International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 20 (2017) pp. 9451-9458

Applicability of Apriori Based Association Rules on Medical Data

Identification of Associations on Medical Data/ Heart disease Dataset using Apriori Based Algorithm

P.Sambasiva Rao[1] and Dr. T.Uma Devi[2]

Ph.D. Scholar[1], Associate Professor[2]
[1,2] Gandhi Institute of Technology and Management (GITAM), University, Visakhapatnam-530 045, Andhra Pradesh, India.

[1]Orcid: 0000-0003-1277-6570

Abstract
With the growing demand of medical data processing and predictive analysis, finding the frequent itemset and association rules are becoming the major focus of the research. The association rule based analysis helps the researchers and analysts to identify the relations and dependencies between various parameters in the dataset. This knowledge can be useful in identifying the cause of the disease for the patient. Voluminous amount of researches are been conducted in order to establish the best association rule mining algorithm similar to Apriori algorithm. Nevertheless, the improvements can be demonstrated. Hence, this work proposes an Apriori association rule discovery based technique and demonstrate the improvements over the existing research methods. Another significant outcome of this work is to establish the relationships between the healthcare parameters with the heart disease symptoms. The work is targeted to improve the precaution measures for the patients in order to save the precious human life.

Keywords: Association Rule Mining, Apriori, medical attributes association, rule discovery, group based rule discovery

INTRODUCTION

The data mining techniques are always highly appreciated by the researchers for extracting information and knowledge from large relational and non-relational datasets. The dataset in general contains huge amount of information and need to be analysed and summarized correctly in order to gain useful information and knowledge. The gained knowledge and information in some cases can be deployed to make significant decisions related to government policies, financial planning, and healthcare related decisions etc. Many researchers are demonstrated the use of various tools and algorithms to find the pattern and dependencies among the parameters of the datasets. These techniques were used to determine the trend in the chemical, financial, pharmaceutical, insurance [1] [2] [3] [4] industries.

The major driving force for the use of data mining techniques into countless domains are the reduced cost of storage and the significant advantages to the business by the analysis and prediction.

Y. Hamuro et al. have demonstrated the improvement in profit [3] for the medical storages located in Japan. The study demonstrated that the increment in sales of pain relievers by 50%. Strong associations such as "People with children tend to buy life insurance policies more often than others" and "Owners of sports utility vehicles are more likely to have wireless phones" can be extremely useful for target marketing [5].

Another popular trend of using data mining techniques is in supermarket or retail chains [6] [7] for improving the sales. R. Agrawal et al. have demonstrated that, finding the relationships between items purchased by the customer can improve the sales [8]. Also similar type of study proves that the arrangements of frequently bought together items can be made next to each other's in order to make it simple for the customers.

The work by C. Bettini et al. [9] and D. J. Cook et al. [10] have illustrated that the use of data mining techniques for information and knowledge extractions from various sources like time series or similar large commercial data sources can be useful.

Healthcare organizations are capable of collecting huge amount of patient information. Alongside with the healthcare organization, other organizations like insurance companies also store the patient information. Thus the gigantic amount of data requires improved algorithms and techniques to extract knowledge to make decisions [11]. The use of computational systems is proven to be useful by the recent advancements in knowledge discovery. Candelieri A et al. have demonstrated that, the reduction in human limitations for subjectivity and errors for analysing the data can be achieved with the use of data mining techniques [12].

Another research method by Bushinak H. et al. have demonstrated that the use of computing techniques provides extensive management of medical knowledge and secure exchange of knowledge or information between the beneficiaries of the system [13]. It's proven that during large

amount of data processing, the human manual approaches are prone to errors [14].

Thus it is recommended by the recent research advancements to incorporate the data mining algorithm with automation in knowledge processing.

The rest of the work is organized as in the Section II the fundamentals of data mining techniques with the definitions to sustain the Apriori algorithm is established and furnished, in Section III the pre-processing techniques such as attribute information and correlations are been established, in Section IV obtained rule sets are discussed, in Section V the results are demonstrated and in Section VI this work presents the conclusion.

APRIORI ASSOCIATION RULE MINING TECHNIQUES

This section of the work elaborates on the standard process of predictive and descriptive analysis of data using data mining. With the understanding of the overall process, this work demonstrates the use of Apriori algorithm for association rule mining.

The predictive and subjective analysis of data using data mining is a process consisting of four major steps as classification & regression, association rule, cluster analysis, text mining and finally the link analysis [15].

The brief descriptions of the processes are described here:

- **Classification & Regression**: This process is deployed on a dataset in order to extract any model. The model is used to analyse the dataset to find the predictive knowledge from the dataset. The same knowledge can be further used for taking various decisions.
- **Association Rule**: This process is deployed on a dataset to find the relation or association on the parameters or attributes in the dataset rather the actual data. The findings from this process can be deployed to find the correlation between the attributes for understanding the dependencies and can be used to make further decisions.
- **Cluster Analysis**: This process is used to group several data items in the dataset in various groups for the benefit of dataset conversion in terms of size and number of attributes.
- **Text Mining**: This process is deployed on unstructured data unlike the previous methods. The data collected from various sources can be unstructured and making it highly difficult to apply any data mining algorithms. Thus a new process called text mining is been introduced.
- **Link Analysis**: The link analysis method deployed on any type of datasets in order to find the links between various data objects. Firstly, this process enables the clustering technique to separate the datasets into various clusters and then secondly, this process enables the links between the clusters or the data objects. These links clearly demonstrate the relation between the clusters or data objects in order to extract the depended knowledge [16].

This brief understanding arranges the foundation for the details understanding of the proposed enhancement in Apriori algorithm by this work.

For the benefit of the demonstration of the Lemmas and concepts a simple and random table consisting of purchases made in a shopping centre is constructed here [Table – 1]. Each and every row in the table denotes a customer id and the items bought by the customer. Each row will be noted as a transaction [5].

Table I: Transaction Example

Customer ID	Item Purchased Together
1	{B1, B2, C1, J1, M1}
2	{C1, J1, M1}
3	{B3, B2, C1, J1, M1}
4	{B1, C1, J2, J1, M1}
5	{B3, J2, J1, M1}
6	{J2, J1, M1}

The item sets bought together are arbitrary for the sake of this example.

Lemma – 1: Considering a small store selling the items {B1, B2, B3, C1, J1, J2, M1}. There are six customers who bought those items in hypothetical order. Each and every record in the table denotes one transaction.

Where,

S denotes set of items in the store

I, S′ denote a non-empty subset of S called itemset

T denotes the transactions

Support(I) denotes the number of transactions containing I as an item

A denotes the support threshold

Proof: Considering S contains {B1, B2, B3, C1, J1, J2, M1} and T denotes {1, 2, 3, 4, 5, 6} transactions, it is natural to denote:

$$I_1 = \{B1,\ B2,\ C1,\ J1,\ M1\} \quad \text{(Eq. 1)}$$

$$I_2 = \{C1,\ J1,\ M1\} \quad \text{(Eq. 2)}$$

$$I_3 = \{B3,\ B2,\ C1,\ J1,\ M1\} \quad \text{(Eq. 3)}$$

$$I_4 = \{B1,\ C1,\ J2,\ J1,\ M1\} \quad \text{(Eq. 4)}$$

$$I_5 = \{B3,\ J2,\ J1,\ M1\} \quad \text{(Eq. 5)}$$

$$I_6 = \{J2,\ J1,\ M1\} \quad \text{(Eq. 6)}$$

Henceforth in order to find the support of I_1, I_3 and I_4 cannot be obtained as those are the largest item sets.

Neverth less, for the smaller item sets as I_2, I_5 and I_6 the support values can be obtained. For example, the support for I_6 is demonstrated as following:

$$Support\ (I_6) = I_3 \cap I_4 = 2 \quad \text{(Eq. 7)}$$

Hence now the dependencies between J1, J2 and M1 are proven to save bought together.

Lemma – 2: Considering a small store selling the items {B1, B2, B3, C1, J1, J2, M1}. There are six customers who bought those items in hypothetical order. Each and every record in the table denotes one transaction.

Where,

L, R being two disjoints item sets, $L \longrightarrow R$ and will be considered as association rule.

I, S′ denote a non-empty subset of S called itemset

T denotes the transactions

Support (L) and Support (R) denotes the support value for the items sets L, R respectively.

Support (L U R) is the total support, both being two disjoint sets

Confidence (L) denotes the confidence of the itemset L as

$$Confidence\ (L) = Support\ (L \cup R) / Support\ (L)$$

Proof: Considering two non-empty item sets as following:

$$S_1' = \{J1, M1\} \quad \text{(Eq. 8)}$$

$$S_2' = \{C1\} \quad \text{(Eq. 9)}$$

Clearly from Eq. 8 and Eq. 9 the item sets are disjoint and perfect to prove this lemma.

Further from Table - 1,

$$Support\ (S_1') = 6 \quad \text{(Eq. 10)}$$

$$Support\ (S_2') = 4 \quad \text{(Eq. 11)}$$

$$Support\ (S_1' \cup S_2') = 4 \quad \text{(Eq. 12)}$$

Henceforth, the calculation of the confidence for S'_1 and S'_2 are demonstrated here:

$$Confidence(S_1') = \frac{Support(S_1' \cup S_2')}{Support(S_1')} \quad \text{(Eq. 13)}$$

And,

$$Confidence(S_2') = \frac{Support(S_1' \cup S_2')}{Support(S_2')} \quad \text{(Eq. 14)}$$

Clearly, from the Eq. 13 and Eq. 14 the confidences are 4/6 and 1 respectively.

It is natural to understand that, the confidence denotes the likelihood of the itemset to be purchase.

PRE – PROCESSING OF MEDICAL DATA

The analysed data for this work demands reduction in the dimension in terms of number of attributes. Also the reductions of numeric values into nominal by the process of clustering algorithm are also proven to reduce time [16].

This work demonstrates the results and discussions on the widely popular UCI heart disease dataset [17][18][19][20][21].

The reduced attribute set is listed here [Table – II].

Table II: List of Attributes and Description

Attribute	**Item Purchased Together/ Description**
ID	Patient identification number
AGE	Age in years
SEX	Patient Sex, Male or Female
PAINLOC	Chest pain location
CP	Chest pain type • Value 1 typical angina • Value 2 atypical angina • Value 3 non-angina pain • Value 4 asymptomatic
SMOKE	Is or is not a smoker
CIGS	Cigarettes per day
YEARS	Number of years as a smoker
CA	Number of major vessels (0-3) colour by fluoroscopy
LOC	Location of the patient, where the data is been registered • VA • Switzerland • Hungary • Cleveland
NUM	Denotes the severity of the heart disease

In the pre-processing process, the following steps are being incorporated for reducing the dimension of the dataset. The process is explained here:

Step-1. Reduction of the dataset based on the age group.

- Rule – 1: Age range from 0 to 20 => GP1
- Rule – 2: Age range from 21 to 45 => GP2
- Rule – 3: Age range from 46 to 65 => GP3
- Rule – 4: Age range from 66 to Max => GP4

Step-2. Reduction of the dataset based on the number of cigarettes per day.

- Rule – 1: Number of Cigarettes per day is 0=> GP1
- Rule – 2: Number of Cigarettes per day is ranging from 1to 20=> GP2
- Rule – 3: Number of Cigarettes per day is ranging from 21to 40=> GP3
- Rule – 4: Number of Cigarettes per day is ranging from 41to Max => GP4

Step-3. Reduction of the dataset based on the number of years with the smoking habit.

- Rule – 1: Number of years is 0=> GP1
- Rule – 2: Number of years is ranging from 1to 20 => GP2
- Rule – 3: Number of years is ranging from 21to 40 => GP3
- Rule – 4: Number of years is ranging from 41to Max => GP4

Step-4. Reduction of the dataset based on the location.

- Rule – 1: VA=> GP1
- Rule – 2: Switzerland => GP2
- Rule – 3: Hungary => GP3
- Rule – 4: Cleveland => GP4

Henceforth, the results of the reduction are presented as following [Table – 3].

Table III: Reduction Attribute Value Description

Attribute	Groups	Modified values in the Dataset
AGE	GP1	1
	GP2	2
	GP3	3
	GP4	4
CIGS	GP1	0
	GP2	1
	GP3	2
	GP4	3
YEARS	GP1	0
	GP2	1
	GP3	2
	GP4	3
LOC	GP1	1
	GP2	2
	GP3	3
	GP4	4

The reduction pre-processing makes the dataset available for Apriori analysis.

RULE SET ANALYSIS

This work performs multiple analyses on the dataset in order to discover the association rule sets. During the process the following rule sets are been identified for each mentioned item sets:

Itemset 1: AGE, SEX, PAINLOC, CP, SMOKE, CA, LOC, NUM

Rule sets Discovered:

I1-R1. CIGS=0 578 ==> SMOKE=0 578 With Confidence (1)

I1-R2. SMOKE=0 578 ==> CIGS=0 578 With Confidence (1)

I1-R3. YEARS=0 578 ==> SMOKE=0 578 With Confidence (1)

I1-R4. SMOKE=0 578 ==> YEARS=0 578 With Confidence (1)

I1-R5. YEARS=0 578 ==> CIGS=0 578 With Confidence (1)

I1-R6. CIGS=0 578 ==> YEARS=0 578 With Confidence (1)

I1-R7. CIGS=0 YEARS=0 578 ==> SMOKE=0 578 With Confidence (1)

I1-R8. SMOKE=0 YEARS=0 578 ==> CIGS=0 578 With Confidence (1)

I1-R9. SMOKE=0 CIGS=0 578 ==> YEARS=0 578 With Confidence (1)

I1-R10. YEARS=0 578 ==> SMOKE=0 CIGS=0 578 With Confidence (1)

Itemset 2:PAINLOC, CP, NUM

Rule sets Discovered:

I2-R1. CP=4 num=0 104 ==> PAINLOC=1 104 With Confidence (1)

I2-R2. CP=4 num=2 108 ==> PAINLOC=1 107 With Confidence (0.99)

I2-R3. CP=4 num=3 99 ==> PAINLOC=1 98 With Confidence (0.99)

I2-R4. CP=4 486 ==> PAINLOC=1 481 With Confidence (0.99)

I2-R5. CP=4 num=1 140 ==> PAINLOC=1 138 With Confidence (0.99)

I2-R6. CP=3 202 ==> PAINLOC=1 195 With Confidence (0.97)

I2-R7. num=3 132 ==> PAINLOC=1 127 With Confidence (0.96)

I2-R8. CP=3 num=0 130 ==> PAINLOC=1 125 With Confidence (0.96)

I2-R9. num=2 131 ==> PAINLOC=1 125 With Confidence (0.95)

I2-R10. num=1 191 ==> PAINLOC=1 181 With Confidence (0.95)

Itemset 3:AGE, SEX, NUM

Rule sets Discovered:

I3-R1. num=3 132 ==> sex=1 122 With Confidence (0.92)

I3-R2. age=3 num=2 100 ==> sex=1 92 With Confidence (0.92)

I3-R3. num=2 131 ==> sex=1 120 With Confidence (0.92)

I3-R4. age=3 num=3 99 ==> sex=1 90 With Confidence (0.91)

Itemset 4:CA, NUM

Rule sets Discovered:

I4-R1. num=0 404 ==> CA=0 404 With Confidence (1)

I4-R2. CA=0 404 ==> num=0 404 With Confidence (1)

I4-R3. num=1 191 ==> CA=1 191 With Confidence (1)

I4-R4. CA=1 191 ==> num=1 191 With Confidence (1)

I4-R5. num=3 132 ==> CA=3 132 With Confidence (1)

I4-R6. CA=3 132 ==> num=3 132 With Confidence (1)

I4-R7. num=2 131 ==> CA=2 131 With Confidence (1)

I4-R8. CA=2 131 ==> num=2 131 With Confidence (1)

Hence a total of 32 rule sets are been generated. This rule sets defines the association property among the attributes in the used dataset for this work.

RESULTS AND DISCUSSION

In this section, the work demonstrates the confidence, lift, leverage and conviction based metric analysis of the rule sets discovered.

Firstly, the Itemset – 1 based rule sets are analysed [Table – 4].

Table IV: Itemset – 1 Rule set Metric Analysis

Rule Name	Confidence	Lift	Leverage	Conviction
I1-R1	1	1.56	0.23	206.8
I1-R2	1	1.65	0.23	206.8
I1-R3	1	1.56	0.23	206.8
I1-R4	1	1.56	0.23	206.8
I1-R5	1	1.65	0.23	206.8
I1-R6	1	1.56	0.23	206.8
I1-R7	1	1.56	0.23	206.8
I1-R8	1	1.65	0.23	206.8
I1-R9	1	1.56	0.23	206.8
I1-R10	1	1.56	0.23	206.8

The results are been analysed graphically [Fig – 1].

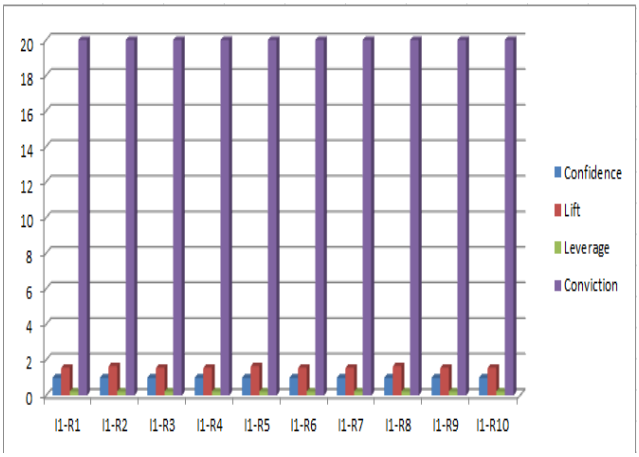

Figure 1: itemset – 1 rule set metric analysis

Secondly, the Itemset – 2 based rule sets are analysed [Table – 5].

Table V: Itemset – 2 Rule set Metric Analysis

Rule Name	Confidence	Lift	Leverage	Conviction
I2-R1	0.99	1.05	0.02	4.41
I2-R2	0.86	1.92	0.08	3.83
I2-R3	0.99	1.04	0.01	2.54
I2-R4	0.76	1.41	0.04	1.98
I2-R5	0.73	1.36	0.04	1.89
I2-R6	0.72	1.35	0.04	1.65
I2-R7	0.97	1.02	0.00	1.37
I2-R8	0.36	1.92	0.08	1.26
I2-R9	0.28	1.1	0.04	1.11
I2-R10	0.29	1.36	0.04	1.10

The results are been analysed graphically [Fig – 2].

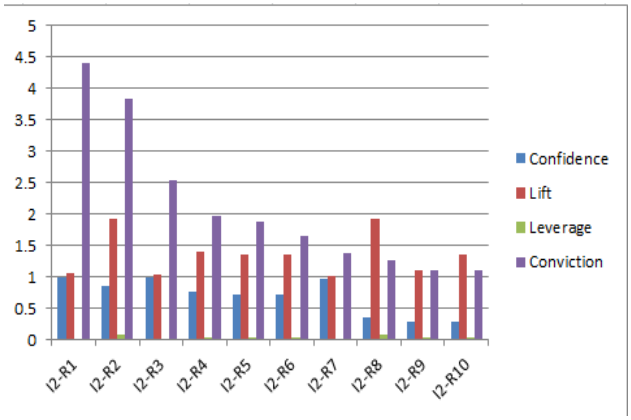

Figure 2: itemset – 2 rule set metric analysis

Thirdly, the Itemset – 3 based rule sets are analysed [Table – 6].

Table VI: Itemset – 3 Rule set Metric Analysis

Rule Name	Confidence	Lift	Leverage	Conviction
I3-R1	0.92	1.17	0.02	2.52
I3-R2	0.92	1.16	0.01	2.33
I3-R3	0.92	1.16	0.02	2.29
I3-R4	0.91	1.15	0.01	2.08

The results are been analysed graphically [Fig – 3].

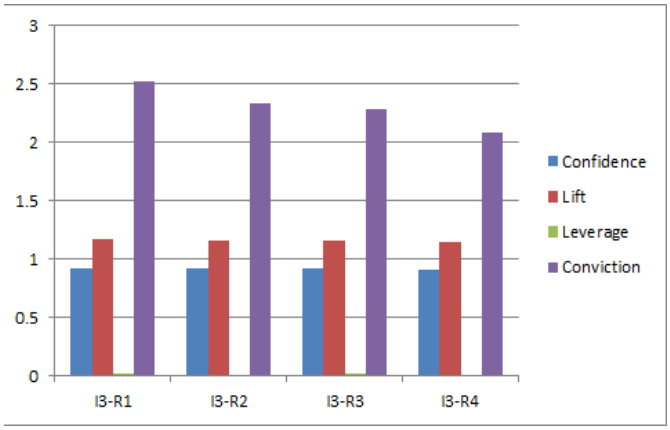

Figure 3: itemset – 3 rule set metric analysis

Finally, the Itemset – 4 based rule sets are analysed [Table – 7].

Table VII: Itemset – 4 Rule set Metric Analysis

Rule Name	Confidence	Lift	Leverage	Conviction
I4-R1	1.00	2.23	0.25	222.65
I4-R2	1.00	2.23	0.25	222.65
I4-R3	1.00	4.71	0.17	150.47
I4-R4	1.00	4.71	0.17	150.47
I4-R5	1.00	6.82	0.13	112.64
I4-R6	1.00	6.82	0.13	112.64
I4-R7	1.00	6.87	0.12	111.93
I4-R8	1.00	6.87	0.12	111.93

The results are been analysed graphically [Fig – 4].

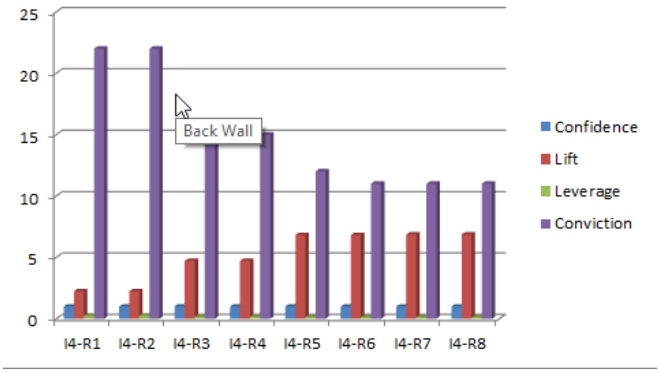

Figure 4: itemset – 4 rule set metric analysis

CONCLUSION

With the unbound growth in the medical diagnosis and use of automation for medical record analysis, a voluminous amount of the data is made available to the researchers. Nevertheless, the data mining techniques are deployed by various researchers in order to analyse and knowledge extraction. This work proposed Apriori based rule set discovery technique for detection of influences of various parameters based on the association rule on the testing datasets. The work is success in demonstrating the relationships between healthcare parameters related to the heart disease symptoms. The work is constructed to improve the precaution measures based on the finding and save the precious human life.

REFERENCES

[1] D. Barbar´a (Editor), Bulletin of the Technical Committee on Data Engineering, Special Issue on Mining of Large Databases, Vol. 21, No. 1, March 1998.

[2] K. M. Decker and S. Focardi, “Technology Overview: A Report on Data Mining”, Technical Report CSCS TR-95-02, Swiss Scientific Computing Center, 1995.

[3] Y. Hamuro, N. Katoh, Y. Matsuda and K. Yada, “Mining Pharmacy Data Helps to Make Profits”, Data Mining and Knowledge Discovery, Vol. 2, 1998, pp. 391–398.

[4] M. Holsheimer and A. P. J. M. Siebes, “Data Mining: The Search for Knowledge in Databases”, Technical Report CS-R9406, CWI, Amsterdam, The Netherlands, 1994.

[5] Doddi, Achla Marathe, SS Ravi, David C. Torney, Srinivas. "Discovery of association rules in medical data." Medical informatics and the Internet in medicine 26.1 (2001): 25-33.

[6] R. Agrawal, T. Imielinski and A. Swami, “Mining Association Rules Between Sets of Items in Large Databases”, Proc. 1993 ACM SIGMOD,Washington, DC, May 1993, pp. 207–216.

[7] H. Toivonen, “Discovery of Frequent Patterns in Large Data Collections”, Ph.D. Thesis, Report A-1996-5, Department of Computer Science, University of Helsinki, Finland, 1996.

[8] R. Agrawal, T. Imielinski and A. Swami, “Mining Association Rules Between Sets of Items in Large Databases”, Proc. 1993 ACM SIGMOD,Washington, DC, May 1993, pp. 207–216.

[9] C. Bettini, X. S. Wang and S. Jajodia, “Mining Temporal Relationships with Multiple Granularities in Time Sequences”, Bulletin of the Technical Committee on Data Engineering, Special Issue on Mining of Large Databases, Vol. 21, No. 1, March 1998, pp. 32–38.

[10] D. J. Cook, L. B. Holder and S. Djoko, “Knowledge Discovery from Structural Data”, CESDIS Technical Report Series No. TR-95-149, Goddard Space Flight Center, Greenbelt, MD, 1995.

[11] Fayyad U., Shapiro G. P., Smyth P. "From Data Mining to Knowledge Discovery in Databases." In American Association for Artificial Intelligence , 37-54. 1996

[12] Candelieri A., Dolce G., Riganello, F., Sannita W. G., "Data Mining in Neurology. In Knowledge- Oriented Applications in Data Mining" pp. 261-276. InTech. 2011.

[13] Bushinak H., AbdelGaber S., AlSharif F. K. "Recognizing The Electronic Medical Record Data From Unstructured Medical Data Using Visual Text

Mining Techniques". IJCSIS International Journal of Computer Science and Information Security, Vol. 9, No. 6 , 25-35. 2011.

[14] Eapen A. G., "Application of Data mining in Medical Applications.", Ontario, Canada, 2004: University of Waterloo. 2004

[15] Weiss, G. M., & Davison, B. D. Data Mining. Handbook of Technology Management, H. Bidgoli Ed., John Wiley and Sons . 2010.

[16] Getoor L., "Link Mining: A New Data Mining Challenge". SIGKDD Explorations Volume 4, Issue 2 . 2003.

[17] P.Sambasiva Rao and Dr. T.Uma Devi, "A Parameter Based Heart Disease Detection Technique using Mining Technique". International Journal of Latest Trends in Engineering and Technology Vol.(7)Issue(3), pp. 077-087. 2016

[18] Zhi-Hua Zhou and Yuan Jiang. NeC4.5: Neural Ensemble Based C4.5. IEEE Trans. Knowl. Data Eng, 16. 2004.

[19] Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J., Sandhu, S., Guppy, K., Lee, S., & Froelicher, V. International application of a new probability algorithm for the diagnosis of coronary artery disease. American Journal of Cardiology, 64,304--310. 1989.

[20] David W. Aha & Dennis Kibler. "Instance-based prediction of heart-disease presence with the Cleveland database." 1989.

[21] Gennari, J.H., Langley, P, & Fisher, D. Models of incremental concept formation. Artificial Intelligence, 40, 11--61. 1989.