Middle-East Journal of Scientific Research 24 (S2): 248-253, 2016
ISSN 1990-9233
© IDOSI Publications, 2016
DOI: 10.5829/idosi.mejsr.2016.24.S2.159

# Mining Association Rules for Early Diagnosis of Diseases from Electronic Health Records

*J. Sabthami, K. Thirumoorthy and K. Muneeswaran*

Department of Computer Science and Engineering, Mepco Schlenk Engineering College, Sivakasi, India

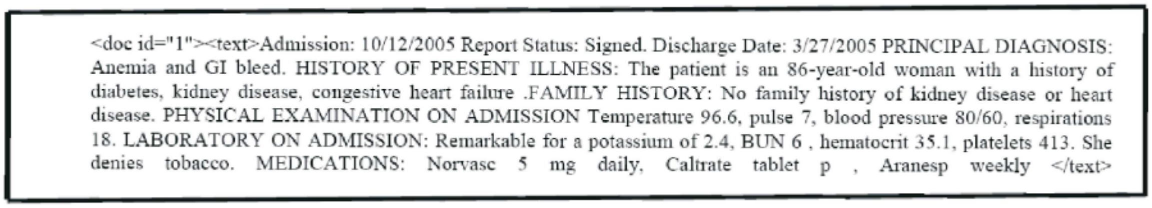
**Abstract:** Clinical documents generated from electronic patient health record system has rich data which addresses the details about patient's disease, injuries, vital sign information, medication, symptoms. To improve the health of the patient, patients are classified based on the presence of disease using Naïve Bayesian and K-Nearest Neighbor (KNN) algorithm. The large amount of missing value is handled by various imputation strategies to improve the accuracy of classification. The association among diseases and association among drugs and diseases are performed to recommend the patients about the possibility of other diseases which may occur and knowledge about the frequent drugs that can be prescribed for the patient with the particular diseases.

**Key words:** Clinical documents · Symptom · Medication · Naïve Bayesian algorithm and K-nearnest neighbor algorithm · Association

## INTRODUCTION

Documenting the patient's information in the form of document is very useful for analyzing the patient's diseases and provide further treatment immediately. Clinical documents which is also called as medical records is in the format of unstructured textual documents, contains the information about the patients like age, gender, drugs prescribed, symptoms, blood pressure, heart rate, hematocrit, blood nitrogen urea, respiratory rate, admission date, injuries, etc [1]. Every day large amount of valuable clinical documents are generated from electronic health record system which is accessed only by the authorized people. The clinical document is used to improve the health of the patients by extracting the information about each patient and to recommend medical advices, drugs. It is difficult and time consuming to read every clinical document and understand the conditions of patients. Therefore, various data mining techniques like clustering, classification, association rule mining are used to analyze every clinical document and infer valuable information at less time and treat the patients accordingly. MedEx tool, is an open source tool built in Java to extract the name, dosage, frequency, generic name, strength, route of drugs consumed by patients from the clinical documents [2]. Among all information the generic drug names are extracted. The regular expression is a sequence of characters and symbols for searching text based on matching pattern that is used for extracting the features like age, blood pressure, weight, hematocrit, heart rate and so on [3]. The result of feature extraction contains large amount of missing values which is handled by single imputation (Average and linear regression) and multiple imputation strategies [4][5] to improve the accuracy of analysis. Multi-label classification is a supervised learning and chain of one or more single label classification. Multi-label classification of diseases is performed because each patient may have more than one disease [6]. The Naïve Bayesian classification handles the data better even if missing values are found. But K-Nearest [7] Neighbor (without missing[8] values) has good accuracy [9] compared to Naïve Bayesian (with missing and without missing values). KNN does not support when missing values are found. Association rule mining is used to find the co-occurring items in the database. The correlation between diseases-diseases [10] and drugs-diseases [11] is performed using association rule mining techniques. The association rule based on disease-disease and drug-disease is generated using Apriori algorithm, which is easy and follow large item set property [12].

**Corresponding Author:** J. Sabthami, Department of Computer Science and Engineering,
Mepco Schlenk Engineering College, Sivakasi, India.



<doc id="1"><text>Admission: 10/12/2005 Report Status: Signed. Discharge Date: 3/27/2005 PRINCIPAL DIAGNOSIS: Anemia and GI bleed. HISTORY OF PRESENT ILLNESS: The patient is an 86-year-old woman with a history of diabetes, kidney disease, congestive heart failure .FAMILY HISTORY: No family history of kidney disease or heart disease. PHYSICAL EXAMINATION ON ADMISSION Temperature 96.6, pulse 7, blood pressure 80/60, respirations 18. LABORATORY ON ADMISSION: Remarkable for a potassium of 2.4, BUN 6 , hematocrit 35.1, platelets 413. She denies tobacco. MEDICATIONS: Norvasc 5 mg daily, Caltrate tablet p , Aranesp weekly </text>

Fig. 1: Example of Clinical Document

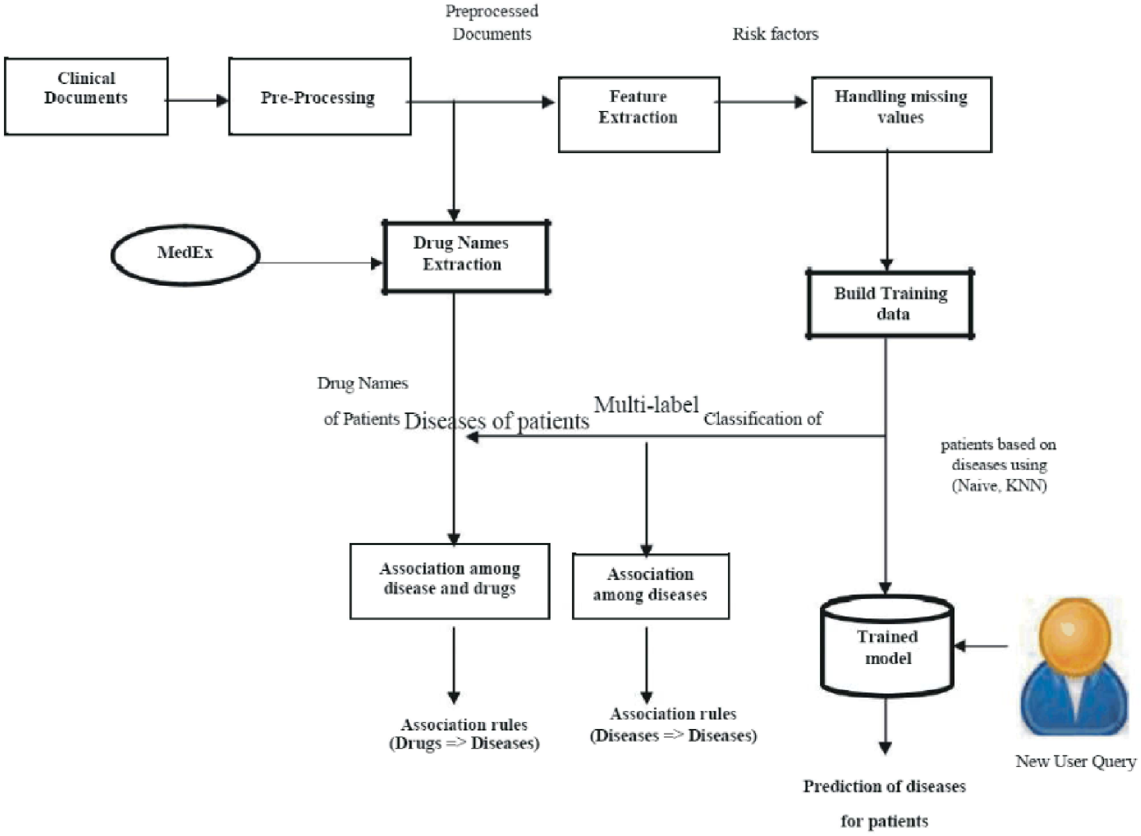

Clinical Documents
Pre-Processing
Preprocessed Documents
Feature Extraction
Risk factors
Handling missing values
MedEx
Drug Names Extraction
Build Training data
Drug Names of Patients
Diseases of patients
Multi-label Classification of patients based on diseases using (Naive, KNN)
Association among disease and drugs
Association among diseases
Trained model
New User Query
Association rules (Drugs => Diseases)
Association rules (Diseases => Diseases)
Prediction of diseases for patients

Fig. 2: Overview of classification and mining association rules from clinical document

### Extraction of Features

**Electronic Record:** Every day a large volume of clinical records are generated from electronic health record. Clinical health record includes the various information about patients like age, gender, blood pressure, weight, hematocrit, blood nitrogen urea, various laboratory test, billing, symptoms, medication prescribed.

The main aim of storing the clinical record is to improve the health care of the patient by eliminating the process of manually tracking the prescriptions of patient's and ensure all data is correct and accurate. The clinical documents are in semi-structured or in unstructured format without pattern of information. Example of Clinical document is given in Fig 1.

**Pre-Processing:** The preprocessing in unstructured clinical document is done by three methods. Fig 2 shows the pre-processing steps such as Removal of Stop words, Removal of Punctuation, Removal of Negation words.

The negation words include pre-negation words, post-negation, family history words. The sentence which contains the negation words are removed. Some of pre-negation words are no change, not necessary, no increase and post-negation words are denies, denied, not have. The words indicating family history includes wives, grandmothers, mother, father's and so on. For Examples, negation words removal - consider a sentence "Her mother has liver cancer". In this example the liver cancer is a disease of the patient's mothers not hers so the sentence is removed.

Table 1: Sample structure format after feature extraction

| Patient id | Age | Gender | Diastolic pressure | Systolic pressure | … | Heart rate |
|---|---|---|---|---|---|---|
| 1 | 86 | Male | 120 | 70 | | 70 |
| 3 | 50 | Male | 150 | 100 | | 80 |
| 5 | 45 | Female | NA | NA | | 73 |
| 7 | 58 | Male | 165 | 102 | | NA |
| 9 | 62 | Female | 127 | 85 | | 65 |

Table 2: Three different strategy to handle the missing values based on percentage of missing values

| % of risk factors data missing | Methods to replace missing value. |
|---|---|
| <3% | Average |
| 3-10% | Linear regression |
| >10% | Multiple Imputation |

**Feature Extraction:** The valuable information like Age, gender blood pressure, temperature, hematocrit, respiration rate, heart rate, Blood nitrogen urea, sodium, potassium value of each patients are extracted using Regular Expressions from the clinical documents.

The medication/drug names are extracted using MedEx tool, which is used to extract the Drug name, generic name of the drug, dosage, brand name and so on. Among them only the generic name of the drug of the patients are extracted. The sample structured format of data after feature extraction from clinical document is given in Table 1.

**Handling Missing Values:**

$$\frac{\text{Total Number of value missed in a feature}}{\times 100\ \text{Total Number of patients}}$$

Based on the percentage of missing value different strategies in Table 2 are applied to handle it.

Average is given as sum of the values of a feature for all patients to the total number of patients. Linear regression is used to predict the missing value from available values. Linear regression is a supervised method in which the regression equation is build based on the patients record without missing values. Regression equation is given as

$$y = b_0 + b_1 x_1 + b_2 x_2 + ...$$

Multiple imputations mean instead of replacing the missing value by a single value, it is replaced with imputed values using specified regression model. This step is repeated m times, resulting in a separate dataset each time. The result of all complete data set is average to solve the missing value problem.

**Multi-label Classification of Patients Based on Diseases:** Some of the features of patients are found to be missing. The missing values are due to if the regular expression fails to recognize the data and if the test is not recorded for the patient.

If age and gender is missing then the entire record has to be removed, since they are considered as mandatory attributes for patient. The replacement of missing value by single imputation like Average, Linear Regression will not provide a better accuracy so multiple imputation method is performed. The percentage of missing value is found by Classification is a supervised learning because the class label is known. The multi label classification is performed to classify the patients based on diseases [Asthma, Coronary Artery Disease (CAD), Depression, Hypertriglyceridemia, Gout, Obstructive Sleep Apnea (OSA), Peripheral Vascular Disease (PVD), Gallstone, Hypertension, Diabetes]. Single label classification algorithms like Naïve Bayesian Algorithm and K-Nearest Neighbor algorithm is performed multiple times for multiple labels to predict the diseases of the patients.

**Naïve Bayesian Algorithm** Naïve Bayesian algorithm is suitable for classification of missing value.

**Input:** D be the training set of tuples with class label.
**Output** Multi-label Classification based on naïve Bayes and predict the class label for unseen data.

**begin**
**repeat** for all label
For $T = t_1, t_2. ..t_n$ set of tuples and $c_1, c_2. ..c_m$ classes.

According to Bayes theorem, classification is done by

$$P(C_i \mid T) = \frac{P(T \mid C_i) P(C_i)}{P(T)}$$

where,
$P(C_i \mid T)$ is the posterior probability
$P(T \mid C_i)$ is the likelihood
$P(C_i)$ is the prior probability
$P(T)$ is the independent probability of T

**End**

**K-Nearest Neighbor (KNN) Algorithm:** KNN algorithm is a simple algorithm in which classification is performed using distance measures. The thumb rule of identifying the value of K is

$$K = \sqrt{Total\ Number\ of\ attributes}$$

**Input:** Data D with training tuple and associated class label, K value.

**Output:** Multi-label Classification based on KNN and predict the class label for unseen data

```
begin

    • Calculate the Euclidean distance measure between the training samples and the new data

      Distance = $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$

    • Sort based on distance value
    • Calculate the nearest neighbor based on K Minimum distances.

    • Find the majority category, which is the predicted value for the unseen data.
    • Perform the steps multiple times to predict multiple label.

end
```

**Association among Diseases-Diseases and Drugs-Diseases:** The association among diseases and diseases and drugs are performed to recommend the patient about the possibility of other disease to occur and medical information about the drugs to be consumed for the particular disease. The Association rule mining is performed using Apriori Algorithm.

**Association among Diseases and Diseases:** The association rule mining algorithm is used to find information of frequently occurring and the possible disease information [10].

Example: - Coronary artery disease -> Hypertension

**Association among Diseases and Drug Names:** The association rules based on disease and drug names is found based on patient's information.

Example: - Metformin -> Coronary artery disease.

The directionality is included in association rule mining to improve the findings of frequent itemsets. The directionality is based on confidence value.

Example: - Insulin -> Diabetes.

Here the insulin is a drug prescribed to all diabetes patient but all diabetes patients will not consume insulin [11].

**Apriori Algorithm:**

```
Input – D: - a database of transaction min_sup (minimum support threshold)

Output – Frequent item set in D

begin repeat

    The set of candidate item set is formed by scanning the D Compare candidate support count with min_sup

        if (candidate support count < min_sup)
        prune

end

    The set of candidate K-itemset is generated by joining by itself

until no frequent itemset is found end
```

**Experimental Result:**

**Dataset:** The clinical documents of 1235 patients is collected from the site www.i2b2.org (informatics for integrating Biology and the Bedside), the 2008 Obesity challenge dataset. The total number of training and testing data after preprocessing and handling missing values is 374 and 228 respectively.

**Performance Metrics**

**Multilabel Classification Result:** The Multilabel classification is evaluated using accuracy. The accuracy of the classification using Naïve Bayesian and KNN is given in Table 4 and 5

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Table 4: Accuracy of classification, when K in KNN is 6

| | Missing | | Average | | All 3 strategy | |
|---|---|---|---|---|---|---|
| Diseases | Naïve | KNN | Naive | KNN | Naive | KNN |
| Asthma | 83.41 | - | 71.18 | 84.24 | 80.35 | 84.72 |
| CAD | 69.87 | - | 68.56 | 58.52 | 68.56 | 61.14 |
| Depression | 65.94 | - | 47.16 | 75.9 | 56.77 | 71.80 |
| Hyper-triglyceridemia | 86.46 | - | 83.84 | 94.32 | 83.84 | 94.32 |
| Gout | 72.93 | - | 42.79 | 84.72 | 52.40 | 86.42 |
| OSA | 86.46 | - | 65.50 | 88.65 | 83.41 | 88.21 |
| PVD | 74.24 | - | 35.37 | 75.11 | 73.36 | 75.98 |
| Gallstone | 66.38 | - | 43.23 | 76.42 | 62.01 | 78.17 |
| Hypertension | 76 | - | 73.36 | 71.62 | 75.55 | 72.49 |
| Diabetes | 69 | - | 65.07 | 64.63 | 65.94 | 67.69 |

Table 5: Accuracy of classification, when K in KNN is 12

| | Missing | | Average | | All 3 strategy | |
|---|---|---|---|---|---|---|
| Diseases | Naïve | KNN | Naive | KNN | Naive | KNN |
| Asthma | 83.41 | - | 71.18 | 85.59 | 80.35 | 85.59 |
| CAD | 69.87 | - | 68.56 | 63.32 | 68.56 | 63.76 |
| Depression | 65.94 | - | 47.16 | 77.29 | 56.77 | 78.60 |
| Hyper-triglyceridemia | 86.46 | - | 83.84 | 94.76 | 83.84 | 94.76 |
| Gout | 72.93 | - | 42.79 | 87.77 | 52.40 | 88.65 |
| OSA | 86.46 | - | 65.50 | 90.83 | 83.41 | 90.83 |
| PVD | 74.24 | - | 35.37 | 80.35 | 73.36 | 79.91 |
| Gallstone | 66.38 | - | 43.23 | 80.35 | 62.01 | 78.60 |
| Hypertension | 76 | - | 73.36 | 74.24 | 75.55 | 75.11 |
| Diabetes | 69 | - | 65.07 | 69.87 | 65.94 | 69.43 |

Table 6: Association among diseases and diseases

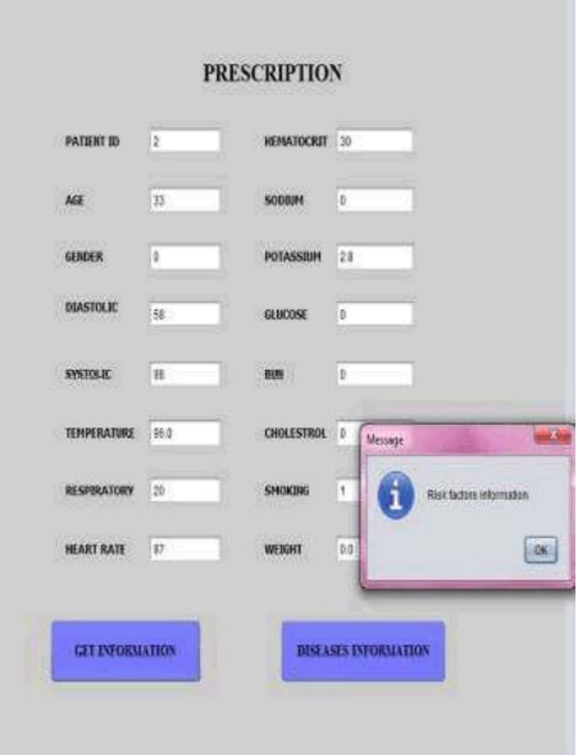
| Rules (when min_sup=37) | Support | Confidence (%) | Lift |
|---|---|---|---|
| PVD, Hypertension => Diabetes | 0.1390 | 91 | 1.29 |
| Cad, PVD => Diabetes | 0.1283 | 91 | 1.28 |
| Cad, PVD => Hypertension | 0.1283 | 91 | 1.12 |

Table 7: Association among diseases and drugs

| Rules (when min_sup=54) | Confide | Support nce (%) | Lift |
|---|---|---|---|
| Insulin=>Diabetes | 0.3490 | 99 | 1.42 |
| Atenolol =>Hypertension | 0.2548 | 99 | 1.42 |
| Aspirin, Clopidogrel =>Cad | 0.12465 | 96 | 1.56 |
| Creatinine, Insulin => Diabetes | 0.2160 | 100 | 1.43 |
| Lisinopril =>Hypertension | 0.3240 | 91 | 1.14 |

According to the classification result, it is observed that Naïve Bayesian has high accuracy before handling missing values but with KNN using all 3 strategies of handling missing value it is accurate. KNN, the accuracy increases with K.

**Association among Diseases and Diseases and Association *among Diseases-drugs*:** Some of the association rules generated using apriori algorithm is evaluated using support, confidence and lift as shown in Table 6 and 7.



PRESCRIPTION

| | | | |
|---|---|---|---|
| PATIENT ID | 2 | HEMATOCRIT | 30 |
| AGE | 33 | SODIUM | 0 |
| GENDER | 0 | POTASSIUM | 2.8 |
| DIASTOLIC | 58 | GLUCOSE | 0 |
| SYSTOLIC | 88 | BUN | 0 |
| TEMPERATURE | 96.0 | CHOLESTROL | 0 |
| RESPIRATORY | 20 | SMOKING | 1 |
| HEART RATE | 87 | WEIGHT | 0.0 |

Message

Risk factors information

Fig. 3: Output of Prescription form

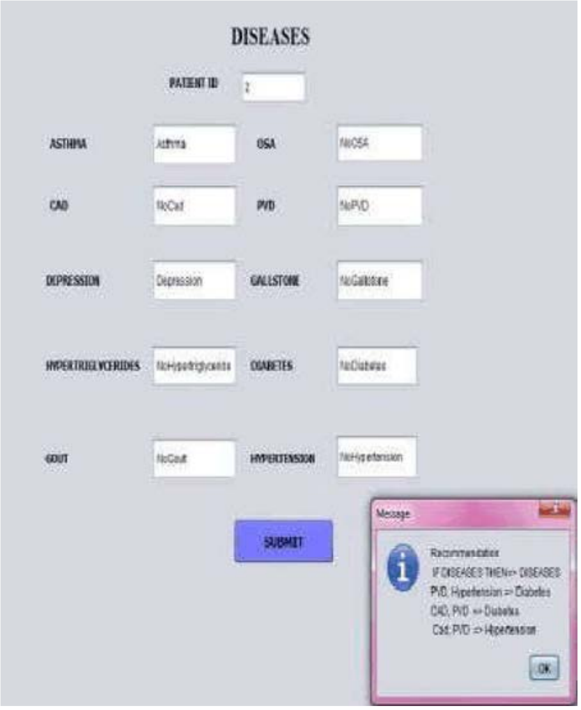
**Support:** Percentage of occurrence of the item frequently in the dataset. And is given as $P(AUB)$

**Confidence:** Conditional probability that, the given X present in a tranaction, Y will also be present. It is given as $P(B / A)$

**Lift:** is an interesting measure. The rules are interesting if the value of lift is >1 and given as $\frac{P(A \cup B)}{P(A)P(B)}$

The recommendation is given based on disease of the patients and association rules. The output of prescription form and recommendation is given in Fig 3 and 4 as screenshot form.

**Conclusion and Future Work:** In this paper, we proposed a method to classify the patients based on the diseases. By performing such a classification we improve the patient's health care by early diagnosing the diseases based on the risk factors. In this work we found the association rule mining to provide information about what are the diseases which may cause in future based on the present diseases and what are the drugs which can be



DISEASES

| | | | |
|---|---|---|---|
| PATIENT ID | 2 | | |
| ASTHMA | Asthma | OSA | NoOSA |
| CAD | NoCad | PVD | NoPVD |
| DEPRESSION | Depression | GALLSTONE | NoGallstone |
| HYPERTRIGLYCERIDES | NoHypertriglycerides | DIABETES | NoDiabetes |
| GOUT | NoGout | HYPERTENSION | NoHypertension |

Recommendation
IF DISEASES THEN=> DISEASES
PVD, Hypetension => Diabetes
CAD, PVD => Diabetes
Cad, PVD => Hypertension

Fig. 4: Output of Recommendation form

prescribed based on the diseases for a patients. Here we compared multilabel Naïve Bayesian and KNN classifier in which KNN classifier is shown to be more accurate by replacing missing values. In future work, it is planned to improve the accuracy of classification and to enhance the method for patients who are in complex condition and to extended it for many diseases.

## ACKNOWLEDGMENT

The authors like to express the gratitude to Management and Principal of Mepco Schlenk Engineering for providing support for carrying out this work.

## REFERENCES

1. Kirk Roberts, Sonya E. Shooshan, Laritza Rodriguez, Swapna Abhyankar, Halil Kilicoglu and Dina Demner-Fushman, 2015. The role of fine-grained annotations in supervised recognition of risk factors for heart disease from EHRs. Journal of Biomedical Informatics-Elsevier 2015.
2. Pakhomov, S.V., A. Ruggieri and C.G. Chute, 2002. Maximum entropy modeling for mining patient medication status from free text in Proceedings of the AMIA Symposium. 2002. American Medical Informatics Association.
3. Jay Urbain, 2015. Mining heart disease risk factors in clinical text with names entity recogntion and distribution semantic models. Published in journal of biomedical informatics in Elseiver. December 2015.
4. Jitendra Jonnagaddala, Siaw-Teng Liaw Pradeep Ray, Manish Kumar, Nai-Wen Chang and Hong-Jie Dai, 2015. Coronary artery disease risk assessment from unstructured electronic health records using text mining. Elseiver 2015.
5. Jeffrey C. Wayman, 2003. Ph.D. Multiple Imputation For Missing Data. Center for Social Organization of Schools Johns Hopkins University. 2003.
6. Grigorios Tsoumakas and Ioannis Katakis, XXXX. Deparatment of informatics, Aristotile University. Multi-label Classification An overview.
7. Yuan Ling, Xuelian Pan, Guangrong Li and Xiaohua Hu, 2015. Member, IEEE, Clinical Documents Clustering Based on Medication/Symptom Names using Multi- View Nonnegative Matrix Factorization, 2015.
8. Parvez Ahmad, Saqib Qamar and Syed Qasim Afser Rizvi, 2015. Techniques of Data Mining In Healthcare, in International journal of computer Application, 2015.
9. Raikwal, J.S. and Kanak Saxena, 2012. Performance Evaluation of SVM and K-Nearest Neighbor Algorithm over medical data. International journal of Computer Application.
10. Gitanjali, J., C. Ranichandra and M. Pounambal, 2014. APRIORI algorithm based medical data mining for frequent disease identification, in International Journal of Information Technology.
11. Adam Wright, Elizabeth S. Chen and Francine L. Maloney, 2010. An automated technique for identifying associations between medications, laboratory results and problems, in Elseiver Journal of Biomedical, 2010.
12. Jiawei Han und Micheline Kamber.Frequent itemset m.