IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-6, Issue-6, June 2018 E-ISSN: 2347-2693

A Comprehensive Study of Various Classification Techniques in Medical Application using Data Mining

Dipti N. Punjani[1], Kishor Atkotiya[2]

[1]National Computer College, Jamnagar, India
[2]Department of Statistics, Saurashtra University, Rajkot, India

Available online at: www.ijcseonline.org

Accepted: 03/Jun/2018, Published: 30/Jun/2018

***Abstract* -** Data mining is a set of techniques to analyze available data to find some hidden truths and unknown facts. The purpose of these techniques is to determine the information from the data. Such information might be in the form of explaining past or predicting future. In recent years, prediction algorithms are used for various applications. From prediction of student performance to product selling, from prediction of diseases to stock market, we try to find out what will happen in future using prediction algorithms. One of the most widely used future predictions is classification. This paper discusses how classification helps in prediction of life threaten diseases like cervical cancer.

***Keywords* -** Data Mining, Prediction, Classification, Decision Tree, Naïve Bayes, Cancer, Cervical Cancer

I. Introduction

People often try to analyse whatever data available in our mind. The human understanding, intuition, common sense and logic help people to understand / explain past and predict / discuss future in a meaningful way. The same way, we can analyse the computerized data for betterment of understanding of relationship among various components of it. The advantage is here is to represent only the hidden information which is not directly visible. Data mining is a set of techniques which are used for this purpose. Classification and Regression are two techniques used to predict one or more unknown fields of a given data. For example, we could develop a classification model to predict grade of a student. We could develop a regression model to predict percentage of a student. Here, if we think of what information we try to predict then it is result. We may think that in both the cases, we are predicting results of a student. Certainly, it is true also. But there is a difference between what kinds of information both the models predict. A classification model predicts grade while a regression model predicts percentage. Subsequently we can generalize the difference by saying; a classification model predicts a categorical value while a regression model predicts a numerical value. A categorical value refers to a set of classes like grade, gender, designation, branch, nationality, state etc. while a numerical value refers to the fields whose values are numerical like percentage, age, salary etc.[1][2]

Several algorithms are proposed for classification as well as for regression. Section 2 discusses fundamentals of classification. Section 3 discusses categories of classification algorithms. Section 4 discusses naïve bayes classification, decision tree classification algorithms while section 5 discusses classification used in medical applications such as cervical cancer. The paper concludes with the comparison and also discusses how these algorithms could be used in medical diagnosis for cervical cancer.

II. Classification

Classification is a process of labelling an unknown class attribute of a data to one of the possible values. The same process could be extended to label more than one unknown class attributes too. The primary objective is to determine categorical values which are at present unknown to us. Classification is performed for prediction purpose. For example, we could use a classifier – classification model to predict type of diseases a patient might be having based on his medical history. As we are predicting future values, accuracy and efficiency are primary requirement. Though future has uncertainty, our model should be enough capable to predict closest to the future reality. The classification process is explained in Section 2.1 in detail. [1][2]

A. Classification Process

Classification is a systematic process of designing a model which is enough accurate to be believed. In reality, we are interested in predicting future based on past and present information. The available information is a set of features while the required future information is a set of outcomes. The sequence of classification steps are shown in Figure 1. [1][2]

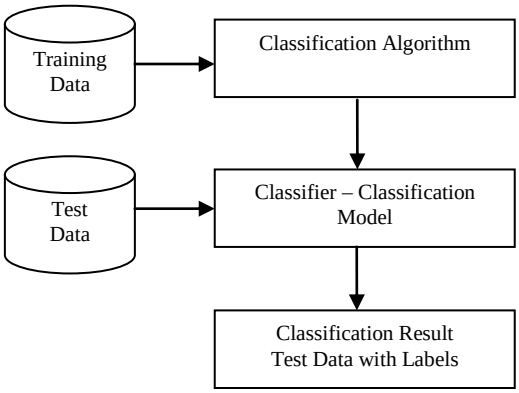

Figure 1 – Classification Process

As shown in Figure 1.1, Classification process required two types of data: Training Data and Test Data. A Training data set is a collection of all records which explain past. These records have all values including the categorical values for which we are interested in designing a classifier. A test data is a collection of all records which explain present and for which we want to predict future. These records do not have values for categorical fields. For example, we could have a training data set composed of medical reports of all past patients and their corresponding diseases. We could imagine that we have a test data set composed of medical reports of current patients which are yet to be analysed for diagnostic purpose. Here the training data must be authentic and complete. Authenticity refers to the correct, accurate, timely and valid values. Completeness refers to the availability of all possible combinations. A training data set is provided to the classification algorithm for data mining purpose. A classification algorithm like naïve bayes, decision tree etc. analysis the training data and apply statistical methods to determine hidden relationships among various features and outcomes. The output of classification algorithm is a statistical model called classifier. This phase is called Training / Learning phase. [1][2]

Test data is fed to the classifier to predict the unknown class labels. Here the outcomes will be decided based on analysis of available features. The output is also called classification result with test data with labels. Here labels refer to the outcomes in the form of values assigned to one or more categorical unknown fields. This phase is called Testing Phase. These are the main components of classification. To improve quality of classification, two other tasks could be done.[1][2]

B. Classification Accuracy

Prior to building a classification model, data cleansing could be done to remove any incomplete, incorrect or inappropriate record which may degrade overall accuracy of classifier. At the same time, analysis of authenticity and completeness could be done to ensure that the training data is utmost accurate and updated. Later on, once a classification model is designed, we could test its accuracy before using it to classify test data. As a part of accuracy checking, a classifier can be used to classify some or all of the training data. Later on, we could compare the classification results with the actual results. We can start using our classifier once satisfactorily accuracy is achieved [3].

III. Classification Categories

Classification is a process of prediction. There are several algorithms which are used to do so. These algorithms are based on various approaches too. This section discusses the primary two categories of classification algorithms.

A. Eager Learners

As the word "Eager" itself says, eager learners are those classification algorithms which have very strong learning phase. Here learning refers to the in depth analysis of data as a part of building a classification model. Such classification algorithms are useful when training data is very large and may not be available all the time. In these methods, the training data set is required only to build a classification model. Once a classification model is developed, there is no further need of training data set. Testing phase requires model and test data for classification purpose. Training phase is time consuming but once model is developed, there is no need to consume space to store a large training data. Decision Tree, Naïve Bayes are examples [4].

B. Lazy Learners

As the word "Lazzy" itself says, lazzy learners are those classification algorithms which have no prior learning phase before moving to the testing phase. Such classification algorithms are useful when training data is small and can be saved and made available all the time. Here no learning phase, does not mean no analysis. The meaning of absence of learning phase is there is no separation like training / learning and testing. The training data is saved. At the time of testing, the training data is analysed to classify test data. The primary difference between analysis done by eager learners and lazzy learners is in the kind of information they find. Eager learners try to build a model while lazzy learners compare test data with training data to find similarities. K-Nearest Neighbour is an example [5].

IV. Classification Algorithms

Many platforms provide API to perform classification. These APIs have various routines to perform classification using algorithms like decision tree, naïve bayes etc. This paper discusses how naïve bayes algorithm is used to perform classification.[5]

A. Probability

The probability is used to measure chance of possibility of a particular outcome of a particular event. As we have discussed that the classification is used to predict future, with every prediction corresponding probability helps us in

defining how much possibility we can expect for a particular prediction. The probability can be described by a value between 0 and 1 which defines possibility of an outcome for an event. For example, probability of rolling a dice and getting 1 is 1/6 as total 6 combinations are possible. At the same time probability of rolling a dice and getting even is 3/6 = ½ as there are three cases (2,4,6) are possible out of 6 combinations. The same concept can be applicable to define probability of one event considering presence of another event. Such probability is called conditional probability [5].

B. Bayes' Theorem

Bayes' theorem is used to define relationship between an event and prior conditions related with the event. This theorem is used to find out probability of an event when certain conditions are satisfied. Such probabilities are called conditional probabilities [5].

$$P(A \mid B) = P(B \mid A) * P(A) / P(B) \quad (1)$$

P(A | B) is probability of event A given event B occurred. P(B | A) is probability of event B given event A occurred. P(A) and P(B) are independent probabilities of events A and B respectively without consideration of any prior events. These probabilities are called marginal probabilities. The same conditional probability can be derived using following formula too [5].

$$P(A \mid B) = P(A \cap B) / P(B) \quad (2)$$

P(A ∩ B) is joint probability of event A and B happening together.

C. Naïve Bayes Classifier

Naïve Bayes classifier is based on using Bayes theorem. Lets say our database is composed of n+1 fields. Every record has n+1 records which are v1,v2,..vn and C. here v1 to vn represent all n fields whose values are known prior to the classification while C is a value being predicted. For example, D={v1,v2,…..,vn,C}. in this case, we will use v1 to vn as n predictors to find value of a target C. We can calculate probability of every value of C using following formula[5].

$$P(C \mid D) = P(v1 \mid C) * P(v2 \mid C) * \ldots * P(vn \mid C) * P(C) \quad (3)$$

For real world applications, we could find probability for each of the possible values of C for given values of v1 to vn as a part of D. The final value of C will be the value corresponding to the highest probability[5].

D. Decision Tree

Naïve Bayes classifier is based on probabilities while decision tree based methods (ID3) are based on entropies. Entropy defines how much same a set of values are. For a set of values which are exactly same, entropy is 0. For a set of values which are drastically different, entropy is 1. Entropy is calculated as below[6].

$$E(T) = \sum_{i=0}^{n} - P_i * log_2 P_i \quad (4)$$

Here T is the attribute which has n+ 1 possible value. Pi refers to the probability of i^{th} value. Various algorithms are developed like ID3 – Iterative Dichotomiser 3. The output of decision tree algorithm is a decision tree which can be presented in the form of if…else… rules. Every interior node of the tree is a decision node – describing a condition and every exterior node defines one of the possible values for the attribute being predicted. One such example is given figure 2. The data is used to predict whether a person will buy a computer or not. So the value being predicted has two possible answers. Yes or No. Age, Student, Credit_rating are the predictors which serve as inputs to the decision tree [6].

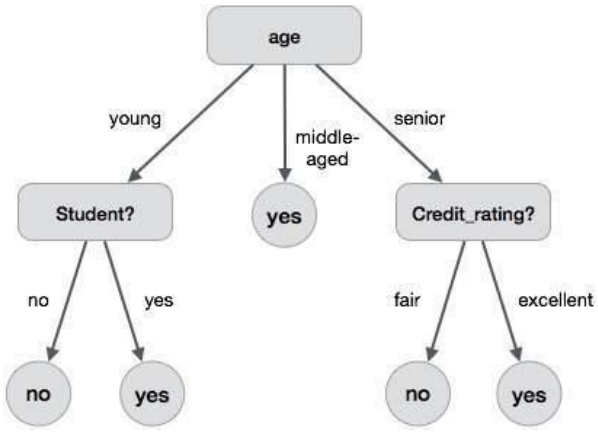

Figure 2 – Decision Tree Sample

V. Classification in Medical Applications

Classification algorithms are widely used in various fields of education, business, finance, marketing and medical science. With the fast growth of computerization, most of the medical tests and diagnosis are computerized. Several laboratories use data analytics methods to compare reports of a patient and summarized in a way which is understandable by doctors as well as patients. Several online forums help patients in diagnosis primarily with suggested therapies, diet plans, symptoms etc. medical science is mainly using image processing and data mining fields to improve analysis. Image processing techniques are useful in analyzing x-rays and MR reports. While data mining techniques are useful in analysis of patients test reports. Various fields of data mining applications in medical science are shown in Figure 3.

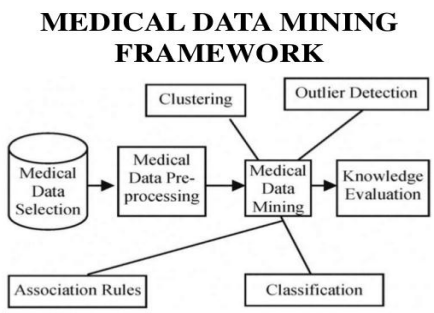

Figure 3 – Medical data Mining Framework

A. Cancers

Cancers are dangerous and lives threaten. It is indeed a requirement to detect them at earliest to increase survival rate post treatment or post surgery. It is obvious that every cancer can not be cure but the early detection can help to the best. The process of cancer detection by a data mining based model can be used to answer either Yes or No. that's means to diagnosed a patient with cancer or not. At present, it is very difficult to answer Yes or No in a computerized way especially for the diseases like cancers. So, most of the research work is focused towards finding risk of patients- like patient at high risk, moderate risk or low risk. So, further treatment can be done accordingly. A way to perform such work is to find out what are the most important and relevant tests a patient should go through. The steps of building a cervical cancer detector are shown in Figure 4.

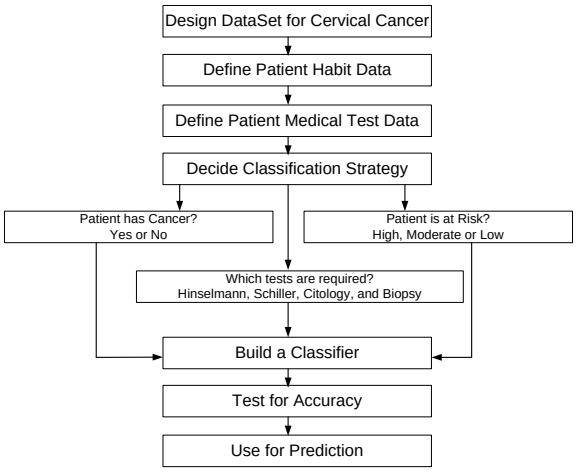

Figure 4 – Cervical Cancer Prediction System

VI. Conclusion

This paper discussed importance of classification in data mining applications. Classification and regression are differentiated. Eagar learners and lazy learners are classified. The overall process of classification is discussed. Two widely used classification algorithms are discussed which are naïve bayes and decision tree method of ID3. It is discussed that classification is one of the most widely required prediction method. In reality, classification is mostly used to sort / order / arrange / group data while in data mining applications, classification is used to predict the category to which an object belongs. Naïve bayes method takes feature independent assumption as it considers every feature independent from others. So it is not suitable in cases where features are closely related and cannot be separated. Decision tree method can be easily implemented too.

Future Work

Further research work can be carried out towards including more classification algorithms like CART, C4.5, KNN and Neural Networks etc. no method is global and accepted for all cases. As discussed in the conclusion, naïve bayes is not suitable when features are dependent on each other. So a method can be developed to check for the appropriateness of an algorithm before used in real life.

References

[1] Han, Jiawei, Jian Pei, and Micheline Kamber. Data mining: concepts and techniques. Elsevier, 2011.
[2] Aggarwal, Charu C. Data mining: the textbook. Springer, 2016.
[3] Jiawei Han, Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, 2001
[4] Kaur H, Wasan SK. Empirical Study on Applications of Data Mining Techniques in Healthcare. J Comput Sci. 2006;2(2):194-200. doi:10.3844/jcssp.2006.194.200.
[5] Venkatadri.M and Lokanatha C. Reddy ,"A comparative study on decision tree classification algorithm in data mining" , International Journal Of Computer Applications In Engineering ,Technology And Sciences (IJCAETS), Vol.- 2 ,no.- 2 , pp. 24- 29 , Sept 2010.
[6] Safavian, S. Rasoul, and David Landgrebe. "A survey of decision tree classifier methodology." IEEE transactions on systems, man, and cybernetics21.3 (1991): 660-674.

