Combining Unsupervised and Supervised Learning for Discovering Disease Subclasses

Pietro Bosoni[12], Svetlana I. Nihtyanova[3], Christopher P. Denton[3], Allan Tucker[2]
[1]University of Pavia, Italy; [2]Dept Computer Science, Brunel University, London. United Kingdom
[3]UCL Division of Medicine, Royal Free Hospital, London
allan.tucker@brunel.ac.uk

***Abstract*— Diseases are often umbrella terms for many subcategories of disease. The identification of these subcategories is vital if we are to develop personalised treatments that are better focussed on individual patients. In this short paper, we explore the use of a combination of unsupervised learning to identify potential subclasses, and supervised learning to build models for better predicting a number of different health outcomes for patients that suffer from systemic sclerosis, a rare chronic connective tissue disorder - but one that shares many characteristics with other diseases. We explore a number of different algorithms for constructing models that simultaneously predict health outcomes and identify subcategories.**

Keywords—Systemic sclerosis, disease subclass, classification

I. INTRODUCTION

Different diseases can affect people in different ways. There are a number of reasons for this. Firstly, disease categories are often "umbrella" terms for a group of subcategories of disease. Take Systemic Sclerosis (SSc), which is a relatively rare disease (overall incidence of up to 56 cases/million/year). It is a chronic connective tissue disorder, affecting the skin, peripheral circulation and multiple internal organs [1]. What is more, it can be classified into two major subsets - limited cutaneous SSc, where skin thickness affects only areas distal to elbows and knees and diffuse cutaneous SSc, where skin involvement can affect the whole body. Of course, these are unlikely to be the only subcategories and discovering others will be essential if we are to make more informed diagnoses.

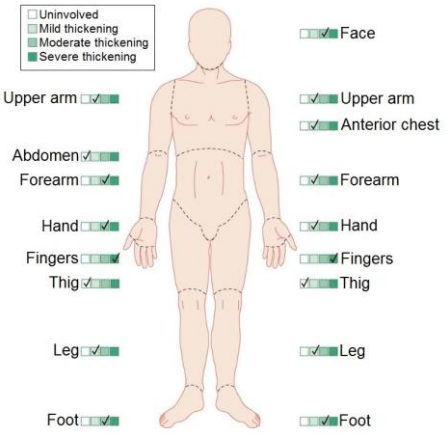

Fig. 1. Rodnan skin score (from Wigley FM: Systemic sclerosis: Clinical features. In Klippel JH, Dieppe PA (eds): Rheumatology, vol. 2, 2nd 2000.)

Secondly, people respond in different ways to the same disease. For example, in SSc some patients are more affected with complications in the lungs, whilst others are in the kidneys, heart or gastro-intestinal system. Patients undergo regular assessments, including history of symptoms, physical examination and a range of blood and internal organ tests. Extent and severity of skin tightness is measured by modified Rodnan skin score - see Figure 1, ranging between 0 (normal) and 3 (severe thickening). Additionally, results from lung function tests are measured regularly. These tests, in combination with other patient characteristics that do not change over time (e.g. serological markers, age at onset, etc.), can be used as predictors of organ complication / mortality. Whilst systemic sclerosis is a rare disease it shares traits that are common to many diseases: firstly, variability in progression between different individuals, including subclasses of disease that can inform how an individual will progress. Secondly, the eventual progression to an advance stage with similar advance-stage symptoms. In this paper we explore the use of classification methods to predict different disease outcomes using a combination of clinical indicators related to SSc. We use unsupervised methods to preprocess patients into different cohorts to identify variations in symptoms and improve classification. This discovery of subtypes of disease is becoming popular [2]. Previously we explored a variation of Naïve Bayes for disovery of glaucoma subtypes [3]. Here we focus on a general approach for any classifier.

II. METHODS

We use data on 700 SSc patients with baseline demographic and clinical characteristics and organ complications including modified Rodnan skin score. We explore three classifiers but focus on CN2 rules [3] in order to identify relevant clinical tests / demographics that are pertinent to disease outcome. This is due to the transparent nature of rule based methods. We explore a number of different experimental architectures:

i) *no_unsup* - Standard k-fold cross-validation classification. This involves using the different disease outcomes as class variables and other test data / demographics as predictors.

ii) *unsup* - K-fold cross-validation classification with unsupervised pre-processing. This involves applying a pre-processing stage whereby all patients are clustered using k-means with no class information.

iii) *unsup-class* - As (ii) but the class information from the training data is used to bias the clustering. The idea is that this will help to further refine the clusters so that they are not overly biased along the class variable. In other words, it is envisaged that the clusters will represent more fine-grain clusters within the different classes.

III. RESULTS

Figure 2 shows the performance (Area Under ROC Curve) of the three different approaches to classification of patients: using standard classification (*no_unsup*), using classification with preprocessing using unsupervised learning on the clinical test data only (*unsup*), and finally using unsupervised learning on the clinical test data and the class data that is available for training (*unsup_class*). These are assessed using three classifiers (Naïve Bayes, Support Vector Machine and CN2) for two disease outcomes: time to death from diagnosis (T2RIP) and time to Pulmonary Hypertension (T2PAH).

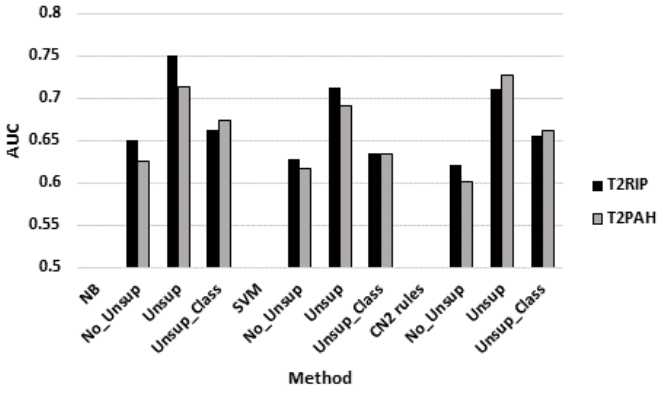

Fig. 2. Classification Area Under Curve (AUC) Comparison

It is clear that the use of unsupervised clustering prior to classification has improved the predictive capabilities of all three classifiers. In particular, the basic unsupervised clustering that doesn't take the class into account (*unsup*) seems to have improved prediction the most. This was unexpected as it was thought that incorporation of class information (*unsup_class*) would focus on more informative clusters, though it could be that class information creates more risk of irrelevant clusters from overfitting. Nevertheless, both improve upon standard classification (*no_unsup*). Figure 3 shows some example rules discovered using CN2 from the *unsup_class* experiments. Notice how the discovered subclasses of disease are exploited in the rules that result in improved accuracy. In particular Cluster C2 is often associated with another confounding feature to predict time to death greater than 147 months (T2RIP=1), whilst Cluster C3 is associated with features that predict shorter times to death (T2RIP=0). It is this incorporation of subcategory information that is improving prediction accuracy.

We now explore the general characteristics of these subcategories in detail. Figure 4 shows the breakdown of the discovered subcategories. Notice that the first cluster is mostly limited SSc, whilst clusters 2 and 3 are predominantly diffuse. Also notice that cluster 2, which appears in many rules (e.g. see Figure 3), is characterized by FVC and DLCO tests with values mostly in the higher quartiles, whilst cluster 3 is characterized by low FVC and DLCO scores. These characteristics are key to assisting the classifiers in improving accuracy and allow us to explore the interaction between features within cohorts that have different FVC and DLCO test results.

Rule quality	Coverage	Rule
0.88	14	IF Cluster=C3 AND Abs=4 THEN T2RIP=0
0.88	27	IF Cluster=C3 AND subset=0 THEN T2RIP=0
0.87	19	IF Cluster=C3 AND Hb=1 AND Cr=4 THEN T2RIP=0
0.86	24	IF Cluster=C3 AND Abs=3 THEN T2RIP=0
0.85	29	IF Cluster=C3 and DLCO=1 THEN T2RIP=0
0.81	36	IF Cluster=C2 AND Cr=3 THEN T2RIP=1
0.81	111	IF Cluster=C3 THEN T2RIP=0
0.79	33	IF age=4 AND Hb=1 THEN T2RIP=0
0.77	47	IF Cluster=C2 AND Hb=4 THEN T2RIP=1
0.76	51	IF Cluster=C2 AND Abs=1 THEN T2RIP=1

Fig. 3. Example CN2 Rules for Predicting Time to Death > 147 months (T2RIP=1)

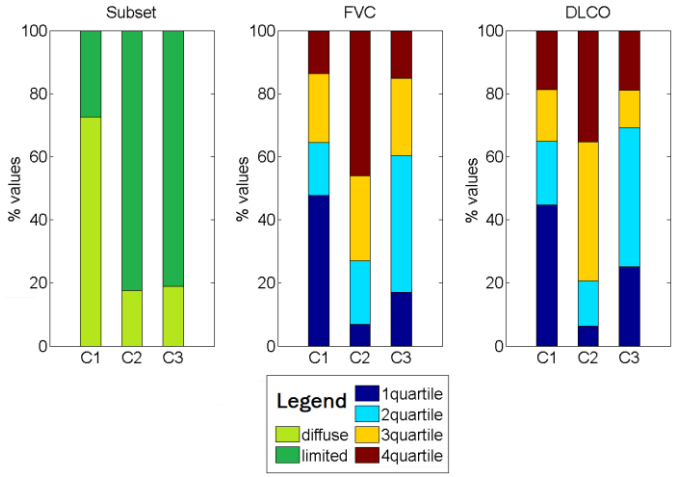

Fig. 4. Cluster Characteristics for Diffuse vs Limited SSC (subtype) and Forced Vital Capacity (FVC) and CO Diffusion Capacity (DLCO)

IV. CONCLUSIONS

In this short paper, we have introduced a simple framework for improving classification models by incorporating subcategory identification in two ways. We have tested these methods for predicting outcome for patients diagnosed with systemic sclerosis and show that subcategory discovery not only improves prediction but that the discovered rules incorporate subcategory information that can be directly interpreted to better understand the meaning of the new subtypes of disease.

REFERENCES

[1] Nihtyanova, S. Burden of disease and predictors of outcome in systemic sclerosis, thesis for Doctor of Medicine, University of London, 2013.

[2] Bailey, P. et al. Genomic analyses identify molecular subtypes of pancreatic cancer, Nature 531, 47–52 2016

[3] Ceccon et al. Exploring early glaucoma and the visual field test: classification and clustering using Bayesian networks, IEEE J Biomed Health Inform. 2014 May;18(3):1008-14.

[4] Clark, P. and Niblett, T (1989) The CN2 induction algorithm. Machine Learning 3(4):261-283.