Identification of Attributes Common for Various Diseases Using Association Analysis

Jothi Prabha A[1] , A. Govardhan[2]

[1] Associate Professor, Jyothishmathi Institute of Technology and Science, Karimnagar, TS, India
[2]Professor in CSE, EC Member and Director of SIT, JNTU Hyderabad, T.S, India

Abstract

Chronic disorders are the dominant cause of disability and death across the world. Chronic diseases have become a big threat for human lives and millions of people are losing their lives, including many young and middle age group people. The number of people dying due to chronic disorders is double that of many virulent diseases (like HIV/AIDS, Malaria and Tuberculosis), parental conditions, perinatal mortalities and malnutrition combined. 80% of chronic disorder deaths occur in low level and average level income countries and half of the deaths are in women. The medication has become a costly affair due to this many poor people are not getting proper diagnosis of their disease. We can use data mining techniques to cut down the cost of diagnosis of illness by evading many tests by selection of only those attributes which are really significant for diagnosis of a disease. Dimensionality reduction plays a vital role in the medical field as it consists of numerous attributes. In this paper we have examined the approach of feature selection for classification and also exhibited a new technique for the feature selection by using association and correlation measures. The aim of our paper is to identify the correlated features or attributes of medical dataset, using indirect association mechanism, so that patient does not require taking many tests and also in future it can be used for designing a clinical decision support system which aids for decision making during disease diagnosis in an inexpensive way.

I. Introduction

Medical data mining has high prospective for evaluating the hidden patterns in the data sets of the medical discipline. These patterns can be employed for clinical investigation. Nevertheless, the available raw medical data are extensively distributed, diversified and voluminous in nature. These data required to be gathered in a systematized pattern. This gathered data can later be integrated to model a hospital information system. Data mining technology bestow a user oriented technique to novel and also hidden patterns in the data.

Chronic disorders are the dominant cause of disability and death across the world. Chronic diseases [5]took the lives of over 35 million people in 2005, including many young and middle age group people. The number of people dying due to chronic disorders is double that of many virulent diseases (like HIV/AIDS, Malaria and Tuberculosis), parental conditions, perinatal mortalities and malnutrition combined. 80% of chronic disorder deaths occur in low level and average level income countries and half of the deaths are in women. Failing to take action to address the causes, death rate from chronic disorder will increase by 17% between the year 2005 and 2015.

Feature selection has been an intensive research area in statistics, pattern recognition and data mining field. The central idea of feature selection is to extract a subset of input variables by excluding features with negligible or no predictive information [6]. Feature selection can considerably enhance the transparency of the emerging classifier models and mostly construct a model that generalizes better to unnoticed points. Moreover, it is likely the instance that identifying the correct subset of predictive features is a crucial problem in its own right. For instance, physician may take a decision based on the selected features whether a risky surgery is required for treatment or not.

In data mining field, association rule learning is a well-known and intensively researched technique for learning interesting relations between variables in huge databases. Various algorithms for generating association rules were presented over the course of time. Few popular algorithms are Apriori, FP-Growth and DHP. Apriori is the well-known algorithm for mining association rules. In diseases, some of the attributes are directly associated or indirectly associated with other disease. Association rule mining helps to identify the association among the diseases for better diagnosis.

II. Literature Review

A. *Feature Selection*

Feature Selection for Classification in Medical Data Mining [1] discusses about the different attributes related to different diseases. Since the expenses occur on the identification of disease by going for more investigations

In this paper [2] we have conferred few of efficient techniques that can be helpful for breast cancer classification. Amongst the various soft computing approaches and data mining classifiers, Decision tree is identified to be the best predictor having 93.62% accuracy on SEER dataset and also on benchmark dataset (UCI machine learning dataset). In future the predictor can be utilized to model a web based application to accept the predictor variables and automated system Decision Tree based prediction can be employed in distant areas like rural regions or countryside, to mimic like human diagnostic expertise for prediction of syndrome. The Bayesian network is also identified to be a well-known method in medical prognosis specifically it has been well utilized for Breast cancer diagnosis and prognosis. In near future we plan to model and implement such system for web based applications.

The emphasis is on using [3]various algorithms and mixtures of various target attributes for efficient and intelligent heart attack prediction using data mining. For predicting heart attack, significantly 15 attributes are recorded and along with primitive data mining techniques other approaches e.g. Time Series, Clustering and Association Rules, ANN and Soft computing approaches etc. shall also be integrated. The result of predictive data mining technique on the same dataset disclose that Decision Tree exceeds and sometime Bayesian classification is having much the same accuracy as of decision tree but alternative predictive methods like Neural Networks, KNN, Classification based on clustering are not performing better. The second conclusion is that the certainty of the Bayesian Classification and Decision Tree further become better after employing genetic algorithm to minimize the actual data size to get the optimal subset of attribute adequate for heart disease prediction

In this paper we have examined three datasets of different diseases like Breast Cancer, Heart Disease and Diabetes. These datasets are downloaded from standard benchmark UCI repository. Nowadays there is high possibility of diabetes and heart attack due of hyper tension and today's lifestyle and many tests are required for diagnosis of that disease. By utilizing data mining approaches and minimizing the number of attributes or selecting only important characteristics of that disease among all available attributes, accuracy of classifier can be improved.

The performance [4]of our proposed approach is assessed on two medical data sets cancer and diabetes data sets by correlating it with the traditional classification algorithms such as Naive Bayes, j48, GNP and neural networks. Accuracy of Pima data has been improvised using genetic network programming (GNP). It has shown 1% improvement than traditional Naive Bayes classification algorithm and 4.6% progress over breast cancer data. The accuracy of heart disease data using j48 approach is 4% more than Naive Bayes. The accuracy has been improved by using NN.J48 which outperformed Naive Bayes and neural networks for Pima and cancer data. Our proposed approach reached 7.5%improvement over GNP with chi square [8]for Pima Indian diabetes data.

III. Problem Definition

The comprehensive work focus on identifying the key attributes which are related for the cause of chronic diseases. Once the important correlated attributes are able to be identified then it becomes easy to generate good prescription [7] for doctor without giving overdose medicine.

A. Preliminary Concepts

Let A = $\{A_1, A_2, A_m\}$ be a set of m attributes. A subset X $\subseteq$ A is called an attribute set. A k-attribute set is an attribute set that contains k attributes. Let D = $\{D_1, D_2, D_n\}$ be a set of n records, called a medical database, where each record $D_{j,}$ j = 1, 2, n, is a set of attributes such that Dj $\subseteq$ A. Each record is associated with a unique identifier. A record D contains an attribute set X if and only if X $\subseteq$ D. The support of an attribute set X is the percentage of records in D containing X. An attribute set X in a medical database D is known as a frequent attribute set if its support is equal to, or greater than a user-stated minimum support threshold, min sup. Accordingly, an infrequent attribute set is an attribute set that does not fulfill the user-stated minimum support threshold.

An *association rule* is an implication of the form X → Y, where X, Y ⊆ A and X ∩ Y = Φ. Support of the rule X→ Y is a fraction of transactions in the database which contain X and Y attributes. In other terms, support measure can be calculated by using the formula $support(X \cup Y) = P(X \cup Y)$ where P (.) is the probability of (.). Confidence of X→ Y is denoted as conf(X→ Y) and can be calculated by using the formula $conf(X \rightarrow Y) = P(Y/X) = P(X \cup Y) / P(X)$ Rules which have at least min support and min confidence are stated to as strong association rules and such framework is known as the support-confidence framework for mining association rules.

The problem of mining association rules is to identify all association rules having confidence at least minconf where minconf is user-stated parameter. The mining task is of two steps: 1) Find all frequent item sets. 2) Generate rules which satisfies minimum confidence. This step is fairly straightforward. Therefore, the complexity of an algorithm depends on the complexity of step 1 only [16].

Indirect association between pair of attributes has been introduced by P.N.Tan.et.al [16]. It examines its utility in various applications. The phases of this algorithm are (1) finding frequent attribute sets and (2) finding all indirect association rules which are satisfying pair of attribute set support and mediator dependency threshold values.

Definition (Indirect Association):

A pair of attributes X and Y is indirectly associated through a mediator M, if the following conditions are satisfied:

1. sup(X, Y) < t_s (Item set pair Support Condition)

2. There is a non-empty set M such that

(a) $\sup(X \cup M) \geq t_f$; $\sup(Y \cup M) \geq t_f$; (Mediator Support Condition)

(b) $dep(X, M) \geq t_d$, $dep(Y, M) \geq t_d$, where dep (P, Q) is a measure of the dependence between item sets P and Q. (Mediator Dependence Condition)

The thresholds above are called item set pair support threshold (t_s), mediator support threshold (t_f), and mediator dependence threshold (t_d), respectively. In practice, it is rational to set $t_f \geq t_s$

Condition 1 is required because an indirect relationship between two attributes is substantial only if both attributes hardly occur together in the same record. Otherwise, it makes more sense to distinguish the pair in terms of their direct association.

Condition 2(a) can be used to assure that the statistical significance of the mediator set. In specific, for medical data, the support of an attribute set affects the disease diagnosed and justifies the feasibility of a taking decision. Also, support has a nice downward closure property which allows pruning the combinatorial search space of the problem. Condition 2(b) ensures that only attributes that are extremely dependent on the presence of x and y will be used to form the mediator set.

Over the course of years, several measures have been proposed to quantify the degree of dependence between attributes of a dataset. From statistics, the Chi-Square test is often used for this purpose. Though, the downside of this approach is that it does not measure the power of dependency between items [14]. Furthermore, the Chi-Square statistics rely on the number of transactions in the database. As a consequence, other statistical measures of association are often used, including Goodman and Krushkal's λ, Pearson's Φ coefficient, Yule's Q and Y coefficients, etc. [12].

Interest factor is alternative measure that has been used quite widely to quantify the power of dependency among items [9][10][11].

Definition: Given a pair of item sets, say X and Y, its' IS measure can be calculated using the following equation.

$$IS(X,Y) = \frac{P(X,Y)}{P(X)P(Y)} \qquad (1)$$

Where P denotes the probability that the given item set appears in a transaction

B. *Problem Description*

The inputs to this problem are I, a set of attributes related to the disease, and D, a database of patient medical data, attribute pair support threshold (t_s), mediator dependency threshold (t_d), mediator support threshold (t_f) and minimum support (m_s). An indirect association rule is of the form < x, y / M > where x and y are attributes (disease symptoms) and M is a mediator set. The attributes (disease symptoms) x and y is indirectly associated through an item set M. Here x and y are attributes and M is a set of attributes. An indirect relationship between two attributes is significant only if both symptoms hardly occur together (i.e., supp (x, y) < t_s) in the same disease data. Or else, it is more meaningful to distinguish the pair in terms of their direct association. Mediator support threshold (t_f) can be used to ensure that the relationship between the mediator set M and the attribute pair (x, y) is statistically significant. In addition, for disease data, a large support between the mediator and each of the attribute pair would justify the feasibility of promoting the attributes together. Using the mediator dependency threshold (t_d) the degree of the dependency of the attribute pair (x, y) on the attributes that belong to the mediator set M is tested. The problem is to identify all indirect association rules between attributes whose attribute pair support is less than t_s, mediator support is greater than t_f and mediator dependency is greater than t_d.

IV. EXPERIMENTATION

The experiments are performed on Intel Core i5 Processor system using java. The results are generated and analyzed with existing system.

A. *Proposed Approach*

Indirect association is a novel kind of infrequent pattern, which gives a new way to understand the value of infrequent patterns and can efficiently lessen the number of uninteresting infrequent patterns. The notion of indirect association is to indirectly join two rarely co-occurred items through a frequent attributes of disease called mediator, and if effectively utilized indirect association can help to recognize real interesting infrequent attributes for disease from databases. Indirect association is closely related to negative association, as both associations deal with attributes that do not have adequately high support. Indirect associations gives an effective way to discover interesting negative associations by discovering only infrequent attributes pairs that are highly expected to be frequent without using negative information or domain knowledge.

B. *Algorithm*

Through references it is known that a miniature work was done on discovering indirect associations between pair of attributes only. In this section, we adopted a technique [15]which generates direct and indirect associations between pair of attributes as well as attribute sets. This method consists of two algorithms. Algorithm1 identifies set of all frequent attribute sets and set of all Valid Candidates (VC). An attribute set V is said to be Valid candidate if sup (V) ≤ t_s and all subsets of V are frequent. Algorithm 2 finds set of all indirect association rules between pairs of attribute sets.

Algorithm 1: Finding Frequent Attribute sets (P), and Valid Candidates (VC)

Input: DDB- Disease Database, m_s, t_s

Output: P- Positive Frequent attributes, VC- Valid Candidates

Method:

```
Find $P_1$, the set all frequent 1-attributes
for($K=2;P_{k-1} != \Phi$ ; K++)
{       $C_K = P_{K-1} \bowtie P_{K-1}$
    // Pruning infrequent attributes
        for each $c \in C_K$   {
```

```
    if any sub-set of c is not a member of $P_{K-1}$ then $C_K = C_K - \{ c \}$
                }
    // find positive frequent attributes $P_k$ and Valid Candidates (VC) in $C_K$
for each c in $C_K$ {
        if support(c) $\geq m_s$ then $P_k = P_k \cup \{ c \}$
        if support(c) $\leq t_s$ then VC= VC U { c }
            }
    P= P U $P_K$
}
return P,VC
```

Algorithm 2: Mining Indirect Association Rules

```
Input: P, VC, $t_f$, $t_d$, IAR= Ø
Output: Indirect Association Rules
Method:
for each l (= X U Y) ∈ VC {
  for each I ∈ P {
    if ( support(X U I ) $\geq t_f$ && support ( Y U I )$\geq t_f$ )
      if ( dependency( X U I ) $\geq t_d$ && dependency (Y U I) $\geq t_d$ )
        IAR= IAR U (X,Y/I )
  }}
return IAR
```

C. Experimental Results and Performance Evaluation

To assess the performance of proposed algorithm experiments were conducted on two seasonal databases containing 2500 and 4000 transactions. We concentrate on mediator support (t_f) which is a support of attributeset and mediator and mediator dependency (t_d) which is estimated by Eq. (4.1).

Data set consisting of 2500 transactions with mediator support as 0.2, 0.25, 0.3, 0.35; mediator dependence as 0.4, 0.45, 0.5, 0.55 is taken and the total number of rules generated is 205,20,63,31 and 13 respectively. Figure 1 shows the graph showing the mediator support and mediator dependency vs. total number of rules.

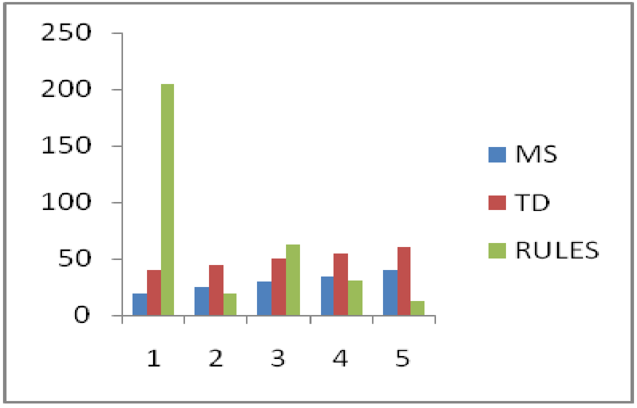

Figure 4-1.Graph showing the mediator support and mediator dependency vs. total number of rules for 2500 records

Figure 2 is generated by considering 4000 transactions with mediator support as 0.2,0.25,0.3,0.35,0.4 mediator dependence as 0.4,0.45,0.5,0.55,0.6 and the total number of rules generated is 35,31,7,7 and 6 respectively

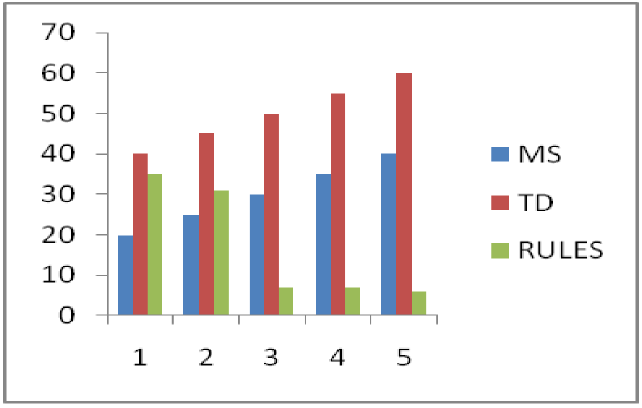

Figure 4-2.Graph showing the mediator support and mediator dependency vs. total number of rules for 4000 records

A seasonal disease data set consisting of 4000 transactions with mediator support as 0.2, 0.25, 0.3, 0.35 and 0.4; mediator dependence as 0.4, 0.45, 0.5, 0.55and 0.6 is taken and the total number of rules generated is 205,20,63,31 and 13 respectively. Figure 4.3 and Figure 4.4 show the graphs for different mediator supports and mediator dependencies vs. total number of rules.

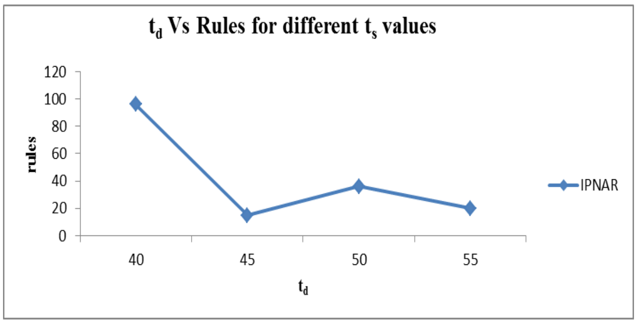

Figure 4-3.Mediator Dependency (t_d) Vs Total Number of Rules

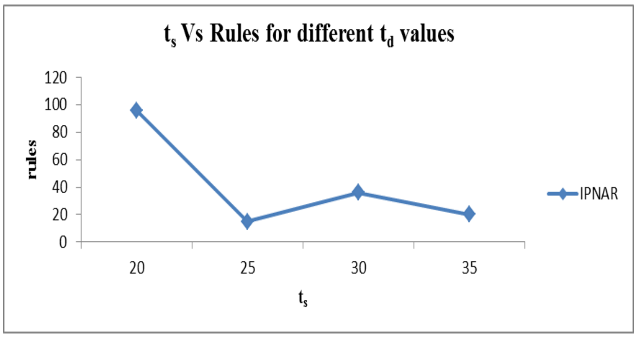

Figure 4-4.Mediator Support (t_s) Vs Total Number of Rules

Regular patient data set consisting of 4000 transactions with mediator support as 0.2,0.25,0.3,0.35and 0.4 mediator dependence as 0.4,0.45,0.5,0.55 and 0.6 and the total number of rules generated is 35,31,7,7 and 6 respectively. Figure 4.5 and Figure 4.6 show the graphs for different mediator supports and mediator dependencies vs. total number of rules.

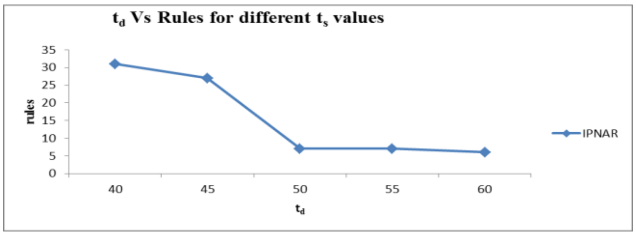

Figure 4-5. Mediator Dependency (t_d) Vs Total Number of Rules

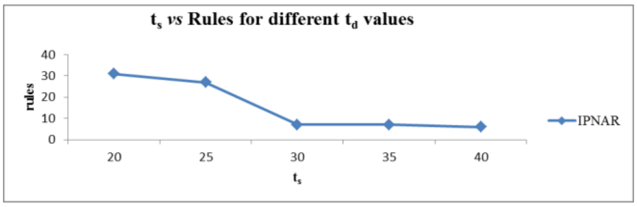

Figure 4-6. F Mediator Support (t_s) Vs Total Number of Rules

D. *Sample Run*

Consider a database consists of 25 records and 16 attributes. For our convenience, we have given numbers to each attribute in the database.

Table 4-1 Item Number and Item Name

Item No	Attribute (Symptom)	Item No	Attribute (Symptom)
1	Fever	9	Diabetic
2	Head ache	10	Eosinophilia
3	Body pains	11	Obesity
4	Fatigue	12	Chest pain
5	Unusual thirst	13	Excess WBC count
6	Cold	14	Less Platelet count
7	Allergy	15	Breathing problem
8	Blood pressure	16	Kidney Pain

Table 4-2.Transactional Database

Tid	Attributes (Symptoms)	Tid	Attributes (Symptoms)
T1	1,2,4,6,16	T2	2,3,4,6
T3	4,5,7,9,10	T4	2,3,4,6,9
T5	2,3,5,7,9	T6	14,15
T7	1,2,4,6,10	T8	8,10,15
T9	2,3,4,5,6	T10	2,3,5,7,9
T11	2,4,9	T12	2,4,6,7,9
T13	1,2,3	T14	3,4,5,7,9
T15	5,6	T16	7
T17	7,8,9	T18	1,2,4,6
T19	1,3,5,7,9	T20	4,5,7,9
T21	10,15,16	T22	1,3,4,6
T23	5,7,9,10,11	T24	1,12,13
T25	13,14,15		

Frequent Attribute (Symptoms) F= {1}, {2}, {3}, {4}, {5},{6},{7},{9},{10},{1,2},{1,3},{1,4}, {1, 5}, {1, 6}, {2, 3}, {2, 4}, {2, 6}, {2, 9}, {3, 4}, {3, 5}, {3, 9}, {4, 5}, {4, 6}, {4, 9}, {5, 7}, {5, 9}, {7, 9}, {2, 4, 6}, {5, 7, 9}}.

Valid Candidates (VC) = {{1 ,7}, {1, 9}, {2, 5}, {2, 7}, {2, 10}, {3, 6}, {3, 7}, {3, 10},{ 4, 7}, {4, 10},{5, 6}, {5, 10}, {6, 7}, {6, 9},{ 6 ,10},{ 7, 10}, {9 ,10}, {1, 2, 3}, {1, 2, 4}, {1, 2, 6}, {2, 3, 4}, {2, 3, 9}, {2, 4, 9}, {3, 4, 5}, {3, 5, 9}, { 4, 5, 9}}.

Table 4-3 Indirect Association Rules

1—2—6 Fever-Head ache-Cold	4—5—7 Fatigue—Unusual thirst—Allergy
1—10—6-7 Fever-Allergy-Cold-Eosinophilia	4—9—7 Fatigue—Diabetic—Allergy
1—15—9 Fever-Breathing Problem-Chest pain	5—1—6 Unusual thirst—Fever—Cold
12—4—5 Chest pain—Fatigue—Unusual thirst	5—4—16 Unusual thirst—Fatigue—Kidney pain
8—11—9 Blood pressure—Obesity, Diabetic	6—2—9 Cold—Head ache—Diabetic
2—1—14 Fever—Head ache—Less platelet count	6—11—12 Cold—Obesity—Chest pain
1-2—3—14 Fever—Head ache—Body pains Less platelet count	1—2—3-6-7 Fever—Cold—Head ache, Body pain, Allergy
3—8—9--12 Body pains—Blood pressure—Diabetic---Chest pain	1—4—16 Fever—Fatigue, Kidney pain
2—7—15 Head ache—Allergy—Breathing Problem	3—1—6 Body pain—Fever—Cold
2—9—15 Head ache—diabetic—breathing problem	3—2—6 Body pain—Head ache—Cold
9-13-14 Diabetic, Excess WBC count, Less platelet count	3—7—6 Body pain—Allergy—Cold

CONCLUSIONS

In this work the seasonal disease dataset are used to identify the common attributes which are causing diseases. An approach called indirect association rule mining adopted to identify the attributes which are directly and indirectly appearing in a disease. The attributes show the symptoms of disease and the various combinations of symptoms found in different patients diagnosis was recorded. By using the above algorithms we checked the common attribute in various diseases. Once if the common attribute for many diseases is identified then the prescription generations becomes very easy for a doctor and the patient also can reduce the investigations and consumption of medicine. In our future work we wish to take many datasets belongs to diseases and identification the main symptom for the cause of various diseases.

REFERENCES

[1] K.Rajeswari, V.Vaithiyanathan and Shailaja V.Pede, Feature Selection for Classification in Medical Data Mining, IJEttS, Volume 2, Issue 2(2013), 492-497

[2] Kharya, Shweta. "Using data mining techniques for diagnosis and prognosis of cancer disease." *arXiv preprint arXiv:1205.1923* (2012).

[3] Soni, Jyoti, et al. "Predictive data mining for medical diagnosis: An overview of heart disease prediction." *International Journal of Computer Applications* 17.8 (2011): 43-48.

[4] Jabbar, M. Akhil, Bulusu Lakshmana Deekshatulu, and Priti Chandra. "Heart disease prediction system using associative classification and genetic algorithm." *arXiv preprint arXiv:1303.5919* (2013).

[5] "Chronic diseases are the major cause of death and disability worldwide." , www.who.int/chp

[6] Srinivas, K., B. Kavihta Rani, and A. Govrdhan. "Applications of data mining techniques in healthcare and prediction of heart attacks." *International Journal on Computer Science and Engineering (IJCSE)* 2.02 (2010): 250-255.

[7] Srinivas, K., G. Raghavendra Rao, and A. Govardhan. "Rough-fuzzy classifier: a system to predict the heart disease by blending two different set theories." *Arabian Journal for Science and Engineering* 39.4 (2014): 2857-2868.

[8] R.Winkler and W.Hays. Statistics: Probabilty, Inference and Decision. Holt, Rinehart &Winston, New York, second edition, 1975.

[9] S.Brin,R.Motwani, and C.Silverstein. Beyond market baskets: Generalizing association rules to correlations. In Proc. ACM SIGMOD Intl.Conf.Management of Data, pages 265-276. Tuscon, AZ, 1997.

[10] T.Brijs, G.Swinnen, K.Vanhoof, and G.Wets. Using association rules for product assortment decisions: A case study. In Proc.of the fifth ACM SIGKDD Conf on Knowledge Discovery and Data Mining, pages 254-260, San Diego, Calif, August 1999.

[11] Robert Cooley, Chris Clifton.Topcat: Data mining for topic identification in a text corpus. In Procedings of the 3rd European Conference of Principles and Practice of Knowledge Discovery in Databases, 1999.

[12] H.T.Reynolds. The Analysis of Cross-Classifications. Macmillan Publishing Co., New York, 1997.

[13] B.Ramasubbareddy,A.Govardhan,A.Ramamohanreddy, Mining Indirect Association between Itemsets, proceedings of Intl conference on Advances in Information Technology and Mobile Communication-AIM-2011 published by Springer LNCS , April 21-22, 2011, Nagapur, Maharastra, India

[14] B.Ramasubbareddy,A.Govardhan,A.Ramamohanreddy, ―Indirect Positive and Negative Association between Itemsets", International conference on Advances in Computing and Communication 2011 (ACC-2011), Rajagiri College of Engineering Kochi Kerala to be held 22-24 July 2011. Proceedings in Springer -LNCS, Berlin, Germany

[15] Jiawei Han**,** and Micheline Kamber,"Data Mining: Concepts and Techniques", Second Edition, Morgan Kaufmann Publishers.

[16] Tan.P, Kumar.V,and Srivastava.J, ―Indirect Association: Mining Higher Order Dependencies in Data", In *Proceedings of the 4th European Conference on Principles and Practice of Knowledge Discovery in Databases,* Lyon, France, pp. 632–637.