Trends in Bioinformatics

ISSN 1994-7941

science alert

ANSInet
an open access publisher
http://ansinet.com

Trends in Bioinformatics 4 (1): 35-46, 2011
ISSN 1994-7941 / DOI: 10.3923/tb.2011.35.46

Extracting Association Rules from Hiv Infected Patients' Treatment Dataset

K. Rameshkumar
Department of Computer Applications, Karunya University, Coimbatore, India

ABSTRACT

In recent days, the data mining techniques are fascinatingly applied in healthcare domain. It proved that these techniques are suitable to extract knowledge from medical domain. Association Rule Mining (ARM) is mined valuable information from large voluminous databases. But most of ARM algorithms are mined many uninteresting or unrelated knowledge. We proposed new n-cross validation based Apriori (nVApriori) algorithm to mine domain irrelevant rules. Acquired Immuno Deficiency Syndrome (AIDS) is a challenge infection in the field of medical domain. This work proposes a new dataset for AIDS/HIV infected patients' case history. The data were collected from Midwest clinics, London. The nVApriori algorithm applies with the proposed dataset. It mines many interesting rules, provides much useful information to domain experts. The proposed algorithm is performed better than traditional Apriori, most interesting rule mining algorithm, Non redundant rule mining algorithm.

Key words: Association rule, AIDS dataset, medical dataset, rule validation, data mining

INTRODUCTION

Data mining is to discover unseen knowledge from big amount of data. Discovered knowledge has been applied in to real database like medicine, astronomy, stock market and other areas (Al-Shalabi, 2011). Association analysis is an interesting and well established research area in data mining. It can discover hidden knowledge from large amount of databases. Studies on mining association rules have evolved from techniques for discovery of functional dependencies, strong rules, classification rules, causal rules, clustering to disk based, efficient methods for mining association rules in large sets of transaction data (Thakur *et al*., 2007).

Till now this field requires further studies to handle real life datasets. It needs to manage the following issues: reduce rules redundancy and unwanted rules, handling different data type such as numeric, categorical, XML, multimedia and composite data; capacity to manage the data dimension such as multidimensional, high dimensional data and multiple sequences; competence with data relation such as multirelational data and linkage record; also manage data quality such as missing data, noise, uncertainty and incompleteness and sensitivity such as mixing with sensitive information.

In recent days, the association analysis increasingly admits with medical datasets. This work is applied association rule mining algorithm with proposed HIV infected patients' treatment dataset. This work used nVApriori algorithm for mining interesting rules which is developed by Ramaraj and Kumar (2009). This nVApriori algorithm is used to mine and reduce unwanted or irrelevant rules. The algorithm used n-cross validation technique to find the unwanted rules. This work also proposes a new dataset for HIV patients' treatment data. The dataset is called HIV/AIDS patients' Treatment (HAT) Dataset.

Association rule mining: ARM finds interesting association or correlation relationship among a large set of data items with massive amounts of data continuously being collected and stored, many industries are becoming interested in mining association rules from their databases. Let D be a set of n transactions such that $D = \{T_1, T_2, T_3,..,T_n\}$, Where $T_i = I$ and I is a set of items, $I = (i_1, i_2, i_3, .. ,i_m\}$. A subset of I containing k items is called a k-itemset. Let X and Y be two itemsets such that $X \subset I$, $Y \subset I$ and $X \cap Y = \phi$. An association rule is an implication denoted by X=>Y where X is called antecedent and Y is called the consequent. This section proceeds to define association rule metrics. Given an itemset X, support supp (X) is defined as the fraction of transactions $T_i \in D$ such that $X \subseteq T_i$. Consider P(X) the probability of appearance of X in D and P(Y | X) the conditional probability of appearance of Y given X. P(X) can estimated as P(X) = supp(X). The support of a rule X=>Y is defined as supp(X=>Y) = supp(XUY). An association rule X=>Y has a measure of reliability called the confidence, defined as conf (X=>Y) = supp(X=>Y)/supp(X). Confidence can be used to estimated P(Y | X): P(Y | X) = P(XUY)/P(X) = conf(X=>Y). This algorithm uses third metric called lift, defined as lift (X=>Y) = P(XUY)/(P(X) P(Y)) = conf(X=>Y)/ supp(Y). Lift quantifies the relationship between X and Y.

ARM in medical field: QuantMiner is a new tool which is developed by Salleb *et al.* (2004). It is genetic-based algorithm software for mining quantitative association rules on atherosclerosis datasets. The authors mined quantitative rules from the atherosclerosis dataset. These rules could handle both categorical and numerical attributes. QuantMiner is an interesting tool for mining descriptive rules from medical and other datasets.

The association rule mining is applied into high dimensional medical domains by Ordonez *et al.* (2006, 2000). The authors applied a greedy algorithm to compute rule covers in order to summarize rules having the same consequent. The significance of association rules is evaluated using support, confidence and lift. They are used association rules on a real dataset to predict absence or existence of heart disease. Rule covers summarized a large number of rules by producing a brief set of rules with high-quality metrics. The constraints are reduced the number of discovered rules and improved running time.

The differential association rule mining techniques is developed by Besemann *et al.* (2004) to identify rules that directly show the differences in annotations across interactions and between different types of interactions. The goal of this technique is to highlight differences between items belonging to different interacting nodes or different networks. Those differences can not be identified by the application of standard relational association rule mining techniques. The developed technique followed the association rule mining spirit by gaining its efficiency from a pruning step that is included even before the frequent item set generation step. They applied their framework to real examples of annotations and interactions. The results confirmed the expected biological knowledge and as identified as yet unknown associations.

The association rules are extracted from SAGE dataset by Gasmi *et al.* (2005). They stress the extraction of the generic basis of association rules from the sage data generated in different biological situations. The generic basis of association rules is a subset of all association rules, from which the remaining association rules are generated. They guaranties extra value knowledge usefulness and reliability. This is useful while handling highly dense sage data. They compare and assess frequent closed itemset algorithm performances on sage data. They also exploit the IGB generic basis of association rules. IGB are informative and more compact than other generic bases.

The association rules are merged with genetic algorithm for medical databases developed by Kwasnicka and Switalski (2006). They also studied possibilities of association rules generation from medical databases by means of genetic algorithms and compared the usefulness of the developed methods with the classical approach. They developed new a genetic method, called Extended Genetic Association Rules. They used real medical data from the Wroclaw Clinic. Their new genetic method uses the FP-growth approach to generate rules from real datasets with discrete values. The new method however did not guarantee finding all such rules.

The association rule mining is applied Protein-Protein Interaction (PPI) database. (Besemann *et al*., 2004). They mined the associations of functional regions of two interacting proteins to helpful in PPI prediction. The data are collected from Database of Interaction Proteins (DIP) and Interaction Proteins and downloaded from functional regions of proteins from Uniprot. A web-based system was designed to integrate the process and mine data to create some rules based on functional region association. PPIs of other species were used to evaluate these rules. In result, over 80% of the association rules produced from yeast PPI data worked in other species too. This indicates that the rules learnt from known PPI provide good bases for PPI prediction.

The fuzzy based association rules are generated from medical domains by Lopez *et al*. (2007). They constructed a fuzzy methodology for the integration and analysis of heterogeneous biological data. The main aspect of this fuzzy methodology is a fuzzy association rule mining algorithm based on the Top-Down (TD)-FP-Growth method. The results showed interesting associations between the structural and functional features of the yeast genome. It proved fuzzy association rules to be an intuitive tool to describe biological relations by using linguistic labels and a few easy understandable parameters like support, confidence and certainty factor.

The risk patterns are mined from medical domain by Li *et al*. (2005). They discussed the problem of finding risk patterns in medical data. They defined risk patterns by a statistical metric, relative risk which has been widely used in epidemiological research. They studied an anti-monotone property for mining optimal risk pattern sets and presented an algorithm to make use of the property in risk pattern discovery. They applied it to a real world dataset to find patterns associated with an allergic event that involved Angiotensin-Converting Enzyme (ACE) inhibitors. The algorithm has generated some useful results in medical research.

The association rule mining based duplicate detection method is developed by Li *et al*. (2005). They presented a novel method for duplicate detection using association rule mining. This method is used in biological duplicate detection. It explored scoring functions and criteria for matching sequence records. This method was evaluated using rules defined manually by domain experts. The authors focused on duplicate detection in a representative biological dataset using the Apriori method for rule mining.

The temporal rule mining is applied with medical datasets by Nehemiah *et al*. (2007). The authors developed a temporal rule mining and prediction system that aids the physician in decision-making. The system was tuned to mine rules from a time series hepatitis and thrombosis data sets. The temporal rules are generated, validated and stored in the knowledge base.

The rule mining is applied chronic hepatitis dataset by Ohsaki *et al*. (2003). It discovered medically interesting rules from the chronic hepatitis B and C diagnosis dataset from the Chiba University Hospital. The dataset contained time series data. The new framework supported for

preprocessing and rule discovering in time series data. The rules are different from the common ones in medical science.

Quantitative association rule mining is applied medical domain by Gupta *et al.* (2006), Gupta and Agrawal (2009). They applied to decipher the nature of associations between different amino acids that are present in a protein. The association rules have enhanced the understanding of protein composition and hold the potential to give clues regarding the interactions among particular sets of amino acids occurring in proteins. It has discovered rules based not only on the presence of amino acids but also on their absence. This is the first systematic study to discover global associations among amino acids.

Microsoft Research (MSR) (Microsoft Research, 2005) applied machine learning, data mining and other software techniques to comb through millions of strains of Human Immuno deficiency Virus (HIV) to find the genetic patterns necessary to train a patient's immune system to fight the virus. MSR technologies weren't initially conceived as medical research tools which may prove to be critical to the ongoing battle to slow down or halt HIV and other deadly viruses. The MSR-aided vaccine designs are undergoing laboratory testing at the University of Washington. The tests are conducted on samples of immune cells taken from HIV-infected patients to discover how effectively the models uncover the appropriate genetic patterns. MSR has generated many interesting patterns which can assist in developing new vaccines and assay kits.

Some statistical methods for data mining are applied in Acquired Immuno Deficiency Syndrome (AIDS) research, using the information about categories of transmission between patient's state-residence and municipality (Von-Borries *et al.*, 1996). The Brazilian National Program on Sexually Transmitted Diseases (STD)/Acquired Immuno Deficiency Syndrome (AIDS) maintains a database of notified AIDS cases that is reviewed every three months. They have analyzed and reported cases of AIDS in Brazil, from 1980 to 1995. The authors used correspondence and cluster-analysis to identify the main categories of transmission and its relation with the State concerned and the 20 highest-incidence municipalities in Brazil. The authors applied Weighted Principal Component Analysis (WPCA) to contingency table. It identified a low dimensional graphical representation of association between rows and columns of the table, while the cluster procedure found hierarchical clusters of observations in a dataset using Euclidean distances. Using statistical techniques for data mining they found very precise and easy-to-read ways of investigating new tendencies and finding new relations among data.

The data mining technique also used to investigate in HIV/AIDS-patient-data (Vararuk *et al.*, 2008). These patterns can be used for better management of the disease and more appropriate targeting of resources. A total of 250,000 anonymised records from HIV/AIDS patients in Thailand were imported into a database. The authors used clustering and association rule discovery methods through the IBM's intelligent miner, clustering highlighted groups of patients with common characteristics. Association rules identified associations that were not expected in the data and were different from those based on traditional reporting mechanisms utilized by medical practitioners. Symptoms that co-exist were identified and others serving as precursors observed. The identification of symptoms that are precursors of other symptoms can allow the targeting of the former so that the later symptoms can be avoided. A pragmatic and targeted approach to the management of resources available for HIV/AIDS treatment can also be arrived at this will provide a much better service, while at the same time reducing the expenses on service.

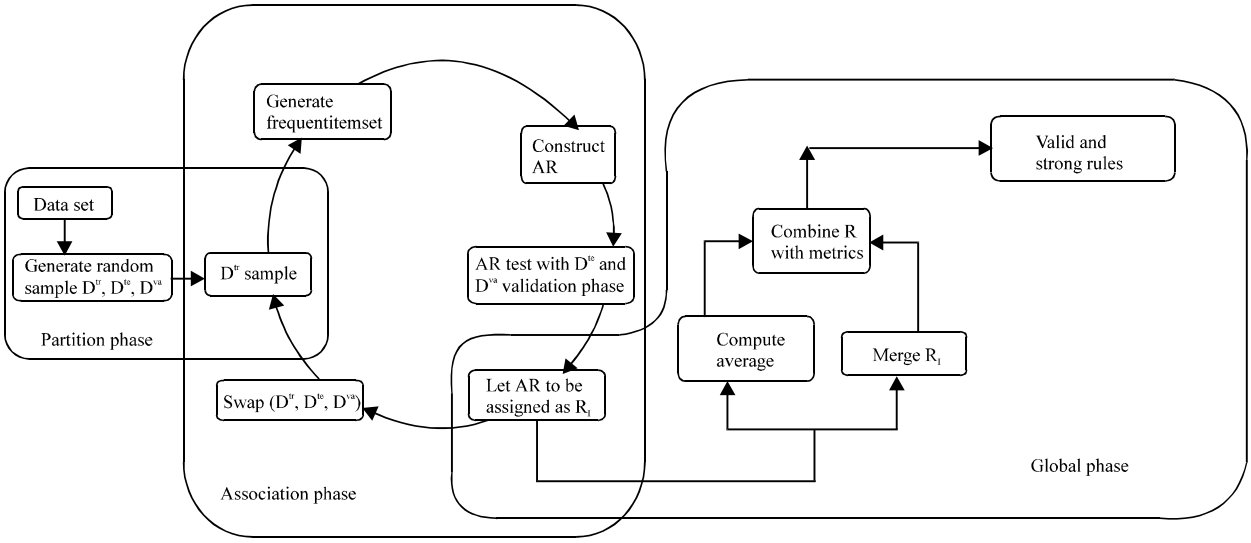

Fig. 1: Diagrammatic representation of proposed n-cross validation technique

The data mining technique is used to examine the association between health workforce, particularly nursing workforce and their achievement, taking into account other factors known to influence health status, such as socio-economic indicators (Vararuk *et al*., 2008). A merged dataset that includes country-level Millennium Development Goals (MDG) outcomes, workforce statistics and general socio-economic indicators were utilized for their study. Data were obtained from the world Health Organization (WHO) and United Nations AIDS (UNAIDS) Database. The main factors in understanding HIV/AIDS prevalence rates are physician-density followed by female literacy rates and nursing density in the country. Using general linear model approaches, increased physician and nurse density was associated with lower adult HIV/AIDS prevalence rate, even when providing for other socio-economic indicators. Increased nurse and physician-density is associated with improved health outcome, suggesting that countries aiming to attain the MDGs related to HIV/AIDS would do well to invest in their health workforce.

nVApriori algorithm: The data mining techniques usually generate a large amount of patterns and rules. However most of these patterns are not interesting from a user's point of view (Sharma *et al*., 2007a, b). Because Ramaraj and Kumar (2009) developed a new machine learning based method for irrelevant rule elimination technique. That is called n-cross validation technique. They merged the above technique with association rule mining algorithms which is called nVApriori algorithm. The nVApriori algorithm is divided into three phases as follows: Partition phase, Association phase and Global phase. The partition phase creates dataset partitions. The association phase mines association rules from dataset partitions and validates mined rules. The association rule mining phase is divided into two phases, the first is to find frequent itemset; the second is to use the frequent itemset to generate association rules (Gupta and Agrawal, 2009). The second phases is straight forward but first one is very computation expensive (Thakur *et al*., 2007). Finally, the global phase merges mined rules and computes average metric of merged rules. The diagrammatic representation of the proposed technique is given in the Fig. 1.

The pseudo code of the proposed algorithm is given below:

Algorithm 1: Pseudo code for proposed nVApriori algorithm

Input: Dataset (D),
Minimum Support (α),
Minimum Confidence (β),
Minimum Lift (γ),
Number of times to validate (n),
Train sample fraction (χ),
Test sample fraction (ψ)

Output: R rules

Step 1: For I = 1 to n do

//---------- Partition phase----------

Step 2: Create the partition D^{tr} , D^{te} and D^{va} based on χ and ψ

//---------- Association phase----------

Step 3: Generate 1- itemset

Search frequent k-itemset X on D^{tr} for k $\in$ {1...k}

Compute train_support Supp (X, D^{tr}) using $\text{train\_Supp}(x \Rightarrow y) = \frac{|D^{tr}_{xy}|}{|D^{tr}|}$

Step 4: Generate rules from the generated frequent itemset X.

For each rule $x \Rightarrow y \in D^{tr}$

Compute train_confidence $\text{Conf}(x \Rightarrow y)$ on D^{tr} using

$$\text{train\_conf}(x \Rightarrow y) = \frac{|D^{tr}_{xy}|}{|D^{tr}_{x}|}$$

Compute train_lift $\text{Lift}(x \Rightarrow y)$ on D^{tr} using $\text{train\_lift}(x \Rightarrow y) = \frac{|D^{tr}| \times |D^{tr}_{xy}|}{|D^{tr}_{x}| \times |D^{tr}_{y}|}$

Step 5: Let the rules set be R^{tr}.

Eliminate rules from R^{tr}

Such that

$\text{Supp}(x \Rightarrow y) < \alpha$ or $\text{Conf}(x \Rightarrow y) < \beta$ or $\text{Lift}(x \Rightarrow y) < \gamma$

// validate the rules using test set //

Step 6: Validate rules R^{tr} on D^{te}.

Let set $R^{te} = R^{tr}$

For each frequent itemset X means $X = (x \cup y \in R^{te})$

Compute test_support $\text{Supp}(X, D^{te})$ using $\text{test\_Supp}(x \Rightarrow y) = \frac{|D^{te}_{xy}|}{|D^{te}|}$

For each rule $x \Rightarrow y \in R^{te}$

Compute test_confidence $\text{Conf}(x \Rightarrow y)$ on D^{te} using

$$\text{test\_conf}(x \Rightarrow y) = \frac{|D^{te}_{xy}|}{|D^{te}_{x}|}$$

Compute test_lift $\text{Lift}(x \Rightarrow y)$ on D^{te} using $\text{test\_lift}(x \Rightarrow y) = \frac{|D^{te}| \times |D^{te}_{xy}|}{|D^{te}_{x}| \times |D^{te}_{y}|}$

Eliminate rules from R^{te}

Such that

$\text{Supp}(x \Rightarrow y) < \alpha$ or $\text{Conf}(x \Rightarrow y) < \beta$ or $\text{Lift}(x \Rightarrow y) < \gamma$

// validate the rules using validate set //

Step 7: Validate rules R^{te} on D^{va}

Let set $R^{va} = R^{te}$

For each frequent itemset X means $X = (x \cup y \in R^{va})$

Algorithm 1: Continued

Compute validate_support $\text{Supp}(X, D^{va})$ using

$$\text{validate\_Supp}(x \Rightarrow y) = \frac{|D_{xy}^{va}|}{|D^{va}|}$$

For each rule $x \Rightarrow y \in R^{va}$

Compute validate_confidence $\text{Conf}(x \Rightarrow y)$ on D^{va} using $\text{validate\_conf}(x \Rightarrow y) = \frac{|D_{xy}^{va}|}{|D_x^{va}|}$

Compute validate_lift $\text{Lift}(x \Rightarrow y)$ on D^{va} using $\text{validate\_lift}(x \Rightarrow y) = \frac{|D^{va}| \times |D_{xy}^{va}|}{|D_x^{va}| \times |D_y^{va}|}$

Eliminate rules from R^{va}

Such that

$\text{Supp}(x \Rightarrow y) < \alpha$ or $\text{Conf}(x \Rightarrow y) < \beta$ or $\text{Lift}(x \Rightarrow y) < \gamma$

Finally

Let the rules set be $R_I = R^{va}$

Next I

//---------- Global phase----------

Step 8: Get intersection of n rule sets and compute the average rule metrics with

$$\text{Supp}(x \Rightarrow y) = \frac{1}{n}\sum_{i=1}^{n} \text{Supp}[(x \Rightarrow y), D_i]$$

$$\text{Conf}(x \Rightarrow y) = \frac{1}{n}\sum_{i=1}^{n} \text{Conf}[(x \Rightarrow y), D_i] \text{ and}$$

$$\text{Lift}(x \Rightarrow y) = \frac{1}{n}\sum_{i=1}^{n} \text{Lift}[(x \Rightarrow y), D_i]$$

$R = R_1 \text{ n } R_2 \text{ n } R_3 \text{ n } \ldots \text{ n } R_n$

Proposed HAT dataset: The patients' treatment data are collected from Midwest hospitals around London. This work is having 50,000 records selected from over 6,000 HIV patients. The HIV /AIDS Treatment (HAT) dataset describes treatment of HIV patients' laboratory results and drugs. The above data are stored in SQL Server 2005 database. This work develops Java procedure to convert database into flat file. It consist 40 attributes and 108 instances that are related to drug list at each treatment, number of days between one to another clinical trial and laboratory results like CD4, RNA level of each treatment. The missing values are applicable in the HAT dataset. The Table 1 shows the sample collected data.

The converted data show below:

- 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40
- 65556, 2, G, JLT, 2, 9, JLT, 1, 1, A, JLT, 3, 1, C, JLT, 1, 2, A, DL, 1, 2, A, DL, 3, 3, C, JLT, 3, 3, C, DL, 3, 2, C, JLT, 1, 4, C, JLT, 2

Experimental study and analysis: This section details the results of nVApriori algorithm with HAT dataset. These experimental results are discussed in the following three ways. First is to analysis performance of nVApriori, number of rules mined by nVApriori and comparative study between nVApriori, MIR, NRRM and Apriori. The HAT dataset contains 108 transaction and 40 attributes. There are 14% percentage of missing values in their transactions, are filled by "?".The number of frequent itemsets mined at different k-itemsets from HAT dataset using nVApriori algorithm summaries in Table 2. This experiment set n = 10, support threshold is 10%, confidence threshold is 25% and lift threshold is 10%.

Table 1: Sample collected data

Patient Id	CD4 count	RNA level	Treatments	Clinical trial date
65556	160	22000	Lamivudine, Nelfinavir, Zidovudine	3/26/1997
65556	194	400	Lamivudine, Nelfinavir, Zidovudine	4/30/1997
65556	96	1000	Nelfinavir, Zidovudine, Lamivudine	2/4/1998
65556	92	2000	Nelfinavir, Lamivudine, Zidovudine	5/6/1998
65556	126	1000	Combivir, Nelfinavir	8/5/1998
65556	178	700	Combivir, Nelfinavir	3/10/1999
65556	261	3000	Lamivudine, Zidovudine, Nelfinavir	12/22/1999
65556	252	2000	Nelfinavir, Combivir	8/2/2000
65556	192	2000	Lamivudine, Nelfinavir, Zidovudine	10/18/2000
65556	322	2000	Lamivudine, Nelfinavir, Zidovudine	3/9/2001

Table 2: Number of frequent itemset mined in train, test and validate sets

		Number of frequent itemset		
k-itemsets (No)	Support (%)	Train set	Test set	Validate set
5	10	759	701	673
10	10	420	400	362
15	10	310	290	255
20	10	200	185	160
21	10	195	167	143
22	10	179	151	139
23	10	167	147	125
24	10	160	142	116
25	10	155	130	109
26	10	150	121	99
27	10	143	118	84
28	10	130	111	80
29	10	115	90	72
30	10	97	82	61

Table 3: Number of rule mined in various k-itemsets

K-itemset	5	10	15	20	21	22	23	24	25	26	27	28	29	30
nVApriori	580	325	209	141	130	125	109	98	96	87	70	68	63	53

This result shows the importance of filtering frequent itemset on the test and validate by varying k. The Table 1 summary results. The reduction in the number of frequent itemset is small, with a reduction about 8-14%. For k = 20, the reduction is about 8-20%. At high k = 25 to 30, the reduction is about 16-37%. The train set is reduced the frequent about 8 to 16%. Totally the validate set is reduced the frequent itemset about 11 to 37%. In low k level, the frequent itemset reduction is low and k is high, the frequent itemset reduction is high.

Number of rules mined: The results of number of rule mined by nVApriori in vary k using HAT dataset is shown in Table 3. Here this experiment set n = 10, support is 10%, confidence is 25% and lift is 10%. This result is summarized after 10 times of verify with test and validate set. The train_set rules are reduced by test and validate sets. It has reduced about 10-35%. In low k value, the nVApriori produces large number of rules. At same time, the k is high; the numbers of rules are small.

The proposed nVApriori algorithm is mined many interesting rules from HAT dataset. Most of the rules are medically dependent to the dataset. Some of the interesting rules are given below:

R1: days1 = 2 days3 = 5 days5 = 4 days7 = 4
[Supp = 23.5%, Conf = 52%, Lift = 50%]

R2: days2 = 4 days3 = 3 days6 = 5 days7 = 4 days8 = 3
[Supp = 35%, Conf = 50%, Lift = 50%]

R3: days1 = 2 days2 = 5 days3 = 3 days4 = 3 days6 = 3 days7 = 4
[Supp = 25%, Conf = 50%, Lift = 50%]

R4: days4 = 3 days5 = 6 days6 = 3 days7 = 5
[Supp = 35%, Conf = 50%, Lift = 50%]

R5: days7 = 2 days8 = 5 days9 = 2
[Supp = 25%, Conf = 50%, Lift = 40%]

R6: drugs1 = JLT drugs2 = JLT drugs5 = JLT drugs6 = JLT drugs8 = JLT
[Supp = 30%, Conf = 55%, Lift = 60%]

R7: drugs1 = IJP drugs2 = IJP drugs4 = JLT drugs5 = JLT drugs6 = JLT drugs8 = JLT
[Supp = 35%, Conf = 50%, Lift = 50%]

R8: drugs1 = JT drugs2 = JLT drugs6 = IJMT drugs8 = IJMT
[Supp = 35%, Conf = 50%, Lift = 60%]

R9: drugs4 = JP drugs6 = JMP
[Supp = 30%, Conf = 50%, Lift = 60%]

R10: drugs3 = IJT drugs4 = IJT drugs6 = EJT drugs8 = EJT
[Supp = 25%, Conf = 50%, Lift = 50%]

R11: initial_cd4_cell_count = 5 dfn3_cd4 = 5
[Supp = 40%, Conf = 40%, Lift = 70%]

R12: dfn4_cd4 = 3 dfn4_cd4 = 2 dfn5_cd4 = 3 dfn7_cd4 = 5 dfn8_cd4 = 5
[Supp = 50%, Conf = 50%, Lift = 75%]

R13: dfn6_cd4 = 2 dfn7_cd4 = 2 dfn8_cd4 = 2 dfn9_cd4 = 3 dfn10_cd4 = 2
[Supp = 53%, Conf = 50%, Lift = 60%]

R14: dfn4_rna = 4 dfn5_rna = 3 dfn7_rna = 2 dfn8_rna = 3
[Supp = 52%, Conf = 50%, Lift = 65%]

R15: initial_rna_level = 9dfn2_rna = 3 dfn6_rna = A dfn9_rna = 3 dfn8_rna = A
[Supp = 50%, Conf = 50%, Lift = 50%]

R16: initial_cd4_cell_count = 5 drugs1 = IJP dfn3_cd4 = 5 drugs3 = JLT dfn5_cd4 = 3
drugs5 = JLT
[Supp = 25%, Conf = 40%, Lift = 50%]

R17: dfn4_cd4 = 3 dfn5_rna = 3 drugs5 = IJT drugs6 = IJT drugs7 = IJT dfn7_cd4 = 5
dfn8_rna = 6 drugs8 = JLT
[Supp = 35%, Conf = 50%, Lift = 60%]

R18: dfn2_cd4 = 2 dfn3_rna = 3 drugs3 = IJT dfn4_cd4 = 3 dfn6_rna = 3 drugs5 = IJT
drugs7 = EJT drugs8 = JLT
[Supp = 25%, Conf = 45%, Lift = 50%]

R19: dfn4_cd4 = 3 dfn5_rna = 3 drugs5 = IJT days5 = 6 drugs6 = IJT drugs7 = IJT
days7 = 5 dfn7_cd4 = 5 dfn8_rna = 6 drugs8 = JLT days8 = 3
[Supp = 33%, Conf = 46%, Lift = 50%]

R20: dfn4_cd4 = 3 days5 = 5 drugs5 = IJT drugs6 = EJT drugs7 = IJT
[Supp = 26%, Conf = 50%, Lift = 40%]

Table 4: Number of rule mined from HAT dataset using Apriori, MIR, NRRM and nVApriori with varying support

	Support (%)								
Algorithm	10	20	30	40	50	60	70	80	90
Apriori	1008	903	875	804	780	750	735	721	699
MIR	954	900	850	801	760	720	692	680	670
NRRM	590	575	542	519	491	472	453	430	401
nVApriori	325	300	285	270	255	242	231	219	207

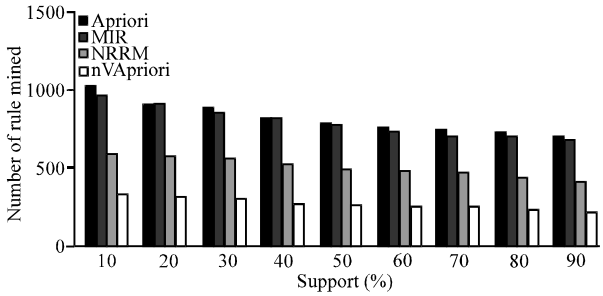

Fig. 2: Comparison of execution time between Apriori, MIR, NRRM and nVApriori

Comparative study: Also the HAT dataset is employed with Apriori, MIR, NRRM. They are produced different amount of rules with varying support values. These are summarized in the Table 4. These four algorithms are produced large amount of rule in low support value. If the support is increase, the mined rules are reduced (Table 4). From these results, totally the rules are reduced minimum 45% to maximum 70%. The proposed nVApriori is compared with Apriori, the reduction about 68 to 70%. The rule reduction is about 66 to 69%, compared with MIR and nVApriori. The proposed nVApriori reduced 45 to 48% of rules compared with NRRM. The nVApriori is successfully reduced large set of unwanted rules which are not related to HAT dataset.

The comparative study of execution time is plotted in the following Fig. 2. The proposed nVApriori is performed well in low and high support level. Also it is performed better than other algorithms. The hashing technique is reduced the search time and helps to improve the performance of nVApriori algorithm. The proposed transaction reduction mechanism is performed better than other techniques.

CONCLUSION AND FUTURE ENHANCEMENT

This study is implemented association rule mining algorithm with AIDS/HIV dataset. We developed a novel methodology for avoiding irrelevant rules. It also proved with rule mining technique. This work presented a new dataset for AIDS/HIV infected patients case history. The nVApriori algorithm is used to mine rules from the HAT dataset. The algorithm mined interestingness and surprising rules which are helped to understand the relationship between treatment records. The rules revealed many interesting information about CD4 cell counts, RNA levels, drugs between treatment and various patients. The number of days between each treatment played a major role in these mined rules. It needs a domain expert analysis to understand the mined rules. In future, the other data mining techniques may test with proposed dataset.

ACKNOWLEDGMENT

This study is financially supported by University Grant Commission, New Delhi, India.

REFERENCES

Al-Shalabi, L., 2011. Knowledge discovery process: Guide lines for new researchers. J. Artifi. Intell., 4: 21-28.

Besemann, C., A. Denton and A. Yekkirala, 2004. Differential association rule mining for the study of protein-protein interaction networks. Proceedings of the 4th Workshop on Data Mining in Bioinformatics, Seattle, Aug. 22, Washington, USA., pp: 72-80.

Gasmi, G., T. Hamrouni, S. Abdelhak, S. Ben Yahia and E.M. Nguifo, 2005. Extracting generic basis of association rules from SAGE data. Proceedings of the 8th International ECML/PKDD Workshop Discovery Challenge, Oct. 7, Porto, Portugal, pp: 1-6.

Gupta, N., N. Mangal, K. Tiwari and P. Mitra, 2006. Mining Quantitative Association Rules in Protein Sequences. In: Data Mining, Lecture Notes on Artificial Intelligence 3755, Williams, G.J. and S.J. Simo (Eds.). Springer-Verlag, Berlin Heidelberg, pp: 273-281.

Gupta, R.K. and D.P. Agrawal, 2009. Improving the performance of association rule mining algorithms by filtering insignificant transactions dynamically. Asian J. Inform. Manage., 3: 7-17.

Kwasnicka, H. and K. Switalski, 2006. Discovery of association rules from medical data-classical and evolutionary approaches. Proceedings of the 21th Autumn Meeting of Polish Information Processing Society Conference, (AMPIPS'06), Department of Computer Science, Wroclaw University of Technology, Poland, pp: 163-177.

Li, J., A. Wai-Chee Fu, H. He, J. Chen and H. Jin, 2005. Mining risk patterns in medical data. Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Aug. 21-24, Chicago, Illinois, USA., pp: 770-775.

Lopez, F.J., A. Blanco, F. Garcia and A. Marin, 2007. Extracting biological knowledge by fuzzy association rule mining. Proceedings of the IEEE International Conference on Fuzzy Systems, July 23-26, London, UK., pp: 1-6.

Microsoft Research, 2005. Microsoft scientists search for breakthroughs in HIV vaccine design. http://www.microsoft.com/presspass/press/2005/feb05/02-23hivvaccinepr.mspx.

Nehemiah, H.K., A. Kannan, K. Vijaya, Y.N. Jane and J.B. Merin, 2007. Employing clinical data dets for intelligent temporal rule mining and decision making, a comparative study. ICGST-BIME Int. J. Bioinform. Med. Eng., 7: 37-45.

Ohsaki, M., Y. Sato, H. Yokoi and T. Yamaguchi, 2003. A rule discovery support system for sequential medical data-in the case study of a chronic hepatitis dataset. proceedings of the Workshop on Discovery Challenge during the ECML/PKDD 2003, Sept. 22-26, Cavtat-Dubronok, Croatia, pp: 1-12.

Ordonez, C., C. Santana and L. Braal, 2000. Discovering interesting association rules in medical data. Proceedings of the ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, May 14, Dallas, Texas, USA., pp: 78-85.

Ordonez, C., N. Ezquerra and C. Santana, 2006. Constraining and summarizing association rules in medical data. Knowledge Inf. Syst., 9: 1-2.

Ramaraj, E.T. and K.R. Kumar, 2009. NVApriori: A novel approach to avoid irrelevant rules in association rule mining using n-cross validation technique. Int. J. Adv. Soft Comput. Appl., 1: 132-150.

Salleb, A., T. Turmeaux, C. Vrain and C. Nortet, 2004. Mining quantitative association rules in a atherosclerosis dataset. Proceedings of the 6th European Conference on Principles and Practice of Knowledge Discovery in Databases, Sept. 20-24, Pisa, Italy, pp: 98-103.

Sharma, S., A. Tiwari, S. Sharma and K.R. Pardasani, 2007a. Design of algorithm for frequent pattern discovery using lattice approach. Asian J. Inform. Manage., 1: 11-18.

Sharma, S., S. Khare and S. Sharma, 2007b. Measuring the interesting of classification rules. Asian J. Inform. Manage., 1: 43-49.

Thakur, R.S., R.C. Jain and K.R. Pardasani, 2007. Fast algorithms for mining multi-level association rules in large databases. Asian J. Inform. Manage., 1: 19-26.

Vararuk, A., I. Petrounias and V. Kodogiannis, 2008. Data mining techniques for HIV/AIDS data management in Thailand. J. Enterprise Inform. Manag., 21: 52-70.

Von-Borries, G.F., M.G. Fonseca, E. Castilho and P. Chequer, 1996. AIDS data mining using statistical techniques. Int. Conf. AIDS, 11: 152-152.