

# Experiments with SVM and stratified sampling with an imbalanced problem: Detection of intestinal contractions

Fernando Vilariño, Panagiota Spyridonos, Petia Radeva, Jordi Vitrià  
Computer Vision Center. Universitat Autònoma de Barcelona. Spain  
{[fernando@cvc.uab.es](mailto:fernando@cvc.uab.es)}

## ABSTRACT

In this paper we show some preliminary results of our research in the fieldwork of classification of imbalanced datasets with SVM and stratified sampling. Our main goal is to deal with the clinical problem of automatic intestinal contractions detection in endoscopic video images. The prevalence of contractions is very low, and this yields to highly skewed training sets. Stratified sampling together with SVM has been reported in the literature to behave well in this kind of problems. We applied both the SOMOTE algorithm developed by Chawla et al. and under-sampling, in a cascade system implementation to deal with the skewed training sets in the final SVM classifier. We show comparative results for both sampling techniques using precision-recall, which appear to be useful tools for performance testing.

## 1. Introduction

Automatic detection of intestinal contractions is one paradigmatic example of classification with imbalanced datasets. Its prevalence is very low, and the data analysis requires high amounts of expert time.

Both the number of intestinal contractions, and their temporal distribution along the intestinal tract, characterize small bowel motility patterns that are indicative of the presence of different malfunctions. Different techniques have been applied for intestinal motility analysis in several medical imaging modalities. A good review about this issue can be found in (Hansen; 2002).

The novelty of our research in this fieldwork relies on the use of Wireless Capsule Video Endoscopy images (WCVE) (Schulmann et al.; 2005; Brodsky; 2003; Eliakim; 2004). In this clinical domain, the specialist has to analyse a video, and manually label each frame where a contraction event happens. Usually, each video analysis may last one or two hours, and among a typical quantity of 20.00 frames, only 700 contractions are reported.

We focused our efforts on the automatic detection of intestinal contractions using video as data source. We have trained a SVM system with contraction and non-contractions frames from several videos, previously labelled by hand by the experts. The choice of SVM (Vapnik; 1995) is underpinned by the fact that empirical results show a good behaviour of this technique with moderate skewed datasets. Recently, several methods have been developed to improve the performance of SVM classifiers on imbalanced problems (Akbari et al ; 2004; Brank et al; 2003). Stratified sampling is based on re-sampling the original datasets in different ways: under-sampling the majority class or over-sampling the minority class. In this work, we show the preliminary results of a comparative study of under-sampling vs. SMOTE over-sampling technique –developed by (Chawla et al; 2002)-.Both techniques SVM and other popular single classifiers have been applied. With the purpose of reducing the imbalanced character of the datasets, we use a 2-steps cascade methodology that prunes false positives without losing sensitivity.

In order to assess the performance of the different methods implemented, precision-recall curves are proposed. The main advantage of these plots is that they show both the sensitivity of the

system and the noise introduced. A detailed explanation of the utility of these plots for imbalanced dataset is also developed.

The rest of the paper is organised as follows: Section 2 introduces the methodology used: the data description and the feature set, and describes the classification system. Section 3 shows the comparative results for several single classifiers and SVM, using different sampling methodologies. Finally, Section 4 is devoted to the discussion of results and suggestions for further research.

## 2. Methodology

### 2.1 Intestinal contractions

Intestinal contractions are a dynamic event. The expert labels a frame in a video as a contraction if it corresponds to a temporal pattern that spans 9 frames as an average –corresponding to 5 seconds in real time-. The frame labelled as contraction is, therefore, the central frame of a sequence of 9 frames. Several examples of intestinal contractions can be found in Fig 1.

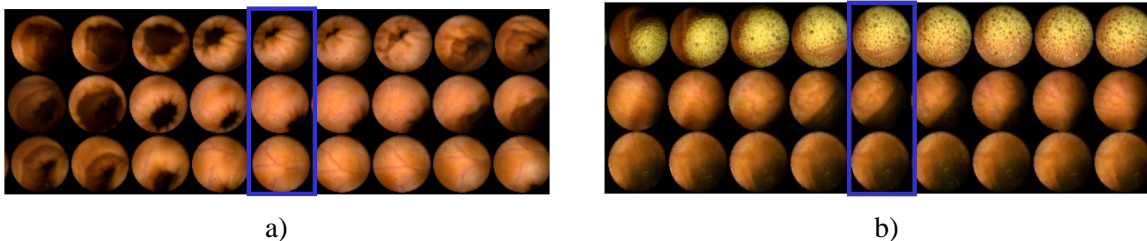


Fig 1: Three typical sequences of contractions a) and non-contractions b). The central frame on each sequence is highlighted.

With the aim of automatically detecting intestinal contractions for each frame, 6 features were calculated based on well known intensity and texture descriptors: normalised intensity, contrast, hole detection based on a Laplacian filtering and 3 values from concurrence matrices. We build up a feature vector of 6x9 features, associating with each frame the dynamic information of the whole sequence. All the chosen features are popular standard descriptors exhaustively used and referenced in the image analysis and computer vision literature (Gonzalez et al.; 2002; Russ; 1994). As far as we know, no previous works have been published to describe the kind of intestinal events we are dealing with in this kind of image acquisition technology. We are developing our project in a continuous contact with the clinical experts, and on each research step, new information is added from the expert's knowledge about different typologies of contraction events. So, our aim is to get the most general and flexible feature set, so as to be adaptable to this dynamic scenario. We reached this optimal set through an exhaustive heuristic search achieved through the repetition of several experiments, and the analysis of the relevance of each feature for classification.

### 2.2 Classification System

Each video typically has 20.000 frames, among which only around 500 contractions are to be found. Since the system should classify all these 20.000 frames as contractions or as non-contractions, a highly skewed problem is presented. We adopted a pre-classification step with the aim of eliminating as many false positives as possible from the original video, without losing many true contractions. This first stage was implemented in the following way: we

consider each one of the features as single values that can be used to divide video frames into contractions and non-contractions using a single threshold. When a threshold is applied, all the video frames are divided into two separated sets: the positives, that will pass to the next step of the system, and the negatives that will be definitely rejected. We expect that most of the contractions of our video will lay in the positive set, and that many of the non contractions will fall on the negative set. In order to guarantee this, we looked for the threshold value of every feature that kept the 99% of contractions in the positive set. Next, among all thresholds, we selected the one less non-contractions in the positive set. With this simple method, a ratio of 1:5 in the proportion between contractions and non-contractions was achieved. This strategy helps us to bound the knowledge that the classifier acquires from the video frames, in an attempt to reduce the variability of the samples in the decision space.

The second stage of the classification system was the SVM, trained only with the frames that passed the first stage. We used a radial basis function kernel (RBF) with  $\gamma=0.01$ . Most classifiers tend to classify everything as the majority class –negative- as the imbalanced character is accented, so this is the optimal solution in terms of global error. For this kind of imbalanced problems, the performance of the SVM is less sensitive, as the classification strategy is based only on some samples –the support vectors- and the rest of the samples have no influence. However, several studies seem to show that stratified sampling is a useful tool for improving the final decision border in SVM (Crone et al; 1997; Akbani et al; 2004). We have tested two different strategies of sampling: under-sampling the majority class and over-sampling with the SMOTE technique.

### 3 Experimental Results

For our experiments, 8 videos of 1-2 hours of length were used. These videos were labelled by the experts, identifying the frames where the contractions were present. An average of 500 contractions were labelled by the expert in each video. We applied the leave-one-out method: each run, one video was used for testing and the rest for training, in a resulting number of 8 runs. On each run, the threshold for the first step was calculated, and the SVM was trained exclusively with the frames that passed the first stage. We also trained a set of popular single classifiers for comparison purposes: linear, logistic, and 1,5, and 10 nearest neighbour (1-NN, 5-NN, 10-NN) (Duda et al.; 2001).

Two sampling methods were tested: under-sampling and SMOTE. For under-sampling, we trained the classifiers choosing randomly from the training set a number of non-contractions equal to that of contractions, so the final training set for each leave-one-out run was build up with  $500*7$  contractions and the same number of non-contractions –around 7.000 training samples. For SMOTE, we replicated the positive samples up to the number of non-contractions. Since around 2500 frames typically pass the first stage, this yields to a training set of around 35.000 samples. The discussion section analyses this point, which deserves special attention.

We carried out our experiments in order to answer two main questions: 1) In what measure the sampling technique used affects the performance of the classifiers tested, and 2) In what measure SVM outperforms the rest of classifiers for our dataset. With the aim of illustrating both questions, we use the precision-recall curves (*pr-curve*), that is a standard evaluation technique in the information retrieval community (Van Rijsbergen; 1979). *Precision* is a measure in the interval [0,1] defined as the ratio of true positives detected over all the system output. It will be *one* when the system detects only positives, and *zero* when the system detects only negatives. In this way, a low precision value is associated with a high number of false positives, and it can be viewed as a measure of noise at the output stage. *Recall* –also known as *sensitivity*- is a measure in the interval [0,1] defined as the ratio of true positives detected over the total number of positives. It will be 1 when the system detects all the existing positives, and

zero when no positive is detected. For imbalanced problems, both measures together give more information than ROC curves. In ROC curves (Metz; 1978; Swets et al 1982), precision is substituted by 1-specificity, or *false positive ratio*, and so the proportion between positives and negatives is not taken into account .

In order to give response to question 1, comparative plots are presented in Figure 2. We can see that no substantial improvement is achieved by any of both sampling techniques, neither for single classifiers nor for the SVM, and the classifiers performance seems to be quite invariant to the sample method. Regarding the performance of SVM with respect to the rest of the classifiers, both for under-sampling and SMOTE, SVM outperforms their competitors. Figure 3 shows that only for low levels of recall –when almost no positive is detected- NN classifiers appear to be competitive. A quantitative approach can be analysed by means of the area under the PR curve (AUPRC). The AUPRC gives us a metric for assessing the classifier performance. We calculated it using the polygonal rule, in the same way as it is done for the calculus of the area under a ROC curve (Bradley; 1997). Table 1 shows the AUPRC related to the graphs in Figure 3.

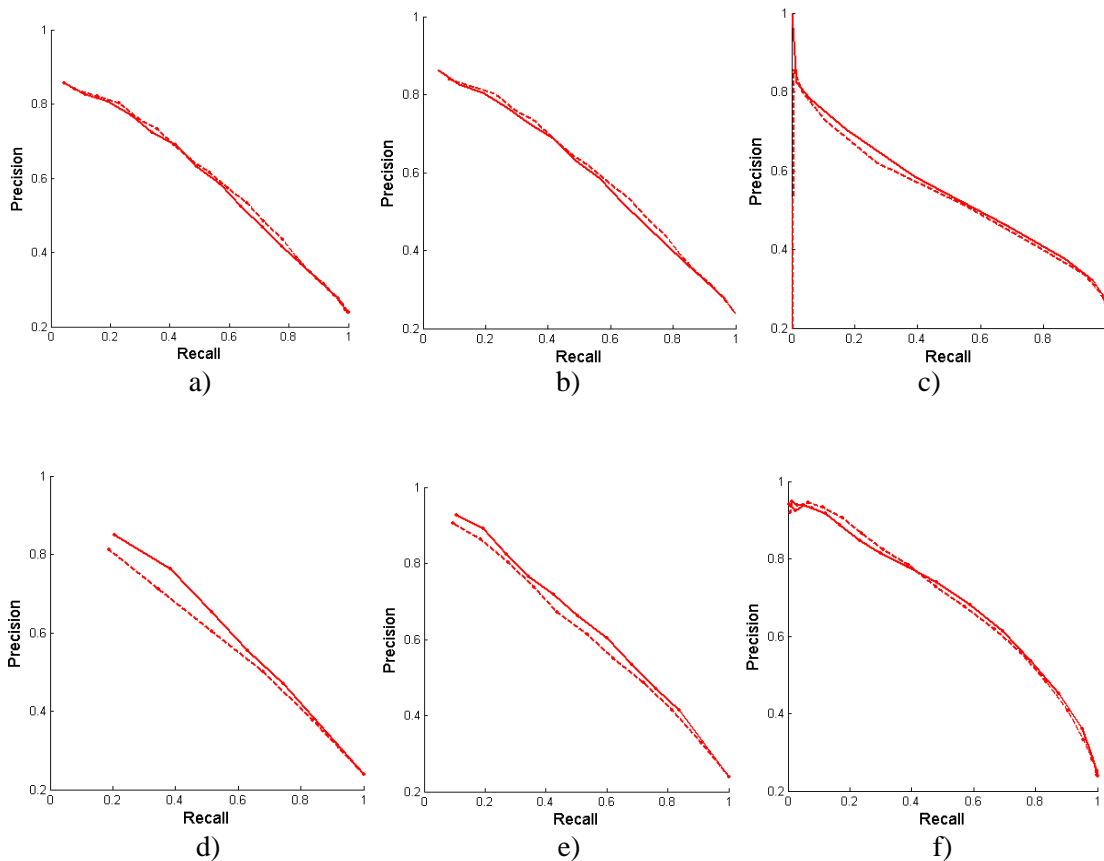


Figure 2: Under-sampling (dashed) vs. SMOTE over-sampling (solid) for detection of intestinal contractions with several classifiers. a) linear, b) logistic, c) 1-NN, d) 5-NN, e) 10-NN, f) SVM.

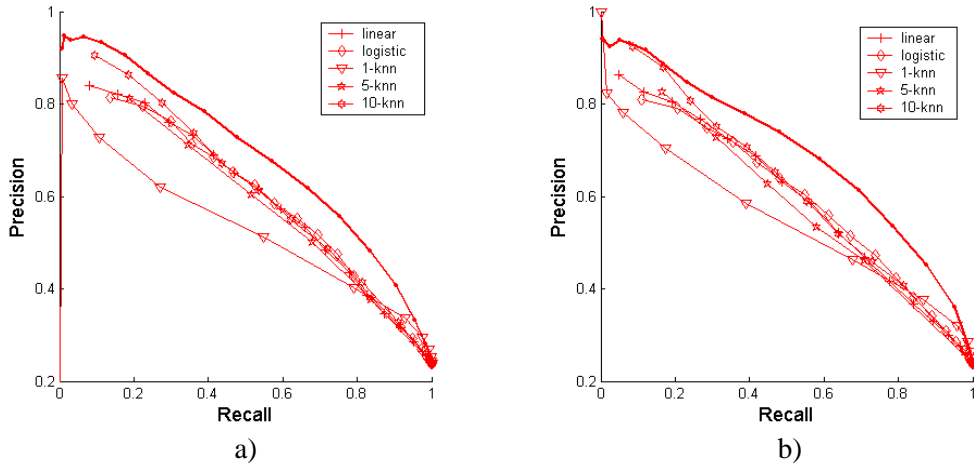


Figure 3: Single classifiers vs. SVM (solid line) for a) under-sampling and b) SMOTE.

Table 1: AUPRC for under-sampling and SMOTE experiments. SVM outperforms all the rest in both cases.

Classifier	AUPRC Under-sampling	AUPRC SMOTE
Linear	0.6100	0.6021
Logistic	0.6067	0.6006
1-NN	0.5324	0.5485
5-NN	0.5946	0.5849
10-NN	0.6282	0.6216
SVM	0.6934	0.6946

Several important conclusions can be inferred from these results, which deserve special attention so next section is devoted to their analysis.

#### 4 Discussion

We have presented some preliminary results of our research in the automatic detection of intestinal contractions in video endoscopy, a highly imbalanced problem due to the low prevalence of contractions in video. In order to implement a classification system, a set of classical image descriptors based on luminance, contrast and texture was associated to each frame. The dynamic behaviour was taken into account building up feature vectors containing all these basic descriptors for each one of the frames in a  $\pm 4$  frame neighbourhood.

One of the main characteristics of our system is the cascade implementation. In a first step, we make an exhaustive search of a single threshold in the feature space that keeps the 99% of contractions and rejects the majority of the non-contractions. The video frames passing this threshold are sent to the second step, consisting of a SVM classifier. It can be affirmed that this way of operating with imbalanced problems constitutes a rapid and efficient way of reducing the total number of false positives, i.e., reducing the ratio of the imbalanced datasets.

The plots of section 3 show a comparative study between two different methods of sampling: under-sampling the majority class and over-sampling the minority class, both for single

classifiers and SVM. The choice of these single classifiers is due to two important reasons: 1) on the one hand, linear and logistic classifiers are simple and fast; K-NN is a non-parametric technique that makes no assumption over the probability density function; we aimed at finding out if there was the same difference in behaviour with these classifiers in order to infer conclusion for the SVM. 2) On the other hand, for the SMOTE implementation the training set typically has around 35.000 samples in a 54 feature space. With this data, most of the software implementations –PRTools were used (Duin et al; 2004)- did not present the possibility of testing some common classifiers such as Parzen or decision trees owing to this huge amount of data.

In this point, recent works have tested SMOTE with several classifiers in common public databases, such as UCI (Chawla et al; 2003; Akbani et al; 2004), showing improvement with respect to under-sampling in some experiments. The results seem to show that for our problem, not a hard assessment can be done in order to affirm that SMOTE outperforms under-sampling –maybe, slightly in the best of cases-. It must be pointed out that the aim of SMOTE is to provide a sufficient quantity of minority class samples to equilibrate the number of samples for the negative class. But this is not the only important issue: special attention must be paid to the fact that when we have very few samples for the minority class, they may not be enough to reconstruct the probability density function underlying them, which is basically the role of SMOTE. We suggest that this is the case of our data set, as well as of other examples presented in the papers mentioned above. Our efforts are focused on the research of this point and the theoretical and empirical confirmation of this hypothesis.

The final results of our SVM classifier outperform all the remaining classifiers. This fact has been illustrated by means of the precision-recall curves that are presented as a useful tool for the assessment of imbalanced problems. In terms of operating points, the plot in Figure 3 b) shows that with an 80% of detection of all contractions, a 40% of false positives is achieved. This appears to be a promising result that motivates us to follow this line of research improving the system, testing new classifiers and sampling methodologies.

The aim of this paper is twofold: 1) On the one hand to present our preliminary results showing that SVM outperforms the best of single classifiers used for this particular imbalanced problem of intestinal contractions. 2) On the other hand, to show that both sampling techniques used in this work yield to similar results for SVM. We cannot avoid deepening in the analysis of the consequences associated with our results. In fact, if the different sampling techniques used appear to have no effect in the final performance of the system, it is due to some relevant characteristics of our specific problem, closely related to the statistical properties underlying our datasets. Moreover, the sampling techniques we use, which seem to work for other studies – some of them already referenced in this paper-, appear to give no performance improvement for our problem. This is a very challenging and interesting line of work, in which our group is completely involved.

We could re-formulate last paragraph into the next question: "Why is performance improving using one sampling method problem dependent in SVM with imbalance datasets?". In this sense, the next experiment is to be of high interest and can help us to shed light in this scenario: We must consider that we have two imbalanced datasets with samples  $N_1$  for the minority class –positives, contractions-, and  $N_2$  for the majority class –negatives, non-contractions- and, consequently  $N_1 \ll N_2$  (this is the typical scenario of an imbalance problem). Now, keep  $N_1$  fixed and under-sample from  $N_2$  several times, to generate smaller number of samples for training the classifier. Our purpose focuses on seeing the extent to which SVM is better as the skew or imbalance gets worse. The same experiment can be repeated over-sampling  $N_1$ . This analysis is extremely important to draw any firm conclusions about classification performance. Nevertheless, the inference that can be obtained from it is not so straightforward, and it is worthy of special and subtle attention. In this point, and in order to understand why a classifier is achieved, two main questions should be stated: how is the data distribution *around the*

*decision border* for *both* datasets, and how *descriptive* the positive samples are for the probability density function (PDF) of its class. Why? The first question is strictly related to the decision of the margins for the SVM, the second one is strictly related to the way in which over-sampling will perform. This analytical study is of high relevance, but it also holds intrinsic difficulties related to the statistical description and the geometrical analysis necessary to understand how the real distribution of the whole data in the n-feature space is, and how well the distribution provided by the samples we use for training fits the real PDF of the minority class. Finally an equally interesting question is open: how the choice of the sigma parameter in the SVM may affect the final performance of the classification in this scenario. We are preparing a more extended work with all these points of study, including the proposed experiment for a further piece of research.

Finally, one of the main objectives of our future work aims at addressing the feature extraction problem, in order to achieve better descriptors for the intestinal contraction events, as well as deepening into the achievement of the optimal set of features that we need in order to properly describe an intestinal contraction –i.e., the feature selection problem-. We expect that our collaboration with the clinical experts will help us to advance in this challenging fieldwork, which is of vital importance for an optimal result.

## References

- [Akbari et al; 2004] Akbari, R. Kwak, S. Japkowicz, N. Applying Support Vector Machines to Imbalanced Datasets. European Conference on Machine Learning. 2004, pp:39-50
- [Bardley ; 1997] Bradley, A. The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms. Pattern Recognition. 1997. 30 pp:1145-1159.
- [Brank et al; 2003] Brank, J. Grobelnik, M. et al. Training text classifiers with SVM on very few positive examples. April 2003. Technical Report MSR-TR-2003-34
- [Brodsky; 2003] Brodsky, L.M. Wireless capsule endoscopy. Issues in Emerging Health Technologies. CCOHTA. 53. 2003.
- [Chawla et al; 2003] Chawla, N., Hall, L., Kegelmeyer, W. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research
- [Crone et al; 1997] Crone, S. F., Lessmann, S., Stahlbock, R. Empirical Comparison and Evaluation of Classifier Performance for Data Mining in Customer Relationship Management. 3<sup>rd</sup> International Conference on KDDM. 1997.
- [Duda et al; 2001] Duda, R., O., Hart, P., E., Stork, D., G. Pattern Classification. Willey Inter-Science. Willey & Sons, Inc. 2001
- [Duin et al.; 2004] Duin, R.P.W., Juszczak, P., Paclik, P., Pekalska, E., Ridder, D.M.J. Tax, PRTools4, A Matlab toolbox for pattern recognition, Delft University of Technology. 2004.
- [Eliakim; 2004] R. Eliakim. Wireless capsule video endoscopy: Three years of experience. World journal of Gastroenterology. 2004. 10, pp: 1238-1239.
- [Gonzalez et al.; 2002] Gonzalez, R.C., Woods, R.E. Digital Image Processing. Prentice Hall. 2<sup>nd</sup> ed. 2002.

[Hansen; 2002] Hansen, M.B. Small Intestinal Manometry. *Physiological Research*. 2002. 51, pp: 541-556.

[Metz; 1978] Metz, C. Basic Principles of ROC Analysis. *Seminars on Nuclear Medicine*. VII. 1978, pp: 283-298.

[Russ; 1995] Russ, J., C. *The Image Processing Handbook*. IEEE Press. 2<sup>nd</sup> ed. 1994.

[Schulmann et al.; 2005] Schulmann, S.K., Hollerbach M.D., et al. Feasibility and diagnostic of video capsule endoscopy for small bowel polyps. *American Journal of Gastroenterology*. 2005.

[Swets et al; 1982] Swets, J., Pickett, R. *Evaluation of Diagnostic Systems: Methods for Signal Detection Theory*. New York. Academic Press. 1982.

[Van Rijsbergen; 1979] Van Rijsbergen, C. *Information Retrieval*, Dept. of Computer Science, University of Glasgow, 1979.

[Vapnik; 1995] Vapnik, V. *The nature of Statistical Learning Theory*. Springer\_Verlag, Ny.