2013 IEEE International Conference on Systems, Man, and Cybernetics

A Bayesian Association Rule Mining Algorithm

David Tian*, Ann Gledson*, Athos Antoniades‡, Aristo Aristodimou‡, Ntalaperas Dimitrios§,
Ratnesh Sahay†, Jianxin Pan¶, Stavros Stivaros‖, Goran Nenadic*, Xiao-jun Zeng*, John Keane*
*School of Computer Science, University of Manchester, UK
tiand@cs.man.ac.uk; anngledson@gmail.com; g.nenadic@cs.man.ac.uk; x.zeng@cs.man.ac.uk; jak@cs.man.ac.uk
†DERI, National University of Ireland
ratnesh.sahay@deri.org
‡University of Cyprus
athos@cs.ucy.ac.cy; aristodimou.aristos@ucy.ac.cy
§UBITECH Ltd, Greece
dntalaperas@ubitech.eu
‖Manchester Medical School, Royal Manchester Children's Hospital, UK
stavros.stivaros@postgrad.manchester.ac.uk
¶School of Mathematics, University of Manchester, UK
Jianxin.Pan@manchester.ac.uk

***Abstract*—This paper proposes a Bayesian association rule mining algorithm (BAR) which combines the Apriori association rule mining algorithm with Bayesian networks. Two interestingness measures of association rules: Bayesian confidence (BC) and Bayesian lift (BL) which measure conditional dependence and independence relationships between items are defined based on the joint probabilities represented by the Bayesian networks of association rules. BAR outputs best rules according to BC and BL. BAR is evaluated for its performance using two anonymized clinical phenotype datasets from the UCI Repository: Thyroid disease and Diabetes. The results show that BAR is capable of finding the best rules which have the highest BC, BL and very high support, confidence and lift.**

***Index Terms*—Bayesian association rules, Bayesian networks, joint probability distribution, Bayesian confidence, Bayesian lift**

I. Introduction

Linked2Safety (L2S) [10] is an EU-funded research project which aims to build a next-generation, semantically-interlinked, secure medical and clinical information space. This information space should facilitate healthcare professionals and pharmaceutical experts etc. at pan-European level to dynamically discover, combine and easily access medical resources and information contained in spatially distributed Electronic Health Records (EHRs). At the same time, this information space must respect patient anonymity, data ownership, privacy and European and national legislation. A primiary application of the L2S project is adverse drug events (ADE) detection.

An adverse drug event (ADE) is any unfavorable and unintended sign (including, for example, an abnormal laboratory finding), symptom e.g. high blood pressure, or disease e.g. heart attack temporally associated with the use of a medicinal product, whether or not considered related to the medicinal product [5]. ADE detection is concerned with early detection of adverse events for patients who have been taking some drugs. The main objective of pharmacovigilance is early detection of novel ADEs with minimal patient exposure. The impact of ADE results in significant social costs estimated in several billion dollars annually, and inflicts unnecessary, sometimes fatal, harm to patients. Hence, their identification is paramount to health care. There are two types of associations: binary (bivariate) associations appear in pairs, including only one drug and one ADE, such as Vioxx relates to heart attack; multi-item associations are associations between two drugs and one or more ADEs, such as Aspirin and Warfarin cause Bleeding [5]. Association rule mining algorithms have been used to mine new ADE associations [5].

The United States Food and Drug Administrations (FDA) Adverse Event Reporting System (AERS) is a database which contains over 4M ADE reports, from 1969 to the present. Recently, the Apriori association rule mining algorithm has been extended to more efficiently mine multi-item ADE associations from the FDA AERS [5]. Constraints were added to Apriori and drugs/ADEs-based indexing techniques were implemented to reduce significantly the search space of possible multi-item ADE associations present in the FDA AERS and to reduce the number of reports to be examined for each rule. Based on a set of 162,744 reports of suspected ADEs reported to AERS and published in the year 2008, the extended Apriori algorithm identified 1167 multi-item ADE associations of which 33% are novel associations.

Multi-objective evolutionary algorithms [6], [7] have been proposed to find interesting rules which Apriori cannot normally find. These algorithms find Pareto optimal solutions which maximize multiple interestingness measures of rules in a single run, whereas Apriori finds rules that maximize one interestingness measure in a single run. Fuzzy association rules algorithms [8], [9] have been proposed to classify gene expression data and predict stock market indices. This work introduces a Bayesian association rule mining algorithm (BAR) which combines Apriori with Bayesian networks and applies BAR to two anonymized public-available clinical phenotype datasets Thyroid disease and diabetes because at the time of doing this work, the L2S ADE datasets were not available.

978-1-4799-0652-9/13 $31.00 © 2013 IEEE
DOI 10.1109/SMC.2013.555

The rest of the paper is structured as follows: Section II presents the basic concepts of association rules and Bayesian networks; Section III presents BAR; Section IV presents data anonymization of clinical data; Section V evaluates results of BAR applied to the example datasets; Section VI presents conclusions and future work.

II. Basic Concepts

A. Association Rules

An association rule is an implication expression of the form A⇒B, where A and B are disjoint itemsets. The *support* denoted $S(A \cup B)$ of an association rule A⇒B is the percentage of instances that contain all the items included in the association rule:

$$S(A \cup B) = \frac{k}{n} \tag{1}$$

where k is the number of instances containing all the items of A and B; n is the total number of instances of the dataset.

The *confidence* of an association rule is a fraction that shows how frequently B occurs among all the instances containing A:

$$\text{confidence} = \frac{S(A \cup B)}{S(A)}, \tag{2}$$

where S(A∪B) is support of the rule; S(A) is support of A. The confidence value indicates how reliable the rule is. Confidence provides an estimate of $P(B|A)$ the conditional probability of B given A assuming B statistically depends on A, and is used to measure the reliability or interestingness of the rule.

The *lift* value of an association rule is the ratio of the confidence of the rule to the support of B:

$$\text{lift} = \frac{\text{confidence}}{S(B)} \tag{3}$$

where S(A) is the support of A; S(B) is the support of B. Lift is a measure of the deviation of the rule from the statistical independency of A and B. The lift is a value between 0 and infinity [4]: a value greater than 1 indicates that A and B appear together more often than expected; a value less than 1 indicates that A and B appear together less often than expected; a value close to 1 indicates that A and B appear together almost as often as expected.

Apriori mines association rules as follows:

1) Generate itemsets and select those itemsets whose supports ≥ the minimum support threshold.
2) Generate rules from selected itemsets. To generate a rule $A \Rightarrow B$ from an itemset, a subset of the itemset forms B and the remaining items forms A.
3) Output the rules with the highest confidence and lift.

B. Bayesian Networks

Let X and Y be two disjoint subsets of random variables such that the probability of Y is P(Y)>0. Then, the conditional probability distribution (CPD) of X given Y is defined as:

$$P(X|Y) = \frac{P(X,Y)}{P(Y)}. \tag{4}$$

The joint probability over variables $X_1, ..., X_n$ is defined as:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|X_1, \ldots, X_{i-1}). \tag{5}$$

A set of variables $X_1, \ldots, X_n$ are independent to one another if and only if:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i). \tag{6}$$

Bayesian networks [11], [12], [13] are graphical representation of probabilistic relationships among a set of variables. Given a finite set $X = \{X_1, ..., X_n\}$ of variables, a Bayesian network G is an annotated directed acyclic graph (DAG) which represents a joint probability distribution over X. The nodes of the graph correspond to the variables $X_1, ..., X_n$. The links of the graph correspond to the direct influence from one variable to the other. If there is a directed link from variable X_i to variable X_j, X_i is a parent of X_j. Each node is annotated with a conditional probability distribution (CPD) that represents $P(X_i|P_a(X_i))$, where $P_a(X_i)$ denotes the parents of X_i in G. A Bayesian network G represents an unique joint probability distribution $P(X_1, \ldots, X_n)$ over X:

$$P(X_1, \ldots, X_n) = \prod_{i} P(X_i|Pa(X_i)). \tag{7}$$

Figure 1 [12] shows an example of a Bayesian Network and some of its conditional probabilities.

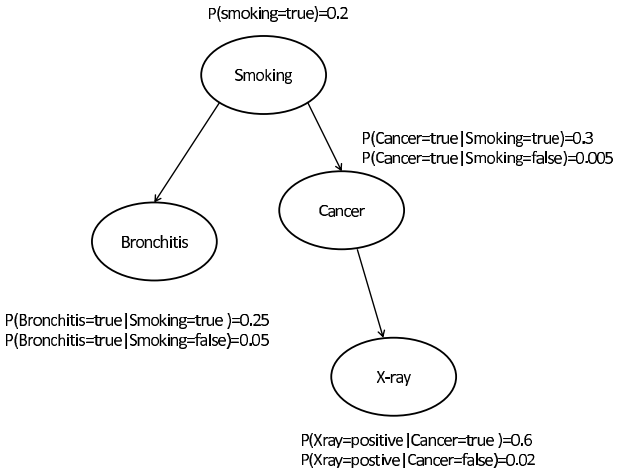


Fig. 1. A Bayesian network for detecting breast cancer. An arc represents a causal relationship between two variables. Conditional probabilities of each variable are attached.

III. Bayesian Association Rules Mining Algorithm

The BAR algorithm is presented in figure 2. Bayesian confidence and Bayesian lift are defined in Sections III-A and III-B.

Algorithm: Bayesian Association Rule Mining
Input: a dataset consisting of instances and attributes
Output: association rules
1. Discretize any continuous variables of the input dataset.
2. Generate association rules using Apriori.
3. For each rule, construct a Bayesian network.
4. Compute the Bayesian confidence (BC)
   and Bayesian lift (BL) of each rule.
5. Output those rules with the highest BC and BL.

Fig. 2. Pseudo-code of the Bayesian Association Rule Mining Algorithm

A. Bayesian Confidence

Let A and B be itemsets such that A consists of items $I_1, I_2, \ldots, I_m$ and B consists of items $I_{m+1}, I_{m+2} \ldots, I_n$. Then, an association rule $A \Rightarrow B$ can be represented as the Bayesian network in figure 3, where $x_1, \ldots, x_m$ and B are Boolean variables corresponding to $I_1, I_2, \ldots, I_m$ and $x_{m+1}, \ldots, x_n$ are Boolean variables corresponding to $I_{m+1}, I_{m+2} \ldots, I_n$. The joint probability distribution represented by the network is the following:

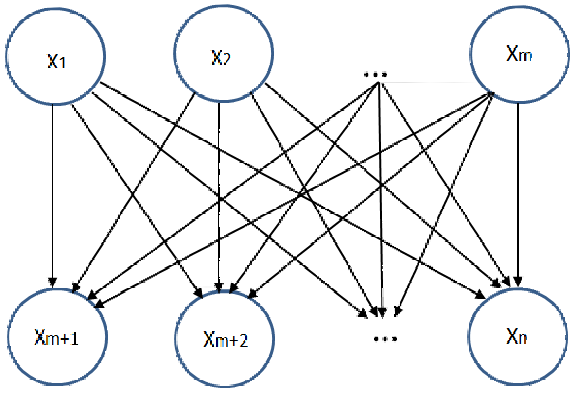


Fig. 3. The Bayesian network representing the rule A⇒ B.

$$
\begin{aligned}
P(x_1, \ldots, x_m, x_{m+1}, \ldots, x_n) &= \prod_{i=1}^{n} P(x_i | parents(x_i)) \\
&= \prod_{i=1}^{m} P(x_i) \prod_{j=m+1}^{n} P(x_j | x_1, x_2, \ldots, x_m) \\
&= \prod_{i=1}^{m} S(\{x_i\}) \prod_{j=m+1}^{n} \frac{P(\{x_j, x_1, x_2, \ldots, x_m\})}{P(\{x_1, x_2, \ldots, x_m\})} \\
&= \prod_{i=1}^{m} S(\{x_i\}) \prod_{j=m+1}^{n} \frac{S(\{x_j, x_1, x_2, \ldots, x_m\})}{S(\{x_1, x_2, \ldots, x_m\})} \qquad (8)
\end{aligned}
$$

where S is the support of an itemset.

Definition 1: Bayesian confidence (BC) of A⇒B is defined as $P(B|A)$ computed using the Bayesian network representing A⇒B:

$$
\begin{aligned}
BC(A \Rightarrow B) = P(B|A) &= \frac{P(A,B)}{P(A)} \\
&= \frac{P(x_1, \ldots, x_m, x_{m+1}, \ldots, x_n)}{P(\{x_1, x_2, \ldots, x_m\})} \\
&= \frac{P(x_1, \ldots, x_m, x_{m+1}, \ldots, x_n)}{S(\{x_1, x_2, \ldots, x_m\})} \qquad (9)
\end{aligned}
$$

Short rules generalize better than long ones because the shorter rules match more instances than long rules; additionally, short rules are easier to interpret by humans than long ones. To penalize long rules, length L of rule is incorporated:

$$
BC = \left(\frac{P(x_1, \ldots, x_m, x_{m+1}, \ldots, x_n)}{S(\{x_1, x_2, \ldots, x_m\})} \right)^L \qquad (10)
$$

For example given an association rule:

{male,binge_drinking,obese,smoking,age≥50}
⇒ {diabetes,heart_disease,hypertension},

its corresponding Boolean variables are Male (M), Smoking (S), Binge drinking (B), Obese (O), Age>50 (A), Diabetes (D), Heart Disease (HD) and Hypertension (H) and BC is:

$$
\left(\frac{P(M=t, S=t, B=t, O=t, A=t, D=t, HD=t, H=t)}{S(\{M=t, S=t, B=t, O=t, A=t\})} \right)^8
$$

B. Bayesian Lift

Definition 2: Given a rule A⇒ B, the Bayesian lift (BL) is defined as $\frac{BC}{P(B)}$ computed using the Bayesian network representing A⇒B:

$$
\begin{aligned}
BL &= \frac{BC}{P(B)} = \frac{P(B|A)}{P(B)} = \frac{P(A,B)}{P(A)P(B)} \\
&= \frac{P(X_1,\ldots,X_m,X_{m+1},\ldots,X_n)}{P(X_1,\ldots,X_m)P(X_{m+1},\ldots,X_n)} \qquad (11) \\
&= \frac{P(X_1,\ldots,X_m,X_{m+1},\ldots,X_n)}{\prod_{i=1}^{m} X_i \prod_{j=m+1}^{n} X_j} \qquad (12)
\end{aligned}
$$

BL is a value between 0 and infinity. BL=1 $\Leftrightarrow \frac{P(B|A)}{P(B)} = 1 \Leftrightarrow \frac{P(A,B)}{P(A)P(B)} = 1 \Leftrightarrow P(A,B) = P(A)P(B) \Leftrightarrow$ A and B are independent [5]. $BL > 1 \Leftrightarrow \frac{P(B|A)}{P(B)} > 1 \Leftrightarrow P(A,B) > P(A)P(B) \Leftrightarrow$ B positively depends on A i.e. A and B are positively correlated [5]. $BL < 1 \Leftrightarrow \frac{P(B|A)}{P(B)} < 1 \Leftrightarrow P(A,B) < P(A)P(B) \Leftrightarrow$ B negatively depends on A i.e. A and B are negatively correlated [5].

IV. Data Anonymization

BAR has been applied to the Thyroid disease and Diabetes datasets from the UCI Repository. These datasets have been anonymized using the L2S data cube approach [14] before input to BAR. In the L2S approach, a subject can be identified if it has unique properties. On the other hand, if a number of subjects have identical properties, they cannot be distinguished from each other; they are un-identifiable and anonymised subjects. Identification of subjects from datasets violates privacy laws. As subjects can be identified from the raw clinical data, only allowed users are granted access to the raw data while third parties can only access the anonymised data.

Fig. 5. The data matrix obtained by transforming the example anonymized data cube in Figure 4(b)

Sex,	BMI>10,	Dyslipidemia
Male,	Yes,	No
Male,	Yes,	No
Male,	Yes,	No
Male,	Yes,	No
Male,	Yes,	No
Male,	Yes,	No
Female,	Yes,	No
Female,	Yes,	No
Female,	Yes,	No
Female,	Yes,	No
Female,	Yes,	No
Female,	Yes,	No
Female,	Yes,	No
Female,	Yes,	No
Female,	No,	No
Female,	No,	No
Female,	No,	No
Female,	No,	No
Female,	No,	No

A. The L2S Data Annonymisation: Data Cubes Generation

The steps to anonymize data are as follows:

1) Data Discretization: The continuous variables are discretized. Discrete values are represented using integers starting from 0.
2) Data Cube Creation: a data cube is created with with columns corresponding to dimensions (attributes) and rows representing the combinations of values on the dimensions. An example data cube is shown in Figure 4(a).
3) Count Perturbation: The counts of values combinations are perturbed by adding noise (integers) in a predefined set of integers e.g. {-1,1}.
4) Cell Suppression (Data Anonymization): A count threshold is set by the user so that the cells with counts lower than the threshold are excluded from further analyses - this is done by replacing the counts with 0 (Figure 4(b)).

V. Evaluation

An anonymized data cube is transformed into a data matrix before input to BAR. The anonymized data cube in figure 4(b) is transformed into the data matrix shown in figure 5.

A. Thyroid Disease Dataset

1) Data Preparation: The thyroid dataset contains 30 attributes (Table I) and consists of a number of smaller datasets (Table II). Each smaller dataset consists of numerous thyroid disease categories such as 'increased binding protein', 'decreased binding protein' etc and a negative category. To mine rules from the whole Thyroid disease dataset, all the smaller datasets were merged into one dataset as follows. All thyroid disease categories of the smaller datasets were merged into a single positive category. In data mining, models obtained from balanced datasets output unbiased predictions and vice versa [1]. In order to obtain class association rules with unbiased classification, a balanced dataset was obtained as follows; the same number of negative instances as the positive instances were extracted from the smaller datasets and the positive and negative categories were merged together to give a balanced dataset consisting of 6792 instances [1].

TABLE I
Thyroid disease dataset

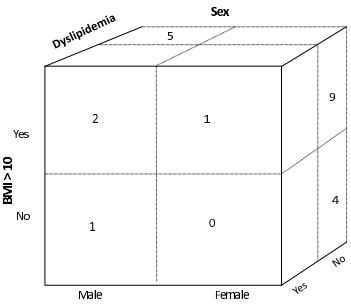
Attributes	Values
age	continuous
sex	M, F
on thyroxine	f, t
query on thyroxine	f, t
on antithyroid medication	f, t
sick	f, t
pregnant	f, t
thyroid surgery	f, t
I131 treatment	f, t
query hypothyroid	f, t
query hyperthyroid	f, t
lithium	f, t
goitre	f, t
tumor	f, t
hypopituitary	f, t
psych	f, t
TSH measured	f, t
TSH	continuous
T3 measured	f, t
T3	continuous
TT4 measured	f, t
TT4	continuous
T4U measured	f, t
T4U	continuous
FTI measured	f, t
FTI	continuous
TBG measured	f, t
TBG	continuous
referral source	WEST, STMW, SVHC, SVI, SVHD, other
class	positive, negative

TABLE II
Class distributions of the smaller Thyroid Disease datasets

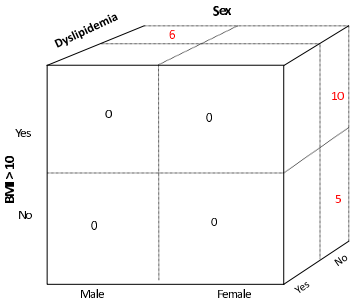
datasets	class distributions (positive : negative)
allbp.data	133 : 2666
allbp.test	30 : 941
allhypo.data	220 : 2579
allhypo.test	71 : 900
allrep.data	29 : 1038
allrep.test	37 : 934
dis.data	45 : 2754
dis.test	13 : 958
hypothyroid.data	150 : 3012
sick.data	171 : 2628
sick.test	60 : 911
sick-euthyroid.data	292 : 2870
thyroid0387	2401 : 6770

2) Data Cleaning: Inconsistent values of the attributes were detected and treated as missing values. Outliers of the attributes were detected using box-and-whisker plots (Figure 6) and were treated as missing values. TBG has the most missing values (99%), hence, it was removed. Then, missing values were replaced with the mean or mode of the attribute. After data cleaning, a data cube was created and association rules were mined from the data cube using BAR.

3) Two-items Associations: Table III illustrates the top 10 mined rules ranked by BC. Each rule consists of two items.



(a) Example data cube: Dimensions are attributes; values on each dimension are the values of a attribute; each cell contains the count of a combination of values of the dimensions.



(b) The anonymized example data cube obtained after count perturbation and cell suppression using 5 as threshold

Fig. 4. Data Cubes

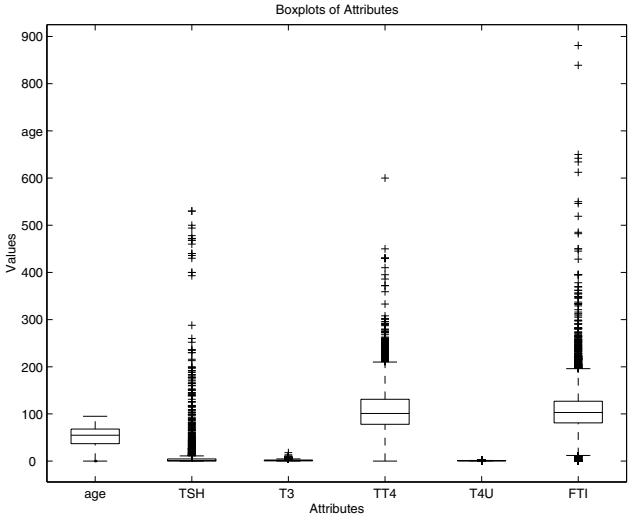


Fig. 6. Boxplots of attributes of Thyroid Disease: each box shows the distribution of values where the bottom line of the box is the 1st quartile, the middle line is the median and the top line of the box is the 3rd quartile; outliers are the points outside the whiskers extending from the box.

For each rule, the values of support, confidence, Bayesian confidence and lift are the maximum i.e. 1. BL of these rules is also the highest (1.116) amongst all mined rules. This indicates that rules with the highest confidence have the highest BC and BL. The BLs of the rules have the same value.

4) Multi-items Associations: Multi-items association rules are also output by BAR. Table IV illustrates the top 10 multi-items association rules ranked by BC. For each rule, the values of support, confidence, BC and lift are 1. This indicates that multi-items rules with the highest confidence also have the highest BC and BL. BLs of the rules are the same value.

5) Class Association Rules: Rough set feature selection [15], [16] was used to reduce the dimensionality of the input data cube from 29 dimensions (attributes) to 21.

TABLE III
THE TOP 10 RULES FROM THYROID DISEASE RANKED BY BAYESIAN CONFIDENCE AND BAYESIAN LIFT

rule	S	C	BC	L	BL
goitre=0 ⇒ lithium=0	1	1	1	1	1.116
lithium=0 ⇒ goitre=0	1	1	1	1	1.116
hypopituitary=0 ⇒ lithium=0	1	1	1	1	1.116
lithium=0 ⇒ hypopituitary=0	1	1	1	1	1.116
TSH=0 ⇒ lithium=0	1	1	1	1	1.116
lithium=0 ⇒ TSH=0	1	1	1	1	1.116
T3=0 ⇒ lithium=0	1	1	1	1	1.116
lithium=0 ⇒ T3=0	1	1	1	1	1.116
TT4=0 ⇒ lithium=0	1	1	1	1	1.116
lithium=0 ⇒ TT4=0	1	1	1	1	1.116

TABLE IV
THE TOP 10 MULTI-ITEMS RULES FROM THYROID DISEASE RANKED BY BAYESIAN CONFIDENCE AND BAYESIAN LIFT

rule	S	C	BC	L	BL
goitre=0 hypopituitary=0 ⇒ lithium=0	1	1	1	1	1.105
lithium=0 hypopituitary=0 ⇒ goitre=0	1	1	1	1	1.105
lithium=0 goitre=0 ⇒ hypopituitary=0	1	1	1	1	1.105
hypopituitary=0 ⇒ lithium=0 goitre=0	1	1	1	1	1.105
goitre=0 ⇒ lithium=0 hypopituitary=0	1	1	1	1	1.105
lithium=0 ⇒ goitre=0 hypopituitary=0	1	1	1	1	1.105
goitre=0 TSH=0 ⇒ lithium=0	1	1	1	1	1.105
lithium=0 TSH=0 ⇒ goitre=0	1	1	1	1	1.105
lithium=0 goitre=0 ⇒ TSH=0	1	1	1	1	1.105
TSH=0 ⇒ lithium=0 goitre=0	1	1	1	1	1.105

Then, the attributes were ranked using information gain and the top 10 ranked attributes were used to reduce the 21-dimensional data cube (Table V). Finally, class association rules are mined from the 10-dimensional data cube using BAR. The top 10 ranked class association rules are shown in Table VI - these have very high confidence, where the highest confidence amongst all the mined rules is 0.62 and corresponds to the following rule:

T4U=0 sex=1 query_hypothyroid=0 ⇒ class=0
0.167 (support) 0.62 (confidence) 1.141 (lift) 0.006 (BC) 0.730 (BL).

TABLE V
TOP 10 RANKED ATTRIBUTES OF THYROID DISEASE BY INFORMATION GAIN

Features	information Gain
T4U	0.0390063
referral_source	0.0232155
TBG_measured	0.0122331
pregnant	0.0105625
TSH_measured	0.0100687
psych	0.0056991
sex	0.0044499
TT4_measured	0.0028364
query_hypothyroid	0.0028238
T4U_measured	0.0025371

TABLE VI
THE TOP 10 CLASS ASSOCIATION RULES FROM THYROID DISEASE RANKED USING BAYESIAN CONFIDENCE AND BAYESIAN LIFT

rule	S	C	L	BC	BL
pregnant=0 sex=1 $\Rightarrow$ class=0	0.169	0.61	1.123	0.026	0.850
TBG_measured=0 sex=1 $\Rightarrow$ class=0	0.169	0.61	1.123	0.026	0.847
psych=0 sex=1 query_hypothyroid=0 $\Rightarrow$ class=0	0.153	0.6	1.104	0.014	0.813
TBG_measured=0 pregnant=0 sex=1 $\Rightarrow$ class=0	0.169	0.61	1.123	0.014	0.803
pregnant=0 sex=1 query_hypothyroid=0 $\Rightarrow$ class=0	0.167	0.61	1.123	0.014	0.805
TBG_measured=0 sex=1 query_hypothyroid=0 $\Rightarrow$ class=0	0.167	0.61	1.123	0.014	0.803
pregnant=0 sex=1 T4U_measured=1 $\Rightarrow$ class=0	0.160	0.61	1.123	0.013	0.802
TBG_measured=0 sex=1 T4U_measured=1 $\Rightarrow$ class=0	0.160	0.61	1.123	0.013	0.800
T4U=0 sex=1 $\Rightarrow$ class=0	0.169	0.61	1.123	0.013	0.771
pregnant=0 psych=0 sex=1 query_hypothyroid=0 $\Rightarrow$ class=0	0.153	0.6	1.104	0.001	0.794

B. Diabetes Dataset

The diabetes dataset contains 8 continuous features and a class variable (Table VII). The dataset has no inconsistent values. Outliers were detected using boxplots (Figure 7) and were treated as missing values. Missing values were replaced with the mean or mode of the attribute. Table VIII illustrates the top 10 association rules ranked using BC and BL - these have very high confidence, where the highest confidence amongst all the mined rules is 0.99.

TABLE VII
DIABETES DATASET

Attributes	Description
preg	Number of times pregnant
plas	Plasma glucose concentration
pres	Diastolic blood pressure (mm Hg)
skin	Triceps skin fold thickness (mm)
insu	2-Hour serum insulin
mass	Body mass index
pedi	Diabetes pedigree function
age	years
class	Whether a patient has sign of diabetes(yn)

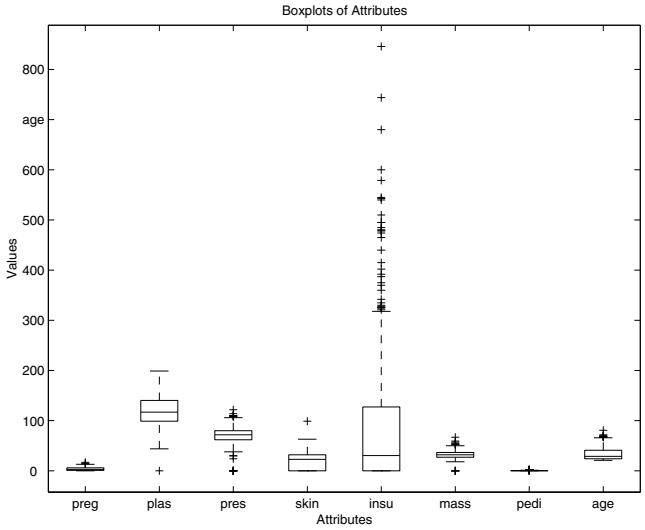


Fig. 7. Boxplots of attributes of Diabetes: each box shows the distribution of values where the bottom line of the box is the 1st quartile, the middle line is the median and the top line of the box is the 3rd quartile; outliers are the points outside the whiskers extending from the box.

TABLE VIII
THE TOP 10 RULES FROM DIABETES RANKED USING BAYESIAN CONFIDENCE AND BAYESIAN LIFT

rule	S	C	BC	L	BL
pedi=0 $\Rightarrow$ skin=0	0.965	0.99	0.703	1.002	1.087
skin=0 age=0 $\Rightarrow$ insu=0	0.878	0.98	0.703	1.007	1.082
skin=0 $\Rightarrow$ pedi=0	0.965	0.98	0.703	1.005	1.083
skin=0 insu=0 $\Rightarrow$ pedi=0	0.941	0.98	0.692	1.005	1.075
skin=0 pedi=0 age=0 $\Rightarrow$ insu=0	0.859	0.98	0.688	1.007	1.074
skin=0 age=0 $\Rightarrow$ pedi=0	0.876	0.98	0.683	1.005	1.079
insu=0 $\Rightarrow$ skin=0	0.962	0.99	0.679	1.002	1.083
skin=0 $\Rightarrow$ insu=0	0.962	0.97	0.679	0.996	1.079
pedi=0 age=0 $\Rightarrow$ skin=0	0.876	0.99	0.676	1.002	1.081
insu=0 pedi=0 $\Rightarrow$ skin=0	0.941	0.99	0.654	1.002	1.073

1) Class Association Rules: Table VIII illustrates the top 10 class association rules - these have very high confidence. The highest confidence amongst all the mined rules is 0.98.

VI. CONCLUSION

The best rules output by BAR have both the highest BC and BL. The best rules also have very high support, confidence and lift values. Hence, BC and BL appear to be promising measures of the quality of association rules. BC and BL are defined based on the joint probability distributions of the association rules. The joint probability distributions are represented by the Bayesian networks which model the possible conditional dependence and independence relationships between the items of the association rules. Therefore, a high joint probability value indicates that there are strong conditional dependence and independence relationships between the items of the association rule and vice versa. The current work shows that the statistical conditional dependence and independence relationships between the items of association rules can be used as quality measures for association rules. Unknown

TABLE IX
THE TOP 10 CLASS ASSOCIATION RULES FROM DIABETES RANKED USING BAYESIAN CONFIDENCE AND BAYESIAN LIFT

rule	S	C	L	BC	BL
preg=0 plas=0 pres=1 $\Rightarrow$ class=0	0.145	0.95	1.45	0.588	1.348
preg=0 plas=0 pres=1 pedi=0 $\Rightarrow$ class=0	0.142	0.95	1.45	0.528	1.291
preg=0 plas=0 pres=1 skin=0 $\Rightarrow$ class=0	0.145	0.95	1.449	0.495	1.280
preg=0 plas=0 pres=1 skin=0 pedi=0 $\Rightarrow$ class=0	0.142	0.95	1.449	0.438	1.236
preg=0 plas=0 pres=1 insu=0 $\Rightarrow$ class=0	0.145	0.95	1.449	0.432	1.261
preg=0 plas=0 pres=1 insu=0 pedi=0 $\Rightarrow$ class=0	0.142	0.95	1.449	0.376	1.218
preg=0 plas=0 pres=1 skin=0 insu=0 $\Rightarrow$ class=0	0.145	0.95	1.449	0.350	1.207
preg=0 plas=0 $\Rightarrow$ class=0	0.212	0.96	1.465	0.255	1.200
preg=0 plas=0 pres=1 age=0 $\Rightarrow$ class=0	0.145	0.95	1.449	0.233	1.181
preg=0 plas=0 pres=1 mass=0 $\Rightarrow$ class=0	0.101	0.98	1.495	0.223	1.225

associations between drugs and adverse events can be mined so that necessary actions could be taken for those patients who took the drugs.

ACKNOWLEDGMENT

This work has been funded by the Linked2Safety project.

REFERENCES

[1] J. Han and M. Kamber, Data Mining: Concepts and Techniques, 3rd edition, Morgan Kaufmann, 2011

[2] S. Hettich, L. C. Blake and J. C. Merz, UCI Repository of machine learning databases, http://www.ics.uci.edu/∼mlearn/MLRepository.html, University of California, Irvine, Dept. of Information and Computer Sciences, 1998

[3] I. Guyon and A. Elisseeff, An Introduction to Variable and Feature Selection. Journal of Machine Learning Research, vol. 3, pp. 1157-1182, 2003

[4] DB2 Business Intelligence Online Manual at http://publib.boulder.ibm.com/infocenter/db2luw/v8/index.jsp?topic=/com.ibm.im.model.doc/c_lift_in_an_association_rule.html

[5] R. Harpaz, H. S. Chase, and C. Friedman, Mining multi-item drug adverse effect associations in spontaneous reporting systems, BMC Bioinformatics, 11:S7, 2010

[6] B. Minaei-Bidgoli, R. Barmaki and M. Nasiri, Mining Numerical Association Rules via Multi-objective Genetic Algorithms, Information Science, vol. 233, pp. 15-24, 2013

[7] D. Martin, A. Rosete, J. Alcala-Fdez and F. Herrera, A multi-objective evolutionary algorithm for mining quantitative association rules, 11th International Conference on Intelligent Systems Design and Applications (ISDA), pp. 1397-1402, 2011

[8] K. Kianmehr, M. Kaya, A. M. ElSheikh, J. Jida and R. Alhajj, Fuzzy association rule mining framework and its application to effective fuzzy associative classification, Data Mining and Knowledge Discovery, vol.1, pp. 477-495, 2011

[9] G.T.S. Ho, W.H. Ip, C.H. Wu and Y.K. Tse, Using a fuzzy association rule mining approach to identify the financial data association, Expert Systems with Applications, vol. 39, pp. 9054-9063, 2012

[10] Linked2Safety: A Next-Generation, Secure Linked Data Medical Information Space For Semantically-Interconnecting Electronic Health Records and Clinical Trials Systems Advancing Patients Safety In Clinical Research, http://www.linked2safety-project.eu/

[11] S. Russell and P. Norvig, Artificial Intelligence: A Modern Approach, Prentice Hall, 3rd edition, 2009

[12] J. Gadewadikar, O. Kuljaca, K. Agyepong, E. Sarigul, Y. Zheng and P. Zhang, Exploring Bayesian networks for medical decision support in breast cancer detection, African Journal of Mathematics and Computer Science Research, vol. 3, pp. 225-231, 2010

[13] C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2007

[14] A. Antoniades, J. Keane, A. Aristodimou, C. Philipou, A. Constantinou, C.Georgousopoulos, F. Tozzi, K. Kyriacou, A. Hadjisavvas, M. Loizidou, C.Demetriou, and C. Pattichis, The effects of applying cell-suppression and perturbation to aggregated genetic data, 12th IEEE International Conference on Bioinformatics Bioengineering (BIBE), pp. 644-649, 2012.

[15] W. R. Swiniarski and A. Skowron, Rough set methods in feature selection and recognition, Pattern Recognition Letters, vol. 24, pp. 833-849, 2003

[16] Q. Shen and A. Chouchoulas, A rough-fuzzy approach for generating classification rules, Pattern Recognition, 35(11), pp. 341-354, 2002