Tumor Detection and Classification

Bhupesh Deka[1], Santosh Ku. Satapathy[2], Dr. Aurobindo Kar[3]

[1,3]Associate Professor, Department of Computer Science Engineering, Gandhi Institute For Technology (GIFT), Bhubaneswar
[2]Assistant Professor, Department of Computer Science Engineering, Gandhi Engineering College, Bhubaneswar

Publishing Date: August 27, 2015

Abstract

Data mining is a powerful method for mining useful patterns or data from image and textual data sets. Medical data mining is very important field as it has significant utility in healthcare domain in the real world. Clustering and Classification are the popular data mining methods used to understand the different features of the health data set. This paper is focused on understanding different techniques for thedetection of tumor which is an essential decision making feature and is a part of healthcare application. Tumor is a life threatening disease which produces problems like brain damage, loss of memory etc. There exist various data mining techniques for early assessment of tumor from datasets. The main idea of project is to present an overview about tumor detection system and various data mining methods.

Keywords: Tumor detection, data mining, clustering, K nearest neighbor.

1.INTRODUCTION

1.1 TumorDetection

Tumor detection system is one of the health care applications and it is essential for early stage detection of tumor. It is a software based application and it is used for better decision making in health care industry. Tumor detection system will make an early diagnosis of the disease based on several methods like data mining, machine learning etc. Most of the existing system consists of training part and testing part for detecting the disease. And it uses datasets as input data and train data. The system may consist of pre- processing stage and diagnosis stage. In pre-processing stage the training and testing datasets are subjected to various classification techniques for enhancing their quality. After that this enhanced datasets are subjected to featureextraction.

1.2 Data mining

Data Mining is defined as extracting information from huge sets of data. In other words, we can say that data mining is the procedure of mining knowledge from data. The information or knowledge extracted so can be used for any of the following applications −

- MarketAnalysis
- FraudDetection
- CustomerRetention
- ProductionControl
- ScienceExploration
- Medicalfield

1.3. Classification

Classification is the process of predicting the class of given data points. Classes are sometimes called as targets/ labels or categories. Classification predictive modeling is the task of approximating a mapping function (f) from input variables (X) to discrete output variables (y)..Classification belongs to the categoryof supervised learning where the targets also provided with the input data. There are many applications in classification in many domains such as in credit approval, medical diagnosis, target marketingetc.

There are two types of learners in classification as lazy learners and eager learners.

Lazy learners

Lazy learners simply store the training data and wait until a testing data appear. When it does, classification is conducted based on the most related data in the stored training data. Compared to eager learners, lazy learners have less training time but more time in predicting . Ex. k-nearest neighbor, Case-based reasoning

Eager learners

Eager learners construct a classification model based on the given training data before receiving data for classification. It must be able to commit to a single 6 hypothesis that covers the entire instance space. Due to the model construction, eager learners take a long time for train and less time topredict.

Ex. Decision Tree, Naive Bayes, Artificial Neural Networks

Here we have used two datasets:

- Pima Indian Datasets
- P300Datasets

2.OVERVIEW OF PROPOSED ALGORITHM

2.1 KNN

The k-nearest neighbor algorithm(k-NN) is anon-parametric method used for classification and regression In both cases, the input consists of the k closest training examples in the feature space . The output depends on whether k-NN is used for classification orregression:

- In k-NN classification, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer , typically small). If k = 1, then the object is simply assigned to the class of that single nearestneighbor.

- In k-NN regression , the output is the property value for the object. This value is the average of the values of its k nearest neighbors. k-NN is a type of instance-based learning, or lazy learning , where the function is only approximated locally and all computation is deferred until classification. The k-NN algorithm is among the simplest of all machinelearning

algorithms. Both for classification and regression, a useful technique can be used to assign weight to the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones. Forexample,

a common weighting scheme consists in giving each neighbor a weight of 1/d, where d is the distance to theneighbor.

The neighbors are taken from a set ofobjects for which the class (for k-NN classification) or the object property value (for k-NN regression) is known. This can be thought of as the training set for the algorithm, though no explicit training step is required. A peculiarity of the k-NN algorithm is that it is sensitive to the local structure of thedata.

K-Nearest Neighbors is one of the most basic yet essential classification algorithms in Machine Learning. It belongs to the supervised learning domain and finds intense application in pattern recognition, data mining and intrusiondetection.

It is widely disposable in real-life scenarios since it is non-parametric, meaning, it does not make any underlying assumptions about the distribution of data (as opposed to other algorithms such as which assume a Gaussian distribution of the givendata.

We are given some prior data (also called training data), which classifies coordinates into groups identified by an attribute.

KNN Algorithm :

Let m be the number of training data samples. Let p be an unknown point.

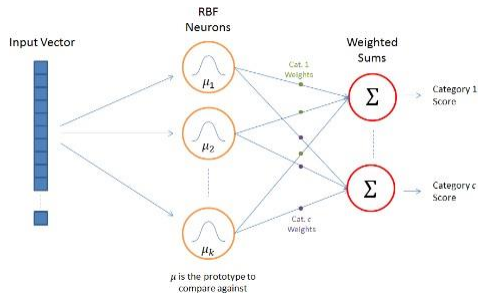
1. Store the training samples in an array of data points arr[]. This means each element of this array represents a tuple (x, y).
2. For i=0 tom:
3. Calculate Euclidean distance d(arr[i], p).
4. Make set S of K smallest distances obtained. Each of these distances correspond to an already classified datapoint.
5. Return the majority label amongS.

2.2 RBFN

Overview

A Radial Basis Function Network (RBFN) is a particular type of neural network. In this project, it's use as a non-linear classifier. Generally, when people talk about neural networks or "Artificial Neural Networks" they are referring to the Multilayer Perceptron (MLP). Each neuron in an MLP takes the weighted sum of its input values. That is, each input value is multiplied by a coefficient, and the results are all summed together. A single MLP neuron is a simple linear classifier, but complex non-linear classifiers can be built by combining these neurons into a network.The RBFN approach is more intuitive than the MLP. An RBFN performs classification by measuring the input's similarity to examples from the training set. Each RBFN neuron stores a "prototype", which is just one of the examples from the training set. When we want to classify a new input, each neuron computes the Euclidean distance between the input and its prototype. Roughly speaking, if the input more closely resembles the class A prototypes than the class B prototypes, it is classified as classA.

RBF Network Architecture



RBF Neuron Activation Function

Each RBF neuron computes a measure of the similarity between the input and its prototype vector (taken from the training set). Inputvectors which are more similar to the prototype return a result closerto 1. There are different possiblechoices of similarity functions, but the most popular is based on the Gaussian . Below is the equation for a Gaussian with a one-dimensionalinput.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi^e}} - \frac{(x-\mu)^2}{2\sigma^2}$$

Where x is the input, mu is the mean, and sigma is the standard deviation. This produces the familiar bell curve shown below, which is centered at the mean, mu (in the below plot the mean is 5 and sigma is 1).

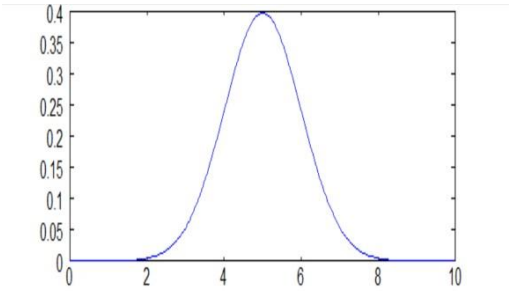


Fig.6: Activation Function

The RBF neuron activation function is slightly different, and is typically written as:

$$\emptyset(x) = e^{-\beta||x-\mu||^2}$$

k-means clustering algorithm

$$J(V) = \sum_{i=1}^{c}\sum_{j=1}^{c_i}(||x_i - v_j||)^2$$

where,

'||xi - vj||' is the Euclidean distance between xi and vj.

'ci' is the number of data points in ith cluster.

'c' is the number of cluster centers.

2.3 NaiveBayes

Naive Bayes is among one of the most simple and powerful algorithms forclassification based on Bayes' Theorem with an assumption of independence among predictors. Naive Bayes model is easy to build and particularly useful for very large data sets.

The Naive Bayes classifier assumes that the presence of a feature in a class is unrelated to any other feature. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that a particular fruit is an apple or an orange or a banana and that is why it is known as"Naive".

CONCLUSION

In this project, we are detecting and classifying tumor by using datasets. We have applied data mining techniques in this project for tumor detection. we have implemented three classification techniques (k-nearest-neighbors,Radial
Basis Function and Naïve Bayes) by using of two different datasets. After testing the model , we found accuracy of each techniques. The accuracy of KNN,RBFNand Naïve Bayes are 50-55%,90- 95% and 80-85% respectively. We got highest accuracy when are using RBFNalgorithm.
Then after we are plotting graphs for two datasets and three classification techniques.

FUTURE SCOPE

The use of data mining in medical field is fairly new development. Our current model is done primarily on simple numeric and categorical data. In the future model will include more complex data types. In addition, our model that has been designed, further refinement is possible by examining other variable and their relationship.

Future study of this project will focus on different evolutionary learning algorithms for Neural Network, which may further increase the classification accuracy of the model.
Further the performance of the network will be compared with some other known classifiers like SVM, KNN, Bayesian network etc.

Given a Hypothesis H and evidence E, Bayes' Theorem states that the relationship between the probability of Hypothesisbefore getting the evidence P(H) and the probability of the hypothesis after getting the evidence P(H|E) is:

$$P(H|E) = \frac{P(E|H).P(H)}{P(E)}$$

This relates the probability of the hypothesis before getting the evidence P(H), to the probability of the hypothesis after getting the evidence, P(H|E). For this reason, is called the prior probability, while P(H|E) is called the posterior probability. The factor that relates the two, P(H|E) / P(E), is called the likelihood ratio. Using these terms, Bayes' theorem can be rephrasedas:
"The posterior probability equals the prior probability times the likelihood ratio."

Reference

[1] Han, amber and Pei, "DataMining: Concept and Techniques",2013

[2] Mitchell T.M,"MachineLearning", McGraw-Hill,1997

[3] JemalA,MurrarT,WardE,SamuelsA,TiwariRC,GhafoorA,FeuerEJ,ThunMJ.Cancer statistics,2005."CA:a cancer journal fpr clinicians",2005 Jan1;55(1):10- 30

[4] Egyptian Informatics Journal, "Feature selection and classification system for chronic diseaseprediction".

[5] S. K. Nanda, D. P. Tripathy, S. S. Mahapatra, "Application of Legendre Neural Network for Air Quality Prediction", the 5th PSU - UNS International Conference on Engineering and Technology, pp. 267-272,2011