Expert Advice Ensemble for Thyroid Disease Diagnosis

Muhammad Anjum Qureshi, Kubilay Eksioglu
Department of Electrical and Electronics Engineering, Bilkent University, Ankara, Turkey
{qureshi, eksioglu}@ee.bilkent.edu.tr

***Abstract*—Thyroid gland influences the metabolic processes of human body due to the fact that it produces hormones. Hyperthyroidism in caused due to increase in the production of thyroid hormones. In this paper a methodology using an online ensemble of decision trees to detect thyroid-related diseases is proposed. The aim of this work is to improve the diagnostic accuracy of thyroid disease. Initially, feature rejection method is applied to discard 10 irrelevant and redundant features from 29 features. Then, it's shown that the offline ensemble of decision trees provides higher performance than state-of-the-art methodologies. Afterwards, the exponential weights based online ensemble method is implemented which reaches comparable classification performance with offline methodology. The proposed system consists of three stages: feature rejection, training decision trees with different cost schemes and the online classification stage where each classifier is weighted using an exponential weight based algorithm. The performance of online algorithm increases as the number of samples increases, because it continuously updates the weights to improve accuracy. The achieved classification accuracy proves the robustness and effectiveness of online version of proposed system in thyroid disease diagnosis.**

Keywords. thyroid, feature rejection, decision trees, Offline Ensemble, Online Ensemble, exponential weights.

I. Introduction

The health care environment contains intensive data, handling this plenty of information, specifically searching the relevant data, in reduced time for physician is critical. Medical diagnosis is increasingly becoming dependent in variety of techniques of data mining. Gathering data from patient, and analysis by artificial intelligence expert system, and finally displaying the processed data to expert for decision making are vital factors [8], [11]. The real-time classification algorithms have emerged that analyze the decision making process and make improvements in real-time. To provide an expert system, one needs to answer the following questions online[14]: What will be the cost for wrong decision? What will be the weights for different classifiers? How the system will adapt and learn these weights and costs to maximize their performance?

In this study, an online learning method is used which continuously learns and updates the weights for classification in real-time. First, all the features are normalized and the unnecessary features are removed using feature rejection method. Two decision trees are trained, one is a simple decision tree based classifier while second classifier is the cost aware version of first. The system is initially proposed in offline setting. It's shown that the offline ensemble of these classifiers provides better results than the individual classifiers. In online version, this final classification stage is continuously updated using EXP4 [9]. Each arrival of test data updates these weights based on the decision accuracy of the contesting classifiers.

Main contributions of this paper are:

- We propose a pipeline of decision trees to classify various thyroid-related diseases and provide numerical comparison with state-of-the-art classification methods for thyroid disease data set and superiority is shown.
- This work uses the EXP4 for online classification. The accuracy achieved with online solution is comparable with the offline benchmark performance.

II. Related Work

In medical diagnosis work, researchers aimed to provide solutions based on state-of-the-art algorithms including k-Nearest Neighbors (k-NN), Support Vector Machines (SVM), and Neural Networks [8], [11]. For datasets with unbalance class structure, decision tree classification provided significant improvement over class based accuracies. These cost-sensitive decision tree methods provided effective imbalanced classification [2], [7].

The ensemble system contains two or more classifiers, and based on their predictions, combined decision is provided for classification [12]. Recent works in the field of online learning proven to be acceptable in every field. These algorithms learns with the real time data, online performance is calculated with the expert advice, and then system adapts to these changes to improve the accuracy [9]. In this paper, ensemble of two decision trees is used, one in simple form and other in cost sensitive approach. Furthermore, EXP4 is used to provide online classification for thyroid disease diagnostics.

Feature extraction and selection are considered as preprocessing steps in the field of classification [6]. In this paper, feature rejection method based on scree plot is used, which is independent of both classifiers and expert knowledge of related field. In medical diagnosis, original features presented to the expert have more value than the projected features in weighted representation.

III. Problem Formulation

Let X be a d-dimensional dataset, where $x_i = (x_{i1}, x_{i2}, ..., x_{id})\ \forall x_i \in X$. L be the set of labels and $l_i \in L$ be the ith label in L. Set of experts (classifiers) is denoted by E and C is the cost matrix, where $C(i, j)$ is the cost of

978-1-5090-6494-6/17/$31.00 2017 IEEE

classifying label l_i as l_j. The goal is to decrease the number of misclassifications in thyroid disease diagnosis.

IV. Classification System

A. Feature Normalization

Feature normalization is preliminary step, to equalize the features in terms of distance. The linear scaling to unit range is adopted to normalize the ranges. Let a be the continuous attribute vector received from all samples for a single attribute. Then normalized a, denoted as $\tilde{a}$, is calculated as:

$$\tilde{a} = \frac{a - a_{min}}{a_{max} - a_{min}}$$

Where, a_{max}, a_{min} are the lower and upper limits respectively for attribute vector a.

B. Feature Rejection

In classification problem, feature extraction is treated as pre-processing technique. In this paper, feature rejection technique is used, to target both extraction and selection phases in one step. The original data is preserved in this rejection, the features are rejected from the original data without altering the selected features. The irrelevant and redundant information is rejected using the scree plot of eigenvalues. The eigenvalues and vectors are obtained via Singular Value Decomposition in sorted form. The singular value decomposition of an $m \times n$ matrix $\mathbf{Q}$ is a factorization of the form $\mathbf{U\Sigma V}^*$. The diagonal matrix $\mathbf{\Sigma}$ contains eigenvalues of matrix. The columns of $\mathbf{U}$ and $\mathbf{V}$ are vectors corresponding to eigenvalues. These values are used to obtain scree plot, which is used to find the point K, after which eigenvalues become negligible, or sharp trends in slope become diminishing (typical value is between 0.01 and 0.02). The vector corresponding to the smallest eigenvalue is selected. The component with maximum contribution in this vector is discarded. This process is repeated till the K components are obtained. Hence, reduced dimension is reached and also discriminative features are selected [5].
Fig. 1 provides the scree plot for 4 different data files available in thyroid dataset, features are same for all the files with same structure, that is the reason scree plot is similar for all four files. The graph indicates that the threshold is 0.015, the components below this threshold value are rejected, the selected features are $K = 19$ from 29 attributes. The main drawback in this method is that the one feature may correspond to two or more discarded components. In the selected dataset, rejecting 11 components corresponds to rejection of 10 features.

C. Offline Ensemble Classifier

After experts are trained, each expert provides advice in the form of a probability mass function mapping labels to probabilities. For a sample in the training dataset, experts provide advices and these advices are combined into a row vector. For each sample this process is repeated and row vectors are combined into a $M \times |E||L|$ matrix. Then this matrix is used to train one-vs-all binary SVMs for each class with linear kernel function and error connecting output codes is used to convert binary SVM to multiclass [4].

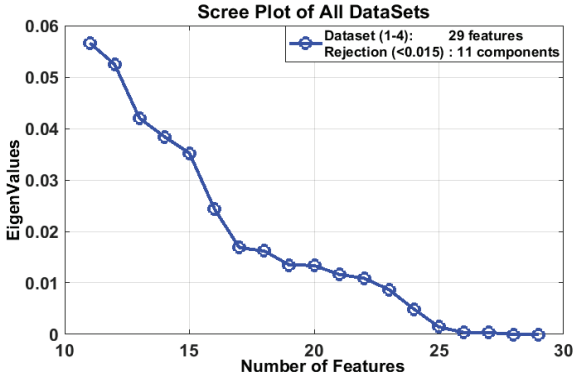

Fig. 1. Scree plot of thyroid dataset: 5 different files

As it's visible from the Table II this offline ensemble methodology outperforms all well known methodologies. For systems where the sample arrival is online, EXP4 is used to provide online learning solution.

D. Online Ensemble with EXP4

Assume N experts, whose predictions are probability mass vectors, providing the posterior probability of each class. Such predictions of expert i, at time t is denoted by $f_{i,t}$. Algorithms predicts p_t and outcome y_t is revealed. Let γ be the learning rate parameter, and p_t is calculated as:

$$p_t = (1-\gamma)\frac{\sum_{i=1}^{N} w_{i,t-1} f_{i,t}}{\sum_{i=1}^{N} w_{i,t-1}} + \frac{\gamma}{|L|} \tag{1}$$

Once the outcome is revealed, all experts suffer a loss according to how different the actual labeling is than the advice. This loss function should penalize the expert based on if the labeling is correct or not and then how confident the expert is of its advice. Loss function should ensure that out of two wrong experts, the one more certain about its decision will be penalized further. Similarly, if the expert is correct but not very confident of the advice, it should be rewarded less then correct and confident expert. Certainty term can be defined as ratio of highest two probabilities; the higher this term, more certain the expert is. To reach such loss function, y_t is defined as a $|L| \times 1$ 1-hot encoded vector of the revealed label, all values are zero except the correct label.

Then, the loss function ℓ is proportional to the square of L_2 norm of the distance between two decisions:

$$\ell(f_{i,t}, y_t) = \frac{\|f_{i,t} - y_t\|^2}{|L|} \tag{2}$$

This loss function $\ell(x)$ must be convex to apply EXP4[9].

$$\ell(\lambda x + (1-\lambda)y) \leq \lambda\ell(x) + (1-\lambda)\ell(y). \quad \textit{for } 0 < \lambda < 1$$

It is easy to prove this convexity with euclidean norm following three properties.
1. Positivity: $0 \leq ||x||^2 < \infty$.
2. Homogeneity: $||\lambda x||^2 = |\lambda|\ ||x||^2$.
3. Triangular Inequality: $||A + B||^2 \leq ||A||^2 + ||B||^2$.

Let $A = \lambda x$ and $B = (1-\lambda)y$, then using triangular inequality property.

$$||\lambda x + (1-\lambda)y||^2 \leq ||\lambda x||^2 + ||(1-\lambda)y||^2$$

Applying Homogeneity property to right hand sight of the above equation.

$$||\lambda x + (1-\lambda)y||^2 \leq |\lambda|\ ||x||^2 + |(1-\lambda)|\ ||y||^2$$

Loss function is just scalar division of $|L|$ with the norm. Hence,

$$\ell(\lambda x + (1-\lambda)y) \leq \lambda\ell(x) + (1-\lambda)\ell(y)$$

For exponential weights algorithm, when $\gamma = \sqrt{\frac{8lnN}{T}}$ it's shown in the literature that the total regret will be bounded by $\sqrt{\frac{T}{2}lnN}$. Since how many times a new patient will be received is unknown, doubling trick is used, whenever $log_{10}(t)$ is an integer, previous knowledge is reset and algorithm is re-trained with the samples that are received before time t.

Algorithm 1 Exponential Weights Based on Probabilities

```
function EXP4
    for each t=1,2,... do
        T ← 10^⌈log10(t)⌉
        γ ← sqrt(8lnN / T)
        Get advice vectors f_{i,t} for i = 1...N
        Get p_t using f_{i,t}, γ
        label ← arg max_{l∈L} p_t[l]
        Observe y_t
        w_{i,t} ← w_{i,t-1} exp(γ ℓ(f_{i,t}, y_t))
```

V. Illustrative Results

In this section, performance of proposed method is evaluated and compared with numerous other state-of-the-art classification methods on thyroid diagnosis dataset.

Dataset: The data used for thyroid disease classification task is available in UCI (Center of Machine Learning and Intelligent Systems, University of California) Machine Learning Repository. Few people were belonging to the primary hypothyroid or compensated hypothyroid group for all tasks ('allhypo','allhyper','allbp' and 'allrep'), providing 92.5%, 95%, 95% and 97% of cases belong to the healthy group respectively[3]. From the tests applied to these patients, for all tasks 29 features are extracted, of which 7 are continuous and 22 are binary. For these classification tasks, total of 2800 cases are given for training and 972 cases for testing. Feature rejection method is applied before performing experiments.

A. Decision trees

In this work, two CART decision trees are used as described by Breiman et. al [1]. First decision tree was trained without any cost, only prior class probabilities are provided to decrease impacts of class imbalance. Second decision tree was trained with cost, to penalize labeling an ill patient as negative. In cost matrix calculations it's assumed the cost of labeling two diseases as each other are equal, though this information may be updated by further domain knowledge.

To calculate cost, α-level Neyman-Pearson test is used [10]. After decision tree without cost is trained, the threshold τ setting alpha to $\alpha = 0.05$ is found, and $C(ill, negative)$ is set as $1/\tau$.

TABLE I
Comparison of feature selection/rejection Methods

Method	Time (sec)	Accuracy
GA	44.76578	97.22
MDLM	6.36728	96.91
Scree-plot	**0.00375**	**96.91**

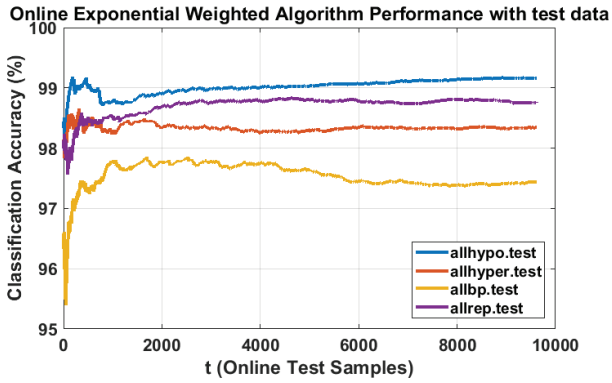

Fig. 2. Online Algorithm learning experiment

B. Feature Selection (Table I)

This experiment compares the scree-plot feature rejection with two other feature selection methods, Genetic Algorithms (GA) and Minimum Description Length Method (MDLM). All the methods are implemented to select best subset of 4 features from dataset. The performance of Neural Network classification for 'allbp' thyroid file is provided in Table I.

C. Online Learning Experiment (Fig. 2)

This experiment evaluates the performance of the classifier with time of test sample arrival. The experts are trained with training dataset, then EXP4 algorithm is implemented on test data in real-time. The test data is up-sampled and 10,000 samples with replacement is generated from the test data and EXP4 is performed. This experiment is repeated 50 times (before each run test data is shuffled), and average of these experiments is obtained. The average performance on all thyroid dataset files is shown in Fig. 2.

D. Spread Calculation Experiment (Fig. 3)

This experiment calculates the mean and spread of classification accuracy of system. This experiment performs the 3-fold validation over full dataset. This experiment is performed 50 times, each time the mergedd data is shuffled. Finally, mean

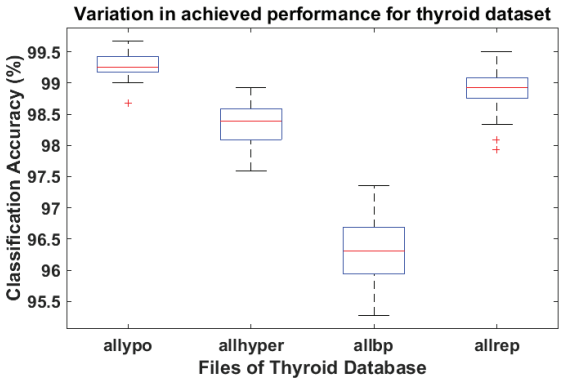

Fig. 3. Performance for 3-folded Validation Experiment

TABLE II
COMPARISON OF CLASSIFICATION ACCURACY % OVER SEPARATE TEST DATA (972 SAMPLES)

Method	'allhypo'	'allhyper'	'allbp'	'allrep'
SVM	96.40	98.05	97.33	96.91
k-NN	96.50	97.33	95.27	96.81
Neural Network	94.96	97.94	97.43	96.19
Decision Tree	98.87	97.33	94.24	98.77
Proposed System	**99.18**	**98.46**	**97.43**	**98.97**

and standard deviation of these experiments for all the data files in thyroid dataset are calculated and plotted in Fig. 3.

E. Comparison with state-of-the-art algorithms (Table II)

As Table II shows, the classification accuracy of proposed system is compared with Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), Decision Tree algorithm and Neural Network (Single hidden layer with 10 units). Initially, all the classifiers are trained from full training data, then performance is evaluated on separate test data (with 972 test samples) available in the repository. The superiority of the proposed algorithm is validated.

VI. CONCLUSION

In this paper we propose a pipeline of well known decision trees to classify different thyroid-related diseases. The proposed method reaches an accuracy between 97.43% and 99.18% for different thyroid tasks, which proves the universal performance validation. The feature rejection is also applied on the dataset, rejecting nearly half of the irrelevant features to reduce the computational complexity. The offline version of the proposed algorithm sets the performance benchmark, which implies that superior performance is reached. EXP4 is implemented for online version of the system. The extensive numerical results show the comparable performance of online version with offline performance benchmark, and superiority over all state-of-the-art classification methods.

REFERENCES

[1] Stone Olshen Breiman, Friedman. *Classification and Regression Trees*. 1984.

[2] Xiao Lin Chen, Yan Jiang, Min Jie Chen, Yong Yu, Hong Ping Nie, and Min Li. A dynamic cost sensitive support vector machine. In *Advanced Materials Research*, volume 424, pages 1342–1346. Trans Tech Publ, 2012.

[3] Włodzisław Duch, Rafał Adamczak, and Krzysztof Grabczewski. Extraction of crisp logical rules from medical datasets. 1997.

[4] Sergio Escalera, Oriol Pujol, and Petia Radeva. Separability of ternary codes for sparse designs of error-correcting output codes. *Pattern Recognition Letters*, 30(3):285–297, 2009.

[5] Imola K Fodor. A survey of dimension reduction techniques, 2002.

[6] Akanksha Juneja, Bharti Rana, and RK Agrawal. A combination of singular value decomposition and multivariate feature selection method for diagnosis of schizophrenia using fmri. *Biomedical Signal Processing and Control*, 27:122–133, 2016.

[7] Bartosz Krawczyk, Michał Woźniak, and Gerald Schaefer. Cost-sensitive decision tree ensembles for effective imbalanced classification. *Applied Soft Computing*, 14:554–562, 2014.

[8] Fatma Latifoğlu, Kemal Polat, Sadık Kara, and Salih Güneş. Medical diagnosis of atherosclerosis from carotid artery doppler signals using principal component analysis (pca), k-nn based weighting pre-processing and artificial immune recognition system (airs). *Journal of Biomedical Informatics*, 41(1):15–23, 2008.

[9] Nick Littlestone and Manfred K Warmuth. The weighted majority algorithm. In *Foundations of Computer Science, 1989., 30th Annual Symposium on*, pages 256–261. IEEE, 1989.

[10] Jerzy Neyman and Egon S Pearson. On the problem of the most efficient tests of statistical hypotheses. In *Breakthroughs in statistics*, pages 73–108. Springer, 1992.

[11] Kemal Polat and Salih Güneş. An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease. *Digital Signal Processing*, 17(4):702–710, 2007.

[12] Michael L Raymer, Leslie A Kuhn, and William F Punch. Knowledge discovery in biological data sets using a hybrid bayes classifier/evolutionary algorithm. In *Bioinformatics and Bioengineering Conference, 2001. Proceedings of the IEEE 2nd International Symposium on*, pages 236–245. IEEE, 2001.

[13] E Ke Tang, Ponnuthurai N Suganthan, Xin Yao, and A Kai Qin. Linear dimensionality reduction using relevance weighted lda. *Pattern recognition*, 38(4):485–493, 2005.

[14] Cem Tekin, Jinsung Yoon, and Mihaela van der Schaar. Adaptive ensemble learning with confidence bounds. *IEEE Transactions on Signal Processing*, 65(4):888–903, 2016.