

A Semi-Supervised Learning Method for Motility Disease Diagnostic

No Author Given

No Institute Given

Abstract. This work tackles the problem of learning a robust classification function from a very small sample set when a related but unlabeled data set is provided. To this end we define a new semi-supervised method that is based on a stability criterion. We successfully apply our proposal in the specific case of automatic diagnosis of intestinal motility disease using video capsule endoscopy.

1 Introduction

In many pattern classification problems, the acquisition of labeled training data is costly and/or time consuming, whereas unlabeled samples can be obtained more easily. Semi-supervised learning addresses this problem by using unlabeled data, together with the labeled data, to build better classifiers. Because semi-supervised learning requires less human effort and gives higher accuracy, it is of great interest both in theory and in practice [1, 2].

Semi-supervised methods can be seen as a solution to the problem of generalizing from small samples. Successful learning from a very small number of training samples requires the introduction of a certain hypothesis bias using additional information, and one such source of information may be unlabeled data. Semi-supervised learning methods use unlabeled data to either modify or re-prioritize hypotheses obtained from labeled data alone. Although not all methods are probabilistic, it is easier to look at methods that represent hypotheses by $p(y|x)$, and unlabeled data by $p(x)$. Generative models have common parameters for the joint distribution $p(x, y)$. It is easy to see that $p(x)$ influences $p(y|x)$. Mixture models with EM are in this category, and to some extent self-training. Many other methods are discriminative, including transductive SVM, Gaussian processes, information regularization, and graph-based methods.

We formulate our problem as follows: given a reduced set of labeled samples belonging to two different classes and a set of unlabeled samples, the task is to learn a binary classification function that generalizes well when tested by a Leave one out Cross-Validation method. The problem is challenging because the number of labeled samples is small and we cannot use classical re-sampling methods to assess generalization capacity of the classifier when considering unlabeled data. To this end we introduce a stability criterion based on data perturbation methods that allows us to select which unlabeled samples are added to the learning set in order to modify the generated hypothesis. This method is applied

to a Support Vector Machine Classifier trained with the problem of intestinal motility disease diagnosis.

This paper is organized as follows. In the next section we describe the proposed semi-supervised learning method and the stability criterion. In Section 3 we explain the problem, the diagnostic of intestinal motility dysfunction using the video capsule endoscopy. In Section 4 we expose the experimental results, comparing the results when using a supervised learning method with the obtained results when using the proposed semi-supervised learning method. Finally, Section 5 ends this paper with our conclusions.

2 The Method

The simplest method of incorporating unlabeled data into a new model is self-training. In self-training a classifier is first trained with the small amount of labeled data. The classifier is then used to classify the unlabeled data. Typically the most confident unlabeled points, together with their predicted labels, are added to the training set. The classifier is re-trained and the procedure repeated. Self-training has been applied to several natural language processing tasks like word sense disambiguation [3] or to classify dialogues as 'emotional' or 'non-emotional' [4], and also to computer vision problems like object detection from images [5].

Co-training is another way to train models from unlabeled data [6]. Unlike self-training, co-training requires multiple learners, each with a different "view" of the data. When one learner is confident of its predictions about the data, we apply the predicted label of the data to the training set of the other learners.

A variation suggested by Dasgupta et al. [7] is to add data to the training set when multiple learners agree on the label. If this is the case, we can be more confident that the data was labeled correctly than if only one learner had labeled it.

In the case of supervised learning, a learned function generalizes well if it does about as well on new inputs as on the old ones. Given an appropriate measure for the 'cost' of an error, the most common approach for assuring that a learning algorithm holds the previous property is to choose the least expensive function over the set of training samples, an approach to learning called *empirical risk minimization*. A classical result in learning theory shows that the functions learned through empirical risk minimization generalize well only if the 'hypothesis space' from which they are chosen is simple enough. The classical definition of a 'simple enough' hypothesis space is based on the *Vapnik-Chervonenkis* dimension [8], but although this approach has generated powerful learning algorithms, the complexity of hypothesis spaces for many realistic scenarios quickly becomes too hard to measure with this definition.

Recently Poggio et al. [9] proposed an elegant solution to this difficulty that shifts attention away from the hypothesis space. Instead, they require the learning algorithm to be stable if it is to produce functions that generalize well. An

algorithm is stable if the removal of any one training sample from any large set of samples results almost always in a small change in the learned function.

In view of these considerations, we propose a semi-supervised method that uses a stability criterion to add samples to the training data set. The proposed stability criterion differs slightly from the one defined by Poggio et al. in order to be used with small sample sets: "An algorithm is stable *if adding one training sample* from a set of samples results almost always in a small change in the learned function".

In order to use the stability criterion during the application of a semi-supervised method, we need define how to measure *small changes in the learned function* when adding a training sample. To this end we adopt the following strategy:

1. Given a binary classification problem with two class labels $\{-1, +1\}$, a training set with an additional training sample, and a learning method that is able to assign a confidence level to the classification of any sample, perform a Leave one out Cross-Validation test over the training set, including the new training sample.
2. At each stage of the Cross-Validation test, repeat n times the following procedure:
 - (a) Add some random noise to the samples, creating a perturbed data set.
 - (b) Build a classifier and classify the "leave one out" sample.
 - (c) Return the class and the confidence level of the sample classification.
3. Compute the mean confidence classification level for each sample and rank all training samples from both classes by assigning the lowest rank to the most confident sample from class -1 and the highest rank to the most confident sample from class +1.
4. Compute a measure of stability of the algorithm for the additional training sample by comparing the produced rankings (a learned function is stable if the sample ranking does not change a lot). To compare two rankings we can use Spearman's rank coefficient [10], this is a non-parametric measure of correlation that assesses how well an arbitrary monotonic function could describe the relationship between two variables, without making any assumptions about the frequency distribution of the variables. The coefficient is defined as:

$$r_{ab} = 1 - 6 \sum_d \frac{(d_{r_a} - d_{r_b})^2}{N(N^2 - 1)} \quad (1)$$

where the sum is taken overall the samples d and d_{r_a} , d_{r_b} , are the obtained ranks of the d sample using two different perturbed data sets a and b , respectively, and N is the number of samples. At the end we average the pairwise rank correlation coefficient overall the n iterations. We will consider that the example with the highest coefficient correlation is the most stable for the classifier.

Once we know how to measure the stability criterion we can define our stability based semi-supervised learning algorithm. The basic idea is to add at each

Table 1. Stability based semi-supervised learning algorithm

1. Pick one classification method.
2. Train a classifier f using the labeled data set $(x, y) \in (X_l, Y_l)$.
3. Use f to classify all unlabeled items $x \in X_u$.
4. Compute the stability criterion for each $x \in X_u$.
5. Pick from the unlabeled data set the sample x^* with the highest stability on the results of the classification, and add $(x^*, f(x^*))$ to labeled data.
6. Repeat this procedure until all samples from the unlabeled data set are added to the labeled data set.

step the sample from the unlabeled data set which presents the highest stability with respect to the classifier. Table 1 summarizes the proposed algorithm.

3 The problem

With the recently appeared Wireless Capsule Video Endoscopy (WCVE) a new field of research is opened to study small intestine affections. Currently, the most extended diagnosis test for small bowel motility disorders is intestinal manometry [11]. This technique has some drawbacks: it is highly invasive, require the hospitalization of the patient and monitorization of the whole process by medical staff, and also the acquired information is limited to examination of some kind of information.

The WCVE is an ingestible device provided with all the suitable technologies for image acquisition, including auto illumination and radio frequency emission. For more details see [12]. The capsule is swallowed by the patient and emits a radio frequency which is received and stored by an external machine. The result is a video which records the "travel" of the capsule along the intestine with a frame ratio of two frames per second.

The human intestine is the portion of the alimentary canal extending from stomach to the anus. It is divided in two segments, the small intestine and the large intestine. The small intestine is subdivided into duodenum, jejunum and ileum, and the large intestine is subdivided into cecum, colon and rectum. We are interested in the part of the intestine comprised between post-duodenum and cecum, because in the other part of the video we can not extract reliable information. In the portion of the video corresponding to this part, we can observe three different elements or events: the intestine wall, the intestinal content, and some artefacts related with the movement of the intestine. In Fig.1 we can see an example of the intestine image visualized by the WCVE. In this image the lumen and the intestinal wall are visualized. The lumen is the cavity where digested food goes through and from where the nutrients are absorbed. It can be recognized in the video images as a dark area.

In order to quantify the information that is visualized in the video, we have to extract the values related to several features. All these analyzed features are, according to medical experts, important in the diagnostic of the intestinal motility disfunctions. The features that we analyze from the video are:

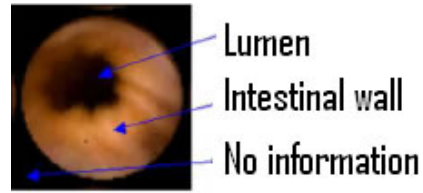


Fig. 1. Example of intestinal video image

Static rate. Static frames appears when the camera has a null apparent motion and the visualized frames are almost the same. The experts believe that a high level of static frames could be considered as a symptom of intestinal disease. An example of this kind of sequences can be seen in Fig.2 - 1st row. We characterize this feature with Earth Mover Distance method (EMD) [13]. From this feature we are interested in quantifying the next information: *percentage of static frames in the video*; *percentage of static frames in those parts of the video that turbid is not present*; *the mean length of the static sequences*; and the *static level of the video*, all of them computed with the mean of the static level of each frame.

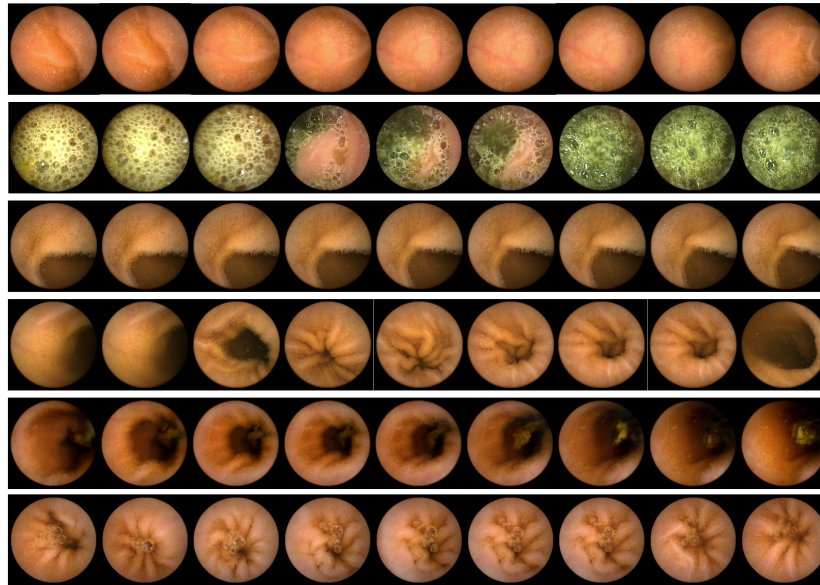


Fig. 2. Some different intestinal video images. From top to bottom: Static Sequence, Turbid frames examples, Tunnel Sequence, Occlusive contraction, Non-Occlusive contraction and Sustained Contraction.

Turbid. The turbid is food in digestion or intestinal juices. The food appears as small pieces through the gut and the intestinal juices could be presented as liquid or bubbles (Fig.2 - 2nd row). For details about turbid detection using WCVE see [14]. Related with turbid we get the next values: *percentage of turbid frames in the video; percentage of turbid frames that are apparently stopped; and static level in the parts that the turbid is not present.*

Tunnel. A tunnel is a sequence of frames where the lumen appear static for a long period of time (Fig.2 - 3rd row). The interpretation of the tunnel is that the intestine is relaxed, and there is not contractile movement.

The tunnel is described in terms of the lumen area of a sequence of nine frames. In order to estimate the area of the lumen, a Laplacian of Gaussian filter is applied (LoG)[15]. For further details about detecting these frames see [16]. We are interested in assessing: *percentage of tunnel in the video; percentage of tunnel in those frames that turbid is not present; the mean length of the tunnel sequences; and the level of static level of all tunnel frames.*

Contractions. Intestinal contractions are the result of muscular stimulation produced by the nervous system. The analysis of these contractions has been proved to be a meaningful method for diagnosis several intestinal dysfunctions [17]. From a physiological point of view, intestinal contractions can be divided into the following three groups:

- Occlusive contractions: the lumen in the central frame of these contractions is completely closed. These are considered as the classical intestinal contractions (Fig.2 - 4th row).
- Non-occlusive contractions. In this type of contraction the lumen never appears completely closed. The origin of this kind of contraction is based in the physiological fact that the intestinal wall do not perform enough pressure during the contractile activity (Fig.2 - 5th row).
- Sustained contractions. They are produced by the muscular tone, and can be visualized as a continuous closing of the intestinal lumen (Fig.2 - 6th row).

The method to classify frames as contractions has been described elsewhere and is called Cascade [?]. Each step of the Cascade receives as input the output of the previous step. The initial input is all the frames of the video, and the output is the set of frames where a contraction is suggested to appear.

We use a Relevance Vector Machine Classifier [18] in order to detect the non-occlusive contractions from all the contraction set detected by the system . The principal feature to characterize this type of contraction is the area of the lumen in each frame. Generally, in non-occlusive contractions the lumen area of the central frame is bigger than it is in the rest of the contractions. This area is estimated in the same way as is done in the tunnel frame detection. The resulting vector of nine lumen areas is used as feature vector to the classification of occlusive or non-occlusive contractions.

We are interested in analyzing: *the number of contractions per minute in the video; the number of contractions per minute in those parts that the turbid is not present; and the percentage of the contractions that are non-occlusive.*

Wrinkle pattern level presence. The wrinkle star pattern is an omnipresent characteristic of the sustained contractions. In the frames where the wrinkle pattern appears we observe strong edges of the folded intestinal wall, distributed in a radial way around the intestinal lumen. In order to localize this pattern in video frames an accurate wrinkle detection is essential. For more information about this method see [19]. The output is the level of wrinkle presence in each frame. Related with this feature we get: *the percentage of video frames with wrinkles; the percentage of the frames without turbid where the wrinkle star pattern appears; the mean length of the wrinkle sequences; the percentage of frames with a very low level of wrinkle presence; the percentage of frames with a high level of wrinkle presence; and the percentage of contractions with wrinkles.*

All these features are assembled in a 21-dimensional feature vector that is used to diagnose subjects in intestinal motility disfunctions.

4 Results

In this section we evaluate the proposed method over our problem. Our video set is provided by the Digestive Diseases Department of General Hospital Vall d’Hebron in Barcelona, Spain. This video set was obtained using the wireless endoscopic capsule developed and provided by Given Imaging, Ltd., Israel. All videos were created at the same conditions, patients and healthy volunteers were in fasting (without eating and drinking in the previous 12 hours). There are three different group of videos, one from patients, another from healthy volunteers, and finally another one from persons with non conclusive manometric diagnose that are used as unlabeled data. Healthy volunteers were randomly selected from a bigger pool of subjects without any symptom. We consider these subjects as healthy subjects without performing a manometry test, because the probability to be patient is too small and anyone of them has any symptom to be patient, statistically the probability of error is very low. Our data set is composed by 50 healthy volunteers, 15 patients and 17 unlabeled subjects. It is important to notice that the intestinal motility diseases are very odd ones. This is the reason why the number of available patient cases is very low.

We evaluate the performance of this methodology using a Leave one out Cross-Validation Method [20] with the data set. Before performing this test we normalize the data with zero mean and standard deviation set to one. The results are validated using several measures that are described in terms of true positives (TP), true negative (TN), false positives (FP), and false negative (FN) as follows: $Error = FP + FN$, $Sensitivity = TP / (TP + FN)$, $Specificity = TN / (TN + FP)$, $Precision = TP / (TP + FP)$ and $False Alarm Ratio, FAR = FP / (TP + FN)$.

In order to perform our classification test we use the Support Vector Machine Classifier (SVM) [21]. SVM Classifier looks for the hyperplane which separates positive and negatives samples, maximizing the distance to the hyperplane. The original algorithm proposed by V. Vapnik was a linear classifier, however there is a way to create non-linear classifiers by applying the kernel trick to the SVM framework. We used the SVM Classifier with the radial basis function kernel [22]

represented by:

$$K_{rbf}(x, x_i) = \exp\frac{-|x - x_i|^2}{2\sigma^2} \quad (2)$$

where the parameter σ will be found by Cross-Validation in each test.

Our first test is performed only using the supervised data (healthy volunteers and patients). In the first row of Table 2 we display the obtained results for this test. We get a 4.61% of error, 80.00% of sensitivity, and specificity, precision and FAR are: 100.00%, 100.00% and 0.00% respectively. It means that all healthy subjects are correctly classified, but three patients are considered as healthy subjects.

The previous test present an important drawback, the limited size of the data set. In order to overcome this problem and be able to get better results we perform another test using both, the supervised and unsupervised data set in the proposed semi-supervised learning method. Before use this algorithm we have to define two important issues:

- Which binary classifier is used and how we estimate the confidence level of the sample classification.
- And how we perturb the original data.

For the first issue we again use an SVM Classifier. As stated before SVM Classifier is a binary classifier which looks for an hyperplane that maximizes the distance between two classes. To get the classification confidence value of a given classified sample we take the distance from this sample to the SVM hyperplane. To create the perturbed data set we choose to add a percentage f of Gaussian noise to the original data set. This percentage was fixed to $f = 20\%$ after some experimental tests.

The semi-supervised learning method labels 8 examples of the unsupervised data set (of 17 elements) as patients and 9 of these examples as healthy volunteers. We use this new training set in order to perform the validation for proposed method of our problem. As we can be seen in the second row of Table 2 we get 1.53% of error, 93.00% of sensitivity, and specificity, precision and FAR are: 100.00%, 100.00% and 0.00% respectively. It means that only one patient is considered as healthy subject and all healthy subjects are well classified.

Table 2. Classification Results

	Error	Sensitivity	Specificity	Precision	FAR
supervised learning	4.61%	80.00%	100.00%	100.00%	0.00%
semi-supervised learning	1.53%	93.33%	100.00%	100.00%	0.00%

5 Conclusions

In this paper we propose a new semi-supervised learning method for the diagnostic of intestinal motility dysfunctions. The principal difficulty of this problem

is the low number of patients samples available. However, we have another set of unlabeled data and we propose to use it with the semi-supervised learning method.

We propose the stability as the criterion used in the semi-supervised learning method. The stability measure that has been defined is based on the Spearman's Rank Correlation.

The performed experiments show that using this semi-supervised learning method with an unlabeled set of samples results in a improvement of results. It is important to notice that we obtain this improvement of the results avoiding the computation of the error (on the unsupervised data) as the criterion to be optimized.

Even if the test have been performed on a very small data set the WCVE and this method are very promising in order to diagnose intestinal motility dysfunctions. The good results demonstrate that this method could be a candidate to replace, in a close future, the most used diagnosis tests for intestinal motility dysfunctions, the manometry test.

Acknowledgements

This work was supported in part by a research grant from Given Imaging Ltd., Yoqneam Israel, and Hospital Universitari Vall dHebron Barcelona, Spain, as well as the project TIN2006-15308-C02 project, MEC, Spain. The technology and methods embraced by this disclosure has been filed for patent protection.

References

1. Chappelle, O., Schölkopf, B., Zien, A.: Semi-supervised learning. Unpublished book. MIT Press (To appear)
2. M. Seeger: Learning with Labeled and Unlabeled Data. Technical report, Edinburgh (2000)
3. Yarowsky, D.: Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In: Proceedings of Association for Computational Linguistics. (1995)
4. Maeireizo, B., Litman, D., Hwa, R.: Co-training for predicting emotions with spoken dialogue data. In: Companion Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL). (2004)
5. Rosenberg, C., Hebert, M., Schneiderman, H.: Semi-supervised self-training of object detection models. (2005) I: 29–36
6. Blum, A., Mitchell, T.: Combining Labeled and Unlabeled Data with Co-Training. In: Proceedings of Computational Learning Theory. (1998)
7. Dasgupta, S., Littman, M.L., Mcallester, D.A.: Pac generalization bounds for co-training. In Dietterich, T.G., Becker, S., Ghahramani, Z., eds.: NIPS, MIT Press (2001) 375–382
8. V. Vapnik: The Nature of Statistical Learning Theory. Springer (2000)
9. Poggio, T., Rifkin, R., Mukherjee, S., Niyogi, P.: General conditions for predictivity in learning theory. *Nature* **428** (2004) 419–422
10. Callan, J.P., Connell, M.E.: Query-based sampling of text databases. *Information Systems* **19**(2) (2001) 97–130

11. Hansen, M.B.: Small intestinal manometry. *Physiological Research* **51** (2002) 541–556
12. Iddan, G., Meron, G., et al.: Wireless capsule endoscopy. *Nature* **405** (2000) 417
13. Y., R.: The earth movers distance as a metric for image retrieval. *International Journal of Computer Vision* **40** (November 2000) 99–121(23)
14. Author: title. *Journal X* (year) XX–XX
15. Russ, J.: *The Image Processing Handbook*. CRC Press (1999)
16. Author: title. *Journal X* (year) XX–XX
17. Quigley, E.M.: Gastric and small intestinal motility in health and disease. *Gastroenterology Clinics of North America* **25** (1996) 113–145
18. Tipping, M.: The relevance vector machine. In: *Advances in Neural Information Processing Systems*, Morgan Kaufmann (2000)
19. Author: title. *Journal X* (year) XX–XX
20. Stone, M.: Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society B* **36** (1974) 111–147
21. Vapnik, V.: *The Nature of Statistical Learning Theory*. Springer, New York (1995)
22. Keerthi, S.S., Lin, C.J.: Asymptotic behaviors of support vector machines with gaussian kernel. *Neural Computation* **15** (2003) 1667–1689