International Journal of Advance Research in Science and Engineering
Vol. No.6, Issue No. 08, August 2017
www.ijarse.com
IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

Survey on Needs, Applications and Agorithms of Data Mining for Healthcare

Sharad Mathur[1], Dr. Bhavesh Joshi[2]

ABSTRACT

Data mining is one of the essential areas of research that is more popular in health organization. Data mining plays an effective role for uncovering new trends in healthcare organization which is helpful for all the parties associated with this field. This paper gives brief summary about data mining and highlights about its need for healthcare sector. It explores selected data mining algorithms like Naive Bayes , Artificial neural networks and decision tree. This paper also highlights applications of Data Mining in healthcare.

Keywords: Data mining,Applications, Naïve Bayes, ANN, Decision tree.

I. INTRODUCTION

Data mining lies at the interface of statistics, database technology, pattern recognition, machine learning, data visualization, and expert systems. A database is a collection of data that is organized so that its contents can easily be accessed, managed, and updated. Databases usually include a query facility, and the database community has a tendency to view data mining methods as more complicated types of database queries. For example, standard query tools can answer questions such as, "How many surgeries resulted in hospital stays longer than 10 days?" Data mining is valuable for more complicated queries such as, "What are the important preoperative predictors of excessive length of stay?" Data mining techniques can be implemented retrospectively on massive data in an automated matter, whereas traditional statistical methods used in epidemiology require custom work by experts.

Data mining encompasses a wide variety of analytical techniques and methods, and data mining tools reflect this diversity. Decision tree is a predictive data mining techniques often used in clinical medicine to easily visualize, and understand resistant to noise in data. And is applicable in both regression and association data mining tasks [1] capable of handling continuous attributes, which are essential in case of medical data e.g. blood pressure, temperature, etc.A Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong independent assumption. In basic terms, a Naive Bayes classifier assumes that the presence of a particular feature of a class is unrelated to the presence of any other feature [2]. Artificial neural networks (ANNs) are widely used in science and technology with applications in various branches of chemistry, physics, and biology. To streamline the diagnostic process in daily routine and avoid misdiagnosis, artificial neural networks can be employed.

Data mining is becoming increasingly popular and essential in healthcare sector [3]. Data mining applications can provide advantage to all parties involved in the healthcare industry [4] [5]. For example, data mining can help healthcare insurer detect fraud and abuse, physicians identify effective treatments and best practices and patients receive better and more affordable healthcare services [6].

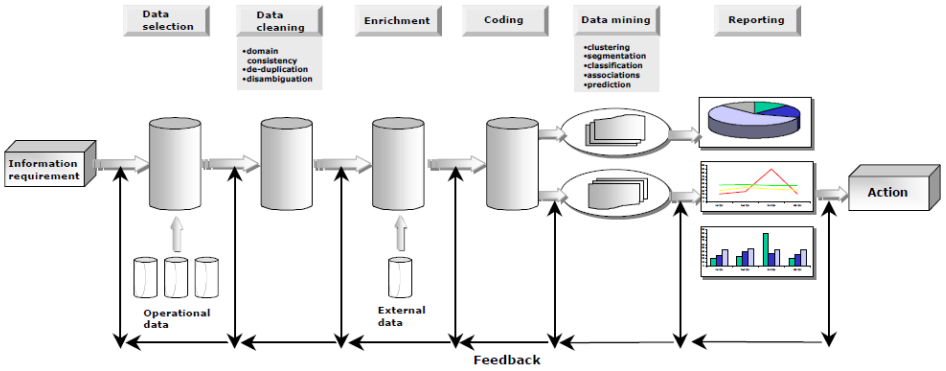
International Journal of Advance Research in Science and Engineering
Vol. No.6, Issue No. 08, August 2017
www.ijarse.com
IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

I. WHAT IS DATA MINING

Data mining refers to extracting or mining" knowledge from large amounts of data". There are many other terms related to data mining, such as knowledge mining, knowledge extraction, data/pattern analysis, data archaeology, and data dredging. Many people treat data mining as a synonym for another popularly used term, Knowledge Discovery in Databases", or KDD.

In general, a knowledge discovery process consists of an iterative sequence of the following step

1. Data cleaning, which handles noisy, erroneous, missing, or irrelevant data.
2. Data integration, where multiple, heterogeneous data source may be integrated into one.
3. Data selection, where data relevant to the analysis task are retrieved from the database.
4. Data transformation, where data are transferred or consolidated into forms appropriate for mining by performing summary or aggregation operations.
5. Data mining, which is an essential process where intelligent methods are applied in order to extract data patterns.
6. Pattern evaluation which is to identify the truly interesting patterns representing knowledge based on some interestingness measures.
7. Knowledge presentation, where visualization and knowledge representation technique are used to present the mined knowledge to the user



II. NEED OF DATA MINING

There is an unstoppable growth in the amount of electronic health records or EHRs being collected by healthcare facilities. It has been the norm for nurses to take responsibility in handling patient data input that was traditionally recorded in paper-based forms. Accuracy is extremely important when it comes to patient care and computerizing this massive amount of data enhances the quality of the whole system. It is well known that healthcare is a complex area where new knowledge is being accumulated daily in a growing rate. Big part of this knowledge is in the form of paperwork, resulting from a studies conducted on data and information collected from the patient's healthcare records. There is a big tendency today to make this information available in electronic form, converting information to knowledge, which is not an easy thing to do [7].

III. DATA MINING ALGORITHMS FOR HEALTHCARE

Data mining also recognized as Knowledge Discovery in databases is very frequently utilized in the field of

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

medicine. The process of supporting medical diagnoses by automatically searching for valuable patterns undergoes evident improvements in terms of precision and response time [8]. Every data mining technique serves a diverse purpose depending on the modeling objective. The two most common modeling objectives are classification and prediction. Classification models predict categorical labels (discrete, unordered) while prediction models predict continuous-valued functions [9]. Decision Trees and Neural Networks use classification algorithms while Regression, Association Rules and Clustering use prediction algorithms [10, 11].Here are some of the data mining algorithms which are successfully used in healthcare.

4.1. Naive Bayes

The Naive Bayesian classifier is based on Bayes' theorem with independence assumptions between predictors. A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. Despite its simplicity, the Naive Bayesian classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods.

Bayes' Theorem is stated as:

$P(h|d) = (P(d|h) * P(h)) / P(d)$

Where

P(h|d) is the probability of hypothesis h given the data d. This is called the posterior probability.

P(d|h) is the probability of data d given that the hypothesis h was true.

P(h) is the probability of hypothesis h being true (regardless of the data). This is called the prior probability of h.

P(d) is the probability of the data (regardless of the hypothesis).

After calculating the posterior probability for a number of different hypotheses, you can select the hypothesis with the highest probability. This is the maximum probable hypothesis and may formally be called the maximum a posteriori (MAP) hypothesis.

This can be written as:

$MAP(h) = max(P(h|d))$

or

$MAP(h) = max((P(d|h) * P(h)) / P(d))$

or

$MAP(h) = max(P(d|h) * P(h))$

The P(d) is a normalizing term which allows us to calculate the probability. We can drop it when we are interested in the most probable hypothesis as it is constant and only used to normalize.

Back to classification, if we have an even number of instances in each class in our training data, then the probability of each class (e.g. P(h)) will be equal. Again, this would be a constant term in our equation and we could drop it so that we end up with:

$MAP(h) = max(P(d|h))$

This is a useful exercise, because when reading up further on Naive Bayes you may see all of these forms of the theorem.

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

Naive Bayes is a classification algorithm for binary (two-class) and multi-class classification problems. The technique is easiest to understand when described using binary or categorical input values.

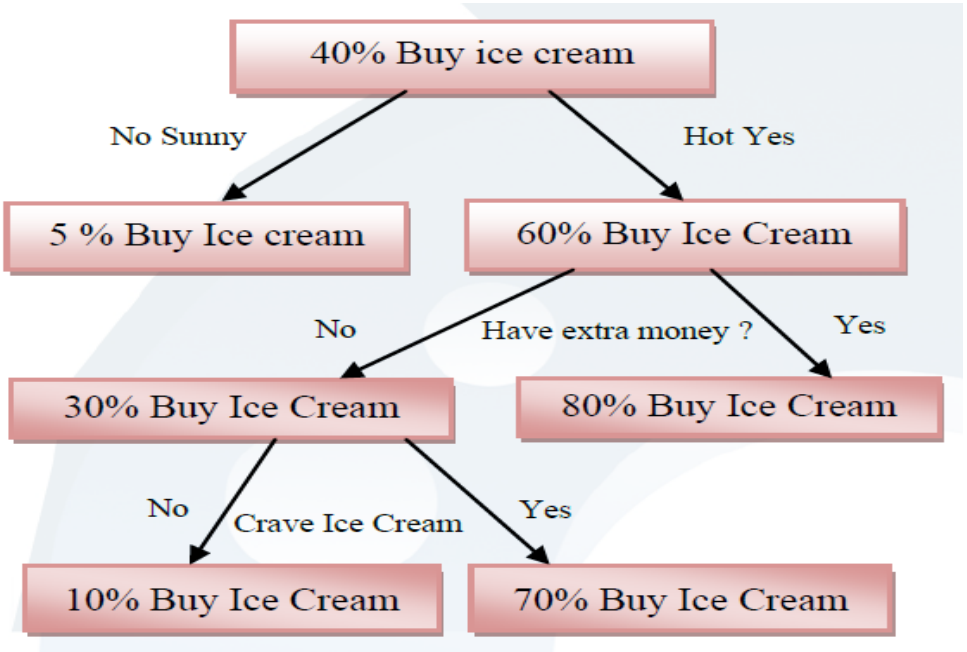
It is called naive Bayes or idiot Bayes because the calculation of the probabilities for each hypothesis are simplified to make their calculation tractable. Rather than attempting to calculate the values of each attribute value P(d1, d2, d3|h), they are assumed to be conditionally independent given the target value and calculated as P(d1|h) * P(d2|H) and so on.

This is a very strong assumption that is most unlikely in real data, i.e. that the attributes do not interact. Nevertheless, the approach performs surprisingly well on data where this assumption does not hold.

4.2 Decision Trees

A decision tree depicts riles for Classifying data into groups. Splits entire data set into some number of pieces and then another rule may be applied to a piece, different rules to different pieces forming a second generation of pieces. The tree depicts the first split into pieces as branches emanating from a root and subsequent splits as branches emanating from nodes on older branches. The leaves of the tree are the final groups, the unsplit nodes. For some perverse reason, trees are always drawn upside down, like an organizational chart. For a tree to be useful, the data in a leaf must be similar with respect to some target measure, so that the tree represents the segregation of a mixture of data into purified groups [12].

Consider an example of data collected on people in a city park in the vicinity of a hotdog and ice cream stand. The owner of the concession stand wants to know what predisposes people to buy ice cream. Among all the people observed, forty percent buy ice cream. This is represented in the root node of the tree at the [13] top of the diagram. The first rule splits the data according to the weather. Unless it is sunny and hot, only five percent buy ice cream. This is represented in the leaf on the left branch. On sunny and hot days, sixty percent buy ice cream. The tree represents this population as an internal node that is further split into two branches, one of which is split again.



Example of Decision Tree

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

4.3 Artificial Neural Networks (ANNS)

An artificial neural network is inspired by attempts to simulate biological neural systems. The human brain consists primarily of nerve cells called neurons, linked together with other neurons via stand of fiber called axons. Axons are used to transmit nerve impulses from one neuron to another whenever the neurons are stimulated. A neuron is connected to the axons of other neurons via dendrites, which are extensions from the cell body of the neurons. The contact point between a dendrite and an axon is called a synapse. Neural networks provide a very general way of approaching problems. When the output of the network is continuous, such as the appraised value of a home, then it is performing prediction. When the output has discrete values, then it is doing classification. A simple rearrangement of the neurons and the network becomes adept at detecting clusters. The fact that neural networks are so versatile definitely accounts for their popularity. The effort needed to learn how to use them and to learn how to massage data is not wasted, since the knowledge can be applied wherever neural networks would be appropriate [14].

Multilayer is feed-forward neural networks trained with the standard back-propagation algorithm. It is supervised networks so they require a desired response to be trained. It learns how to transform input data in to a desired response, so they are widely used for pattern classification. With one or two hidden layers, they can approximate virtually any input-output map. It has been shown to approximate the performance of optimal statistical classifiers in difficult problems. The most popular static network in the multilayer. The multilayer is trained with error correction learning, which is appropriate here because the desired multilayer response is the arteriographic result and as such known [15]. Error correction learning works in the following way from the system response at neuron j at iteration t, yj (t),and the desired response dj (t)for given input pattern an instantaneous error ej (t)is defined by

$$e_j(t) = d_j(t) - y_j(t)$$

Using the theory of gradient descent learning, each weight in the network can be adapted by correcting the present value of the weight with a term that is proportional to the present input and error at the weight, i.e.

$$w_{jk}(t + 1) = w_{jk}(t) + \eta\delta_j(t)x_k(t)$$

The η(t) is the learning-rate parameter. The wjk(t) is the weight connecting the output of neuron k to the input neuron j at iteration t. The local error δj(t) can be computed as a weighted sum of errors at the internal neurons.

V. APPLICATIONS IN HEALTHCARE

Healthcare industry today generates large amounts of complex data about patients, hospital resources, disease diagnosis, electronic patient records, medical devices etc. Larger amounts of data are a key resource to be processed and analyzed for knowledge extraction that enables support for cost-savings and decision making. Data mining applications in healthcare can be grouped as the evaluation into broad categories.

5.1 Fraud and Abuse

To identify fraud and abuse data mining applications often set up norms and then recognize unusual patterns of claims by physicians, clinics, laboratory or some others. These data mining applications can also throw a light on unsuitable prescriptions or referrals and false insurance and health claims [16].

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

5.2 Treatment Effectiveness

Data mining applications can develop to evaluate the effectiveness of medical treatments. Data mining can deliver an analysis of which course of action proves effective by comparing and contrasting causes, symptoms, and courses of treatments. The use of classification algorithms to help in the early detection of heart disease, a major public health concern all over the world.The use of data mining as a tool to aid in monitoring trends in the clinical trials of cancer vaccines. By using data mining and visualization, medical expertscould find patterns and anomalies better than just looking at a set of tabulated data [13].

5.3 Cost Effective Treatment

Data mining allows organizations and institutions to get more out of existing data at minimal extra cost. KDD and data mining have been applied to discover fraud in credit cards and insurance claims [16]. By extension, these techniques could also be used to detect anomalous patterns in health insurance claims, particularly those operated by PhilHealth, the national healthcare insurance system for the Philippines.

5.4. Pharmaceutical Industry

The technology is being used to help the pharmaceutical firms manage their inventories and to develop new product and services. A deep understanding of the knowledge hidden in the Pharma data is vital to a firm's competitive position and organizational decision-making.

5.5. Evidence-Based Medicine and Prevention of Hospital Errors

When medical institutions apply data mining on their existing data, they can discover new, useful and potentially life-saving knowledge that otherwise would have remained inert in their databases. For instance, an ongoing study on hospitals and safety found that about 87% of hospital deaths in the United States could have been prevented, had hospital staff (including doctors) been more careful in avoiding errors [17]. By mining hospital records, such safety issues could be flagged and addressed by hospital management and government regulators.

5.6. Hospital Management

Organizations including modern hospitals are capable of generating and collecting a huge amount of data. Application of data mining to data stored in a hospital information system in which temporal behavior of global hospital activities is visualized [18]. Three layers of hospital management:

- Services for hospital management
- Services for medical staff
- Services for patients

VI. CONCLUSION

Data Mining is a powerful tool to help physicians perform diagnosis and other enforcements. In this regard it has several advantages including:

- The ability to process large amount of data
- Reduced likelihood of overlooking relevant information

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

- Reduction of diagnosis time

Data mining have proven suitable for satisfactory diagnosis of various diseases. In addition, their use makes the diagnosis more reliable and therefore increases patient satisfaction. However, despite their wide application in modern diagnosis, they must be considered only as a tool to facilitate the final decision of a clinician, who is ultimately responsible for critical evaluation of its outputs.

REFERENCES

[1] K. Aftarczuk, "Evaluation of selected data mining algorithms implemented in Medical Decision Support Systems," Blekinge, 2007.

[2] Patrick Breheny, Kernel density classification, STA 621: Nonparametric Statistics October 25.

[3] T Koh, HianChye, and Gerald Tan, "Data mining applications in healthcare."Journal of healthcare information management, Vol. 19, No. 2, 2011, pp. 65-68.

[4] H. Kaur and SiriKrishanWasan, "Empirical study on applications of data mining techniques in healthcare." Journal of Computer Science Vol. 2, No. 2, 2006, pp. 194-200.

[5] M. K. Obenshain, "Application of data mining techniques to healthcare data." Infection Control & Hospital Epidemiology Vol. 25, No.08, 2004, pp. 690-695.

[6] S. H. Liao, Pei-Hui Chu, and Pei-Yuan Hsiao, "Data mining techniques and applications–A decade review from 2000 to 2011." Expert Systems with Applications, Vol. 39, No.12, 2012, pp. 11303-11311

[7] Ceusters, W. (2001). Medical Natural Language Understanding as a Supporting Technology for Data Mining in Healthcare.KJ Cios (ed.) Medical Data Mining and Knowledge Discovery, Physica-verlag Heidelberg, (pp. 41-67). New York

[8] KamilaAftarczuk, Evaluation of selected data mining algorithms implemented in Medical Decision Support Systems, Thesis no: MSE-2007-21, September 2007, School of Engineering, Blekinge Institute of Technology, Sweden.

[9] Han, J., Kamber, M., Data Mining Concepts and Techniques (Morgan Kaufmann Publishers, 2006).

[10] Charly, K., Data Mining for the Enterprise, 31st Annual Hawaii Int. Conf. on System Sciences, IEEE Computer, 7, 295-304, 1998.

[11] V. Krishnaiah, G. Narsimha& N. Subhash Chandra, A Study on Clinical Prediction using Data Mining Techniques, International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR) ISSN 2249-6831 Vol. 3, Issue 1, Mar 2013, 239-248 TJPRC Pvt. Ltd..

[12] B. Rajasekhar et al., "QUALITY OF CLUSTER INDEX BASED ON STUDY OF DECISION TREE", International Journal of Research in Computer Science, eISSN 2249-8265 Volume 2 Issue 1 (2011) pp. 39-43.

[13] Jensen, D. D. and Cohen, P. R (1999), "Multiple Comparisons in Induction Algorithms," Machine Learning (to appear). Excellent discussion of bias inherent in selecting an input. Explore http://www.cs.umass.edu/~jensen/papers.

[14] AnchanaKhemphila, VeeraBoonjing, "Parkinsons Disease Classification using Neural Network and Feature selection", World Academy of Science & Tech, 64, 2012

[15] Khemphila A., Boonjing V., "Parkinsons Disease Classification using Neural Network and Feature selection", World Academy of Science, Engineering and Technology,Vol:6 2012-04-25.

IJARSE
ISSN (O) 2319 - 8354
ISSN (P) 2319 - 8346

[16] Kou, Y., Lu, C.-T., Sirwongwattana, S., and Huang, Y.-P. (2004). Survey of fraud detection techniques. In Networking, Sensing and Control, 2004 IEEE International Conference on Networking, Sensing and Control. (2) 749-754.

[17] Health Grades, Inc. (2007). The Fourth Annual HealthGrades Patient Safety in American Hospitals Study.

[18]ShusakuTsumoto and Shoji Hirano, ―Temporal Data Mining in Hospital Information Systems.