INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATION & MANAGEMENT

IJRCM

IJRCM

A Monthly Double-Blind Peer Reviewed Refereed Open Access International e-Journal - Included in the International Serial Directories
Indexed & Listed at:
Ulrich's Periodicals Directory ©, ProQuest, U.S.A., EBSCO Publishing, U.S.A., Index Copernicus Publishers Panel, Poland,
Open J-Gage, India [link of the same is duly available at Inflibnet of University Grants Commission (U.G.C.)]
as well as in Cabell's Directories of Publishing Opportunities, U.S.A.
Circulated all over the world & Google has verified that scholars of more than Hundred & Thirty Two countries/territories are visiting our journal on regular basis.
Ground Floor, Building No. 1041-C-1, Devi Bhawan Bazar, JAGADHRI – 135 003, Yamunanagar, Haryana, INDIA
www.ijrcm.org.in

CONTENTS

CALL FOR MANUSCRIPTS

We invite unpublished novel, original, empirical and high quality research work pertaining to recent developments & practices in the area of Computer, Business, Finance, Marketing, Human Resource Management, General Management, Banking, Insurance, Corporate Governance and emerging paradigms in allied subjects like Accounting Education; Accounting Information Systems; Accounting Theory & Practice; Auditing; Behavioral Accounting; Behavioral Economics; Corporate Finance; Cost Accounting; Econometrics; Economic Development; Economic History; Financial Institutions & Markets; Financial Services; Fiscal Policy; Government & Non Profit Accounting; Industrial Organization; International Economics & Trade; International Finance; Macro Economics; Micro Economics; Monetary Policy; Portfolio & Security Analysis; Public Policy Economics; Real Estate; Regional Economics; Tax Accounting; Advertising & Promotion Management; Business Education; Management Information Systems (MIS); Business Law, Public Responsibility & Ethics; Communication; Direct Marketing; E-Commerce; Global Business; Health Care Administration; Labor Relations & Human Resource Management; Marketing Research; Marketing Theory & Applications; Non-Profit Organizations; Office Administration/Management; Operations Research/Statistics; Organizational Behavior & Theory; Organizational Development; Production/Operations; Public Administration; Purchasing/Materials Management; Retailing; Sales/Selling; Services; Small Business Entrepreneurship; Strategic Management Policy; Technology/Innovation; Tourism, Hospitality & Leisure; Transportation/Physical Distribution; Algorithms; Artificial Intelligence; Compilers & Translation; Computer Aided Design (CAD); Computer Aided Manufacturing; Computer Graphics; Computer Organization & Architecture; Database Structures & Systems; Digital Logic; Discrete Structures; Internet; Management Information Systems; Modeling & Simulation; Multimedia; Neural Systems/Neural Networks; Numerical Analysis/Scientific Computing; Object Oriented Programming; Operating Systems; Programming Languages; Robotics; Symbolic & Formal Logic and Web Design. The above mentioned tracks are only indicative, and not exhaustive.

Anybody can submit the soft copy of his/her manuscript **anytime** in M.S. Word format after preparing the same as per our submission guidelines duly available on our website under the heading guidelines for submission, at the email addresses: infoijrcm@gmail.com or info@ijrcm.org.in.

GUIDELINES FOR SUBMISSION OF MANUSCRIPT

1. **COVERING LETTER FOR SUBMISSION**:

DATED: _____________

THE EDITOR
IJRCM

Subject: **SUBMISSION OF MANUSCRIPT IN THE AREA OF** ____.

(e.g. Finance/Marketing/HRM/General Management/Economics/Psychology/Law/Computer/IT/Engineering/Mathematics/other, please specify)

DEAR SIR/MADAM

Please find my submission of manuscript entitled '____' for possible publication in your journals.

I hereby affirm that the contents of this manuscript are original. Furthermore, it has neither been published elsewhere in any language fully or partly, nor is it under review for publication elsewhere.

I affirm that all the author (s) have seen and agreed to the submitted version of the manuscript and their inclusion of name (s) as co-author (s).

Also, if my/our manuscript is accepted, I/We agree to comply with the formalities as given on the website of the journal & you are free to publish our contribution in any of your journals.

NAME OF CORRESPONDING AUTHOR:
Designation:
Affiliation with full address, contact numbers & Pin Code:
Residential address with Pin Code:
Mobile Number (s):
Landline Number (s):
E-mail Address:
Alternate E-mail Address:

NOTES:
a) The whole manuscript is required to be in ***ONE MS WORD FILE*** only (pdf. version is liable to be rejected without any consideration), which will start from the covering letter, inside the manuscript.
b) The sender is required to mention the following in the **SUBJECT COLUMN** of the mail:
New Manuscript for Review in the area of (Finance/Marketing/HRM/General Management/Economics/Psychology/Law/Computer/IT/Engineering/Mathematics/other, please specify)
c) There is no need to give any text in the body of mail, except the cases where the author wishes to give any specific message w.r.t. to the manuscript.
d) The total size of the file containing the manuscript is required to be below **500 KB**.
e) Abstract alone will not be considered for review, and the author is required to submit the complete manuscript in the first instance.
f) The journal gives acknowledgement w.r.t. the receipt of every email and in case of non-receipt of acknowledgment from the journal, w.r.t. the submission of manuscript, within two days of submission, the corresponding author is required to demand for the same by sending separate mail to the journal.

2. **MANUSCRIPT TITLE:** The title of the paper should be in a 12 point Calibri Font. It should be bold typed, centered and fully capitalised.

3. **AUTHOR NAME (S) & AFFILIATIONS**: The author (s) **full name**, **designation**, **affiliation** (s), **address**, **mobile/landline numbers**, and **email/alternate email address** should be in italic & 11-point Calibri Font. It must be centered underneath the title.

4. **ABSTRACT**: Abstract should be in fully italicized text, not exceeding 250 words. The abstract must be informative and explain the background, aims, methods, results & conclusion in a single para. Abbreviations must be mentioned in full.

5. **KEYWORDS**: Abstract must be followed by a list of keywords, subject to the maximum of five. These should be arranged in alphabetic order separated by commas and full stops at the end.

6. **MANUSCRIPT**: Manuscript must be in ***BRITISH ENGLISH*** prepared on a standard A4 size ***PORTRAIT SETTING PAPER***. It must be prepared on a single space and single column with 1" margin set for top, bottom, left and right. It should be typed in 8 point Calibri Font with page numbers at the bottom and centre of every page. It should be free from grammatical, spelling and punctuation errors and must be thoroughly edited.

7. **HEADINGS**: All the headings should be in a 10 point Calibri Font. These must be bold-faced, aligned left and fully capitalised. Leave a blank line before each heading.

8. **SUB-HEADINGS**: All the sub-headings should be in a 8 point Calibri Font. These must be bold-faced, aligned left and fully capitalised.

9. **MAIN TEXT**: The main text should follow the following sequence:

   INTRODUCTION

   REVIEW OF LITERATURE

   NEED/IMPORTANCE OF THE STUDY

   STATEMENT OF THE PROBLEM

   OBJECTIVES

   HYPOTHESES

   RESEARCH METHODOLOGY

   RESULTS & DISCUSSION

   FINDINGS

   RECOMMENDATIONS/SUGGESTIONS

   CONCLUSIONS

   SCOPE FOR FURTHER RESEARCH

   ACKNOWLEDGMENTS

   REFERENCES

   APPENDIX/ANNEXURE

   It should be in a 8 point Calibri Font, single spaced and justified. The manuscript should preferably not exceed ***5000 WORDS***.

10. **FIGURES &TABLES**: These should be simple, centered, separately numbered & self explained, and **titles must be above the table/figure**. **Sources of data should be mentioned below the table/figure**. It should be ensured that the tables/figures are referred to from the main text.

11. **EQUATIONS**: These should be consecutively numbered in parentheses, horizontally centered with equation number placed at the right.

12. **REFERENCES**: The list of all references should be alphabetically arranged. The author (s) should mention only the actually utilised references in the preparation of manuscript and they are supposed to follow **Harvard Style of Referencing**. The author (s) are supposed to follow the references as per the following:

- All works cited in the text (including sources for tables and figures) should be listed alphabetically.
- Use (**ed.**) for one editor, and (**ed.s)** for multiple editors.
- When listing two or more works by one author, use --- (20xx), such as after Kohl (1997), use --- (2001), etc, in chronologically ascending order.
- Indicate (opening and closing) page numbers for articles in journals and for chapters in books.
- The title of books and journals should be in italics. Double quotation marks are used for titles of journal articles, book chapters, dissertations, reports, working papers, unpublished material, etc.
- For titles in a language other than English, provide an English translation in parentheses.
- The location of endnotes within the text should be indicated by superscript numbers.

PLEASE USE THE FOLLOWING FOR STYLE AND PUNCTUATION IN REFERENCES:

BOOKS

- Bowersox, Donald J., Closs, David J., (1996), "Logistical Management." Tata McGraw, Hill, New Delhi.
- Hunker, H.L. and A.J. Wright (1963), "Factors of Industrial Location in Ohio" Ohio State University, Nigeria.

CONTRIBUTIONS TO BOOKS

- Sharma T., Kwatra, G. (2008) Effectiveness of Social Advertising: A Study of Selected Campaigns, Corporate Social Responsibility, Edited by David Crowther & Nicholas Capaldi, Ashgate Research Companion to Corporate Social Responsibility, Chapter 15, pp 287-303.

JOURNAL AND OTHER ARTICLES

- Schemenner, R.W., Huber, J.C. and Cook, R.L. (1987), "Geographic Differences and the Location of New Manufacturing Facilities," Journal of Urban Economics, Vol. 21, No. 1, pp. 83-104.

CONFERENCE PAPERS

- Garg, Sambhav (2011): "Business Ethics" Paper presented at the Annual International Conference for the All India Management Association, New Delhi, India, 19–22 June.

UNPUBLISHED DISSERTATIONS AND THESES

- Kumar S. (2011): "Customer Value: A Comparative Study of Rural and Urban Customers," Thesis, Kurukshetra University, Kurukshetra.

ONLINE RESOURCES

- Always indicate the date that the source was accessed, as online resources are frequently updated or removed.

WEBSITE

- Garg, Bhavet (2011): Towards a New Natural Gas Policy, Political Weekly, Viewed on January 01, 2012 http://epw.in/user/viewabstract.jsp

A I R E P: A NOVEL SCALED MULTIDIMENSIONAL QUANTITATIVE RULES GENERATION APPROACH

SAPNA JAIN
RESEARCH SCHOLAR
JAMIA HAMDARD UNIVERSITY
DEPARTMENT OF COMPUTER SCIENCE
NEW DELHI

DR. M. AFSHAR ALAM
PROFESSOR
JAMIA HAMDARD UNIVERSITY
DEPARTMENT OF COMPUTER SCIENCE
NEW DELHI

DR. RANJT BISWAS
PROFESSOR
JAMIA HAMDARD UNIVERSITY
DEPARTMENT OF COMPUTER SCIENCE
NEW DELHI

ABSTRACT
This paper is aimed to propose an AIREP algorithm which uses association rule mining (ARM) discovery technique which is widely used in various data mining applications. The task of discovering scalable rules from the multidimensional database with reduced support is an area for exploration for research . In this paper we have proposed an algorithm AIREP to generate scaled rules using the pruning technique. The algorithm also prunes the database at each step in order to reduce the search space and to reduce the unnecessary frequent subset generation at each step. IREP induces a set of rules in disjunctive form growing and pruning phases and help in generating the scaled and efficient association rule. Experimental on real world datasets show that the proposed approach improves performance over existing approaches by minimizing the explosion of number of rules involving frequent items and without missing the frequent itemsets involving rare items.

KEYWORDS
Association rules, Data Mining algorithms, Frequent itemsets, Pre-pruning and Post-pruning.

INTRODUCTION
Due to increase in the amount of information available for analysis is increasing, scalability of data mining applications is becoming a critical factor. The scalability of data mining techniques is very important due to the rapid growth in the size of databases. Scalability in data mining can be done in large datasets by using scalable input /output architecture, minimizing transaction time. We can scale the data mining algorithms by using data reduction techniques such as aggregation, dimensional reduction, compression, and discretization. We can use reduce algorithm complexity by using parallelization methods (M.S.Danessh 2010, Jyothi Pillai 2010).

Data mining represents techniques for discovering knowledge patterns hidden in large databases (R Uday Kiran 2009). Data mining tasks can be classified into two categories, Descriptive Mining and Predictive Mining. The Descriptive Mining techniques such as Clustering, Association Rule Discovery, Sequential Pattern Discovery, is used to find human-interpretable patterns that describe the data. The Predictive Mining techniques like Classification, Regression, Deviation Detection, use some variables to predict unknown or future values of other variables (Jyoti Pillai 2010).

The discovery of association rules in transaction databases is an important data-mining problem because of its wide application in many areas, such as market basket analysis, decision support, financial forecast, collaborative recommendation, and prediction. For data mining approach, the association rule set is usually used for prediction. However, traditional association rule algorithms typically generate a large number of rules, most of which are unnecessary when used for prediction (Anurag Choubey 2011). The problem of mining association rules is to generate a set of potentially interesting association rules in a data set of sessions that have support higher than the specified minimum support threshold and assign an interestingness value to all rules based on an interestingness measure (Maja Dimitrij 2011) .

Association rules are an commonly used representation to describe the effective discovery of correlations among the underlying data in large databases. ..In this model, the set I = {il, i2 , . . . , im} is a collection of items or attributes. The database DB consists of a set of transactions, where each transaction is a subset of items in I. An association rule is an implication of the form $X \rightarrow Y$ with $X, Y \subset I$ and $X \cap Y = 0$. The meaning of the rule is that a transaction containing items in X will likely contain items in Y.(Neelu Khare 2009, Tarek F. Gharib 2010). To determine whether an association rule is interesting, two measures are used: support and confidence. An association rule, $X \rightarrow Y$, has support s% in DB if s% of transactions in DB contains items in X U Y (C.S.Kanimozhi 2009). The same association rule is said to have confidence c% if among the transactions containing items in X, there are c% of them containing also items in Y. So, the problem is to find all association rules which satisfy predefined minimum support and minimum confidence constraints (S. Lofti 2009).

Many interesting and efficient algorithms have been proposed for mining association rules for these Boolean attributes, for examples, Apriori (R. Agrawal 1994), DHP (J S. Park 1995), and partition algorithms (A Savasere 1995). However, in a real database, attributes can be quantitative and the corresponding domains can have multiple values or a continuous range of values, for examples, age, and salary. A common approach to the QAR mining problem is to transform it into a problem of conventional BAR mining (R. Agrawal 1995 , Agrawal R 1993). Existing algorithms involve discretizing the domains of quantitative attributes into intervals so as to discover quantitative association rules. For each distinct value of a quantitative or categorical attribute, the pair <attribute, value> is mapped to a Boolean attribute and then algorithms for mining BARs are applied.

In many cases, the number of intervals associated with an attribute is large hence when we join the attributes in the mining process, the number of itemsets (i.e., a set of <attribute, interval> pairs) can become prohibitively large. As a result, effective techniques to prune the large search space of QAR mining and avoid the costly generation of a large number of candidate sets are necessary in order to develop an efficient algorithm for the problem. Also these intervals may not be meaningful enough for human experts to easily obtain nontrivial knowledge (Suraj Srivastava 2010).

We may be able to build a decision tree which perfectly reflects the data but the tree may not be generally applicable. Pruning is a technique for simplifying and hence generalising a decision tree. Error-Based Pruning replace sub-trees with leaves .It uses decision class is the majority. The pruning is based on predicted error rates. It prunes sub trees which result in lower predicted error rate (Fadi Thabtah 2006). The two common techniques are:

1) Cost Complexity Pruning: The predicted error rate modelled as weighted sum of complexity and error on training set and the test cases used to determine weighting .

2) Reduced Error Pruning: In this approach we use test set to assess error rate directly.

Pruning mechanisms are an important component of practical learning algorithms for decision trees and lists and are essential for learning comprehensible and accurate classifers in the presence of noise (Tzung-Pei Hong 2008).

This research work is the extension of the previous work where we have proposed algorithm for Aprori-UB which uses multidimensional access method UB-tree to generate better association rules with high support and confidence. The Aprori-Ub approach reduces not only the number of item sets generated but also the overall execution time of the algorithm. In this paper we have used the incremental reduced error pruning method to minimize the error during the generation of the scaled multidimensional association rule (Przemyslaw Kazienko 2009).

The rest of the paper is organized as follows. Section 2 gives the overview of the previous work done in the same field. Section 3 explains the concepts used in this paper. Section 4 gives the proposed work. Section 5 gives the experimentation details. Section 6 and Section 7 discusses the conclusion and future scope.

RELATED WORK

In the previous section we have introduced the basic concept of scalable Data Mining, Association Rule mining, Pruning. A brief overview of various algorithms, concepts and techniques defined in different research papers have been given in this section.

The popular approaches like Apriori discover association rules based on frequent itemsets which are extracted by fixing the same (or single) minimum support for all items (J.Shahrabi 2009). Let's start by defining association rule mining that finds correlation between the items in the data sets (] Michael Hahsler 2009, Laszlo Szathmary 2010). ARM is two step process i.e. first finds all frequent itemsets having minimum support and then generate strong association rules having minimum confidence, from the frequent itemsets. i.e. ARM finds out association rules which satisfy the predefined minimum support and confidence from a given database. Let J = {i1, i2…im} be a set of items, D be a set of database transactions and T is transaction contains set of items such that (. Mart´inez-Ballesterosa 2011). Let A, B be set of items and T is said to contain A, iff A⊂T. An association rule is an implication of the form A⇒B holds, where A⊂ J also B⊂J and A∩B is NULL. The rule holds in transactions set D with support s, where s is the percentage of both A and B. The A⇒B rule has confidence c in the transaction set D if c is the percentage of transactions in D containing A that also contains B (Rakhi Garg 2011).

Association analysis is an important data mining technique, widely used in commercial, financial, telecommunications, medical fields and so on, but rarely applied in the bibliometric analysis [12]. Our research shows that association rules can discover information hidden in the keywords, publications, authors, research institutions and other materials. In particular, it can instruct researchers to find research fields and techniques related to its research direction. At the same time, it helps them broaden research ideas and even discover new research fields by providing relevant publications, authors and research institutions. This has significant instructive value to research works.

CONCEPT USED

Associative algorithms normally derive a large set of rules since (R. Uday kiran 2009) classification data sets are typically highly correlated and (Anurag Choubey 2011) association rule mining approaches that consider all attribute values combinations in the database are used for rules discovery. As a result, there have been many attempts to reduce the size of their classifiers, mainly focused on preventing rules that are either redundant or misleading from taking any role in the prediction process of test data objects. The removal of such rules can make the classification process more effective and accurate (Fadi Thabtah 2006).

Several pruning methods have been used effectively to reduce the size of the classifiers, some of which have been adopted from decision trees, like pessimistic estimation, others from statistics such as chi-square testing (x^2). These pruning techniques are utilised during either rule discovery or the construction of the classifier. For instance, a very early pruning step, which eliminates rule items that do not pass the support threshold, may occur in the process of finding frequent rule items[3,2]. Another pruning such as chi-square testing may take place when generating the rules, and a late pruning method like database coverage may be used after discovering all potential rules.

There are various pruning methods used and each method is cast in the framework of the search in the state space (Floriana Esposito 1999) .Some of the pruning methods are :-

REDUCED ERROR PRUNING (REP)

This method uses the pruning set in order to evaluate the goodness of a sub tree of tree T_{ax}. Search is accomplished in the any-depth branch pruning state pace, Ẏp(T),{πT }) according to the first-better search strategy and a post-order traversal. The evaluation function f is defined as follows:

$$f(T) = -\sum_{t€r} e(T)$$

where e(t) is the number of errors made by node t during the classification of the examples in the pruning set. The search in the space moves from a state ¹ to a state T^1 € T\`t($Ẏ_r$) if the inequality

f(T\`) >= f(T) holds or equivalently if

$$\sum_{t€r} e(T) <= \sum e(T)$$

The states to be explored are generated according to the order defined by bottom-up methods, hence there is no choice of the best state to be reached, starting from another state (Kahayan Lal 2010).

PESSIMISTIC ERROR PRUNING

This pruning method, proposed in (Quinlan J.R 1987) as well, is characterized by the fact that it avoids using an independent pruning set. Search is accomplished in the any-depth branch pruning state space, ({Ẏp(T),{πT }), according to thefirst-better search strategy and a pre-order traversal. The evaluation function f is defined as follows:

f (t) = -∑n\`t

t€ Ẏp where n\`(t) = [e(t) + ½]and e(t) is the number of errors made by node t during the classification of the examples in the training set. Indeed, let T\` be the arrival state of an edge out coming from T such that it is obtained by pruning a node t€T. Then it can be proved that f(T\`) – f(T) = n(T_t) – n1(t)

where n(T_t) = -∑ e(s) +| ẎT_t | . Pruning is accomplished also when the following condition holds: -SE(n\`-$T_t$)) <= f(T\`) – f(T).

where SE is the standard error. This is equivalent to prune when

as stated in Quinlan's original formulation. Therefore, there is no evaluation of the best pruning to perform among the possible ones, and the first pruning operation that turns out to be good is performed. It follows that the search strategy adopted is the first-better with a pre-order traversal (M. Mart´inez-Ballesterosa 2010, Maja Dimitrijevic 2011).

It should also be noted that the top-down approach to tree pruning used in PEP is not justified when there is no guarantee that all subtrees of a pruned branch T_t have to be pruned. Indeed, it may happen that by pruning a node t other nodes that should not be pruned according to the same criterion are actually discarded. However, this top-down approach gives the pruning method a high run-speed, with a computational complexity being linear in the number of nodes (Fang Li 2009, Preetham Kumar 2008).

MINIMUM ERROR PRUNING (MEP)

This method was proposed as a bottom-up approach for searching a single tree that minimizes the expected error rate. For a k-class problem, the expected probability that an observation reaching a node t belongs to the ith class is the following:

$$pi(t) = \frac{n_i(t) + p_{aim}}{N(t) + m}$$

where n_i (t) is the number of training examples in t classified into the ith class, p_{ai} is the apriori probability of the ith class, m is a parameter of the estimate method, N(t) is the number of training examples reaching t. When a new observation reaching t is classified, the expected error rate is given by

$$EER = \min(i)\ \{1 - p_i - (t)\}$$

In the MEP method, the choice of m is critical. We have to decide to choose the value m using an independent pruning set. More precisely, given a set of possible values for m, we select that returning the smallest tree with the lowest empirical error rate on an independent pruning set. Therefore, this is an example of two-phased pruning method (Pratima Gautam 2010).

CRITICAL VALUE PRUNING (CVP)

It is a pruning method that searches in the one-depth branch pruning state space, (ŸP(T),π). The evaluation function associated with this reduced space is given by the sum of the values taken by the selection measure in each internal node of the tree (f (T = 0) [17]. Therefore, if GR(t) is the gain ratio at node t, the evaluation function can be defined as $f(T) = \sum GR(T)/T_t$.A tree T` will be generated from a tree [1] if it happens that f(T`) = min $f(T`)/t\pi_r$.

The search goes on according to a hill-climbing strategy until the minimum tree $T^1$is reached (Jianying Hu 2007).

At the end of the search, the number of the explored states will be $|\dot{Y}T_{max}|$, denoted as $T_{max}...T_{max-1}$. This method is two-phased as the previous one. However, in this case not all states traversed are considered in the second phase. A traversed state Ti, i<=1 <= $\dot{Y}_x$is considered to be transient if it happens that

$f(Tj)-f(T_{j-1}) >= f(Tj)-f(T_{j-1})$, j<=1.

For the second phase, we have to choose the best tree among the sequence of the pruned trees by measuring both the significance of the tree as a whole and its predictive ability. The significance of the tree is estimated by means of the G statistics, which evaluates the degree of interdependence between the leaves of a tree and the classes of the problem: It will be higher for fully expanded trees that correctly classify the whole set of examples. The weakness of this measure is that a test on this statistics is only able to establish whether the predictive ability of a tree is meaningful, but it cannot be used to choose among trees that pass the test (Maja Dimitrijevic 2011).

COST –COMPLEXITY PRUNING

This pruning method is characterized by two phases:

(1) Selection of a family of sub trees of T_{max} according to some heuristics.

(2) Choice of the best tree in the family according to an accurate estimate of the actual error rate.

For what concerns the first phase, search is performed in the any depth branch pruning state space according to a hill-climbing strategy and a post-order traversal. The evaluation function can be defined as follows: $f(T) = -\sum_{t \in \dot{Y}_r} e(T)$

where e(t) is the number of errors made by node t on the training, growing set. It is possible to move from T to T` =πr(t) if it happens that

$$\frac{f(T) - F(T`)}{|\dot{Y}_T| - |\dot{Y}_{T-1}|} = f(T) - F(T`)$$

For each reached state, the next state that gives the lowest value of the ratio and apparent error rate increase on and number of leaves decrease is detected. The search goes on until the smallest tree $T_{1\ is}$ reached(E. Chandra 2010, Przemyslaw Kazienko 2009).

The second phase of the method aims at selecting the best among the trees traversed in the first phase. Once again, not all states are considered, and a transient state can be defined as follows: let

$T_{max} = T^{m,}-T^{m-1}$,....T_0...T_1...T_m be the states followed by the search process and let ^{T}the complexity parameter of a state T^i,T^m then it is transient if бi = бi-1 Jyothi Pillai 2010, Fan Lilin 2010).

PROPOSED WORK

We define ,C_k as a candidate itemset of size k ,Z_k as a frequent itemset of size k, An AIREP algorithm is

1. Find frequent set L_{k-1}
2. Join step: C_k is generated by joining L_{k-1} with itself (cartesian product L_{k-1} x L_{k-1})
3. Prune step : Use theIncremental Reduced Error pruning to generate scalable single rule.
4. Frequent set L_k has been achieved.

The proposed AIREP (Aprori Incremental Reduced Error Pruning) pseudo code:

```
AIREP (T, ụ)
Z1 ← large multidimensional itemsets that appear in more than
Of large item set ụ transactions
K ← 2
While ( Zk-1 ≠ 0 )
Ck ← Generate (Zk-1 ) // join and prune step
// using IREP
procedure I-REP (Examples, SplitRatio)
Theory = Ө ;
While Positive (Examples) ≠ Ө;
Clause = Ө;
Split Examples (Split Ratio, Examples, Growing Set, Pruning Set)
Cover = Growing Set
While Negative (Cover) ≠ Ө ;
Clause = Clause ᴜ Find Literal (Clause; Cover)
Cover = Cover (Clause, Cover)
loop
NewClause = BestSimplification (Clause, PruningSet)
if Accuracy(NewClause,PruningSet) < Accuracy(Clause,PruningSet)
exit loop
Clause = NewClause
if Accuracy(Clause,PruningSet)<=Accuracy(fail,PruningSet)
exit while
Theory = Theory ᴜ Clause
Examples = Examples -Cover
return (Theory)
```

```
// end of IREP
//frequent set generation
        for transaction t € Z
        Ck ← Subset(Ck,t)
        for candidates c € Ct
                count[c] =count[c + 1]
        Zk ← { c € Ck| count[c] >= e}
k ← k+1
return Zk
```

FIGURE 1: PSEUDOCODE OF PROPOSED AIREP ALGORITHM

The basic idea of Incremental Reduced Error Pruning (IREP) is that instead of first growing a complete concept description and pruning it thereafter, each individual clause will be pruned right after it has been generated. This ensures that the algorithm can remove the training examples that are covered by the pruned clause before subsequent clauses are learned thereby preventing these examples from influencing the learning of subsequent clauses.

Figure 1 shows pseudo-code for this algorithm. As usual, the current set of training examples is split into a growing (usually 2/3) and a pruning set (usually 1/3). However, not an entire theory, but only one clause is learned from the growing set. Then, literals are deleted from this clause in a greedy fashion until any further deletion would decrease the accuracy of this clause on the pruning set. Single pruning steps can be performed by submitting a one-clause theory to the same BestSimplification subroutine used in REP or, as in our implementation, one can use a more complex pruning operator that considers every literal in a clause for pruning. The best rule found by repeatedly pruning the original clause is added to the concept description and all covered positive and negative examples are removed from the training growing *and* pruning set. The remaining training instances are then redistributed into a new growing and a new pruning set to ensure that each of the two sets contains the predefined percentage of the remaining examples. From these sets the next clause is learned. When the predictive accuracy of the pruned clause is below the predictive accuracy of the empty clause (i.e., the clause with the body fail), the clause is not added to the concept description and I-REP returns the learned clauses. Thus,
the accuracy of the pruned clauses on the pruning set also serves as a stopping criterion. Post-pruning methods are used as pre-pruning heuristics.

In figure 2 the attributes of the dataset are divided into instances and converted into divided attributes. In order to build a rule, IREP uses the following strategy. First the uncovered examples are randomly partitioned into two subsets, a growing set and a pruning set. Next, a rule is grown. It begins with an empty conjunction of conditions, and considers adding to this any condition of the form $Z_n = U_i$, $Z_n <= \Theta$ or $Z >= \Theta$ where Z_n is a nominal attribute and u is a legal value for Z_n, or Z_c is a continuous variable and Θ is some value for Z_c that occurs in the
training data. After growing a rule, the rule is immediately pruned. After growing a rule, the rule is immediately pruned To prune a rule, our implementation considers deleting any final sequence of conditions from the rule and chooses the deletion that maximizes the function

$$u(Rule,PrunePos,PruneNeg) = \frac{X + (N - n)}{X + N}$$

where X (respectively N), is the total number of examples in PrunePos ,PruneNeg and p ,n, is the number of examples in PrunePos ,PruneNeg covered by Rule.This process is repeated until no deletion improves the value of u.

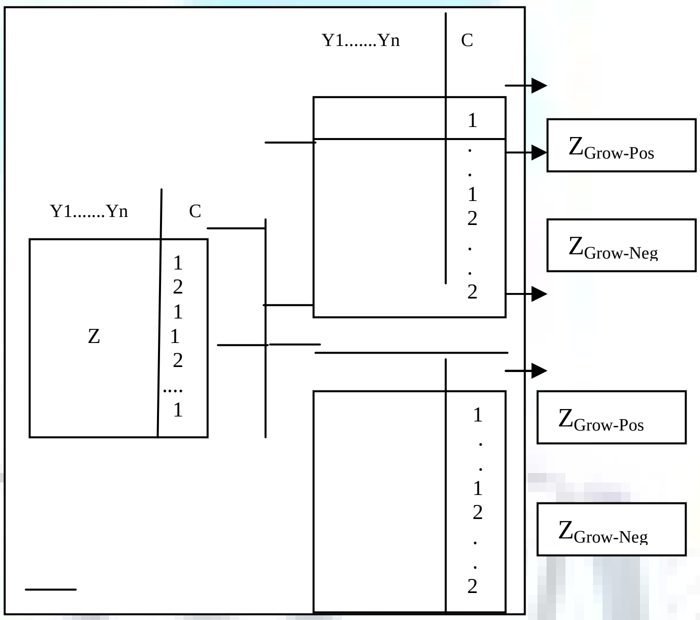
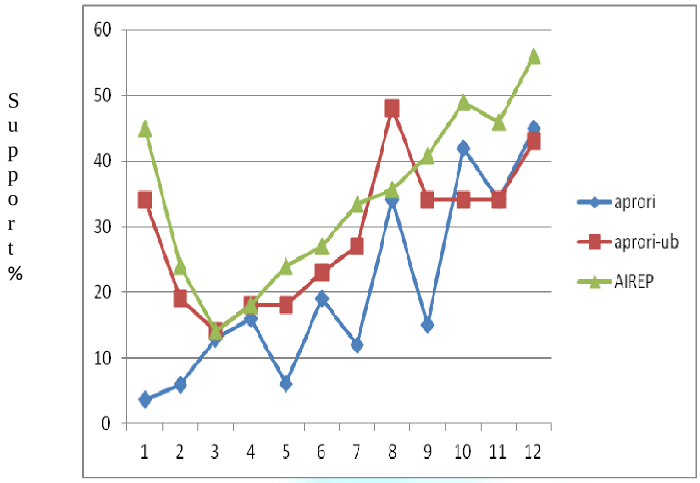


FIGURE 2: PARTITIONING OF ORIGINAL DATA SET OF LABELLED INSTANCES

EXPERIMENTATION

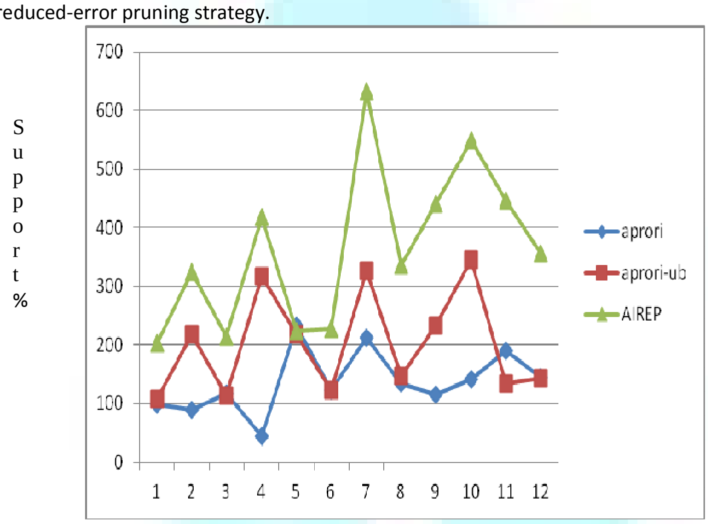
The experiments with proposed AIREP showed that the rules generated were fast and efficient. We used java programming language to implement the AIREP algorithm. We used the synthetic and real world dataset to test the efficiency of the algorithm. The experiment were conducted on breast-cancer real life dataset and other synthetic dataset .The dataset is belongs to UCI Machine repository datasets.The features of the dataset is computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. The attributes of the dataset includes ID number, Diagnosis, radius, texture, area, smoothness.



Items sorted based on support.

FIGURE 3: SUPPORT% OF REAL WORLD DATASET USING AIREP

AIREP learns the clauses in the order in which they will be used by a PROLOG interpreter. Before subsequent rules are learned, each clause is completed (learned *and* pruned) and all covered examples are removed. Therefore, the AIREP approach eliminates the problem of incompatibility between the separate-and conquer learning strategy and the reduced-error pruning strategy.



Item sorted based on support

FIGURE 4: SUPPORT% OF SYNTHETIC WORLD DATASET USING AIREP

The experiment show that AIREP's asymptotic complexity is *O*(*n* log2 *n*), *n* being the size of the training set. The cost of growing one clause in REP is*O*(*n* log *n*), because for selecting each of the £(log *n*) literals. Thus, a constant number of conditions is tested against *O*(*n*) examples. AIREP considers *every* literal in the clause for pruning. Therefore, each of the £(log *n*) literals has to be evaluated on the £(*n*) examples in the pruning set until the final clause has been found, i.e., at most *O*(log *n*) times. Thus, the cost of pruning one clause is *O*(*n* log2 *n*). Assuming that AIREP stops when the correct theory of constant size has been found, the overall cost is also *O*(*n* log2 *n*). This is significantly lower than the cost of growing an over fitting theory which has been shown to be -(*n*2 log *n*) under the same assumptions .

No.	Label	Count
1	10-19	0
2	20-29	1
3	30-39	36
4	40-49	90
5	50-59	96
6	60-69	57
7	70-79	6
8	80-89	0
9	90-99	0

TABLE 5: SUPPORT COUNT OF THE ATTRIBUTES OF DATASET

We tested our proposed method that contained a wide range of item numbers (more than 900 items) in a transaction, to find only multi dimensional quantitative association rules. To use data, we appended the two attribute fields namely tumor size and age using data generating programs based on randomization functions, in order to ensure that the final test data remained unbiased. The data contained 49,100 transactions. The number of rules obtained under different values of constraints, such as minimum support and minimum size of item-sets generated scaled association rules.

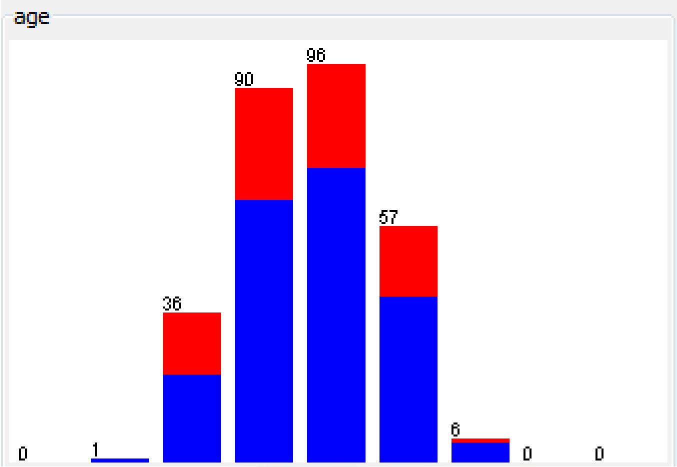


FIGURE 6: AGE ATTRIBUTE RULE GENERATION OF BREAST-CANCER DATASET

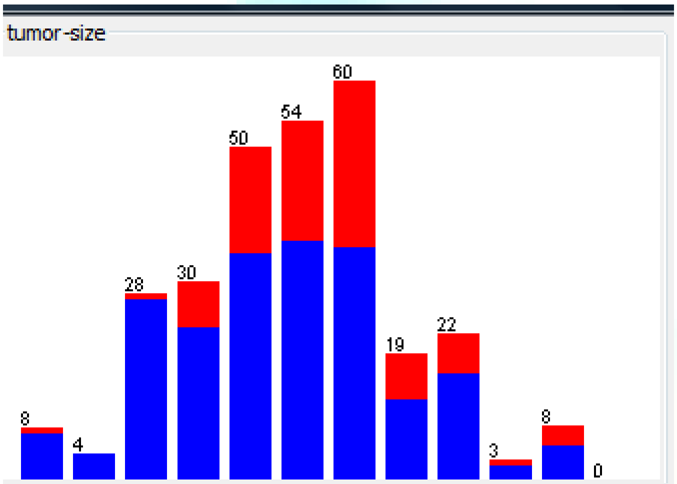
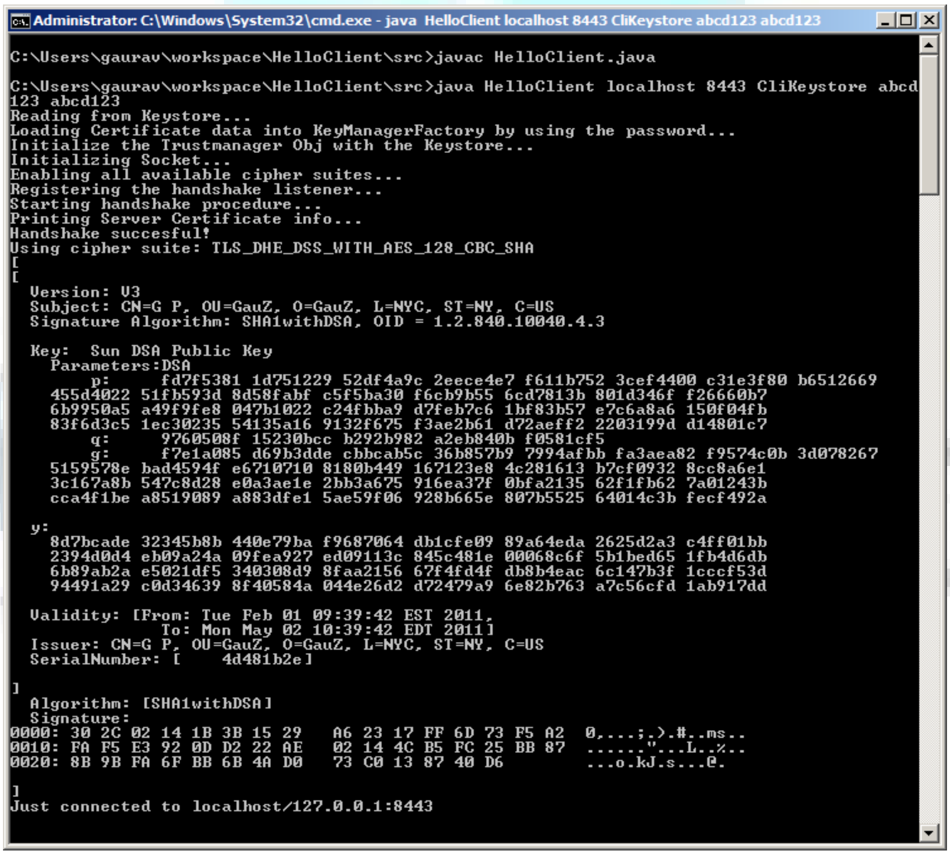


FIGURE 7: TUMOR-SIZE ATTRIBUTE RULE GENERATION OF BREAST-CANCER DATASET



```
Administrator: C:\Windows\System32\cmd.exe - java  HelloClient localhost 8443 CliKeystore abcd123 abcd123

C:\Users\gaurav\workspace\HelloClient\src>javac HelloClient.java

C:\Users\gaurav\workspace\HelloClient\src>java HelloClient localhost 8443 CliKeystore abcd
123 abcd123
Reading from Keystore...
Loading Certificate data into KeyManagerFactory by using the password...
Initialize the Trustmanager Obj with the Keystore...
Initializing Socket...
Enabling all available cipher suites...
Registering the handshake listener...
Starting handshake procedure...
Printing Server Certificate info...
Handshake succesful!
Using cipher suite: TLS_DHE_DSS_WITH_AES_128_CBC_SHA
[
[
  Version: V3
  Subject: CN=G P, OU=GauZ, O=GauZ, L=NYC, ST=NY, C=US
  Signature Algorithm: SHA1withDSA, OID = 1.2.840.10040.4.3

  Key:  Sun DSA Public Key
    Parameters:DSA
        p:     fd7f5381 1d751229 52df4a9c 2eece4e7 f611b752 3cef4400 c31e3f80 b6512669
    455d4022 51fb593d 8d58fabf c5f5ba30 f6cb9b55 6cd7813b 801d346f f26660b7
    6b9950a5 a49f9fe8 047b1022 c24fbba9 d7feb7c6 1bf83b57 e7c6a8a6 150f04fb
    83f6d3c5 1ec30235 54135a16 9132f675 f3ae2b61 d72aeff2 2203199d d14801c7
        q:     9760508f 15230bcc b292b982 a2eb840b f0581cf5
        g:     f7e1a085 d69b3dde cbbcab5c 36b857b9 7994afbb fa3aea82 f9574c0b 3d078267
    5159578e bad4594f e6710710 8180b449 167123e8 4c281613 b7cf0932 8cc8a6e1
    3c167a8b 547c8d28 e0a3ae1e 2bb3a675 916ea37f 0bfa2135 62f1fb62 7a01243b
    cca4f1be a8519089 a883dfe1 5ae59f06 928b665e 807b5525 64014c3b fecf492a

  y:
    8d7bcade 32345b8b 440e79ba f9687064 db1cfe09 89a64eda 2625d2a3 c4ff01bb
    2394d0d4 eb09a24a 09fea927 ed09113c 845c481e 00068c6f 5b1bed65 1fb4d6db
    6b89ab2a e5021df5 340308d9 8faa2156 67f4fd4f db8b4eac 6c147b3f 1cccf53d
    94491a29 c0d34639 8f40584a 044e26d2 d72479a9 6e82b763 a7c56cfd 1ab917dd

  Validity: [From: Tue Feb 01 09:39:42 EST 2011,
               To: Mon May 02 10:39:42 EDT 2011]
  Issuer: CN=G P, OU=GauZ, O=GauZ, L=NYC, ST=NY, C=US
  SerialNumber: [    4d481b2e]

]
  Algorithm: [SHA1withDSA]
  Signature:
0000: 30 2C 02 14 1B 3B 15 29   A6 23 17 FF 6D 73 F5 A2  0,....;.).#..ms..
0010: FA F5 E3 92 0D D2 22 AE   02 14 4C B5 FC 25 BB 87  ......"...L..%..
0020: 8B 9B FA 6F BB 6B 4A D0   73 C0 13 87 40 D6        ...o.kJ.s...@.

]
Just connected to localhost/127.0.0.1:8443
```

FIGURE 8: RULE GENERATION AFTER PRUNING

The efficiency increases with an increase in minsupport because an itemset now needs to be present in a larger number of transactions to eventually make it to the large set. The candidate set contains itemsets expected to be large as well as those expected to be small.

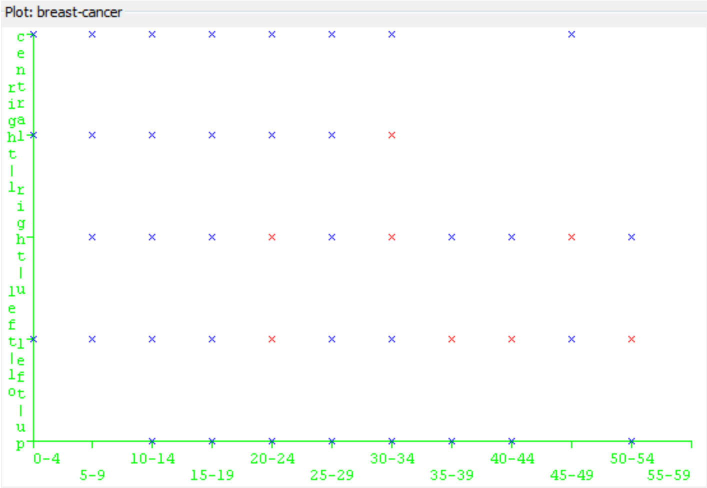


FIGURE 9: COMPARISON OF BREAST-CANCER DATASET ATTRIBUTES

Scaled Figure 8 bears out that most of the candidate itemsets expected to be large indeed turn out to be large. Initially, there is a large increase in the fraction of itemsets expected to be small in the candidate set as minsupport increases. Candidate itemsets as minsupport increases. Figure 7,8 shows efficiency of the pruning function optimization, with the remaining tuple optimization turned . It plots the fraction of new itemsets pruned due to this optimization .Figure 9 shows the results of the rule generated after pruning. The effectiveness of the optimization increases with an increase in minsupport as we can use a smaller value for. We also measured the pruning of new and old item sets when both the remaining tuple and pruning function optimizations were turned on. The curves for combined pruning tracked closely the two curves for the remaining tuple optimization. The pruning function optimization does not prune old candidate itemsets. The scaled rules obtained were classified on the basis of the attributes such as and generated with high confidence and support.

CONCLUSION

The goal of mining association rules is to discover important associations among items in a database of transactions such that the presence of some items will imply the presence of other items. The problem of mining association rules has been decomposed into two sub-problems: discovering the large Item-sets, and then generate rules based on these large Item-sets. The attention has been placed on the first sub-problem since the second sub-problem is quite straightforward up to some extent. Thus, there have been several algorithms proposed to solve the first sub-problem. These researches in algorithms of mining association rules are basically motivated by the fact that the amount of the processed data in mining association rules is huge; thus it is crucial to devise efficient algorithms to conduct mining on such data. The rules that we discover have one item in the consequent and a union of any number of items in the antecedent. We tested the effectiveness of our algorithm by applying it to sales data obtained from a large retailing company. For this data set, the algorithm exhibited excellent performance. The estimation procedure exhibited high accuracy and the pruning techniques were able to prune out a very large fraction of itemsets without measuring them.

FUTURE WORK

As a part of future work, we are going to investigate appropriate methodology for assigning confidence values in a dynamic manner to generate rare association rules in an efficient manner.

The existing approaches extract frequent item sets involving rare items by assigning minimum support to each item and employing an iterative process to discover frequent item sets. We are going to investigate how popular noniterative approaches like frequent-pattern growth approaches can be extended to assign minimum support to each item to extract frequent itemsets involving rare items. In future we discuss and propose a method to generate conditional hybrid dimension association rules using fuzzy logic, whereas hybrid dimension association rule is hybridization between inter-dimension and intra-dimension scaled association rules.

REFERENCES

JOURNAL AND OTHER ARTICLES

1. A.Meenakshi, K.Alagarsamy (2010), “A Novelty Approach for Finding Frequent Itemsets in Horizontal and Vertical Layout- HVCFPMINETREE”, International Journal of Computer Applications Volume 10– No.5.
2. Anurag Choubey, Ravindra Patel, J.L. Rana (2011), “A Survey of Efficient Algorithms and New Approach for Fast Discovery of Frequent Itemset for Association Rule Mining (DFIARM)”, International Journal of Soft Computing and Engineering (IJSCE), Volume-1, Issue-2.
3. C.S.Kanimozhi Selvi and A.Tamilarasi,(2009) , “An Automated Association Rule Mining Technique with Cumulative Support Thresholds”, International. Journal of Open Problems in Compt. Math, Vol. 2, No. 3.
4. E. Chandra, K. Nandhini(2010), “Knowledge Mining from Student Data “, European Journal of Scientific Research No-1, pg156-163
5. Fadi Thabtah (2006), “Pruning Techniques in Associative Classification: Survey and Comparison “, Journal of Digital Information Management, Volume 4 Number 3.
6. Fan Lilin (2010), “Research on Classification Mining Method of Frequent Itemset”, Journal of Convergence Information Technology Volume 5, Number 8.
7. Floriana Esposito, Donato Malebra, Giovanni Semeraro, Valentina Tamma(1999),”The effects of pruning methods on the predictive accuracy of induced decision trees.” ,Applied Stochastic. Models Business and Industry, pg 277-299.
8. J.Shahrabi and R. S. Neyestani (2009), “Discovering Iranians’ Shopping Culture by considering Virtual Items Using Data Mining Techniques “, Journal of Applied Sciences Volume-9 Issue-13, pg 2351-2361.
9. Jianying Hu, Aleksandra Mojsilovic (2007), “High-utility pattern mining: method for discovery of high-utility item sets”, Science Direct, The Journal of pattern recognition society, pg 3317 – 3324.
10. Jyothi Pillai, O.P.Vyas (2010), “Overview of Itemset Utility Mining and its applications”, International Journal of Computer Applications, Volume 5- No -11.
11. Kahayan Lal, N.C.Mahanti (2010), “Mining Association Rules in Large Database by Implementing Pipelining Technique in Partition Algorithm”, International Journal of Computer Applications Volume 2 – No.4.
12. M.S.Danessh, C. Balasubramanian and K. Duraiswamy (2010), “Similarity Data Item Set Approach: An Encoded Temporal Data Base Technique”, Journal of Computing, Volume 2, Issue 3.
13. Maja Dimitrijevic, Zita Bosnjak (2011), “Web Usage Association Rule Mining System “, Interdisciplinary Journal of Information, Knowledge, and Management Volume 6.

14. M. Mart´inez-Ballesterosa, A. Troncosob, F. Martinez A´ lavers and J. C. Riquelmea(2010), "Mining quantitative association rules based on evolutionary computation and its application to atmospheric pollution", Integrated Computer-Aided Engineering , 227–242.
15. Neelu Khare ,Neeru Adlakha ,K. R. Pardasani(2009), "An Algorithm for Mining Multidimensional Fuzzy Association Rules, International Journal of Computer Science and Information Security,Vol. 5, No. 1.
16. Quinlan J.R.(1987), "Simplifying decision trees", International Journal of Man-Machine Studies; pg 221-234.
17. Pratima Gautam, Neelu Khare K. R. Pardasani (2010), "A model for mining multilevel fuzzy association rule in database ", Journal of computing, Volume 2, Issue 1.
18. Przemyslaw Kazienko (2009), "Mining Indirect Association rules for web recommendation ", International Journal of Applied Mathematical Computing Science, Vol. 19, No. 1, pg165–186.
19. Rakhi Garg, P. K. Mishra (2011), "Exploiting Parallelism in Association Rule Mining", International Journal of Advancements in Technology Vol No- 2.
20. R Uday kiran, P Krishna Reddy (2009), "An Improved Multiple Minimum Support Based Approach to Mine Rare Association Rules", IEEE Symposium on Computational Intelligence and Data Mining.
21. Suraj Srivastava, Deepti Gupta, Harsh K Verma (2010), "Comparative Investigations and Performance Evaluation for Multiple-Level Association Rules Mining Algorithm", International Journal of Computer Applications Volume 4– No.10.
22. Tarek F. Gharib, Hamed Nassar, Mohamed Taha, Ajith Abraham(2010), "An efficient algorithm for incremental mining of temporal association rules", Science Direct Data & Knowledge Engineering, Data & Knowledge Engineering , pg 800-815.
23. Tzung-Pei Hong, Chun-Wei Lin, Yu-Lung Wu(2008) , "Incrementally fast updated frequent pattern trees", Journal of Expert Systems with Applications, pg 2424–2435.

CONFERENCE PAPERS

24. A Savasere, E. Ommcinskl and S Navathe (1995) "An efficient algorithm for mining association rules in large databases", International Conference on Very Large Databases, Zurich, Switzerland, pg 432-444.
25. Agrawal R, Imielinski T, Swami A (1993), "Mining association rules between sets of items in large databases". Buneman P, Jajodia S, Proceedings of the ACM SIGMOD international conference on management of data, Washington DC, pg 207–216.
26. Fang Li, Chengyao Li, and Yangge Tian (2009), "Applying Association Rule Analysis in Bibliometric Analysis", Proceedings of the Second Symposium International Computer Science and Computational Technology China, pg 431-434.
27. J S. Park and M.S. Chen and PS. Yu (1995), "An effective hash-based algorithm for mining association rules", International Conference on Management of Data", pg 175-186.
28. K. P. Bennett (1992), "Decision Tree Construction Via Linear Programming." Proceedings of the 4th Midwest Artificial Intelligence and Cognitive Science Society, pp. 97-101.
29. Laszlo Szathmary, Petko Valtchev, and Amedeo Napoli2 (2010), "Finding Minimal Rare Itemsets and Rare Association Rules", Proceedings of the 4th International Conference on Knowledge Science, Engineering and Management pg 16-27.
30. Niblett T, Bratko I (1986), "Learning decision rules in noisy domains. Proceedings of Expert Systems "Cambridge University.
31. Preetham Kumar, Ananthanarayana V. S (2008), "Discovery of Multi Dimensional Quantitative Closed Association Rules by Attributes Range Method", Proceedings of the International MultiConference of Engineers and Computer Scientists Vol-I,pg 19-21, Hong Kong.
32. R. Agrawal and R Smkant (1994), "Fast algorithms for mining association rules", In Proceedings of the 20th International Conference on Very Large Databases Chile, pg 487-499.
33. S. Lofti, M.H. Sadreddini (2009), "Mining Fuzzy Association Rules Using Mutual Information", Proceedings of the International MultiConference of Engineers and Computer Scientists, pg 18 – 20, Hong Kong.
34. UNPUBLISHED DISSERTATIONS AND THESES
35. Michael Hahsler and Bettina Grun and Kurt Hornik and Christian Buchta (2009), "Introduction to a rules: A computational environment for mining association rules and frequent item sets", PhD thesis.

WEBSITE

36. http://www.scribd.com/doc/23918301/Association-Rule-Pruning-based-on-Interestingness-Measures-with-Clustering.
37. http://edu.eap.gr/pli/pli10/info/CVs/Kotsiantis_CV.htm
38. http://tr.ietejournals.org/article.asp.
39. http://perweb.firat.edu.tr/default.asp

REQUEST FOR FEEDBACK

Dear Readers

At the very outset, International Journal of Research in Computer Application and Management (IJRCM) acknowledges & appreciates your efforts in showing interest in our present issue under your kind perusal.

I would like to request you to supply your critical comments and suggestions about the material published in this issue as well as on the journal as a whole, on our E-mails i.e. **infoijrcm@gmail.com** or **info@ijrcm.org.in** for further improvements in the interest of research.

If you have any queries please feel free to contact us on our E-mail **infoijrcm@gmail.com**.

I am sure that your feedback and deliberations would make future issues better – a result of our joint effort.

Looking forward an appropriate consideration.

With sincere regards

Thanking you profoundly

Academically yours

Sd/-

Co-ordinator

ABOUT THE JOURNAL

In this age of Commerce, Economics, Computer, I.T. & Management and cut throat competition, a group of intellectuals felt the need to have some platform, where young and budding managers and academicians could express their views and discuss the problems among their peers. This journal was conceived with this noble intention in view. This journal has been introduced to give an opportunity for expressing refined and innovative ideas in this field. It is our humble endeavour to provide a springboard to the upcoming specialists and give a chance to know about the latest in the sphere of research and knowledge. We have taken a small step and we hope that with the active co-operation of like-minded scholars, we shall be able to serve the society with our humble efforts.

Our Other Journals