Data-based Prediction System

Cancer research as a case study

By Uria Noiman and Mordechai Ben Zecharia | Industrial Engineering and Management department in JCT | Supervisor, Prof Avi Rosenfeld

Subject

In this project, we investigated a specific type of cancer.
We tried to determine whether:
The main causes of the disease can be characterized?
It is possible to classify with high accuracy, the probability that a given person will be ill or not based on software that studies and analyzes big databases?

Motivation

Early detection of cancer can promote healing, therefore learning technology for detection is desirable.
In addition, the prediction system can also be used in other areas, whether for security needs such as early identification of potential terrorist, business needs such as customer characterization.

Challenges

1. How to prepare the initial data so it can be analyzed?
2. Is it possible to find the main attributes and classify with high probability into one of two groups based on the given databases?
3. How to implement the prediction system to find the best results?
4. How to implement the prediction system in a generic manner that enables implementation in a wide range of fields?

Conclusions

1. We found that it is possible to classify with high accuracy based on those databases.
2. It is possible to find key parameters with high effect on the classification type.
3. Although this prediction system was developed for this field it is reasoneble to assume the system will work for other fields, therefore, it is worth developing and marketing it in the future.
4. An article should be published by the researcher who provided the databases.

1 Pre-Processing

The new data required pre-processing, and therefore a preliminary stage was needed to eliminate the irrelevant information and turn the structure into a uniform structure that could be worked with.

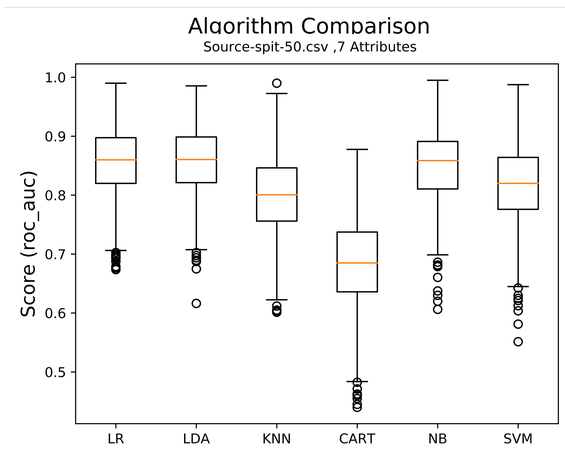
2 Data Mining

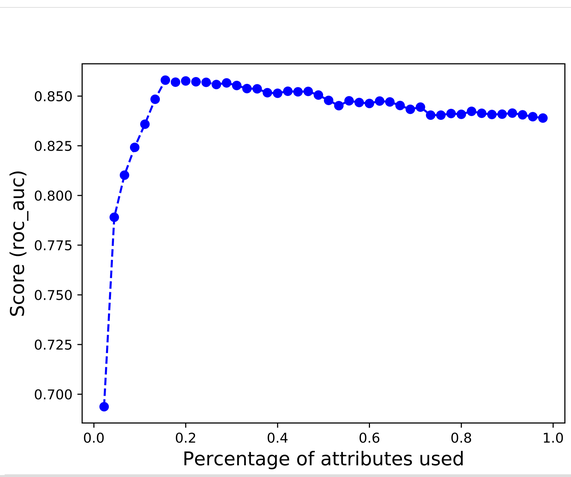
A sophisticated algorithm studies the data and generates an analysis report. On a parallel track, manual research by data scientist is performed using manual software for data mining.

3 Analysis Results

Each track, manual and algorithmic, performs an independent analysis of the data, understands the most important attributes and chooses the most appropriate model for cancer predicting.

$$f(x) = \frac{1}{1 + e^{-(\beta_0 - \beta_{1x})}}$$





4 Model Verification

Comparison of the results of the two tracks (from the previous stage) is performed, and afterwords, a prediction is performed on a test group using both tracks, if there are delta between the two tracks, an analysis is performed and another attempt is made to improve the model.

		Accuracy		ROC	
Learning Task	Attributes	Weka	Python Algo	Weka	Python Algo
Spit12->Spit12	12	77.12%	78.37%	0.84	0.85
Spit->Best	12	83.12%	82.22%	0.90	0.89
Best12->Best12	12	84.02%	84.28%	0.91	0.91
Best->Spit	12	74.87%	75.38%	0.85	0.85
Spit25->Spit25	25	79.65%	77.13%	0.86	0.84
Best25->Best25	25	85.18%	84.81%	0.91	0.91

5 Prediction System

Based on the chosen model, a user-friendly web application is built, enabling data entry and forecasting according to the data entered.