

An Efficient Rule-Mining for Medical Diagnosis: A Market-Basket Approach

¹AWODUN, Mojirade Adejumo & ²ADEDARA, Raphael Oluwadare

¹Department of Computer Science, University of Ilorin, Ilorin, Kwara State, Nigeria.

² Department of Computer Science, College of Education, Ikere-Ekiti, Ekiti State, Nigeria.

Received: February 01, 2020

Accepted: March 03, 2020

ABSTRACT: Medical diagnosis is an important and also delicate and time involving task that needs to be carried out in order to accurately identify a disease a patient might be suffering from. Due to the inaccurate diagnosis given to patients which could result in a wrong prescription that may lead into unwelcome consequences therefore to help in this direction an automated system is developed which is able to generate the association rules among different symptoms which will also assist the medical personnel in diagnosing patient's diseases accurately. In actualizing this aim, an analysis technique in data mining which is Market Basket Analysis using bitmap representation of Transaction table is adopted in revealing the association rule among the symptoms and diseases given a support and confidence level. The rules generated are compared with a doctor's prescription to determine its relevance in healthcare, the irrelevant ones are then filtered out to ascertain the ones which are acceptable for diagnosis of the various diseases.

Key Words: Rule-Mining, Medical Diagnosis, Market-Basket and Market-Basket Approach

Introduction

Good medical care relies solely on well- trained doctors and nurses and on high-quality facilities and equipment. Good medical care also relies on adequate record keeping because without accurate, up-to- date and accessible patients case notes, medical personnel may not offer the best treatment or may misdiagnose a condition which can leads to a very severe consequences like death.

In our society today, a large number of deaths are recorded daily most especially in the underdeveloped and urban areas which is due to inadequate medical facilities. However in recent years especially this 21st century, some hospitals have introduced computer in their services. Notwithstanding, the use of computer in the health services is mainly on some housekeeping functions such as accounting, costing, store control, billing, medical record keeping, payroll and so on. Much emphasis has not be made in the area of medical diagnosis.

For several years now most especially in Nigeria, the idea of using a computer for medical diagnosis has been partially left out in the process of automating or computerizing health services. Medical activities in Nigeria are still mostly by manual method. "Health is wealth" goes a popular saying, and this is one of the reasons the medical aspect of life should not be taken with levity.

Misdiagnosis has been explained as a failure or delay in diagnosis. In diagnosis, errors can occur as a result of single or various flaws within the healthcare system such flaw can include:

- Patients inability to express themselves which leads to faulty data gathering;
- Wrong interpretation of results;
- Not considering all other possible diagnosis;
- Failure to verify diagnosis and gather new data e.t.c.

These and many other flaws have resulted in misdiagnosis which has caused inappropriate or inadequate treatment which in turn has led to different fatal consequences. It has been discovered that no healthcare in the world is free from the unfortunate issue of making wrong diagnosis of diseases.

Due to the increase of data in our society today, most organizations and big companies have large databases that contain a great amount of undiscovered and valuable information. The huge amount of data in the database makes it very difficult to examine and discover knowledge of interest with the manual process. This is due to the fact that the quantity of data in the world roughly doubles every year, but the astonishing fact is that the amount of relevant and useful information decreases rapidly which has now led to emergence of data mining technologies. Jackson (2002) opined that just like the real world mining, the "information gold nuggets" or "diamonds" can be automatically found by sifting out massive quantities of data- wreckage from our database through data mining with the computer.

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information. Data mining is supported by a variety of techniques such as statistical methods, rule induction, neural networks, conceptual clustering, association discovery, etc. Information extraction can be enhanced by the cooperative use of several data mining techniques (Fawzi et al, 2011). Association rule mining which is a technique in data mining, finds interesting associations and/or correlation relationships among large set of data items. Association rules show attributes value conditions that occur frequently together in a given dataset. A typical and widely-used example of association rule mining is Market Basket Analysis.

Market-Basket Analysis is an advanced predictive model used to determine associations between items or events and deriving rules that indicate the likelihood of items occurring together in groups of specific types. It also identifies items that are unlikely to occur together. Apart from discovering the associating rules, Market-basket analysis also assist in visualizing the degree of attraction between items, view charts representing item associations, rank association rules and applying them to new data to produce recommendations. Market basket analysis is a data mining methods used in retail companies which focused on discovering purchasing patterns by extracting associations or co-occurrences from a store's transactional data.

In healthcare, data mining is becoming increasingly trendy and also indispensable. So many factors have motivated the use of data mining applications in health services. For example, the existence of medical insurance fraud and abuse has led numerous healthcare insurers to attempt to reduce their losses by using data mining tools to help them find and track offenders. Detecting fraud using data mining applications is prevalent in the commercial world, for example, by detecting fraudulent credit card transactions. The huge amounts of data generated by healthcare transactions are too complex and voluminous to be processed and analyzed by traditional methods.

Fawzi et al (2011) affirmed that there are various natural arguments among the traditional ways of data mining strategies and medicine in past researches. He also affirmed that Data Mining in the field of medical research starts with an assumptions thereby altering the outputs so as to be able to achieve the previous assumptions. This process is different in the case of standard data mining because here the datasets does not begins with assumptions this is because standard data mining lacks clarifying development and models and focuses mainly on relating items. He also stated that Healthcare institutions should utilize the process of data mining to be able to gain more facts and knowledge from the data that is present in their institution records.

The impact of data mining in health care nowadays is great which has made data mining essential. This is because the health organization generates and gather large amounts of data on daily basis. This gives room for data mining which makes it possible to extract data that helps in getting interesting rules, knowledge and regularities. The knowledge extracted in the data can help in improving the work efficiently and also enhance quality decision making. Data mining assist healthcare organization in predicting trends in patient conditions and their behaviors. It also helps in early detection of contagious diseases and their treatments. Data mining can be limited in data access this is owed to the fact that the unprocessed inputs for data mining frequently exist in different settings and systems (Borris, 2012).

Rani (2011) demonstrated in his research the use of neural networks in the analysis of Heart diseases dataset to generate efficient classification rules. The neural network is skilled using Back propagation algorithm to be suitable in performing classification task of medical data. It is also stated in his research that data mining algorithms can gain knowledge from previous examples in clinical records and model the often times non-linear relationships between the independent and dependent variables. Neural networks have been applied to various areas of healthcare like diagnosis areas, image analysis, drug application etc. It is also stated in the research that neural networks are more vigorous in noisy and incomplete data. As advantageous as the neural network is the structure of the neural network gotten for a particular problem cannot be specify due to the fact that the structure is often learnt by experiment and it is also difficult to understand how it works.

Mohammed et al (2013) asserted in a paper which focuses on analysis of different data mining techniques needed for medical data mining especially in discovering locally frequent diseases like heart ailments, cancer of the lungs and breast cancer etc. that data mining is the process of digging inside data for the purpose of discovering latent patterns which can be translated into useful information. The various data mining techniques which has been used in medical data are Apriori Algorithm, FP Growth, unsupervised neural networks, Decision trees, linear programming, Association rule mining. The latter (i.e Association rule mining) is used in discovering frequently occurring items in the given dataset. In his research the Apriori Algorithm is used for discovering frequent patterns for finding locally frequent disease and also

visualizing the experimental results in various view points. Medical mining yields required business intelligence to support well informed diagnosis and decisions making.

Sona et al (2014) affirmed that the Healthcare database consists of different records of patient information from the hospitals. When these data are mined properly, it reveals hidden patterns which can lead to better diagnosis of different diseases and also helps in medicines treatment. According to Jacquelin et al (2013), applying data mining on medical institution’s existing data helps to discover new, useful and potentially life- saving knowledge which would have otherwise remained hidden in their databases. He also proposed a system which will help in detecting discrepancies in medical data. The data mined is cleaned to remove unwanted, inconsistent, noisy and corrupt data thereby making the data to be of high quality data which is potentially more accurate and also prevents occurrences of fraud in case of false health sale and billings. Data mining can also help in detecting medical fraud such as providing false and intentionally misleading statements to patients, falsify medical records or reports, unnecessary medical treatment or drug prescription and also detecting false medicine composition. Data mining has the ability to generate a knowledge rich environment which can help to significantly improve the quality of clinical decisions.

Srinivas, Rao and Govardhan (2012) affirmed that data mining aims at discovering knowledge out of data and presenting it in a form that is easily compressible to humans. It is a process that is developed to examine large amounts of data routinely collected. Similarly, Rani (2011) affirmed that data mining is an interdisciplinary field and its main aim is to uncover relationship in data and to predict outcomes. Diverse data mining problems can be solved adequately by soft computing techniques. These techniques are Clustering, Decision tree, fuzzy logic, neural network, genetic algorithms and rough sets which will lead to an intelligent, interpretable, low cost solution than traditional techniques.

Association Rule Mining in Health Care

Association Rule is one of the most important techniques of data mining that is used to detect the frequent patterns or rules, interesting relationships among a set of data items in the data repository. It is also known as market basket analysis due to its ability to discover the association among purchased item or unknown patterns of sales of customers in a transaction database. Association rule also plays a vital role in the healthcare field to detect the relationships among diseases, health state and symptoms. Association Rule is widely used by Healthcare organization for discovering relationships between various diseases and drugs. It is also used in detecting fraud and abuse in health insurance. It is also used with classification techniques to enhance the analysis capability of Data Mining.

Han et al (2006) proposed in his research that Association rule mining process could be divided into two main phases to enhance the implementation of the algorithm.
The phases are:

1. **Frequent Item Generation:** This is to find all the itemsets that satisfy the minimum support threshold. The itemsets are called frequent itemsets.
2. **Rule Generation:** This is to extract all the high confidence rules from the frequent itemsets found in the first step. These rules are called strong rules.

Reasons for using Support count and Confidence Rule

Support count is an important measure because a rule that has very low support may occur simply by chance. A rule with a low support count is likely to be uninteresting from a business perspective because it may not be profitable to promote items that are frequent together. Therefore, support count is often used to eliminate uninteresting rules. It also has a desirable property that can be exploited for the efficient discovery of association rules.

Confidence Rule on the other hand, measures the reliability of the inference made by a rule. For a given $A \rightarrow B$, the higher the confidence, the more likely it is for B to be present in transactions that contain A. confidence also proves an estimate of the conditional probability of B given A.

Importance and Uses of Data Mining in Medicine

Data Overload: The health record is rich of knowledge, but the data is so huge that it seems that one is being drowned inside the data and is actually starving for knowledge.

Evidence based medicine and prevention of hospital errors: Medical institutions that applies data mining on their existing data discover new, useful and potentially life saving knowledge that would have remain lifeless in the database if not discover.

Policy making in public health: GIS and data mining combined tools like (free, open source, java- based data mining tools), to analyze similarities between different aspect of the medical record.

Objectives of the Study

The objectives of this study are:

1. To study the healthcare data i.e. the various occurrences of diseases and the symptoms that comes with it;
2. To enhance the system used for diagnosis for better performance of medical personnel in the diagnosis of disease; and
3. To evaluate the efficacy of the system developed.

Misdiagnosis has been said to occur due to cognitive error, that is, failure to consider all likely diagnoses. With the increase in biomedical knowledge it is humanly impossible for a physician to remember all the various symptoms associated with all diagnostic possibilities. Diagnosis errors are usually errors of omission which are difficult to identify unlike most medical errors which are usually errors of commission. Sometimes misdiagnosis cases surface through malpractice litigation, morbidity and mortality conferences and uncertain feedbacks from patients and physicians, (Kenneth et al, 2013).

Misdiagnosis is a serious and common occurrence in the healthcare in which the repercussions can damage a patient's health and cost money, or even a life. Not all misdiagnosis are due to physician's negligence, skillful doctors can and do make diagnosis error even when using reasonable care.

In the Healthcare, many facts can be achieved from the past data stored in computers, since the data are huge in quantity it is very strenuous for an individual to examine the whole data and gain awareness through the traditional method of viewing files. Therefore, this research is applying a Market- basket analysis technique which is a technique in data mining into medical records to identify association between diseases and symptoms in order to assist in solving the problem of misdiagnosis.

The aim of this study was to design a system to assist the medical practitioners in the diagnosis of preventable disease. This was done by using market- basket analysis technique with the bitmap based algorithms which would help to analyze different diseases and determine patterns in the symptoms that always occurred together.

Methodology

The nature of data used is based on diseases and symptoms. The information is gathered under supervision of expertise doctors and from the patients records. The data used are based on different patients suffering from ten (10) specified diseases which are:

1. Influenza
2. Measles
3. Malaria
4. Typhoid
5. Cholera
6. Typhus
7. Yellow fever
8. Pneumonia
9. Gastroenteritis
10. Dengue fever

The data contain hundred (100) different cases of each diseases and their symptoms.

The Healthcare data consists of a set of patients' records or data such as the name, sex, address and some other categorical information. The diverse information contained in the patients record might not be suitable for data mining. So therefore, there is need for data preparation before the data mining process can begins. The data preparation process in this study includes: data cleaning and data transformation.

Data Management: the data gotten are grouped according to their types for easy accessed. The patients record used are stored differently due to their transaction type (i.e. disease classification). Ten (10) different diseases are used for this research and for each disease one hundred (100) different cases are chosen.

Data Mining: The association rule mining algorithm is applied on the medical data sets generated. Here the Basket analysis technique is applied to be able to identify the relationships between data. It detects significant and concise medical association rules to be used in medical diagnosis.

Association Rule mining technique works in two main steps:

1. It searches through the items to discover the ones having frequency count which surpass the user-defined support count. This set is referred to as Frequent itemset that is a set of items that has a support higher than a user-supplied minimum support.
2. It discover item sets (or patterns or rules) meeting the second user-defined frequency count. This are item set meeting the specified confidence level. The item set generated at this step are the final rules generated.

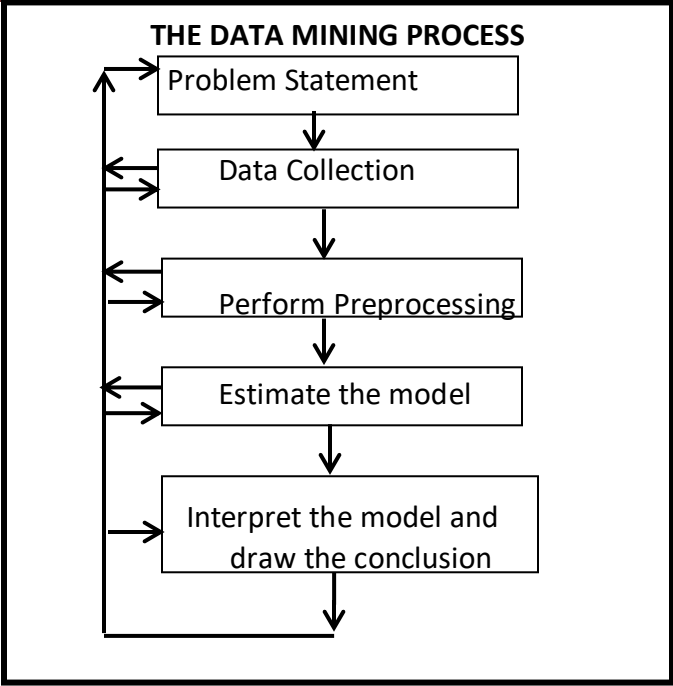


Fig. 1: The data mining Process

Apriori Algorithm

The main aspect of the Apriori algorithm which is the basis of this analysis is grounded on the rule of identifying associations between items in a transactional database by applying the rule of sufficient repeat occurrence of a particular item in transaction data set which is enough to sustain the minimum support count. The Algorithm works as follow:

Apriori Generate Algorithm

C_k : Candidate item set of size k
 L_k : Frequent item set of size k
 $L_1 = \{\text{Frequent items}\};$
For $(k= 1; L_k \neq \emptyset; K++)$ do begin
 C_{k+1} = candidates generated from L_k ;
 For each transaction t in database do increment the
 count of all candidates in C_{k-1} that are contained in t
 L_{k+1} = candidates in C_{k+1} with min_support
End
return $\bigcup_k L_k$;

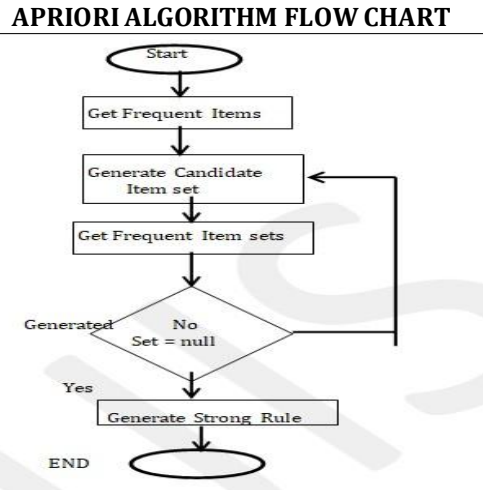


Fig 2: Apriori Flow chart

- STEP 1: The Algorithm first check through the transaction dataset to detect the number of items in singularity and the frequency of occurrence for the item.
- STEP 2: The Items are identified and the first candidate sets are made and frequency counts are also list down
- STEP 3: The candidate set is then checked with the minimum support count provided in order to identify the Items with the required support count.
- STEP 4: The items that fall below the requirements are eliminated, while the remaining items that meet up the requirement are process again to create another candidate set that will check their occurrence in different group (e.g. group of two, three and so on).
- STEP 5: Each time a candidate set is created it is check with the provided support count to see if its surpass the minimum threshold value and if it does not then the process of generating new candidate set is terminated
- STEP 6: The Association Rule generated from the Item set is then compared to the threshold value of the particular transactional Dataset to identify the Strong and Weak association rules.

Technique used for generating the Rules

Market Basket Analysis : This analysis method makes use of the horizontal database (i.e. each transaction has its items listed out in rows). The itemset are represented in a bitmap format (i.e. each item has a bit position assign to it to indicate the presence and absence of the items (“1” or “0”) respectively. Items are therefore generated from the bitmap by performing logical AND operation on each column (itemset) and the support count of each column is calculated wherefore any column less than the given support count is dropped or discarded. This analysis method make use of algorithm which only scan the database once unlike the Aprior algorithm which takes more time in finding the frequent itemsets, because it needs to scan the database again and again which makes it to be a time consuming process.

The basic concepts of data mining can be explained using a supermarket data as an example.

No.	Items purchased
001	MILK, TEA, SUGAR, CORNFLAKES
002	MILK, CORNFLAKES, SUGAR
003	MILK, TEA, SUGAR
004	MILK, TEA, BREAD, CORNFLAKES
005	MILK, TEA, BREAD, CUSTARD

Table 1: a small set of transaction

The above table is used to show the transaction performed by five different customers. Each row of the table is referred to as a transaction. This table will be used in illustrating the concepts involved in data mining.

1. An **Association rule** is a pair of disjoint itemsets. If A and B denotes the two disjoint itemsets, the association rule is written as $A \rightarrow B$
2. The **support count** of the association rule $A \rightarrow B$ with respect to a transaction set “C” is the support of the itemsets $A \cup B$ with respect to C.
3. The **confidence** of the rule $A \rightarrow B$ with respect to a transaction set “C” is the ratio $\text{support}(A \rightarrow B) / \text{support}(C)$.

The given association rules $A \rightarrow B$ is consider significant if it has a high support count and high confidence. For a stated the support count, the rules with higher confidence values are more significant than those with lower confidence values. It can be seen from the given definition above that the confidence value is at most 1, therefore they are measured in percentage rather than ratio.

This analysis method is implemented in Java language. This algorithm divides the database into different units and after which the results of the different processing units are combine to produce a single results. In medical analysis, market basket analysis is applied on the data separately in order to get accurate results.

A preprocessing step is used to convert the raw data into a form similar to market basket data. According to the market basket data discussed above, the set of all the symptoms and disease diagnosis codes referred to the set of items. The set of the symptoms and disease codes for each patient is considered to be a transaction. In other words, each itemset constitute of one or more disease codes. Therefore, the support count of an itemset C is the number of patients (transaction) whose set of items corresponds to the items in C. The major goal of this research is to discover the relationships between diagnosed diseases and

their corresponding symptoms. A natural way of discovering this rule is finding the association rules of the form $A \rightarrow B$, where A is a set of symptoms codes and B is a set of disease diagnosis code. For this rule to be significant it must constitute a high support count and high confidence rule.

System Requirements

The system Requirement specifies the hardware and software requirements for carrying out this project. The following are the hardware and the software requirements.

- Hardware Requirements
 - a. Processor: Intel(R) core(TM) i7- 2640M CPU @ 2.80 GHz
 - b. RAM: 4GB
 - c. System type: 64-bit Operating S
- Software Requirement
 - The researcher made use of
 - a. Windows 7
 - b. 64- bit Operating System
 - c. Microsoft Office (Windows Excel)
 - d. Java Net bean version 8.02

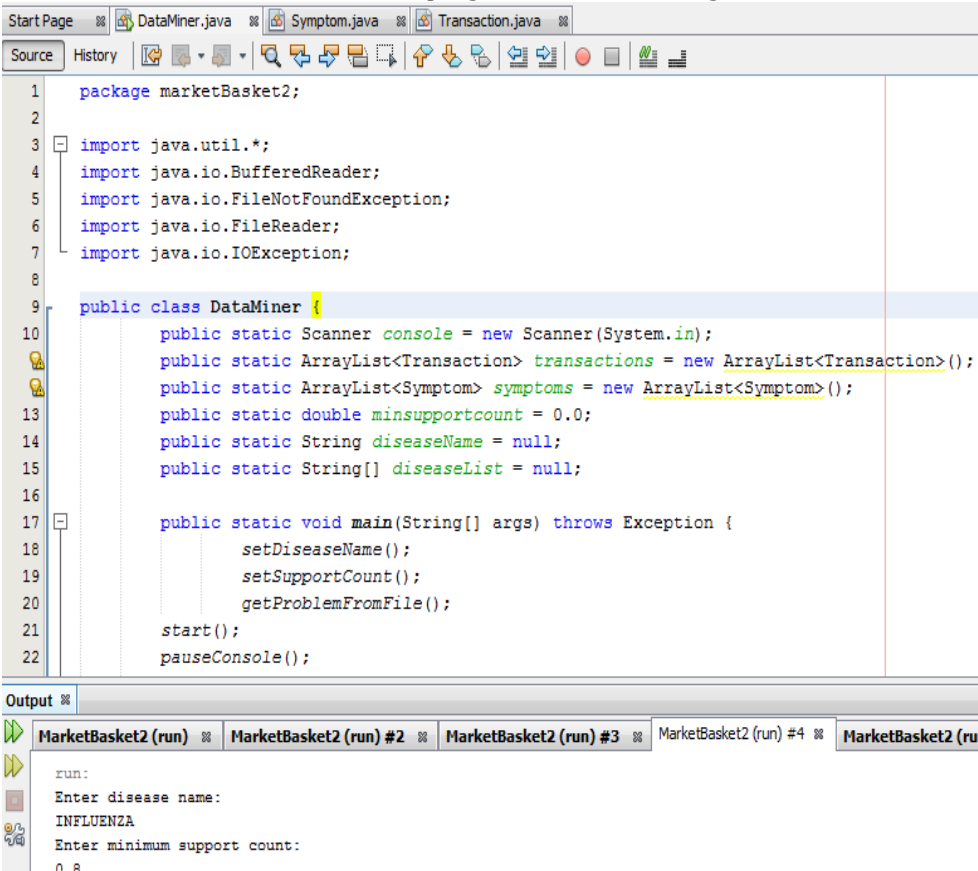
Results and Discussion

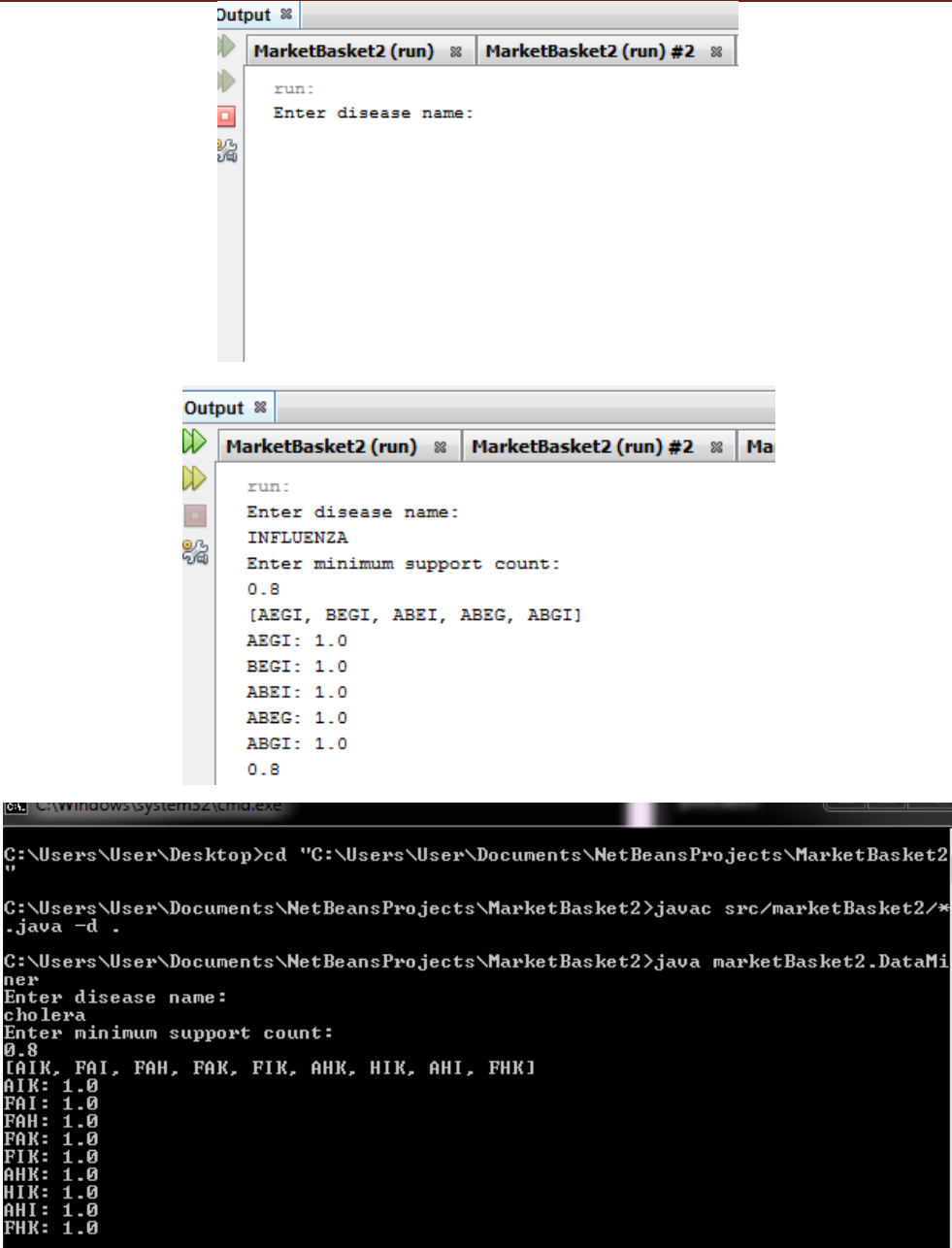
The data is represented in two formats; the generated format and the transformed format

- The generated format is the original format in which the data is generated.
- The transformed format is the transaction format (that is the bitmap mode) which makes it suitable for discovering the frequent items in them.

The Program is designed in such a way that, there is an enter items command which will request for the disease that the physician wants to get information about with the minimum support and minimum confidence threshold for the generation of the rules.

The below shows the screenshots of the program and the results generated:





Original Format

One hundred cases of patients suffering from the selected diseases are gotten with the different symptoms which each of them is suffering from.

NOTE: The digits in the symptoms column each represents a particular symptoms which the patients is suffering from.

Discussion

The above tables show the different results given by the system after which the irrelevant rules are filtered out. After running the program on the disease table with the support count given that is 0.8 which is equivalent to 80% the confidence level is set to be able to ascertain the rules which meet up with the set count. The rules are then validated with the medical personnel in order to establish the relevant ones.

Most of the rules in the result indicated a justify correspondence between a set of symptoms and disease diagnosis. According to doctor’s prescription, the above rules are 80% acceptable in the healthcare system and can assist the medical personnel in diagnosis. It will also reduce stress on medical personnel and also assist patients to get adequate attention and care.

The results obtained also shows the list of the unaccepted rules. However, the doctors contacted states that this rules cannot be accepted because some important symptoms are missing which can be as a result of the data collected. However, it is possible for two different diseases to have more than one similar symptom, but their degree of intensity may differ for example in the case of malaria and typhoid fever, headache is a major symptom, it occurs once in a while and mostly slightly in malaria. But in the case of typhoid, the intensity of headache is severe. In other to ascertain these two diseases; each symptom should be assigned a weighing factor which has value between zero and one. It is also advice that in the case of these two diseases; blood test can be carried out to have adequate result.

For all its peril, medical data mining can be the most rewarding. This is due to the fact that Thousands of data elements can be gathered in order to find solutions to some scientific questions. In the process of finding an answer to a medical problem, a life can be extended and comfort can be given to an ill person.

Summary

Owning to the danger surrounding misdiagnosis in Healthcare System, as medical diagnosis is essential and also delicate task that if not carried out effectively and carefully can lead to death. Therefore a system needs to be developed to safeguard against the inaccurate prescription given to patients which sometimes leads to misdiagnosis.

In data mining, association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases. There is a lot of hidden data available within the healthcare system which makes the health care environment information rich but knowledge poor. There are many algorithms available for generating association rules over time, and one of the techniques of association rule mining is Market- basket analysis which is meant for identifying association between items in a transactional database.

The research work has study the manual operation of diagnosis and treatments of some selected diseases affecting mankind. It explained the general diagnosis procedures and techniques involves.

In this research, the analysis technique (Market- basket analysis) is applied on medical data (i.e. Symptoms and diseases) in other to detect the association between a group of symptoms and a disease. Each patient's data is referred to as a transaction. The rules generated in the research provide a method of measuring joint frequencies for different symptoms and the corresponding diseases.

Recommendations

The result generated in this research warrant further application in the Health System and also a data mining technique for medical informatics can be developed. Further research may be carried out on how to use some knowledge of the underlying domain to decide the types of association rules which benefit the medical personnel. A research can also be carried out on a set of restricted data which includes a specific group of diseases (Such as Cancer) and their different diagnosis procedures and treatments using data mining technique. This type of study can provide useful information concerning the effectiveness of a group of method in diagnosing a particular disease in the data and a set of treatments which is very accurate in curing the diseases.

This market basket technique can also be applied in detecting fraudulent behavior among physicians and patients, such as Association of medical personnel with frequent prescription of expensive and un-intruding diagnosis of patients. This study serves as advancement toward the technological development of our medical services.

Conclusion

Since the ability of computer system is known; one can conclude on the best method which favors the diagnosis of diseases in healthcare. The system will support and supplement human judgment, that is the physician, by reproducing the logical usage of data to the skilled observed. And also make diagnosis very fast and accurate. Doctors and Nurses should be encouraged to learn how to use computer so as to ease the work of system developer.

REFERENCES

1. Boris Milovic and Milan Milovic. (2012). Prediction and Decision Making in Health Care Using Data Mining. Kuwait Chapter of Arabian Journal of Business and Management Review, 1(12), 126-136.
2. Fawzi Elias Bekri, Govardhan A. (2011). Association of Data Mining and Healthcare Domain: Issues and Current State of the Art. Global Journal of Computer Science and Technology, 11(21), 1- 9.

3. Jackson, J. (2002). Data Mining: A Conceptual Overview. Communications of the Association for Information Systems, 8, 267-296.
4. Kenneth H. Falchuk, Evan Falchuk (2013): The Misdiagnosis epidemic; Five root causes and the growing demand for more patient- centric care. International Journal for Healthcare Manangement. Volume 5, Issue 1, Pages 61-65.
5. Mohammed Al-Maolegi and Bassam Arkok. (2014). An Improved Apriori Algorithm for Association Rules. International Journal on ANtural language Computing, 3(1), 21-29.
6. Rani, K. U. (2011). Analysis Of Heart Diseases Dataset Using Neural Network Approach. International Journal of Data Mining and Knowledge Management Process, 1-8.
7. Sona Baby and Ariya T.K. (2014). A survey paper of data mining in medical diagnosis. International Journal of Research in Computer and Communication Technology, 98- 101.
8. Srinivas, K., Rao, G. R., & Govardhan, A. (2012). Mining Association rules from Large Datasets Towards Disease Prediction. 2012 International Conference on Information and Computer Networks, 27, 22- 26.