MARMARA UNIVERSITY FACULTY OF ENGINEERING 2020-2021 SPRING SEMESTER

CSE4062

Introduction to Data Science and Analytics

Group 6

Explore your data Part 2 (Revised)

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

Explore your data Part 2

The gene expressions of 54,675 genes have been investigated to select the best attributes between diseased patients and healthy patients . The genes were investigated using the Random Forest algorithm [1] with checking the entropy information of the genes [2] as a decision tree classifier [3]. This gives the genes that are differentially expressed.

First of all, the data set normalized between 0, 1. Then, the dataset was preprocessed. The labels of the instances were differentiated as diseased and healthy for the exposed group and the control group, and not-exposed group and control group. Then, the dataset was splitted in two sets for each investigation group. 70% of the data was splitted for the training set, 30% of the data was splitted for the test set.

The decision tree algorithm checks the expression levels of the genes and constructs trees by checking the entropy values. Then, the trees are added to random forest with the accuracy level higher than 90% of the accuracy level. Then, the samples were classified based on the trees in the random forest.

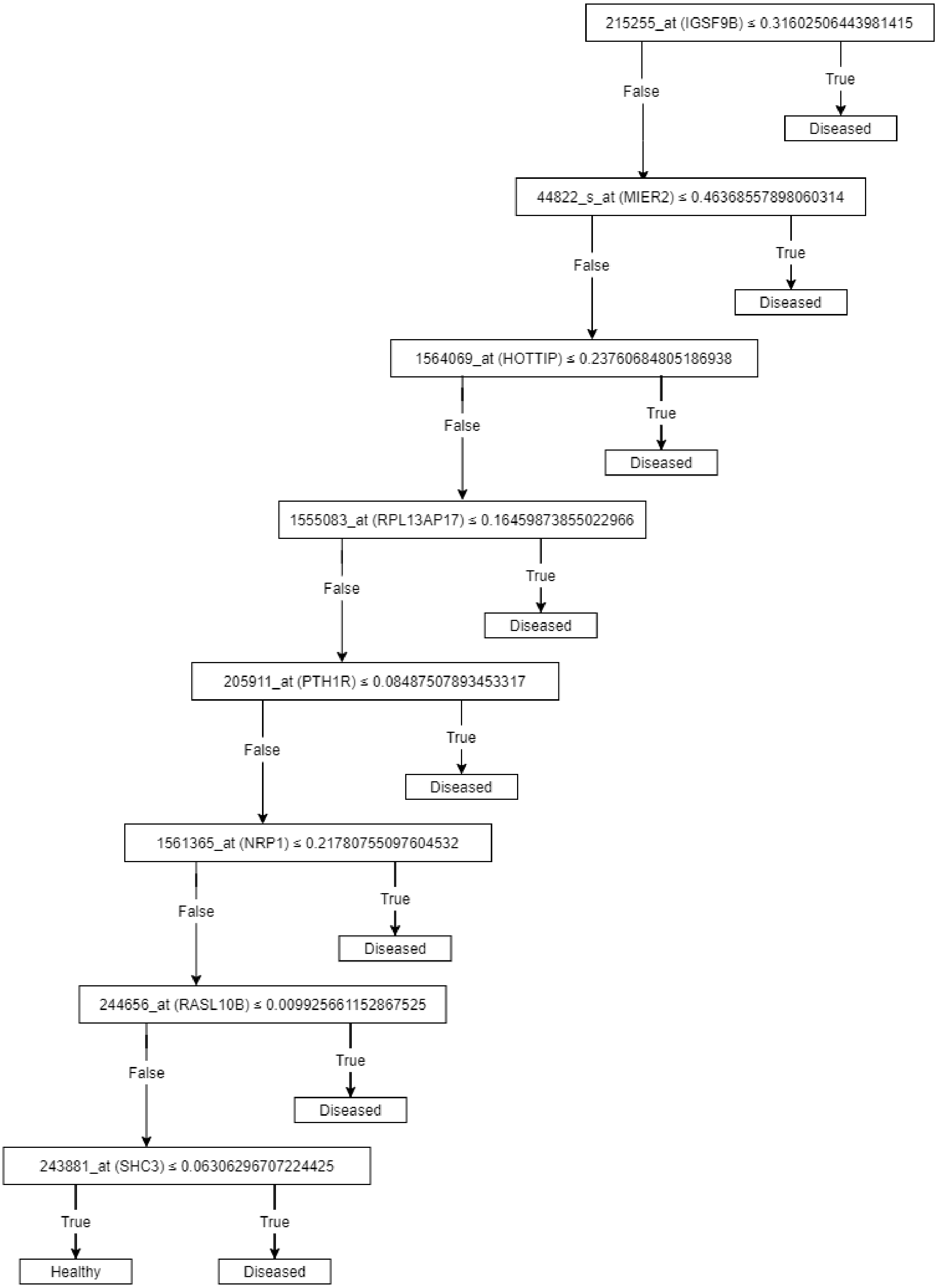
215255_at (IGSF9B) ≤ 0.31602506443981415
False
True
Diseased
44822_s_at (MIER2) ≤ 0.46368557898060314
False
True
Diseased
1564069_at (HOTTIP) ≤ 0.23760684805186938
False
True
Diseased
1555083_at (RPL13AP17) ≤ 0.16459873855022966
False
True
Diseased
205911_at (PTH1R) ≤ 0.08487507893453317
False
True
Diseased
1561365_at (NRP1) ≤ 0.21780755097604532
False
True
Diseased
244656_at (RASL10B) ≤ 0.009925661152867525
False
True
Diseased
243881_at (SHC3) ≤ 0.06306296707224425
True
True
Healthy
Diseased

Figure 1: *Decision tree of the best attributes selected for exposed/control group.*

	Gene ID	Gene Symbols
1	215255_at	IGSF9B
2	44822_s_at	MIER2
3	1564069_at	HOTTIP
4	1555083_at	RPL13AP17
5	205911_at	PTH1R
6	1561365_at	NRP1
7	244656_at	RASL10B
8	243881_at	SHC3

***Table 1:** Differentially expressed genes.*

Differentially expressed genes found by the algorithm mentioned above are searched for their relevance to cancer by reviewing the literature and using servers such as Human Protein Atlas and GeneCard. Some of the genes have been shown to play a role in different types of cancer. The gene IGSF9B is ,immunoglobulin superfamily member 9B, a kinase binding gene coding transmembrane protein which is abundantly expressed in interneurons. Studies have shown that the altered expression of IGSF9B is found in ovarian cancer. MIER2, is a member of the Mesoderm Induction Early Response gene family. MIER2 gene is a prognostic marker in renal cancer, endometrial cancer and colorectal cancer with a low cancer specificity. HOTTIP is an RNA gene and affiliated with the lncRNA class. The studies have shown that the HOTTIP gene is strongly related with the regulation of the proliferation and the apoptosis of papillary thyroid carcinoma cells (Yuan et. al.) alongside the tongue squamous cell carcinoma, gastric cancer and ovarian cancer. RPL13AP17 is a pseudogene (Ribosomal Protein L13a Pseudogene 17) showing no cancer related finding. Parathyroid hormone 1 receptor, PTH1R, is a protein coding gene. The protein encoded by the PTH1R gene is a receptor for parathyroid hormone (PTH) and for parathyroid hormone-like hormone (PTHLH). The gene shows enhanced cancer specificity in renal cancer. High expression of PTH1R is found to be related to Crohn's disease. NRP1 is a protein coding gene and the encoded protein is a transmembrane receptor and acts as a co-receptor for SARS-CoV-2 (which causes COVID-19) to infect host cells. The gene is a prognostic marker in stomach cancer, cervical cancer and renal cancer. RASL10B is a member of RAS like family 10. The gene shows unfavorable prognosis in endometrial cancer. SHC3 of SHC adaptor protein 3 is a protein coding gene showing gastric cancer relation among its pathways. It was shown that the SHC3 gene was increasingly expressed in malignant hepatocellular carcinoma (HCC) cell lines associated with HCC invasion and metastasis.

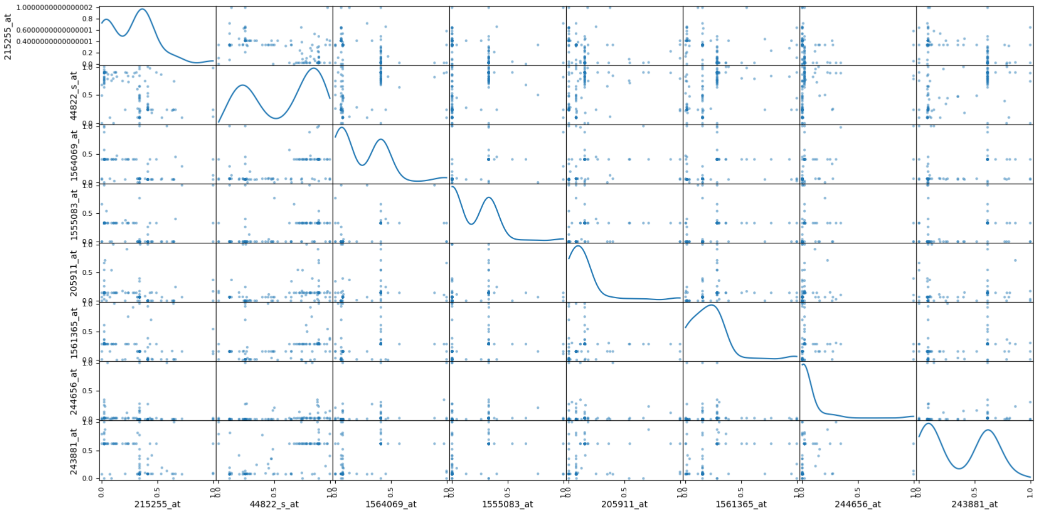

Figure 2: *Scatter matrix plotting of the differentially expressed genes.*

The relationships of the genes were plotted and shown in Figure 2. X-axis and Y-axis show the particular genes that have been found as the best attributes shown in Figure 3 and 4. Situations where the graph forms a positively inclined diagonal are positive correlations and situations where it forms a negatively inclined diagonal are negative correlations. However, in our case, the presence of correlation was not observed.

References:

[1] Hsueh, H. M., Zhou, D. W., & Tsai, C. A. (2013). Random forests-based differential analysis of gene sets for gene expression data. Gene, 518(1), 179-186.

[2] Butte, A. J., & Kohane, I. S. (1999). Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. In Biocomputing 2000 (pp. 418-429).

[3] Polaka, I., Tom, I., & Borisov, A. (2010). Decision tree classifiers in bioinformatics. Applied Computer Systems, 42(1), 118-123.

[4] Yuan Q, Fan Y, Liu Z, Wang X, Jia M, Geng Z, Zheng J, Lu X. miR-744-5p mediates lncRNA HOTTIP to regulate the proliferation and apoptosis of papillary thyroid carcinoma cells. Exp Cell Res. 2020 Jul 1;392(1):112024.