ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

Prediction of Most Risk Factors in Hepatitis Disease using Apriori Algorithm.

K Shankar*.

Department of Computer Science and Information Technology, Kalasalingam University, Krishnankoil Tamilnadu, India–626126

ABSTRACT

In today's world, a lot of people are frequently affected by various diseases. Day by day, the count of patients affected by hepatitis disease is increasing. Hepatitis is one of the leading dangerous diseases affecting human. In most hospitals, the medical records of patients with various diseases are maintained in electronic medium. It is very difficult to extract the useful information from the vast volume of records manually. The main objective of this paper is to find the most important factors causing hepatitis disease is the need of the day as the death rate is on the rise. It is a time consuming process to identify patients at risk and take remedial action in time due to large volume of data. Data mining techniques are used to determine buried information that is useful to healthcare practitioners in effective decision making. In this paper, identify the most risk of hepatitis disease affecting majority of patients by using Apriori algorithm with the aid of Weka data mining tool.

Keywords: Data Mining Technique, Medical data mining, Association Rule Mining, Apriori Algorithm, Hepatitis Disease, Prediction.

**Corresponding author*

INTRODUCTION

Every day a lot of information is generated and stored in most of the areas like finance, banking, hospital, etc. A lot of important information may be present in the stored data. This information can be used for decision making. It is a Herculean task and very time consuming to extract that valuable information from the huge amount of data by humans. From the exponentially growing data, it is too hard to discover the useful knowledge without using data mining techniques. Data mining is the nontrivial process of discovering knowledge from a huge volume of data. Such knowledge can be useful in making important decisions [1].

Mining association rule is one of the important technique in data mining which is used to generate frequent itemsets with a user defined minimum support value from large dataset [2]. It was first proposed for market basket analysis. Nowadays it plays a significant role in the field of healthcare industry [3]. Hepatitis disease is a one of the major cause of deaths in the globe. The diagnosis of disease is a difficult and tedious task in medical field. Nowadays, healthcare industry generates huge volume of patient data. These large volumes of dataset contain hidden information, which used for decision making. The death rates of human beings are increased day by day due to hepatitis disease. Diagnosis is usually given on the physical examination of a patient, signs and symptoms. Healthcare experts have their own experience on the bases of which they predict about particular disease of patient which may sometime lead to the false results. The diagnosis of disease is a phenomenal task in medical field. For the purpose of decision making and prediction in regard to hepatitis disease, data mining techniques have shown significant improvement in medical industry.

Every year more than thousand people are affected by hepatitis in the United States from the survey of Centres for Disease Control and Prevention (CDC) [4]. According to the World Health Organization (WHO), report reveals that a large majority of an estimated 325 million people living with the disease lacks access to life-saving testing and treatment [5]. The above survey motivated to predict the risk factors of hepatitis disease. In this research work, applying association rule mining method over hepatitis disease dataset to discover the risk factors (symptoms) affected majority of the patients for decision making and prediction.

ASSOCIATION RULE MINING

Association rule mining is a one of the popular technique in data mining. It finds out frequent patterns, correlations among sets of items (attributes), associations in transactional databases, relational databases, and other information repositories [6]. Association rule mining was first presented in 1993 by R. Agrawal. After that several algorithms have been suggested and developed [7].

Association rule mining is to discover association rules that satisfy the user-defined minimum support and confidence from a given database. Support is defined as the fraction of transactions that contain both A and B [8]. Confidence is a measures how often items in B appear in transactions that contain A. In general, association rule mining can be viewed as two-step process.

Find all frequent itemsets: By definition, an itemset whose support value is greater than or equal to a minimum support threshold value.

Generate strong association rules from the frequent itemsets: Rules that satisfy both a minimum support threshold value and a minimum confidence threshold value [9]. Association rules are first used to find out the frequently occurring pattern which will help in market basket analysis. Association also has great impact in the healthcare field to detect the relationships among diseases, symptoms and health state [10]. Association rules can also be used in many application areas including medical, web usage mining, intrusion detection, library etc [11]. There are many algorithms for generating association rules. One of the well known algorithms is Apriori.

Apriori algorithm is the most classical and important algorithm for mining frequent itemsets. It finds out the relationships among itemsets using two input value minimum support and minimum confidence. Apriori is used to discover all the frequent itemsets in a given database. Frequent itemsets are used to generate the association rules [14]. It uses a Level-wise search, where n-itemsets (An itemset that contains n items) are used to explore (n+1)-itemsets, to extract frequent itemsets from transactional database. First, the set of frequent 1-itemsets is found. This set is denoted as L1. L1 is used to find L2, the set of frequent 2-itemsets, which is used to find L3, and so on, until no more frequent n-itemsets can be found [12]. Apriori algorithm has two steps:

σ Join step
σ Prune step

Join step: To find L_n, joining L_{n-1} with itself to generate C_n. C_n is collection of candidate itemset of size n. L_n is collection of frequent itemsets of size n.

Prune Step: In the prune step find all the frequent itemsets. L_n is the subset of C_n but it is not essential that all items in C_n are frequent. To find out the support count of each candidate in C_n scans the database. If support count of itemsets is greater than or equal to user-defined minimum support then itemsets are frequent which belongs to L_n. If support count of itemsets is not greater than or equal to user-defined minimum support then it is not frequent. These infrequent itemsets are eliminates from C_n. Continue this process until all the frequent items occur in L_n [15]. The step by step procedure of the Apriori Algorithm as follows [12]:

Apriori Algorithm

```
$C_n$: Medical data itemset of size n
$L_n$ : frequent itemset of size n
$L_1$ = {frequent items};
for (n = 1; $L_n$ !=$\emptyset$; n++) do begin
    $C_{n+1}$ = Medical data generated from $L_n$;
    each transaction t in database do
        increment the count of all medical data in $C_{n+1}$ that are contained in t
    $L_{n+1}$ = medical data in $C_{n+1}$ with min_support
end
return $\cup_n L_n$;
```

DATA SOURCE

Hepatitis disease patient actual data for 155 patients were collected from UCI Repository and simulated data of 10393 patient's records are used for this research work [13]. These chosen dataset contains 19 attributes as shown in Table 1. The given attributes are the medical indicators of hepatitis disease symptoms.

Table 1: Attributes of Hepatitis Disease Dataset

Attribute ID	Symptoms (Attributes) Name	Attribute ID	Symptoms (Attributes) Name
1.	Age	11.	Spiders
2.	Sex	12.	Ascites
3.	Steroid	13.	Varices
4.	Antivirals	14.	Bilirubin
5.	Fatigue	15.	ALK Phosphate
6.	Malaise	16.	SGOT
7.	Anorexia	17.	Albumin
8.	Liver Big	18.	Protime
9.	Liver Firm	19.	Histology
10.	Spleen Palpable		

PROPOSED MINING METHOD

Hepatitis is one of the dangerous diseases which frequently affects majority of the patients. The hepatitis patient's medical records have been maintained by hospital information system. Every hospitals has large amount of patients records in their electronic filing. Electronic Medical Records (EMR) systems do not help medicinal practitioners much to identify the patients in dangerous conditions and initiate obstacle in time. EMR contains large volume of patients' data. It would be rather difficult to analyze the huge volume of medical records manually. Hence, data mining algorithms are to be devised to mine the EMRs and identify the attributes causing the disease for the majority of patients. Apriori data mining technique is used in this research work to identify most risk factors (symptoms) of hepatitis disease by which patients are mostly affected. The flow chart in figure 1 reveals that the work flows of the proposed mining method.

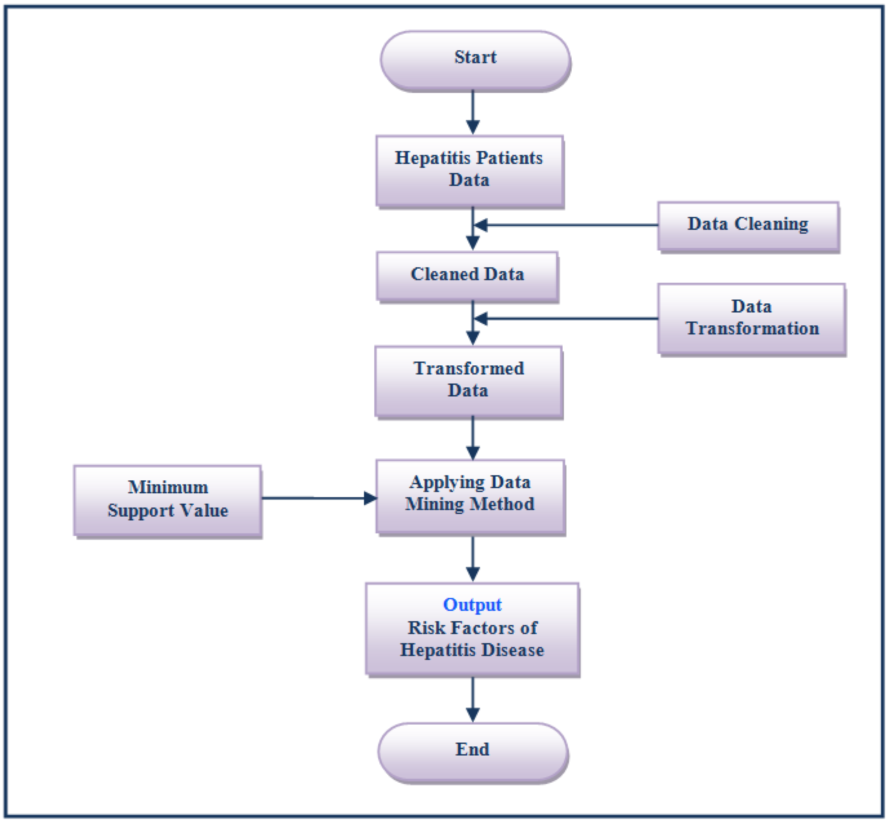

Figure 1: Data flow diagram of the mining process

As can be seen from the diagram the process involves: Data cleaning, Data Transformation and mining the risk factors of hepatitis disease. Every algorithm requires submission of data in a specified format. The conversion of raw data into machine understandable format is called preprocessing. The data preparation phase covers all activities to construct the final dataset from the initial raw data. These raw data can be stored in several formats including text, spreadsheets or other database files. The raw data stored as Attribute Relation File Format (ARFF) are used for experimenting the proposed method.

In step 1, actual data of hepatitis patients from hepatitis disease dataset are taken for the analysis. In step 2, input data from EMRs may contain noise and needs to be cleaned up before mining. It cannot be used directly for processing, with the machine-learning algorithms. Data cleaning can be applied to remove noise and correct inconsistencies in the data. In this process, all the missed attribute values are replaced by value zero (0). In the next step, it needs to be transformed, to make the data appropriate for the mining process. Data transformation through coding normalizes the values to binary 0 or 1. Convert the dataset into binary format denoting the presence or absence of symptom that causes hepatitis diseases as 1 or 0 respectively. Finally, transformed data and minimum support threshold value are input to the proposed mining method for analysis and identification of attributes affecting most of the patients that cause hepatitis disease are identified.

RESULTS AND DISCUSSION

Hepatitis patients' data of 155 from UCI Repository [13] and simulated dataset of 10393 patients with 19 attributes are used in this research work. The proposed mining method is implemented with aid of WEKA data mining tool. The results obtained are promising. Figure 2 shows that the statistical information of hepatitis clinical indicator data of 155 patients. X axis indicates symptoms (attributes) and Y axis indicates the number of patients affected by various symptoms of hepatitis disease. Majority of the male patients are affected by varices symptom and no patients are affected by the symptom protime (values between 10 and 20) and also the symptom SGOT (values between 301 and 400).

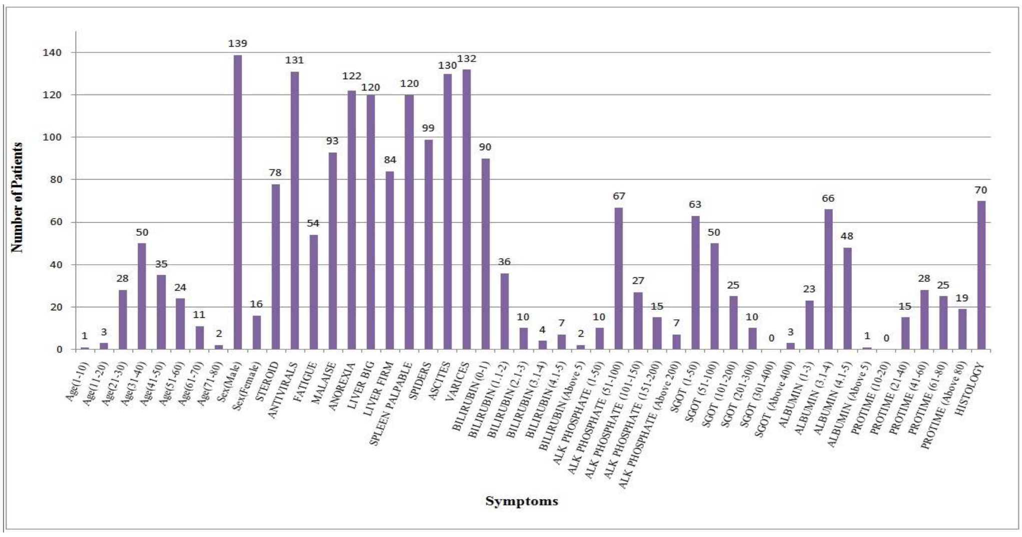

Figure 2: Number of patients affected by various symptoms over dataset contains 155 patients' record

Figure 3 shows that the statistical information of hepatitis clinical indicator data of 10393 patients. Majority of the male patients are affected by Liver Firm symptom and no patients are affected by the symptom protime (values between 10 and 20) and also the symptom SGOT (values between 301 and 400).

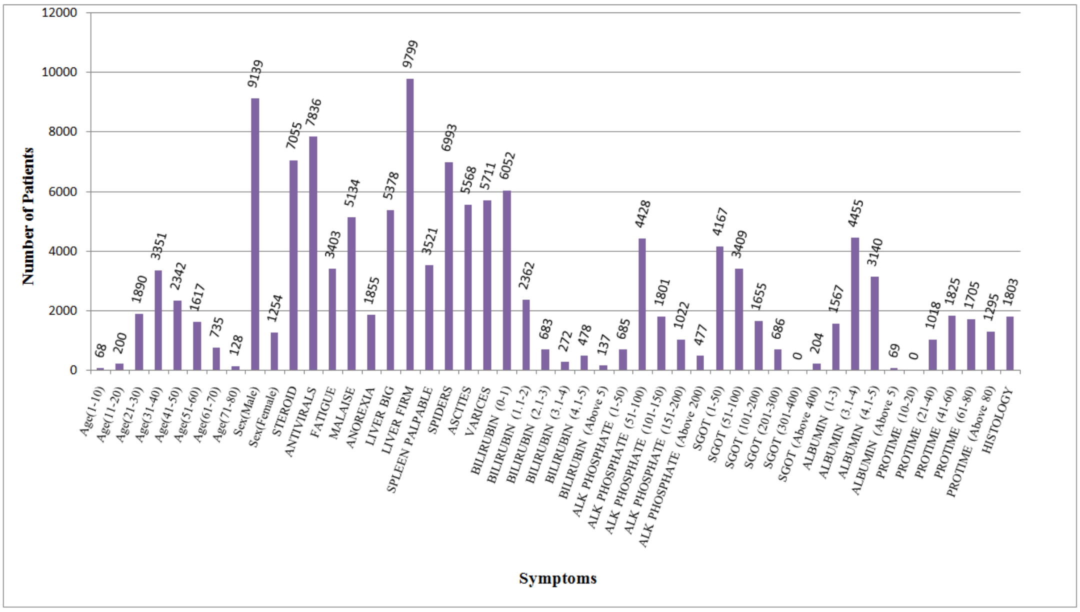

Figure 3: Number of patients affected by various symptoms over dataset contains 10393 patients' records

The developed method is experimented with hepatitis patient dataset containing 155 patients' records with various support values ranging from 0.3 to 0.5 to find risk factors (symptoms) of majority of the patients affected frequently and the results are plotted in figure 4.

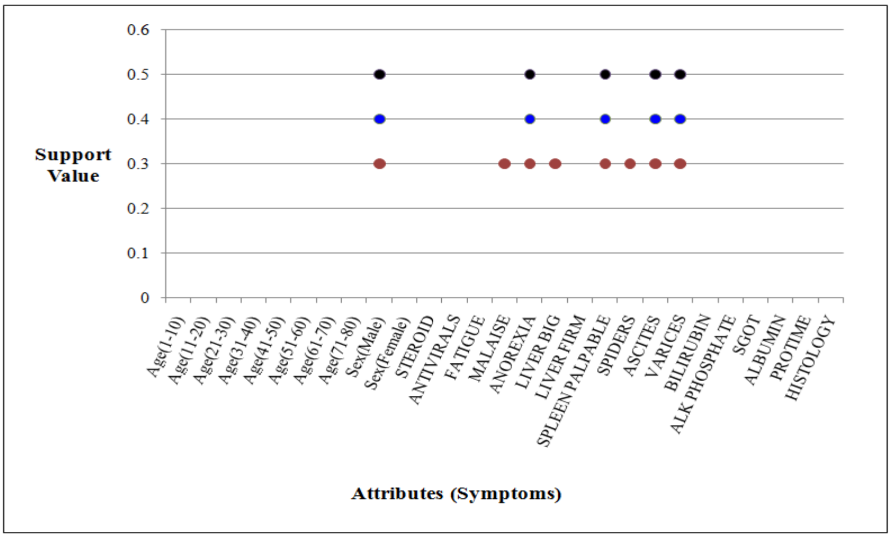

Figure 4: Frequently occurring risk factors with various support value over 155 hepatitis patients' records

The developed method is also applied over a large simulated dataset containing 10393 patient records, with various support values ranging from 0.3 to 0.5. Risk factors of hepatitis affected majority of the patients frequently as shown in figure5.

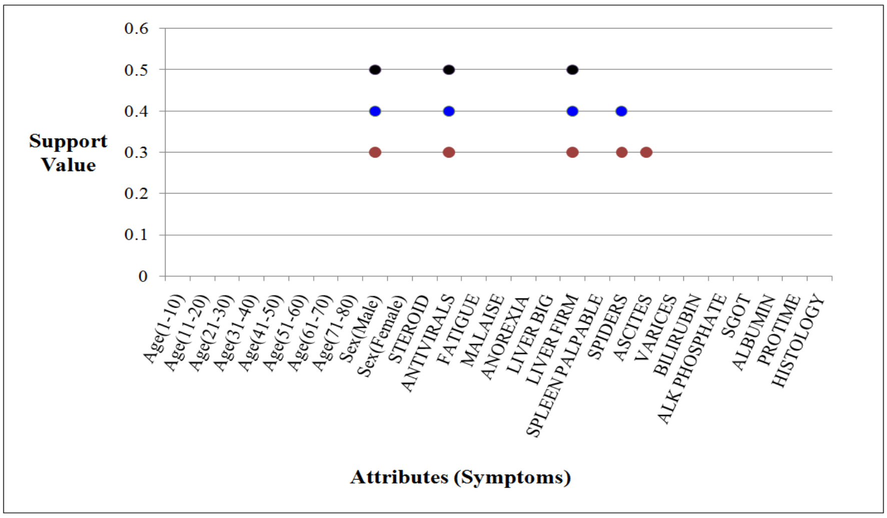

Figure 5: Predict the frequently occurring risk factors with various support value from simulated dataset of 10393

From the result, it reveals that the proposed method is efficient to mine the risk factors of hepatitis which affect majority of the patients frequently. Atleast half of the male patients are affected frequently by the symptoms Anorexia, Spleen Palpable, Ascites and Varices for the chosen dataset includes 155 patients' records. Antivirals and Liver Firm are affecting atleast half of the male patients frequently for the chosen dataset contains 10393 patients' data.

Plot in figure 6 shows the number of patients is affected frequently by most risk of hepatitis disease for two different datasets data with various minimum support values.

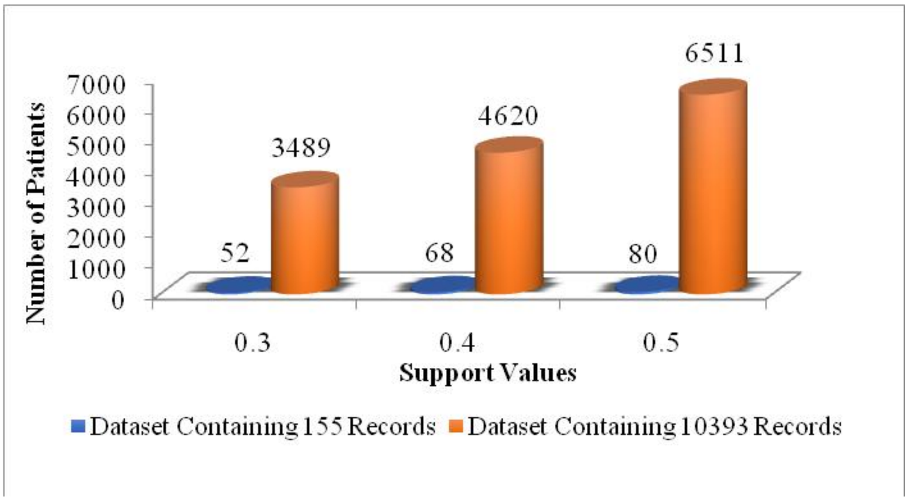

Figure 6: Number of patients is affected by most risk of hepatitis disease for two different datasets with various minimum support values

CONCLUSION

Medical data mining plays an important role in the diagnosis of various diseases and in life saving decisions. It is necessary to find frequently affecting symptoms from patient data to predict the most risk factors causing dangerous diseases. In this paper, proposed work that finds most risk factors causing hepatitis disease affecting majority of the patients using Apriori algorithm. It provides a hasty aid to the medical practitioner in making emergency decisions to save the lives of patients. The proposed method is applied over a hepatitis disease datasets of 155 and 10393. The prediction results are encouraging and the efficiency of the method in identifying most important factors (symptoms) which causes the hepatitis disease.

REFERENCES

[1] Xue Z. Wang. Data Mining and Knowledge Discovery for Process Monitoring and Control, Springer-Verlag, London, 1999.

[2] Ilayaraja, M. and Meyyappan, T. Medical Data Mining Method to Predict Risk Factors of Heart Attack and Raise Early Warning to Patients, International Journal of Applied Engineering Research, Vol.10, No.55, 2015.

[3] Ilayaraja M, Meyyappan T. Efficient Data Mining Method to Predict the Risk of Heart Diseases Through Frequent Itemsets. Procedia Computer Science. 2015; 70: 586-92.

[4] Centres for Disease Control and Prevention, https://www.cdc.gov/hepatitis/statistics/index.htm

[5] World Health Organization, http://www.who.int/hepatitis/en/

[6] Chandrakar I, Kirthima AM. A Survey on Association Rule Mining Algorithms. International Journal of Advanced Research in Computer Science. 2013 Nov 1;4(11): 270-272.

[7] Prithiviraj, P., and R. Porkodi. A comparative analysis of association rule mining algorithms in data mining: a study. Open J. Comput. Sci. Eng. Surv 3.1 (2015): 98-119.

[8] Kotsiantis S, Kanellopoulos D. Association rules mining: A recent overview. GESTS International Transactions on Computer Science and Engineering. 2006 Jan;32(1):71-82.

[9] Tiwari M, Singh R, Singh SK. Association–Rule Mining Techniques: A general survey and empirical comparative evaluation. International Journal of Advanced Research in Computer and Communication Engineering. 2012; 1.10: 858-860.

[10] Tomar D, Agarwal S. A survey on Data Mining approaches for Healthcare. International Journal of Bio-Science and Bio-Technology. 2013 Oct;5(5):241-66.

[11] Kaur J, Madan N. Association Rule Mining: A Survey. International Journal of Hybrid Information Technology. 2015;8(7):239-42.

[12] Ilayaraja M, Meyyappan T. Mining medical data to identify frequent diseases using Apriori algorithm. In Pattern Recognition, Informatics and Mobile Engineering (PRIME), 2013 International Conference on 2013 Feb 21 (pp. 194-199). IEEE.

[13] Lichman, M. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, 2013.

[14] Adeniji, I. A. Saheed, Y.K. Oladele, T.O. and Braimah, J.O. "Comparative Analysis of Association Rule Mining Techniques for Monitoring Behavioural Patterns of Customers in a Grocery Store", African Journal of Computing & ICT, Vol.8, No.3, 2015.

[15] Jeetesh Kumar Jain, Nirupama Tiwari and Manoj Ramaiya. "A Survey: on Association Rule Mining", International Journal of Engineering Research and Applications (IJERA), Vol.3, Issue.1, pp.2065-2069, 2013.