Data Mining Algorithms And Medical Sciences

Irshad Ullah

Irshadullah79@gmail.com

Abstract

Extensive amounts of data stored in medical databases require the development of dedicated tools for accessing the data, data analysis, knowledge discovery, and effective use of sloretl knowledge and data. Widespread use of medical information systems and explosive enlargement of medical databases require conventional manual data analysis to be coupled with methods for competent computer-assisted analysis. In this paper, I use Data Mining techniques for the data analysis, data accessing and knowledge discovery procedure to show experimentally and practically that how consistent, able and fast are these techniques for the study in the particular field? A solid mathematical threshold (0 to 1) is set to analyze the data. The obtained outcome will be tested by applying the approach to the databases, data warehouses and any data storage of different sizes with different entry values. The results shaped will be of different level from short to the largest sets of tuple. By this, we may take the results formed for different use e.g. Patient investigation, frequency of different disease.

Key Words

Knowledge Discovery, Medical Database, Association rule mining Techniques, Analysis, Transformation.

1 Introduction: Data Mining.

Data analysis is a process in which raw data is prepared and structured so that valuable information can be extracted from it. The process of organizing and thinking about data is way to accepting what the data does and does not contain. There are a variety of ways in which public can approach data analysis, and it is notoriously easy to direct data during the analysis phase to push certain conclusions or agendas [12].Analysis of data is a process of inspecting, cleaning, transforming, and modeling data with the objective of highlighting useful information, suggesting conclusions, and supporting decision making. Data analysis has multiple facets and approaches, encompassing diverse techniques under an array of names, in different business, science, and social science domains [9]

Data Mining is the discovery of unknown information found in databases [15] [20]. Data mining functions include clustering, classification, prediction, and associations. One of the most important data mining applications is that of mining association rules. Association rules, first introduced in 1993 [18], are used to identify relationships among a set of items in databases. These relationships are not based on inherent properties of the data themselves, but rather based on co-occurrence of the data items. Emphasis in this paper is on the basket market analysis data. Various algorithms have been proposed to discover frequent itemsets in transaction databases. Data mining presents new perspectives for data analysis. The purpose of data mining is to extract and discover new knowledge from data. Over the past few decades, new methods have been developed about the capabilities of data collection and data generation. Data collection tools have provided us with a huge amount of data. Data mining processes have integrated techniques from multiple disciplines such as, statistics, machine learning, database technology, pattern recognition, neural networks, information retrieval and spatial data analysis. Data mining techniques have been used in many different fields such as, business management, science, engineering, banking, data management, administration, and many other applications.

DOI : 10.5121/ijcsit.2010.2612

Data mining is a repetitive process consisting of several steps. Starting with the understanding and definition of a problem and ending with the analysis of results and determine a strategy with using the result [13].

2 Related work

The AIS algorithm is the first published algorithm developed to produce all large itemsets in a transaction database [18]. This algorithm has targeted to discover qualitative rules. This technique is limited to only one item in the consequent. This algorithm makes multiple passes over the entire database. The SETM algorithm is proposed in [14] and motivated by the desire to use SQL to calculate large itemsets [19]. In this algorithm each member of the set large itemsets, Lk, is in the form <TID, itemset> where TID is the unique identifier of a transaction. Similarly, each member of the set of candidate itemsets, Ck, is in the form <TID, itemset>. Similar to [18], the SETM algorithm makes multiple passes over the database.

The Apriori algorithm [17] is a great success in the history of mining association rules. It is by far the most well-known association rule algorithm. This technique uses the property that any subset of a large itemset must be a large itemset. The Off-line Candidate Determination (OCD) technique is proposed in [8], and is based on the idea that small samples are usually quite good for finding large itemsets. The OCD technique uses the results of the combinatorial analysis of the information obtained from previous passes to eliminate unnecessary candidate sets.

Sampling [7] reduces the number of database scans to one in the best case and two in the worst. A sample which can fit in the main memory is first drawn from the database. The set of large itemsets in the sample is then found from this sample by using a level-wise algorithm such as Apriori. Each association rule mining algorithm assumes that the transactions are stored in some basic structure, usually a flat file or a TID list, whereas actual data stored in transaction databases is not in this form. All approaches are based on first finding the large itemsets. The Apriori algorithm appears to be the nucleus of all the association rule mining algorithms. In this work my focus is on association rule mining technique. I take two algorithms, first the well known Apriori and then our own developed SI [11] algorithm.

3 Problem of association rules

A formal statement of the association rule problem is as follows:
Definition 1**:** [18] [4] Let I = $\{i_1, i_{2,}…., i_m\}$ be a set of m distinct attributes. Let D be a database, where each record (tuple) T has a unique identifier, and contains a set of items such that $T \subseteq I$. An association rule is an implication of the form of $X \Rightarrow Y$, where X, $Y \subseteq I$ are sets of items called itemsets, and $X \cap Y = \phi$. Here, X is called antecedent while Y is called consequent; the rule means $X \Rightarrow Y$. Two important measures for association rules, support (s) and confidence (α), can be defined as follows. The support (s) of an association rule is the ratio (in percent) of the records that contain $X \cup Y$ to the total number of records in the database. For a given number of records, confidence (α) is the ratio (in percent) of the number of records that contain $X \cup Y$ to the number of records that contain X.

Association rules can be classified based on the type of vales, dimensions of data, and levels of abstractions involved in the rule. If a rule concerns associations between the presence or absence of items, it is called Boolean association rule. And the dataset consisting of attributes which can assume only binary (0-absent, 1-present) values is called Boolean database.

4 Logical analysis of data

The logical analysis of data was originally developed for the analysis of datasets whose attributes take only binary (0-1) values [2, 3, 6].Since it turned out later that most of the real-life applications include attributes taking real values, a “binarization" method was proposed in [1]. The purpose of binarization is the transformation of a database of any type into a “Boolean database”.

Table 1. Original Database

I	Age 20….2	Age 30…3	M.Statu s	M.Statu s	. .
1	1	0	1	0	..
2	0	1	0	1	..

LAD is a methodology developed since the late eighties, aimed at discovering hidden structural information in Boolean databases. LAD was originally developed for analyzing binary data by using the theory of partially defined Boolean functions. An extension of LAD for the analysis of numerical data sets is achieved through the process of “binarization” consisting in the replacement of each numerical variable by binary “indicator” variables, each showing whether the value of the original variable is present or absent, or is above or below a certain level. LAD has been applied to numerous disciplines, e.g. economics and business, seismology, oil exploration, medicine etc. [16].

Logical Analysis of Data (LAD) is one of the techniques used in data analysis. Unlike other techniques, which involve probabilistic and geometric analysis, LAD uses logical rules to analyze observations. Its main purpose is to detect hidden patterns in the data set that distinguish observations of one class from the rest of the observations.

4.1 Binarization

The methodology of LAD is extended to the case of numerical data by a process called binarization, consisting in the transformation of numerical (real valued) data to binary (0, 1) ones. In this [5] transformation we map each observation u = (uA, uB,…) of the given numerical data set to a binary vector x(u) = (x1, x2,…) Є {0, 1}n by defining e.g. x1 = 1 iif uA ≥ α1, x2 = 1 iif uB ≥ α2, etc, and in such a way that if u and v represent, respectively, a positive and negative observation point, then x(u) ≠ x(v). The binary variables xi, i = 1,2, …, n associated to the real attributes are called indicator variables, and the real parameters αi, i = 1, 2, …, n used in the above process are called cut points.The basic idea of binarization is very simple. It consists in the introduction of several binary attributes associated to each of the numerical attributes; each of these binary attributes is supposed to take the value 1 (respectively, 0) if the numerical attribute to which it is associated takes values above (respectively, below) a certain threshold. Obviously the computational problem associated to binarization is to find a minimum number of such threshold values (cutpoints) which preserve the essential information contained in the dataset, i.e. the disjointness of the sets of (binarized) positive and negative observations.In order to illustrate the binarization of business datasets, let us consider the examples presented in Table 2. A very simple binarization procedure is used for each variable “age” and “marital status”. Quantitative attributes such as “age” is divided into different ranges like age: 20..29, 30..39, ect. The “marital status” variable is divided into binary values by converting its domain values into attributes.

Table 2. Boolean Database

ID	Age	M.Status	#cars
1	23	Single	0
2	31	Married	2

4.2 Binary Variables

A binary variable has only two states: 0 or 1, where 0 means that the variable is absent, and 1 means that it is present. If all binary variables are thought of as having the same weight, we have the 2-by-2 contingency table where a is the number of variables that equal 1 for both items i and j, b is the number of variables that equal 1 for item i but that are 0 for item j, c is the number of variables that equal 0 for item i but equal 1 for item j, and d is the number of variables that equal 0 for both item i and j. The total number of variables is z, where z = a + b + c + d.

Table 3.A contingency table for binary variables

	Item j			
Item i		**1**	**0**	**Sum**
	1	A	B	A + b
	0	C	D	C+ d
	Sum	A + c	b+ d	Z

For noninvariant similarities, the most well-known coefficient is the Jaccard dissimilarity coefficient, where the number of negative matches d is considered unimportant and thus is ignored in the computation: $d(I,J)=\dfrac{b+c}{a+b+c}$

The measurement value 1 suggests that the objects i and j are dissimilar and the measurement value 0 suggests that the objects are similar. This method is used in SI algorithm while the Apriori algorithm works using similarity measures.

4.3 Practical implementation.

Table 4. Medical data in Boolean format

	$\mathbf{I_1}$	$\mathbf{I_2}$	$\mathbf{I_3}$	$\mathbf{I_4}$	$\mathbf{I_5}$
$\mathbf{T_1}$	1	1	0	0	1
$\mathbf{T_2}$	0	1	0	1	0
$\mathbf{T_3}$	0	1	1	0	0
$\mathbf{T_4}$	1	1	0	1	0
$\mathbf{T_5}$	1	0	1	0	0
$\mathbf{T_6}$	0	1	1	0	0
$\mathbf{T_7}$	1	0	1	0	0
$\mathbf{T_8}$	1	1	1	0	1
$\mathbf{T_9}$	1	1	1	0	0

First column of the above Table indicates the transactions from 1 to 9 and subsequent five columns indicate the disease is present. Zero (0) means absence of the disease and one (1) means it is present. In order to use Jacquard’s coefficient to find frequent itemsets we use K maps to arrange *a*, *b* and *c*. To find whether I_1, I_2 are frequent items we arrange K-Map for a, b and c as shown in the table below

Table 5: K-Map for two items

	I1’	I1
I2’	0	2
I2	3	4

Where d (I1, I2) = (2+3/2+3+4) = 0.55. If d < ϕ then I1, I2 are declared frequent itemsets. Similarly for other itemsets of size greater then 2, K-Maps of different sizes are constructed and their distances are computed respectively according to the technique presented in algorithm given in

4.4 SI Algorithm

Two algorithms are used for the implementation purpose. One which we develop SI algorithm and the well known Apriori algorithm to check the accuracy and efficiency.

Input

Φ User specified threshold between 0 And 1

T Binary transactional Database

Output

Frequent itemsets

Step

p = { i_1,i_2,…..in} set of data items in transactional database.

Create K Map for all the permutation in row.

Scan the transactional database and put the presence for every combination of data items in corresponding K Map for every permutation of row.

For every permutation of p:

a) Calculate dissimilarity using K Map Constructed for every permutation using the following Jacquard’s dissimilarity equation.

$$d\,(i1,i2,\ldots..in) = \sum_{i1=0}^{1}\sum_{i2=0}^{1}\sum_{i3=0}^{1},\ldots\ldots\ldots.\sum_{in=0}^{1} f(i_1,i_2,\ldots..in)f(0,0,\ldots..0) -$$

$$f(1,1,\ldots\ldots1)/\,f\sum_{i1=0}^{1}\sum_{i2=0}^{1}\sum_{i3=0}^{1},\ldots\ldots\sum_{in=0}^{1} f(i_1,i_2,\ldots..in) - f(0,0,\ldots..0).$$

b) If d < Φ then i_1,i_2,…..in are frequent .

In step first the data in binary transactional database and a user specified threshold value is provided to the algorithm (input for the algorithm). In second step k-map is generated for all the permutation in row on the base of formula given above. In step three the whole database scan process is performed to put the presence for every combination of data item in corresponding k-map for every permutation of row. In step four the dissimilarity is calculated based on jacquard’s dissimilarity coefficient. Now in the last step the dissimilarity value is compared with the user supplied threshold. If dissimilarity is less than the user specified threshold then it will be added to the frequent item list otherwise the value will be discarded. And finally the algorithm will display the frequent item list.

4.5 Apriori Algorithm

Input

Φ User specified threshold between 0 And 100

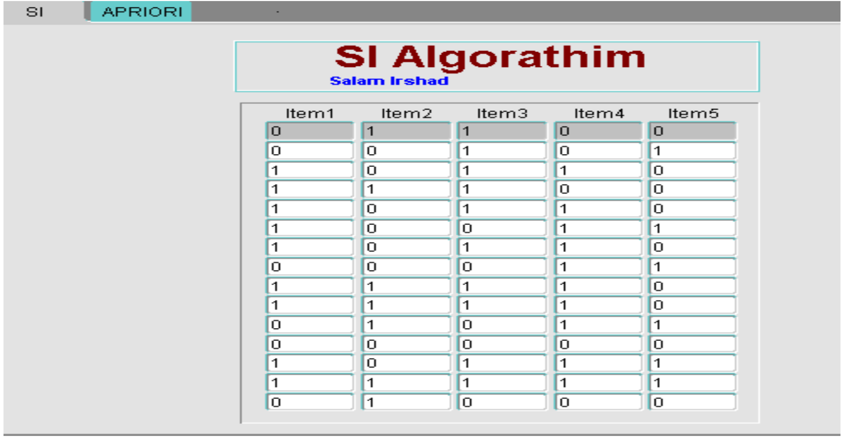
T Binary transactional Database

Output
Frequent itemsets
Apriori algorithm work on similarity measure while the SI
algorithm works on dissimilarity measure.

5 Results.

Different experiments are performed to check the results and efficiency of the technique. The data required in database should be in binary format. I downloaded the dataset transa from the net [10]. We may use Binarization technique to transform the data of any format to binary format. The data was stored in a format like:
0 1 0 0 1 0 0 1
1 1 0 0 1 1 0 0
I coded the algorithms in ORACLE 10g using laptop computer having 20GB hard drive and 1.6MH processor. I create a table in the database to store the data for the purpose of experiment. To load the data to the database the oracle provide a facility by making a control file and then by using SQL loader. We first convert the data into a format that the item now is separated by commas instead of spaces. Now the data is loaded to the table with the help of SQL loader and became as:
0,1,0,0,1,0,0,1
1,1,0,0,1,1,0,0
After loading the data into table the algorithms are implemented on the database having hundred records, initially.

Figure 1

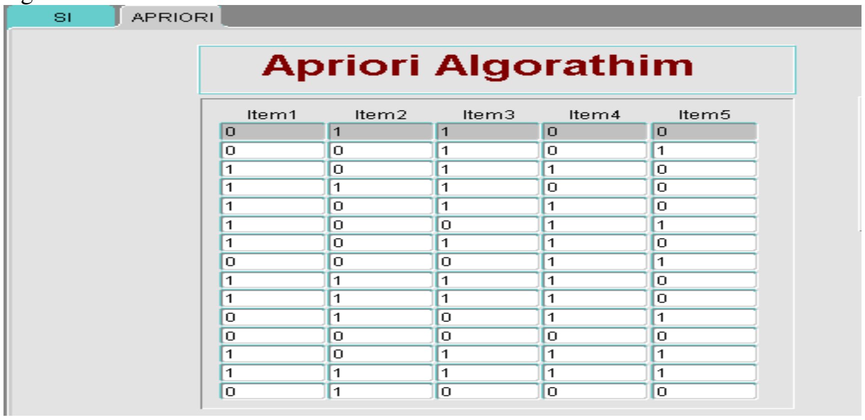


SI Algorathim

Salam Irshad

Item1	Item2	Item3	Item4	Item5
0	1	1	0	0
0	0	1	0	1
1	0	1	1	0
1	1	1	0	0
1	0	1	1	0
1	0	0	1	1
1	0	1	1	0
0	0	0	1	1
1	1	1	1	0
1	1	1	1	0
0	1	0	1	1
0	0	0	0	0
1	0	1	1	1
1	1	1	1	1
0	1	0	0	0

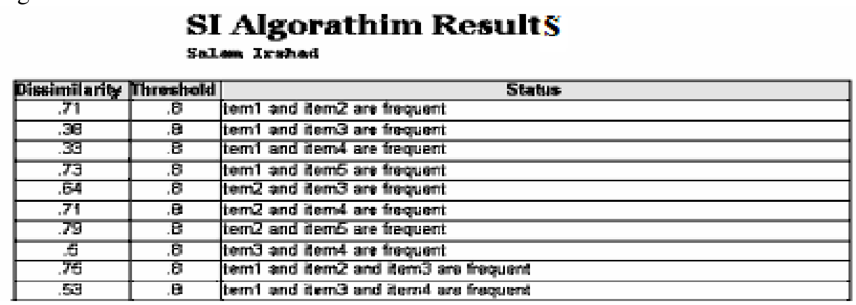
Figure 2



Apriori Algorathim

Item1	Item2	Item3	Item4	Item5
0	1	1	0	0
0	0	1	0	1
1	0	1	1	0
1	1	1	0	0
1	0	1	1	0
1	0	0	1	1
1	0	1	1	0
0	0	0	1	1
1	1	1	1	0
1	1	1	1	0
0	1	0	1	1
0	0	0	0	0
1	0	1	1	1
1	1	1	1	1
0	1	0	0	0

Figure 3

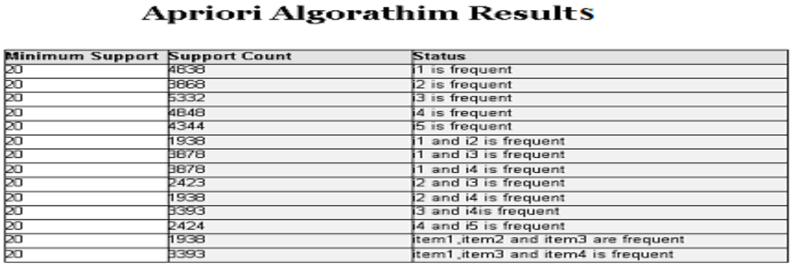


SI Algorathim ResultS

Salem Irshad

Dissimilarity	Threshold	Status
.71	.8	tem1 and item2 are frequent
.38	.8	tem1 and item3 are frequent
.33	.8	tem1 and item4 are frequent
.73	.8	tem1 and item5 are frequent
.64	.8	tem2 and item3 are frequent
.71	.8	tem2 and item4 are frequent
.79	.8	tem2 and item5 are frequent
.5	.8	tem3 and item4 are frequent
.75	.8	tem1 and item2 and item3 are frequent
.53	.8	tem1 and item3 and item4 are frequent

Figure 4



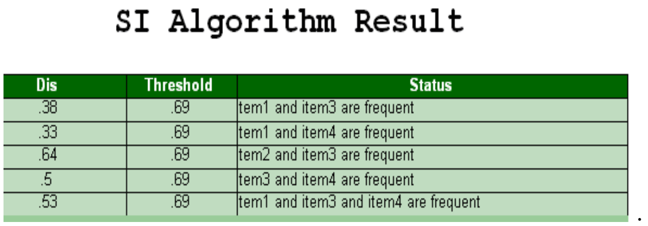
Apriori Algorathim ResultS

Minimum Support	Support Count	Status
20	4838	i1 is frequent
20	3868	i2 is frequent
20	5332	i3 is frequent
20	4848	i4 is frequent
20	4344	i5 is frequent
20	1938	i1 and i2 is frequent
20	3878	i1 and i3 is frequent
20	3878	i1 and i4 is frequent
20	2423	i2 and i3 is frequent
20	1938	i2 and i4 is frequent
20	3393	i3 and i4is frequent
20	2424	i4 and i5 is frequent
20	1938	item1 ,item2 and item3 are frequent
20	3393	item1 ,item3 and item4 is frequent

The largest frequent list generated by the algorithm is
I_1, $I_{2,}$ I_3 „ I_1, I_3, I_4
After giving the data to Apriori algorithm it also produced the same results with the same largest frequent sets contain,
I_1, $I_{2,}$ I_3
I_1, I_3, I_4
After loading more data, the total records in the database are 1000. Applying Apriori and SI algorithms on the updated database, the results produced are given.
The largest frequent list
I_1, $I_{2,}$ I_3
I_1, I_3, I_4
We see that again the algorithms produce the same results. After loading more data the total number of records became 1500. And again applying the algorithms on the database, the results produced are given below.
I_1, $I_{2,}$ I_3
I_1, I_3, I_4
So again the algorithms produced the same results. After loading more data to the database the total records in the database are 2000. Again applying Apriori and SI algorithm on the database the results produced are given.
I_1, $I_{2,}$ I_3
I_1, I_3, I_4
We see that again the algorithms produced the same results. Now we have to load more data to the database the total number of records become 4000. And once again applying both the algorithm on the database the results produced are given below.
I_1, $I_{2,}$ I_3
I_1, I_3, I_4
We see that again the algorithm produced the same results.
Up to this we analyzed the performance, efficiency and accuracy of the algorithms by changing size of the database. And this is clear that the results produced from this medical database are

very consistent and reliable for the diagnoses of patients and different diseases. A disease may be identified by providing its symptoms to the algorithms. After this we change the input threshold to analyze the performance, efficiency and accuracy at different threshold values. The input threshold changes from .80% to .70% (dissimilarity) for SI algorithm and from 20% to 30% (similarity) for Apriori algorithm. The database contains 4000 records and after applying both the algorithms the results produces are given below.

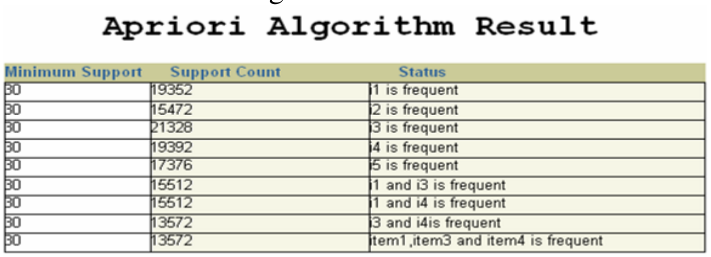
Figure 5



SI Algorithm Result

Dis	Threshold	Status
.38	.69	tem1 and item3 are frequent
.33	.69	tem1 and item4 are frequent
.64	.69	tem2 and item3 are frequent
.5	.69	tem3 and item4 are frequent
.53	.69	tem1 and item3 and item4 are frequent

Figure 6



Apriori Algorithm Result

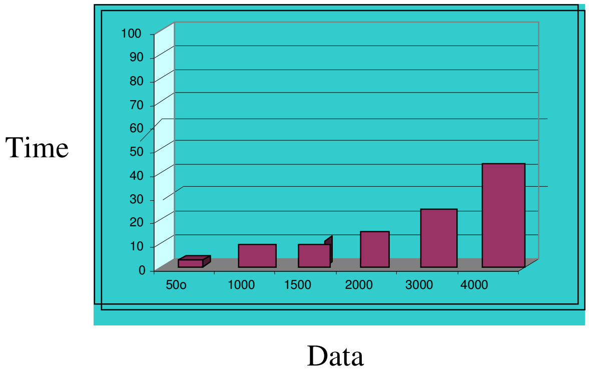
Minimum Support	Support Count	Status
30	19352	i1 is frequent
30	15472	i2 is frequent
30	21328	i3 is frequent
30	19392	i4 is frequent
30	17376	i5 is frequent
30	15512	i1 and i3 is frequent
30	15512	i1 and i4 is frequent
30	13572	i3 and i4is frequent
30	13572	item1,item3 and item4 is frequent

The largest frequent list produced is I_1, I_3, I_4

Now this is clear that both algorithms produced the same results at different threshold. Up to this, we analyze that these techniques are very reliable for the analysis and discovery of hidden pattern and information in any type of database, just like in this medical database and any medical database regarding patients, Staff, Books, and Students may be analyzed very clearly.

6 Results analysis through Graph

Figure 7: Graph for the results produced at different database size by algorithms



7 Conclusion and future work

In this research, I study that how data mining techniques are used for the data analysis and Knowledge discovery in medical sciences. The output produced was based on realistic reasons and values so it is reliable, efficient and precise for the experts. Here we produced the results by performing different experiments and proved that such techniques are very consistent. By this patients may be categorized for the treatment purposes. Results may be used for the analysis of medical staff to improve the performance. Students and subjects may be grouped to make the work more consistent. Further we may perform experiments for other algorithms from different point of views on different data storage. Experiments can be performed for clustering measures.

8 References.

[1] Boros E., P.L. Hammer, T. Ibaraki, A. Kogan.(1997). Logical Analysis of Numerical Data. Mathematical Programming, 79:163-190.

[2] Boros E., P.L. Hammer, T. Ibaraki, A. Kogan, E. Mayoraz, I. Muchnik.(2000). An Implementation of Logical Analysis of Data. IEEE Transactions on knowledge and Data Engineering, 12(2):292-306.

[3] Crama Y., P.L. Hammer, T. Ibaraki. (1988). Cause-effect Relationships and Partially Defined Boolean Functions. Annals of Operations Research, 16:299-325.

[4] David Wai-Lok Cheung, Vincent T. Ng, Ada Wai-Chee Fu, and Yongjian Fu. (December 1996). Efficient Mining of Association Rules in Distributed Databases, IEEE Transactions on Knowledge and Data Engineering, Vol. 8, No. 6, pp. 911-922.

[5] E. Boros, P. L. Hammer, T. Ibaraki, A. Kogan, E. Mayoraz and I. Muchnik(December 1996) An implementation of logica analysis of data, RUTCOR Research Report RRR 22-96, Rutgers University, 1996., pp. 911-922.

[6] Hammer P.L.(1986). The Logic of Cause-effect Relationships, Lecture at the International Conference on Multi-Attribute Decision Making via Operations Research-based Expert Systems, Passau, Germany.

[7] Hannu Toivonen (1996). Sampling Large Databases for Association Rules, Proceedings of the 22nd International Conference on Very Large Databases, pp. 134-145, Mumbai, India.

[8] Heikki Mannila, Hannu Toivonen, and A. Inkeri Verkamo (July 1994). Efficient Algorithms for Discovering Association Rules, Proceedings of the AAAI Workshop on Knowledge Discovery in Databases (KDD-94), pp. 181-192.

[9] H.J.Adèr,. (2008). Chapter 14: Phases and initial steps in data analysis. In H.J. Adèr & G.J. Mellenbergh (Eds.) (with contributions by D.J. Hand), Advising on Research Methods: A consultant's companion (pp. 333-356). Huizen, the Netherlands: Johannes van Kessel Publishing.

[10] http://www2.cs.uregina.ca/~dbd/cs831/notes/itemsets/item set prog1.htm

[11] Irshad Ullah, Abdus Salam and Saif-ur-Rehman (2008),Dissimilarity Based Mining for Finding Frequent itemsets. Proceedings of 4th international conference on statistical sciences Volume (15), University of Gujrat Pakistan, 15: 78

[12] Irshad Ullah (2010). Data Analysis by Data Mining Algorithms A Practical Apprach. International Conference on Word Statistics Day. Superior University Lahore

[13] Jiawe Han & Micheline Kamber, Data mining concepts and techniques San Franciso Moraga Kaufman 2001.

[14] M. Houtsmal and A. Swami (1995). Set-Oriented Mining for Association Rules in Relational Databases,Proceedings of the 11th IEEE International Conference on Data Engineering, pp. 25-34, Taipei,Taiwan.

[15] Ming-Syan Chen, Jiawei Han and Philip S. Yu.(1996). Data Mining: An Overview from a Database Perspective, IEEE ransactions on Knowledge and Data Engineering, Vol. 8, No. 6, pp. 866-883

[16] Peter L. Hammer Tiberius Bonates.(2005). Logical Analysis of Data: From Combinatorial Optimization to Medical Applications, RUTCOR Research Report RRR 10 - 2005.

[17] Rakesh Agrawal and Ramakrishnan Srikant.(1994). Fast Algorithms for Mining Association Rules in Large Databases, Proceedings of the Twentieth International Conference on Very Large Databases, pp. 487-499, Santiago, Chile.

[18] R. Agrawal, T. Imielinski, A. Swami.(1993). Mining Associations between Sets of Items in Massive Databases, Proc. of the ACM-SIGMOD Int'l Conference on Management of Data,Washington D.C.

[19] Ramakrishnan Srikant (1996). Fast Algorithms for Mining Association Rules and Sequential Patterns,Ph. D. Dissertation,University of Wisconsin, Madison.

[20] Usama M. Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth.(1996). From Data Mining to knowledge Discovery: An Overview, Advances in Knowledge Discovery and Data Mining, AAAI Press, , pp 1-34.

Authors:

Irshad Ullah MS-CS (Database Systems)

City university of Science and information Technology Peshawar.

B.Ed

University of Peshawar Pakistan.

Lecturer in Computer Science at Tameer-e-Seerat College Peshawar:2003 to 2004

Subject Specialist (Sr.IT Teacher) at GHSS Ouch since 2004.