supp}(X \to Y) = \mathsf{supp}(X).$$

The *confidence* of $X \longrightarrow Y$ is the number

$$\mathsf{conf}(X \to Y) = \frac{\mathsf{supp}(XY)}{\mathsf{supp}(X)}.$$

Clearly, the confidence of $X \longrightarrow Y$ is an estimation of the probability that a record that contains the items of X, chosen at random, will contain the items of Y.

Example 3.1 For the table

	A	B	C	D	R
t_1	1	1	0	0	0
t_2	1	0	1	0	0
t_3	0	0	0	1	0
t_4	1	1	1	0	0
t_5	1	0	0	0	1
t_6	1	0	0	0	1
t_7	1	1	0	0	0

and the association rule $AB \to C$, support equals 3 and confidence is

$$\mathsf{conf}(AB \to C) = \frac{\mathsf{supp}(ABC)}{\mathsf{supp}(AB)} = \frac{1}{3} = 0.33$$

□

An AR $X \to Y$ holds with support μ and confidence c if $\mathsf{supp}(XY) \geq \mu$ and $\mathsf{conf}(X \to Y) \geq c$.

Association rules of the form $X \to Y$ with $Y \subseteq X$ are called *vacuous* because

$$\mathsf{conf}(X \longrightarrow Y) = \frac{\mathsf{supp}(XY)}{\mathsf{supp}(X)} = 1$$

regardless of the actual data set; such rules are not informative and they are also referred to as *trivial rules*.

Despite its simple formulation, finding association rules with a prescribed support and confidence can offer formidable computational challenges.

For data sets having n attributes there are $3^n - 2^n$ possible association rules (see Section D). Even for modest values of n the number of possible nontrivial association rules is very large and the number of collection of possible rules is immense. For $n = 20$ there exist more that one billion non-trivial AR and more that $10^{300000000}$ sets of AR (for comparison, there are 10^{80} atoms in the known universe!). Thus, a considerable effort in DM has been invested in designing efficient algorithms for computing association rules embedded in data sets.

To find an AR $X \to Y$ with support μ and confidence c we need to:

- find an item set U that is at least μ-frequent, that is, $\mathsf{supp}(U) \geq \mu$;
- find a subset V of U such that $\mathsf{supp}(V) \leq \frac{\mathsf{supp}(U)}{c}$.

The item sets U and V define the AR $X \to Y$, where $X = U - V$ and $Y = V$, having support at least equal to μ and confidence ar least equal to c. Thus, computing association rule amounts to computing frequent item sets.

The most common algorithm is the *Apriori algorithm* by Agrawal, Imielinski and Swami, which consists of the following main steps (see Section C):

- detect all items that that have a support at least equal to μ;
- for successive numbers $i \geq 2$ join item sets that contain i items with individual item sets (candidate generation phase);
- evaluate the resulting item sets and retain only those who have sufficient support (evaluation phase).

Without entering details, we mention that the algorithm raises non-trivial issues of memory management because often large data sets cannot be accommodated entirely in the main memory of computers. There are numerous references on Apriori implementations that examine these problems [24, 3, 31, 13, 2, 1].

4 Association Rules and Nosocomial Infections

The study of development of drug resistance of bacteria involved in intra-hospital infections has been pursued in [8, 10] and many other reports.

Among the Gram-negative bacteria which are notorious for their drug resistance, *Pseudomonas aeruginosa* is a common cause of infections in humans and its transmission is caused by medical equipment, including catheters.

The data collection includes records that describe single *Pseudomonas aeruginosa* isolates. The attributes of the records are

- date reported;
- source of isolate (sputum, blood);
- location of patient in the hospital;
- patient's home zip code;
- resistant (R), intermediate resistance (I), susceptible (S) for piperacillin, ticarcillin/clavulanate, ceftazidime, imipenem, amikacin, gentamicine, tobramycine, ciprofloxacin.

Records passed through a pre-processing phase, when duplicate records were removed, so each patient had one isolate per month. The system was designed to detect patterns of increasing resistance to antimicrobials; therefore, items of the form S-antimicrobial were removed.

Data is partitioned horizontally in time-slices; in each slice identification of association rules with high support is performed and the confidence of these rules is computed. Then, the variation of in confidence of a rule $X \rightarrow Y$ between the current time-slice and the confidence of the same rule in previous time slices is calculated. If a substantial increase in the confidence occurs (as verified using a statistical test described in Section A) relative to the previous partition(s), this finding constitutes an *event*.

More specifically, data was partitioned horizontally in

- A. 12 one-month fragments: 2,000 ARs;

- B. 4 three-month fragments: 12,000 ARs;
- C. 2 six-month fragments: 20,000 ARs.

Minimum support for an item was 2 and minimum support for an AR was 10.

The investigators Patterns sought short-lived interesting patterns in slices of type A, and long-lived interesting patterns in slices of type C.

A relatively small number of ARs were presented to the user as shown below:

Experiment	A	B	C
	34	57	28

Note that for AR of the form $\emptyset \rightarrow Y$ have

$$\mathsf{supp}(\emptyset \rightarrow Y) = \mathsf{supp}(\emptyset) = n;$$

and

$$\mathsf{conf}(\emptyset \rightarrow Y) = \frac{\mathsf{supp}(Y)}{n},$$

which shows that only confidence is significant and equals the probability of Y.

Thus, conf(R-antimicrobial) gives the probability that *Pseudomonas aeruginosa* develops resistance to the antimicrobial; variation in the level of confidence are evaluated on an monthly, quarterly, and semestrial basis.

The selection of association rules was based on the variance in their confidence as follows:

- For each AR $X \rightarrow Y$ the confidence in P_c, $\mathsf{conf}(X \rightarrow Y, P_c)$ was compared with $\mathsf{conf}(X \rightarrow Y, P_d)$, the confidence of $X \rightarrow Y$ in the last data set P_d in which $X \rightarrow Y$ was found which precedes P_c.
- The comparison of confidences is done using a χ^2-square comparison of two proportions, or when the number of expected value is small, by the Fisher exact test.
- If $\mathsf{conf}(X \rightarrow Y, P_c) \geq \mathsf{conf}(X \rightarrow Y, P_d)$ and the probability that the difference between the proportions occurred by chance is less than 5%, then this finding is presented to the user.

Among the AR found are the following:

$\emptyset \rightarrow$ R-ticarcillin/clavulanate R-ceftazidime R-piperacillin

a jump from 4%(Oct) to 8%(Nov) to 11%(Dec)suggests that the isolate is resistant to ticarcillin/clavulanate, ceftazidime and piperacillin

R-ceftazidime R-piperacillin $\rightarrow$ sputumR-ticarcillin/clavulanate

8%(Feb)-32%(Aug) it is likely that the isolate is from sputum and is ticarcillin resistent *given that* is resistant to ceftazidime and piperacillin

R-piperacillin $\rightarrow$ sputumR-ticarcillin/clavulanateR - ceftazidime

an increase from 6% (Q3) to 26% (Q4) in the probability that the isolate is from sputum, is ticarcillin/clavulanate and ceftazidime resistant *given that* is piperacillin resistant

R-ticarcillin/clavulanate $\longrightarrow$ sputumR-ceftazidimeR-piperacillin

an increase from 7% (Q3) to 24% (Q4) in the probability that isolate is from sputum, is ceftazidime and piperacillin resistant *given that* is ticarcilline/clavulanate resistant

R-ticarcillin/clavulanateR-ceftazidimeR-piperacillin $\longrightarrow$ sputum

an increase from 12% (Q3) to 42% (Q4) in the probability that the isolate is from sputum *given that* it is resistent to ticarcillin/clavulanate, ceftazidime, and piperacillin

5 Association Rules and Adverse Drug Reactions

Adverse drug reactions (ADE) pose a serious problem for the health of the public and cause wasteful expenses [30]. It is estimated that ADEs account for 5% of hospital admissions [18], 28% of af emergency department visits [20], and for 5% of hospital deaths [14]. In US only, ADEs result in losses of several billion dollars annually.

Due to their impact, ADE are monitored internationally in multiple sites. The Uppsala Monitoring Center in Sweden, a unit of the World Health Organization (WHO), mines data originating from individual case safety reports (ICSRs) and maintains *Vigibase*, a WHO case safety reporting database. Its activity started in 1978 and access to Vigibase is allowed for a fee.

At the Food and Drug Administration (FDA), a US federal unit, the AERS database (Adverse Event Reporting System) is maintained where access is free. Besides, proprietary ADE databases exists at various pharma entities who, by US law, must record adverse reactions to drugs.

We discuss the study performed in [11] and the observations of [30] on using association rules for mining ADE databases.

ADE can involve single or multiple drugs and describe single or multiple adverse reactions. The simplest association rule describing an ADE has the form Vioxx $\rightarrow$ heart attack and involves one drug and one reactions. Clearly, rule of this form cannot capture ADE that result from undesirable drug interactions and this is the focus of [11]. This study is based on a set of 162,744 reports of suspected ADEs reported to AERS and published in the year 2008. A total of 1167 multi-item ADE associations were identified.

An ADE database has certain unique characteristics that allow for more efficient mining algorithms. Namely, the set of items is partitioned into two classes: drugs and symptoms; association rules have the form $X \longrightarrow Y$, where X is a set of drugs and Y is a set of symptoms.

Given a set of drugs X it is important to find the largest set of symptoms Y such that $X \longrightarrow Y$ has a certain level of support and confidence. Indexing based on drugs and on symptoms was used to speed up searches in the data.

The general architecture of the AERS database is shown in Figure 3.

The attribute that binds various parts of the AERS database is *ISR*: the unique number for identifying an EARS report.

A taxonomy that characterizes the associations was developed based on a representative sample, as shown in Tables 1 and 2

67 percentages of potential multi-item ADE associations identified were characterized and clinically validated by a domain expert as previously recognized ADE associations.

Filtering of the rules was done based on interestingness measures (confidence is just one of them). Actually, in this case confidence is inappropriate since rules of the form $X \rightarrow$ NAUSEA will have high confidence due to the high frequency of NAUSEA.

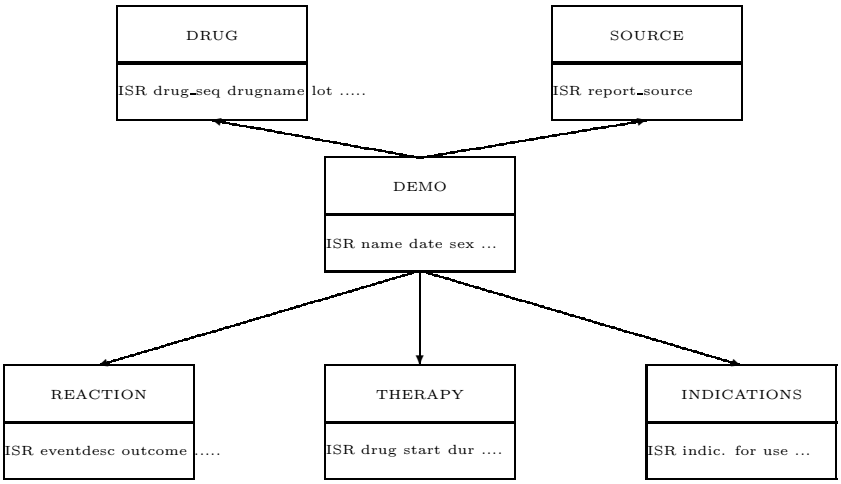

Figure 3: Architecture of AERS database

Table 1: Taxonomy of multi-item sets of drugs

1a	Drug-drug interactions found that are known	4%
1b	Drug-drug combinations known to be given together or treat same indication	78%
1c	Drug-drug combinations that seem to be due to confounding	9%
1d	Drug-drug interactions that are unknown	9%

Table 2: Taxonomy of multi-item ADE associations rules

2a	Associations (drug[s]-event) that are known	67%
2b	Associations (drug[s]-event) that are unknown	33%

Various alternatives for choosing an interest measure for association rules are studied extensively in data mining [25, 7, 17, 16, 12] and [15].

Let $X \to Y$ be an association rule. Denote the supports of the item sets $X \cap Y$, $X \cap \bar{Y}$, $\bar{X}X \cap Y$, and $\bar{X} \cap \bar{Y}$ by a, b, c, d, respectively The most commonly used interestingness measures for $X \to Y$ are given next.

Int. Measure	Formula
support	a
confidence	$\frac{a}{a+c}$
χ^2	$\frac{(ad-bc)^2(a+b+c+d)}{(a+b)(a+c)(b+c)(b+d)}$
interest (lift)	$\frac{a(a+b+c+d)}{a+b)^2}$
conviction	$\frac{(a+c)(b+d)}{(a+b+c+d)c}$

In [11] the interestingness measure used was the Relative Reporting Ratio (RR), defined by

$$\begin{aligned} \mathsf{RR} &= \frac{n \cdot \mathsf{supp}(X \cup Y)}{\mathsf{supp}(X) \cdot \mathsf{supp}(Y)} \\ &= \frac{(a+b+c+d)(n-d)}{(a+b)(a+c)}, \end{aligned}$$

where $n = a + b + c + d$ is the total number of records.

Note that RR can be written as

$$\mathsf{RR} = \frac{n \cdot \mathsf{supp}(X \cup Y)}{\mathsf{supp}(X) \cdot \mathsf{supp}(Y)} = \mathsf{conf}(X \to Y) \cdot \frac{n}{\mathsf{supp}}(Y)$$

and can be regarded as as the confidence of the rule $X \longrightarrow Y$ normalized by the relative support of the consequent Y. RR is symmetric relative to X and Y.

Large values of RR suggest that the occurrence of drugs-adverse reactions is larger than in the general collection of drugs.

A sample of multi-item ADE associations found in [11] is:

1a-2a	metformin metoprolol → NAUSEA	50	7.4
1b-2a	cyclophosphamide, prednisone, vincristine → FEBRILE NEUTROPENIA	78	45
1c-2a	cyclophosphamide, doxorubicin, prednisone, rituximab → FEBRILE NEUTROPENIA	63	59
1b-2b	atorvastatin, lisinopril → DYSPNOEA	55	3.5
1a-2b	omeprazole simvastatin → DYSPNOEA	58	12
1d-2b	varenicline darvocet → ABNORMAL DREAMS, FATIGUE, INSOMNIA,MEMORY IMPAIRMENT, NAUSEA	52	2668

Since each metformin and metoprolol cause nausea, association rules of the form metformin metoprolol → NAUSEA is forseable. The rule

cyclophosphamide prednisone vincristine → FEBRILE NEUTROPENIA

involves a drug combination used in cancer treatment and describes a known complication.

Similar conclusions are obtained for a variety of combination of other drugs.

6 Transitivity of Association Rules

A study of interactions between medications, laboratory results and problems using association rules was done by Wright, Chen, and Maloney at BWH in Boston [28]. The data examined included 100,000 patients. Encoding of problems, laboratory results, and medications was done using proprietary terminologies.

The importance of this study is that it highlighted difficulties of inferences involving probabilistic implications expressed by association rules. The authors noted that certain association rule occur with un unjustified high level of confidence. A typical example is the AR

$$\text{insulin} \rightarrow \text{ hypertension},$$

which involves unrelated terms. The explanation is the existence of co-morbidities, in this case, diabetes and hypertension, which highlights the need of mining for co-morbidities. It is shown that item sets such that

p_1	{lisinopril, multivitamin, hypertension}
p_2	{insulin, metformin, lisinopril,diabetes, hypertension}
p_3	{insulin, diabetes}
p_4	{metformin, diabetes}
p_5	{metformin, polycystic ovarian syndrome}
$\vdots$	$\vdots$

occur with a high level of support.

The difficulty of analyzing such association rules comes from the fact that association rules do not enjoy transitivity. This means that if $X \rightarrow Y$ and $Y \rightarrow Z$ are association rules with known confidence no conclusion can be drawn about the confidence of $X \rightarrow Z$.

Example 6.1 For the data set

	A	B	C
t_1	1	1	0
t_1	0	1	1

and the association rules $A \rightarrow B$ and $B \rightarrow C$ we have

$$\mathsf{supp}(A \rightarrow B) = 50\%, \mathsf{conf}(A \rightarrow B) = 100\%,$$
$$\mathsf{supp}(B \rightarrow C) = 100\%, \mathsf{conf}(B \rightarrow C) = 50\%.$$

but

$$\mathsf{supp}(A \rightarrow C) = 50\% \text{ and } \mathsf{conf}(A \rightarrow C) = 0\%.$$

On the other hand, for the data set

	A	B	C
t_1	1	0	1
t_1	0	1	0

and $A \to B$ and $B \to C$ we have

$$\begin{aligned} \mathsf{supp}(A \to B) = 50\%, \mathsf{conf}(A \to B) = 0\%, \\ \mathsf{supp}(B \to C) = 50\%, \mathsf{conf}(B \to C) = 0\%. \end{aligned}$$

but

$$\mathsf{supp}(A \to C) = 50\% \text{ and } \mathsf{conf}(A \to C) = 100\%.$$

So, the confidence of $A \to C$ is unrelated to either $\mathsf{conf}(A \to B)$ or to $\mathsf{conf}(B \to C)$. □

To deal with the lack of transitivity, it is necessary to investigate association rules of the form $X \to Z$ starting from existent AR $X \to Y$ and $Y \to Z$ which have a satisfactory medical interpretation. This is the point of view espoused in [27] who present their software *TransMiner*.

The reverse approach is adopted in [28]: starting from an association rule $X \to Z$ (e.g. insulin $\to$ hypertension they seek to identify candidate item sets Y such that $X \to Y$ and $Y \to Z$ are plausible association rules. Y could be diabetes or other co-morbidities of hypertension; once these cases are excluded the confidence of insulin $\to$ hypertension decreases sharply.

7 Conclusions and Open Problems

DM cannot replace the human factor in medical research; however it can be a precious instrument in epidemiology, pharmacovigilance. Interaction between DM and medical research is beneficial for both domains; biology and medicine suggest novel problems for data mining and machine learning.

Many open problems remain to be resolved. We estimate that extending association mining to unstructured data (progress reports, radiology reports, operative notes, outpatient notes), integration of "gold standards" in evaluation of AR extracted from medical practice, developing information-theoretical techniques for AR evaluation will attract the interest of both data miners and medical researchers because of their potential benefits in the practice of medicine. We conclude with a quotation from [29] written 29 years after the motto of this paper:

> Knowledge should be held in tools that are kept up to date and used routinely–not in heads, which are expensive to load and faulty in the retention and processing of knowledge.
>
> *L.L. Weed, M.D.: New connections between medical knowledge and patient care, British Medical Journal, 1997*

A Fisher Exact Test and the χ^2-Test

Let X and Y be two categorical random variables that assume the values $x_1, \ldots, x_m$ and $y_1, \ldots, y_n$. Consider a matrix A with m rows and n columns, where $a_{ij} \in \mathbb{N}$ is the number of times the pair (x_i, y_j) occurs in an experiment.

Let R_i and C_j be random variables (for $1 \leq i \leq m$ and $1 \leq j \leq n$) that represent the sum of the elements of row i and the sum of the elements of column j, respectively. Clearly,

$$\sum_{i=1}^{m} R_i = \sum_{j=1}^{n} C_j = \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij} = N.$$

The conditional probability $P(A = (a_{ij}) \mid R_i = r_i, C_j = c_j)$ is given by

$$P(A = (a_{ij}) \mid R_i = r_i, C_j = c_j) = \frac{r_1! \cdots r_m! c_1! \cdots c_n!}{N! \prod_{i=1}^{m} \prod_{j=1}^{n} a_{ij}!}$$

This discrete distribution is a generalization of the hypergeometric distribution.

In the special case $m = n = 2$ we have the matrix

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix},$$

$r_1 = a_{11} + a_{12}$, $r_2 = a_{21} + a_{22}$, and $c_1 = a_{11} + a_{21}$, $c_2 = a_{12} + a_{22}$. The probability $P(A|(R_i = r_i, C_j = c_j)$ is

$$\begin{aligned} & P(A|R_1 = r_1, R_2 = r_2, C_1 = c_1, C_2 = c_2) \\ = \; & \frac{r_1! r_2! c_1! c_2!}{N! a_{11}! a_{12}! a_{21}! a_{22}!} \\ = \; & \frac{(a_{11} + a_{12})!(a_{21} + a_{22})!(a_{11} + a_{21})!(a_{12} + a_{22})!}{N! a_{11}! a_{12}! a_{21}! a_{22}!} \\ = \; & \frac{\binom{a_{11}+a_{12}}{a_{11}} \binom{a_{21}+a_{22}}{a_{21}}}{\binom{n}{a_{11}+a_{21}}} = \frac{\binom{r_1}{a_{11}} \binom{r_2}{a_{21}}}{\binom{n}{c_1}}. \end{aligned}$$

The probability of getting the actual matrix given the particular values of the row and column sums is known as the *cutoff probability* P_{cutoff}.

Example A.1 On a certain day two urology services $U1$ and $U2$ use general anesthesia and IV sedation in lithotripsy interventions as follows

	$U1$	$U2$	
gen. anesthesia	5	0	$r_1 = 5$
iv sedation	1	4	$r_2 = 5$
	$c_1 = 6$	$c_2 = 4$	

The null hypothesis here is that there exists a significant association between the urology department and the type of anesthesia it prefers for lithotripsy.

The matrices that correspond to the same marginal probability distributions and their corresponding probabilities are

$\begin{pmatrix} 5 & 0 \\ 1 & 4 \end{pmatrix}$	$\begin{pmatrix} 4 & 1 \\ 2 & 3 \end{pmatrix}$	$\begin{pmatrix} 3 & 2 \\ 3 & 2 \end{pmatrix}$	$\begin{pmatrix} 2 & 3 \\ 4 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & 4 \\ 5 & 0 \end{pmatrix}$
0.0238	0.2381	0.4762	0.2381	0.0238

The sum of these probabilities are 1, as expected equals 1 and the cutoff probability is 0.0238. The probability that results shown by the matrix

$$\begin{array}{ccccc} & U1 & U2 & \\ \text{gen. anesthesia} & 5 & 0 & r_1 = 5 \\ \text{iv sedation} & 1 & 4 & r_2 = 5 \\ & c_1 = 6 & c_2 = 4 & \end{array}$$

are randomly obtained is not larger than 0.0238, which allows us to conclude that U1 has indeed a strong preference for using general anesthesia, while the preference in U2 is for intravenous sedation. □

Example A.2 The exact Fisher test outlined in Example A.1 can be applied only if the expected values are no larger than 5. If this is not the case, we need to apply the approximate χ^2-test.

In a hospital the number of isolates resistent to ticarcillin/clavulanate, ceftazidime, and piperacillin during the third quarter equals 29; two of these isolates originate from sputum. In the third quarter, the number of isolates resistent to all three antibiotics is 34 and 8 of these originate from sputum.

This is presented by the matrix

$$\begin{array}{ccccc} & Q3 & Q4 & \\ \text{non-sputum} & 27 & 26 & r_1 = 53 \\ \text{sputum} & 2 & 8 & r_2 = 10 \\ & c_1 = 29 & c_2 = 34 & 63 \end{array}$$

We need to ascertain whether the larger proportion of resistant bacteria in sputum in the 4^{th} quarter reflect something other that statistical variability.

The expected values of the observations computed from the marginal values are

$$\begin{array}{ccccc} & Q3 & Q4 & \\ \text{non-sputum} & \frac{53*29}{63} & \frac{53*34}{63} & r_1 = 53 \\ \text{sputum} & \frac{10*2}{63} & \frac{10*8}{63} & r_2 = 10 \\ & c_1 = 29 & c_2 = 34 & \end{array} =$$

$$\begin{array}{ccccc} & Q3 & Q4 & \\ \text{non-sputum} & 24.39 & 28.60 & r_1 = 53 \\ \text{sputum} & 0.31 & 1.27 & r_2 = 10 \\ & c_1 = 29 & c_2 = 34 & \end{array}$$

The χ^2-square criterion is computed as

$$\chi^2 = \sum_{i,j} \frac{(|o_{ij} - e_{ij}| - 0.5)^2}{e_{ij}},$$

where the term 0.5 is a correction for continuity. In our case

$$\begin{aligned}\chi^2 &= \frac{(|27-24.39|-0.5)^2}{24.39}+\frac{(|26-28.60|-0.5)^2}{28.60}\\ &\quad+\frac{(|2-0.31|-0.5)^2}{0.31}+\frac{(|8-1.27|-0.5)^2}{1.27}\\ &= \frac{1.71^2}{0.31}+\frac{2.10^2}{28.60}+\frac{1.19^2}{0.31}+\frac{6.23^2}{1.27}=35.40.\end{aligned}$$

In this case the χ^2 variable has one degree of freedom and the value obtained is highly significant at 0.001 level. Thus, we can conclude that the variation in the confidence level of the rule from the third to fourth quarter can be accepted with a high degree of confidence. □

B Enumeration of Subsets of Sets

A systematic technique for enumerating the subsets of a set was introduced in [22] by R. Rymon in order to provide a unified search-based framework for several problems in artificial intelligence; this technique is especially useful in data mining.

Let S be a set, $S = \{i_1, \ldots, i_n\}$. The *Rymon tree* of S is defined as follows:

1. the root of the tree is the empty set, and
2. the children of a node P are the sets of the form $P \cup \{s_i \mid i > \max\{j \mid s_j \in P\}$.

Example B.1 Let $S = \{i_1, i_2, i_3, i_4\}$. The Rymon tree for $\mathcal{C}$ and d is shown in Figure 4. □

The key property of a Rymon tree of a finite set S is that every subset of S occurs exactly once in the tree.

Also, observe that in the Rymon tree of a collection of the form $\mathcal{P}(S)$, the collection of sets of $\mathcal{S}_r$ that consists of sets located at distance r from the root denotes all $\binom{n}{r}$ subsets of size r of S.

C Frequent Item Sets and the Apriori Algorithm

Suppose that I is a finite set; we refer to the elements of I as *items*.

Definition C.1 A *transaction data set on* I is a function $T : \{1, \ldots, n\} \longrightarrow \mathcal{P}(I)$. The set $T(k)$ is the *k-th transaction of* T. The numbers $1, \ldots, n$ are the transaction identifiers (*tid*s). □

Example 2.2 shows that a transaction is the set of items present in the shopping cart of a consumer that completed a purchase in a store and that the data set is a collection of such transactions.

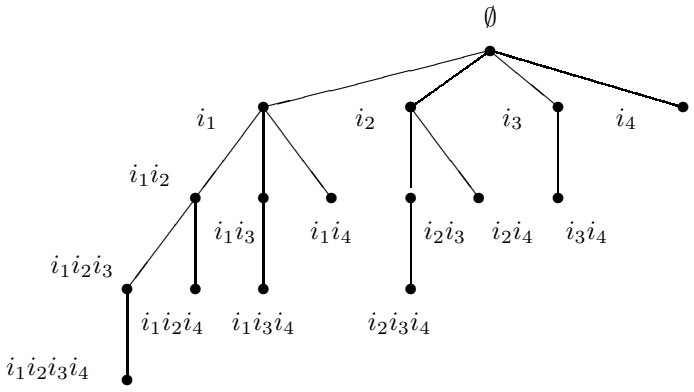

Figure 4: Rymon Tree for $\mathcal{P}(\{i_1, i_2, i_3, i_4\})$

Example C.2 Let $I = \{i_1, i_2, i_3, i_4\}$ be a collection of items. Consider the transaction data set T given by:

$$\begin{array}{rcl} T(1) & = & \{i_1, i_2\}, \\ T(2) & = & \{i_1, i_3\}, \\ T(3) & = & \{i_1, i_2, i_4\}, \\ T(4) & = & \{i_1, i_3, i_4\}, \\ T(5) & = & \{i_1, i_2\}, \\ T(6) & = & \{i_3, i_4\}. \end{array}$$

Thus, the support of the item set $\{i_1, i_2\}$ is 3; similarly, the support of the item set $\{i_1, i_3\}$ is 2. Therefore, the relative supports of these sets are $\frac{1}{2}$ and $\frac{1}{3}$, respectively. □

The following rather straightforward statement is fundamental for the study of frequent item sets.

Theorem C.3 *Let $T : \{1, \ldots, n\} \longrightarrow \mathcal{P}(I)$ be a transaction data set on a set of items I. If K and K' are two item sets, then $K' \subseteq K$ implies $\mathsf{supp}_T(K') \geq \mathsf{supp}_T(K)$.*

Proof. Note that every transaction that contains K also contains K'. The statement follows immediately. ∎

If we seek those item sets that enjoy a minimum support level relative to a transaction data set T, then it is natural to start the process with the smallest non-empty item sets.

The support of an item set enjoys the property of *supramodularity* [23]. Namely, if X, Y are two sets of items then

$$\mathsf{supp}(X) + \mathsf{supp}(Y) \leq \mathsf{supp}(X \cup Y) + \mathsf{supp}(X \cap Y).$$

Definition C.4 An item set K is *μ-frequent* relatively to the transaction data set T if $\mathsf{supp}_T(K) \geq \mu$.

We denote by $\mathcal{F}_T^\mu$ the collection of all μ-frequent item sets relative to the transaction data set T, and by $\mathcal{F}_{T,r}^\mu$ the collection of μ-frequent item sets that contain r items for $r \geq 1$. ∎

Note that

$$\mathcal{F}_T^\mu = \bigcup_{r \geq 1} \mathcal{F}_{T,r}^\mu.$$

If μ and T are clear from the context, then we may omit either or both adornments from this notation.

Let $I = \{i_1, \ldots, i_n\}$ be an item set that contains n elements.

Denote by $\mathcal{G}_I = (\mathcal{P}(I), E)$ the Rymon tree of $\mathcal{P}(I)$. The root of the tree is $\emptyset$. A vertex $K = \{i_{p_1}, \ldots, i_{p_k}\}$ with $i_{p_1} < i_{p_2} < \cdots < i_{p_k}$ has $n - i_{p_k}$ children $K \cup \{j\}$ where $i_{p_k} < j \leq n$.

Let $\mathcal{S}_r$ be the collection of item sets that have r elements. The next theorem suggests a technique for generating $\mathcal{S}_{r+1}$ starting from $\mathcal{S}_r$.

Theorem C.5 *Let $\mathcal{G}$ be the Rymon tree of $\mathcal{P}(I)$, where $I = \{i_1, \ldots, i_n\}$. If $W \in \mathcal{S}_{r+1}$, where $r \geq 2$, then there exists a unique pair of distinct sets $U, V \in \mathcal{S}_r$ that has a common immediate ancestor $T \in \mathcal{S}_{r-1}$ in $\mathcal{G}$ such that $U \cap V \in \mathcal{S}_{r-1}$ and $W = U \cup V$.*

Proof. Let u, v be the largest and the second largest subscript of an item that occurs in W, respectively. Consider the sets $U = W - \{u\}$ and $V = W - \{v\}$. Both sets belong to $\mathcal{S}_r$. Moreover, $Z = U \cap V$ belongs to $\mathcal{S}_{r-1}$ because it consists of the first $r-1$ elements of W. Note that both U and V are descendants of Z and that $U \cup V = W$.

The pair (U, V) is unique. Indeed, suppose that W can be obtained in the same manner from another pair of distinct sets $U', V' \in \mathcal{S}_r$, such that U', V' are immediate descendants of a set $Z' \in \mathcal{S}_{r-1}$. The definition of the Rymon tree $\mathcal{G}_I$ implies that $U' = Z' \cup \{i_m\}$ and $V' = Z' \cup \{i_q\}$, where the letters in Z' are indexed by number smaller than $\min\{m, q\}$. Then, Z' consists of the first $r-1$ symbols of W, so $Z' = Z$. If $m < q$, then m is the second highest index of a symbol in W and q is the highest index of a symbol in W, so $U' = U$ and $V' = V$. ∎

Example C.6 Consider the Rymon tree of the collection $\mathcal{P}(\{i_1, i_2, i_3, i_4)$ shown in Figure 4.

The set $\{i_1, i_3, i_4\}$ is the union of the sets $\{i_1, i_3\}$ and $\{i_1, i_4\}$ that have the common ancestor $\{i_1\}$. ∎

Next we discuss an algorithm that allows us to compute the collection $\mathcal{F}_T^\mu$ of all μ-frequent item sets for a transaction data set T. The algorithm is known as the *Apriori Algorithm*.

We begin with the procedure `apriori_gen` that starts with the collection $\mathcal{F}_{T,k}^\mu$ of frequent item sets for the transaction data set T that contain k elements and generates a collection $\mathcal{C}_{k+1}$ of sets of items that contains $\mathcal{F}_{T,k+1}^\mu$, the collection the frequent item sets that have $k+1$ elements. The justification of this procedure is based on the next statement.

Theorem C.7 *Let T be a transaction data set on a set of items I and let $k \in \mathbb{N}$ such that $k > 1$.*

If W is a μ-frequent item set and $|W| = k+1$, then, there exists a μ-frequent item set Z and two items i_m and i_q such that and $|Z| = k-1$, $Z \subseteq W$, $W = Z \cup \{i_m, i_q\}$ and both $Z \cup \{i_m\}$ and $Z \cup \{i_q\}$ are μ-frequent item sets.

Proof. If W is an item set such that $|W| = k+1$, then we already know that W is the union of two subsets U, V of I such that $|U| = |V| = k$ and that $Z = U \cap V$ has $k-1$ elements. Since W is a μ-frequent item set and Z, U, V are subsets of W it follows that each of theses sets is also a μ-frequent item set. ∎

Note that the reciprocal statement of Theorem C.7 is not true, as the next example shows.

Example C.8 Let T be the transaction data set introduced in Example C.2. Note that both $\{i_1, i_2\}$ and $\{i_1, i_3\}$ are $\frac{1}{3}$-frequent item sets; however,

$$\mathsf{supp}_T(\{i_1, i_2, i_3\}) = 0,$$

so $\{i_1, i_2, i_3\}$ fails to be a $\frac{1}{3}$-frequent item set. □

The procedure `apriori_gen` mentioned above is the algorithm 1. This procedure starts with the collection of item sets $\mathcal{F}_{T,k}$ and produces a collection of item sets $\mathcal{C}_{T,k+1}$ that includes the collection of item sets $\mathcal{F}_{T,k+1}$ of frequent item sets having $k+1$ elements.

Data: a minimum support μ, the collection $\mathcal{F}_{T,k}^\mu$ of frequent item sets having k elements
Result: the set of candidate frequent item sets $\mathcal{C}_{T,k+1}^\mu$
$j = 1$;
$\mathcal{C}_{T,j+1}^\mu = \emptyset$;
for $L, M \in \mathcal{F}_{T,k}^\mu$ *such that* $L \neq M$ *and* $L \cap M \in \mathcal{F}_{T,k-1}^\mu$ **do**
  add $L \cup M$ to $\mathcal{C}_{T,k+1}^\mu$;
end
remove all sets K in $\mathcal{C}_{T,k+1}^\mu$ where there is a subset of K containing k elements that does not belong to $\mathcal{F}_{T,k}^\mu$.

Algorithm 1: The Procedure `apriori_gen`

Note that in `apriori_gen` no access to the transaction data set is needed.

The *Apriori* Algorithm 2 operates on "levels". Each level k consists of a collection $\mathcal{C}^{\mu}_{T,k}$ of candidate item sets of μ-frequent item sets. To build the initial collection of candidate item sets $\mathcal{C}^{\mu}_{T,1}$ every single item set is considered for membership in $\mathcal{C}^{\mu}_{T,1}$. The initial set of frequent item set consists of those singletons that pass the minimal support test. The algorithm alternates between a candidate generation phase (accomplished by using `apriori_gen` and an evaluation phase which involve a data set scan and is, therefore, the most expensive component of the algorithm.

Data: a transaction data set T and a minimum support μ
Result: the collection $\mathcal{F}^{\mu}_{T}$ of μ-frequent item sets
$\mathcal{C}^{\mu}_{T,1} = \{\{i\} \mid i \in I\}$;
$i = 1$;
while $\mathcal{C}^{\mu}_{T,i} \neq \emptyset$ **do**
  /* evaluation phase */ $\mathcal{F}^{\mu}_{T,i} = \{L \in \mathcal{C}^{\mu}_{T,i} \mid \mathtt{supp}_T(L) \geq \mu\}$;
  /* candidate generation */ $\mathcal{C}^{\mu}_{T,i+1} = \mathtt{apriori\_gen}(\mathcal{F}^{\mu}_{T,i})$;
  $i++$;
end
return $\mathcal{F}^{\mu}_{T} = \bigcup_{j<i} \mathcal{F}^{\mu}_{T,j}$;

Algorithm 2: The Apriori Algorithm

Example C.9 Let T be the data set given by:

	i_1	i_2	i_3	i_4	i_5
$T(1)$	1	1	0	0	0
$T(2)$	0	1	1	0	0
$T(3)$	1	0	0	0	1
$T(4)$	1	0	0	0	1
$T(5)$	0	1	1	0	1
$T(6)$	1	1	1	1	1
$T(7)$	1	1	1	0	0
$T(8)$	0	1	1	1	1

The support counts of various subsets of $I = \{i_1, \ldots, i_5\}$ are given below:

i_1	i_2	i_3	i_4	i_5
5	6	5	2	5

i_1i_2	i_1i_3	i_1i_4	i_1i_5	i_2i_3	i_2i_4	i_2i_5	i_3i_4	i_3i_5	i_4i_5
3	2	1	3	5	2	3	2	3	2

$i_1i_2i_3$	$i_1i_2i_4$	$i_1i_2i_5$	$i_1i_3i_4$	$i_1i_3i_5$	$i_1i_4i_5$	$i_2i_3i_4$	$i_2i_3i_5$	$i_2i_4i_5$	$i_3i_4i_5$
2	1	1	1	1	1	2	3	2	2

$i_1i_2i_3i_4$	$i_1i_2i_3i_5$	$i_1i_2i_4i_5$	$i_1i_3i_4i_5$	$i_2i_3i_4i_5$
1	1	1	1	2

$i_1i_2i_3i_4i_5$
0

Starting with $\mu = 0.25$ and with $\mathcal{F}^{\mu}_{T,0} = \{\emptyset\}$ the Apriori Algorithm computes the following sequence of sets:

$$\begin{aligned}
\mathcal{C}^{\mu}_{T,1} &= \{i_1, i_2, i_3, i_4, i_5\},\\
\mathcal{F}^{\mu}_{T,1} &= \{i_1, i_2, i_3, i_4, i_5\},\\
\mathcal{C}^{\mu}_{T,2} &= \{i_1i_2, i_1i_3, i_1i_4, i_1i_5, i_2i_3, i_2i_4, i_2i_5, i_3i_4, i_3i_5, i_4i_5\},\\
\mathcal{F}^{\mu}_{T,2} &= \{i_1i_2, i_1i_3, i_1i_5, i_2i_3, i_2i_4, i_2i_5, i_3i_4, i_3i_5, i_4i_5\},\\
\mathcal{C}^{\mu}_{T,3} &= \{i_1i_2i_3, i_1i_2i_5, i_1i_3i_5, i_2i_3i_4, i_2i_3i_5, i_2i_4i_5, i_3i_4i_5\},\\
\mathcal{F}^{\mu}_{T,3} &= \{i_1i_2i_3, i_2i_3i_4, i_2i_3i_5, i_2i_4i_5, i_3i_4i_5\},\\
\mathcal{C}^{\mu}_{T,4} &= \{i_2i_3i_4i_5\},\\
\mathcal{F}^{\mu}_{T,4} &= \{i_2i_3i_4i_5\},\\
\mathcal{C}^{\mu}_{T,5} &= \emptyset.
\end{aligned}$$

Thus, the algorithm will output the collection:

$$\begin{aligned}
\mathcal{F}^{\mu}_{T} &= \bigcup_{i=1}^{4} \mathcal{F}^{\mu}_{T,i}\\
&= \{i_1, i_2, i_3, i_4, i_5, i_1i_2, i_1i_3, i_1i_5, i_2i_3, i_2i_4, i_2i_5, i_3i_4, i_3i_5, i_4i_5,\\
&\quad i_1i_2i_3, i_2i_3i_4, i_2i_3i_5, i_2i_4i_5, i_3i_4i_5, i_2i_3i_4i_5\}.
\end{aligned}$$

□

D Association Rules

Definition D.1 An *association rule* on an item set I is a pair of non-empty disjoint item sets (X, Y). □

Note that if $|I| = n$, then there exist $3^n - 2^{n+1} + 1$ association rules on I. Indeed, suppose that the set X contains k elements; there are $\binom{n}{k}$ ways of choosing X. Once X is chosen, Y can be chosen among the remaining $2^{n-k} - 1$ non-empty subsets of $I - X$. In other words, the number of association rules is:

$$\sum_{k=1}^{n} \binom{n}{k} (2^{n-k} - 1) = \sum_{k=1}^{n} \binom{n}{k} 2^{n-k} - \sum_{k=1}^{n} \binom{n}{k}.$$

By taking $x = 2$ in the equality:

$$(1 + x)^n = \sum_{k=0}^{n} \binom{n}{k} x^{n-k}$$

we obtain

$$\sum_{k=1}^{n} \binom{n}{k} 2^{n-k} = 3^n - 2^n.$$

Since $\sum_{k=1}^{n} \binom{n}{k} = 2^n - 1$, we obtain immediately the desired equality. The number of association rules can be quite considerable even for small values of n. For example, for $n = 10$ we have $3^{10} - 2^{11} + 1 = 57,002$ association rules.

An association rule (X, Y) is denoted by $X \to Y$. The confidence of $X \to Y$ is the number

$$\mathsf{conf}_T(X \to Y) = \frac{\mathsf{supp}_T(XY)}{\mathsf{supp}_T(X)}.$$

Definition D.2 An association rule *holds in a transaction data set* T with support μ and confidence c if $\mathsf{supp}_T(XY) \geq \mu$ and $\mathsf{conf}_T(X \to Y) \geq c$. □

Once a μ-frequent item set Z is identified we need to examine the support levels of the subsets X of Z to insure that an association rule of the form $X \to Z - X$ has a sufficient level of confidence, $\mathsf{conf}_T(X \to Z - X) = \frac{\mu}{\mathsf{supp}_T(X)}$. Observe that $\mathsf{supp}_T(X) \geq \mu$ because X is a subset of Z. To obtain a high level of confidence for $X \to Z - X$ the support of X must be as small as possible.

Clearly, if $X \to Z - X$ does not meet the level of confidence, then it is pointless to look rules of the form $X' \to Z - X'$ among the subsets X' of X.

Example D.3 Let T be the transaction data set introduced in Example C.9. We saw that the item set $L = i_2i_3i_4i_5$ has the support count equal to 2 and, therefore, $\mathsf{supp}_T(L) = 0.25$. This allows us to obtain the following association rules having three item sets in their antecedent which are subsets of L:

rule	$\mathsf{supp}_T(X)$	$\mathsf{conf}_T(X \to Y)$
$i_2i_3i_4 \to i_5$	2	1
$i_2i_3i_5 \to i_4$	3	$\frac{2}{3}$
$i_2i_4i_5 \to i_3$	2	1
$i_3i_4i_5 \to i_2$	2	1

Note that $i_2i_3i_4 \to i_5$, $i_2i_4i_5 \to i_3$, and $i_3i_4i_5 \to i_2$ have 100% confidence. We refer to such rules as *exact association rules*.

The rule $i_2i_3i_5 \to i_4$ has confidence $\frac{2}{3}$. It is clear that the confidence of rules of the form $U \to V$ with $U \subseteq i_2i_3i_5$ and $UV = L$ will be lower than $\frac{2}{3}$ since $\mathsf{supp}_T(U)$ is at least 3. Indeed, the possible rules of this form are:

rule	$\mathsf{supp}_T(X)$	$\mathsf{conf}_T(X \to Y)$
$i_2i_3 \to i_4i_5$	5	$\frac{2}{5}$
$i_2i_5 \to i_3i_4$	3	$\frac{2}{3}$
$i_3i_5 \to i_2i_4$	3	$\frac{2}{3}$
$i_2 \to i_3i_4i_5$	6	$\frac{2}{6}$
$i_3 \to i_2i_4i_5$	5	$\frac{2}{5}$
$i_5 \to i_2i_3i_4$	5	$\frac{2}{5}$

Obviously, if we seek association rules having a confidence larger than $\frac{2}{3}$ no such rule $U \to V$ can be found such that U is a subset of $i_2i_3i_5$.

Suppose, for example, that we seek association rules $U \to V$ that have a minimal confidence of 80%. We need to examine subsets U of the other sets:

$i_2i_3i_4$, $i_2i_4i_5$, or $i_3i_4i_5$, which are not subsets of $i_2i_3i_5$ (since the subsets of $i_2i_3i_5$ cannot yield levels of confidence higher than $\frac{2}{3}$. There are five such sets:

rule	$\mathsf{supp}_T(X)$	$\mathsf{conf}_T(X \rightarrow Y)$
$i_2i_4 \rightarrow i_3i_5$	2	1
$i_3i_4 \rightarrow i_2i_5$	2	1
$i_4i_5 \rightarrow i_2i_3$	2	1
$i_3i_4 \rightarrow i_2i_5$	2	1
$i_4 \rightarrow i_2i_3i_5$	2	1

Indeed, all these sets yield exact rules, that is, rules having 100% confidence. □

Many transaction data sets produce huge number of frequent item sets and, therefore, huge number of association rules particularly when the levels of support and confidence required are relatively low. Moreover, it is well known (see [26]) that limiting the analysis of association rules to the support/confidence framework can lead to dubious conclusions. The data mining literature contains many references that attempt to derive interestingness measures for association rules in order to focus data analysis of those rules that may be more relevant (see [19, 6, 7, 9, 16, 12]).

References

[1] Ramesh C. Agarwal, Charu C. Aggarwal, and V. V. V. Prasad. Depth first generation of long patterns. In R. Bayardo, R. Ramakrishnan, and S. J. Stolfo, editors, *Proceedings of the 6th Conference on Knowledge Discovery in Data, Boston, MA*, pages 108–118. ACM, New York, 2000.

[2] Ramesh C. Agarwal, Charu C. Aggarwal, and V. V. V. Prasad. A tree projection algorithm for generation of frequent item sets. *Journal of Parallel and Distributed Computing*, 61(3):350–371, 2001.

[3] J.-M. Adamo. *Data Mining for Association Rules and Sequential Patterns*. Springer-Verlag, New York, 2001.

[4] R. Agrawal, T. Imielinski, and A. N. Swami. Mining association rules between sets of items in large databases. In P. Buneman and S. Jajodia, editors, *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington, D.C.*, pages 207–216, 1993.

[5] R. Agrawal and J. Schaffer. Parallel mining of association rules. *IEEE Transactions on Knowledge and Data Engineering*, 8:962–969, 1996.

[6] C. C. Aggarwal and P. S. Yu. Mining associations with the collective strength approach. *IEEE Transactions on Knowledge and Data Engineering*, 13:863–873, 2001.

[7] R. Bayardo and R. Agrawal. Mining the most interesting rules. In S. Chaudhuri and D. Madigan, editors, *Proceedings of the 5th KDD, San Diego, CA*, pages 145–153. ACM, New York, 1999.

[8] S. E. Brossette and P. A. Hymel. Data mining and infection control. *Clinics in Laboratory Medicine*, 28:119–126, 2008.

[9] S. Brin, R. Motwani, and C. Silverstein. Beyond market baskets: Generalizing association rules to correlations. In J. Pekham, editor, *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 265–276, Tucson, AZ, 1997. ACM, New York.

[10] S. E. Brosette, A. P. Sprague, J. M. Hardin, K. B. Waites, w. T. Jones, and S. A. Moser. Association rules and data mining in hospital infection control and public health surveillance. *Journal of the American Medical Information Association*, 5:373–381, 1998.

[11] R. Harpaz, H. S. Chase, and C. Friedman. Mining multi-item drug adverse effect associations in spontaneous reporting systems. *BMC Bioinformatics*, 11, 2010.

[12] R. Hilderman and H. Hamilton. Knowledge discovery and interestingness measures: A survey. Technical Report CS 99-04, Department of Computer Science, University of Regina, 1999.

[13] J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. In Weidong Chen, Jeffrey F. Naughton, and Philip A. Bernstein, editors, *Proceedings of the ACM-SIGMOD International Conference on Management of Data, Dallas, TX*, pages 1–12. ACM, New York, 2000.

[14] L. Juntti-Patinen and P. J. Neuvonen. Drug-related death in a university central hospital. *European Journal of Clinical Pharmacology*, 58:479–482, 2002.

[15] S. Jaroszewicz and D. A. Simovici. A general measure of rule interestingness. In *Principles of Data Mining and Knowledge Discovery, LNAI 2168*, pages 253–266, Heidelberg, 2001. Springer-Verlag.

[16] S. Jaroszewicz and D. Simovici. Interestingness of frequent item sets using bayesian networks as background knowledge. In Won Kim, Ron Kohavi, Johannes Gehrke, and William DuMouchel, editors, *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, WA*, pages 178–186. ACM, New York, 2004.

[17] S. Jaroszewicz and D. A. Simovici. Interestingness of frequent itemsets using bayesian networks as background knowledge. In *Proceedings of KDD*, pages 178–186, 2004.

[18] M. Pirmohamed, A. M. Breckenridge, N. R. Kitteringham, and B. K. Park. Adverse drug reactions. *British Medical Journal*, 316:1295–1298, 1998.

[19] G. Piatetsky-Shapiro. Discovery, analysis and presentation of strong rules. In G. Piatetsky-Shapiro and W. Frawley, editors, *Knowledge Discovery in Databases*, pages 229–248. MIT Press, Cambridge, MA, 1991.

[20] P. Patel and P. J. Zed. Drug-related visits to the emergency department: how big is the problem? *Pharmacotherapy*, 22:915–923, 2002.

[21] R. Srikant H. Toivonan A. I. Verkamo R. Agrawal, H. Mannila. Fast discovery of association rules. In *Advances in Knowledge Discovery and Data Mining*. MIT Press.

[22] R. Rymon. Search through systematic set enumeration. In Bernhard Nebel, Charles Rich, and William R. Swartout, editors, *Proceedings of the 3rd International Conference on Principles of Knowledge Representation and Reasoning, Cambridge, MA*, pages 539–550. Morgan Kaufmann, San Mateo, CA, 1992.

[23] D. Simovici and C. Djeraba. *Mathematical Tools for Data Mining*. Springer-Verlag, London, 2008.

[24] M. Steinbach, G. Karypis, and V. Kumar. A comparison of document clustering techniques. In M. Grobelnik, D. Mladenic, and N. Milic-Freyling, editors, *KDD Workshop on Text Mining, Boston, MA*, 2000.

[25] P. Tan, V. Kumar, and J. Srivastava. Selecting the right interestingness measure for association patterns. In *KDD*, pages 32–41, 2002.

[26] P. N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Addison-Wesley, Reading, MA, 2005.

[27] S. Mukhopadhyay V. Narayanasamy, M. Palakal, and J. Mostafa. Transminer: Mining transitive associations among biological objects from medline. *Journal of Biomedical Science*, 11:864–873, 2004.

[28] A. Wright, E. S. Chen, and F. L. Maloney. An automated technique for identifying associations between medications, laboratory results and problems. *Journal of Biomedical Informatics*, 43:891–901, 2010.

[29] L. L. Weed. New connections between medical knowledge and patient care. *British Nedical Journal*, 315:231–235, 1997.

[30] A. M. Wilson, L. Thabane, and A. Holbrook. Application of data mining techniques in pharmacovigilance. *British Journal of Clinical Pharmacology*, 57:127–134, 2003.

[31] M. J. Zaki and C.J. Hsiao. Efficient algorithms for mining closed itemsets and their lattice structure. *IEEE Transactions on Knowledge and Data Engineering*, 17:462–478, 2005.