A Survey on Data Mining Approaches in Medicine

Vahid Rafe¹, Roghayeh Hashemi Farhoud²

¹. Department of Computer Engineering, Faculty of Engineering, Arak University, Arak, Iran.
². Department of Computer Engineering, Hamedan University of Technology, Hamedan, Iran.

Corresponding author email: rhff90@yahoo.com, v-rafe@araku.ac.ir

ABSTRACT: Widespread use of medical information systems and explosive growth of medical databases require traditional manual data analysis to be coupled with methods for efficient computer assisted analysis. This paper presents data mining survey, selected data mining techniques that can be applied in medicine, method validation techniques, also application of data mining in medicine.

Key words: Data mining in medicine, KDD, data mining techniques in medicine, application data mining in medicine.

INTRODUCTION

Medical data mining has great potential for exploring the hidden patterns in the data sets of the medical domain. These patterns can be utilized for clinical diagnosis. However, the available raw medical data are widely distributed, heterogeneous in nature, and voluminous. These data need to be collected in an organized form. This collected data can be then integrated to form a hospital information system. Data mining technology provides a user oriented approach to novel and hidden patterns in the data. Medical diagnosis is regarded as an important yet complicated task that needs to be executed accurately and efficiently. The automation of this system would be extremely advantageous.

Regrettably all doctors do not possess expertise in every subspecialty and moreover there is a shortage of resource person at certain places. Therefore, an automatic medical diagnosis system would probably be exceedingly beneficial by bringing all of them together. Appropriate computer-based information and/or decision support systems can aid in achieving clinical tests at a reduced cost. Efficient and accurate implementation of automated systems needs a comparative study of various techniques available. (Soni J, Ansari U. 2011) Knowledge discovery in databases is well-defined as a process consisting of several distinct steps. Data mining is the core step, which results in the discovery of hidden but useful knowledge from massive databases. A formal definition of Knowledge discovery in databases is given as follows: “Data mining is the non-trivial extraction of implicit previously unknown and potentially useful information from data” (Frawley J, Piatetsky-Shapiro. 1996). The discovered knowledge can be used by the healthcare administrators to improve the quality of service. The discovered knowledge can also be used by the medical practitioners to reduce the number of adverse drug effects, to suggest less expensive therapeutically equivalent alternatives. Anticipating patient’s future behavior on the given history is one of the important applications of data mining techniques that can be used in health care management.

Hospitals must also minimize the cost of clinical tests. They can achieve these results by employing appropriate computer-based information and/or decision support systems. Health care data is massive. It includes patient-centric data, resource management data and transformed data. Health care organizations must have the ability to analyze data. Treatment records of millions of patients can be stored and computerized and data mining techniques may help in answering several important and critical questions related to health care. The availability of integrated information via the huge patient repositories, there is a shift in the perception of clinicians, patients and payers from qualitative visualization of clinical data by demanding a more quantitative assessment of information with the supporting of all clinical and imaging data. For instance it might now be possible for the physicians to compare diagnostic information of various patients with identical conditions. Likewise, physicians can also confirm their findings with the conformity of other physicians dealing with an identical case from all over the world (Miller A, Blott B, Hames T. 1992). Clinical decisions are often made based on doctor’s intuition and experience rather than on the knowledge rich data hidden in the database. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. Wu, et al proposed that integration of clinical decision support with computer based patient records could reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome. (Chen J, Greiner R. 1999) In particular, data mining may accomplish class description, association, classification, clustering, prediction and time series analysis. (Han J, Kamber. 2006).
**Knowledge discovery and data mining**

The terms Knowledge Discovery in Databases (KDD) and Data Mining are often used interchangeably. KDD is the process of turning the low-level data into high-level knowledge. Hence, KDD refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and KDD are often treated as equivalent words but in real data mining is an important step in the KDD process. The following fig. 1 shows data mining as a step in an iterative knowledge discovery process.

![Diagram of the process of knowledge discovery](image)

**Figure 1.** Data mining as a step in the process of knowledge discovery.

Knowledge discovery as a process is depicted in Figure 1 and consists of an iterative sequence of the following steps:

1. Data cleaning (to remove noise and inconsistent data).
2. Data integration (where multiple data sources may be combined).
3. Data selection (where data relevant to the analysis task are retrieved from the database).
4. Data transformation (where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance).
5. Data mining (an essential process where intelligent methods are applied in order to extract data patterns).
6. Pattern evaluation (to identify the truly interesting patterns representing knowledge based on some interestingness measures).
7. Knowledge presentation (where visualization and knowledge representation techniques are used to present the mined knowledge to the user)

Steps 1 to 4 are different forms of data preprocessing, where the data are prepared for mining. The data mining step may interact with the user or a knowledge base. The interesting patterns are presented to the user and may be stored as new knowledge in the knowledge base. Note that according to this view, data mining is only one step in the entire process, albeit an essential one because it uncovers hidden patterns for evaluation. (Han j, Kamber. 2006).

**Data Mining Process**

In the KDD process, the data mining methods are for extracting patterns from data. The patterns that can be discovered depend upon the data mining tasks applied. Generally, there are two types of data mining tasks: Descriptive data mining tasks that describe the general properties of the existing data, and predictive data mining tasks that attempt to do predictions based on available data. Data mining can be done on data which are in Quantitative, textual, or multimedia forms.

Data mining applications can use different kinds of parameters to examine the data. They include association (patterns where one event is connected to another event), sequence or path analysis (patterns where one event leads to another event), classification (identification of new patterns with predefined targets) and clustering (grouping of identical or similar objects). Data mining involves some of the following key steps: (the Data Mining Process.[Online]).

1. **Problem definition:** The first step is to identify goals. Based on the defined goal, the correct series of tools can be applied to the data to build the corresponding behavioral model.
2. **Data exploration:** If the quality of data is not suitable for an accurate model then recommendations on future data collection and storage strategies can be made at this. For analysis, all data needs to be consolidated so that it can be treated consistently.
3. **Data preparation:** The purpose of this step is to clean and transform the data so that missing and invalid values are treated and all known valid values are made consistent for more robust analysis.
4. **Modeling:** Based on the data and the desired outcomes, a data mining algorithm or combination of algorithms is selected for analysis. These algorithms include classical techniques such as statistics, neighborhoods and clustering but also next generation techniques such as decision trees, networks and rule based algorithms. The specific algorithm is selected based on the particular objective to be achieved and the quality of the data to be analyzed.
(5) Evaluation and Deployment: Based on the results of the data mining algorithms, an analysis is conducted to determine key conclusions from the analysis and create a series of recommendations for consideration.

**Selected data mining techniques in medicine**

There are various data mining techniques available with their suitability dependent on the domain application. Statistics provide a strong fundamental background for quantification and evaluation of results. However, algorithms based on statistics need to be modified and scaled before they are applied to data mining. We now describe a few Classification data mining techniques in medicine.

**Decision Trees**

It is a knowledge representation structure consisting of nodes and branches organized in the form of a tree such that, every internal non-leaf node is labeled with values of the attributes. The branches coming out from an internal node are labeled with values of the attributes in that node. Every node is labeled with a class (a value of the goal attribute). Tree-based models which include classification and regression trees, are the common implementation of induction modeling[5]. Decision tree models are best suited for data mining. They are inexpensive to construct, easy to interpret, easy to integrate with database systems, and they have comparable or better accuracy in many applications. There are many Decision tree algorithms such as HUNTS algorithm (this is one of the earliest algorithms), CART, ID3, C5.0 (a later version of ID3 algorithm). The decision tree shown in Fig. 3 is built from the very small training set (Table 1). In this table each row corresponds to a patient record. We will refer to a row as a data instance. The data set contains three predictor attributes, namely Age, Gender, Intensity of symptoms and one goal attribute, namely disease whose values (to be predicted from symptoms) indicates whether the corresponding patient have a certain disease or not.

![Decision Tree Diagram](image)

**Figure 3. A decision tree built from the data in Table 1**

<table>
<thead>
<tr>
<th>Age</th>
<th>Gender</th>
<th>Intensity of symptoms</th>
<th>Disease (goal)</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>Male</td>
<td>Medium</td>
<td>Yes</td>
</tr>
<tr>
<td>32</td>
<td>Male</td>
<td>High</td>
<td>Yes</td>
</tr>
<tr>
<td>24</td>
<td>Female</td>
<td>Medium</td>
<td>Yes</td>
</tr>
<tr>
<td>44</td>
<td>Female</td>
<td>High</td>
<td>Yes</td>
</tr>
<tr>
<td>30</td>
<td>Female</td>
<td>low</td>
<td>No</td>
</tr>
<tr>
<td>21</td>
<td>Male</td>
<td>low</td>
<td>No</td>
</tr>
<tr>
<td>18</td>
<td>Female</td>
<td>low</td>
<td>No</td>
</tr>
<tr>
<td>34</td>
<td>Male</td>
<td>Medium</td>
<td>No</td>
</tr>
<tr>
<td>55</td>
<td>Male</td>
<td>Medium</td>
<td>No</td>
</tr>
</tbody>
</table>
Decision tree can be used to classify an unknown class data instance with the help of the above data set given in the Table 1. The idea is to push the instance down the tree, following the branches whose attributes values match the instance attributes values, until the instance reaches a leaf node, whose class label is then assigned to the instance. For example, the data instance to be classified is described by the tuple (Age=23, Gender=female, Intensity of symptoms=medium, Goal=?), where “?” denotes the unknown value of the goal instance. In this example, Gender attribute is irrelevant to a particular classification task.

The tree tests the intensity of symptom value in the instance. If the answer is medium, the instance is pushed down through the corresponding branch and reaches the Age node. Then the tree tests the Age value in the instance. If the answer is 23, the instance is again pushed down through the corresponding branch. Now the instance reaches the leaf node, where it is classified as yes. (Harleen Kaur, Krishan Wasan S, 2006).

**Support Vector Machine (SVM)**

Support vector machines (SVM) are a classification technique originated from statistical learning theory. (Cristianini N, Shawe-Taylor J. 2006) Depending on the chosen kernel, SVM selects a set of data examples (support vectors) that define the decision boundary between classes. SVM have been proven for excellent classification performance, while it is arguable whether support vectors can be effectively used in communication of medical knowledge to the domain experts. SVMs are well suited to dealing with interactions among features and redundant features. (Vapnik VN. 1998).

**Genetic Algorithms (GAs) / Evolutionary Programming (EP)**

Genetic algorithms and evolutionary programming are algorithmic optimization strategies that are inspired by the principles observed in natural evolution. Of a collection of potential problem solutions that compete with each other, the best solutions are selected and combined with each other. In doing so, one expects that the overall goodness of the solution set will become better and better, similar to the process of evolution of a population of organisms. Genetic algorithms and evolutionary programming are used in data mining to formulate hypotheses about dependencies between variables, in the form of association rules or some other internal formalism. (Ramakrishnan N, etc, 2010).

**Fuzzy Sets**

Fuzzy sets form a key methodology for representing and processing uncertainty. Uncertainty arises in many forms in today’s databases: imprecision, non-specificity, inconsistency, vagueness, etc. Fuzzy sets exploit uncertainty in an attempt to make system complexity manageable. As such, fuzzy sets constitute a powerful approach to deal not only with incomplete, noisy or imprecise data, but may also be helpful in developing uncertain models of the data that provide smarter and smoother performance than traditional systems. [10]

**Neural Networks**

Neural networks (NN) are those systems modeled based on the human brain working. As the human brain consists of millions of neurons that are interconnected by synapses, a neural network is a set of connected input/output units in which each connection has a weight associated with it. The network learns in the learning phase by adjusting the weights so as to be able to predict the correct class label of the input.

**Rough Sets**

A rough set is determined by a lower and upper bound of a set. Every member of the lower bound is a certain member of the set. Every non-member of the upper bound is a certain non-member of the set. The upper bound of a rough set is the union between the lower bound and the so-called boundary region. A member of the boundary region is possibly (but not certainly) a member of the set. Therefore, rough sets may be viewed as with a three-valued membership function (yes, no, perhaps). Rough sets are a mathematical concept dealing with uncertainty in data. They are usually combined with other methods such as rule induction, classification, and clustering methods. (Shelly G, etc. 2011).

**Measures for performance evaluation**

Measures for performance evaluation depend on the learning task. If the task is diagnosis or prognosis, classification accuracy is the most frequently used quality evaluation measure, even in medical diagnosis and prognosis, the classification accuracy is not necessarily the best quality measure of a classifier. Selected evaluation measures are outlined below.

**Classification accuracy**

Assume a two-class classification problem (classes ‘positive’ and ‘negative’). Consider four subsets: True positives (TP): True positive answers of a classifier denoting correct classifications of positive cases; True
negatives (TN): True negative answers denoting correct classifications of negative cases; False positives (FP): False positive answers denoting incorrect classifications of negative cases into class positive; False negatives (FN): False negative answers denoting incorrect classifications of positive cases into class negative. The classification accuracy measures the proportion of correctly classified cases: (Delen D, etc. 2005).

\[
\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)}
\]

**Sensitivity and specificity**

For medical applications, two other measures are more frequently used than the classification accuracy: sensitivity and specificity [29, 52, and 37]. Sensitivity measures the fraction of positive Cases that are classified as positive. Specificity measures the fraction of negative cases classified as negative.

\[
\text{Sensitivity} = \frac{TP}{(TP+FN)}
\]

\[
\text{Specificity} = \frac{TN}{(TN+FP)}
\]

In many medical problems, high classification accuracy is less important than high sensitivity and/or specificity of a classifier’s answers. Sensitivity can be viewed as a detection rate that one wants to maximize. If the goal is to increase the sensitivity of answers, the learner should try to increase the correct classifications of positive cases (TP) and/or decrease the number of incorrect classifications of positive cases into class negative (FN).

On the other hand, in order to increase the specificity, the learner should try to increase the number of correct classifications of negative cases (TN) and/or decrease the number of incorrect classifications of negative cases into class positive (FP). Note that 1-Specificity can be interpreted as a false alarm rate which one wants to minimize. (Sousa TF, etc. 2004).

**Applications of data mining in medicine**

There is vast potential for data mining applications in healthcare.

**Prediction and diagnosis of diseases**

The diagnosis and treatment of patients is very important in medical sciences. (Soni J, etc. 2011) Selection of an inappropriate treatment for a patient can have harmful effects as well as it wastes the time and money. Not all the physicians are expert in all areas of their professions, and thus existence of a system for diagnosis and treatment selection can simultaneously reduce the costs and increase the accuracy and performance. (Naren Ramakrishnan, etc. 2010).

**Treatment effectiveness**

Data mining applications can be developed to evaluate the effectiveness of medical treatments. By comparing and contrasting causes, symptoms, and courses of treatments, data mining can deliver an analysis of which courses of action prove effective. For example, the outcomes of patient groups treated with different drug regimens for the same disease or condition can be compared to determine which treatments work best and are most cost-effective. (Hian Chye K, Gerald T. 2005).

**Healthcare management**

To aid healthcare management, data mining applications can be developed to better identify and track chronic disease states and high-risk patients, design appropriate interventions, and reduce the number of hospital admissions and claims. (Hian Chye K, Gerald T. 2005).

**Customer relationship management**

While customer relationship management is a core approach in managing interactions between commercial organizations—typically banks and retailers—and their customers, it is no less important in a healthcare context. Customer interactions may occur through call centers, physicians’ offices, billing departments, inpatient settings, and ambulatory care settings. (Hian Chye K, Gerald T. 2005).

As in the case of commercial organizations, data mining applications can be developed in the healthcare industry to determine the preferences, usage patterns, and current and future needs of individuals to improve their level of satisfaction. These applications also can be used to predict other products that a healthcare customer is likely to purchase, whether a patient is likely to comply with prescribed treatment or whether preventive care is likely to produce a significant reduction in future utilization.
CONCLUSION

In this paper, we examined the data mining in medicine. In order to do it, firstly the importance of data mining in medicine was briefly talked and then different techniques of data mining and their validation methods were presented. Some instances of the applications of data mining in medicine were finally pointed out. To build a framework for utilization of text mining in this area would be useful as a future work since a great part of medicine related data are saved in the form of texts.

Evaluation of data mining tools and software and the study of their characteristics and capabilities in medicine can greatly help the data miners.

REFERENCES