International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Rejection Threshold Optimization using 3D ROC Curves: Novel Findings on Biomedical Datasets

Asli Uyar[1], Yasemin Atilgan Sengul[2]

Submitted: 23/11/2020 Accepted : 02/03/2021

Abstract: Reject option is introduced in classification tasks to prevent potential misclassifications. Although optimization of error-reject trade-off has been widely investigated, it is shown that error rate itself is not an appropriate performance measure, when misclassification costs are unequal or class distributions are imbalanced. ROC analysis is proposed as an alternative approach to performance evaluation in terms of true positives (TP) and false positives (FP). Considering classification with reject option, we need to represent the tradeoff between TP, FP and rejection rates. In this paper, we propose 3D ROC analysis to determine the optimal rejection threshold as an analogy to decision threshold optimization in 2D ROC curves. We have demonstrated our proposed method with Naive Bayes classifier on Heart Disease dataset and validated the efficiency of the method on multiple datasets from UCI Machine Learning Repository. Our experiments reveal that classification with optimized rejection threshold significantly improves true positive rates in biomedical datasets. Furthermore, false positive rates remain the same with rejection rates below 10% on average.

Keywords: Decision threshold optimization, rejection threshold optimization, 3D ROC curves, Naive Bayes

1. Introduction

Reject option has been widely tackled by various researchers in the field of pattern recognition (e.g. [1-4]) in order to avoid potential misclassifications. A rejection in the classification task occurs when the cost of rejecting classifying an instance is lower than the cost of misclassification. Although it is expected to reduce misclassification rate via reject option, correct classifications may also be converted into rejects. Therefore, it is not easy to find statistically optimal error-reject characteristics of a system.

Typically, the trade-off between the errors and rejects has been used to describe, improve or compare the performance of various classification problems [5]. However, in cases when misclassification costs are unequal and class distributions are imbalanced or skewed, the overall error rate is not an adequate performance measure. Receiver Operating Characteristics (ROC) curve plots true positive rate (TPR) versus false positive rate (FPR) by adjusting the classification decision threshold [6]. ROC analysis provides statistically consistent and more comprehensive performance assessment compared to error rate. Although the ideal case is maximizing TP and minimizing FP rates, classifiers that triggers TPR more often, would also cause an increase in FPR [7]. Decision threshold optimization on typical two-dimensional (2D) ROC curves has been a common machine learning approach to balance the trade-off between TP and FP rate, and has been utilized in various classification problems including software defect prediction and medical diagnosis [8, 9].

Considering classification with reject option, error-reject characteristics should be interpreted in terms of TPR, FPR and reject rate (RR). In this paper, we propose 3D ROC analysis in order to optimize the rejection threshold while taking into account the tradeoff between TPR and FPR. We employ Naive Bayes classifier on eleven medical datasets. Our experimental results show that 3D ROC methodology significantly improves TPs with average RR below 10% while FPs remain statistically indifferent.

Our proposed methodology for 3D ROC analysis can be summarized as follows:

- Binary classification based on Naive Bayes classifier on eleven datasets: Naïve Bayes is easy to interpret and its results are fairly accurate in terms of posterior probabilities of classes. Although this algorithm does not have hyper parameters, the single parameter that could be optimized is the decision threshold on the class posterior probabilities.
- Optimization of decision threshold using 2D ROC analysis: Decision threshold is generally set to 0.5 by default in Naïve Bayes classification. However, the number of instances that belong to one class may be significantly lower than the other. Further, posterior probabilities of this minority class would often be lower than the other. Our objective is to optimize the decision threshold of Naive Bayes in order to avoid misclassification of minority class.
- Interpretation of reject option around the decision threshold: We have introduced a reject option for the instances whose posterior probabilities are less confident than a pre-defined bound. Instead of a single rejection threshold, we have defined a rejection interval whose left and right boundaries are defined as $[t_l, t_r]$.
- Visualization of the tradeoff between three constraints, i.e. TPR, FPR and RR, using 3D ROC curves: We generated 3 dimensional ROC surfaces to illustrate the effects of different rejection intervals on the performance of Naive Bayes classifier.
- Optimization of rejection interval using Euclidean distance on 3D ROC surface: Our approach aims to minimize the distance between (FPR, TPR, RR) and the ideal case (0,1,0).
- Employment of the model on eleven datasets in biomedical

[1] *Computer Eng., Okan University, İstanbul – 34959, TURKEY*
ORCID ID :0000-0002-7913-1083
[2] *Industrial Eng., Doğuş University, İstanbul – 34775, TURKEY*
ORCID ID : 0000-0002-5109-2262
** Corresponding Author Email: yatilgan@dogus.edu.tr*

domain: Paired t-tests between default threshold method and our experimental results on multiple datasets validate that reject option can be confidently used to improve TP rates, while keeping FP rates stable.

Related Work

In a landmark study, Chow investigated the optimum rejection rule using generalized Bayes rule and the costs for making an error, correct classification or rejection [1]. If the cost of rejections is equal to the cost of misclassifications, the reject threshold becomes zero, i.e. no pattern is rejected (simple Bayes rule). Chow's rule relies on the exact knowledge of posterior probabilities, so any algorithm that provides posterior probabilities as an output of its classification could be directly used to evaluate reject option. When $p(C_i|x)$ is the probability of an instance x being in the class i, the optimum rejection rule implies that an instance should be rejected if $p(C_i|x)$ is less than a threshold t. More precisely, its formula is given as follows:

$$\begin{cases} accept\ x\ if\ p(C_i|x) \geq (1-t) \\ reject\ x\ if\ p(C_{i|x}) < (1-t) \end{cases} \quad (1)$$

In this equation, two disjoint regions are defined such that V_a is the acceptance region that should be preserved as large as possible to classify instances and V_r is the rejection region that should be minimized to balance the tradeoff between these two measures. Chow presented a typical error-reject tradeoff curve for all levels of threshold, between 0 and 1 [1].

Error-reject curves have been widely used by other researchers like Gorski [5]. Gorski distinguished "good" lists, where the correct decision is made by a neural network classifier, from "bad" lists by adding the rejection option. He obtained Bayesian optimal error-reject characteristics by using cost sensitive classification and changing the decision threshold. Results of this study on a bank check recognition system showed that a single decision parameter allows to reject "bad", in other words, less confident, patterns and tunes the overall system towards the desired performance [5].

Many authors inspired from Chow's optimum reject rule and extended their work for various classification problems. One of these works has been completed by Hansen *et al.* [10] to discuss the effects of finite and infinite training sets for binary classifiers. Authors focused on setting accurate rejection boundaries in order to reject ambiguous inputs near the decision threshold. They also investigated this rejection boundary by increasing its thickness on different classifiers to form similar error-reject curves. Their results revealed that a scaled error-reject ratio would provide excellent fit to the data in digit recognition systems.

Besides error-reject curves, Tortorella utilized ROC curves in binary classification tasks with reject option, when prior knowledge/probability about the classes are not known in advance [11]. Most frequently, ROC curves evaluate the TPR as the fraction of actual positive classes classified as "positive" against the FPR as the fraction of actual negative classes classified as "positive" [11-14]. Tortorella also assessed the classification performance with a similar manner, but he used a cost-sensitive classification in order to optimize rejection threshold [11]. Our approach is similar to Tortorella's view, since they also searched for two rejection thresholds for two classes by taking distinct costs of true positives, false positives and rejections into account. Ceylan [15] also used ROC curves to show true positive ratio to the false positive rate for different thresholds of the classifier output. According to the experimental results in this study, Bayesian optimization-based K-Nearest Neighbor performs better results with the accuracy of 95.833%. However, assigning costs to misclassifications is very difficult since they are unknown or controversial.

Tosun and Bener have applied decision threshold optimization using Naïve Bayes classifier as the prediction algorithm, to find the optimum threshold for software defect data [16]. Their analysis showed that it is possible to decrease false alarms without changing the true positive rates. Another study [17] showed that using Naïve Bayes classifier resulted in higher accuracy on average 75.7%.

To extend 2D ROC analysis, 3D ROC curves have been utilized in few studies in order to assess the effects of three different parameters on the classification accuracy (e.g. [18-21]). In the area of medical imaging diagnosis, Wang *et al*. [18] used 3D ROC analysis based on three parameters, TPR, FPR and threshold parameter. The authors avoided making hard decisions between 0 and 1 by adding a single threshold parameter resulting from soft decisions. Their experiments were conducted on real breast MR images by controlling normalized detector probability with the optimal threshold. This approach comprehensively evaluated the extensions from 2D ROC to 3D ROC curves for medical diagnosis systems.

In recognition systems, imbalanced datasets with unknown costs, classes can be defined as "ill", meaning that they are poorly sampled, or "well" [21]. If the major concern is to identify those ill-defined classes, such as in the software defect prediction, where the objective is to detect few defective modules, both classification and rejection thresholds should be well studied. Landgrebe *et al*. [21] proposed 3D ROC analysis to evaluate the rejection performance in terms of FPR of rejections, TPR and FPR of positive classification. Their evaluation criteria was derived from traditional ROC analysis, i.e. area under curve measure, and converted to Volume Under Curve (VUC). This approach showed that VUC provided to be a powerful and sensitive performance evaluation measure, when we are interested in comparing both classification and rejection capabilities of different classifiers.

3D ROC analysis could also be used to interpret the relation between TPR, FPR and mis-verifications for example in determining the accuracy, for speaker verification applications [19]. In the biometrics identification systems, modified versions of 3D ROC curves showed more comprehensive information for system accuracy and performance in terms of FPR and false rejection rates, FRR [20].

Overall, the literature suggests that 3D ROC analysis has gained increasing attention in machine learning community. In the present study, we employed 3D ROC approach to deal with the typical error-reject tradeoff problem on biomedical datasets.

The rest of this paper is organized as follows: the size and class distribution of the datasets used in this study are provided in Section 2. In Section 3, we describe the important concepts in the methodology such as 2D ROC curves (3.1) and 3D ROC analysis of rejection threshold optimization (3.2). We present our experimental design and results in Section 4 and we conclude with a discussion and future work for further studies in Section 5.

2. Datasets

Throughout this study, eleven biomedical datasets have been used to evaluate the tradeoff between classification and rejection rates. Datasets are from life sciences field, which are downloaded from UCI data repository [22]. The first public dataset is Heart Disease consisting of four databases concerning heart disease diagnosis. In this dataset, there exists a total of 14 numeric and categorical attributes such as patient age, sex, pain characteristics and patient

Table 1. Characteristics of the datasets used in this study

Dataset	Attributes	Positives	PR (%)	Negatives	NR (%)	Total
Heart Disease	14	138	45.5	165	54.5	303
Pima Indian Diabetes	8	268	34.9	500	65.1	768
Arrhythmia	279	207	45.8	245	54.2	452
Cardiotocography	21	176	9.6	1655	90.4	1831
Immunotherapy	7	19	21.1	71	78.9	90
Breast Cancer Wisconsin	9	239	34.2	460	65.8	699
Heart Failure	12	96	32.1	203	67.9	299
Mammographic Mass	4	403	41.9	558	58.1	961
Diabetic Retinopathy	19	611	53.1	540	46.9	1151
SPECTF Heart	44	55	20.6	212	79.4	267
Mice Protein Expression	77	570	52.8	510	47.2	1080

*PR: Positive Rate, NR: Negative Rate

records, and a class label indicating positive as the presence of heart disease and negative as the healthy status. Our proposed methodology has been demonstrated using the Heart Disease dataset, and efficiency of rejection threshold optimization has been further validated on additional ten datasets. Size and class distributions of the datasets are represented in Table 1.

3. Methodology

Studies show that adequately trained machine learning algorithms are not significantly different than each other [23]. In this study, we used Naive Bayes algorithm because the outputs of the classifier are posterior probabilities which can be directly used in our approach. Naive Bayes classifier assumes that each attribute is independent, normally distributed and equally important. It is derived from the Bayes theorem such that posterior probability of an instance x belonging to class C_i is proportional to prior probability of the class, C_i, and the likelihood $p(x|C_i)$, normalized by the evidence, $p(x)$.

$$p(C_i|x) = \frac{p(x|C_i)*p(C_i)}{p(x)} \quad (2)$$

3.1. 2D ROC Analysis

In the machine learning community, after realization of the weakness of simple error rate as a performance measure, the use of ROC curves has gained an increasing attention [6]. In this study, we use ROC curve analysis to evaluate the performance of binary Naive Bayes classifier, where each instance x is mapped to one of the positive and negative classes labelled as p and n respectively. Given a classifier and an instance, the prediction outcomes depending on actual class labels of instances can be represented as a 2x2 confusion matrix as shown in Table 2.

Table 2: Confusion Matrix

Actual Case	***Predicted***	
	Positive	***Negative***
Positive	TP	FN
Negative	FP	TN

Common classifier performance metrics such as TPR and FPR are derived from the confusion matrix above.

- ***TP rate (TPR)*** is a measure of accuracy for correct prediction of positive instances and is equal to the ratio of number of true positives (TP) over the sum of true positives and false negatives (FN). TPR corresponds to sensitivity in medical diagnosis.

$$TPR = (TP)/(TP + FN) \quad (3)$$

- ***FP Rate (FPR)*** represents the number of positive predictions when the actual is negative (FP), over the sum of true negative (TN) and false positives (FP). FPR corresponds to (*1 - specificity*) in medical domain.

$$FPR = (FP)/(TN + FP) \quad (4)$$

In classification with Naive Bayes algorithm, TPR and FPR have been calculated for a single decision threshold (default: 0.5) that maps to a single point on the ROC curve. Varying TPR and FPR can be calculated by changing the decision threshold in the range of [0:0.1:1]. The resulting set of (TPR, FPR) pairs are plotted in a 2D space that represents all possible classification outcomes as a results of threshold variation.

An example ROC curve has been shown in Figure 1. The lower left point (0,0) in the curve represents assigning all instances to the negative class. Since there are no positive predictions, both TPR and FPR becomes 0. Conversely, upper right corner (1,1) indicates positive prediction for all instances. The upper left point (0,1), where TPR is 1 and FPR is 0 represents perfect classification. Any point on the ROC curve closer to the upper left corner would be closer to the perfect classification. Therefore, the decision threshold that provides the nearest point to (0,1) is accepted as the optimum decision threshold (*t_opt*). The Euclidean distance between (TPR, FPR) values and perfect classification (0,1) can be calculated using the below formula:

$$distance = \sqrt{(1 - TPR)^2 + (0 - FPR)^2} \quad (5)$$

Figure 1 demonstrates the effect of threshold optimization on the variation of TPR and FPR. After classification with the default decision threshold 0.5, the Euclidean distance to the ideal point (0,1) is 0.249, whereas it is 0.238 with $t_{opt} = 0.4$. In this study, we assume equal misclassification costs. However, we can define the desired trade-off between TPR and FPR depending on the requirements of the specific application domains. Then, the distance in 2D ROC would be weighted.

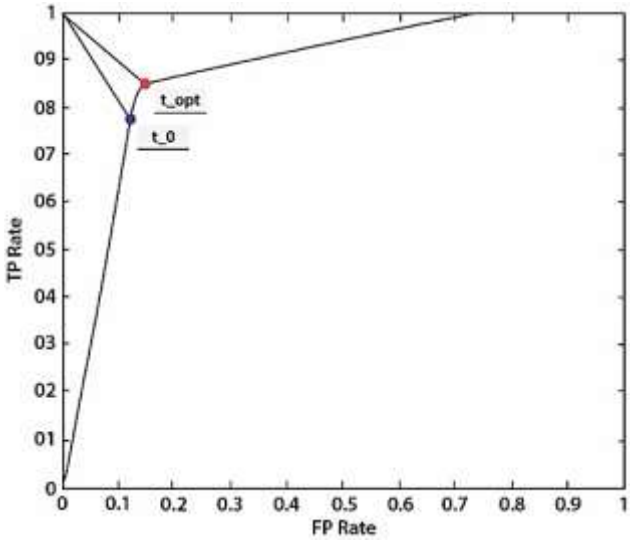

Fig. 1. Example 2D ROC curve for Heart Disease data

3.2. 3D ROC Analysis of Rejection Threshold Optimization

When reject option is introduced, a binary classification task can be assumed to be transformed into a three-class problem, since the rejected samples can be thought to be assigned to a new pseudo-class *r* [2]. In this case, a portion of the instances will be rejected and the FPR and TPR will be evaluated again on the remaining classified samples. In this study, we extend the 2D ROC analysis approach to 3D analysis of TPR, FPR and RR for classifiers with reject option. The aim of classification with reject option is to maximize TPR, while minimizing FPR and RR. This trade-off can be illustrated in a similar manner to 2D ROC curves.

Considering rejection, the initial problem is to decide which samples to reject. For Naive Bayes classifier, the outputs are posterior class probabilities. We want to reject the samples whose posterior probabilities are below a confidence level such that rejecting these samples would avoid potential misclassifications. In other words, when posterior probabilities of two classes are close to the decision threshold, it is likely that classifying those instances are error prone.

After threshold optimization on 2D ROC curve, a rejection interval can be defined in the neighbourhood of t_{opt}, which can also be called as critical interval due to the uncertainty of the classifier's decision. The final decision rule, together with optimized threshold and reject option, is given in (6) where t_r is the rejection interval on the right hand side of the t_{opt}, and t_l is the rejection interval on the left hand side. This rejection principle represents an asymmetric behavior because of the unequal intervals on two sides of the t_{opt}. A graphical view that shows the rejection intervals for each decision threshold between 0 and 1 is shown in Figure 2.

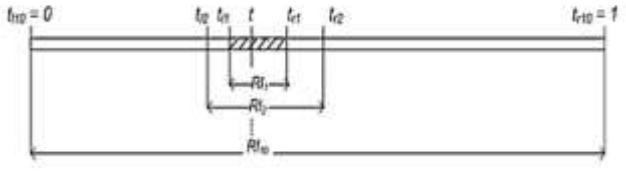

Fig. 2. Demonstration of the search space for rejection intervals

$$\begin{cases} p, if\ p(C_p|x) \geq t_{opt} + t_r \\ n, if\ p(C_p|x) \leq t_{opt} - t_l \\ r, if\ t_{opt} - t_l < p(C_p|x) < t_{opt} + t_r \end{cases} \quad (6)$$

By applying the above decision rule with reject option, a set of (TPR, FPR, RR) tuples have been obtained by varying the reject region around the t_{opt}. The effect of rejection threshold adjustment is illustrated on a sample dataset in Figure 3.

As the final step of our methodology, the objective is to find the optimum rejection interval. This issue has been performed as an analogy to threshold optimization in 2D ROC curves. The ideal case is the point (0,1,0) on the 3D ROC surface corresponding to FPR, TPR and RR, respectively. This point minimizes FPR and RR and maximizes TPR, which is also the desired case in the domains we have been working. In order to find the distance between (FPR, TPR, RR) and the point (0,1,0), Euclidean distance formula is given in (7). The computations for TPR and FPR are shown in (3) and (4).

$$distance(TPR, FPR, RR) = \sqrt{(0 - FPR)^2 + (1 - TPR)^2 + (0 - RR)^2} \quad (7)$$

- ***Rejection Rate (RR)*** Rejection rate (RR) represents the number of instances whose threshold is within the rejection interval. Those instances that are not classified by our model are not represented in the confusion matrix. So this measure is calculated using (8), where *N* is the total number of instances, *NR* is the number of rejected instances.

$$RR = NR/N \quad (8)$$

Finally, (9) can be used in order to find the optimum rejection interval. Pseudocode of our methodology is illustrated in 1 for 2D ROC decision threshold optimization and 2 for 3D ROC rejection threshold optimization.

$$[t_l, t_r]^* = argmin_{[t_l,t_r]} dist(TPR, FPR, RR) \quad (9)$$

Algorithm 1 Pseudocode for 2D ROC Decision Optimization

1: Dataset = (*Heart Disease, Pima Indian Diabetes, Arrhythmia, Cardiotocography, Immunotherapy, Breast Cancer Wisconsin, Heart Failure, Mammographic Mass, Diabetic Retinopathy, SPECTF Heart, Mice Protein Expression)*
2: threshold = [0: 0.1: 1]
3: **for all** *D* in Dataset **do**
4: *Output = load "wekaOutput:txt"*
5: **for all** *t* in threshold **do**
6: **for all** *x* in *Output* **do**
7: **if** $p(C_p|x) >= t$ **then**
8: *Predict* ← C_p
9: **else** {$p(C_p|x)$ is smaller than *t*}
10: *Predict* ← C_n
11: **end if**
12: **end for**
13: calculate *TPR*, *FPR* using Equation 3, 4
14: *DistArray* ← calculate the distance using Equation 5
15: **end for**
16: t_{opt} ← threshold giving the minimum distance in *DistArray*
17: **end for**

Algorithm 2 Pseudocode for 3D ROC Rejection Optimization

1: Dataset = (*Heart Disease, Pima Indian Diabetes, Arrhythmia, Cardiotocography, Immunotherapy, Breast Cancer Wisconsin, Heart Failure, Mammographic Mass, Diabetic Retinopathy, SPECTF Heart, Mice Protein Expression)*
2: threshold = [0: 0.1: 1]
3: **for all** *D* in Dataset **do**
4: *Output = load*"*wekaOutput:txt*"
5: **for all** *t* in threshold **do**
6: **for** *interval* = 0:1 to 1 **do**
7: $[t_l, t_r]$ = [*interval* * *t; interval* * (1 - *t*)]
8: *RejectRegion* ← $[t_l, t_r]$
9: **for all** *x* in *Output* **do**
10: **if** $p(C_p|x) >= t + t_r$ **then**
11: *PredictWithReject* ← C_p;
12: **else if** $p(C_p|x) <= t - t_l$ **then**
13: *PredictWithReject* ← C_n;
14: **else** {$p(C_p|x)$ is within the rejection region}
15: Reject *x*
16: **end if**
17: **end for**
18: calculate *TPR,FPR,RR* using Equation 3,4,8
19: *DistArrayWithReject* ← calculate distance using Equation 7
20: **end for**
21: **end for**
22: find the optimal $[t_l, t_r]$ using Equation 9
23: **end for**

4. Experiments and Results

We used Weka machine learning tool [24] to perform Naive Bayes classification. We conducted our experiments using 10-fold cross validation technique in order to overcome sampling bias. For 10-fold cross validation, each dataset is divided into 10 equal subsets, 9 subsets were used for training and 1 subset was used for testing. Repeating these 10 times ensures that each data sample was used for training and testing. This random splitting has been performed using stratification principle in order to ensure that the proportions of positive and negative classes remain the same in both training and test sets as in the original dataset [25].

We implemented post-processing steps for 2D and 3D ROC analysis using Matlab environment. Our experiments consist of three sequential steps: classification with default threshold, classification with optimal threshold and classification with reject option. 3D ROC representation for Heart Disease can be seen in Figure 3. The results of these steps for each dataset can be seen in Table 3. We have derived important conclusions for the classification problems in biomedical datasets. First, 2D ROC analysis, second column in Table 3, shows that optimum decision threshold is different from its default value 0.5 in each dataset. Overall, decision threshold optimization improves the TPR, on the average, from 68.7% to 85.7%, with a cost of an increase in FPR, on the average, from 10.3% to 15.8%. Paired t-tests with $\alpha = 0.05$ were applied on 10-fold cross validation classification results in order to validate the statistical significance of the findings. Statistical tests revealed that classification with reject option, when the reject region is optimized using the proposed 3D ROC based technique, improves TPR without significantly increasing FPR. Reject option did not change the decision threshold optimization results in two datasets, Immunotherapy and Heart Failure. This suggests, decision threshold optimization provides the optimum balance between TPR and FPR in some decision making problems.

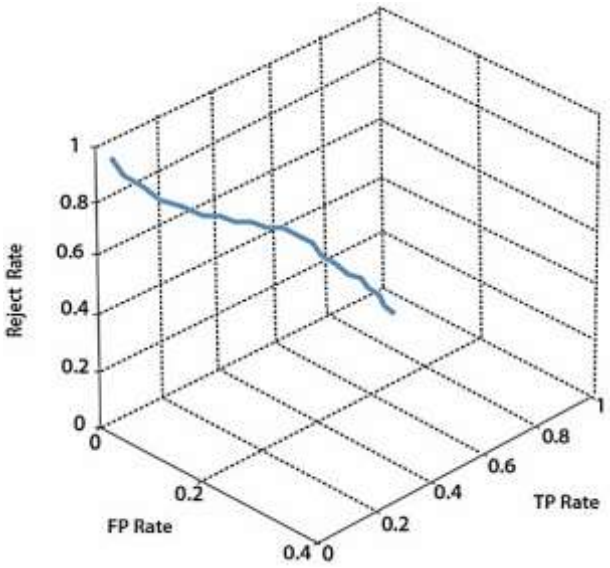

Fig. 3. 3D ROC Analysis for Heart Disease dataset

Table 3. Experimental Results

Dataset	Default threshold - t_0=0.5			2D ROC			3D ROC with rejection			
	AUC	TPR (%)	FPR (%)	t_opt	TPR (%)	FPR (%)	[tl,tr]	TPR (%)	FPR (%)	RR (%)
Heart Disease	0.792	78.9	13.3	0.4	82.6	16.3	[0.24,0.64]	85.7	12.9	9.9
Pima Indian Diabetes	0.731	61.2	15.6	0.3	71.6	24.2	[0.18,0.28]	80.9	27.8	9.6
Arrhythmia	0.747	59.5	10.2	0.6	76.2	16.3	[0.59,0.67]	80	17.8	6.6
Cardiotocography	0.892	80	1.5	0.76	94.3	3.6	[0.76,0.86]	100	3.8	3.6
Immunotherapy	0.667	33.3	0	0.7	100	7.1	[0.68,0.69]	100	7.1	0
Breast Cancer Wisconsin	0.937	89.6	2.2	0.68	93.8	5.6	[0.56,0.89]	97.8	4.7	4.4
Heart Failure	0.809	78.9	17.1	0.53	84.2	17.1	[0.51,0.52]	84.2	17.1	0
Mammographic Mass	0.812	81.3	18.8	0.58	86.3	22.4	[0.66,0.79]	93.3	11.4	15.3
Diabetic Retinopathy	0.702	62.6	22.2	0.57	78.9	31.5	[0.55,0.57]	77.6	28.2	5.2
SPECTF Heart	0.681	45.5	9.3	0.66	81.8	25.6	[0.17,0.3]	86.7	19.4	16.4
Mice Protein Expression	0.909	85.3	3.5	0.56	93.1	4.4	[0.56,0.59]	95	4.5	2.3

*AUC: Area Under the ROC Curve, TPR: True Positive Rate, FPR: False Positive Rate, RR: Rejection Rate

5. Conclusion

In this study, we have presented a novel approach to classification with reject option: rejection threshold optimization based on 3D ROC analysis. We have analysed the trade-off between (TPR, FPR, RR) tuples as an analogy to 2D ROC analysis on (TPR, FPR) pairs. We have projected the error/misclassification rate into TPR and FPR for more appropriate evaluation of the classifier performance in imbalanced datasets. Then, considering rejection option, we have combined error-reject curves and 2D ROC curves as a 3D ROC surface that allows the optimization of a rejection interval. We have demonstrated our method using Naive Bayes on eleven biomedical datasets. Experimental results show that optimization of rejection boundaries using 3D ROC significantly increases TPR. The proposed model is directly applicable to classifiers such as Naive Bayes that produce continuous outputs as an estimate of class posterior probabilities. Other algorithms such as Decision Trees, Support Vector Machines and Artificial Neural Networks produce continuous or discrete outputs that can be converted to class posterior probabilities using post-processing steps. As a future work, we aim to extend the proposed model for other classification techniques.

In this study, we have tackled the problem of finding the optimal rejection interval. As another future direction, we are interested in analysing the rejected samples. There are two approaches to handle rejected samples: multistage approach using the same classifier but increasing information content of data [2], cascading classifiers with different learners [26]. We plan to use either of these methods to re-investigate rejected samples in medical diagnosis domains.

References

[1] C.K. Chow, "On optimum recognition error and reject tradeoff," *IEEE Trans. Information Theory*, vol. 16, no. 1, pp. 41-46, January 1970.

[2] P. Pudil, J. Novovicova, S. Blaha, J. Kittler, "Multistage pattern recognition with reject option." in *Proc. 11th IAPR International Conference on Pattern Recognition Vol.II. Conference B: Pattern Recognition Methodology and Systems*, The Hague, Netherlands, 1992, pp. 92-95

[3] P. Vcelak, M. Kryl, M. Kratochvil, J. Kleckova, "Identification and classification of DICOM files with burned-in text content," *International Journal of Medical Informatics*, vol. 126, pp. 128-137, June 2019.

[4] B. Hanczar, "Performance visualization spaces for classification with rejection option", *Pattern Recognition*, vol. 96, Dec. 2019.

[5] N. Gorski, "Optimizing error-reject trade off in recognition systems," in *Proc. Fourth International Conference on Document Analysis and Recognition*, Ulm, Germany, 1997, pp. 1092-1096 vol.2

[6] T. Fawcett, "An introduction to roc analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861-874, June 2006.

[7] T. Menzies, J. Greenwald and A. Frank, "Data mining static code attributes to learn defect predictors," in *IEEE Transactions on Software Engineering*, vol. 33, no. 1, pp. 2-13, Jan. 2007.

[8] A. Tosun, B. Turhan, A. Bener, "Ensemble of software defect predictors: a case study," in *Proc. ESEM*, Kaiserslautern, Germany, pp. 318-320, Oct. 2008.

[9] A. Uyar, N. Ciray, A. Bener, M. Bahceci, "3p: Personalized pregnancy prediction in ivf treatment process," in *Proc. First Int. Conf. Electronic Healthcare for the 21st Century*, London, UK, Sept. 2008, pp. 58-65.

[10] L. Hansen, C. Liisberg, P. Salamon, "The error-reject tradeoff", Open Systems and Information Dynamics," vol. 4, no.10, 2000

[11] F. Tortorella, "An optimal reject rule for binary classifiers." in *Proc. of the Joint IAPR International Workshops on Advances in Pattern Recognition*, London, UK, Sept. 2000, pp. 611-620.

[12] M. A. Maloof, "On machine learning, ROC analysis, and statistical tests of significance," *Object recognition supported by user interaction for service robots*, Quebec City, Quebec, Canada, 2002, pp. 204-207 vol.2.

[13] M.R. Hassan, M.M. Hossain, J. Bailey, K. Ramamohanarao, "Improving k-nearest neighbour classification with distance functions based on receiver operating characteristics," in *Proc. of the ECML PKDD*, vol. 5211, Springer, Berlin, Heidelberg, 99. 489-504.

[14] C.M. Santos-Pereira, A.M., Pires, "On optimal reject rules and roc curves," *Pattern Recognition Letters*, vol. 26, no. 7, pp. 943-952, May 2005.

[15] Z. Ceylan, "Diagnosis of Breast Cancer Using Improved Machine Learning Algorithms Based on Bayesian Optimization", *IJISAE*, vol. 8, no. 3, pp. 121-130, Sep. 2020.

[16] A. Tosun, A. Bener, "Reducing false alarms in software defect prediction by decision threshold optimization," *2009 3rd International Symposium on Empirical Software Engineering and Measurement*, Lake Buena Vista, FL, 2009, pp. 477-480.

[17] A. Uyar, A. Bener, H. N. Ciray, "Predictive Modeling of Implantation Outcome in an In Vitro Fertilization Setting: An Application of Machine Learning Methods," Med. Decis. Making, vol. 35, no. 6, Aug. 2015, pp. 714-725.

[18] Su Wang *et al.*, "3D ROC analysis for medical imaging diagnosis," *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, Shanghai, 2005, pp. 7545-7548.

[19] F. M. Ham, R. Acharyya and Young-Chan Lee, "Speaker verification using 3-D ROC curves for increasing imposter rejections," *The 2006 IEEE International Joint Conference on Neural Network Proceedings*, Vancouver, BC, 2006, pp. 2561-2565.

[20] Y. Du, C. Chang, "3d combinational curves for accuracy and performance analysis of positive biometrics identification," *Optics and Lasers in Engineering*, vol. 46, no. 6, pp. 477-490, June 2008.

[21] T. Landgrebe, D.M.J. Tax, P. Paclík, R.P.W. Duin, "The interaction between classification and reject performance for distance-based reject-option classifiers." Pattern Recognition Letters, vol. 27, no. 8, pp. 908-917, June 2006, 10.1016/j.patrec.2005.10.015

[22] A. Asuncion, D. Newman, UCI machine learning repository, 2007, Available: http://www.ics.uci.edu/~mlearn/ MLRepository.html

[23] E. Seğmen , A. Uyar, "Performance analysis of classification models for medical diagnostic decision support systems," *2013 21st Signal Processing and Communications Applications Conference (SIU)*, Haspolat, 2013, pp. 1-4.

[24] I. H. Witten, E. Frank, "Data Mining: Practical machine learning tools and techniques," 2nd ed., San Francisco, USA: Morgan Kaufmann, 2005,

[25] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in Int. Joint Conf. on AI, Quebec, Canada, Aug. 1995, pp. 1137-1145.

[26] C. Kaynak, E. Alpaydin, "Multistage cascading of multiple classifiers: One man's noise is another man's data," presented at the *17th International Conference on Machine Learning*, Stanford, CA, USA, June 2000.