Comparative study on classification of Thyroid diseases

Sara Iris Garcia
Faculty of Engineering and Computing
Coventry University
Coventry, UK
garciaas@coventry.ac.uk

***Abstract*—The thyroid disease is a condition that affects millions of people around the globe. An incorrect diagnosis can lead to death. In this paper we applied feature selection techniques and three methods for balancing the dataset**

***Keywords*—Thyroid, SCUT, SVM, Random Forest**

I. Introduction

The thyroid is an endocrine gland located in the neck. It is responsible for the regulation of vital body functions like breathing, heart rate, body temperature, and body weight, and plays a major role in the metabolism and the child's development.

The thyroid produces T4 (Thyroxine) and T3 (Triiodothyronine) hormones which travel in the bloodstream to the cells and bound to proteins. If the level of T4 hormone in the bloodstream is too low or too high, the hypothalamus releases Thyrotropin Releasing Hormone which triggers the pituitary to release more or less Thyroid Stimulating Hormone which in turn, stimulates the thyroid to increase or decrease the production of T4. An underactive thyroid will produce low amounts of T4 and T3 hormones, leading to Hypothyroidism. An overactive thyroid will produce high amounts of T4 and T3 hormones, which leads to Hyperthyroidism.

Thyroid disease is a global health problem that affects nearly one-third of the entire world population and is present in more females than males in a ratio of 5:1. Hypothyroidism is the most frequent thyroid disease in the world.

In this paper/study, three approaches are explored to overcome the class imbalance of the dataset: Random Oversampling, SMOTE (Synthetic Minority Oversampling Technique), and SCUT (SMOTE and Clustered Undersampling Technique). A comparative classification of the thyroid disease is performed using Decision Trees, KNN and Support Vector Machine using Step Forward Selection (SFS) and Principal Component Analysis (PCA) for feature selection.

II. Related Work

Various authors have conducted relevant studies in the classification of thyroid diseases. Razia et al[3] present the performance obtained by Support Vector Machine, Naïve Bayes, Decision Trees and Multiple Linear Regression classifiers were they achieve an accuracy of 99.23% with Decision Trees classifier.

Kousarrizi et al[5] apply Sequential Forward Selection, Sequential Backward Selection and Genetic Algorithm for feature selection and SVM as classifier into two thyroid disease datasets. The higher accuracy obtained from their experiments was 98.62% using 3 feature subset.

G. Rashita Banu[3] applies Linear Discriminant Analysis (LDA) technique to classify hypothyroidism related diseases, reporting 99.62% of accuracy using k=6 folds cross validation.

S. Pandey et al[6] present an ensemble of C4.5 algorithm to generate a decision tree and Random Forest with an accuracy of 99.47%

III. Thyroid Dataset

The dataset was obtained from the UCI machine learning repository of the Garavan Institute in Sydney, Australia. It contains 7200 instances, 21 features and 3 classes (Normal, Hypothyroidism, and Hypothyroidism).

TABLE I. Dataset Description

Attribute	***Data type***	***Value range***
Age	Real	[0.00, 0.9]
Sex	Integer	[0, 1]
On thyroxine	Integer	[0, 1]
Query on thyroxine	Integer	[0, 1]
Antithyroid medication	Integer	[0, 1]
Sick	Integer	[0, 1]
Pregnant	Integer	[0, 1]
Thyroid surgery	Integer	[0, 1]
I131 treatment	Integer	[0, 1]
Query hypothyroid	Integer	[0, 1]
Query hyperthyroid	Integer	[0, 1]
Lithium	Integer	[0, 1]
Goitre	Integer	[0, 1]
Tumor	Integer	[0, 1]
Hypopituitary	Integer	[0, 1]
Psych	Integer	[0, 1]
TSH	Real	[0.0, 0.53]
T3	Real	[0.0005, 0.18]
TT4	Real	[0.002, 0.6]
T4U	Real	[0.017, 0.233]
FTI	Real	[0.002, 0.642]
Class	Integer	{1, 2, 3}

Fig. 1. Example of a figure caption. (*figure caption*)

IV. Methodology

The dataset was split into training set and testing set. 70% of the dataset is used for training and the remaining 30% is reserved for testing.

The class distribution is highly imbalanced (92.5% for the majority class) and a strong correlation exists among features, therefore feature selection techniques are used to reduce the dimension of the dataset.

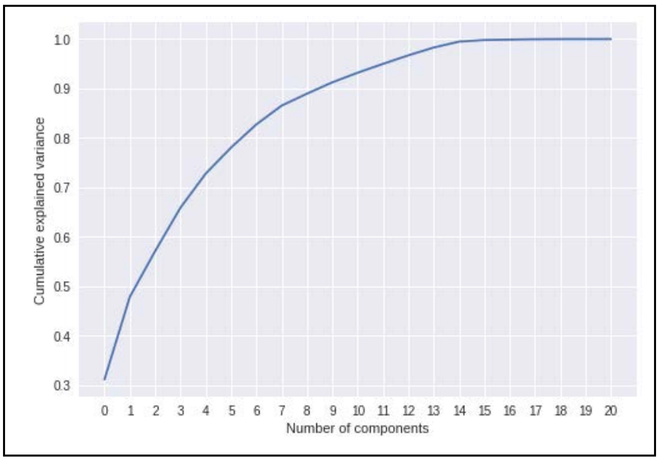

Figure 1 Feature variances

Two techniques are applied for dimensionality reduction. The first technique is Principal Component Analysis (PCA) retaining 95% of variance, which reduces the dimension to 13 components. The second technique selected is Step Forward Selection (SFS) using SVM with radial basis function kernel, reducing the dimensionality of the dataset to 10 features.

Random Oversampling, SMOTE and SCUT are applied to resample the data to address the imbalance of the class distribution in the training set. The resulting dataset is used to train Decision Trees, Random Forest, KNN and Support Vector Machine classifiers. A total of 36 experiments are performed combining a dimensionality reduction technique and a resampling method.

V. Random Oversampling

Random oversampling technique is applied to the minority classes (Normal, Hyperthyroidism) to adjust the class distribution of the training set, resulting in a total of 4451 instances per class.

The best performance was obtained with Random Forest using Step Forward Selection (SFS) scoring 99.8% of accuracy.

TABLE II. Random Forest with SFS

	Precision	*Recall*	*F1-score*
Normal	0.98	0.98	0.98
Hyperthyroidism	1.00	1.00	0.99
Hypothyroidism	1.00	1.00	1.00
		Accuracy	99.8%

Fig. 2. Performance of Random Forest with SFS.

VI. SMOTE

Chawla et al[9] propose a technique for over sampling called SMOTE (Synthetic Minority Oversampling Technique) in which the minority class is oversampled by creating synthetic examples. After applying SMOTE to the training set, a total of 13353 instances are obtained, 4451 per class.

The best performance obtained with this technique is achieved by Random Forest classifier with no feature selection, scoring 99.7% of accuracy.

TABLE III. Random Forest no feature selection

	Precision	*Recall*	*F1-score*
Normal	0.98	0.94	0.96
Hyperthyroidism	0.99	0.99	0.99
Hypothyroidism	1.00	1.00	1.00
		Accuracy	99.7%

Fig. 3. Performance of Random Fores with no feature selection.

VII. SCUT

Agrawal et al[2] propose a technique called SCUT (SMOTE and Clustered Undersampling Technique) which combines SMOTE to resample the minority classes and a cluster-based undersampling on the majority classes.

The first step of the SCUT technique is to split the dataset into *n* subsets where *n* is a class in the dataset. The second step is to calculate the media among all the classes. This is the parameter of comparison for each subset, if the subset is less than the media, an oversampling is perform using SMOTE such that the number of instances in the subset is equal to the media.

If the subset is higher than the media, a cluster-based undersampling is conducted using Expectation Maximization (EM) algorithm (Dempster et al, 1977) to discover the inner class clusters. A random selection is made on each cluster, such that the total number of instances selected from all the clusters equal to the mean. The EM algorithm is used to find soft clusters by returning a maximum likelihood of an instance to belong to a Gaussian distribution.

TABLE IV. Class distribution after SCUT

	Original		***SCUT***
Normal	114		1608
Hyperthyroidism	259		1608
Hypothyroidism	4451	Cluster 1	536
		Cluster 2	536
		Cluster 3	536
Total	4824	***Total***	4824

Fig. 4. Class distribution of the training set after applying SCUT.

The best performance obtained with this technique is achieved by Random Forest classifier with no feature selection, scoring 99.7% of accuracy.

TABLE V. RANDOM FOREST WITH NO FEATURE SELECTION

	Precision	***Recall***	***F1-score***
Normal	0.96	1.00	0.98
Hyperthyroidism	0.97	1.00	0.99
Hypothyroidism	1.00	1.00	1.00
		Accuracy	99.7%

Fig. 5. Performance of Random Forest with no feature selection.

VIII. CONCLUSIONS

The evaluated feature selection techniques and resampling methods show to be effective in the diagnosis of thyroid disease, although they do not show a significant improvement compared with a random oversampling method.

IX. REFERENCES

[1] S. Razia, S. Prathyusha, N. Krishna and N. Sumana, "A Comparative study of machine learning algorithms on thyroid disease prediction", International Journal of Engineering & Technology, vol. 7, no. 28, pp. 315-319, 2018.

[2] A. Agrawal, H. Viktor and E. Paquet, "SCUT: Multi-Class Imbalanced Data Classification using SMOTE and Cluster-based Undersampling", in 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2015), 2015, p. 2.

[3] G. Banu, "Predicting Thyroid Disease using Linear Discriminant Analysis(LDA) Data Mining Technique", Communications on Applied Electronics (CAE), vol. 4, no. 12, pp. 4-6, 2016.

[4] F. Gharehchopogh, M. Molany and F. Mokri, "Using Artificial Neural Network In Diagnosis Of Thyroid Disease: A Case Study", International Journal on Computational Sciences & Applications (IJCSA), vol. 3, no. 4, pp. 49-61, 2013.

[5] M. Kousarrizi, F. Seiti and M. Teshnehlab, "An Experimental Comparative Study on Thyroid Disease Diagnosis Based on Feature Subset Selection and classification", International Journal of Electrical & Computer Sciences IJECS-IJENS, vol. 12, no. 1, pp. 13-19, 2012.

[6] S. Pandey, D. Gour and V. Sharma, "Comparative Study on Classification of Thyroid Diseases", International Journal of Engineering Trends and Technology (IJETT), vol. 28, no. 9, pp. 457-460, 2015.

[7] M. Rahman and D. Davis, "Addressing the Class Imbalance Problem in Medical Datasets", International Journal of Machine Learning and Computing,, vol. 3, no. 2, pp. 224-228, 2013.

[8] E. Dogantekin, A. Dogantekin and D. Avci, "An automatic diagnosis system based on thyroid gland: ADSTG", Expert Systems with Applications, vol. 37, pp. 6368–6372, 2010.

[9] N. Chawla, K. Bowyer, L. Hall and W. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique", Journal of Artificial Intelligence Research, vol. 16, pp. 321-357, 2002.

[10] A. Dempster, N. Laird and D. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm", Journal of the Royal Statistical Society., vol. 39, no. 1, pp. 1-38, 1977.

[11] UCI Repository of Machine Learning Databases, University of California at Irvine, Department of Computer Science. Available: http://archive.ics.uci.edu/ml/machine-learning-databases/thyroid-disease (Accessed: 14 Dec 2018).