AUTOMATIC DETECTION
OF
DIABETES DIAGNOSIS

Being diagnosed with diabetes can be a very scary thing,
and it can easily make your life stand still for a moment.

Pattern Recognition and Machine Intelligence in Medicince
Spring semester, 2014–2015

by

ANIKET BHUSHAN
13MF10006
Manufacturing Science and Engineering
Indian Institute of Technology
Kharagpur

0.1 Introduction

Diabetes has been recognized as a continuing health challenge for the twenty-first century, both in developed and developing countries. It is understood that diabetes prevalence is increased because of modern lifestyles, urbanization, and economic development. So, Diabetes disease diagnosis via proper interpretation of the Diabetes data is an important classification problem. Diabetes develops when the body doesnt make enough insulin or is not able to use insulin effectively, or both. As a result, glucose builds up in the blood instead of being absorbed by cells in the body. The bodys cells are then starved of energy despite high blood glucose levels. Over time, high blood glucose damages nerves and blood vessels, leading to complications such as heart disease, stroke, kidney disease, blindness, dental disease, and amputations. Other complications of diabetes may include increased susceptibility to other diseases, loss of mobility with aging, depression, and pregnancy problems.

Many previous works has been done in this field using various classification techniques Support Vector Machines (SVMs), Neural networks, Bayesian classifcation has been studied increasingly in recent years.And also it was applied to the problem of diagnosis of Diabetes diseases in several works due to outstanding characteristics and excellent generalization performance.

2 class Diabetes Disease Dataset : I have used the UCI Diabetes diseases dataset introduced by Black C.L. (Blake C. L., 1998). This dataset contains 768 samples, where each sample has 8 features which are eight clinical findings:

1. Number of times pregnant
2. Plasma glucose concentration a 2 h in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-hour serum insulin (mu U/ml)
6. Body mass index
7. Age (years)

0.2 Methodology

In my project work i have tried to carry out the detection of diabetes diagnosis through the use of Naive Bayes classifier. But since for applying Naiye Bayes classifier feature vectors should be independent, a proper routine of PCA is carried out which not only insures independency of the features but also carries out dimensionality reduction.

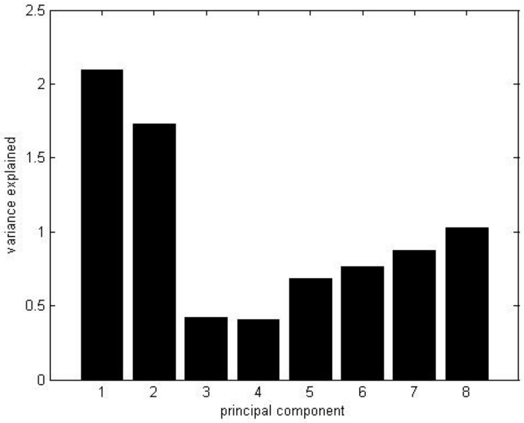
The feature vector revieved from the above database has dependent features and ofcourse some noise. So, a feature evaluation is necessary to extract good features. t-test is carried out for the determination of statically significant features.

0.3 Work-flow chart

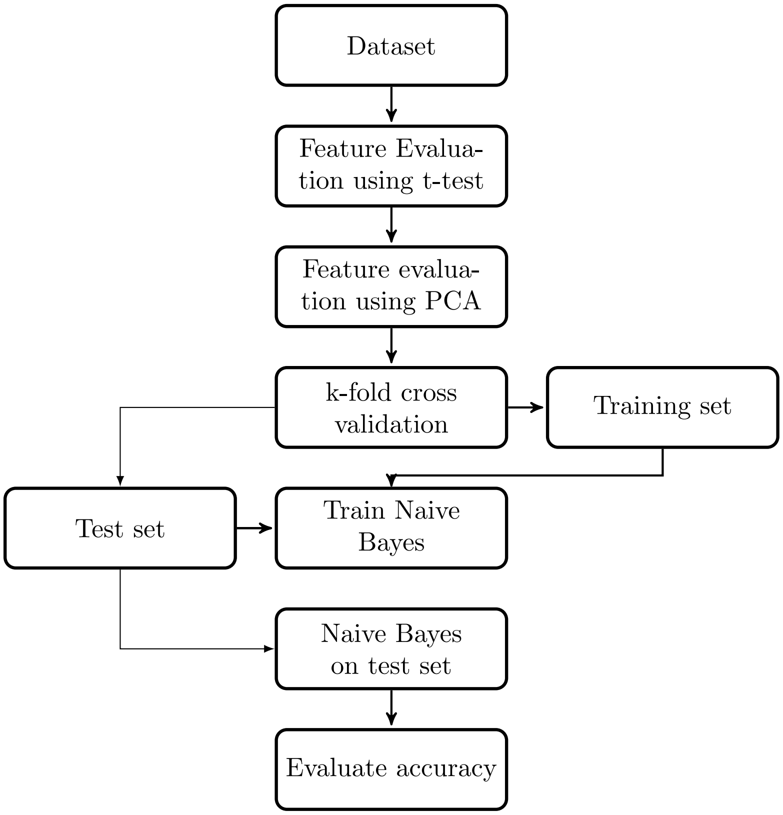
From the given dataset for the determination of statically significant features t-test is being carried out from which ranking of all the features are obtained on their p-value basis. From the correlation matrix of given feature vector it can be concluded that the features are not independent , so Naive Bayes classifier cannot be directly used. Hence to remove this dependency and also to reduce the useless feature vectors PCA is being carried out in my work.

Out of the 8 principal components 6 principal components were chosen to be used for the classifcation purpose. 6 principal components were chosen which have 6 highest variance retained, such that total 89 % of the variance is retained.

After obtaining the reduced dataset Naive Bayes classification can be easily applied because the features are independent.And to evaluate the accuracy of the Naive bayes classfication on the data set whole data set is divided in to 10 partition, such each time 9 partitions are taken as training set and remaining as test set. So, over a total loop of 10 times we get average performance of Naive Bayes classifcation.



1



0.4 Result

The Naive Bayes models has done good classification with classifcation accuracy of : 75.13 %.
Confusion matrix :

	Diabetic(Predicted)	Non - Diabetic(Predicted)
Diabetic(Actual)	43 = TP	7 = FN
Non-Diabetic(Actual)	12 = FP	15 = TN

sensitivity : 78.03 %
specificity : 67.74 %