Patient-Specific Ventricular Beat Classification without Patient-Specific Expert Knowledge: A Transfer Learning Approach

Jenna Wiens and John V. Guttag

Abstract— **We present an adaptive binary classification algorithm, based on transductive transfer learning. We illustrate the method in the context of electrocardiogram (ECG) analysis. Knowledge gained from a population of patients is automatically adapted to patients' records to accurately detect ectopic beats. On patients from the MIT-BIH Arrhythmia Database, we achieve a median sensitivity of 94.59% and positive predictive value of 96.24%, for the binary classification task of separating premature ventricular contractions (PVCs), a type of ectopic beat, from non-PVCs.**

I. INTRODUCTION

Accurately labeled training data for new classification tasks is often expensive and time consuming to obtain. Transductive transfer learning provides a method for extracting useful knowledge from labeled training data and adapting it to a related target task for which there is no labeled data [1]. In contrast to traditional supervised learning, transductive transfer learning does not assume that the input distribution is identical between training and target tasks [2]. However, it often relies on the assumption that a given point will be labeled identically in both the training and target domains. In many situations where transfer learning would be useful, this assumption does not hold. For example, when examining EEG's from two different patients, an expert might give different labels to points that are close or even identical in the feature space [3]. When one uses traditional transfer learning methods, inter-patient differences can render a classifier trained on patient A nearly useless when applied to patient B. In this paper, we present a method that relaxes this assumption.

Our two stage method starts by constructing a highly sensitive minimum enclosing ball (MEB) that includes a subset of the overlapping data from the training tasks. In the case of ECG data, for example, the MEB might enclose typical normal sinus rhythm beats, since these differ the least across patients. Next, we use the labels generated by applying the MEB to the target task to train a task-specific linear SVM. In contrast to training a transductive SVM on the unlabeled data, this two stage approach does not require *a priori* knowledge about the ratio of positive to negative examples (which can vary significantly across patients).

We evaluated our method on a public domain database of ECG signals, for the task of identifying PVCs. We chose this particular task since there exists reliable publicly available hand-coded software that we could use as a benchmark. The classifiers produced by our method performed similarly to the hand-coded classifier achieving a median gross sensitivity of 94.59% and a median gross positive predictive value of 96.24%. This is superior to the performance of a global linear SVM on the same data, even though the global SVM uses more training data.

John V. Guttag is with Faculty of Electrical Engineering and Computer Science at MIT, guttag@csail.mit.edu

Jenna Wiens is with the Department of Electrical Engineering and Computer Science at MIT, jwiens@csail.mit.edu

II. OVERVIEW OF APPLICATION

There exist two popular machine learning approaches to building heartbeat classifiers for ECG analysis. The first, is a global approach that attempts to train a classifier on a population of patients and apply it directly to a specific patient [4], [5], [6]. However, because of inter-patient differences, global classifiers that attempt to directly transfer knowledge often perform poorly [7].

The second approach focuses on patient-specific and patient-adaptable classifiers, where all or part of the training set and test set are drawn from the same patient within the same time frame. [8], [9], [7] each showed that the addition of patient-specific training data can improve the performance of global heartbeat classifiers. Unfortunately, although such classifiers are more accurate, their construction is labor intensive since they require expert knowledge (typically supplied by a cardiologist) to produce a labeled training set for each patient. Moreover, since a patient's ECG often evolves over time, an expert might have to produce such labels at each time of analysis.

Physicians who are trained to read ECGs work by combining what they have learned from a career of reading ECGs with knowledge extracted from ECGs of the current patient. We hypothesize that adaptive transductive transfer learning can provide automated heartbeat classifiers with a similar ability to adapt.

Here we focus on the binary classification task of PVCs vs. non-PVCs. However, we hypothesize that our method can adapt to other linearly separable tasks such as classifying normal sinus rhythm beats vs. ectopic beats.

III. METHOD

The two main classification stages of our method, see Figure 1, are described in Sections III-A and III-B, respectively.

A. Knowledge Transfer

The goal of the first classification stage in Figure 1 is to transfer knowledge about the non-PVC beats from the training data (collected from different patients) to the unlabeled target data. This stage is based on the assumption

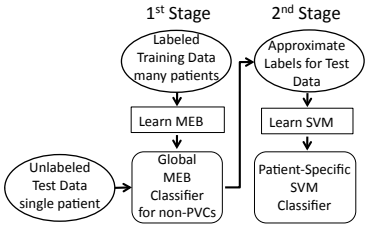

Fig. 1. First, a MEB transfers cross-patient knowledge, providing approximate beat labels for an incoming test record. These labels are then used to learn a linear SVM, resulting in a patient-specific classifier without requiring any patient-specific labeled training data.

that there is some overlap among the non-PVC beats in the population.

The problem of transferring knowledge about the non-PVC beats from the training data can be interpreted geometrically. Given an appropriate feature space the non-PVC beats from the training patients cluster together such that one can learn a hypersphere that encloses this cluster. Applied to the target data, any example that lies on or inside this hypersphere is likely of the same class. This is a form of 1-class classification, described in [10], and referred to here as the minimum enclosing ball (MEB). Based on all n non-PVC examples from the training records, $\underline{x}_i$ where $i = 1...n$, one can learn a hypersphere that encloses a maximum number of beats with a minimum radius:

$$\textbf{MEB:} \quad \min_{R,\underline{a},\xi_1...\xi_n} \quad R^2 + C\sum_i \xi_i \tag{1}$$

$$\text{s.t.} \quad ||\underline{x}_i - \underline{a}||^2 \leq R^2 + \xi_i, \tag{2}$$

$$\xi_i \geq 0, \ \forall i \tag{3}$$

In the MEB problem, the two parameters $\underline{a}$ (center) and R (radius) characterize the hypersphere. The slack variables ξ_i account for the possibility of error in the training set and the C parameter represents the tradeoff between the volume of the ball and the number of errors on the training set [10]. We use the dual of the MEB problem, which allows for the incorporation of kernels. We use a radial basis function (RBF) kernel, since with the correct parameters, it allows for a tight description of the data.

A tight description of the non-PVC beats from the training records will enclose only the typical non-PVC beats. Applied to the target data, anything that lies on or inside the hypersphere is labeled as a non-PVC. One can be relatively confident about the labels of the examples that fall inside the MEB. Conversely, it is expected that many non-PVC beats will fall outside the MEB. *I.e.*, the MEB is used to detect typical rather than anomalous examples.

B. Task Adaptation

The second classification stage involves adapting the "global" knowledge, transferred from the training tasks, to the target task. It involves using the approximate labels produced by the MEB to train a task-specific linear SVM for the target data.

At this stage in the classification process (the beginning of stage 2 in Figure 1) each target example has already been labeled positive or negative by the MEB produced in stage 1. Using these labeled feature vectors, a linear SVM is learned for the target data. A linear kernel is used to reduce the risk of over-fitting to the initial labels, which are known to be only approximations.

At this stage it is unknown which of the approximate labels are correct. However, one can be confident that almost all the examples that fall inside the MEB are labeled correctly. To exploit this, the cost of misclassifying what falls on or inside the global MEB is set to be greater than the cost, C, of misclassifying what falls outside the MEB, by a factor of $R > 1$ as in [11].

The result of this stage is a linear classifier specific to the target task, which did not require any human provided target-specific labels. We expect the classifier to be accurate under the following two conditions:

- There exists some overlap in the feature space, among examples belonging to one class from the related tasks.
- The positive and negative examples for the target task are close to linearly separable.

C. The Data Set

To evaluate the classification scheme presented in Figure 1, we used data from the MIT-BIH Arrhythmia Database available at Physionet.org [12]. The database, contains half-hour recordings from 47 different patients, totaling approximately 109,000 cardiologist labeled heartbeats.

We focused on the 27 recordings, listed in Table 1, that contained a majority of normal sinus rhythm beats. Patients with paced beats, bundle branch block and/or long runs of atrial flutter/fibrillation, were omitted.

D. Pre-Processing

First, baseline wander was corrected as described in [5]. In addition, power line interference was removed using a 60Hz notch filter. Finally, the amplitude of each patient's ECG signal was normalized, to account for differences in gain settings.

Once pre-processed, the RR-intervals (length of time between preceding and following heartbeat peaks) were recorded and the data was segmented into individual beats. A simple segmentation scheme was employed where each beat was constructed from 100 samples (277ms) before the R-peak and 157 samples (436ms) after the R-peak.

E. Feature Extraction

To construct this classifier, each segmented heartbeat was transformed into a feature vector. We characterized each beat using three features:

1) pre RR-interval normalized by patient's average,
2) post RR-interval normalized by patient's average, and
3) morphological distance between the current beat and the patient's median beat (the median of each sample)

The first two features were chosen because PVCs usually have a shorter pre RR-interval and a longer post RR-inteval

than other beats. The third feature was chosen because PVCs typically have a morphology different from non-PVCs. We calculated the morphological distance between two beats using the dynamic time warping algorithm described in [13].

IV. Model Selection & Validation

We split the data into separate training/validation and test sets. The training set (14 patients) was used for model selection while the test set (13 patients) was used for error estimation. The model parameters were selected based on a leave-one-patient-out cross validation performed on the training set.

We used the Statistical Pattern Recognition Toolbox [14] implementation of Tax and Duin's methods for support vector data description [10] to develop the MEB. We selected the parameters for constructing the MEB, by performing leave-one-patient-out cross validation using the 14 patients in the training set. For each patient we trained a MEB on the non-PVCs from all other patients and then tested it on the beats of the remaining patient. We set $C = 10,000$, since we assumed that the number of cardiologist errors was small. We swept the RBF kernel spread, γ and measured the average sensitivity, specificity and positive predictive values using cross-validation. Since the ultimate goal of the two-stage classifier is to detect PVCs rather than non-PVCs, we consider the PVCs the positive class. In the initial classification stage, we wish to transfer knowledge pertaining to only the most typical non-PVCs. Therefore, we trade-off specificity for high sensitivity, and as a result achieve a high negative predictive value. Consequently, we chose a γ that corresponded to a sensitivity of 99.5% on the training data.

Next, we used SVM_{light} to perform a grid search on the training data to select the SVM parameters C and R. For a given R there was little change in the performance as we varied C. In contrast, the change in performance as we varied R for a fixed C was substantial.

Based on the cross-validation, C and R were chosen as the values at which the sensitivity and positive predictive value were closest to equal.

V. Results & Discussion

Table I reports the median PVC detection results for each patient from our data set. The first classifier is the method described here, the second is a global SVM and the last is our benchmark, a hand-coded algorithm from [15]. These are discussed individually in the following subsections.

A. Performance of Transfer Learning Based Method

Given the small size of our data set and the high inter-patient differences, we repeated the validation and evaluation process 100 times with the data split randomly into separate training sets for model selection and test sets for error estimation. The independent cross-validation of the 100 independent trials resulted in a mean $\gamma = 0.05$ with a standard deviation of 0.02. The parameters of the linear SVM were held constant at R=10 and C=100 across all test sets. The results of this evaluation are presented and discussed here.

The median performance of the 100 independent trials is a gross sensitivity of 94.59%, and positive predictive value of 96.24%.These results exclude the first 5-minutes of each record. This was done so that we could compare our results with the results reported by [15] whose algorithm tests on the last 25 minutes of each recording.

TABLE I
Classification performance for each record.

Rec.	#	Total	Test Results					
			MEB+SVM		SVM		Hamilton	
	Beats	PVC	TP	FP	TP	FP	TP	FP
100	1900	1	1	0	1	0	1	0
101	1522	0	0	3	0	3	0	2
103	1727	0	0	0	0	0	0	0
105	2154	29	27	78.5	29	105	18	38
106	1695	460	411	0	427	0	455	1
112	2109	0	0	2	0	0	0	0
113	1504	0	0	0	0	5	0	0
114	1603	30	28	4	30	4	30	5
115	1635	0	0	0	0	0	0	0
116	2015	98	95	0	95	0	97	2
117	1282	0	0	2	0	2	0	0
119	1660	364	364	0	364	0	364	0
121	1558	1	1	1	1	3	1	0
122	2052	0	0	0	0	0	0	0
123	1268	3	3	0	3	0	0	0
200	2166	700	659	3	689	1	669	2
205	2199	65	63	1	62	1	62	0
208	2433	824	795	27.5	797	17	803	2
209	2517	1	0	0	1	3	0	5
213	2698	195	156.5	24	165	36	184	3
215	2793	131	126.5	3	130	1	128	1
220	1692	0	0	0	0	27	0	0
223	2197	455	409.5	5	116	2	403	3
228	1702	302	297	3	302	6	298	2
230	1858	1	0	0	1	6	1	0
233	2559	692	680	3	660	2	679	4
234	2290	3	3	1	3	2	3	0
Tot.	52788	4355	4119.5	161	3876	226	4196	70

The false positive (FP) column for our classification method ("MEB+SVM") in Table I, shows that three records, 105, 208 and 213, account for over 80% of the total false positives. On average, over 78 false positives were detected in record 105 alone. Almost all of the false positives for this record were located in portions of the ECG annotated as "noise" by the cardiologists. For record 213, many of the false positives were beats labeled by cardiologists as a fusion of ventricular and normal beats. The labels of fusion beats are often debatable, even among cardiologists. Likewise for record 208, many of the false positives detected were annotated as either fusion beats or noisy.

Figure 2 shows the cumulative distribution function for both the sensitivity and positive predictive value of the 100 independent trials. Well over half of the trials result in a high sensitivity and positive predictive value. However for about 10% of the trials, an "unfortunate" training and test set combination led to a classifier with a sensitivity and positive predictive value of less than 90%. We believe this can be attributed to the small size of our data set. We hypothesize that given a larger training set, the variance in the performance would decrease, leading to an increase in the average performance.

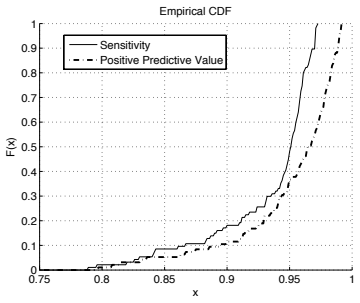

Fig. 2. Empirical CDF of the performance of 100 random trials.

B. Performance of a Global SVM

To compare the performance of our two-stage classifier with traditional global classifiers, we investigated the performance of a global linear SVM. While this many not be the "best" global classifier, linear classifiers are common in global heartbeat classification since they do not require the tuning of additional hyperparameters and maintain a level of interpretability not always achievable with non-linear classifiers. We independently trained 100 linear SVMs with $R = 1$ and $C = 100$ on each training set and then applied each to the corresponding test set. The columns headed "SVM" in Table I report the median test results for each patient. The global SVM resulted in a median gross sensitivity and positive predictive value of 89.00% and 94.49%.

Training a global SVM requires more training data than our method; it requires many labeled examples of PVCs, whereas our method does not. Yet, overall our technique outperforms the global SVM, particularly in terms of gross positive predictive value. The global SVM classifier does particularly poorly in the case of record 223, missing >70% of the PVCs. Record 223 contains ventricular arrhythmias not prominent in other records. We suspect that our method outperforms the global SVM in this case because it is capable of adapting to record 223, while the global SVM is not.

C. Performance of a Hand-Coded Classifier

We started this work hoping to demonstrate that transductive transfer learning could be used to build an automatic heartbeat classifier that performed as well as the best hand-coded classifiers. As a benchmark, we chose the algorithm developed by [15] for EP Limited. This algorithm, for detecting PVCs, was developed and evaluated by its authors on the same MIT-BIH Arrhythmia Database that we used, making a direct performance comparison possible.

We note several similarities between the performance of the two different techniques. For example, record 105, which contains a large amount of noise, is difficult for both techniques, and has the greatest number of false positives.

The pattern matching techniques from [15] achieve a gross sensitivity of 96.35% and a gross positive predictive value of 98.36% on the test set. On the same patients, our technique performs only slightly worse in terms of sensitivity and positive predictive value. A summary of the performance for the three techniques is reported in Table II.

TABLE II

GROSS CLASSIFICATION PERFORMANCE RESULTS.

Gross Median Results		
Classifier	Sensitivity	+ Pred. Value
MEB+SVM	94.59%	96.24%
Global SVM	89.00%	94.49%
Hamilton [15]	96.35%	98.36%

VI. CONCLUSION

We presented a novel binary classification method for PVC classification based on transductive transfer learning. We developed the method after making two key observations: 1) there is considerable overlap in the feature space for non-PVC beats, and 2) the data for each patient is close to linearly separable.

Knowledge from a population of patients with an underlying normal sinus rhythm was adapted to build an accurate patient-specific classifier. The resulting classifiers had a median gross sensitivity of 94.59% and positive predictive value of 96.24%, comparable to the performance of hand coded classifiers. In addition to outperforming traditional global-classifiers our technique requires less training data. Specifically, it requires only non-PVC data.

In this paper we focused on the task of PVCs vs. non-PVCs; however, we speculate that our method is applicable in other heartbeat classification tasks and contexts where the conditions of section III-B hold.

REFERENCES

[1] S. J. Pan and Q. Yang, "A Survey on Transfer Learning," *IEEE T KNOWL DATA EN*, vol. 99, no. 1, 2009.

[2] J. Quinonero-Candela, *Dataset Shift in Machine Learning*. MIT Press, 2009.

[3] H. Qu and J. Gotman, "A Patient-Specific Algorithm for the Detection of Seizure Onset in Long-Term EEG Monitoring," *IEEE T BIO-MED ENG*, vol. 44, pp. 115–122, Feb 1997.

[4] Y. Hu, W. J. Tompkins, J. L. Urrusti, and V. X. Afonso, "Applications of Artificial Neural Networks for ecg Signal Detection and Classification," *Journal of Electrocardiology*, vol. 26, pp. 66–73, 1993.

[5] P. de Chazal, M. O'Dwyer, and R. B. Reilly, "Automatic Classification of Heartbeats Using ECG Morphology and Heartbeat Interval Features," *IEEE T BIO-MED ENG*, vol. 51, pp. 1196–1206, 2004.

[6] I. Christov, I. Jekova, and G. Bortolan., "Premature Ventricular Contraction Classification by the Kth Nearest-Neighbours Rule," *Physiological Measurement*, vol. 26, no. 1, pp. 123–130, 2005.

[7] P. de Chazal and R. Reilly, "A Patient-Adapting Heartbeat Classifier Using ECG Morphology and Heartbeat Interval Features," *IEEE T BIO-MED ENG*, vol. 53, no. 12, pp. 2535–2543, Dec. 2006.

[8] T. Ince, S. Kiranyaz, and M. Gabbouj, "Automated Patient-Specific Classification of Premature Ventricular Contractions," in *Conf Proc IEEE Eng Med Biol Soc*, 2008, pp. 5474–5477.

[9] Y. H. Hu, S. Palreddy, and W. Tompkins, "A Patient-Adaptable ECG Beat Classifier Using a Mixture of Experts Approach," *IEEE T BIO-MED ENG*, vol. 44, no. 9, pp. 891–900, Sept. 1997.

[10] D. Tax and P. Duin, "Support Vector Data Description," *Machine Learning*, vol. 54, pp. 45–66, 2004.

[11] K. Morik, P. Brockhausen, and T. Joachims, "Combining Statistical Learning with a Knowledge-Based Approach ," in *ICML*, 1999.

[12] R. Mark and G. Moody, "MIT-BIH Arrhythmia Database," 1997. [Online]. Available: http://ecg.mit.edu/dbinfo.html

[13] Z. Syed, J. Guttag, and C. Stultz, "Clustering and Symbolic Analysis of Cardiovascular Signals," *EURASIP*, vol. 2007, pp. 97–112, 2007.

[14] V. Franc, "Statistical Pattern Recognition Toolbox," Master's thesis, Czech Technical University in Prague, February 2000.

[15] P. Hamilton, "Open Source ECG Analysis," in *CinC*, vol. 29, 2002, pp. 101–104.