Using Data Mining technique to diagnosis heart disease

Lamia AbedNoor Muhammed
College of Computer and Mathematic Science
University of Al-Qadissiya
Diwaniya, Iraq
bon1491988@yahoo.com

Abstract-Medical diagnose is a promise application that exploits data mining techniques. The physicist diagnose, represented by human expertise, can be incurrence to fail. In contrast the data mining can recruit the extracted knowledge from huge of clinical data though data mining and produce a predictive model, use the classification task to achieve the diagnostic. Different methods exist in this field, to produce the classifier. One of them is naïve bays.
In this paper, we will present and discuss the experiment that was executed with naïve bayes technique in order to built predictive model as an artificial diagnose for heart disease based on data set which contains set of parameters that were measured for individuals previously. Then compare the results with other techniques according to using the same data that were given from UCI repository data.
Keyword: naïve classifier; artificial diagnosis; medical data mining

I. 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[1], where this diagnosis is to reveal the existence of the disease or not. Diagnosis of disease is one of the promise application for several reason; a huge of related data are available and in the same time, the need for accurate diagnosis. So this has encouraged to attract many researchers to work in this field. At early, study in [2] has demonstrated how consistent and complete data mining in medical diagnosis can acquire a set of logical diagnostic rules for computer-aided diagnostic systems. A case study in paper[3], the problem of predicting outcomes in medical and engineering applications is discussed. The problem is solved with a data mining approach. The work in paper[4]; a comparison of different learning models used in Data Mining and a practical guideline how to select the most suited algorithm for a specific medical application is presented and some empirical criteria for describing and evaluating learning methods are given. Paper[5], it focuses on analyzing medical diagnostic data using classification rules in data mining and context reduction in formal concept analysis. Context reduction technique given in Formal Concept Analysis along with classification rules has been used to find redundancies among the various medical examination tests. In paper[6] briefly examine the impact of data mining techniques, including artificial neural networks, on medical diagnostics. The research paper[7] provides a survey of current techniques of KDD, using data mining tools for healthcare and public health. It also discusses critical issues and challenges associated with data mining and healthcare in general. The research in paper[8] intends to provide a survey of current techniques of knowledge discovery in databases using data mining techniques that are in use in today's medical research particularly in Heart Disease Prediction.

II. Data Mining techniques

Data mining is a field related with many ways to provide significant knowledge that are modeled in order to use in different applications. However data is available that contain a hidden knowledge, the data mining technique search for the regularity, relationships, outlier events in this data and present as hidden knowledge. Then this knowledge can be employed in different applications. Predictive is one of the tasks that use the hidden knowledge; extracted and modeled in order to predicate unknown value based on this knowledge in future. Prediction task is classified in one of two types; one can either try to predict some unavailable data values or pending trends, or predict a class label for some data and is tied to classification. Once a classification model is built based on a training set, the class label of an object can be foreseen, based on the attribute values of the object and the attribute values of the classes. Prediction is referred to the forecast of missing numerical values, or increase/ decrease trends in time related data. The major idea is to use a large number of past values to consider probable future values [9].In diagnosis task at most, it will be needed for predication binary logic state i.e.

existence or absence of disease in the person according to his examination record. So The induces classification can be treated as diagnosis whereas give the answer about the chance of existence of disease. The naive Bayes classifier is optimal for any two-class concept with nominal features that assigns class 0 to exactly one example, and class 1 to the other examples, with probability [10]. The naïve classifier employs the probabilistic relationships between the class label(specific variable) and other variables that characterized individual instance.

III. Classification model

Classification model(Classifier) can be built through learning process. It is describing a predetermined set of data classes or concepts[11]. The model is constructed from available data i.e. classified examples. These examples consist of set of features(variables). This model is induced through supervised learning process, the classified examples would be processed as training data. The training example is presented to learning algorithm with its class and let to extract specific knowledge gradually. After, classifier was built, this model would be evaluated according to the answers accuracy that model will give through testing. Whereas, another data, unlabeled examples that are known test data, will provide to the model. Then let the model answer, give the class to each example. At last these answers would be compared with the correct answers. In classification we wish to learn a mapping from a vector of measurements x to a categorical variable Y. The variable to be predicted is typically called the class variable (for obvious reasons), and for convenience of notation we will use the variable C, taking values in the set {c1, ..., cm} to denote this class variable for the rest of this chapter (instead of using Y). The observed or measured variables X1, ..., Xp are variously referred to as the features, attributes, explanatory variables, input variables, and so on , where a correct prediction incurs a loss of 0 and an incorrect class prediction incurs a loss of 1 irrespective of the true class and the predicted class. We will begin by discussing two different but related general views of classification: the decision boundary (or discriminative) viewpoint, and the probabilistic viewpoint[12].

IV. Naïve Bayes Classifier (NBC)

Naïve classifiers are an old and well known type of classifiers. They use a probabilistic approach, i.e, they try to compute conditional class probabilities and then predicate the most probable class. Naïve classifiers exploit as their name already indicates- Bayes rule and a set of conditional independence assumption[13]. Applying probabilistic approaches to classification techniques typically involve modeling the conditional probability distribution P(C |D), where C ranges over classes and D over descriptions data, in some language, of objects to be classified. Given a description d of a particular object, we assign the class arg max
P(C = c|D = d). A Bayesian approach splits this posterior distribution into a prior distribution P(C) and a likelihood P(D |C):

$$\underset{c}{argmax}\, P(C=c|D=d) = \underset{c}{argmax} \frac{P(C=c|D=d)P(C=c)}{P(D=d)} \quad \ldots(1)$$

The denominator P(D = d) is a normalizing factor that can be ignored when determining the maximum a posteriori class, as it does not depend on the class.
The key term in Eq.(1) is P(D = d |C = c),the likelihood of the given description given the class (often abbreviated to P(d |c)). A Bayesian classifier estimates these likelihoods from training data, but this typically requires some additional simplifying assumptions. For instance, in an attribute-value representation (also called propositional or single-table representation), the individual is described by a vector of values a1,……,an for a fixed set of attributes A1,...,An. Determining P(D = d |C = c) here requires an estimate of the joint probability P(A1=a1,...,An=an|C = c), abbreviated to P(a|c). This joint probability distribution is problematic for two reasons: (1) its size is exponential in the number of attributes n, and (2) it requires a complete training set, with
several examples for each possible description. These problems vanish if we can assume that all attributes are independent given the class:

$$P(A_1 = a_1, \ldots, A_n = a_n | C = c) = \prod_{i=1}^{n} P(A_i = a_i | C = c) \quad \ldots(2)$$

This assumption is usually called the naive Bayes assumption, and a Bayesian classifier using this assumption is called the naive Bayesian classifier, often abbreviated to 'naive Bayes'. Effectively, it means that we are ignoring interactions between attributes within individuals of the same class[14].

V. PRACTICAL WORK

In this paper, the experiment was applied on the specific data . It was attempting to induce a classifier in order to use it with diagnosis the heart if it is normal or abnormal.

A. Experiment Data

The data was used, was extracted from UCI repository data set. This dataset describes diagnosing of cardiac Single Proton Emission Computed Tomography (SPECT) images. Each of the patients is classified into two categories: normal and abnormal. The database of 267 SPECT image sets (patients) was processed to extract features that summarize the original SPECT images. As a result, 44 continuous feature pattern was

created for each patient. The pattern was further processed in order to obtain 22 binary feature patterns. The characteristics of SPECT data are shown in Table(I).

Table I. characteristics of experiment data set

no.	parameters	value
1	Number of Instances	267
2	Number of Attributes	23
3	Number of training instances	80
4	Number of test instance	187
5	Number of classes	2

B. Procedure of work

The procedure that was applied in this paper work is to build the classifier and test it:-

- Building classifier: it was performed through training process, where the training data with 80 instances, grouped into two groups; first group(40 instances) with positive (normal i.e. heart with no disease) that are considered as a control examples, while the second group(40 instances) with negative (abnormal i.e. heart with disease). From these data , estimated the prior probabilities and as a result, for each class(positive, negative) has its probabilistic model as shown in Eq.(2).
- Testing the model: it was accomplished with testing data set that contains(187) instances; (15 instances) have positive answer, while(172 instances) have negative answer. The testing was performed by presenting each instances without class attribute then let the model answers based on posterior distribution that was shown in eq.(1).
- Calculate the accuracy of the model by calculating the incorrect answers and compare with the all answer given from classifier.

C. Results

The result that was obtained from practical work was very good, whereas all the answers are correct. So the accuracy of this model achieveed ratio(100%). This result was compared with other works that were performed on the same data set and found in the same site of experiment data. These results are shown in table(II).

Table II. Accuracies for different Classifiers Algorithm

no.	Algorithm type	Accuracy
1	CLIP3	%84.0
2	CLIP4	%86.1
3	ensemble of CLIP4	%90.4
4	Naïve bayes classifier	100%

VI. CONCLUSION

Applying different techniques with this experiment data would reveal different response as shown in Table(II). So the selection of the suitable technique is necessary for the diagnosis to get accurate answers.

In addition, there is another issue with the diagnosis field, where each disease has its private parameters that cooperate in diagnosing, and differentiated from other disease. So in our paper work, good results were given in diagnose heart (normal , abnormal) using cardiac Single Proton Emission Computed Tomography (SPECT) images with naïve bayes classifier in comparing with other, may be not with another disease that has different data space.

REFERENCES

[1] J. Soni, and others, "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", International journal of Computer Application,volume17, No. 8, March 2011.

[2] B. Kovalerchuk, , E. Vityaev, and J. Ruiz, "Consistent and complete data and "expert" mining in medicine'. In: Cios, K. (ed.) Medical Data Mining and Knowledge Discovery, Springer, Heidelberg, pp. 238-280, 2001.

[3] A. Kusiak, and others,"Data Mining: Medical and Engineering Case Studies", Proceedings of the Industrial Engineering Research 2000 Conference, Cleveland, Ohio, pp. 1-7, 2000.

[4] A. Plamena, D. Maya, and R. Petia , DATA MINING LEARNING MODELS AND ALGORITHMS FOR MEDICAL APPLICATIONS, www.cvs.uab.es/ ~petia/maya%20sear3.pdf.

[5] G., Anamika, K. Naveen, and B. Vasudha , "Analysis of Medical Data using Data Mining and Formal Concept Analysis", World Academy of Science, Engineering and Technology 11,pp. 61-64, 2005.

[6] K. Siri, B. Vasudha, and K.,Harleen, "THE IMPACT OF DATA MINING TECHNIQUES ON MEDICAL DIAGNOSTICS', Data Science Journal, Volume 5, pp.119-126, 2006.

[7] D. Ruben, Canlas Jr., "DATA MINING IN HEALTHCARE: CURRENT APPLICATIONS AND ISSUES", www..

[8] S. Jyoti, and others, "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction" , International Journal of Computer Applications (0975 – 8887) , Volume17, pp. 43-48, 2011

[9] B.N. Lakshmi., and G. H. Raghunandhan, "A Conceptual Overview of Data Mining", Proceedings of the National Conference on Innovations in Emerging Technology-2011 Kongu Engineering College, Perundurai, Erode, Tamilnadu, India.17 & 18 February, 2011.pp.27-32.

[10] I. Rish, "An empirical study of the naïve Bayes classifier", proceedings of IJCAI 2001, Workshop on Empirical Methods in Artificial, 2001.

[11] H. Jiawei, and K. Michlin, "Data Mining: Concepts and Techniques", Morgan Kaufmawn Publishers, pp. 279-281, 2001.

[12] H. David, M. Heikk¸ and S. Padhraic, "Principles of Data Mining", The MIT Press ©, 2001.

[13] B. Christian, and G. Jorg, "A naïve Bayes Style Possibilistic Classifier" .

[14] P. A. Flach, and N. Lachiche, "Naïve Bayesian Classification of Structured Data Learning", machine Learning, Kluwer Academic Publishers, vol. 57, pp. 233-269, 2004.