Disease Predictor — Symptom-based Big Data Analysis

VM455 Final Presentation

Group member:
Sheng Cen
Jiajin Wu

2020.12.7

Problem Statement

- Automatic disease predictor given the input of occurring symptoms
- Early diagnosis

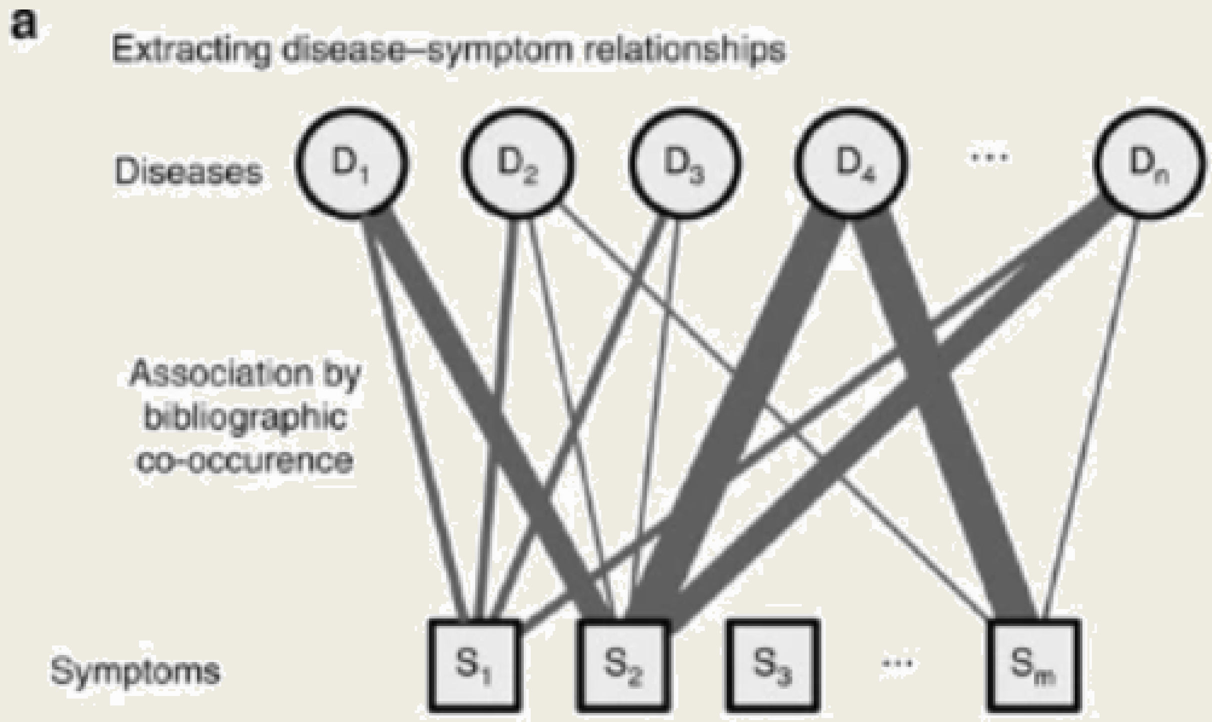


Figure 1: Extracting the disease–symptom relationships from PubMed bibliographic literature database.

Problem Statement

Dataset (Patient Cases)

Disease	Symptom_1	Symptom_2	Symptom_3	Symptom_4
Fungal infection	itching	skin_rash	nodal_skin_eruptions	dischromic_patches
Fungal infection	skin_rash	nodal_skin_eruptions	dischromic_patches	
Fungal infection	itching	nodal_skin_eruptions	dischromic_patches	

- Patients report the disease he/she caught, as well as related symptoms.
- Given such large dataset, predict the most possible disease given a combination of related symptoms.

Methods and Data

Data pre-processing:

- Correlation analysis among symptoms using correlation matrix, to reduce the redundant variables, and transform to continuous ones instead of fuzzy variables.
- Principle Component Analysis: further reduce the dimensionality of variables; faster for fitting later model.

Model training and result analysis:

- Use different training methods (SVM, decision trees w/o boosting/bagging) for classification.
- W/o prior Principle Component Analysis.
- Cross validation parameter tuning.
- Model accuracy; Training speed...

Methods and Data

Raw dataset:

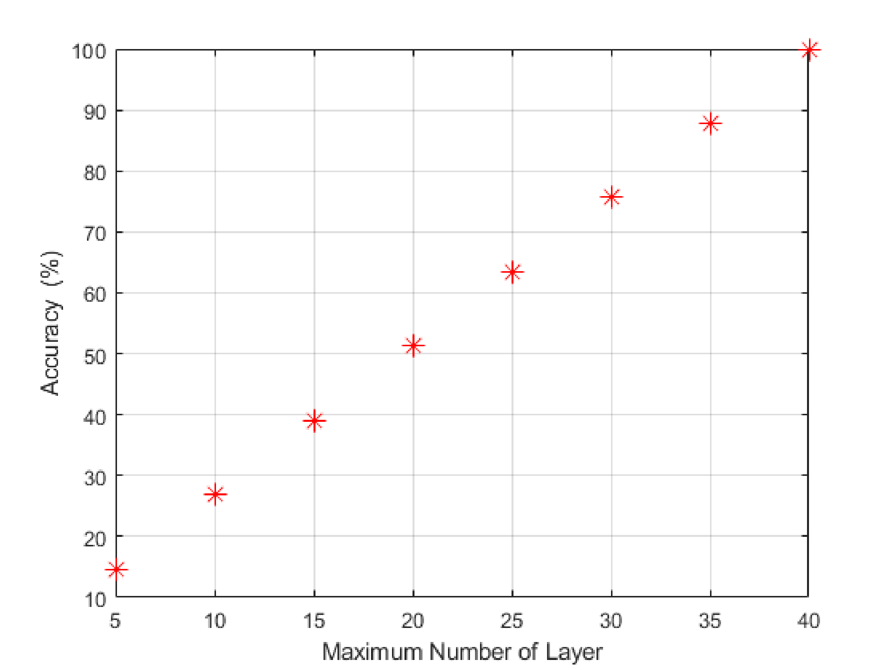
- 4920 patient cases for training
- 133 kinds of diseases
- 42 kinds of symptoms

Disease	Symptom_1	Symptom_2	Symptom_3	Symptom_4
Fungal infection	itching	skin_rash	nodal_skin_eruptions	dischromic _patches
Fungal infection	skin_rash	nodal_skin_eruptions	dischromic _patches	
Fungal infection	itching	nodal_skin_eruptions	dischromic _patches	

Figure 2: Case record for patients

Result

	Accuracy	Training Time
SVM Linear	100 %	83 sec
Decision Tree	12.2 ~ 99.1 % (depending on max. layer #)	0.65 ~ 2 sec (depending on max. split #)
Boosting Tree	99.7 %	16 sec
Bagging Tree	49.1 %	7 sec

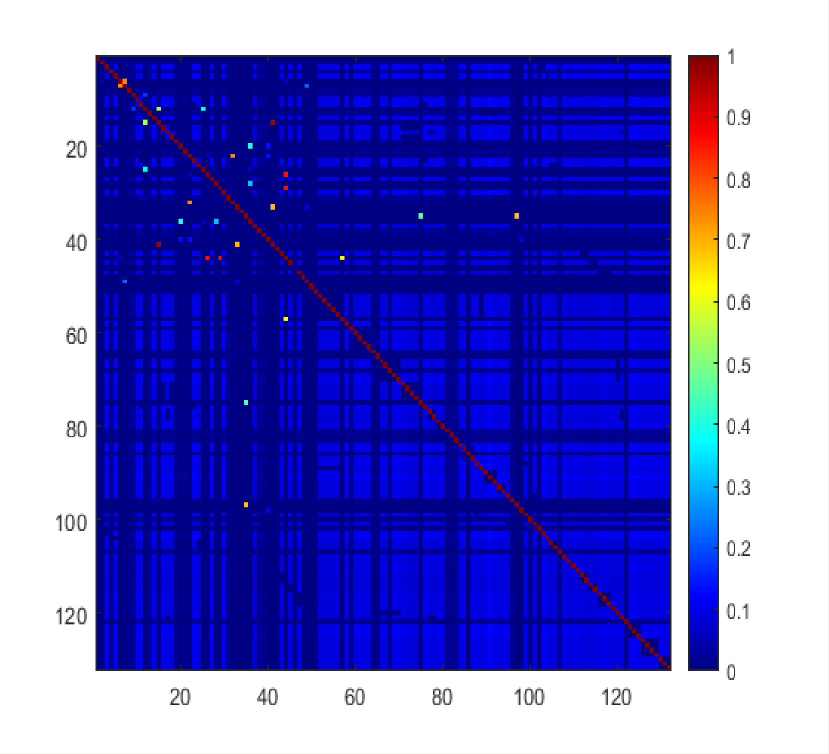


- SVM: accurate, very slow;
- Decision Tree: very fast, flexible to use;
- Boosting Tree: relatively fast, accurate;
- Bagging Tree: inaccurate, fast

- The accuracy of decision tree is proportional to the maximum split number.

Result

- Reduce the variables with high correlation (apply hypothesis test to obtain the p-value. E.g.: eliminating the correlated features with $p > 0.5$)



Matrix of p-value between elements

- To reduce the dimensionality.
(from 133 to 113 symptoms)

- Save the time without too much decrease in accuracy.
(from 49.3% to 49.0%).

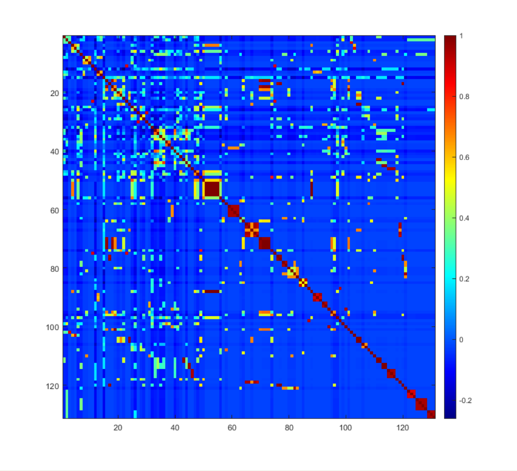
- Similar to the idea of PCA, but they are different methods.

- PCA: from 133 to 36 components, 49.3% to 51.2% accuracy.

Result

Discrete matrix
composed of only 0
(no such symptom)
and 1
(have such symptom)

×



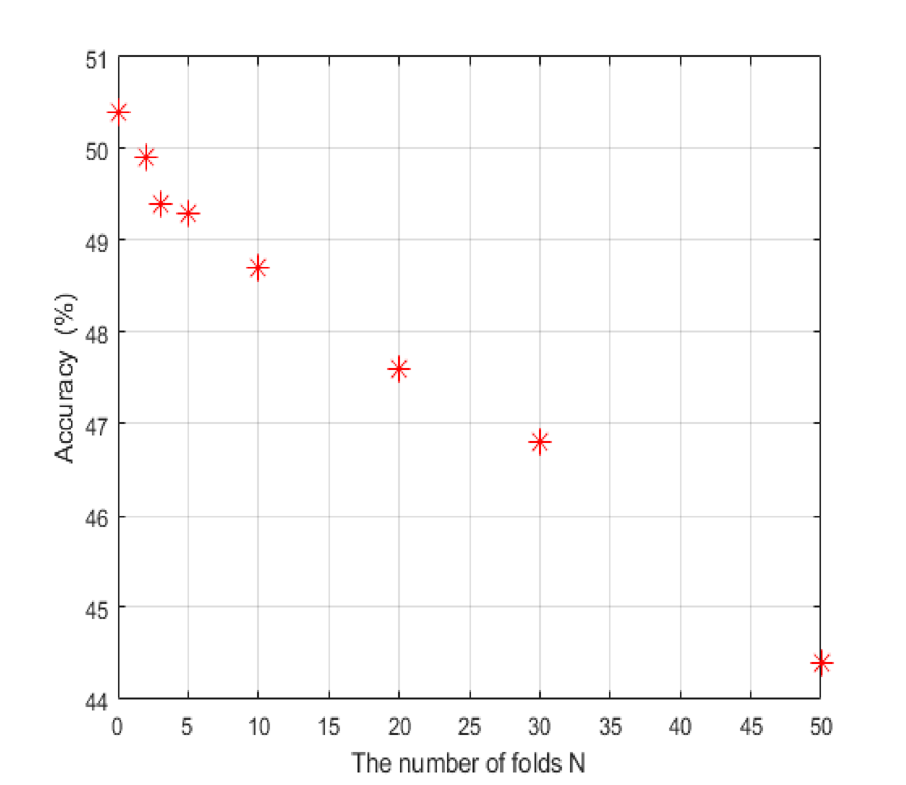
Continuous matrix of
decimals between 0.00
(not possible to have such
symptom)
and 1.00
(almost sure to have
such symptom)

Correlation matrix of symptoms

- To make the data more easy to process by converting discrete data to continuous ones;
- To consider and research the correlation between symptoms (which is common in the real world) to build better models and improve the accuracy.
- For decision trees with maximum layers of 20, the improvement is about 5%.

Result

- The parameter of N-fold in cross validation of the dataset may affect the accuracy of the model due to the error in the prediction of each fold .



- The trade off between accuracy and efficiency.
- Applied in training sets whose size is extremely large to save time.
- Not so useful for this project due to its small size and high demand in accuracy.

Q&A