Decision Support System For Lung Cancer By using Bayesian Classification

Than Than Lin ;Khin Sandar
Computer University, Pathein
minmin1494@gmail.com , drkhinsandar@gmail.com

Abstract

In many application domains, classification of complex measurements is essential in a diagnostic process. Correct classification of measurements may in fact be the most critical part of the diagnostic process. The main feature of the proposed system is to provide a sample and integrated tool for designing diagnostic application. Lung cancer is the second most common malignancy in men and the third most common cancer in women. Usually lung cancer nodules have a multifocal origin and a rather poor prognosis. Therefore, a careful review of the symptoms presented and a detailed physical exam greatly help with the diagnosis occurs. This paper proposes a decision support system for lung cancer classification using Bayesian Analysis to help the physician or the patient who tests herself at home for lung cancer with the most possible result. Bayesian classification is one of the classification methods successfully applied to the cancer diagnostic problems. The system stores the knowledge of the medical experts and the medical records of the previous cases. Based on the knowledge stored, the system will learn the patterns using Bayesian Analysis and decide the probability based on the symptoms of the patients. Classifier accuracy is also estimated to get the better decision support system with the minimum error rate by using Bayesian Analysis that provides a theoretical justification and lower error rate than other classifiers, [6].

Keywords

Decision support system, Bayesian classification, classifier accuracy, Naïve Bayesian Classification, Expert System, Lung Cancer Classification.

1. Introduction

The Simplex Bayesian Classification is widely used in Nave Bayes Classifier. With the increasing data and characteristics of biomedical and health-care ,we need to use methods which allow modeling the uncertainties that come with the problem and explicitly indicate statistical dependence and independence, and allow integrating biomedical and clinical background knowledge. Therefore, computer-based medical systems are playing an increasingly relevant role in assisting both diagnosis and treatment. With the current rapid increase in the amount of biomedical data being collected electronically in critical care and the wide-spread availability of cheap and reliable computing equipment, many researchers have already started, or eager to start, exploring these data.[4,6]

The term cancer covers more than a hundred diseases that share one trait: In all of the diseases, cells grow out to control and destroy healthy tissues. Lung cancer is the most common cancer and leading cause of cancer deaths worldwide. Majority of people, they come to the surgical outpatient department complain of either pain or lump in the lung. This system intends to develop the decision support system whether a patient has a lung cancer or not, and if she has the cancer what stage it is. Naive Bayes classifiers are widely used in Machine Learning. Indeed, they can efficiently be learned, they provide simple generative models of the data and they achieve pretty good results in various classification tasks. These classifications rely on the hypothesis that the attributes of the description domain are independent conditionally to each class. Bayesian Analysis is used to classify the disease and it needs the previously trained data. Those trained data are got from the medical experts and medical books and previous experiences of the system.

This paper organized as follows: Section 2.explains the related work. In Section 3 illustrates the implementation of the system . Section 4 present the model of Bayesian classification accuracy. Finally , we remark the conclusions in section.

2. Related work

A literature survey showed that there have been several studies on the medical problems using classification algorithms such as the classification methods, the linear discrimination analysis, the Bayesian analysis, and etc. For the last few years, researchers have started paying attention to the cancer classification using many theorems. A simple

Bayesian classifier (Duda & Hart, 1973; Langley, 1993) estimates the probability that a test instance is a member of each class and the test instance is assigned to the class with the highest probability [4,5,6].

Former researchers relied their machine learning algorithms on all 13th attributes in own dataset . The GCG guideline with a Naive Bayesian Classifier (NBC) is impleted, using only 13th attributes and tested with its performance in the measurements of accuracy, sensitivity and specificity [9]. For their application, they chose the problem of predicting recurrence in lung cancer, a set of 655 instances of real patient data with a binary outcome (Recurrence / No Recurrence) and 13 possible predictive attributes.

A group of Surveillance proposed system of prediction of survivability rate of lung cancer patients using three data mining techniques: the Naive Bayes, and the back-propagated neural network, algorithms [4].Bayesian classification method selected a subset of the available attributes, with which to build a Naive Bayes classifier . It is shown that such attributes are interdependent , especially when some attributes are redundant .

2.1 Proposed System

This system intends to implement the decision support system for classifying lung cancer and help in diagnosing it. It uses Bayesian classifier to classify the disease. Expert knowledge is collected from the medical experts, medical books and the system also learns from the past experiences. It will prepare the training data using the expert knowledge stored. When the new case (patient) arrives, the system will check the symptoms and find the probabilities of each class based on the attributes (symptoms). For example, the probability Attribute A, to become the class C is 0.2. Then it calculates with multiple attributes and finds the best probability and classifies the disease by the process flow of the proposed system architecture is shown in figure 1.

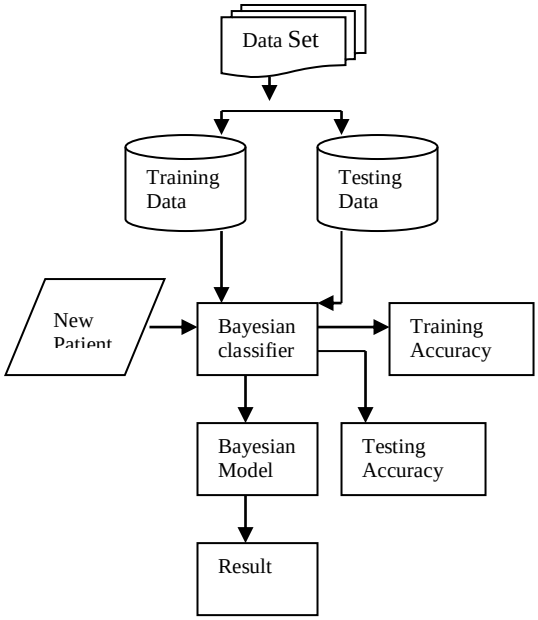

Figure 1. Proposed System Architecture

2.1.1 Bayesian Classification

Bayesian Classifiers are statistical classifiers. They can predict class membership probabilities, such as probability that a given sample belongs to a particular class. Bayesian classification is based on comparing classification algorithms. It can be described as follows:

Let X be a data sample whose class label is unknown and H be some hypothesis, such as that the data sample X belongs to a specific class C.

For classification problems, it needs to determine P (H|X), the probability that the hypothesis H holds the observed data sample X given.

The probability of a disease given by a Symptom P(d|s) is dependent on the probability of that anyone in the population has the disease P(d), has the symptom P(s) and the likelihood that is given by the disease the probability of having the symptom is P(s|d) [4,5,6].

$$P(d/s) = \frac{P(d)*P(s/d)}{P(s)} \qquad \text{- - - - - - - - -(1)}$$

2.1.2 Attributes and their values

In this system, there are 13 attributes, and most of the attributes are categorical values. All Data are collected from the medical books and doctors (medical experts) from Pathein Hospital, Pathein. The training data set has records.

@attribute age {'<35', '>50', '35-50'}
@attribute recurrence {no_rec, rec}
@attribute familyhistory {none, present}
@attribute massposition {lower, middle, upper}
@attribute massduration {longhistory, shorthistory}
@attribute growthrate {fast, slow}
@attribute masssize {'<3cm', '>5cm', '3-5cm'}
@attribute pain {painful, painless}
@attribute painduration {always, none, often, sometimes}
@attribute patientsymptoms {appetiteloss, none, present, weightloss}
@attribute cytology {pleuralfluid, sputum}
@attribute chestxray {BM (Broadening of mediastinum), ECS (Enlarged cardiac shadow), EH (Elevation of a hemidiaphragm) , L2SC (Lung, lobe or segmental collapse), PE (Pleural effusion), PPO (Peripheral pulmonary opacity), RD (Rib destruction), UHE (Unilateral hilar enlargement)}
@attribute percutanea {cancercell, noncancercell}
@attribute result {no, yes}

2.1.3 Sample Train Data

'<35', no_rec, present, upper, longhistory, fast, '<3cm', painless, none, present, sputum, UHE, cancercell, yes

'<35', no_rec, present, upper, longhistory, fast, '<3cm', painless, none, present, pleuralfluid, PPO, noncancercell, no

'35-50', rec, none, lower, shorthistory, slow, '3-5cm', painful, often, present, sputum, L2SC, cancercell, yes

'35-50', rec, none, lower, shorthistory, slow, '3-5cm', painful, often, present, pleuralfluid, PE, noncancercell, no

'>50', rec, present, middle, longhistory, fast, '>5cm', painless, sometimes, none, sputum, BM, cancercell, yes

'>50', rec, present, middle, longhistory, fast, '>5cm', painful, sometimes, none, pleuralfluid, ECS, noncancercell, no

'<35', rec, none, lower, shorthistory, slow, '<3cm', painless, none, present, sputum, EH, noncancercell, no

'<35', rec, none, middle, shorthistory, slow, '<3cm', painful, often, present, pleuralfluid, RD, cancercell, yes.

3. Implementation of the System

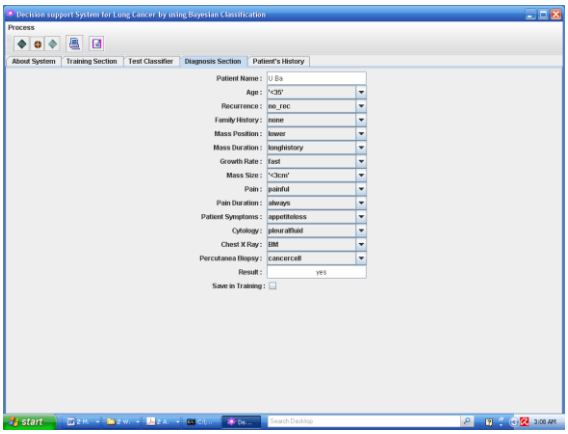

Figure 2. Diagnosis

3.1 Diagnosing the Patients

Figure 3 shows the diagnosis section of the system. When we fill the information needed in the system, we can see the result of the patient. After that we can explain the system for the next patient. The new patient information will be saved in the history records.

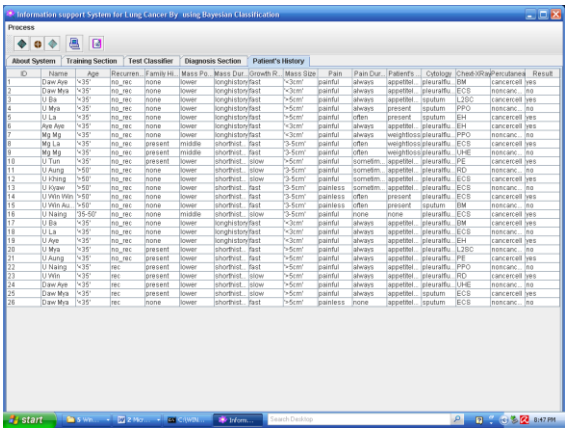

Figure 3. Patient History

4. Classifier Accuracy

Classification accuracy of Lung Cancer is important since it determines to evaluate how accurately a given classifier.

Accuracy estimates also help in the comparison of different classifiers.

Given : a collection of labeled record (training set)

Each record contains a set of attribute and the true class label.

Find : a rule for classification as function of the values of the features.

Goal : previously unseen records should be assigned a class as accurately as possible.

A test set is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it. In addition to accuracy, classifiers can be compared with respect to their speed, robustness (e.g., accuracy on noisy data), scalability, and interpretability.[4]

Sensitivity=t_pos/pos,

Specificity=t_neg/neg,

Precision=t_pos/(t_pos+f_pos)

Accuracy=Sensitivity (pos/(pos+neg)) + Specificity (neg/(pos+neg))

t_pos =the number of true positives

pos =the number of positives

t_neg =the number of true negatives

neg =the number of negatives

f_pos =the number of false positives

Correct classification: The known label of test sample is identical with the class result from the classification model.

Accuracy ratio: the percentage of test set samples that are correctly classified by the model.

Thus we maximize P (Ci | X). The class Ci for which P (Ci | X) is maximized is called the maximum posteriori hypothesis. By Bayes' theorem

$$P (Ci | X)= P (X | Ci) P (Ci) / P (X)$$

In this system, 655 records have been trained and 15 datasets have been ignored as unknown label since there is no data provided for the result field. When training with 'Training set' method, 220 records have been correctly classified and 16 have been incorrectly classified. So there is 97.00% of accuracy achieved. When data set are trained, 98.00% of accuracy have been achieved with cross validation and holdout method.

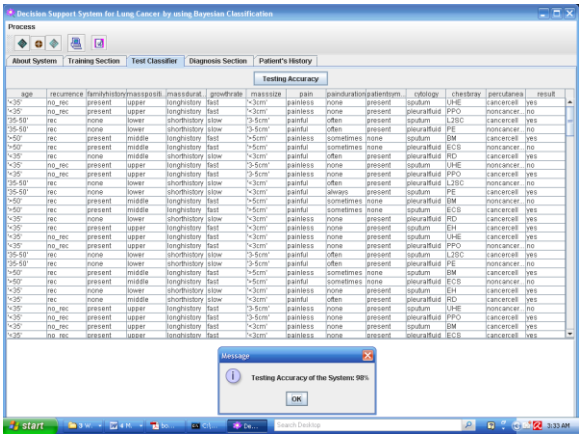

Figure 4. Testing data Accuracy

5. Conclusion

Naive Bayesian Analysis seems to be a suitable technique for medical knowledge based system. We have presented a Bayesian Analysis system, with the aim of supporting patients managements by symptom analysis in medical domain. The propose system can make the decision of a doctors. That means for achieving more accurate and effective medical diagnosis for lung cancer patient. At the present time, data mining is the powerful technology in computer science fields.

References

[1] Alan Rea, email A.Rea@qub.ac.uk Generated with CERN WebMaker All documents are the responsibility of, and copyright, © their authors and do not represent the views of The Parallel Computer Centre, nor of The Queen's University of Belfast.

[2] By Stuart J. Russell and Peter Norvig, Artificial Intelligence A Modern Approach, 2nd Ed, P-2, P-3, P14, P-16, P-17.

[3] Back to Computing Research: Driving Information Technology and the Information Industry Forward (http://cra.org/research.impact) Copyright 1996 by NEC Research Institute and the Computing Research Association. Contributions to this document by Tom Dean and Jon Doyle are appreciated.

[4] Han and Kamber, Jiawei Han, Micheline Kamber, Data Mining Concepts and Techniques, 2nd Ed.

Original ISBN : 978-1-55860-901-3, 2006, Elservier Inc, pp.7, pp.8, pp.48, pp.285, pp.289, pp.290, pp.310, pp.311, pp.319, pp.360.

[5] Jiawei Han Machelin Kamber , Data Mining: Concepts and Techniques, Semon Fraser University.P-296, P-297.

[6] K. T. Lynn, Case-Based Reasoning for Evaluation of Symptom Bronchogenic Carcinoma Patients, University of Computer Studies, Yangon

[7] Monte Carlo Resort, Las Vegas, Nevada, USA June 25-28, 2001

[8] N. Lavraˇc (1), E. Keravnou (2), and B. Zupan (3,1)

Department of Intelligent Systems, J. Stefan Institute Jamova 39, 1000 Ljubljana, Slovenia

Department of Computer Science, University of Cyprus P.O.Box 20537, CY-1678 Nicosia, Cyprus

Faculty of Computer and Information Sciences, University of Ljubljana,Trˇzaˇska 25, 1000 Ljubljana, Slovenia

[9] R. Livi and S. Cagnoni, "Time-Qualified Evaluation of Blood Pressure Excess," Proc. Second Ann. Symp. Computer-Based Medical Systems, CS Press, Los Alamitos, Calif., Order No, 1960, 1989, pp.124-129