Prediction of Thyroid Disease Using Data Mining Techniques

Irina Ioniţă
Informatics, Computer Science, Mathematics and Physics,
Petroleum-Gas University of Ploieşti, Ploieşti, Romania
tirinelle@yahoo.com

Liviu Ioniţă
Informatics, Computer Science, Mathematics and Physics,
Petroleum-Gas University of Ploieşti, Ploieşti, Romania
lionita@gmail.com

Abstract

Recently, thyroid diseases are more and more spread worldwide. In Romania, for example, one of eight women suffers from hypothyroidism, hyperthyroidism or thyroid cancer. Various research studies estimate that about 30% of Romanians are diagnosed with endemic goiter. Factors that affect the thyroid function are: stress, infection, trauma, toxins, low-calorie diet, certain medication etc. It is very important to prevent such diseases rather than cure them, because the majority of treatments consist in long term medication or in chirurgical intervention. The current study refers to thyroid disease classification in two of the most common thyroid dysfunctions (hyperthyroidism and hypothyroidism) among the population. The authors analyzed and compared four classification models: Naive Bayes, Decision Tree, Multilayer Perceptron and Radial Basis Function Network. The results indicate a significant accuracy for all the classification models mentioned above, the best classification rate being that of the Decision Tree model. The data set used to build and to validate the classifier was provided by UCI machine learning repository and by a website with Romanian data. The framework for building and testing the classification models was KNIME Analytics Platform and Weka, two data mining software.

Keywords: data mining, classification model, thyroid diseases, neural network, decision tree, Naïve Bayes

1. Introduction

Prevention in health care is a continuous concern for the doctors and the correct diagnostic at the right time for a patient is crucial, due to the implied risk. Recently, the usual medical report can be accompanied by an additional report given by a decision support system or other advanced diagnosis techniques based on symptoms. Questions such as: “what are the most important factors that affect thyroid?”, “which is the category of the population predisposed to goiter disease?”, “what is the most adequate treatment for a certain disease?” etc. may find answers in applying data mining techniques. Health care data can be processed and after rigorous usage can provide knowledge used in decision making, diagnosing diseases more rapidly and accurately, offering better medication for patients and minimizing the death risk. The authors focus their work on using classification methods and identifying the best algorithm for classification thyroid disorders.

Thyroid is a butterfly-shaped gland, which is located at the bottom of the throat responsible for producing two active thyroid hormones, levothyroxine (T4) and triiodothyronine (T3) that affect some functions of the body such as: stabilizing body temperature, blood pressure, regulating the heart rate etc. Reverse T3 (RT3) is manufactured from thyroxine (T4), and its role is to block the action of T3. An abnormal function of the thyroid implies the occurrence of hyperthyroidism and hypothyroidism, two of the common thyroid affections. Hypothyroidism (underactive thyroid or low thyroid) means that the thyroid gland doesn't produce enough of certain important hormones. Without an adequate treatment, hypothyroidism can cause various health problems such as: obesity, joint pain, infertility and heart disease. Hyperthyroidism (overactive thyroid) refers to a condition in which the thyroid gland produces too much of the hormone thyroxin. In this case, the body's metabolism is accelerating significantly, causing sudden weight loss, a rapid or irregular heartbeat, sweating, and nervousness or irritability (eMedonline, 2016). In Figure 1 are presented the main factors that affect the thyroid function. It is obvious that factors such as stress, infection, toxins, trauma and certain medication are directly responsible for the improper production of thyroid hormones. Symptoms identification and the early detection of abnormal values of thyroid hormones after clinical investigation will help in establishing the proper diagnostic and to prescribe the right medication. The patient must periodically evaluate his clinical state in order to receive the treatment as long as he needs it.

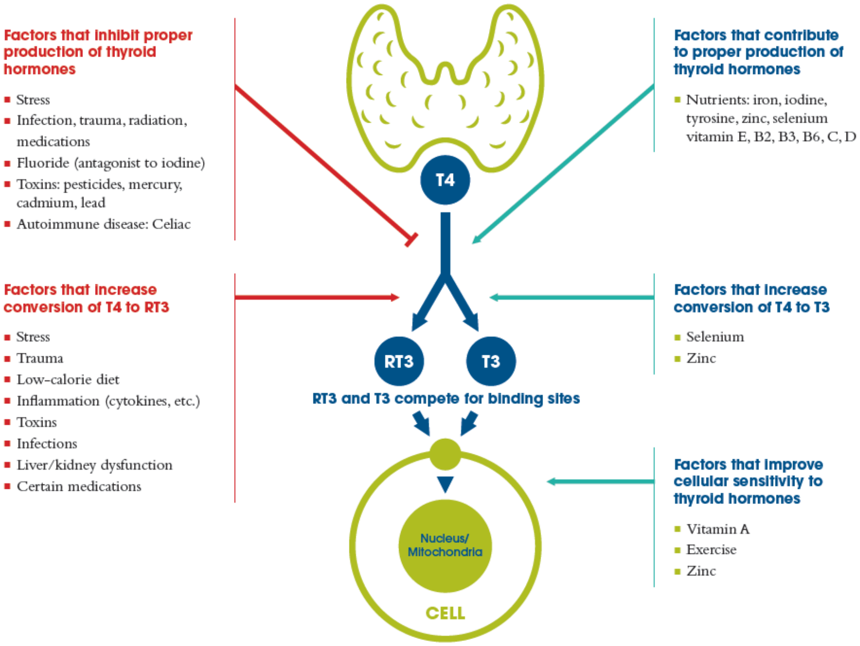

Factors that inhibit proper production of thyroid hormones
▪ Stress
▪ Infection, trauma, radiation, medications
▪ Fluoride (antagonist to iodine)
▪ Toxins: pesticides, mercury, cadmium, lead
▪ Autoimmune disease: Celiac
Factors that contribute to proper production of thyroid hormones
▪ Nutrients: iron, iodine, tyrosine, zinc, selenium vitamin E, B2, B3, B6, C, D
T4
Factors that increase conversion of T4 to RT3
▪ Stress
▪ Trauma
▪ Low-calorie diet
▪ Inflammation (cytokines, etc.)
▪ Toxins
▪ Infections
▪ Liver/kidney dysfunction
▪ Certain medications
Factors that increase conversion of T4 to T3
▪ Selenium
▪ Zinc
RT3
T3
RT3 and T3 compete for binding sites
Factors that improve cellular sensitivity to thyroid hormones
▪ Vitamin A
▪ Exercise
▪ Zinc
Nucleus/ Mitochondria
CELL

Figure 1. Factors that Affect Thyroid Function (The Institute for Functional Medicine, 2014)

The paper is organized in three main sections. After a short introduction, in section 2 are mentioned the solutions given by data mining techniques in health care, highlighting the importance of discovering patterns, relations between various data (initial appearing to be insignificant) and to make predictions based on volumes of data. Section 3 consists in the experiments proposed by the authors, regarding the classification of thyroid diseases and the results obtained after running the data mining algorithms. The discussions on the applied classification methods and the interpretation of evaluation measurements are presented at the end of this section. The paper ends with the authors' conclusions regarding the research work and their future preoccupation.

2. Data mining in health care

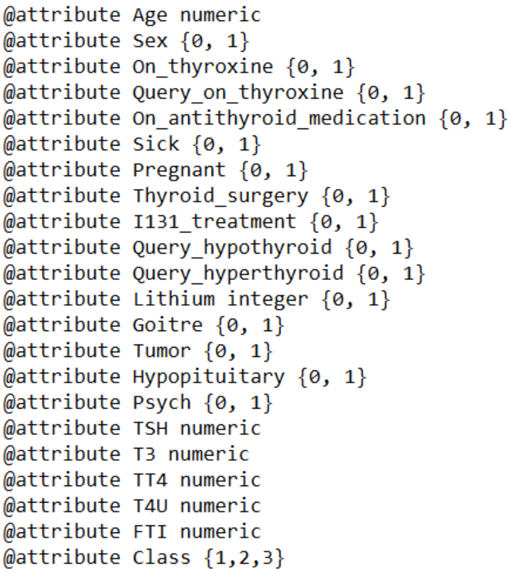
Data mining refers to extracting unknown patterns from an enormous volume of data involving different methods and algorithms which exist at the intersection of fields such as artificial intelligence, machine learning, statistics and database systems (Piatetsky-Shapiro & Parker, 2011). Hospitals, clinics and medical analysis laboratories accumulate a large amount of patient data over the years. These data provide a basis for the analysis of risk factors for many diseases (various types of cancer, heart diseases, diabetes, hepatitis etc.). In literature are mentioned certain applications of data mining techniques in the health domain, some of them being presented in the following paragraphs. The authors have narrowed their research area on thyroid disorders and the examples given below are strictly about the related work described in literature, regarding the application of data mining for these classes of diseases. The majority of examples refer to diagnosing diseases of thyroid using decision trees, artificial neural networks, support vector machine, expert systems etc. For example, the diagnosis of thyroid disorders using ANN's is discussed in Gharehchopogh, Molany & Mokri (2013), Anupama Shukla & Prabhdeep Kaur (2015), Prerana, Parveen Sehgal & Khushboo Taneja (2015). The authors of the research work presented in Margret, Lakshmipathi & Kumar (2012) proposed the diagnosis of thyroid disease using decision tree splitting rules. The classification of the thyroid nodules was performed with support vector machines in Chuan-Yu Chang, Ming-Fang Tsai & Shao-JerChen (2008), while a comparison study by data mining classification algorithms (C 4.5, C5.0) for thyroid cancer set is presented in Upadhayay, Shukla & Kumar (2013). A diagnosis expert system based on fuzzy rules is described in Keleş & Keleş, (2008), while a three-stage expert system based on support vector machines is presented in Hui-Ling Chen et al. (2012). All the mentioned studies have the same goal, namely the diagnosis of thyroid disorders, classifying the stored data in medical databases and predicting the occurrence of a certain disease on population. Data mining can help to design the patient profile prone to develop a thyroid dysfunction and to identify the risk factors for these categories of diseases.

The next section of the paper consists in two subsections. On the first one we present the experiments developed on a database provided by UCI Machine Learning Repository (UCI, 2016), containing data about clinical history of patients with thyroid disorders. On the second subsection we used data about Romanian patients collected from a web site (tiroida.ro, 2016). The authors considered for classification of data four data mining algorithms: Naive Bayes, Decision Trees, Multilayer Perceptron and RBF Network to build a robust classifier. The goal of this study is to find the best classification model in order to make future classification of new patient data more accurately. In this paper more experiments are discussed and an interpretation of results (evaluation measurements) is given.

3. Experiments and results

Experiment A.

The authors used for their experiments a data set (UCI, 2016) containing 756 records about persons with thyroid dysfunctions. The classification model has 22 attributes; the *class attribute* is the target and it has three possible values: *hypothyroidism, hyperthyroidism* and *normal.* The current data set was extracted and preprocessed from the original file. A description of the attributes used in the experiments is given in Figure 2 (an extract from *thyroid.arff* test file).



```
@attribute Age numeric
@attribute Sex {0, 1}
@attribute On_thyroxine {0, 1}
@attribute Query_on_thyroxine {0, 1}
@attribute On_antithyroid_medication {0, 1}
@attribute Sick {0, 1}
@attribute Pregnant {0, 1}
@attribute Thyroid_surgery {0, 1}
@attribute I131_treatment {0, 1}
@attribute Query_hypothyroid {0, 1}
@attribute Query_hyperthyroid {0, 1}
@attribute Lithium integer {0, 1}
@attribute Goitre {0, 1}
@attribute Tumor {0, 1}
@attribute Hypopituitary {0, 1}
@attribute Psych {0, 1}
@attribute TSH numeric
@attribute T3 numeric
@attribute TT4 numeric
@attribute T4U numeric
@attribute FTI numeric
@attribute Class {1,2,3}
```

Figure 2. Attributes of the classification models used in the experiments

6 attributes have numeric values (*Age, TSH, T3, TT4, T4U, FTI),* 15 attributes are categorical with two possible values (0/1). The *class* attribute is categorical too, but it can take three possible values (1 – *hypothyroidism*, 2 – *hyperthyroidism* or 3 – *normal*). The software used to build and test the data mining models is KNIME Analytics Platform 2.12.1 (KNIME, 2015). KNIME (Konstanz

Information Miner) is a user-friendly and comprehensive open-source data integration, processing, analysis, and exploration platform that enables the user to visually create data flows (diagrams), selectively execute some or all the analysis steps, and later investigate and understand the results through interactive views on data and models (KNIME, 2015). According to a survey posted by Gregory Piatetsky (Piatetsky, 2016), KNIME is on the list with the top 10 most popular tools in 2016 as analytic data mining software. Also, the authors worked with other free data mining tools such as Weka (Weka, 2016), but they focused on KNIME because this is a tool blending for Python, R, SQL, Java, Weka, and many more.

The proposed KNIME diagram representing the data mining models is given in Figure 3. The nodes that constitute the model diagram are: *ARFF Reader* – the input node used to load the data set in arff format, *Partitioning* – the node with the role of data set partition (for training and for the validation of the classification model), *Naive Bayes Learner* and *Decision Tree Learner* – the nodes used to build the classification model, *Naive Bayes Predictor* and *Decision Tree Predictor* – the nodes used to validate the model, *Scorer* – the node reports a confusion matrix and the accompanying quality measures in its view, *Normalizer* – the data set are normalized to be able to apply the neural network models, *Multilayer Perceptron* and *RBFNetwork* – the nodes corresponding to the neural network classification models, *Weka Predictor* – a node implemented in Weka to validate the models.

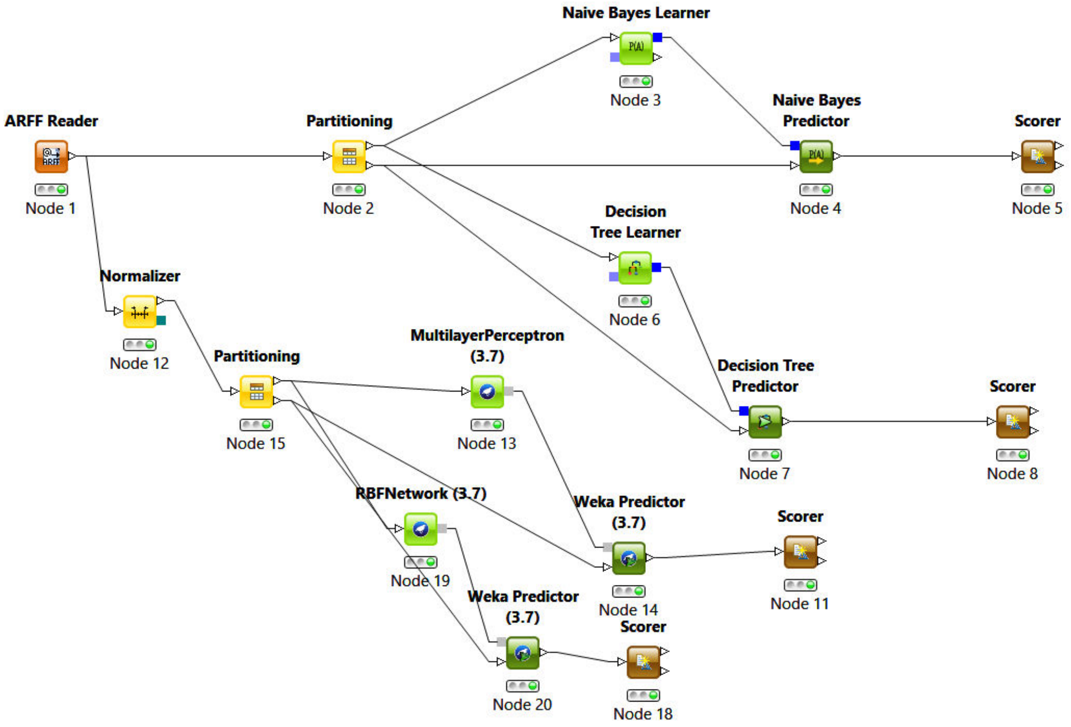


Figure 3. KNIME Diagram

The authors considered for classification of data four data mining algorithms: Naive Bayes, Decision Trees, Multilayer Perceptron and RBF Network.

Naive Bayes methods represent a set of supervised learning algorithms based on applying Bayes' theorem with the "naive" assumption of independence between every pair of features. Given a class variable *y* and a dependent feature vector $(x_1, \dots, x_n)$ Bayes' theorem states the following relationship (Naive, 2016, Md. Faisal Kabir, Chowdhury Mofizur Rahman, Almgir Hossain & Keshav Dahal, 2011):

$$P(y|x_1,\dots,x_n) = \frac{P(y)P(x_1,\dots,x_n|y)}{P(x_1,\dots,x_n)} \quad (1)$$

Using the relation:

$$P(x_i|y,x_1,\dots,x_{i-1},x_{i+1},\dots,x_n) = P(x_i|y) \quad (2)$$

for all i, the relationship may be simplified to:

$$P(y|x_1,\dots,x_n) = \frac{P(y)\prod_{i=1}^{n}P(x_i|y)}{P(x_1,\dots,x_n)} \quad (3)$$

$P(x_1,\dots,x_n)$ is constant given the input. Considering this, the following classification rule can be used:

$$P(y|x_1,\dots,x_n) \propto P(y)\prod_{i=1}^{n}P(x_i|y) \quad (4)$$

$$\Downarrow$$

$$\hat{y} = \arg\max_{y} P(y)\prod_{i=1}^{n}P(x_i|y) \quad (5)$$

The advantages of this classifier are: fast to train (single scan), fast to classify, not sensitive to irrelevant features, handles real and discrete data, handles well streaming data. As a disadvantage we are mentioning that this model assumes independence of features.

Decision Trees represent a classification method used in various fields like education, medicine, financial analysis, industry etc. Algorithms such as ID3, CART, C4.5 are well-known in the data mining field and are characterized as simple and easy to understand. A decision tree induction generates a flow chart like a tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node consists in a class label (Upadhayay, Shukla, Kumar, 2013). In KNIME, the algorithm associated to the Decision Tree learner node provides two quality measures for split calculation (the gini index and the gain ratio). CART uses Gini Index as an attribute selection measure to build a decision tree. C 4.5 builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy. Further, there exists a post pruning method to reduce the tree size and increase prediction accuracy, based on the minimum description length principle (MDL). For the current classification problem the authors have to define a function:

$thyroid\_disease: Data\_set \rightarrow Classes$, where *Data_set* represent the records in a database of patients with tuples $(x_1,\dots,x_n)$ – the values of attributes $(A_1,\dots,A_n)$ and $Classes = \{C1, C2, C3\}$ – where C_1 *is hyperthyroidism*, C_2 *is hyporthyroidism and* C_3 *is normal.*

Gini Index is calculated as:

$$Gini\ (Data_{set}) - 1 - \sum_{i=1}^{n}(p_i)^2 \quad (6)$$

where p_i is the probability that a tuple in *Data_set* belongs to class $C_{i.}$

Gain ratio can be defined as a modification of the information gain, it uses normalized information gain to reduce its bias on large attributes (Kaur, & Wasan, 2006).

The mathematic formula is:

$$Gainration\ (Attribute) = \frac{Gain\,(Attribute)}{SplitInfo\,(Attribute)} \quad (7)$$

$$SplitInfo\,(Attribute) = -\sum_{i=1}^{q}\left(\frac{|Data\_set_i|}{|Data\_set|}\right) \times \log_2\left(\frac{|Data\_set_i|}{Data\_set}\right) \quad (8)$$

Multilayer Perceptron is a feedforward neural network with one or more number of hidden layers between input and output layer. The MLP algorithm consists in several steps such as (Noriega, 2016):

a. The network is initialized with all weights set to numbers generated randomly between -1 and +1.
b. The first training pattern is taken and the activation function applies to obtain the network output.
c. A comparison between the network output and the target output is effectuated.
d. The back-propagation method is applying to calculate the error backwards (during this step both the output layer of weights and the input weights must be corrected).
e. The global error is calculated based on the average difference between the target and the output vector.
f. The first five steps repeat for each pattern in the training data set in order to complete one epoch.
g. The training data set is generated randomly for the next epoch in order to avoid the network being influenced by the order of the data.
h. The stop criterion is the established number of epochs or the moment when the error remains at the same value.

For the study case presented in the current paper, the network inputs are considered all the 21 attributes mentioned above, and the output is the *class* attribute.

Radial Basis Function Network (RBF Network) is known as a supervised feed forward neural network with one hidden layer. It uses relatively smaller number of locally tuned units and its behavior is adaptive. The activation functions for these artificial neural networks are radial basis functions. The output of a RBF network is a linear combination of radial basis functions of the inputs and neuron parameters (Venkatesan & Anitha, 2006).

After the normalization of data set, using the KNIME node namely *Normalizer,* the full data set was split into two partitions: the training set for model construction (66%) and the hold-out set for model evaluation (33%).

The results of the experiment are presented in the table 1.

Table 1. Evaluation measurements for classification models

		Accuracy Statistics	Classification Models			
			Naïve Bayes	*Decision Tree*	*MLP*	*RBF Network*
Classes	*Hyperthyroidism (1)*	Recall	0.444	0.778	0.25	0.5
		Precision	0.4	1	0.5	0.667
		Sensitivity	0.444	0.778	0.25	0.5
		Specifity	0.976	1	0.992	0.992
		F-measure	0.421	0.875	0.333	0.571
	Hypothyroidism (2)	Recall	0.5	0.9	0.333	0.417
		Precision	0.455	0.6	0.571	0.625
		Sensitivity	0.5	0.9	0.333	0.417
		Specifity	0.976	0.976	0.988	0.998
		F-measure	0.476	0.72	0.421	0.5
	Normal (3)	Recall	0.971	0.975	1	0.987
		Precision	0.979	0.978	0.964	0.963
		Sensitivity	0.971	0.975	1	0.987
		Specifity	0.737	0.842	0.55	0.55
		F-measure	0.975	0.981	0.981	0.975
		Accuracy	0.934 (93.4%)	**0.965 (96.5%)**	0.946 (94.6%)	0.946 (94.6%)

The accuracy of the classification models is over 90%, the best model being Decision Tree with 96.5% accuracy. If the number of epochs for MLP model increases to 1500 (initial was set to 500) the accuracy of the classification model increases to 95.34%. The authors initially considered the Decision Tree model Gini Index as quality measure and not pruning method. For the same data set, but choosing MDL as a post pruning method, the accuracy of the decision tree model increases to 96.9%. Other experiment consists in analyzing the modification occurred after changing the partition percent to 70%. The obtained results are presented in table 2.

Table 2. Accuracy for classification models

	Classification Models			
	Naïve Bayes	Decision Tree	MLP	RBF Network
Accuracy	91.63%	**96.91%**	95.15%	96.03%

The authors analyzed the impact of certain attributes over the classification model accuracy. In the next experiment the following attributes were ignored: *Query_on_thyroxine, Query_on_hypothyroid, Query_on_hyperthyroid.* Table 3 presents the accuracy in this case. The classification model based on decision tree obtained the best accuracy (97.35%), while Naïve Bayes obtained the weakest classification.

Table 3. Accuracy of classification model after removing three of the model attributes

	Classification Models			
	Naïve Bayes	Decision Tree	MLP	RBF Network
Accuracy	89.96%	**97.35%**	94.71%	94.27%

As we observed, Decision Tree model was the model with the highest accuracy for the classification of thyroid diseases, followed by MLP and RBF Network models. The results indicate that, *Normal* outcome had been recognized by all the considered models accurately, whereas is not the same for *Hypothyroidism* and *Hyperthyroidism* classes. A possible cause may be the increase number of observations regarding the thyroid diseases (Hypothyroidism, Hyperthyroidism).

Experiment B.

On the first experiment, the authors did not use data of Romanian patients because there is no availability for records of such database due the confidentiality of the patients. However, the authors collected data from a site (tiroida.ro, 2016) on persons with thyroid disorders, but this data refers only to a small range of parameters (TSH, FT4, ATPO - TPO antibodies, T3 etc.). The normal range for each parameter varies from one laboratory to another and is difficult to keep only the adequate values.

After a careful analysis and data processing, a data set with 140 records resulted and consists in row material for the second experiment, discussed in the current paper. The attributes used to apply data mining algorithms are: TSH, FT4, ATPO, AGE, SEX and Class. The target is class with three possible values: 1- hyperthyroidism, 2- hypothyroidism and 3 – normal.

On the Experiment B, Weka was the analytic platform used to build the data mining models Simple CART, J48, Multilayer Perceptron, RBF Network and Naïve Bayes. As test option, the authors chose 10-folds cross validation (table 4) and percentage split=66% (table 5). The results are depicted below.

As we observe, Simple CART, MLP and J48 obtained the best accuracy (over 80%), while RBF Network only 64.28%.

Table 4. Evaluation measurements for classification models applied on Romanian data set
Test option: 10-folds cross-validation

			Classification Models Test option: 10-folds cross-validation				
		Accuracy Statistics	*Simple CART*	*J48*	*MLP*	*RBF Network*	*Naïve Bayes*
Classes	*Hyperthyroidism (1)*	Recall	0.528	0.556	0.528	0.167	0.25
		Precision	0.826	0.741	0.864	0.136	0.474
		ROC Area	0.661	0.761	0.77	0.62	0.687
		F-measure	0.644	0.635	0.655	0.128	0.327
	Hypothyroidism (2)	Recall	0.808	0.731	0.808	0.692	0.692
		Precision	0.808	0.826	0.778	0.643	0.692
		ROC Area	0.848	0.828	0.896	0.908	0.878
		F-measure	0.808	0.776	0.792	0.667	0.692
	Normal (3)	Recall	0.949	0.936	0.949	0.846	0.885
		Precision	0.813	0.811	0.813	0.71	0.726
		ROC Area	0.811	0.808	0.836	0.806	0.85
		F-measure	0.876	0.869	0.876	0.772	0.798
		Accuracy	**81.42%**	80.00%	**81.42%**	*64.28%*	*68.57%*

Table 5. Evaluation measurements for classification models applied on Romanian data set.
Test option: Percentage split = 66%

			Classification Models Test option: Percentage split = 66%				
		Accuracy Statistics	*Simple CART*	*J48*	*MLP*	*RBF Network*	*Naïve Bayes*
Classes	*Hyperthyroidism (1)*	Recall	0.7	0.7	0.3	0.4	0.2
		Precision	1	0.875	0.5	0.5	0.286
		ROC Area	0.824	0.812	0.632	0.655	0.7
		F-measure	0.824	0.778	0.375	0.444	0.235
	Hypothyroidism (2)	Recall	0.75	0.75	0.875	0.875	0.625
		Precision	0.857	0.857	0.875	0.875	0.833
		ROC Area	0.884	0.823	0.975	0.944	0.972
		F-measure	0.8	0.8	0.875	0.875	0.714
	Normal (3)	Recall	1	1	0.9	0.9	0.9
		Precision	0.882	0.909	0.794	0.844	0.771
		ROC Area	0.889	0.917	0.794	0.82	0.907
		F-measure	0.938	0.952	0.844	0.871	0.831
		Accuracy	**89.58%**	**89.68%**	77.08%	*79.16%*	*70.83%*

By setting the percentage split to the value 66%, we improved the classification accuracy (over 89%), namely: the best classification rate was associated to J48 and Simple CART models, followed by RBF Network model.

Compared with the experimental results on the first discussion (Experiment A), on the second experiment the accuracy of classification models is lower and it's influenced by a lack of information contained by the Romanian database and the quality of the data. Also, the classification models from Experiment A use more predictors than data mining models from the experiment B.

The future work of the authors will focus on obtaining answers to questions like: Who are the persons most affected by thyroid diseases? What segment of age concentrates the thyroid cancer? How can we prevent thyroid diseases on the youth? By applying data mining algorithms.

4. Conclusions

As the medical reports show serious thyroid dysfunctions among the population, more affected being women, thyroid classification is a very important subject for researchers in medical science. In literature are mentioned various research works in the field of thyroid classification based on different data mining techniques used to build robust classifier. In this paper the authors discussed about applying four classification models (Naïve Bayes, Decision Tree, MLP and RBF Network) on thyroid data set to identify more accurately the dysfunction of thyroid namely hyperthyroidism and hypothyroidism. The best classification model was the decision tree model in all the effectuated experiments. The future work will focus on the identification of factors that affect the thyroid diseases and on testing more data mining techniques for the classification of different diseases (diabetes, heart diseases etc.).

References

Piatetsky-Shapiro, G. & Parker, G. (2011). *Lesson: Data Mining, and Knowledge Discovery: An Introduction,* Introduction to Data Mining. KD Nuggets.

Gharehchopogh, F.S, Molany, M., & Mokri, F.D. (2013). *Using artificial neural network in diagnosis of thyroid disease: a case study*", International Journal on Computational Sciences & Applications (IJCSA) Vol.3, No.4.

Shukla, A. & Kaur, P. (2009). *Diagnosis of thyroid disorders using artificial neural networks*, IEEE International Advance computing Conference (IACC 2009)– Patiala ,India, pp 1016-1020.

Prerana, Sehgal, P., & Taneja, K. (2015). *Predictive Data Mining for Diagnosis of Thyroid Disease using Neural Network*, International Journal of Research in Management, Science & Technology (E-ISSN: 2321-3264), Vol. 3, No. 2.

Margret, J., Lakshmipathi, B., & Kumar, S.A. (2012). *Diagnosis of Thyroid Disorders using Decision Tree Splitting Rules*, International Journal of Computer Applications (0975 – 8887) Volume 44– No. 8.

Chang, C.Y., Tsai, M.F., & Shao-JerChen (2008). *Classification of the Thyroid Nodules Using Support Vector Machines*, International joint conference on Neural, Networks, pp 3093- 3098.

Upadhayay, A., Shukla, S., & Kumar, S. (2013), *Empirical Comparison by data mining Classification algorithms (C 4.5 & C 5.0) for thyroid cancer data set*, International Journal of Computer Science & Communication Networks,Vol 3(1), 64-68, ISSN:2249-5789.

Keleş, A. & Keleş, A. (2008). *ESTDD: Expert system for thyroid diseases diagnosis.* Expert Systems with Applications 34.1, pp. 242-246.

Hui-Ling Chen et al. (2012). *A Three-Stage Expert System Based on Support Vector Machines for Thyroid Disease Diagnosis*, Journal of Medical Systems, Volume 36 Issue 3, June 2012, pp. 1953-1963.

UCI machine learning repository. (2016), https://archive.ics.uci.edu/ml/machine-learning-databases/thyroid-disease/ , accessed in 5 May 2016.

KNIME software. (2015) https://www.knime.org/, accessed in September 2015.

Naïve. (2016). http://www.cs.ucr.edu/~eamonn/CE/Bayesian%20Classification%20withInsect_examples.pdf, accessed in 20 April 2016

Md. Kabir, F., Rahman, C.M., Hossain, A., & Dahal, K. (2011). *Enhancement Classification Accuracy of Naïve Bayes Data Mining Models*, International Journal of Computer Application (IJCA),vol.28(3).

Venkatesan, P. & Anitha, S. (2006). *Application of RBF Neural Network for diagnosis of Diabetes Mellitus*, Current Science, Vol. 91, No. 9, pp. 1195-1199.

Kaur, H. & Wasan, S. K. (2006). *Empirical Study on Applications of Data Mining Techniques in Healthcare*, Journal of Computer Science 2(2), 194200.

Piatetsky, G. (2016), *R, Python Duel As Top Analytics, Data Science software – KDnuggets 2016 Software Poll Results*, KDnuggets, http://www.kdnuggets.com/2016/06/r-python-top-analytics-data-mining-data-science-software.html

Noriega, L. (2016). *Multilayer Perceptron Tutorial*, http://www.cs.sun.ac.za/~kroon/courses/machine_learning/lecture5/mlp.pdf, accesed in 10 April 2016.

eMedonline. (2016). http://www.emedonline.ro/afectiuni/view.article.php/c1/59/p3, accesed in 10 April 2016.

The Institute for Functional Medicine. (2015), *Factors that Affect Thyroid Function,* http://tulawellnessmd.com/wp-content/uploads/2015/09/Understanding-Factors-that-Affect-the-Thyroid-Diagram.pdf, accesed in May 2016.

http://tiroida.ro/, accessed on August 2016

Weka 3: Data Mining Software in Java, http://www.cs.waikato.ac.nz/ml/weka/, accessed in June 2016