MARMARA UNIVERSITY FACULTY OF ENGINEERING 2020-2021 SPRING SEMESTER

CSE4062

Introduction to Data Science and Analytics

Group 6

Predictive Analysis

Identification of Thyroid Cancer by Using Machine Learning

150116061 Ertuğrul Sağdıç, Computer Engineering - ertugrulsagdic98@gmail.com
150816004 Defne ÇIĞ, Bioengineering - defnecig@marun.edu.tr
150816019 Kasım Anlatır, Bioengineering - kasimanlatirr@gmail.com
150216031 Cem Telliağaoğlu, Environmental Engineering - cem.telliagaoglu@gmail.com
199520021 Lukas Eckerle, Industrial Engineering - lukas.eckerle@gmail.com

Revise Explore your Data Part 2

In the previous part, we have investigated the gene expressions to select the best attributes between exposed group and control group, and not-exposed group and control group. The purpose to do that was to find the differentially expressed genes in order to classify the instances as exposed or not-exposed. However, we could not find a way to accomplish that. Thus, we have changed our methodology with classifying the instances as diseased or healthy.

We have run the Random Forest algorithm by checking the entropy information of the gene expressions between diseased patients and healthy patients. We found 8 differentially expressed genes shown in Table 1.

	Gene ID	Gene Symbols	Cancer related findings
1	215255_at	IGSF9B	Favorable ovarian cancer biomarker
2	44822_s_at	MIER2	Possible biomarker in renal cancer, endometrial cancer and colorectal cancer
3	1564069_at	HOTTIP	Strong relation with the papillary thyroid carcinoma cells
4	1555083_at	RPL13AP17	No cancer related finding
5	205911_at	PTH1R	Enhanced cancer specificity in renal cancer
6	1561365_at	NRP1	Prognostic marker in stomach cancer, cervical cancer and renal cancer
7	244656_at	RASL10B	Unfavorable prognosis in endometrial cancer
8	243881_at	SHC3	Favorable hepatocellular carcinoma and unfavorable gastric cancer related findings

Table 1: *Differentially expressed genes.Cancer relation of each gene was derived using online databases (GeneCards, NCBI and the Human Protein Atlas)*

Predictive Analysis

We have made predictive analysis of our dataset using classification algorithms, including K-Nearest Neighbors (k-NN) and Neural Network by using the differentially expressed genes found in the previous step. We compared the performances of these classification algorithms trained with the training set which is 70% of our dataset and the classification capabilities on the test set by using the evaluation metrics [1]. These metrics are accuracy, sensitivity, specification, precision, and F1 score.

We made experiments with models constructed consisting of 4 of the Neural Network and 4 of the k-NN algorithms.

We have used SGD optimizer for the Neural Network algorithm. We constructed 4 models by changing the learning rate. For each model the learning rate changes as follows: 0.1, 0.01, 0.001, 0.0001.

We constructed 4 models for the k-NN algorithm by using different k values. For each model k value changes as follows: 3, 5, 9, 15.

The evaluation metrics are calculated by using the Confusion Matrix shown in Table 2.

Condition	Being Healthy	Having Disease
Being Healthy	TP	FN
Having Disease	FP	TN

Table 2: *Confusion Matrix.*

If we examine the models constructed using the Neural Network algorithm with different learning rates (LR), we can see that the models with learning rate 0.1 and 0.01 have the highest accuracy, sensitivity, specification, precision, and F1 score among the other Neural network models shown in Table 3.

Both models have:

- 1.0 accuracy level which means that 100% correct predictions
- 1.0 sensitivity (recall) which means that the model identifies the true positive rate with 100%
- 1.0 specification which means that the model predicts the true negative rate with 100%
- 1.0 precision shows that the model predicted the healthy people with 100% rightly in total predicted positive class
- F1 score is a harmonic means of sensitivity and precision shown in Equation 1. F1 score is calculated as 1.0 since the sensitivity and precision values are 1.0

$$\frac{2 \times Precision \times Sensitivity}{Precision + Sensitivity}$$

Equation 1: *F1 Score calculation.*

If we examine the models constructed using the k-NN algorithm with different k values, we can see that the model with k value 3 has the highest accuracy, sensitivity, specification, precision, and F1 score among the other k-NN models shown in Table 3.
The model has:

- 0.987 accuracy level which means that 98% correct predictions
- 0.976 sensitivity (recall) which means that the model identifies the true positive rate with 97%
- 1.0 specification which means that the model predicts the true negative rate with 100%
- 1.0 precision shows that the model predicted the healthy people with 100% rightly in total predicted positive class
- F1 score is a harmonic means of sensitivity and precision shown in Equation 1. F1 score is calculated as 0.987.

Algorithm	Experiment	Sensitivity	Specification	Precision	Precision	F1 Score
Neural Network	LR = 0.1	1.0	1.0	1.0	1.0	1.0
	LR = 0.01	1.0	1.0	1.0	1.0	1.0
	LR = 0.001	0.942	0.714	1.0	1.0	0.947
	LR = 0.0001	0.828	0.684	1.0	1.0	0.812
k-NN	k = 3	0.987	0.976	1.0	1.0	0.987
	k = 5	0.987	0.972	1.0	1.0	0.985
	k = 9	0.987	0.971	1.0	1.0	0.985
	k = 15	0.975	0.972	0.978	0.972	0.972

Table 3: *Evaluation Metrics for each algorithm with different parameters.*

In conclusion, selection of parameters affects the performance of the model in a considerable way.

References:

[1] Hossin, M., Sulaiman, M.N. (2015). A Review on Evaluation Metrics for Data Classification Evaluations. International Journal of Data Mining & Knowledge Management Process (IJDKP). 5(2): 1-11.