Probabilistic Multi-Label Learning for Medical Data

Damien Zufferey
AiSlab Group, Institute of Information Systems, University of Applied Sciences and Arts Western Switzerland
Email: damien.zufferey@hevs.ch
DIVA Group, Department of Informatics, University of Fribourg, Switzerland

***Abstract*—We report on a probabilistic approach for the classification of chronically ill patients. We rely on multi-label learning for its ability to represent in a natural way classification problems involving coexistence of diseases. We use a public clinical database for the evaluation of our proposed algorithm. Preliminary results show the benefits of our approach.**

I. Introduction

Multi-label learning (MLL) is a growing research topic that has received, in last few years, significant contributions from machine learning community [1]. MLL differs from classical machine learning by tackling the learning problem from a different perspective which looks like natural for many problems of the real life, such as this application in the medical domain: prediction of gene function [2]. In our case, we are interested in applying MLL on clinical data for the identification of chronic diseases. This research is motivated by the problem of classifying patients affected by multiple co-morbidities to enhance decision support for physicians. We proposed an algorithm [3] based on bag of words (BoW) and supervised dimensionality reduction methods for the classification of chronically ill patients. We here extend our work following a probabilistic approach as described in the Section IV. For the evaluation of our work, the public-access intensive care unit database (MIMIC-II) [4] has been used.

II. Background

Let X be the domain of observations and L be the finite set of labels. Given a training set $T = \{(x_1, Y_1), (x_2, Y_2), ..., (x_n, Y_n)\}$ $(x_i \in X, Y_i \subseteq L)$ i.i.d. drawn from an unknown distribution D, the goal is to learn a multi-label classifier $h : X \rightarrow 2^L$. However, it is often more convenient to learn a real-valued scoring function of the form $f : X \times L \rightarrow \mathbb{R}$. Given an instance x_i and its associated label set Y_i, a working system will attempt to produce larger values for labels in Y_i than those not in Y_i, i.e. $f(x_i, y_1) > f(x_i, y_2)$ for any $y_1 \in Y_i$ and $y_2 \notin Y_i$. By the use of this function $f(\cdot, \cdot)$, we can obtain a multi-label classifier: $h(x_i) = \{y | f(x_i, y) > \delta,\ y \in L\}$, where δ is a threshold to infer from the training set. The function $f(\cdot, \cdot)$ can also be adapted to a ranking function $rank_f(\cdot, \cdot)$, which maps the outputs of $f(x_i, y)$ for any $y \in L$ to $\{1, 2, ..., |L|\}$ such that if $f(x_i, y_1) > f(x_i, y_2)$ then $rank_f(x_i, y_1) < rank_f(x_i, y_2)$.

III. Data Set

The MIMIC-II clinical database [4] is publicly and freely available after registration. The last release of the database contains around 33,000 patients. We choose to skip the neonates and the children in order to concentrate only on the adult population ($\geq$ 16 years old) which consists of around 24,000 patients, where we extracted a subset of 19,773 patients with chronic diseases. Regarding the restriction to the adult population, we motivate this decision by the divergence which exists between these two groups in term of medical conditions and treatment plans. The average age of the patients in the database is 67 years old. The distribution of the population is: 56% of man and 44% of women. The clinical data we consider are the laboratory tests and the items registered in the chart. By chart, we mean a logbook per patient which records the results of heterogeneous examinations, such as: fluid assessment, physiological measure, or severity score which evaluates vital functions. According to the length of the stay, a patient will make several laboratory tests and various examinations. Thus, clinical data of patients are time series. In order to attenuate the amount of missing values, we take a subset of items, from the laboratory tests and from the chart, that are present at least for 80% of the patients. We end up with 76 items from the laboratory test and from the chart.

As labels we consider 10 chronic diseases where their distributions amongst the 19,773 extracted patients are presented in the Table I. We use the coding scheme of the International Classification of Disease revision 9 (ICD-9)[1] available in the MIMIC-II database for building the 10 chronic diseases.

Label / Chronic disease	No. of patients	%
Hypertensive disease	12,309	62.3%
Fluid electrolyte disease	6,177	31.2%
Diabetes mellitus	6,056	30.6%
Lipoid metabolism disease	5,965	30.2%
Kidney disease	5,828	29.5%
COPD	4,253	21.5%
Thyroid disease	2,246	11.4%
Hypotension	1,962	9.9%
Liver disease	1,088	5.5%
Thrombosis	931	4.7%

TABLE I. Distribution of labels / chronic diseases in the 19,773 extracted patients of the MIMIC-II database.

IV. Method

A. Feature extraction

Laboratory events and chart events of each patient are summarized into one feature vector. Due to the heterogeneity and the different frequencies of the selected medical data, we propose the following approach for the feature extraction according to the type of the measured values:

[1] http://www.who.int/classifications/icd

1) Numerical values: consist of measured values such as blood pressure, creatinine and temperature. When they appear one time, such as the height at the patient admission, they are taken in the feature vector as they are. When they appear several times, the following summary features are computed: mean, median, standard deviation and range.

2) Categorical values: consist of observed values such as cardiovascular function assessment score and urine color. For a patient which did several times a particular examination where results are discrete values which can be divided into mutually exclusive classes, we can represent this information as an histogram. Then, the relative frequency of each category of the histogram is used as feature. There is also the case where only one observation exists for each patient, such as the gender at the patient admission, in that case, we encode as feature the value in a binary variable.

B. Model

We propose a generative model for MLL using Gaussian Mixture Model (GMM) [5] as the based classifier, called ML-GMM hereafter. We use a combination of Label Power-set (LP) and Binary Relevance (BR) as the transformation methods [1]. In particular, we apply the LP method, which considers all possible combinations between labels to transform the MLL problem to a multi-class formulation which can be naturally solved using GMMs. To handle the intractability of LP, we consider, according to a predefined constant n, only combinations of labels that have at least n observations in the training set. In contrast, combinations of labels that have less than n observations are grouped together to form the class *"other"*. Then, in the case of the class *"other"* is relevant, we apply the BR method, which considers each label independently as relevant or not, to transform the MLL problem to a set of binary problems, according to the number of labels. The manner how ML-GMM combines LP and BR allows handling the MLL problem efficiently and at the same time preserving dependency information between labels.

V. Experiment

In the classical learning approach of multiclass problems, the evaluation is done through common metrics such as accuracy, precision, and recall. In multi-label problems, the evaluation is much more complicated and needs extended evaluation metrics. One of the commonly used evaluation metrics is the Hamming loss [6], which is described below.

Let a testing set $S = \{(x_1, Y_1), (x_2, Y_2), ..., (x_m, Y_m)\}$.

Hamming loss evaluates how many times an observation-label pair is misclassified. The score lies between 0 and 1, where 0 corresponds to the best result:

$$hloss_S(h) = \frac{1}{m}\sum_{i=1}^{m}\frac{|h(x_i)\triangle Y_i|}{|L|}, \quad (1)$$

where $\triangle$ represents the symmetric difference.

To evaluate our proposed algorithm (ML-GMM), we divided the MIMIC-II dataset containing 19,773 patients into 3 subsets of 6,591 patients. We used the first two subsets (training and validation) in a grid-search for finding optimal parameters for our algorithm. Finally, using the third subset (testing), we computed the reported results, as presented in the Figure 1. We can see that the Hamming loss is reducing when we allow the algorithm to handle a larger combination of classes using the LP method. The best performance is achieved when considering 14 classes in LP method. Note that the algorithm is purely BR when the number of classes is 0.

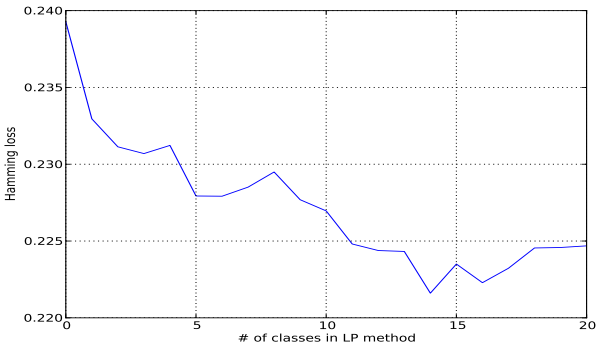

Fig. 1. Results in term of Hamming loss of our ML-GMM algorithm on the MIMIC-II database.

VI. Conclusion

In this abstract, we proposed an approach for a MLL-based algorithm for the classification of chronically ill patients. Our solution elegantly combines the LP method for its ability to consider correlations between labels, and the BR method for its ability to scale well with a large number of labels. For future work, we will conduct additional experiments to evaluate our algorithm by considering additional evaluation metrics and to compare with existing state-of-the-art multi-label algorithms.

References

[1] G. Madjarov, D. Kocev, D. Gjorgjevikj, and S. Džeroski, "An extensive experimental comparison of methods for multi-label learning," *Pattern Recognition*, vol. 45, no. 9, pp. 3084 – 3104, 2012, best Papers of Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA'2011). [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320312001203

[2] Z. Barutcuoglu, R. E. Schapire, and O. G. Troyanskaya, "Hierarchical multi-label prediction of gene function," *Bioinformatics*, vol. 22, no. 7, pp. 830–836, 2006. [Online]. Available: http://bioinformatics.oxfordjournals.org/content/22/7/830.abstract

[3] S. Bromuri, D. Zufferey, J. Hennebert, and M. Schumacher, "Multi-label classification of chronically ill patients with bag of words and supervised dimensionality reduction algorithms," *Journal of Biomedical Informatics*, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1532046414001270

[4] M. Saeed, M. Villarroel, A. T. Reisner, G. Clifford, L.-W. Lehman, G. Moody, T. Heldt, T. H. Kyaw, B. Moody, and R. G. Mark, "Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): A public-access intensive care unit database," *Critical Care Medicine*, vol. 39, pp. 952–960, May 2011.

[5] C. M. Bishop *et al.*, *Pattern recognition and machine learning*. springer New York, vol. 1.

[6] M.-L. Zhang and Z.-H. Zhou, "ML-KNN: A lazy learning approach to multi-label learning," *Pattern Recognition*, vol. 40, no. 7, pp. 2038 – 2048, 2007.