Available online at www.sciencedirect.com
SCIENCE DIRECT®
Pattern Recognition Letters 27 (2006) 875–881
Pattern Recognition Letters
www.elsevier.com/locate/patrec

ROC curves and video analysis optimization in intestinal capsule endoscopy

Fernando Vilariño [a,*], Ludmila I. Kuncheva [b], Petia Radeva [a]

[a] Computer Vision Centre, Universitat Autonoma de Barcelona, 08193 Bellaterra, Barcelona, Spain
[b] School of Informatics, University of Wales, Bangor LL57 1UT, UK

Available online 19 January 2006

Abstract

Wireless capsule endoscopy involves inspection of hours of video material by a highly qualified professional. Time episodes corresponding to intestinal contractions, which are of interest to the physician constitute about 1% of the video. The problem is to label automatically time episodes containing contractions so that only a fraction of the video needs inspection. As the classes of contraction and non-contraction images in the video are largely imbalanced, ROC curves are used to optimize the trade-off between false positive and false negative rates. Classifier ensemble methods and simple classifiers were examined. Our results reinforce the claims from recent literature that classifier ensemble methods specifically designed for imbalanced problems have substantial advantages over simple classifiers and standard classifier ensembles. By using ROC curves with the bagging ensemble method the inspection time can be drastically reduced at the expense of a small fraction of missed contractions.

Keywords: ROC curves; Classification; Classifiers ensemble; Detection of intestinal contractions; Imbalanced classes; Wireless capsule endoscopy

1. Introduction

The number of applications of machine learning to clinical problem solving is growing. New multimedia technologies, including real-time video, present a challenge today because of the considerable size of the databases that they generate. Some of the main difficulties root in the huge amount of data to be analyzed and the difference between the prevalences of the classes of interest within the dataset. One paradigmatic example of this situation is found in the clinical problem addressed by this paper: the detection of *intestinal contractions* in video images.

Both the number of intestinal contractions, and their distribution along the intestinal tract, characterize patterns of low bowel motility that are indicative of presence of different malfunctions (e.g., myopathy, neuropathy, obstruction, etc.). Wireless capsule video endoscopy (WCVE) (Schulmann et al., 2005; Brodsky, 2003; Eliakim, 2004) is a recent technology in which a pill with an attached camera is swallowed by the patient. The camera travels along the intestinal tract and emits a radio signal recorded as a video (Hansen, 2002). Currently, the medical expert looks for contractions by visual inspection, labelling by hand video frames identified as contractions, for further reference, a process that may require more than 1 h. Thus our efforts were focused on automatic detection of contraction frames in the video.

The prevalence of contraction frames in a video is very small; between 1:50 and 1:100. This implies an *imbalanced* or a cost-sensitive problem. In such problems, even a small error rate results in an unacceptably large number of false positive classifications. We propose to use ROC curves to evaluate several classifier models, including classifier ensembles. The aim is to help the expert by identifying locations in the video which contain contractions with high probability thereby greatly reducing the inspection time.

The rest of the paper is organized as follows. Section 2 explains the methodology used: data preparation, classifier

* Corresponding author. Tel.: +34 9358 12301; fax: +34 9358 11670.
E-mail address: fernandov@cvc.uab.es (F. Vilariño).

0167-8655/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2005.10.011

ensembles design, and ROC curve analysis. In Section 3 we present and discuss the results of our experiments. Finally, Section 4 concludes the work and suggests further research directions.

2. Methodology

2.1. Feature extraction from endoscopic sequences

Different medical imaging modalities have been used for intestinal motility analysis, e.g., non-invasive examinations such as plain abdominal X-ray, magnetic resonance imaging (MRI), photon emission computed tomography (PET) and functional MRI (f-MRI) and invasive examinations such as manometry, electromyography and intubation (Hansen, 2002). All are currently in use in medical practice benefiting from simultaneous analysis of different modalities.

The video data of our study is provided by wireless capsule video endoscopy (WCVE) (Adler and Gostout, 2003). This is a recent technology in which a pill with an attached camera is swallowed by the patient. The camera travels along the intestinal tract and emits a radio signal recorded by an external device. The result is a video of approximately 20,000 frames of clinical interest, accounting for about a 3-h real life time span. This examination method has been used successfully in several clinical applications and studies (Schulmann et al., 2005; Brodsky, 2003; Eliakim, 2004).

Fig. 1 shows two sequences of frames, where sequences in (a) represent contractions while sequences in (b) represent non-contractions. As it can be seen, the appearance of a contraction in such a sequence of frames does not have a clear definition. The clinical interest underpinning this study is in the automatic detection of one specific type of intestinal contraction, which corresponds to the top sequence in Fig. 1(a). This contraction type is of special interest to the physician in fasting scenarios. In a video sequence of about nine frames, the contraction is represented as the lumen progressively closing and reopening. In order to describe this paradigm of contraction, we extracted 34 features using basic image descriptors: mean intensity of each frame, $x_1,\ldots,x_9$; hole size of each frame, $x_{10},\ldots,x_{18}$; global contrast of each frame, $x_{19},\ldots,x_{27}$.

The three descriptors measured along the nine frames can be regarded as time sequences. Each sequence is normalized by taking out the mean and dividing by the standard deviation so that we look at the time pattern only.

The seven remaining features are: x_{28} is the correlation between sequences $x_1,\ldots,x_9$ and $x_{10},\ldots,x_{18}$; x_{29} is the correlation between sequences $x_1,\ldots,x_9$ and $x_{19},\ldots,x_{27}$; and x_{30} is the correlation between sequences $x_{10},\ldots,x_{18}$ and $x_{19},\ldots,x_{27}$. Features x_{31}, x_{32}, x_{33} are the correlations between sequences $x_1,\ldots,x_9$, $x_{10},\ldots,x_{18}$ and $x_{19},\ldots,x_{27}$ on the one hand and the corresponding sequences averaged across the objects for the class "contractions". Feature x_{34} is the variance of intensity averaged across the nine frames. This value is then normalized by taking out the mean

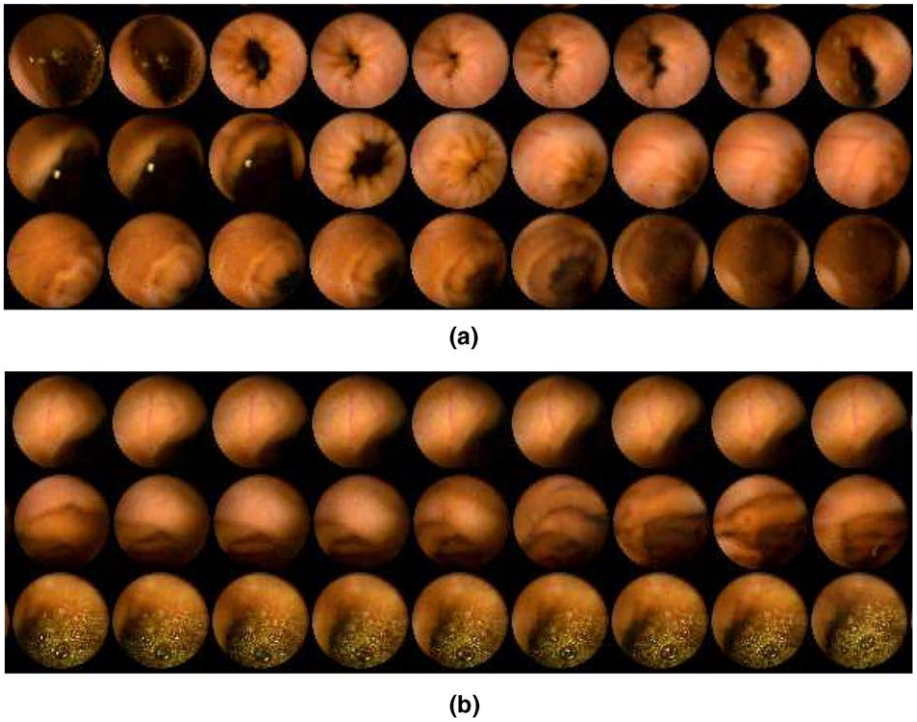

Fig. 1. Video frames obtained from wireless endoscopy: typical patterns of contractions and non-contractions. (a) Three contraction sequences of nine frames and (b) three non-contraction sequences of nine frames.

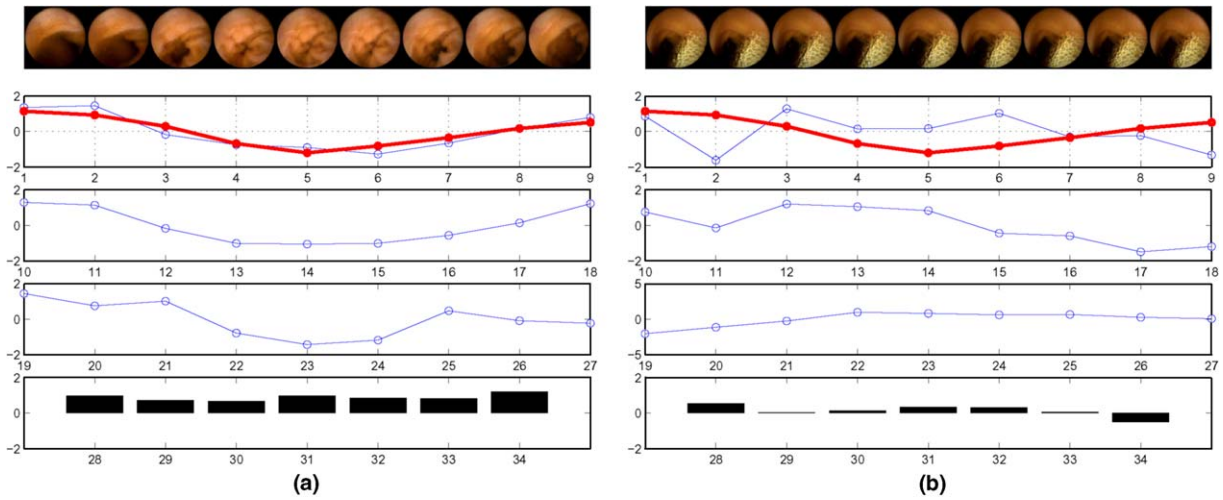

Fig. 2. Feature patterns for (a) contractions and (b) non-contractions. Each feature is identified by its number in the x-axis. The thick line corresponds to the average value of the first descriptor for class contraction.

across the whole video and dividing by the standard deviation. Features $x_1,\ldots,x_{34}$ are plotted in Fig. 2 for examples of contractions and non-contractions. The sequences are joined by lines so as to see the shape patterns. The sequence $x_1,\ldots,x_9$ averaged across class contraction is overlaid in both subplots (the thick line).

In order to mitigate the imbalanced property of the problem, a pre-processing stage was applied with a simple feature selection procedure. For each of the 34 features, we looked for a threshold that kept 99% of contractions (*true positives*). Then we selected the feature which preserved 99% of class contractions within the minimum number of frames. We found that among all frames with $x_5 \leqslant -0.4$, we had 99% of class contraction. The set of all such frames constituted about 25% (5000) of the video. This pre-processing stage implies that all frames which have $x_5 \geqslant -0.4$ will be labeled as non-contractions and the remaining frames will be run through our classifier for labeling.

We carried out experiments without the pre-processing stage. The resultant classifier was able to recognize the same number of true contractions but had to process an unnecessarily large amount of frames. Thus a large amount of false positives was generated. Consider the following example. In a video of 20,000 frames, there will be about 30 contractions. A threshold of 99% will most likely put all 30 in the reduced set, so the large number of frames left out after the pre-processing will be free of contractions anyway.

2.2. Single classifiers and ensembles

The prevalence of contraction frames in a video is very small: there are typically 30–50 contractions in a video sequence of 5000 frames (a ratio less than 1:100). This low prevalence of contractions implies an *imbalanced problem* or a *cost-sensitive problem*. There are several approaches to solving imbalanced problems, such as stratified sampling (Chawla, 2003) and cost manipulation (Domingos, 1999; Maloof, 2003; Ling and Li, 1998). Decision trees (Zadrozny and Elkan, 2001) and classifier ensembles (Tan et al., 2003) have been adapted to imbalanced problems too. Semi-supervised techniques have also been proposed (See-Kiong et al., 2004), where the test samples with the highest probability for the minority class are added to the training set.

Stratified sampling was adopted for our imbalanced problem. We used eight individual classifiers and two classifier ensemble methods. The individual classifiers were linear discriminant classifier (LDC), quadratic discriminant classifier (QDC), logistic classifier (LOGLC), nearest neighbor (k-NN) with $k \in \{1,5,10\}$, decision trees (DT), and Parzen classifier (Parzen) (Duda et al., 2001). The two ensemble methods were: heterogeneous ensembles and bagging. The heterogeneous ensembles were built by taking a set of single classifiers of different types and aggregating their outputs. As we applied eight classifiers, there are $2^8 - 1$ (empty set) $- 8$ (single classifiers) $= 247$ possible heterogeneous ensembles. Bagging produces a classifier ensemble whereby each classifier is trained on a bootstrap sample. We constructed bagging ensembles of 25 decision trees. In this study we used the average of the classifier outputs to be the ensemble output for both ensemble methods. This was done because we need a continuous-valued output as the ensemble decision. Our choice of bagging over AdaBoost (used for imbalanced problems in (Viola and Jones, 2001)) was based on the findings in the recent literature that bagging is the better of the two models for datasets with substantial amount of noise (Bauer and Kohavi, 1999).

2.3. ROC curves

ROC curve (receiving operating characteristic) analysis has been widely used as a method for medical decisions making. We assume that one of the classes is the class of

interest and the objects labeled in this class will be called "positive". We also assume that each classifier gives a continuous-valued output which is cut at a certain threshold. All objects for which the classifier output exceeds the threshold are labeled as positive and the remaining objects are labeled as negative. By varying the threshold from the minimum to the maximum value of the classifier output, we construct a ROC curve for this classifier. The curve shows true positive rate (*sensitivity*) versus false positive rate (1-*specificity*). The user is then able to decide upon a compromise between sensitivity and specificity achievable simultaneously by the classifier. This compromise may be based upon prior class probabilities or different misclassification costs (Breiman et al., 1984; Maloof, 2003). ROC curves have been especially useful for imbalanced or cost-sensitive two-class problems (Kubat et al., 1998; Monard and Batista, 2003; Mac Namee et al., 2002).

The area under the ROC curve (AUC) is deemed to be a better measure of classifier performance than accuracy (Bradley, 1997; Rosset, 2004; Ling et al., 2003). However, in an imbalanced problem such as detection of contractions in endoscopy videos we are looking for operation points on the curve which will present the user with the best time-accuracy compromise. The overall performance of the classifier is of secondary importance.

3. Experimental results

Our experiments were built in the following way: The specialist analyzed 10 videos and manually labeled all contractions. A subset of 305 typical examples was then selected to be our class 'contraction' (positive). For the non-contraction class (negative), 3050 examples were randomly chosen from all the videos, taking special care that the selected sequences did not belong to class contraction.

All eight classifiers, the 247 heterogeneous ensembles and the bagging ensemble (25 decision trees) were trained and tested 100 times and the results were averaged. For each run we used the 305 contraction objects and a random bootstrap sample of size 305 from the class 'non-contraction'. This set of 610 objects was split randomly into 80/20 proportion for training and testing, respectively.

For all classifiers we used the Matlab toolbox PRTOOLS developed by Professor R.P. Duin and his group at the Delft University of Technology (Duin et al., 2004). We used the implementation of the single classifiers (LDC, QDC, LOGLC, 1-NN, 5-NN, 10-NN, decision tree and Parzen) and built our own ensembles and bagging routines. The continuous-valued outputs of all classifiers were used (these are available within PRTOOLS). For the calculus of the AUC, we used the trapezoidal rule, approximating the underlying function using linear interpolation.

The ROC curves for all classifiers were calculated on the testing set. The ensemble with the largest area under the curve appeared to be the one using just two classifiers: decision tree and Parzen, AUC = 0.9603. Fig. 3 plots the ROC curve for this ensemble and also the ROC curves for the two component classifiers. The remaining single classifiers were very similar to one another and slightly worse than the Parzen classifier. The hybrid ensemble outperforms all single classifiers and follows the best behavior of its components in the different areas of the ROC curve. Table 1 displays the AUC for the individual classifiers as well as for the best eight hybrid ensembles.

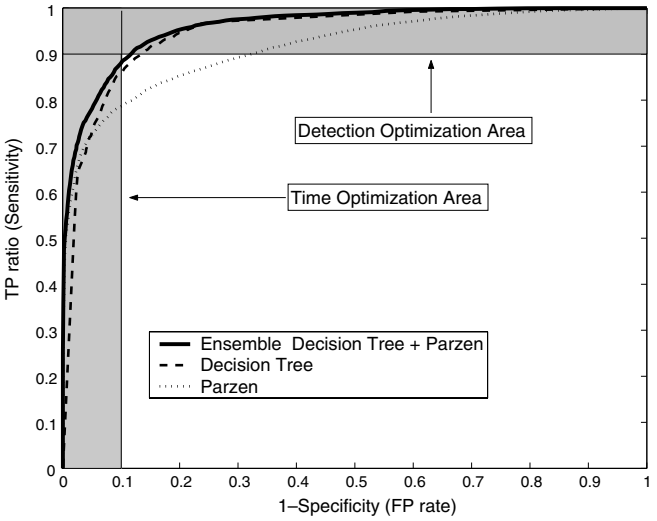

Fig. 3. The ensemble with largest AUC (Parzen + decision tree) outperforms both single classifiers and follows the best behavior of its components in the ROC curve.

The physicians are interested in two different operation points on the ROC curve: *accuracy* of positive detection over a 98%, and *minimization of visualization time*, with a guarantied positive detection over 80%. This brings to the fore two different areas of the ROC curve as marked in Fig. 3. The shaded vertical stripe shows an example of a desirable time-optimization area. Its width denotes the maximum FP rate we are prepared to accept. To see how this is related to time-optimization, consider an example of an (unthresholded) video of 20,000 frames with 30 contractions in it. Assuming that all contractions were correctly labeled, the total number of frames which the system will leave to the expert to inspect is approximately $30 + 0.1 \times 19{,}970 = 2027$. A lower acceptable FP rate, e.g., 0.05, will leave just over 1000 frames for inspection. Thus, in such a heavily imbalanced problem, the inspection time will depend exclusively upon the false positive rate. The interpretation of the shaded horizontal stripe is trivial:

Table 1
AUC for single classifiers and the best eight ensembles

Classifier	AUC	Ensemble	AUC
LDC	0.9040	PARZEN + DT	0.9603
QDC	0.8878	DT + 10-NN	0.9599
LOQLC	0.9033	DT + 1-NN + 10-NN	0.9598
PARZEN	0.9160	DT + 5-NN	0.9591
DT	0.9463	DT + 1-NN + 5-NN	0.9591
1-NN	0.8938	DT + 1-NN	0.9583
5-NN	0.9582	PARZEN + DT + 1-NN	0.9582
10-NN	0.9567	LOQC + DT + 1-NN	0.9567

its height measures the amount of accuracy we are prepared to sacrifice. In the example in Fig. 3, we accept at least 90% True Positive, i.e., the maximum number of missed contractions in the example above should be three or less.

Classifiers that perform best in one area may not be the best in the other area. As the plot shows the decision tree classifier is very good for the accuracy optimization, while the Parzen classifier is more successful for minimization of visualization time. The hybrid ensemble outperforms both which re-confirms the well accepted now claim that ensembles are superior to single classifiers.

Bagging ensemble was constructed from 25 decision trees as the base classifiers, with a resulting AUC = 0.9647. Fig. 4(a) shows the ROC curves for the bagging ensemble and the best heterogeneous ensemble at TP rate of 98% (accuracy zone). The best hybrid ensemble for this zone appeared to be the one consisting of a decision tree and 10-NN, with AUC = 0.9599. Bagging shows superior performance at the point of entering this zone. The heterogeneous ensemble outperforms bagging for FP rate over 50% which renders large inspection time. The same analysis was applied for TP rate of 80% and the result is plotted in Fig. 4(b). In this case, the best heterogeneous ensemble consists of a decision tree, 1-NN and 5-NN, with an AUC = 0.9591. The bagging ensemble only slightly outperforms this ensemble at the desired point. In contrast to the previous case, the heterogeneous ensemble is better for low accuracy rates, e.g., under 70%. Even by small differences, Fig. 4(b) favors the bagging ensemble as the best classifier for both operation points.

Since the main objective of this study is to look for a compromise between inspection time and accuracy, we suggest a variant of the ROC curve. On the x-axis we plot the inspection time required and on the y-axis, the sensitivity of the classification. The inspection time is calculated in the following way: The output of a classifier is a set of frames with suspected contractions (true positive and false positive classifications). Each frame must be visualized as a middle frame in the sequence of nine frames in order to create the dynamic impression. The typical visualization rate is 5 frames per second. That implies 1.8 s for each sequence (i.e., for each output frame), and a bound of 2 s can be used. The total visualization time for one video will be, therefore, the number of output frames multiplied by 2. The x-axis is close to but not a mere rescaling of the FP rate. Consider a thresholded video of 5000 frames with an estimated number of 30 contractions in it. Take an (x, y) point from the standard ROC curve. To calculate the corresponding x' on our ROC variant, we use $x' = (4970x + 30y) \times 2$.

We used the best ROC curve for any point, so different classifiers are responsible for different parts of the curve. This was done in the following way. Suppose that the ROC curves for all classifiers and ensembles are drawn on the same plot. For each value of FP we selected the curve with the maximum TP (the highest curve). In different parts of the ROC curve, different classifiers or ensembles might be the best. A system operating in a real environment should keep the collection of classifiers and ensembles which make up the overall "best" ROC curve. The operation point selected by the physician will translate into running the classifier or the ensemble responsible for this point.

Fig. 5 shows the ROC variant (solid line). Two more ensemble ROC curves are shown for comparison, demonstrating that both ROC curves are inferior to the combined one. Table 2, shows a summary of the results for TP accuracy and visualization time. For the MANUAL method, time is calculated assuming that the physician inspects the video at 5 frames per second. For the rest of the cases, time is calculated as explained above. For 80% sensitivity, only 9 min and 10 s are needed, while 67 min are needed for manual labelling.

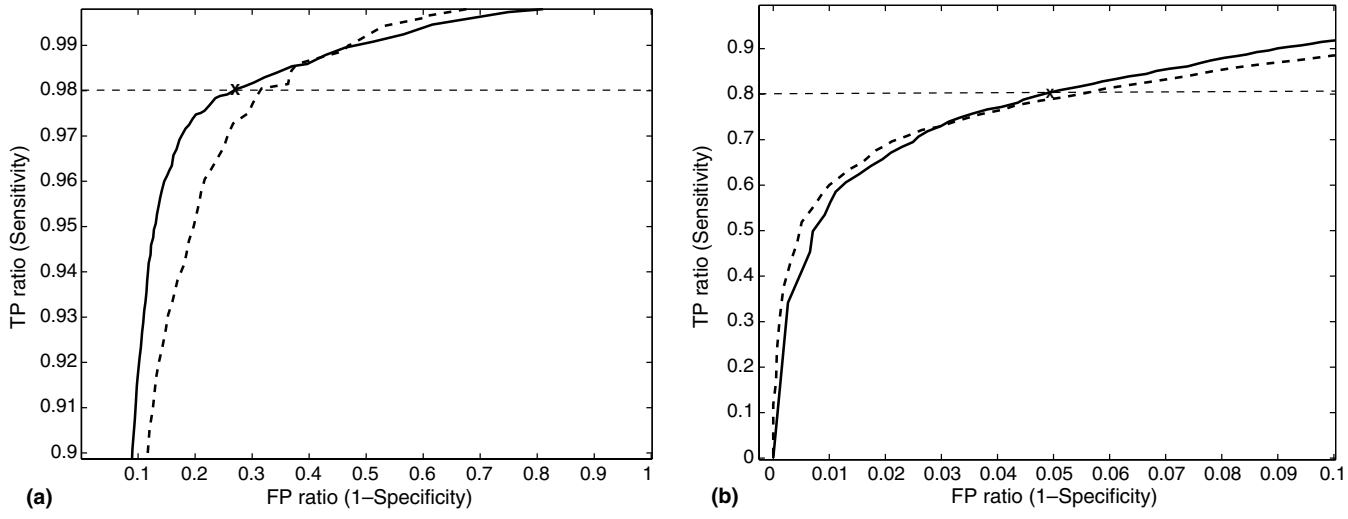

Fig. 4. Bagging (solid) versus best ensemble (dashed) for each zone of interest. (a) Bagging versus hybrid ensemble (decision tree and k-NN-10) in the sensitivity optimization area. (b) Bagging versus hybrid ensemble (decision tree, k-NN-1 and k-N-5) in the time optimization area.

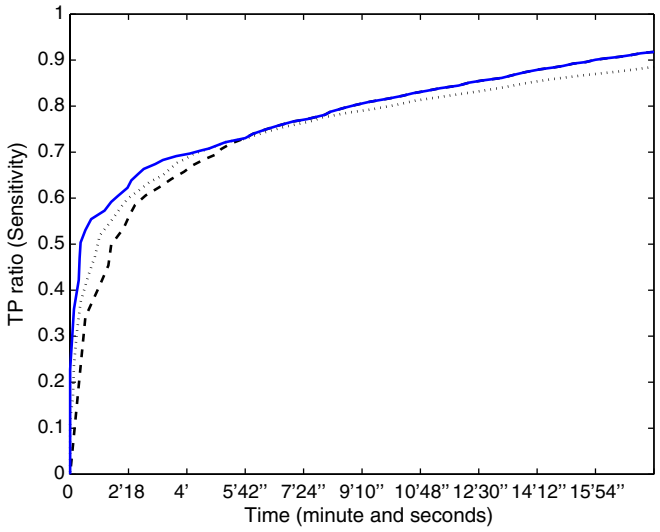

Fig. 5. The time-accuracy variant of ROC curve (solid line). TP are plotted against visualization time. For comparison, two more ensemble curves are shown: (i) bagging (dashed line) and (ii) decision tree, 1-NN and 5-NN (dotted line).

Table 2
Analysis of a 20,000-frame video with 30 contractions

Method	Positives	# Frames	Visualization time
MANUAL	30 (100%)	20,000	1 h 7 min
Threshold	29 (99%)	5000	2 h 45 min
98% Detection	28 (98%)	500	16 min 40 s
80% Detection	23 (80%)	275	9 min 10 s

4. Conclusions and further research directions

In this work, we show that ROC analysis shortens significantly the time required for a qualified professional to inspect videos of intestinal wireless capsule endoscopy with a minimal loss of performance. We tested eight single classifiers, a hybrid ensemble model based on all combinations of these (247 ensembles) and a bagging ensemble of 25 decision trees. The best model according to the AUC criterion was the bagging ensemble. As we were interested in finding a compromise between accuracy and inspection time, a variant of a ROC curve was designed plotting the best achievable sensitivity versus inspection time. Operation points can be picked from this curve, which offer significant reduction of the inspection time with reasonable sensitivity.

This methodology can be applied to other types of intestinal findings (different types of contractions, ulcera, tumors, etc.).

Future plans are focused on feature selection, feature extraction and especially developing new features. Pattern recognition literature abounds with exquisite and powerful feature selection and extraction methods (Aha and Bankert, 1995; Blum and Langley, 1997; Dash and Liu, 1997; Jain and Zongker, 1997; Scott et al., 1998), which should be carefully examined for their suitability for imbalanced problems. We expect that collaboration with domain experts will help us design new features which account for a number of subtleties used by the experts to identify a contraction.

References

Adler, D.G., Gostout, C.J., 2003. Wireless capsule endoscopy. Hospital Physician, 12–14.

Aha, D.W., Bankert, R.L., 1995. A comparative evaluation of sequential feature selection algorithms. In: Proc. 5th Internat. Workshop on AI and Statistics, 1995, pp. 1–7.

Bauer, E., Kohavi, R., 1999. An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. Mach. Learning 36, 105–142.

Blum, A., Langley, P., 1997. Selection of relevant features and examples in machine learning. Artificial Intell. 97, 245–271.

Bradley, A., 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recogn. 30, 1145–1159.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth and Brooks.

Brodsky, L.M., 2003. Wireless capsule endoscopy (Issues in Emerging Health Technologies). CCOHTA 53.

Chawla, N.V., 2003. Data duplication: An imbalanced problem. In: Workshop on Learning from Imbalanced Datasets. II. International Conference on Machine Learning.

Dash, M., Liu, H., 1997. Feature selection for classification. Intell. Data Anal. 1, 131–156.

Domingos, P., 1999. Metacost: A general method for making classifiers cost-sensitive. In: Proc. Fifth ACM SIGKDD Internat. Conf. on Knowledge Discovery and Data Mining, pp. 155–164.

Duda, R.O., Hart, P.E., Stork, D.G., 2001. Pattern Classification. Wiley-Interscience.

Duin, R.P.W., Juszczak, P., Paclik, P., Pekalska, E., De Ridder, D., Tax, D.M.J., 2004. PRTools4, A Matlab Toolbox for Pattern Recognition. Delft University of Technology.

Eliakim, R., 2004. Wireless capsule video endoscopy: Three years of experience. World J. Gastroenterol. 10, 1238–1239.

Hansen, M.B., 2002. Small intestinal manometry. Physiol. Res. 51, 541–556.

Jain, A.K., Zongker, D., 1997. Feature selection: Evaluation, application and small sample performance. IEEE Trans. Pattern Anal. Mach. Intell. 19 (2), 153–158.

Kubat, M., Holte, R.C., Matwin, S., 1998. Machine learning for the detection of oil spills in satellite radar images. Mach. Learning 30, 195–215.

Ling, C.X., Li, C., 1998. Data mining for direct marketing: Problems and solutions. Knowledge Discovery Data Mining, 73–79.

Ling, C.X., Huang, J., Zhang, H., 2003. AUC: A statistically consistent and more discriminating measure than accuracy. In: Proc. 18th IJCAI.

Mac Namee, B., Cunningham, P., et al., 2002. The problem of bias in training data in regression problems in medical decision support. Artificial Intell. Med. 24, 51–70.

Maloof, M.A., 2003. Learning when data sets are imbalanced and when costs are unequal and unknown. in: Proc. Internat. Conf. on Machine Learning. Workshop on Learning from Imbalanced Data Sets II.

Monard, M.C., Batista, G., 2003. Learning with skewed class distributions. Cadernos de Computaçao.

Rosset, S., 2004. Model selection via the AUC. In: Proc. 21st ICML.

Schulmann, S.K., Hollerbach, M.D., et al., 2005. Feasibility and diagnostic of video capsule endoscopy for small bowel polyps. Amer. J. Gastroenterol.

Scott, M.J.J., Niranjan, M., Prager, R.W., 1998. Parcel: Feature subset selection on variable cost domains. Technical Report. Cambridge University, Engineering Department.

See-Kiong, N., Zexuan, Z., Yew-Soon, O., 2004. Whole-genome functional classification of genes by latent semantic analysis on microarray data. Conf. Res. Practice Inform. Technol. 29, 123–129.

Tan, A.C., Gilbert, D., Deville, Y., 2003. Multi-class protein fold classification using a new ensemble machine learning approach. Genome Inform. 14, 206–217.

Viola, P., Jones, M., 2001. Rapid object detection using a boosted cascade of simple features. IEEE Comput. Soc. Conf. on Computer Vision and Pattern Recognition 1, 511–518.

Zadrozny, B., Elkan, C., 2001. Learning and making decisions when costs and probabilities are both unknown. In: Proc. 7th ICKDDM.