Classification of Brain Cancer by Using Naïve Bayesian Classifier

Zin Mar Win Kyaw, Aung Htein Maw
University of Computer Studies, Yangon
good.wise6789@gmail.com

Abstract

Brain tumors typically behave in a unique fashion compared to neoplasm occurring elsewhere in the body. In particular, they rarely spread to other parts of the body and typically produce symptoms due to localized growth within the brain. For this reason, neoplasm in the brain, even malignant tumors, are typically referred to as "tumors" as opposed to "brain cancer". Classification is the process of finding a set model that describe and distinguished data classes or concepts for the purpose if being able to use the model to predict the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data. This system focuses on the brain cancer diseases by using Naïve Bayesian Classifier that classifies Benigh Tumor of Glial Cells and Glioma. This system can be mentioned treatment information for the patient.

1. Introduction

Classification is one of data mining approaches which are based on supervised learning. It is suited predicting or describing data set with binary or nominal categories. The objective of classification is to reduce detail and diversity of data and resulting information overwork by grouping similar data. A classification model can be used to predict the class label of unknown instants. The major classification approaches consists of decision tree, decision rules, K-nearest neighbors, Bayesian approaches, neural networks, regression-based methods and vector-based methods. In supervised machine learning, an induction algorithm is typically presented with a set of training instances where each instance is described by a vector of feature (of attribute) values, and a class label.

For the aim of classification the objects are mostly described by the values or certain attributes like education, punished, total services, etc. The value of an attribute will also be called a feature.

"Benign", brain tumors are typically grow in a well-circumscribed fashion and do not invade surrounding brain tissue or other structures. Despite their name , "benign" brain tumors can still be life or function threatening, particularly when they occur at the base of skull where they can affect normal structures like the brainstem, eyes and cranial nerves and where treatment with surgery or radiation may be difficult. "Malignant" brain tumors are those tumors with a propensity to grow quickly, invade normal brain and/or spread to the other parts of the neuroaxis. Low grade gliomas may be slow growing neoplasm but often recur despite conventional treatments and can exhibit malignant transformation over time [4].

In this paper, Naïve Bayesian Classifier for brain cancer diagnosis system is proposed. The rest of this paper is organized as follows. Section 2 describes the related work. Section 3 describes proposed system. Section 4 describes experimental result. Finally, concludes this paper in section 5.

2. Related Work

The basic feature selection problem is an optimization problem, with a performance measure for each subset of features to measure its ability to classify the samples. The problem is to search through the space of feature subsets to identify the optimal or near-optimal one with respect to the performance measure. Many successful feature selection algorithms have been devised. Yang and Honavar classify many existing approaches into three groups: exhaustive search, heuristic search and randomized search. Exhaustive search is a brute force approach where every possible subset is tested with the performance measure, and the best one is chosen. It guarantees the optimal subset as a result. However, if the number of feature is large, this approach is intractable. Heuristic search is where certain heuristics are used to greedily but intelligently search through the subset space to identify a subset with a reasonable performance measure. Forword Selection and Backward Elimination are two examples of heuristic search [5].

Langly and Sage use the forward sequential selection method to select a subset of the available attributes, with which to build a Naïve Bayes Classifier. It is shown that such attributes are interdependent, especially when some attributes are interdependent, especially when some attributes are redundant. Searching in the space of feature subsets has been studied for many years [7]. Sequential backward elimination, sometimes called sequential backward selection. Several authors have examined the use of heuristic search for feature subset selection. Others have used randomized and randomized population based heuristic search techniques such as genetic algorithms to select feature subsets for use with decision tree or nearest neighbor classifiers. The solution to the feature selection problems is neither trivial, nor unique. The set of optimal features can be different for different hypothesis space. In distinguishes between two models of selection an 'optimal' set of feature under some objective function [2]. Langley & Sage used the wrapper approach to select features for Naïve Bayes and used it to select features for nearest-neighbor algorithm. Pazzani used the wrapper approach to select features and join features for Naïve-Bayes and showed that it indeed finds correct combinations when features interact [7].

3. Proposed System

3.1. Data Mining System

Data Mining refers to extraction or mining knowledge from large amount of data. Data mining is a process that uses a variety of data analysis tools to discover patterns and relationships in data that may be used to make valid predictions. Data mining is a task of discovering interesting patterns from large amount of data can be stored in databases, data warehouses or other information repositories, drawing from area such as machine learning, information retrieval and high performance computing. This can be viewed as a result of the natural evolution of information technology.

3.2 Bayesian Classification

Bayesian classifiers are statistical classifier. They can predict class membership probabilities, such as the probability that a given sample belongs to a particular class. Bayesian classification is based on Bayes Theorem. Bayesian classifiers have also exhibited high accuracy and speed when applied to large datasets. Naïve Bayesian classifier assumes that the effect of an attribute value on a given class is independent of the values of the other attributes. It is called class conditional independence.

3.3 Naïve Bayesian Classification

The probability of a disease given 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 given the disease the probability of having the symptom is P(s|d).

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

3.4 Process Flow of the Naïve Bayesian Classifier

The processes of this system using Naïve Bayesian analysis is as follows:

1. User symptoms are applied to the system through the user interface.
2. Those symptoms are changed to attribute vector in order to apply to Naïve Bayesian classifier to classify the disease.

   P(Ci|X)>P(Ci|X)for i <=j <= m , j!=i

3. Thus we maximize P(Ci|X). The class Ci for which P(Ci|X) is maximized is called the maximum posterior hypothesis. As P(X) is constant for all classes, only P(X|Ci)P(Ci) need be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is P(C1)=P(C2)=…=P(Cm) and we would therefore maximize P(X|Ci)P(Ci). Note that the class prior probabilities may be estimated by P(Ci)=si/s, where si is the number of the training samples of class Ci, and s is the total number of training samples.
4. Given data set s with many attributes, it would extremly computationally expensive to compute P(X|Ci). In order to reduce computation in evaluating P(X|Ci), the naïve assumption of the class conditional independence is made.. Thus,

   $$P(X \mid C_i) = \prod_{k=1}^{n} P(X \mid C_i)$$

   The probabilities P(X|Ci), P(X2|Ci),…, P(Xn|Ci) can be estimated from the training samples, where
   (a) If Ak is categorical , then P(Xk|Ci)= ,where sik is the number of training samples of class Ci having the value xk for Ak and si is the number of training samples belonging to Ci .
   (b) If is continuous-valued, then the attribute is typically assumed to have a Gaussian distribution so that

   $$P(X_k \mid C_i) = g(x_k, \mu_{C_i}, \sigma_{C_i}) = \frac{1}{\sqrt{2\pi}\sigma_{C_i}} e^{\frac{(x_2-\mu_{c_i})^2}{2\sigma^2 c_i}}$$

5. In order to classify an unknown sample X, P(X|Ci) P(Ci) is evaluated for each class Ci. Sample X is then assigned to the class Ci if and only if

   $$P(X \mid C_i)P(C_i) > P(X \mid C_i)P(C_i) \quad for\ 1 \le f \le m, j \ne i$$

   In order words, it is assigned to the class Ci for which P(X|Ci) P(Ci) is the maximum.

3.4.2. Apply Naïve Bayesian Classifier

Using Naïve Bayesian Classification

(1)

X=(Headache=”Migraine”,Seizure=“Cyanosis”, Vision=”Diplopia”,CNSProblem=”Hemiplegia”, Brain Location =”Frontal Lobe”, Cerebral Edema =”Congnitive Impairment”, Fever =”None”, Risk Factor =”Severe Head Injury”, Differentiation =”Poor”, Deep Vein Thrombosis =”Swelling”)

For C1, P(PDX=”Glioma”) =6/10 =0.6

For C2, P(PDX=”Benigh Tumor of Glial Cells”) =4/10 =0.4

For (X| Ci) ,

P(Headache=”Migraine”|PDX=”Glioma”) =1/6=0.167

P(Headache =”Migraine”| PDX=”Benigh Tumor of Glial Cells”)=2/4=0.5

P(Seizure=”Cyanosis”|PDX=”Glioma”) =2/6=0.333

P(Seizure =”Cyanosis” | PDX=”Benigh Tumor of Glial Cells”)=1/4=0.25

P(Vision=”Diplopia”|PDX=”Glioma”) =2/6=0.333

P(Vision =”Diplopia” | PDX=”Benigh Tumor of Glial Cells”)=3/4=0.75

P(CNSProblem=”Hemiplegia”|PDX=”Glioma”) =3/6=0.5

P(CNSProblem =”Hemiplegia” | PDX=”Benigh Tumor of Glial Cells”)=0/4=0

P(BrainLocation=”FrontalLobe”| PDX=”Glioma”)=5/6=0.833

P(BrainLocation=”Frontal Lobe”| PDX=”Benigh Tumor of Glial Cells”)=1/4=0.25

P(Cerebral Edema =”Congnitive Impairment” | PDX=”Glioma”)=6/6=0.833

P(CerebralEdema=”CongnitiveImpairment”| PDX=”BenighTumorofGlialCells”) =1/4=0.25

P(Fever =”None” | PDX=”Glioma”)=6/6=1

P(Fever =”None”| PDX=”Benigh Tumor of Glial Cells”)=1/4=0.25

P(Risk Factor =”Severe Head Injury” | PDX=”Glioma”)=3/6=0.5

P(RiskFactor=”SevereHeadInjury”| PDX=”BenighTumorof Glial Cells”)=0/4=0

P(Differentiation=”Poor”|PDX=”Glioma”) =3/6=0.5

P(Differentiation=”Poor”| PDX=”Benigh Tumor of Glial Cells”)=0/4=0

P(Deep Vein Thrombosis =”Swelling” | PDX=”Glioma”)=2/6=0.333

P(Deep Vein Thrombosis =”Swelling” | PDX=”Benigh Tumor of Glial Cells”)=1/4=0.25

P(X|PDX=”Glioma”)=0.167*0.333*0.333*0.5*0 .833*1*1*0.5*0.5*0.333=0.001

P(X| PDX=”Benigh Tumor of Glial Cells”)=0

P(X|PDX=”Glioma”) P(PDX=”Glioma”)=0.001*0.6=0.001

P(X|PDX=”Benigh Tumor of Glial Cells”)P(PDX=”Benigh Tumor of Glial Cells”)=0

Result =Glioma

3.4.3 Classifier Accuracy

Estimating classifier accuracy is important since it determines to evaluate how accuracy agiven classifier will label future data, data on which the classifier has not been trained. Accuracy estimates also help in the comparison of different classifiers. The following Classification features are used to train and test the classifier.

Given: a collection of labeled records (training set). Each record contains a set of features (attributes) and the true class (label).

Find: a model for the class as a function of the values of the features.

Goals: 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.

The Sensitivity and Specificity measures can be used to determine the accuracy measures. Precision may also be used to access the percentage of samples labeled as example, ”cancer” that actually are “cancer” samples. These measures are defined as

Sensitivity = t-pos/pos

Specificity = t-neg/neg

Precision = t-pos/(t-pos+f-pos)

Where,

t-pos = the number of true positives(“cancer” samples that were correctly classified as such)

pos = the number of positive (“cancer”) samples

t-neg = the number of true negative(“not cancer” samples that were correctly classified as such)

neg = the number of negative samples

f-pos = the number of false postive(“not cancer” samples that were incorrectly classified as such)

Accracy= sensitivity [pos/(pos+neg)] + specificity [neg/(pos+neg)]

4. Design and Implementation

System Design

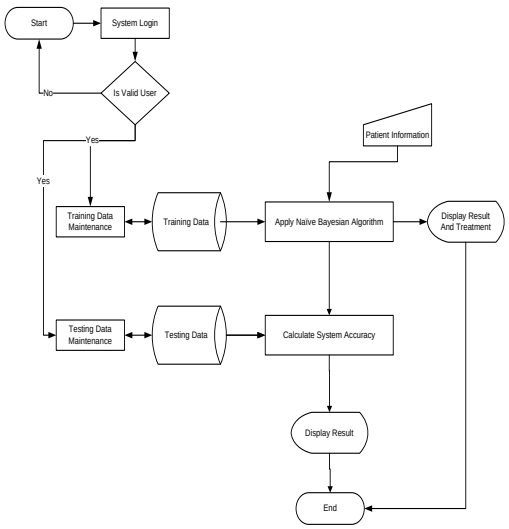

Figure 1. System Flow Diagram

This system has ten categories and two classes. Ten categories are Headache, Seizure, Vision, CNS Problem, Brain Location, Cerebral Edema, Fever, Risk Factor, Differentication, and Deep Vein Thrombosis. Two classes are Glioma and Benigh Tumor of Glail Cells. Training

dataset collects from Yangon General Hospital, North Okkalapa Hospital and Internet download. Training dataset has over two thousands records. Testing dataset has under one thousand of records. This system can help the medical students and the junior doctors when they classified diagnosis problem in brain cancer cases. In this cases, medical students and doctors can save time in diagnosis problems.

5. Conclusion

Solving the cancer diagnosis problem is a complex task. Computer-based classification systems will play an increasingly important role in solving brain cancer diagnosis problem. Cancer classifications are certainly important to the future of medical fields as it might brighten the medical perspective in cancer treatment. Naïve Bayesian Classifier can be used to test on unseen data and doctors can save time in disease diagnosis. When disease is able to be diagnosed earlier, it also means that the patients' chances of survival are higher. The system improves the quality of the diagnosis process in accuracy. In addition, this system can help the junior doctor when they classified diagnosis problem in emergency cases. The proposed system is simple to use and can helpful in real world Brain Cancer diagnosis system.

References

[1] Christina Wallin," TJHSST Computer Systems Lab SeniorResearch Project Naive Bayes Classification"

[2] D.W.Aha ,& Bankert, R.L,"A comparative evaluation of sequential feature selection algorithms", in D.fisher & H.Lenz,eds, Proceedings of the fifth International Workshop on Artifical Intelligence and statistics, FT.Lauderdale, FL, pp.starry. Stanford.EDU:pub/ronnyk/tables.ps,1995.

[3] Funny introduction to Bayesian statistics http://yudkowsky.net/bayes/bayes.html

[4] J.Yang and Honavar, "Feature subset selection using a genetic algorithm", in processings of the genetic programming conference.1997,pages 380.Stanford, CA.

[5] Karteek Popuri, Dana Cobzas, Martin Jagersand, Sirish L. Shah and Albert Murtha," 3D variational brain tumor segmentation on a clustered feature set"

[6] P.Langley, "selection of Relevant Features in Machine Learning", in Proceedings of the AAAI Fall Symposium or Relevance, pages,1-5,1994.

[7] Pompe, U., & Kononenko, I. (1995). Naive Bayesian classifier within ILP-R. Proceedings of the 5th International Workshop *on* Inductive Logic Programming (pp. 417–436). Department of Computer Science, Katholieke Universiteit Leuven.

[8] Robert and Casella 2004, Monte Carlo Statistical Methods, Springer.

[9] "Yoshimasa Tsuruoka and Jun'ichi Tsujii**,"** Training a Naive Bayes Classifier via the EM Algorithm with a Class Distribution Constraint"

[10] http://www.cancerinfotips.com/brain tumors sysptoms.html