Latent Class Multi-Label Classification to Identify Subclasses of Disease for Improved Prediction

Awad A Alyousef[1], Svetlana Nihtyanova[2], Christopher P. Denton[2], Pietro Bosoni[3], Riccardo Bellazzi[3,] Allan Tucker[1]

[1] Dept. Computer Science, Brunel University London, UK
[2] UCL Royal Free Hospital, London, UK
[3] University of Pavia, Italy
awad.alsaidalyousef@brunel.ac.uk

Abstract

Disease subtyping, which helps to develop personalised treatments, remains a challenge in data analysis because of the many different ways to group patients based upon their data. However, if we can identify subclasses of disease then it will help to develop better models that are more specific to individuals and should therefore improve prediction and understanding of the underlying characteristics of the disease in question. This paper proposes a new algorithm that integrates latent class models with classification. The new algorithm uses latent class models to cluster patients within groups that results in improved classification as well as aiding the understanding of the underlying differences of the discovered groups. The methods are tested on data from patients with Systemic Sclerosis (SSc), a rare potentially fatal condition. Results show that the "Latent Class Multi-Label Classification Model" improves accuracy when compared with competitive similar methods.

Keywords— Multi-Label Classification; Latent Class Model; Naïve Bayes.

I. INTRODUCTION

Healthcare organisations need to find better methods to assist diagnosis that are both accurate and explainable. Machine learning classifiers that can exploit the huge amounts of historical patient data are clearly a promising technology to achieve this. Their aim is to accurately predict a class label for new patients (e.g. diagnosis or risk factor) based on historical data. However, in some situations, patients might belong to more than one class. For example, a patient might have diabetes and cancer at the same time [1]. In this case *Multi-Label (ML) classification* can be employed, where multiple class labels can be assigned to a single patient's data. ML classification is a challenging task in machine learning as it requires the prediction of more than one class. Previous research shows the effectiveness and robustness of ML classification. In [2] the authors observed that better accuracy is obtained when *algorithm adaption* is used (see Section B). Also, in [3] multi-label classification and associated *evolution metrics* have been introduced, suggesting that algorithm adaptation is the best option for ML classification. Dhongade et al. review ML classification and observe that the approach is mostly used in text categorisation and medical diagnosis [4]. This paper introduced ML-KNN and showed that ML-KNN is better than other established algorithms [4]. ML classification aims to predict all classes that belong to any patient and may also help to find the relationships between classes. A novel model for ML classification based on Bayesian networks was introduced in order to predict all classes whilst simultaneously finding correlations between classes. The model performed better compared to binary classification methods [5]. Although many researchers are working on enhancing Naïve Bayes performance in classification problems, less research has been done in ML Naïve Bayes. Shouman et al. explored in [6] the effectiveness of k-means as a clustering method in improving supervised learning techniques like Naïve Bayes. The results showed that integrating a clustering method with Naïve Bayes could enhance accuracy [6]. Naïve Bayes has been used as an important method in medical diagnoses fields as it can give accurate results and reveal hidden information between the variables [7]. Kabir et al. claim higher accuracy can be produced when datasets are split into sub-groups, where each group has similar intra-group characteristics [8]. They focus on the improvement of the classification accuracy for Naïve Bayes by clustering the dataset using K-means [8]. Eshghi compared traditional clustering methods and latent class models and found that using different clustering methods might produce different groups, suggesting that each methodology could lead to different interpretations [9].

Our research deals with ML classification of a disease where patients have multiple comorbidities. We explore the effectiveness of ML classification for Naïve Bayes classifiers but when a Latent Class Model is used to cluster patients. The latent classes within the model can help us to explain the relationship between the clusters and the comorbidities.

A. Latent Class Analysis: A Brief Overview

Latent class analysis (LCA) proposes that there is an unobserved variable that explains all the relationships between the observed variables. In medical diagnosis the observed variables are signs, symptoms, and test results. Latent class models have attracted much attention in the medical statistics community in the past few years [10]. They can be used for clustering categorical data based upon the assumption that observed variables are mutually independent given the class

variable. Many latent class models have been explored including the *Hierarchical Latent Class Model (HLC)*, which is a more general form of the Latent Class model based upon Bayesian networks (trees), where the leaf nodes are observed, and others are latent [11]. Dey et al. proposed a latent class model for identifying subgroups according to health effects in a large population. It used LCA to produce a clustering of patients [12] by using random sampling without replacement, assuming local independence between observed variables, and each patient is assigned to one class where *Bayesian Information Criteria* is used to select the best number of classes in the model.

B. Multi Label Classification: A Brief Overview

Multi-label classification is a classification task where each record can be associated with one or more classes. This is common in medical diagnosis, for example diabetes and prostate cancer could be found in one patient. Most of the works on ML classification deal with document classification problems. Carvalho et al. proposed three different methods that deal with ML classification. The first integrates single-label classifiers to be suitable for multi-label classification tasks. The second adapts internal mechanisms of single label classifiers to allow their use in multi label problems. The third method was a new algorithm designed to deal with ML classification [13]. There are two main approaches in ML classification: *Problem Transformation Methods* and *Algorithm Adaption Methods*. The first one converts ML classification problems into single label classification problems. The second one modifies single label classification algorithms to be applied on multi-label data [1]. Ceylan et al. studied the efficacy of using ML classification techniques for prediction of cervical cancer. Naïve Bayes, decision trees and others were compared in term of their accuracy, hamming loss, subset accuracy and ranking loss performance evaluation metrics [14]. The purpose of this study was to help physicians, academics and cancer researchers to create fast and accurate diagnosis. The results found similar accuracy that is over 80% for all applied methods [14].

II. METHODOLOGY

In this paper we explore a combination of latent class modelling in combination with ML classification when applied to a dataset that measures a number of complications associated with a rare disease, S*ystemic Sclerosis*.

The algorithm consists of four phases: In the first phase patient's data are collected along with the labels (classes). The latent class model is then applied to the data set in order to separate the dataset into three subgroups. We have classified each group using the Naïve ML Classifer (MLC) model in order to predict the labels (classes). Finally, we have evaluated the proposed algorithm and compared it to other related algorithms.

A. Data Collection

SSc is an uncommon connective tissue disorder with multisystem involvements and a chronic and often progressive course [15]. The diagnosis of SSc is made on clinical grounds and the ACR/EULAR 2013 Classification Criteria for SSc are widely accepted and used by clinicians and researchers [16]. The presence of skin thickening, affecting the fingers and spreading proximal to the metacarpophalangeal joints is sufficient to make the diagnosis. In patients with skin changes only affecting the fingers, additional features, including fingertip pitting scars or ulcers, telangiectasia, abnormal nailfold capillaries, Raynaud's phenomenon, SSc-specific antibody positivity or pulmonary involvement need to be present for a patient to be classified as having SSc.

We explore data on 677 SSc patients attending the Scleroderma unit at the Royal Free Hospital, including basic demographic and clinical characteristics, blood results and pulmonary function tests. Our purpose was to predict time to death, time to pulmonary fibrosis (PF) and time to pulmonary hypertension (PH) – two common complications in SSc. The aim of our proposed algorithm is to cluster the patients within three groups and to predict time to development of PF, PH and death for each group. The patients have been selected as follows: Select all patients from the above dataset who developed at least one of the above classes within the first 5 years and all patients who have not developed over 5 years. The novel algorithm was applied on the resulting dataset in order to predict time to development of PF, PH and death.

B. Latent Class Model

The latent class model uses a discrete latent variable that clusters patients within groups based on unobservable subgroups of individuals. In our algorithm we use one nominal variable with *K* categories representing the number of subgroups. Each subgroup (or cluster) of patients contains individuals that share common characteristics. The latent class model is estimated by the maximisation of the log-likelihood.

$$P(Y_i = y) = \sum_{u=1}^{k} \pi_u \, P_u(y) \quad (1)$$

The above equation is a manifest distribution of response vector Y_i.

π_u is the probability weight that patient *i* belongs to class *u* (*u*=1,……, *k*) and is calculated from the following equation:

$$P(U_i = \varepsilon_u) = \frac{\exp(\psi_{ou})}{1+\exp(\psi_{ou})} \text{ u.c. } \sum_u \pi_u = 1; \pi_u > 0 \quad (2)$$

The log-likelihood $\ell^*(\theta)$ may be efficiently maximised through an Expectation-Maximisation (EM) algorithm. The EM algorithm is based on the complete log-likelihood

$$\ell^*(\theta) = \sum_{i=1}^{n} \sum_{u=1}^{k} \lambda_{ui} \, [log \, \pi_u + log \, p_u \, (y_i)] \quad (3)$$

C. MLC Naïve Bayes

We have adapted the single Naïve Bayes algorithm in order to deal with sets of labels. The aim of this method is to predict all labels' values and discover if this improves the performance of the classification compared to multiple single class models (by assuming conditional relationships between the different labels). The conditional probability of patient $p_i$with relate to each class label l_j is defined as follows:

$$P(l_j \backslash p_i) = \frac{P(l_j)P(p_i \backslash l_j)}{P(p_i)} \quad (4)$$

$$P(L = l_j) = \frac{N_j}{N} \quad (5)$$

N_j is the amount of values having the label l_j.

$$P(a_k \backslash l_j) = \frac{1+ N_{kj}}{m+ \sum_{k=1}^{m} N_{kj}} \quad (6)$$

N_{kj} is the total frequency of values a_k appearing in individual case in category l_j.

To calculate the average of posterior probability of patient p_i in each class as follows:

$$P_{app} = \frac{1}{n} \sum_{k=0}^{n} P(l_j \backslash p_i) \quad (7)$$

The SSc dataset has three classes time to PF, time to PH and time to death. The aim of the new algorithm is to cluster patients within subgroups using the latent class model whilst predicting the three classes using MLC Naïve Bayes. We have selected all the patients who have at least one comorbidity in order to predict whether the patient might suffer further comorbidities within the following five years. The following Pseudocode explains the steps that are used to build the new algorithm.

Algorithm1 Pseudocode of Latent Model Multi Label Classification

Input: Dataset of Patient Features and Labels.
Output: Clusters of patients and a multi label naïve Bayes model for each group.
Begin
1: Build latent class model for the dataset using EM algorithm and cross validation method.
2: Output *LC* (Patients groups).
3: Assign each test data example to one of the above groups by using scoring formula.
4: Compute the conditional probability between the labels and features.
5: Compute the conditional probability between the labels.
6: Build the Multi label Naïve Bayes from the dataset.
7: Predict the labels by computing the posterior probability for the test data for each group.
8: Compute accuracy and other metrics.
End

D. Results

We now document the comparison of our Latent Class Model Multi Label Naïve Bayes Classifier to other methods when predicting the time to PF, time to PH and time to death for each discovered subgroup. We have run standard Naïve Bayes (NB), Multi Label Naïve Bayes Classification (MLNB), Standard Naïve Bayes with latent class model (LCNB), and Multi label Naïve Bayes Classification with Latent Class model (LCMLNB) in order to predict Time to death, Time to develop PF and time to PH. The following plots show the results of these methods on all of the test data. They show that multi label classification with latent class model performs better than standard Naïve Bayes and Multi label classification. Error rates are generally lower in LCMLNB.

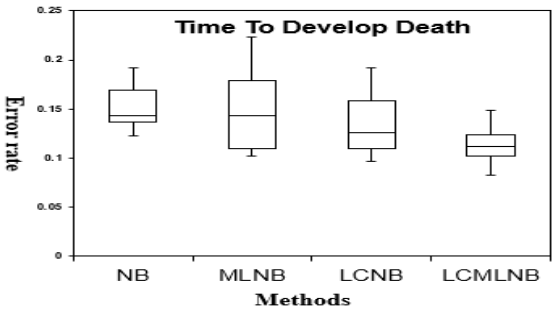

Fig. 1. Comparison between Latent Class Multi Label Classification Model with other methods to predict time to death

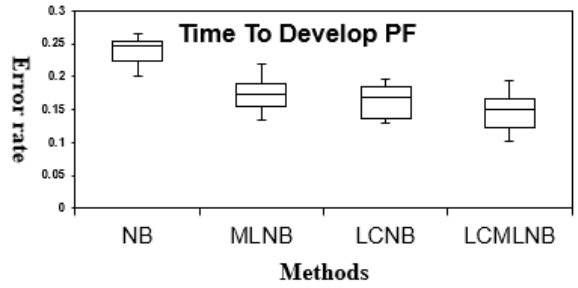

Fig. 2. Comparison between Latent Class Multi Label Classification Model with other methods to predict time to develop PF.

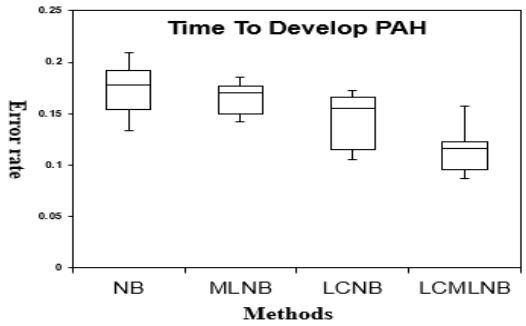

Fig. 3. Comparison between Latent Class Multi Label Classification Model with other methods to predict time to develop PH.

It is necessary to discover the meaning of the latent class model in our dataset. The latent class model in our paper has split the dataset into three groups. All the patients within the same group have similar characters. The following tables show the difference between percentages for blood test results and antibody information within the groups. It shows that most of the patients in group 2 suffer skin thickening and are female. It can be seen from the results that haemoglobin levels (Hb) are lower in group 3 compared to other groups while serum creatinine (Cr) levels are higher. Finally, regarding Lung function, it shows that FVC and DLCO are higher in group2 compared to other groups. All these results can help clinicians to better identify the characteristics of an individual in order to personalise their care plan whilst improving the prediction of their longer term outcomes.

TABLE I. SUBSET AND GENDER VARIABLES DISTRIBUTION AS PERCENTAGE WITHIN GROUPS

	Subset		Gender	
	Skin thickening	Not skin thickening	Males	Females
Group 1	0.47	0.52	0.17	0.82
Group 2	0.98	0.016	0.075	0.925
Group 3	0.27	0.72	0.25	0.74

TABLE II. HB AND CR VARIABLES DISTRIBUTION AS VALUES WITHIN GROUPS

	Hb	Cr
Group 1	12.83	74.37
Group 2	12.98	76.89
Group 3	**10.87**	**257.59**

The following graph shows that the percentage of Group 1 patients who develop both *PH* and *PF* and die within 5 years is higher than for other groups. Also, the percentage of Group 2 and 3 patients who develop *PF* is low compared to those who die or develop *PH*.

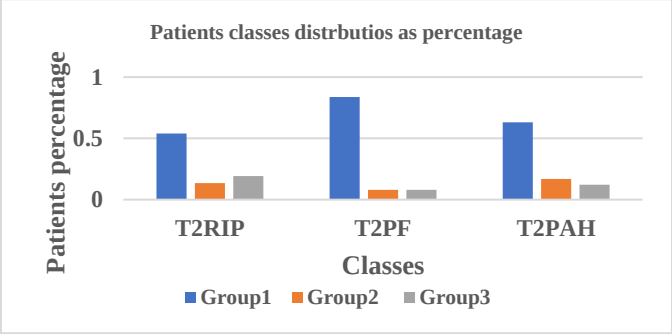

Fig. 4. Patients Classes Distributions within the groups

III. CONCLUSION

In this paper, a set of algorithms were tested on SSc data identifying subgroups of patients and predicting all classes together (time to death, time to develop PAH, and time to develop PF). We have proposed a new algorithm that uses latent class models to cluster patients within three groups having similar characteristics and Naïve Bayes multi-label classification in order to predict the above classes for each group. Our method has improved the classification and also identified patients within meaningful groups. We compared the proposed algorithm with other algorithms including single naïve Bayes and single label classification. It is envisaged that this new model can be used by clinicians to cluster patients and discover key features in each group for classifying more confidently.

REFERENCES

[1] Nareshpalsingh, M.J. and Modi, N.H. (2017). Multi-Label Classification Methods: A Comparative Study. *International Research Journal of Engineering and Technology (IRJET)*, 4(12), pp. 263-270.

[2] Dharmadhikari, S.C., Ingle, M. and Kulkarni, P. (2012). Learning Deep Latent Spaces for Multi Label Classification. *International Journal of Advanced Computer Science and Applications (IJACSA)*, USA.

[3] Prajapati, P., Thakkar, A. and Ganatra, A. (2012). A Survey and Current Research Challenges in Multi-Label Classification Methods. *International Journal of Soft Computing and Engineering (IJSCE)*, 2(1).

[4] Dhongade, P., Longadge, R. and Kapgate, D. (2014). A Review on Classification of Multi-Label Data in Data Mining. *International Journal of Computer Science and Mobile Computing(IJCSMC)*, 3(12), pp. 189-196.

[5] Alessandro, A., Corani, G., Mauá, D. and Gabaglio, S. (2013). An Ensemble of Bayesian Networks for Multilabel Classification. In: *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI)*. Beijing, China: pp. 1220–1225.

[6] Shouman, M., Turner, T.and Stocker, R. (2012). Using Data Mining Techniques in Heart Disease Diagnosis and Treatment. In: *Proceedings of IEEE Japan-Egypt Conference on Electronics Communications and Computers*, vol. 2, pp. 174-177, 2012.

[7] Vembandasamy, K., Sasipriya, R. and Deepa, E. (2015). Heart Diseases Detection Using Naive Bayes Algorithm. *International Journal of Innovative Science, Engineering & Technology*, 2 (9), pp. 441-444.

[8] Kabir, M.F., et al. (2011). Enhanced Classification Accuracy on Naive Bayes Data Mining Models. *International Journal of Computer Applications*, 3(2).

[9] Eshghi, M., Haughton, D., Legrand, P., Skaletsky, M. and Woolford, S. (2011). Identifying Groups: A Comparison of Methodologies. *Journal of Data Science*, 9, pp. 271–291.

[10] Rindskopf, D. and Rindskopf, W. (1986). The Value of Latent Class Analysis in Medical Diagnosis. *Stat Med*, 5, pp.21-27.

[11] Zhang, N. L. (2004). Hierarchical Latent Class Models for Cluster Analysis. *Journal of Machine Learning Research*, 5, pp.697-723.

[12] Dey, A., Chakraborty, A., Majumdar, K. and Mandel, A. (2016). Application of Latent Class Analysis to Estimate Susceptibility to Adverse Health Outcomes based on Several Risk Factors. *International Journal of Community Medicine and Public Health*, 3(12), pp. 3423-9.

[13] Tsoumakas, G., Zhang, M. L. and Zhou, Z. H. (2009). Tutorial on Learning from Multi-Label Data. In: *Proceedings of ECML PKDD'09 Conference*. Bled, Slovenia.

[14] CEYLAN, Z. and PEKEL, E. (2017). Comparison of Multi-Label Classification Methods for Prediagnosis of Cervical Cancer. *International Journal of Intelligent Systems and Applications in Engineering*, 5(4), pp. 232-236.

[15] Kasper, D.L., Fauci, A.S., Hauser, S.L., Longo, D.L., Jameson, J.L. and Loscalzo, J. (2015). *Harrison's Principles of Internal Medicine*. McGraw-Hill Education, 19th edition.

[16] Van den Hoogen, F. et al. 2013 Classification Criteria for Systemic Sclerosis: An American College of Rheumatology/European League Against Rheumatism Collaborative Initiative. *Arthritis and Rheumotology*, 65(11), 2013.