AN EXPERIMENTAL STUDY ON HYPOTHYROID USING ROTATION FOREST

Sheetal Gaikwad[1] and Nitin Pise[2]

[1]PG Scholar, Department of Computer Engineering,Maeers MIT,Kothrud,Pune,India
[2]Associate Professor, Department of Computer Engineering,
Maeers MIT, Kothrud, Pune,India

ABSTRACT

This paper majorly focuses on hypothyroid medical diseases caused by underactive thyroid glands. The dataset used for the study on hypothyroid is taken from UCI repository. Classification of this thyroid disease is a considerable task. An experimental study is carried out using rotation forest using features selection methods to achieve better accuracy. An important step to gain good accuracy is a pre- processing step, thus here two feature selection techniques are used. A filter method, Correlation features subset selection and wrappers method has helped in removing irrelevant as well as useless features from the data set. Fourteen different machine learning algorithms were tested on hypothyroid data set using rotation forest which successfully turned out giving positively improved results.

KEYWORDS

Filter method, Hypothyroid, Rotation forest, Wrapper method.

1. INTRODUCTION

With increasing number of population in the world, the medical field is facing a big challenge to keep the fitness of every human being. Thus the scope of medical science has increased tremendously, lead to accurate medical diagnosis. Using machine learning techniques, given a medical dataset as input it should accurately diagnose the datasets. The major focus in this paper is for hypothyroid diseases medical data set. Thyroid hormones produced by the thyroid gland helps control the body's metabolism. Hyperthyroidism or an overactive thyroid can cause inflammation of the thyroid. The seriousness of thyroid disorders should not be underestimated as thyroid storm (an episode of severe hyperthyroidism) and myxoedema coma (the end stage of untreated hypothyroidism) may lead to death in a significant number of cases [1, 3]. Thyroid function diagnosis is an important classification issue. Proper interpretation of the thyroid data, besides clinical examination and complementary investigation, is an important problem in the diagnosis of thyroid disease [1, 2, 4].

2. PRE-PROCESSING

Pre-processing is the import step need to be followed to achieve good classification of data. Pre-processing is nothing but cleaning of data to remove unwanted or bad data. A data set is said to be a bad data when it has incomplete values or missing data also the information which doesn't helps in classification are unwanted data. The major two techniques for getting good data from the thyroid data set which are used in this paper are Correlation based feature subset selection and wrappers technique.

DOI : 10.5121/ijdkp.2014.4603

2.1. Thyroid dataset

The hypothyroid data set has been taken from UCI machine learning repository. The data set has about 3772 instances and 11 attributes (features).It has 4 number of classes out of which 8 are nominal and 3 are numeric type.

2.2. Filter method

One of the most commonly used filter technique is Correlation based feature subset selection (CFS) which helps in removing irrelevant features. This CFS algorithm tries to calculate the relevance of the attributes with the class i.e. a feature is evaluate to be good or bad based on its high or low relevance with the class. If a feature is highly correlated to the class and inverse uncorrelated to any other features then that feature is said to be a good feature and may be considered for classification, else the feature is said to be irrelevant feature and can be discarded in classification process. In our study a Best First Search (BFS) searching technique is been used. Thus using this method we get optimal selected features than heuristic features to test for its relevance which is very important in medical data sets to be evaluated. Based on these techniques of CFS using BFS the scores of each subset of features are calculated with Eq. (1)

$$Merit_{S_k} = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}}. \qquad (1)$$

Where k defines number of feature, r_{cf} is correlation between class and features, r_{ff} is correlation between feature and $merits_k$ defines value of correlated features k.

2.3. Wrapper method

This is also a feature subset selection technique where it finds the feature subsets based on the performance of the preselected classification algorithm and a training data set. This preselected classification algorithm is also known as induction algorithm. In this paper J48 algorithm is been selected as an induction algorithm for wrapper subset selection method where the data set learning over J48 algorithm and selects the best possible features using Best First Search (BFS) approach

2.4. Proposed combining feature method

This is the very important step where a new and final features subset will be created as this feature subset will have all relevant attributes as well is trained on J48 classifier and is created. In short this new subset will be a union subset of CFS and Wrapper using BFS searching technique. Thus advantages from both the techniques are achieved i.e.by using CFS, irrelevant features are eliminated which enhances the speed of classification and by using Wrapper, increasing the accuracy the feature which is precisely selected but evaluating them on J48 classifier.

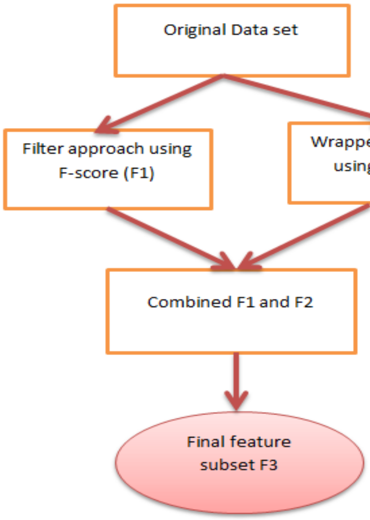

Original Data set
Filter approach using F-score (F1)
Wrapper Approach using J48 (F2)
Combined F1 and F2
Final feature subset F3

Figure 1. Wrapper method

3. ROTATION FOREST

Rotation Forest is an ensemble method which trains L decision trees independently, using a different set of extracted features for each tree. Let x = [x_1, . . . ,x_n]T be an example described by n features (attributes) and let X be an N x n matrix containing the training examples. We assume that the true class labels of all training examples are also provided. Let D = D_1, . . . , D_L be the ensemble of L classifiers and F be the feature set. Rotation Forest aims at building accurate and diverse classifiers. Bootstrap samples are taken as the training set for the individual classifiers, as in bagging. The main heuristic is to apply feature extraction and to subsequently reconstruct a full feature set for each classifier in the ensemble.

To do this, the feature set is split randomly into K subsets, principal component analysis (PCA) is run separately on each subset, and a new set of n linear extracted features is constructed by pooling all principal components. The data is transformed linearly into the new feature space. Classifier D_i is trained with this data set. Different splits of the feature set will lead to different extracted features, thereby contributing to the diversity introduced by the bootstrap sampling. We chose decision trees as the base classifiers because they are sensitive to rotation of the feature axes and still can be very accurate. The effect of rotating the axes is that classification regions of high accuracy can be constructed with fewer trees than in bagging and AdaBoost. Our previous study reported an experiment whose results were favourable to Rotation Forest compared to bagging, AdaBoost and Random Forest with the same number of base classifiers. The design choices and the parameter values of the Rotation Forest were picked in advance and not changed during the experiment. The prerequisite followed were number of features in a subset is set to 3, number of classifiers in the ensemble is set to 10, extraction method used is principal component analysis (PCA) and base classifier model is decision tree (hence the name forest).

3. EXPERIMENTATION

This figure 2 shows the experimentation followed to achieve the results for hypothyroid data set.

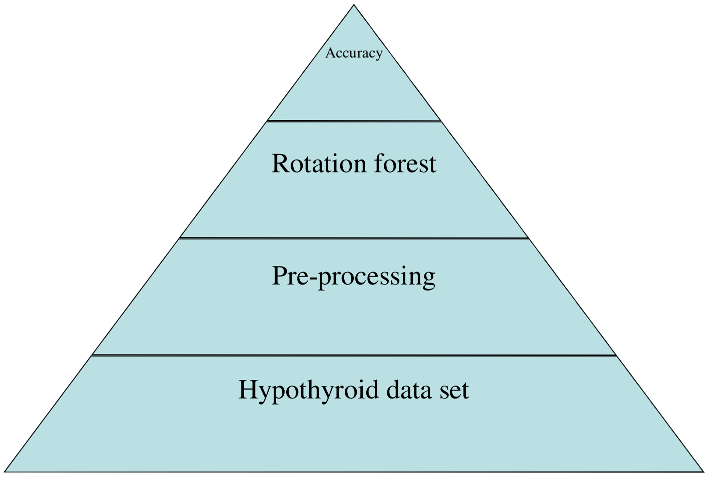


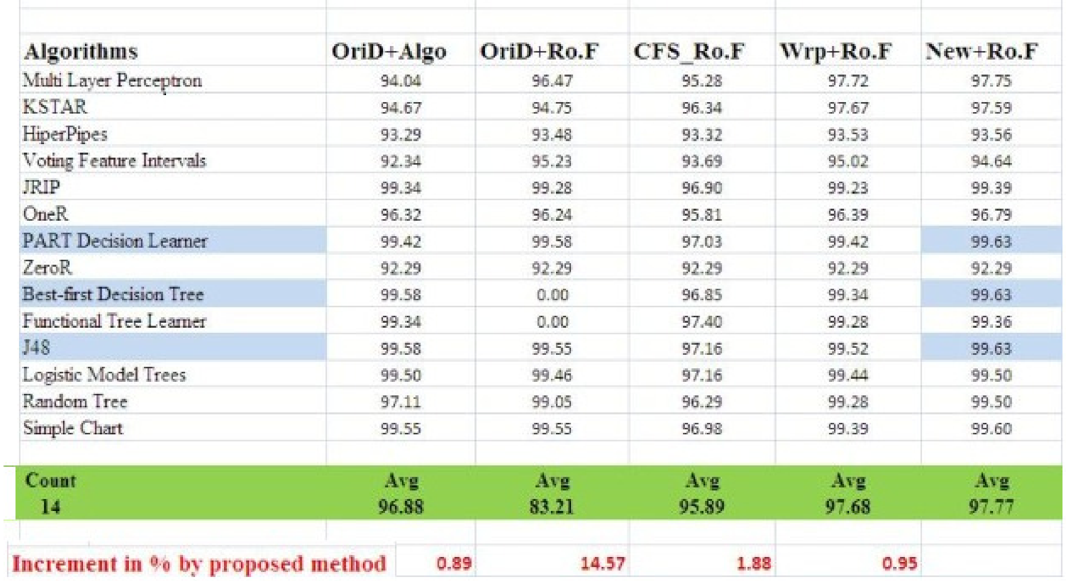
Figure 2. Experimentation process

The details process is described as follows:

- Hypothyroid medical data set in (dot).arff format is given as an input to the system.
- Execute the original data set through filter technique. The techniques used is Correlation based feature subset selection, where a feature subset is given out as the output of this algorithm, let's say that subset A.
- Execute the original hypothyroid data set through Wrappers techniques using the inducer as J48 classifier and get a new feature subset B.
- Later the union of both filter and wrapper feature subset is used to create a new data set in arff (Attribute relation file format) format.
- This new data set with selected features is given as input to rotation forest ensemble with different base classifiers/algorithms.
- The accuracy was noted down with combination of fourteen algorithms as base classifier to rotation forest.

4. RESULTS

The figure 3 shows different experimentation combination followed to justify the better classification accuracy achieved by this method on hypothyroid data set using rotation forest. The result is captures in five rows for all 14 algorithms where first row shows Original Hypothyroid Data set with respective algorithm individually (OriD+Algo), Original Data set with rotation forest (OriD+Ro.F), Dataset with CFS on rotation forest (CFS+Ro.F), Dataset with Wrapper Subset on rotation forest (Wrpper+Ro.F) and New generated feature subset after combination with ensemble rotation forest (New+Ro.F)



Algorithms	OriD+Algo	OriD+Ro.F	CFS_Ro.F	Wrp+Ro.F	New+Ro.F
Multi Layer Perceptron	94.04	96.47	95.28	97.72	97.75
KSTAR	94.67	94.75	96.34	97.67	97.59
HiperPipes	93.29	93.48	93.32	93.53	93.56
Voting Feature Intervals	92.34	95.23	93.69	95.02	94.64
JRIP	99.34	99.28	96.90	99.23	99.39
OneR	96.32	96.24	95.81	96.39	96.79
PART Decision Learner	99.42	99.58	97.03	99.42	99.63
ZeroR	92.29	92.29	92.29	92.29	92.29
Best-first Decision Tree	99.58	0.00	96.85	99.34	99.63
Functional Tree Learner	99.34	0.00	97.40	99.28	99.36
J48	99.58	99.55	97.16	99.52	99.63
Logistic Model Trees	99.50	99.46	97.16	99.44	99.50
Random Tree	97.11	99.05	96.29	99.28	99.50
Simple Chart	99.55	99.55	96.98	99.39	99.60
Count 14	Avg 96.88	Avg 83.21	Avg 95.89	Avg 97.68	Avg 97.77
Increment in % by proposed method	0.89	14.57	1.88	0.95	

Figure 3. Hypothyroid results

From the above figure 3 it makes clear that result obtained by our technique of combining feature of filter and wrapper as an input to rotation forest has given improved results. The classification accuracy gain has improved compared to other data set combinations as well as have shown good results over three algorithms as base classifier selected for rotation forest. The figure 4 gives a clear difference of the achieved results by our method for hypothyroid than original hypothyroid data set.

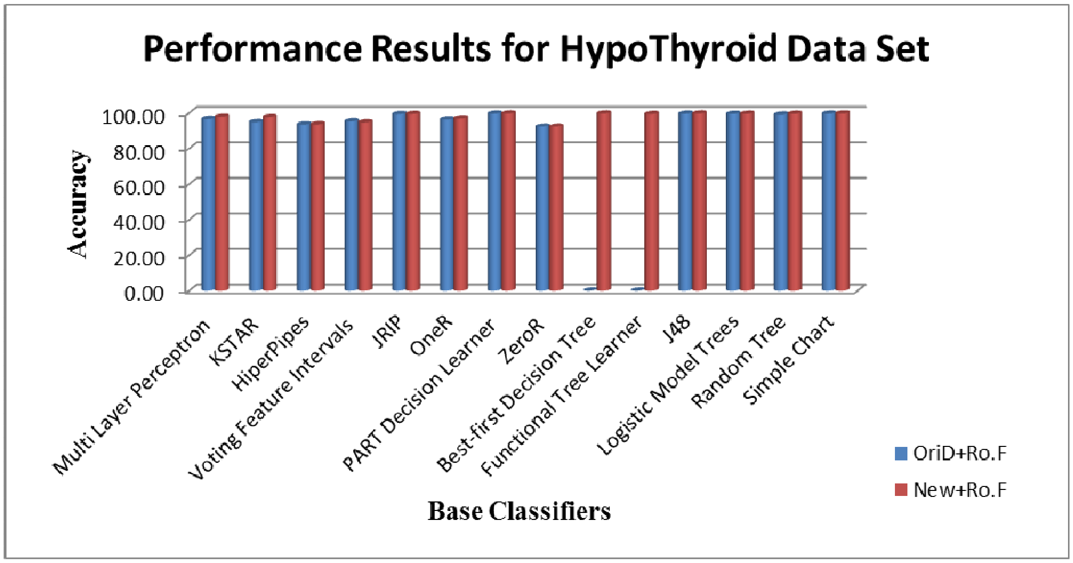


Figure 4. Hypothyroid Performance results

It is found that hypothyroid data set execution with rotation forest gave good accuracy than dataset execution only with the classifier, but the accuracy results for new subset gave even higher results than both of the above methods. This experimentation was carried out for fourteen medical datasets and it gave all successful and improved results for all. The proposed method for hypothyroid using rotation forest has improved by an average percentage of 14.57% which is really a big achievement. Also it has shown positively good results and improvement in comparison with other combination methods. Maximum of 96.63% classification accuracy has been gained by considering three algorithms like PART Decision learner, Best first decision tree and J48 respectively.

5. CONCLUSIONS

In this experimental study on hypothyroid using rotation forest has given good classification accuracy. As hypothyroid has large features which indirectly increase classification execution time as well as may lead to misclassification , a combining feature subset method has proved very useful pre-processing step to remove unwanted features. Further the performance of the classification was boosted by the use of rotation forest ensemble technique. For about fourteen algorithm used as base classifier for rotation forest gave improved results over all other combinations methods. Most importantly this experimental study showed 14.57% more improved accuracy over the original data set with rotation forest. Thus this proposed method has shown up to 99.63% classification accuracy for hypothyroid data set which will boost the medical diagnosis efficiently.

REFERENCES

[1] L. Ozyilmaz and T. Yildirim (2002). “Diagnosis of thyroid disease using artificial neural network methods,” in: Proceedings of ICONIP’02 9th international conference on neural information processing (Singapore: Orchid Country Club) pp. 2033–2036.

[2] K. Polat, S. Sahan and S. Gunes (2007). “A novel hybrid method based on artificial immunerecognition system (AIRS) with fuzzy weighted pre-processing for thyroid disease diagnosis,” Expert systems with Applications, vol. 32, pp. 1141-1147.

[3] G. Zhang, L.V. Berardi(2007). “An investigation of neural networks in thyroid functiondiagnosis,” Health Care Management Science, 1998, pp. 29-37.
Available: http://www.endocrineweb.com/thyroid.html, (Accessed: 7)

[4] Baris Senliol, et al (2008). “Fast Correlation Based Filter (FCBF) with a different searchstrategy”, Computer and Information Sciences, ISCIS'08. 23rd International Symposium on. IEEE.

[5] Blake.C and Merz.L (1998). UCI repository of machine learning databases, Department of Information and Computer Science, University of California at Irvine, Irvine CA.

[6] Brown, G., Pocock, A., Zhao, M.-J., Lujan, M (2012).Conditional Likelihood Maximisation: AUnifying Framework for Information Theoretic Feature Selection. The Journal of Machine Learning Research (JMLR).

[7] Dash .M, Liu .H (1997). “Feature Selection for Classification”, Intelligent Data Analysis, pp 131–156.

[8] Deng H and Runger G. (2012). “Feature Selection via Regularized Trees”, in International Joint Conference on Neural Networks (IJCNN), IEEE.

[9] Dietterich, T (2000). “Ensemble methods in machine learning”, Proceedings of the 1st International Workshop on Multiple Classifier Systems.

[10] Dr. Bhargava Neeraj, Sharma Girja, Dr. Bhargava Ritu and Mathuria Manisha (2013). “Decision Tree Analysis on J48 Algorithm for Data Mining”, International Journal of Advanced Research in Computer Science and Software Engineering (JARCSSE), Volume 3, Issue 6.

[11] Duch, W (2000). “Similarity based methods: a general framework for classification approximation and association”, Control and Cybernetics.

[12] Guyon, I., Elisseeff, A (2003). “An Introduction to Variable and Feature Selection”, Journal of Machine Learning Research 3, Volume 3,pp 1157-1182

[13] Hall Mark A (2009). “Correlation based Feature Selection for Machine Learning”, Thesis for the degree of Doctor of Philosophy at The University of Waikato.

[14] Hsu Hui-Huang, Hsieh Cheng-Wei, Lu Ming-Da (2011). “Hybrid feature selection by combining filters and wrappers”, Elsevier Expert Systems with Applications Journal, ISSN 09574174.

[15] http://www.cs.waikato.ac.nz/ml/weka/ WEKA 3 tool. A Weka Tool of The University of Waikato for classifiers comparison.

[16] Kohavi Ron and John George H. (1997). “Wrappers for feature subset selection. Elsevier”, Artificial Intelligence Volume 97 Issue 1-2, pp 273 - 324.

[17] Kotsiantis S. B. and Pintelas P. E. (2004). “Hybrid Feature Selection instead of Ensembles of Classifiers in Medical Decision Support”, Proceedings of Information Processing and Management of Uncertainty in Knowledge Based Systems.

[18] Ludmila I. Kuncheva and Juan J. Rodríguez (2007). “An Experimental Study on Rotation Forest Ensembles”, In Springer Book Title Multiple Classifier Systems pp 459-468, 7th International Workshop, Proceedings.

AUTHORS

Sheetal Gaikwad received the B.E Degree in Information Technology from Cummins College of Engineering for Women, Pune, India in 2011. She is pursuing M.E in Computer Engineering at Maeers MIT College, Kothrud, Pune, India. She has cleared GATE 2012 with good score. She has worked as a visiting faculty in Cummins College of Engineering as well as a teaching assistantship in MIT College for 2 years. She has published papers in international conferences proceedings. Her research areas are Data Mining, Security and Artificial Intelligence.

Prof. Nitin Pise is currently working as an Associate professor in Maeers MIT College, Kothrud, Pune, India. He is pursuing Ph.D from COEP, Pune, India. He has about 19 year of experience in teaching industry. He has published about more than 20 research articles/papers in national as well as in international journal and conferences. His specialization is in Data Mining, Distributed Systems and Artificial Intelligence.