ABSTRACT

The diagnosis of Chronic Obstructive Pulmonary Disease (COPD) is based on symptoms, clinical examination, exposure to risk factors (smoking and certain occupational dusts) and confirming lung airflow obstruction (on spirometry). However, most people with COPD remain undiagnosed and controversies regarding spirometry persist. Developing accurate and reliable automated tests for the early diagnosis of COPD would aid successful management. We evaluated the diagnostic potential of a non-invasive test of chemical analysis (volatile organic compounds - VOCs) from exhaled breath. We applied 26 individual classifier methods and 30 state-of-the-art classifier ensemble methods to a large VOC data set from 109 patients with COPD and 63 healthy controls of similar age; we evaluated the classification error, the F measure and the area under the ROC curve (AUC). The results show that classifying the VOCs leads to substantial gain over chance but of varying accuracy. We found that Rotation Forest ensemble (AUC 0.825) had the highest accuracy for COPD classification from exhaled VOCs.

Keywords: Automatic diagnosis from exhaled breath, Chronic Obstructive Pulmonary Disease (COPD), Classification, Classifier ensembles, Pattern recognition, Rotation Forest
INTRODUCTION

Chronic Obstructive Pulmonary Disease (COPD) is characterised by airflow limitation, which is not fully reversible. The causes are largely attributed to inhaling tobacco smoke, occupational exposure to dust and chemicals, and indoor and outdoor pollution (Rabe et al., 2007). COPD is a major public health problem and is the only one of the top five causes of death in the first world that is still rising. It is predicted to become the third leading cause of death by 2030, according to a study published by the World Bank/World Health Organization (WHO, 2008), and accounts for much chronic illness and morbidity. Yet, the Global initiative Obstructive Lung Disease (GOLD) report by Rabe et al., (2007) admits that COPD remains relatively unknown or ignored by the public as well as public health and government officials.

The current diagnosis of COPD is based on reported symptoms, patient’s medical history (particularly exposure to risk factors), clinical examination, and then confirming lung air-flow obstruction (spirometry) where Forced Expiratory Volume in 1 second (FEV1) divided by Forced Vital Capacity is less than 0.80 and FEV1 predicted is less than 0.7 . (Rabe et al., 2007)

Developing accurate and reliable automatic tests for early diagnosis of COPD is crucial for disease management as removing risk factors and early inhaled treatments has been shown to prevent progression, chronic ill health and premature death. (Rabe et al., 2007). The current main test, spirometry, is effort dependent and often performed poorly. It can lead to over diagnosis in the young and underdiagnosis in the elderly. Moreover, it has not been validated in ethnic minorities. (Rabe, 2007). The quest for a reliable biomarker in COPD is ongoing.

The smell of breath has long been linked with illness or physical conditions. Can volatile organic compounds (VOCs), measured from the exhaled breath, be used to identify COPD? Following on from Pauling’s (1971) initial description of around 200 volatile organic compounds (VOCs) in exhaled breath, the trapping, detection and analysis of breath VOCs have been further developed. VOC analysis has been used to distinguish smokers from non-smokers (van Berkel et al., 2008), recognition of asthma (Ibrahim et al., 2011; Fens et al., 2009), lung cancer (Ulanowska et al., 2011; Machado et al., 2005; Philips et al., 2003; Bajtarevic et al., 2010, Barkar 2006) and tuberculosis (Philips et al., 2007). Diagnosis of COPD from VOCs has also been attempted (Basanta et al., 2010, Fend et al., 2009, Van Berkel et al., 2010, Philips et al., 2012).

Here we study the diagnostic potential of the chemical signature of the exhaled breath for distinguishing between patients with COPD and healthy controls. We apply a large collection of state-of-the-art classification methods developed within the areas of pattern recognition, machine learning and data mining, with a special focus on classifier ensembles. We applied these methods to the largest data set so far derived from our previous work (Philips et al., 2012). We demonstrate that the ensemble methods are superior to the individual classifier methods, resulting in better classification accuracy, F measure and the area under the ROC curve (AUC).

MATERIAL AND METHODS

Related Work

Table 1 shows a summary of the classification methods and techniques used in the recent literature on diagnosis of lung disease based on breath samples. While the collection of sources is by no means comprehensive, it reveals that the possibilities offered by modern pattern recognition (Duda et al., 2001; Bishop, 2006), machine learning (Hastie et al., 2011; Schapire and Freund, 2012) and data mining (Witten & Frank, 2001) remain largely unexplored.

We applied the software package Weka (Hall et al., 2009); a collection of machine learning algorithms for data mining tasks. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited
A Review of Kernel Methods Based Approaches to Classification and Clustering of Sequential Patterns: Sequences of Discrete Symbols
www.igi-global.com/chapter/review-kernel-methods-based-approaches/76057?camid=4v1a

Gene Set- and Pathway- Centered Knowledge Discovery Assigns Transcriptional Activation Patterns in Brain, Blood, and Colon Cancer: A Bioinformatics Perspective