A Knowledge-Based Data Mining System for Diagnosing Malaria Related Cases in Healthcare Management

Olugbenga Oluwagbemi 1*, Uzoamaka Ofoezie 2, Nwinyi, Obinna 3

1Rochester Institute of Technology, 28 Lomb Memorial Drive, Rochester NY 14623, Rochester, New York, USA
2(Bioinformatics Unit) Department of Computer and Information Sciences
School of Natural and Applied Sciences
College of Science and Technology, Covenant University, Ogun State, Nigeria.

3Department of Biological sciences, School of Natural and Applied Sciences
College of Science and Technology, Covenant University, Ogun State, Nigeria.

gbemiseun@yahoo.com

Abstract

Data mining a process for assembling and analyzing data into useful information can be applied as rapid measures for malaria diagnosis. In this research work we implemented (knowledge-base) inference engine that will help in mining sample patient records to discover interesting relationships in malaria related cases. The computer programming language employed was the C#.NET programming language and Microsoft SQL Server 2005 served as the Relational Database Management System (RDBMS). The results obtained showed that knowledge-based data mining system was able to successfully mine out and diagnose possible diseases corresponding to the selected symptoms entered as query. With this finding, we believe the development of a Knowledge-based data mining system will not only be beneficial towards the diagnosis of malaria related cases in a more cost effective means but will assist in crucial decision making and new policy formulation in the malaria endemic regions.

Keywords: diagnosis, data mining, malaria

1. Introduction

Data mining as a process for analyzing data from different perspectives and summarizing it into useful information can generate information that can be used to increase revenue, cut costs, or both [1]. Data mining identifies trends within data that go beyond simple analysis. Through the use of sophisticated algorithms, non-statistician users have the ability to identify key attributes of business processes and target opportunities [2]. Data mining refers to extracting or “mining” knowledge from large amounts of data [3]. In some hospitals in Nigeria, it is difficult to select or extract very important information from the database because it is practically cumbersome and this makes it increasingly difficult to support decision making and also to detect abnormal patterns of disease enlistments. Traditionally, analysts have performed the task of extracting useful information from recorded data. But, the increasing volume of data in modern business and science most especially the health sector
calls for computer-based approaches. Data mining is the process of applying computer-based methodology, including new techniques for knowledge discovery, from data.

The aim of this research work is to implement an algorithm that will help in mining large patient records to discover interesting relationships in malaria related cases, which will assist in crucial decision making and new policy formulation. This aim will be achieved through the following objectives; (i) assist a doctor in detecting abnormal patterns in the data base of the patients’ records, (ii) Increase time efficiency of physicians by reducing time spent on documentation and time spent sorting through the large database during and after shifts, (iii) capture comprehensive data on patient for clinical research, (iv) to provide an insight into the symptom patterns among patients of the hospital, (v) to provide the Hospital Authorities with data mining solutions and services to allow them understand health related behaviors and the pathologies encountered.

2. Relevant work

Several research efforts have been made on the control of malaria. A recent study discussed how information or data is being targeted efficiently and effectively to control and justifies investment in research of malaria [4]. Data mining operations have recently been performed on the malaria causing parasites “Plasmodium falciparum” to define organellar function [5]. In another work, a systematic database has been used to identify anti-malaria drug treatment trials for patients [6]. The availability of reliable health data are widely recognized as essential for efforts to strengthen health-care systems in resource-poor settings worldwide; a local space-time kriging approach was applied [7]. It was observed in another study that reliable and timely information on disease-specific treatment burdens within a health system is critical for the planning and monitoring of service provision; a modeling approach was applied [8]. Recently, a new computational approach for mining the malaria transcriptome was highlighted in a study [9]. Another study revealed the application of mining for traditional Chinese medical knowledge discovery [10]. A research work was recently carried out on assembling a global database of malaria parasite prevalence for the Malaria Atlas Project. Results obtained then revealed the Malaria Atlas Project database held 3,036 Plasmodium falciparum and 1,347 Plasmodium vivax estimates in 74 countries. The database was seen to become a foundation for the development of global malaria endemicity models [11]. In the field of Bioinformatics, a data mining system was recently applied for predicting vertebrate genes in genomic DNA sequences [12]. Finally, a computer-based approach was applied to malaria research [13].

3. Methodology

3.1 Data gathering, mining and knowledge discovery

Data collection was obtained by survey from about four hospitals within the Lagos metropolis in Nigeria. Data cleaning was performed on the data collated to remove noise and inconsistent data. Second, data integration was performed in cases where multiple data sources were combined. Third, data relevant to the analysis task were retrieved and selected
for processing. Data transformation was the fourth step, where data were transformed or consolidated into forms appropriate for mining. Data mining was the fifth process carried out. Pattern evaluation was the sixth step, as can be seen from Fig. 6, our knowledge-based data mining system was able to identify truly interesting patterns representing knowledge based on some interestingness measures.

Knowledge presentation was the seventh step where visualization and knowledge representation techniques were used to present the mined knowledge to the user.

Data obtained was thus subjected to thorough analysis in mining interesting disease patterns which will ultimately assist for proper decision making in the health sector based on different data of patients with different cases of malaria and other kind of diseases.

3.2 System architecture

![System Architecture for the Knowledge Based Data Mining System](image)

**Figure 1.**

3.3 Components of the Knowledge based data mining system architecture

(i) **Knowledge Base:** Knowledge Base is a collection of fact (data) stored in such a way that new knowledge can be inferred. It makes room for inferences and enables reasoning.

(ii) **Inference Engine:** An inference engine is a computer program that tries to derive answers from a knowledge base. It is the "brain" that expert systems use to reason about the information in the knowledge base for the ultimate purpose of formulating new conclusions. Inference engines are considered to be a special case of reasoning engines, which can use more general methods of reasoning. The inference engine actually reasons based on the rules and data present within the system. The result of such reasoning leads to the decision of the system.
• **Rules:** Relationships between records follow certain rules. The enforcement of rules in a database requires that database modification operations must pass the rules before they can be allowed to proceed.

Structured query language (SQL) provides three modification operations namely: insert, delete, and update. Rules are in different categories; some are simple and apply in a similar way to many records within a database, while others are more complex and set exclusive to certain tables and to particular relationships between records. Declarative referential integrity can be used in a relational database to implement many simple and repetitive rules [14]. This is implemented as simple declarations in a SQL create table statement. Procedural referential integrity enforcement, on the other hand, requires more complex and exclusive rules. This however, involves the use of a database programming language, unique to each Database management systems (DBMS). Programs written in such language to enforce such rules are referred to as triggers.

• **Decisions:** They are the decisions made in building the system and are also applied within the system for desired input and output which assists in making decisions based on the quality of data.

(iii) **Application/User:** An application is a system built to enable users to be able to achieve desired goals as applied to the system.

(iv) **Database:** A database implies a persistent and integrated storage allowing concurrent access to it by many users. It is a collection of records related by referential integrity. Thus, a database is an organized collection of structured data, to serve many applications with minimum redundancy. Database technology actually helps to alleviate many of the problems associated with conventional file organization methods. These include: data duplication, inflexibility and difficulties associated with accessing files by on-line users. Characteristics of a database are as follows:

(i) Data integrity: helps to guarantee the integrity of data within the database which helps to avoid data redundancy.

(ii) Flexibility: data stored within databases can serve the requirements of many users and applications in a variety of ways. This depicts a dynamic nature of such data with respect to these characteristics.

(iii) Data independence: data within databases are not used for one specific purpose, but for general purposes and such data is independent of the application that uses it.

For the experiment performed in this research, moderate volumes of data were used based on the hardware facilities available for now. Files containing the data for initial testing were in the range of 26Kb to 1Mb. The database contains the object tables such as the Patient Information table (with 10 Fields: Medical Identification Number, Patient Identification Number, Patient Name, Address, Date of birth, Sex, Status, Treatment_ID, Physician_ID, Date of Visit), Next of kin table (5 Fields: Patient ID, Name, Relationship, Phone Number and Address), diagnosis table (7 Fields: Diagnosis_ID, Symptoms, Patient_ID, Date of Visit, Physician_ID, Treatment_ID) and the treatment table (4 attributes: Treatment_ID, Date of
Visit, Physician_ID, Treatment Notes). These tables contain their respective corresponding data.

3.4 System design

The system design defines the structure of the system. It involves the analysis and determination of data processing requirements of a company and the designing of systems to fulfill them. The system consists of a form and a database. The form enables an administrator to enter transactions into the database, update or modify the records and view the records in the database. An administrator can also view frequent items in the database and the rules generated from these frequent item sets. This would be achieved by creating a form using the Microsoft Visual C# and connecting it to a database created with Microsoft SQL Server.

3.5 System modeling

System modeling was presented through the class diagram and the use case diagrams. Class diagram depicts the system’s object structure. They show object classes that the system is composed of as well as the relationship between those object classes, while the use case graphically depicts the interactions between the internal parts of the system, external system and the users.(as shown in Figure 2 and 3).

Figure 2. Class diagram for the knowledge-based system
4. Implementation

Implementation is a realization of a technical specification or algorithm as a program, software component, or other computer system. Implementation involves coordinating the user department and the data processing department in getting the new system into operation. The implementation process involved programming, system specification and testing. Implementation tools are:

(i) C#.NET programming language was used to develop the system. The software was implemented using C# because it has extra features. It inherits some of the best features of C++, Microsoft Visual Basic and the development features provided by Microsoft Visual Studio 2008.

(ii) Microsoft SQL Server 2005 is the Relational Database Management System (RDBMS) used because it is a robust RDBMS that can manage the large amount of data and can be integrated with our development platform. It also has tools that facilitate the design and implementation of databases.

Choice of C# as the Programming language in use

Microsoft C# programming language was used for the implementation of this application. It was used due to the following reasons among others: It has many in-built functions that makes programming easier, it is a flexible high-level programming language, it has object oriented programming features, and it has a cross platform programming capability.
5. Results & Discussion

Figure 4: The Login page is the page that allows access to medical officers (doctors and nurses). Each user has a unique password and are granted access into this application.

Figure 5 shows the Patient information form. This form is used by medical doctors and nurses to enter to key-in patients’ data into the database. Such data include: the Patient Identification Number (Patient_ID), the medical record number, Address, Date of Birth, Gender, Treatment Identification Number (Treatment_ID), Physician Identification Number (Physician_ID), Date of Visit, and Status.

The Diagnosis form and the Result form was depicted in Figure 6. This form serves as a user interface for entering patient’s diagnosis data into the database by the physician and the inference engine within the system makes the result of likely diseases affecting the patient available. After data mining had been performed, our software was able to identify truly interesting patterns. Here, malaria and typhoid were truly interesting patterns diagnosed based on the rules acted upon by the inference engine.

Figure 7. The Treatment Note form serves as an interface for the medical personnel to enter the possible treatments and prescriptions that a patient need.
Figure 5. Patient Information Form

Figure 6. Diagnosis Form

Figure 7. Treatment Note Form
Conclusion

From the study, knowledge-based data mining system in Healthcare management hospitals presents as one of the best applications for deriving meaningful diagnosis malaria related cases/ other human health challenges. From our study, this application serves as a model tool that will enable hospitals to effectively monitor patients’ medical records without ambiguity. This will provide a great reduction in the man hours wasted in most conventional hospitals without this tool. It is our believe that for the attainment of the millennium development goals (MDG) in terms of healthy living for the masses adoption of this knowledge based data mining system must be a priority in the hospitals of countries with such an aspiration for its masses.

References


-9-


