An Empirical Study on the Effect of Imbalanced Data on Bleeding Detection in Endoscopic Video

Farah Deeba, Shahed K. Mohammed, Francis M. Bui, Khan A. Wahid

Abstract— **In biomedical applications including classification of endoscopic videos, class imbalance is a common problem arising from the significant difference between the prior probabilities of different classes. In this paper, we investigate the performance of different classifiers for varying training data distribution in case of bleeding detection problem through three experiments. In the first experiment, we analyze the classifier performance for different class distribution with a fixed sized training dataset. The experiment provides the indication of the required class distribution for optimum classification performance. In the second and third experiments, we investigate the effect of both training data size and class distribution on the classification performance. From our experiments, we found that a larger dataset with moderate class imbalance yields better classification performance compared to a small dataset with balanced distribution. Ensemble classifiers are more robust to the variation in training dataset compared to single classifier.**

I. Introduction

With its introduction in 2000 [1], capsule endoscopy (CE) has played a significant role in the diagnosis and management of gastrointestinal (GI) tract diseases. As a patient friendly alternative to traditional wired endoscopy, CE has eased the endoscopic process by replacing the insertion of tubular structures containing fiberoptic bundles and multiple channels through the body cavity opening with a swallowable capsule. Typically at a frame rate of 2-6 fps, a commercial capsule transmits 14,000-72,000 images of GI tract during its battery life span of 8-12 hours [2]. Examining these frames poses significant burden on the clinicians, requiring the investment of time and undivided attention.

In the last 15 years after the inception of CE, a large number of research works have been published proposing computer-aided screening and decision systems based on CE images [3]. The majority of these works are based on supervised learning, since supervised approaches have been reported to achieve higher accuracy compared to an unsupervised one [3]. The performance of the supervised learning depends on the quality of training data. For medical images, class imbalance is a common phenomenon where a positive instance can typically only be seen after observing thousands of negative instances [4].

*Research supported by NSERC.

Farah Deeba is with ECE department at the University of Saskatchewan, Canada (phone: 1-306-8505487; e-mail: farah.deeba@ usask.ca).

Shahed K. Mohammed is with ECE department at the University of Saskatchewan, Canada (e-mail: shahed.mohammed@ usask.ca).

Francis M. Bui is with ECE department at the University of Saskatchewan, Canada (e-mail: francis.bui@ usask.ca).

Khan A. Wahid is with ECE department at the University of Saskatchewan, Canada (e-mail: khan.wahid@ usask.ca).

To deal with class imbalance problem, different approaches have been proposed in the literature including data sampling and cost sensitive learning framework [5], [6], [7]. Data sampling methods involve the modification of an imbalanced data set by some mechanisms to provide a balanced distribution, which is typically performed by undersampling the majority class or oversampling the minority data following specific rules. On the other hand, the cost-sensitive learning methods use different cost matrices to describe the misclassification cost for different classes. Although these techniques have exhibited significant improvement in case of class imbalance, it would be interesting to investigate whether there is any advantage of using a certain classifier over others, in absence of any of these techniques to address the class imbalance problem.

Besides, in this paper, we have tried to address the important question of how to select class distribution of training set to get optimum classification performance in case of limited training data size. It is believed that for a fixed sized training set, balanced dataset would give more accurate classification results [8]. However, for an available dataset with class imbalance, we can select a smaller training set to achieve balanced class distribution, or a larger training set keeping the original class distribution constant. There is no empirical study in the field of endoscopic image classification to show the relative importance of these two factors, namely training data size and class distribution. In this paper, we have experimented with five different classifiers with different training data size and class distribution to address the above mentioned issues. We have selected the extensively visited problem of bleeding detection in CE images for our experiment. The classifiers we have selected for our experiments are: SVM, decision tree, kNN, and neural network. In a previous work, it was established that ensemble classifiers are more robust against imbalanced training dataset compared to single classifier for bleeding detection [9]. Therefore, RUSBoost, an ensemble classifier is included to provide a reference for the performance measure. The detailed empirical study should be helpful in providing a better understanding of the role of specific classifier, class distribution and training data size on classification performance in case of class imbalance for bleeding detection in capsule endoscopic images.

The rest of the paper is organized as follows: section II describes the proposed methodology, section III details the experimental results and section IV concludes the study.

978-1-4577-0220-4/16/$31.00 ©2016 IEEE

II. METHODOLOGY

A. Image Feature Extraction and Feature Selection

From the 256x256 capsule endoscopic images, bleeding and non-bleeding regions have been extracted following the steps originally proposed in [10], [11]. From the extracted regions, histogram based first order features (i.e., mean, standard deviation, energy, entropy, and skewness) have been selected from each of RGB and HSV planes, yielding a feature space of dimension 30. Applying ant colony optimization to select the optimum feature subset [12], we get a feature subset consisting of mean of red, mean of green, mean of saturation and standard deviation of red color channel.

B. Classifiers

For comparison of different classifiers, we selected four popularly used classifiers: Support Vector Machine, Neural Network, decision tree, and k-Nearest Neighbor. RUSBoost is also included to serve as a reference classifier.

Support Vector Machine (SVM) has been extensively used in classification problem of endoscopic videos [4],[13]. However, performance of SVM drops drastically in case of training data imbalance, as by design of the underlying algorithm, SVM learns a decision boundary which is skewed towards the minority class to maximize the margin and minimize the classification error [6]. Neural network is another classifier with frequent application for classification of endoscopic images [14], [11] and also shows progressively detrimental effect with increasing class imbalance [15]. Decision tree, on the other hand, partitions the training set successively into smaller subsets, which are then used to form disjoint rules leading to a final hypothesis to minimize classification error [5]. In case of imbalanced data, successive partitioning results in fewer observations of minority class leading to fewer leaves representing minority concepts. Thus the rules based on confidence are biased towards the majority class. k-nearest neighbor (kNN) classifier is an instance-based classifier learning algorithm, which finds the k training set examples closest to the test example based on a certain distance metric (Mahalanobis distance has been used in our experiments) and the label of the test example is determined by the predominance of a class in the neighborhood. Thus, kNN effectively uses the prior class information to estimate class labels, resulting in suboptimal performance for minority class in case of imbalanced dataset [16]. RUSBoost, unlike the above mentioned classifiers, is an ensemble learning method, which combines weak learners and aggregates their predictions to form a new classifier, outperforming each of the weak learners. In this paper, we have used decision tree as the weak learner, the most commonly used weak or base classifier [17]. Ensemble classifiers generally outperform single classifier in case of class imbalance [18]. RUSBoost is specifically designed to handle class imbalance problem [19], and thus is expected to yield better performance compared to single classifiers. In this paper, RUSBoost has been included to provide a reference for classification performance in case of class imbalance.

TABLE I. DESCRIPTION OF DATASET

Dataset	Total	Bleeding	Non-bleeding
Total	8248	836	7448
Test	2071	209	1862
Training for Experiment 1	660	Varied between 5% to 95% of total training set with 5% increment	Varied between 5% to 95% of total training set with 5% increment
Training for Experiment 2	Varied	627 (All bleeding samples available for training)	Varied with a ratio between 8:1 to 1:8 with increment 1
Training for Experiment 3	660×n n=0.1:0.1:1	Varied between 5% to 95% of total training set with 5% increment	Varied between 5% to 95% of total training set with 5% increment

III. EXPERIMENTS

C. Data Acquisition and Experimental Setup

For our experiment, we compiled a dataset containing a total of 8248 images, in which bleeding samples constitute the minority class. These images were taken from resources available at [18, 19] and captured using Pillcam SB, SB2, SB3 (Given Imaging) and EndoCapsule1 (Olympus). We ensure that the dataset represents the wide variation of CE images arising from different technology, patients and clinical condition. We have performed all our experiments using MATLAB 2015a. For each experimental iteration, the test set is formed by randomly selecting 25% of the original dataset (25% of total bleeding and 25% of total non-bleeding examples). The remaining data are allocated for training. For the first experiment, the training set size is kept constant while the class distribution is varied such that bleeding class accounts for 5% to 95% of total training data. To ensure that the training set size is constant and to make the best use of available training data, we set the training set size such that,

$$training\ set\ size = \frac{bleeding\ sample\ available\ for\ training}{0.95}$$

For the second experiment, we keep the size of the minority class, i.e., bleeding class constant. The size of non-bleeding class is varied so that bleeding to nonbleeding example ratio is varied between 1:8 to 8:1. For the third experiment, first experiment is iterated for different training data size while maintaining the similar class distribution as of first experiment. The description of the dataset and the training dataset are given in Table 1. For each classifier, we used 5-fold cross-validation of the training sets to optimize the classifier parameters. All the results are based on 10 iterations performed on each of the training class distribution for each experiment. To identify the optimal range of class distribution for each distribution for all the experiments, we perform paired t-tests to compare the performance of the

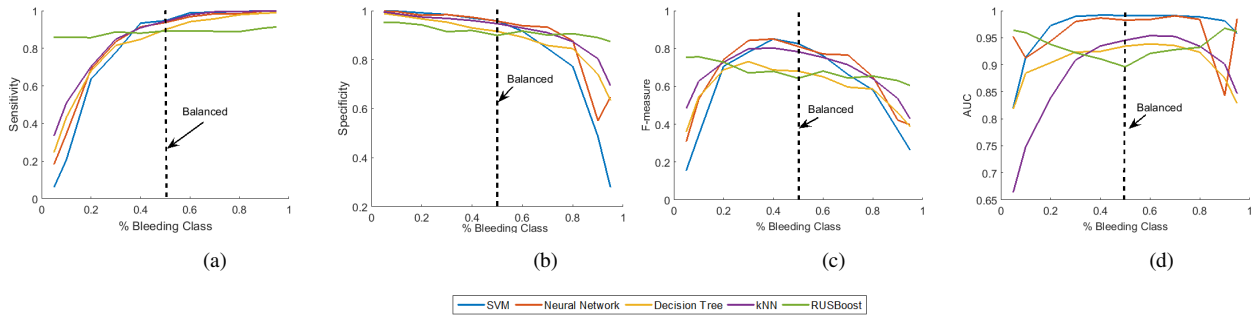

Fig. 1. Variation of classifier performance with variation in class distribution of training data for fixed size training data; (a) Sensitivity; (b) Specificity; (c) F-measure; (d) AUC; Here the x-axis indices correspond to percentage of bleeding examples in the total training set.

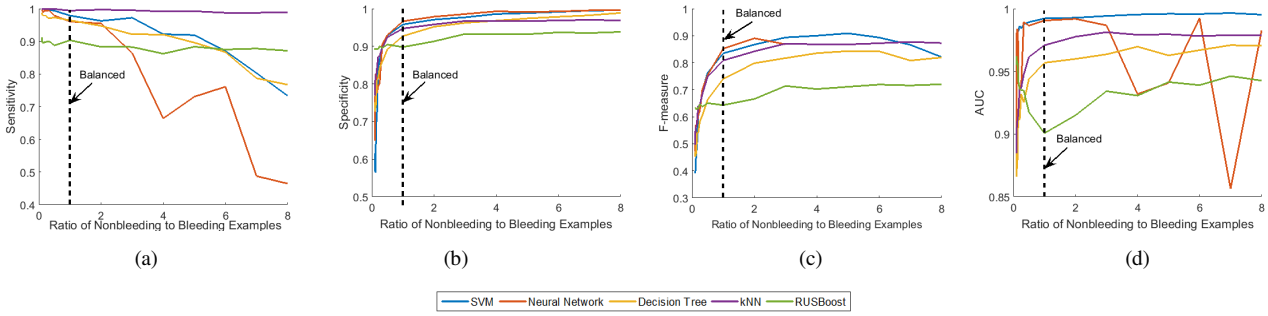

Fig. 2. Variation of classifier performance with variation in class distribution and size of training data for fixed number of bleeding examples; (a) Sensitivity; (b) Specificity; (c) F-measure; (d) AUC; The x-axis indices correspond to nonbleeding to bleeding example ratio in training set.

best average classifier with the performance of other class distributions. A t-test giving p-value<0.05 is considered to be statistically significant.

D. *Performance Metric*

For quantitative performance evaluation, we adopted sensitivity, specificity, and F-measure as follows:

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$F - Measure = \frac{2 \cdot sensitivity \cdot precision}{sensitivity + precision}$$

Here, TP and FP are respectively the correctly and incorrectly labeled bleeding examples, and TN and FN are respectively the correctly labeled and incorrectly labeled non-bleeding examples. Precision is the ratio of correctly labeled bleeding examples to all the examples labeled as bleeding. We omitted accuracy to measure the performance as it largely depends on the class distribution and can be quite deceiving in case of data imbalance. Sensitivity and specificity are not sensitive to changes in data distribution. F-measure, however is sensitive to class distribution, provides more insight into the functionality of a classifier compared to accuracy metric [5]. We also use area under the curve (AUC) obtained from the ROC curve, which is another evaluation criteria useful to assess the performance of a classifier, especially for imbalanced dataset [5].

E. *Experimental Results and Discussions*

In experiment 1, we have kept the training data size the same and varied the class distribution to see the effect on classification performance for different classifiers. As expected, for all the classifiers, sensitivity improves as the bleeding class percentage increases in training set and the specificity corresponds to percentage of non-bleeding examples in the similar fashion. From Fig. 1, we find that RUSBoost is the most robust classifier, which can be justified as it is specifically designed to handle class imbalance. The remaining classifiers can be ranked from best to worst in term of robustness as: kNN, Decision Tree, Neural Network and SVM. However, from careful examination of AUC curve and F-measure curve from Fig. 1, it is evident that SVM and neural network are the two best classifiers in case of balanced class distribution. Therefore, the classifier selection should depend on class distribution of available dataset. On the other hand, in case a classifier is chosen in advance, the class distribution of training set should be selected accordingly.

In experiment 2, we simulate the situation where the best use of minority class is ensured. We want to find the optimum class distribution in that case given that the minority class size is fixed at its highest possible number. This experiment will also give an idea of the effect of training data size on classifiers. From Fig. 2, we can find that though RUSBoost is insensitive to class distribution and class size, it gives sub-optimum classification results compared to other classifiers. Another important point is that the training data size for experiment 2 is larger than experiment 1 due to specific class distribution choice. The

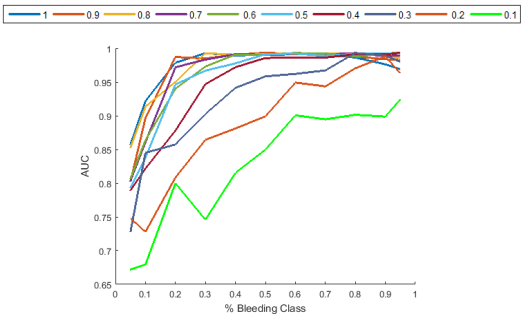

Fig. 3. Effect of class distribution and training-set size on AUC for SVM classifier. X-axis indices correspond to percentage of bleeding examples in the training set. The curves correspond to different training set size 660×n, where n=0.1:0.1:1.

overall performance improvement for experiment 2 compared to experiment 1 is evident from Fig. 1 and Fig. 2. This emphasizes the importance of large training set over certain class distribution. The stable performance of SVM, decision tree and specifically kNN in terms of AUC and F-measure indicates that it is desirable to utilize all the available data rather using downsampling to achieve balanced class distribution for ensuring optimum classification performance.

In experiment 3, we have re-emphasized the generally accepted idea that for a specific class distribution, a large dataset would provide better result. Due to space limitation, we only demonstrate a representative graph for SVM in Fig. 3. Here, it is evident that the performance for a certain class distribution improves as the training size increases and for an optimum training data size (n>0.4), balanced training gives the optimum result.

IV. Conclusion

From our comprehensive analysis, we provide a useful insight regarding the choice of classifier, training class distribution and training data size in absence of any measure to handle class imbalance. From our experiments, we have reached the following conclusions: (1) Choice of classifier is important. Ensemble classifier should be chosen in case of extreme class imbalance. Among the single classifiers, kNN is most insensitive to the variation of class distribution. SVM is most suitable in case of balanced distribution. (2) For a training set which is fixed and not very small (n>0.4 in our case), balanced class distribution provides the best classification performance regardless of classifiers. (3) For a specific class distribution of training dataset, increasing the data size will improve the classification performance. (4) If class distribution and data size both vary, there is no certain rule according to which classification performance changes. Empirically, we have shown that by keeping minority class constant and increasing the majority class, better classification performance can be achieved using classifiers, for example, SVM and kNN. Therefore, it can be concluded that downsampling of majority class to achieve balanced class distribution is not an optimum choice to handle class imbalance problem. In most cases, a large dataset is preferable compared to a small dataset with balanced class distribution. This paper should serve as useful guideline to select the most suitable classifier, training data distribution and size to achieve optimum classification performance for imbalanced class distribution, in context of CE image classification.

References

[1] G. Iddan, G. Meron, A. Glukhovsky, and P. Swain, "Wireless capsule endoscopy.," *Nature*, vol. 405, no. May, p. 417, 2000.

[2] A. Koulaouzidis, E. Rondonotti, and A. Karargyris, "Small-bowel capsule endoscopy: A ten-point contemporary review," *World J. Gastroenterol.*, vol. 19, no. 24, pp. 3726–3746, 2013.

[3] D. K. Iakovidis and A. Koulaouzidis, "Software for enhanced video capsule endoscopy: challenges for essential progress," *Nat. Rev. Gastroenterol. Hepatol.*, vol. 12, no. 3, pp. 172–186, 2015.

[4] B. Giritharan, X. Yuan, J. Liu, B. Buckles, J. Oh, and S. J. Tang, "Bleeding detection from capsule endoscopy videos," *30th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, vol. 2008, pp. 4780–4783, 2008.

[5] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, 2009.

[6] R. Akbani, S. Kwek, and N. Japkowicz, "Applying Support Vector Machines to Imbalanced Datasets," pp. 39–50, 2004.

[7] V. Lopez, A. Fernandez, S. Garcia, V. Palade, and F. Herrera, "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics," *Inf. Sci. (Ny).*, vol. 250, pp. 113–141, 2013.

[8] G. M. Weiss and F. J. Provost, "Learning When Training Data are Costly: The Effect of ClassDistribution on Tree Induction.," *J. Artif. Intell. Res.*, vol. 19, pp. 315–354, 2003.

[9] F. Deeba, S. K. Mohammed, F. M. Bui, and K. A. Wahid, "Learning from Imbalanced Data: A Comprehensive Comparison of Classifier Perfromance for Bleeding Detection in Endoscopic Video," in *The 5th International Conference on Informatics, Electronics & Vision (ICIEV 2016)*, 2016.

[10] F. Deeba, F. M. Bui and K. A. Wahid, "Performance Assessment of a Bleeding Detection Algorithm for Endoscopic Video based on Classifier Fusion Method and Exhaustive Feature Selection," *J. Med. Syst. (submitted)*

[11] S. Sainju, F. M. Bui, and K. A. Wahid, "Automated bleeding detection in capsule endoscopy videos using statistical features and region growing," *J. Med. Syst.*, vol. 38. 4, p. 25, Apr. 2014.

[12] S. K. Mohammed, F. Deeba, F. M. Bui, and K. A. Wahid, "Application of Modified Ant Colony Optimization for Computer Aided Bleeding Detection System," in *The 2016 International Joint Conference on Neural Networks (IJCNN 2016)*, 2016.

[13] Y. Yuan and M. Q. H. Meng, "Polyp classification based on Bag of Features and saliency in wireless capsule endoscopy," *Proc. - IEEE Int. Conf. Robot. Autom.*, pp. 3930–3935, 2014.

[14] G. Pan, G. Yan, X. Qiu, and J. Cui, "Bleeding detection in wireless capsule Endoscopy based on probabilistic neural network," *J. Med. Syst.*, vol. 35, no. 6, pp. 1477–1484, 2011.

[15] M. A. Mazurowski, P. A. Habas, J. M. Zurada, J. Y. Lo, J. A. Baker, and G. D. Tourassi, "Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance," *Neural Networks*, vol. 21, no. 2–3, pp. 427–436, 2008.

[16] W. Liu and S. Chawla, "Class confidence weighted kNN algorithms for imbalanced data sets," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 6635 LNAI, no. PART 2, pp. 345–356, 2011.

[17] M. Galar, A. Fern, E. Barrenechea, and H. Bustince, "A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 42, no. 4, pp. 463–484, 2012.

[18] J. Ghosh and Y. Park, "Ensembles of alpha-Trees for Imbalanced Classification Problems," *IEEE Trans. Knowl. Data Eng.*, vol. 99, no. 1, p. 1, 2012.

[19] C. Seiffert, T. Khoshgoftaar, J. Van Hulse, and A. Napolitano, "Rusboost: A hybrid approach to alleviating class imbalance," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 40, no. 1, pp. 185–197, 2010.