Multi-Label Classification of Chronically Ill Patients with Bag of Words and Supervised Dimensionality Reduction Algorithms

Stefano Bromuri[a,*], Damien Zufferey[a], Jean Hennebert[b], Michael Schumacher[a]

[a] *University of Applied Sciences Western Switzerland*
Institute of Business Information Systems
TechnoArk 3,
CH-3960, Sierre, Switzerland
[b] *University of Applied Sciences Western Switzerland,*
Institute of Information and Communication Technologies,
Bd de Pérolles 80, CH-1705, Fribourg, Switzerland

Abstract

Objective:

This research is motivated by the issue of classifying illnesses of chronically ill patients for decision support in clinical settings. Our main objective is to propose multi-label classification of multivariate time series contained in medical records of chronically ill patients, by means of quantization methods, such as bag of words (BoW), and multi-label classification algorithms. Our second objective is to compare supervised dimensionality reduction techniques to state-of-the-art multi-label classification algorithms. The hypothesis is that kernel methods and locality preserving projections make such algorithms good candidates to study multi-label medical time series.

Methods: We combine BoW and supervised dimensionality reduction al-

[*] Corresponding Author
Email address: stefano.bromuri@hevs.ch (Stefano Bromuri)

gorithms to perform multi-label classification on health records of chronically ill patients. The considered algorithms are compared with state-of-the-art multi-label classifiers in two real world datasets. Portavita dataset contains 525 diabetes type 2 (DT2) patients, with co-morbidities of DT2 such as hypertension, dyslipidemia, and microvascular or macrovascular issues. MIMIC II dataset contains 2635 patients affected by thyroid disease, diabetes mellitus, lipoid metabolism disease, fluid electrolyte disease, hypertensive disease, thrombosis, hypotension, chronic obstructive pulmonary disease (COPD), liver disease and kidney disease. The algorithms are evaluated using multi-label evaluation metrics such as *hamming loss*, *one error*, *coverage*, *ranking loss*, and *average precision.*

Results: Non-linear dimensionality reduction approaches behave well on medical time series quantized using the BoW algorithm, with results comparable to state-of-the-art multi-label classification algorithms. Chaining the projected features has a positive impact on the performance of the algorithm with respect to pure binary relevance approaches.

Conclusions: The evaluation highlights the feasibility of representing medical health records using the BoW for multi-label classification tasks. The study also highlights that dimensionality reduction algorithms based on kernel methods, locality preserving projections or both are good candidates to deal with multi-label classification tasks in medical time series with many missing values and high label density.

Keywords:

Multi-Label Classification, Complex Patient, Diabetes Type 2, Clinical Data, Dimensionality Reduction, Kernel Methods.

1. Introduction

The average lifespan has increased considerably due to the invention of better drugs and improvement of healthcare, but the rate of chronic illnesses per patient has also increased, becoming a burden for the economy of industrialized and emerging countries [1].

The interaction between chronic illnesses and multiple drugs intake make the patient treatment complex to handle for caregivers. The possibility of taking informed decisions about complex patients is important to slow down the development of their illnesses.

Unfortunately, doctors have to take decisions whose consequences will be evaluated only after years of treatment. Furthermore, given the growth in number of chronically ill patients, caregivers are often in charge of hundreds of patients [2]. In addition, patient electronic health records (EHR) often contain the evolution in time of the patient clinical data, which are high dimensional multivariate time series of physiological values.

As reported in [3], physicians would use services that improve their understanding of an illness even if these involve more cognitive effort than in the standard practice. In particular, in the medical informatics and data mining community [4, 5] it has already been discussed that classifying patients given their physiological values and laboratory tests may help caregivers' decision making process.

This paper is motivated by the problem of classifying patients affected by multiple illnesses to enhance the decision support of medical doctors. There are two challenges to overcome in order to define a system capable to correctly classify the multiple illnesses that may affect a chronically ill patient: a) dealing with irregular multivariate time series; b) dealing with the interaction of multiple co-morbidities in a heterogeneous population of patients.

The presence of high dimensional and multivariate data presents a big challenge to standard classification algorithms due to the curse of dimensionality [6]. Clinical time series are often irregular, a patient may present different number of records with respect to another patient and the periods of time in which the values are collected may not be aligned. The challenge is even more difficult if we consider the inherently multi-label properties of medical data, where a patient may present multiple co-morbidities at once.

Concerning irregular time series, quantization algorithms, such as the Bag of Words (BoW) model, have proven successful in several medical tasks [7].

As a matter of fact, BoW is often used in biomedical time series. In [7], Wang et al. present an application of the BoW model to EEG and ECG signals. Similarly to us, the authors of [7] are faced with the issue of time series of different length with possibly heterogeneous patients at hand.

Jiu et al. present a supervised approach towards BoW codebook generation using neural networks in [8]. In particular, the approach uses Multi-Layer Perceptrons (MLP) and the backpropagation algorithm to update the weights of the codewords according to their discrimination capabilities with respect to a set of classes.

Similarly to [8], in [9] Ordonez et al. present a modification of the BoW model to classify medical time series. Such a model uses continuous multivariate time series to compute a symbolic representation of the signals that is then used as the codebook for the classification of the patients.

Concerning multi-label classification algorithms, an extensive review can be found in [10]. Multi-label learning [11] implies training sets where each instance has a labelset and the task is to predict the labelset of unseen instances. As reported in [10], there exist works that combine supervised dimensionality reduction with multi-label learning [12, 13, 14]. Furthermore, most of these works focus on applying multi-label techniques on text analysis with static datasets [15].

In general terms multi-label classification of complex patients in discrete medical time series is quite an unexplored issue. Firstly, we think that the main contribution of this paper is to propose the combination of BoW, to quantize irregular time series present in patient health records, and multi-label classification algorithms, to classify the chronic illnesses that a patient may present. These are two established techniques, but in medical settings their combination is quite novel.

Secondly, we believe that this contribution is interesting to biomedical informatics as we evaluate linear and non linear supervised dimensionality reduction approaches with respect to multi-label classification in medical time series, and we compare these approaches with state-of-the-art multi-label classification algorithms. In doing this, we aim at identifying the most effective supervised dimensionality reduction techniques with respect to medical time series. We aim to confirm the hypothesis that, given the nature of the data at hand, non-linear supervised dimensionality reduction algorithms have a behaviour comparable to state of the art multi-label classifiers.

Thirdly, our contribution is also of interest to biomedical research because we perform our evaluation against two real world medical datasets:

the Portavita dataset, provided for this study by the Portavita company[1], containing 525 diabetic patients presenting, sometimes simultaneously, hypertension, dyslipidemia or microvascular and macrovascular complications of diabetes type 2 (DT2) [16]; an extraction of 2635 patients from the public MIMIC II database [17], where we consider patients affected simultaneously by thyroid disease, diabetes mellitus, lipoid metabolism disease, fluid electrolyte disease, hypertensive disease, thrombosis, hypotension, chronic obstructive pulmonary disease (COPD), liver disease and kidney disease.

The rest of this paper is organized as follows: Section 2 presents a background on multi-label classification, kernel methods, and supervised dimensionality reduction algorithms; Section 3 presents the Portavita and MIMIC II datasets and their properties; Section 4 presents the training schema for the attempted multi-label classification algorithms; Section 5 presents an evaluation for the multi-label algorithms considered in this paper; finally, Section 6 concludes this paper and draws the lines for future work.

2. Background

In this Section we present the concepts of multi-label classification, kernels, locality preserving projections and multi-class Fisher discriminant analysis. In Section 4 we show how we combined these concepts in a system for classification of multi-label chronically ill patients.

[1] www.portavita.eu

2.1. Multi-Label Classification

Let X be the domain of observations and let 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 $h : X \to 2^L$. However, it is often more convenient to learn a real-valued scoring function of the form $f : X \times L \to \mathbb{R}$. 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 that are 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 the 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 $rank_f(x_i, y_1) < rank_f(x_i, y_2)$.

Furthermore, there exist several approaches to train multi-label classifiers (see [10] for a comprehensive review on the subject). The simplest approach, known as **binary relevance** (BR), is to train one binary classifier per label with traditional classification algorithms, considering each label as a separated problem. BR has the disadvantage of not taking into consideration the relationships existing amongst labels. To overcome this issue, several *ensemble methods* have been defined in the past, amongst which the most popular ones are **classifier chains** (CC) and **label powersets** approaches (LP). CC methods work by recursively training classifiers with the label predicted by the previous classifier as new features. LP methods focus on training classifiers defining classes by means of subsets of the labelset. Despite having been demonstrated effective, CC and LP methods present computational disadvantages with respect to BR methods, whose complexity is linear in respect to the number of labels. In addition, CC methods are difficult to train with classifier presenting many parameters, as each classifier in the chain needs to be optimized differently. Secondly, LP method are computationally infeasible due to the large number of possible classes in a labelset.

Other approaches use the probabilistic distribution of the labels and their dependencies within a neighbourhood to tune the classifier output. MLkNN [18] is a successful example of such a method.

Within this paper we will show the effect of using dimensionality reduction algorithms with a BR approach, considering the output of each classifier as separated, or as CC classifiers by concatenating the projected features.

2.2. Kernels

Non-linear subspaces may be suitable to describe clinical datasets as due to their high dimensionality they may lie in complex manifolds. Therefore, we may need to map our input data in terms of clinical datasets to a higher dimensional space using a linearization function. If we consider a set of m samples $x_1, x_2, \ldots, x_m \in \mathbb{R}^n$, belonging to c classes, then we can consider a non linear mapping $\phi : \mathbb{R}^n \to \mathcal{F}$, where we choose ϕ so that $\langle \phi(x_i), \phi(x_j) \rangle = K(x_i, x_j)$, where $K(.,.)$ is a positive semi-definite kernel function.

Performing this map explicitly can be computationally expensive, to avoid it we can apply the *Kernel Trick* [19], and calculate the Gram matrix $K(.,.)$, containing the inner product between the input vectors in the linearization space. This then allows us to modify linear techniques using the inner product with appropriate kernel functions, opening up the possibility of applying well known approaches in non-linear spaces.

Within this paper we will use the RBF kernel and the *histogram intersection kernel* [20]. The RBF kernel is defined as:

$$K(x, y) = \frac{exp^{-\|x-y\|^2}}{2\sigma^2} \tag{1}$$

The histogram intersection kernel can be defined starting from two histograms x and y consisting both of m features. We denote the ith features of x as x_i and for y as y_i. Then we can define the kernel as:

$$K(x, y) = \sum_{i}^{m} min(x_i, y_i) \tag{2}$$

A big advantage of this kernel is that it is parameterless.

2.3. Locality Preserving Projections

As explained in [21], a LPP projection is a linear transformation for which the data residing in a space $\mathbb{R}^n$ are mapped in a subspace $\mathbb{R}^r$, with $r < n$, such that nearby data pairs in the original n-dimensional space are also close in the identified subspace. More formally, if we consider a square matrix $A \in \mathbb{R}^{d \times d}$, where $A_{i,j} \in [0, 1]$, representing the affinity between the elements x_i and x_j in a dataset with d elements, the T_{LPP} transformation matrix can be defined as follows:

$$T_{LPP} = \underset{T \in \mathbb{R}^{d \times r}}{\arg\min} \left(\frac{1}{2} \sum_{i,j=1}^{n} A_{i,j} \|T^T x_i - T^T x_j\|^2 \right) \tag{3}$$

Within this paper we are interested in the usage of such a projection within the KLFDA technique, further details on how to calculate T_{LPP} can be found in [21].

2.4. Linear Discriminant Analysis and Local Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) [22] is a widely used supervised dimensionality reduction technique that can find the linear transformation which best separates elements of different classes. To achieve this, LDA makes use of the within-class scatter matrix $S^{(w)}$ and of the between-class scatter matrix $S^{(b)}$. These can be defined as:

$$S^{(w)} = \sum_{i=1}^{C} \sum_{x \in E_i} (x - \mu_i)(x - \mu_i)^T \tag{4}$$

where μ_i is the mean of class E_i, and

$$S^{(b)} = \sum_{i=1}^{C} N_i (\mu_i - \mu)(\mu_i - \mu)^T \tag{5}$$

where μ is the global mean and N_i is the number of elements belonging to class E_i. $S^{(w)}$ is a measure of the variance between the elements belonging to the same class, while $S^{(b)}$ is a measure of the variance of the elements belonging to different classes. Ideally, we want the scatter to be minimized for elements of the same class and maximized for elements of different classes. The transformation matrix T_{LDA} that achieves this is defined as:

$$T_{LDA} = \arg\max \frac{det(T^T S^{(w)} T)}{det(T^T S^{(b)} T)} \tag{6}$$

As explained in [23], to specify a Locality Sensitive LDA (LSDA), we can define the local within-class scatter matrix $\tilde{S}^{(w)}$ and the local between class scatter matrix $\tilde{S}^{(b)}$

$$\tilde{S}^{(w)} = \frac{1}{2} \sum_{i,j=1}^{n} \tilde{W}_{i,j}^{(w)} (x_i - x_j)(x_i - x_j)^T \tag{7}$$

$$\tilde{S}^{(b)} = \frac{1}{2} \sum_{i,j=1}^{n} \tilde{W}_{i,j}^{(b)} (x_i - x_j)(x_i - x_j)^T \tag{8}$$

where

$$\tilde{W}_{i,j}^{(w)} = \begin{cases} A_{i,j}/N_i & \text{if } y_i = y_j = c, \\ 0 & \text{if } y_i \neq y_j \end{cases} \tag{9}$$

$$\tilde{W}_{i,j}^{(b)} = \begin{cases} A_{i,j}(1/N - 1/N_i) & \text{if } y_i = y_j = c, \\ 1/N & \text{if } y_i \neq y_j \end{cases} \tag{10}$$

which implies that we are weighting the pairwise values according to their affinity matrix $A_{i,j} \in [0,1]$, with $A_{i,j}$ closer to 1 if x_j is close to x_i and to 0 if they are far apart.

Then, the objective function can be expressed again as a generalized eigenvalue problem:

$$T_{LSDA} = \underset{T \in \mathbb{R}^{d \times r}}{\arg\max} \left[tr((T^T \tilde{S}^{(w)} T)^{-1} T^T \tilde{S}^{(b)} T)) \right] \tag{11}$$

we refer the interested reader to [23], for further details on how to compute LSDA.

2.5. Kernel Local Discriminant Analysis

KLFDA [23] is a generalization of the previously presented LSDA using kernel functions. If we consider $\tilde{S}_b^{\phi}, \tilde{S}_w^{\phi}$ and $\tilde{S}_t^{\phi}$ as the local between-class, within-class and total scatter matrices respectively in the space identified by a kernel mapping, then KLFDA seeks to find:

$$T_{opt} = \arg\max \frac{T^T \tilde{S}_b^{\phi} T}{T^T \tilde{S}_w^{\phi} T} = \arg\max \frac{T^T \tilde{S}_b^{\phi} T}{T^T \tilde{S}_t^{\phi} T} \tag{12}$$

We can justify the use of supervised techniques based on LPP and kernel methods with the considerations in [21], for which LPP is particularly useful in applications where by preserving the structure of the neighbourhood in the lower dimensional space, nearest neighbour based approaches can still perform well, and the curse of dimensionality is mitigated. Kernel methods are useful in cases where the classes are non-linearly separable. In our case, we apply the version of KLFDA specified in [23] using regularization.

3. Materials

In this Section, we present the descriptive statistics of the two datasets taken into consideration. For multi-label datasets, amongst the descriptive statistics it is important to also consider *label cardinality* and *label density.* Given a dataset D, and a set of labels L, where the labels of an example are denoted with Y_i we can define label cardinality and label density as below.

Label Cardinality: Label cardinality of a dataset D is the average number of labels of the examples in D:

$$LC(D) = \frac{1}{|D|} \sum_{i=1}^{|D|} |Y_i| \tag{13}$$

Label Density: Label density of D is the average number of labels of the examples in D divided by $|L|$

$$LD(D) = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|Y_i|}{|L|} \tag{14}$$

Label cardinality quantifies the average number of alternative labels that characterize the examples in the dataset. With respect to label cardinality, label density also considers the number of labels. The two metrics are important because multi-label algorithms may present a different behaviour in datasets with similar cardinality, but different density.

3.1. The Portavita Dataset

The Portavita dataset is a medical dataset collected during the standard care of DT2 patients. Such a dataset includes 525 diabetic patients affected by four complications which are: hypertension, dyslipidemia, microvascular and macrovascular diseases. A summary showing the distribution of the labels amongst the patients in such a dataset is shown in Table 1.

Label	Number of Patients	%
Hypertension	280	53.33%
Dyslipidemia	280	53.33%
Microvascular	280	53.33%
Macrovascular	280	53.33%

Table 1: Distribution of Labels in the Portavita Dataset.

The Portavita dataset presents a label cardinality of 2.13, a label density of 0.532, with a total of 15 possible symbols (combination of co-occurring labels), all occurring in the dataset. All patients have multiple health records (> 3), for an average number of records per patient equal to 6.72 and a total number of records equal to 3528, comprising a set of common laboratory tests and physical examinations that are part of normal routine tests in DT2. Table 2 gives a summary of the descriptive statistics of such laboratory tests.

Depending on the stability of the diabetic patient physiological values, the data may be collected once every six months, or once every three months, to check for the presence of microvascular or macrovascular complications.

Test Name	Frequency	MEAN	MEDIAN	MIN	MAX	SD	%Missing values
BMI (kg/cm^2)	each visit	29.79	29.00	16.00	49.00	5.20	0.00%
Body Weight (g)	each visit	85.17	84.00	35.00	189.00	16.99	0.00%
Heart-rate (bpm)	3/6 months	72.83	72.00	60.00	360.00	11.26	35.44%
Height (cm)	once	169.04	169.00	130.00	270.00	9.95	52.33%
Abdominal circumference (cm)	3/6 months	103.03	102.00	75.00	189.00	13.57	84.17%
Diastolic blood pressure (mmHg)	3/6 months	78.24	80.00	60.00	180.00	9.81	7.61%
Systolic blood pressure (mmHg)	3/6 months	138.07	138.00	100.00	270.00	17.69	7.90%
HDL-cholesterol (mg/dL)	3/6 months	1.22	1.20	0.09	9.30	0.34	2.58%
LDL-cholesterol (mg/dL)	3/6 months	2.72	2.60	0.30	17.20	0.98	5.09%
HbA1c (mg/dL)	6 months	52.29	50.00	5.30	180.00	12.91	16.27%
Total Chol/HDL-Chol ratio	3/6 months	4.15	3.90	1.30	27.30	1.35	4.05%
Albumine/Creatinine ratio	3/6 months	5.05	1.10	0.40	939.20	18.56	49.04%
Aspartate Transaminase (IU/L)	3/6 months	33.60	24.00	5.00	2985.00	71.38	91.52%
Natrium (mmol/L)	3/6 months	139.73	140.00	116.00	165.00	2.84	60.26%
Kalium (mmol/L)	3/6 months	4.35	4.30	2.20	7.80	0.47	39.74%
Creatinine in Urine (mg/dL)	3/6 months	79.90	75.00	8.00	1085.00	29.52	18.19%
Albumine in Urine (mg/dL)	3/6 months	30.81	8.00	0.60	1000.00	82.54	46.27%
Hemoglobin (g/dl)	3/6 months	8.40	8.50	2.90	12.30	1.07	80.82%
Fasting Glucose (mmol/L)	3 months	7.61	7.20	1.30	42.60	2.15	14.48%
Alanine Transaminase (IU/L)	3/6 months	32.00	25.00	4.00	2510.00	42.81	76.53%
GammaGT (IU/L)	3/6 months	61.99	35.00	4.00	2855.00	104.82	87.24%
Creatinine Kinase (IU/L)	3/6 months	128.42	91.00	8.00	5750.00	176.33	92.46%
Total Cholesterol (mmol/L)	3/6 months	4.80	4.70	1.30	33.30	1.16	2.55%
Triglyceride (mmol/L)	3/6 months	1.94	1.62	0.12	87.05	1.49	2.84%
Cockcroft (mL/min)	variable	90.35	84.00	6.00	4356.00	94.03	53.32%
Glucose after Meal (mmol/L)	3 months	9.03	8.20	1.20	61.00	3.68	90.68%
Modification of Diet in Renal Disease (mL/min)	variable	85.02	84.00	3.00	974.00	26.29	33.29%

Table 2: Portavita Dataset Descriptive Statistics.

As this is a real world dataset, the presence of a label may simply point towards a suspected issues, requiring further laboratory tests before it can be confirmed. In other cases, the label is assigned at the beginning of the treatment, and then it is never removed even if the patient does not present the complication any more.

We can calculate that the prior probability for a patient to present a label to be 53.33% for each label, which represents a base average precision to compare against with the attempted classifiers. In the Portavita dataset, the tests are performed with a frequency of 3/6 months for most of the features, which are conducted at the same time for each patient, and consequently, it is quite easy to produce a set of vectors and to go from the relational model to the multivariate time series associated to a patient for this dataset.

3.2. MIMIC II Dataset

As a second dataset for our study, we decided to use an extraction of 2635 patients from the MIMIC II database. Since MIMIC II is a large database, we decided to select patients that had more than 40 records. Our selection has an average of 60.39 records and a total of 159 127 records. The chronic illnesses and number of patients per illness in MIMIC II dataset are shown in Table 3.

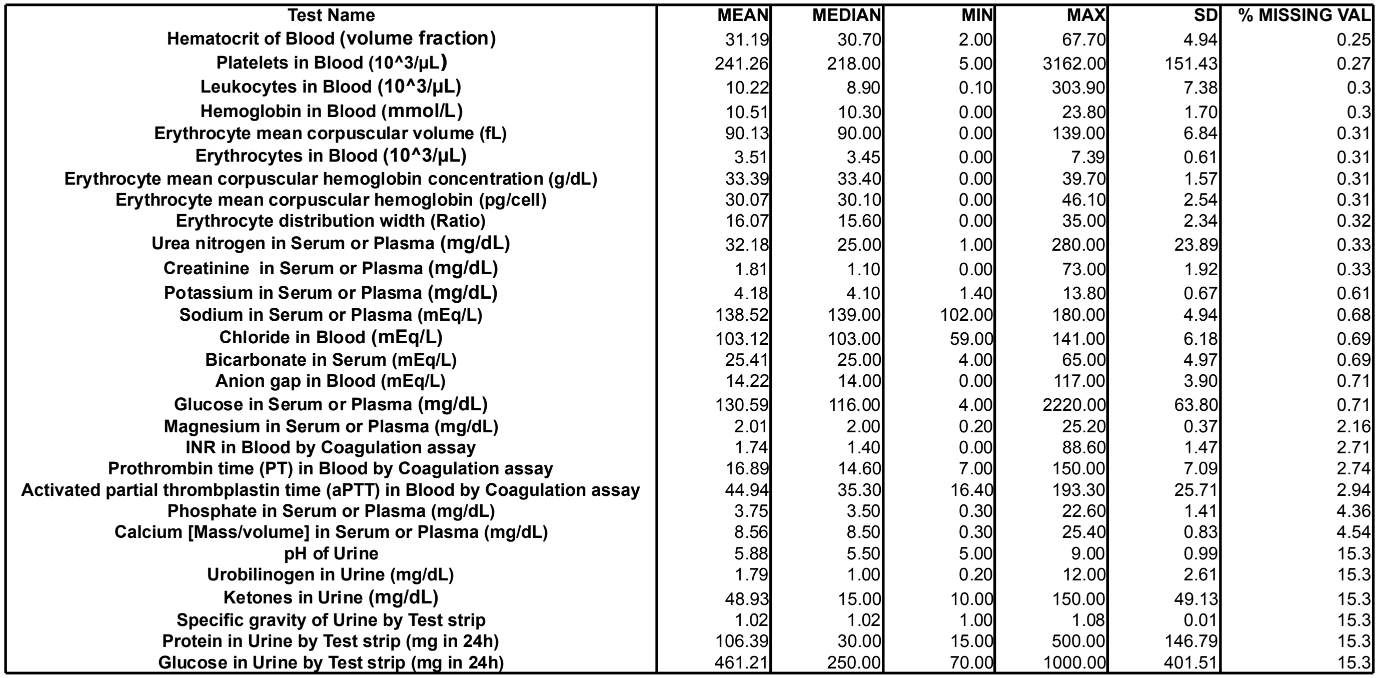
The MIMIC II dataset has a label cardinality of 2.54 and a label density of 0.254, with 1023 possible symbols, of which 194 are present in the dataset. The patients of MIMIC II are very different from those of Portavita, as MIMIC II is focused on intensive care patients, while Portavita's patients are standard care patients. This also implies that there are more laboratory tests collected per patient in MIMIC II than in Portavita.

Label	Number of Patients	%
Thyroid disease	297	11.2%
Diabetes mellitus	875	33.2%
Lipoid metabolism disease	671	25.4%
Fluid electrolyte disease	1014	38.4%
Hypertensive disease	1568	59.5%
Thrombosis	180	6.8%
Hypotension	294	11.1%
COPD	573	21.7%
Liver Disease	208	7.8%
Kidney Disease	1013	38.4%

Table 3: Distribution of Labels in the MIMIC II Dataset.

In MIMIC II the frequencies of the laboratory tests depend on the gravity of the patient and not on a treatment. We transformed the patients' records in multivariate time series by taking the sample frequency of the most frequent laboratory tests for each patient (for example, glucose in serum) and we applied a last observation carried forward (LOCF) to the less frequent measurements considering them as constant between two measurements. We are aware that LOCF underestimates the variability of the data. Our simplifying assumption in applying LOCF is that if the variability between measurements of such values was not crucial for the caregivers of the intensive care units in the first place, then it is acceptable to underestimate variability in our classification analysis. Validating approaches to handle data sampled with different frequencies is an interesting problem that we cannot exhaust within a single contribution, and therefore will be subject of future work.

For those data that are completely missing, we applied the imputation approach explained in the next Section.



Test Name	MEAN	MEDIAN	MIN	MAX	SD	% MISSING VAL
Hematocrit of Blood (volume fraction)	31.19	30.70	2.00	67.70	4.94	0.25
Platelets in Blood (10^3/µL)	241.26	218.00	5.00	3162.00	151.43	0.27
Leukocytes in Blood (10^3/µL)	10.22	8.90	0.10	303.90	7.38	0.3
Hemoglobin in Blood (mmol/L)	10.51	10.30	0.00	23.80	1.70	0.3
Erythrocyte mean corpuscular volume (fL)	90.13	90.00	0.00	139.00	6.84	0.31
Erythrocytes in Blood (10^3/µL)	3.51	3.45	0.00	7.39	0.61	0.31
Erythrocyte mean corpuscular hemoglobin concentration (g/dL)	33.39	33.40	0.00	39.70	1.57	0.31
Erythrocyte mean corpuscular hemoglobin (pg/cell)	30.07	30.10	0.00	46.10	2.54	0.31
Erythrocyte distribution width (Ratio)	16.07	15.60	0.00	35.00	2.34	0.32
Urea nitrogen in Serum or Plasma (mg/dL)	32.18	25.00	1.00	280.00	23.89	0.33
Creatinine in Serum or Plasma (mg/dL)	1.81	1.10	0.00	73.00	1.92	0.33
Potassium in Serum or Plasma (mg/dL)	4.18	4.10	1.40	13.80	0.67	0.61
Sodium in Serum or Plasma (mEq/L)	138.52	139.00	102.00	180.00	4.94	0.68
Chloride in Blood (mEq/L)	103.12	103.00	59.00	141.00	6.18	0.69
Bicarbonate in Serum (mEq/L)	25.41	25.00	4.00	65.00	4.97	0.69
Anion gap in Blood (mEq/L)	14.22	14.00	0.00	117.00	3.90	0.71
Glucose in Serum or Plasma (mg/dL)	130.59	116.00	4.00	2220.00	63.80	0.71
Magnesium in Serum or Plasma (mg/dL)	2.01	2.00	0.20	25.20	0.37	2.16
INR in Blood by Coagulation assay	1.74	1.40	0.00	88.60	1.47	2.71
Prothrombin time (PT) in Blood by Coagulation assay	16.89	14.60	7.00	150.00	7.09	2.74
Activated partial thrombplastin time (aPTT) in Blood by Coagulation assay	44.94	35.30	16.40	193.30	25.71	2.94
Phosphate in Serum or Plasma (mg/dL)	3.75	3.50	0.30	22.60	1.41	4.36
Calcium [Mass/volume] in Serum or Plasma (mg/dL)	8.56	8.50	0.30	25.40	0.83	4.54
pH of Urine	5.88	5.50	5.00	9.00	0.99	15.3
Urobilinogen in Urine (mg/dL)	1.79	1.00	0.20	12.00	2.61	15.3
Ketones in Urine (mg/dL)	48.93	15.00	10.00	150.00	49.13	15.3
Specific gravity of Urine by Test strip	1.02	1.02	1.00	1.08	0.01	15.3
Protein in Urine by Test strip (mg in 24h)	106.39	30.00	15.00	500.00	146.79	15.3
Glucose in Urine by Test strip (mg in 24h)	461.21	250.00	70.00	1000.00	401.51	15.3

Table 4: Mimic II Dataset Descriptive Statistics.

The descriptive statistics of MIMIC II dataset, are shown in Table 4. In MIMIC II case, the descriptive statistics for the physiological values are calculated before the LOCF procedure. The missing values rates are calculated after LOCF. A difference between Portavita and MIMIC II datasets is that Portavita has a balanced distribution of labels, whereas in MIMIC II the patient populations are imbalanced. Additionally the two datasets differ in label density. Another difference is that Portavita has a time granularity of months, whereas the tests are performed multiple times per day in MIMIC II. Finally, Portavita has way more missing values than MIMIC II. These differences will allow us to evaluate the considered algorithms in diverse settings and thus also highlight their strengths and weaknesses.

3.3. Missing Value Imputation

For both of the datasets, the multivariate time series present missing values. In medical datasets, the missing at random assumption does not hold, since if a patient presents missing values for a test, it is often because there was no medical reason to perform it. Thus, removing patients with many missing values would bias the study towards patients with more recognized medical conditions. Similarly, removing features with many missing values implies losing information about the status of the patients.

In the Portavita dataset, some of the features are missing more than 90% of the values. This is quite a normal situation in real world standard care datasets, as the patients considered may have different treatments and needs. To be useful, classification algorithms must be robust to large amounts of missing values and still be able to generalize with respect to unseen data.

It is well known that there is not a single universal approach to deal with missing values [24] in medical datasets. One of the most used approaches is to substitute the mean for the missing values [25], but this is rarely considered acceptable [26]. A more acceptable approach is to use medical knowledge to substitute with values within a likely range [26]. With respect to the mean imputation, this avoids the misleading effect of considering ill someone due to imputing values out of normal ranges.

Given these considerations, we performed plausible physiological values imputation in our multi-label classification analysis. We either impute physiological values in ranges that are likely for the given patient illnesses (putting high blood pressure if the patient has hypertension) or we impute physiological values of a healthy person when the related illness label is absent (normal blood pressure if the patient has not hypertension).

4. Methods

In this section we illustrate how we apply a set of multi-label classification algorithms to the selected medical discrete time series datasets.

For each of the algorithms selected we apply the following steps on the data: after transforming our data from medical records to multivariate time series as described in Section 3, we standardize the data to have the same contribution for each feature, we apply a BoW quantization and we standardize the data again to have the same contribution for each codeword. Then, for dimensionality reduction approaches, we apply a dimensionality reduction algorithm and we use a nearest centroid classifier based on the cosine distance. For standard multi-label classifiers, we apply the multi-label classification algorithm after the second standardization step. Fig. 1, inspired by the work of Wang et al. in [7], illustrates the main steps applied by our system in the specific case of KLFDA. For the comparison, we chose the following algorithms, all applied on the model calculated with BoW:

- BoW Cosine: This technique applies the cosine distance on the patient histograms and it represents the baseline for the comparison.

- LDA-BR, KDA-BR and KLFDA-BR: Linear Discriminant Analysis [27], Kernel Discriminant Analysis and Kernel Local Fisher Discriminant Analysis, with a binary relevance approach, where the classes of the patients are those explained in Section 3.

- LDA, KDA and KLFDA: The same as above, but concatenating the features.

- MLkNN, DMLkNN, BPMLL, BR-SVM: Multi-label k-nearest neighbours [18], dependent multi-label k-nearest neighbours [28], back propagation multi-label learning [29] and multi-label support vector machines with a binary relevance approach [30].

We purposely decided to use multi-label algorithms capable of handling non-linearly separable data to confirm our hypothesis that supervised dimensionality reduction algorithms such as KLFDA and KDA are suitable candidates for multi-label learning in medical time series. In the rest of this Section we explain how we apply the BoW algorithm, the nearest centroid classifier and finally the metrics used for the evaluation of the multi-label classifiers.

4.1. From Irregular Multivariate Time Series to Bag of Words

The BoW model was originally introduced for text document analysis [31]. In document retrieval a *codebook* is defined as a set of pre-selected words, also called codewords. The BoW method counts the codewords per document, reducing each document to a histogram. Adapted versions of the BoW model have been recently applied in the field of computer vision for image classification [32, 33], and for biomedical time series classification [7]. When the entities to be analyzed are not documents, but are irregular time series of continuous physiological values, the codebook of the BoW model can be defined using a clustering algorithm. In this paper, the k-means algorithm

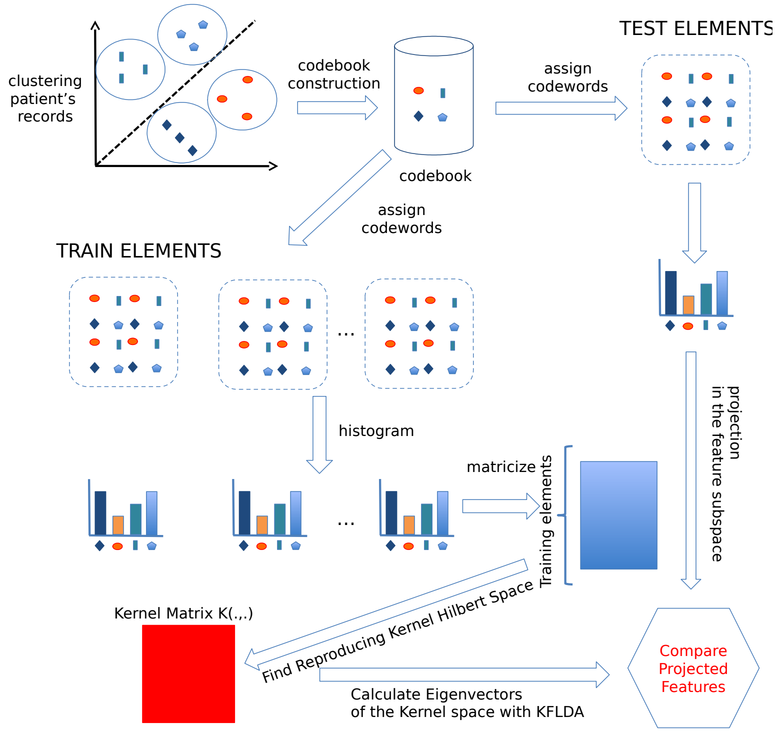


Figure 1: BoW with Kernel Methods.

[34] is used to cluster the multivariate time series obtained from the health records as explained in Section 3, associated to the patients to create a set of centroids. These centroids then become the codewords retained in the codebook.

More formally, if we have a set of health records $X = [x_1, x_2, \ldots, x_n]$, with $x_i \in \mathbb{R}^d$, where d are numerical features of each record, associated to a set of patients P, where each patient can have more than one record, and a set of clustering centers $c_i \in [c_1, ...c_h]$ calculated with k-means and representing the codebook, then we can set an un-normalized feature f, in an un-normalized histogram $\mathbf{u}$, for $f = 1, \ldots, |P|$, as:

$$u_f = \sum_{i=1}^{P_r} ||x_i - c_f||_2 \tag{15}$$

where P_r is the number of records associated to a patient, and $|| \cdot ||_2$ is the euclidean norm. After calculating $\mathbf{u}$, we can calculate a normalized histogram $\mathbf{h}$, as:

$$h_f = \frac{u_f}{||\mathbf{u}||_2} \tag{16}$$

for $f = 1, \ldots, |P|$. Each patient is then represented in terms of a normalized histogram, allowing us to compare patients even if they have a different number of records.

4.2. Nearest Centroid Classification and Label Ranking

To classify a new element x we first use the eigenvectors computed with KLFDA to project the testing sample in the identified subspace for a given label k:

$$\hat{x}^{(k)} = T_{opt}^{(k)} * \phi(x) \tag{17}$$

where $\hat{x}^{(k)}$ is the projected testing sample using the transformation matrix $T_{opt}^{(k)}$, calculated for label k, on the mapping $\phi(x)$.

Second, we concatenate all the test sample projections for each of the labels in a single vector:

$$\hat{\hat{x}} = (\hat{x}^{(1)}|\hat{x}^{(2)}|\ldots|\hat{x}^{(k)}) \tag{18}$$

The possibility to concatenate features is a big advantage of dimensionality reduction approaches such as KDA and KLFDA as it allows us to define an easy way to chain the features calculated by the different classifiers, without the need to train another classifier recursively as it happens with classifier chains (CC).

Third, we calculate a *cosine* distance between the mean of the projected training elements and the projected testing element for each of the labels.

$$d_k = \cos(\hat{\hat{x}}, \mu_k) = \frac{\hat{\hat{x}} \cdot \mu_k}{\parallel \hat{\hat{x}} \parallel\parallel \mu_k \parallel} \tag{19}$$

Where μ_k represent the mean of the concatenation of the features of the training elements in the projected space belonging to label k. To decide whether an element has a label or not, we perform the following test:

$$\hat{y}_k = \begin{cases} 1 & \text{if } d_k < d_{\sim k}, \\ 0 & \text{otherwise} \end{cases} \tag{20}$$

Finally, we can define the ranking function $rank_f(\hat{\hat{x}}, k)$ for label k as:

$$rank_f(\hat{\hat{x}}, k) = 1 - \frac{d_k}{d_k + d_{\sim k}} \tag{21}$$

4.3. Multi-label Metrics

As stated in [18, 35], multi-label performance metrics differ from single label ones. Following the same approach presented in [36, 18], we propose the following five evaluation metrics for multi-label learning.

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 is the best:

$$hloss_S(h) = \frac{1}{m} \sum_{i=1}^{m} \frac{|h(x_i) \triangle Y_i|}{|L|}. \tag{22}$$

One-error: evaluates how many times the top-ranked label is not in the set of proper labels of the observation. The score lies between 0 and 1, where 0 is the best:

$$one\text{-}error_S(f) = \frac{1}{m} \sum_{i=1}^{m} \gamma(\arg\max_{y \in L} f(x_i, y)), \tag{23}$$

where

$$\gamma(y) = \begin{cases} 1 & \text{if } y \notin Y_i, \\ 0 & \text{otherwise.} \end{cases} \tag{24}$$

Coverage: evaluates how far on average we need to traverse the list of labels in order to cover all the proper labels of the observation. A score as small as possible is better:

$$coverage_S(f) = \frac{1}{m} \sum_{i=1}^{m} \max_{y \in Y_i} rank_f(x_i, y) - 1. \tag{25}$$

Ranking loss: evaluates the average part of label pairs that are ordered in reverse for the observation. The score lies between 0 and 1, where 0 is the best:

$$\begin{aligned} rloss_S(f) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{|Y_i||(L \setminus Y_i)|} \\ \times |\{(y_1, y_2)|f(x_i, y_1) \leqslant f(x_i, y_2), \\ (y_1, y_2) \in Y_i \times (L \setminus Y_i)\}|. \end{aligned} \tag{26}$$

Average precision: evaluates the average fraction of labels ranked above a particular label $y \in Y_i$ which actually are in Y_i. The score lies between 0 and 1, where 1 is the best:

$$\begin{aligned} avgprec_S(f) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{|Y_i|} \\ \times \sum_{y \in Y_i}\frac{|\{y'|rank_f(x_i, y') \leqslant rank_f(x_i, y), y' \in Y_i\}|}{rank_f(x_i, y)}. \end{aligned} \tag{27}$$

Where $\triangle$ represents the symmetric difference, and $\setminus$ is the set-theoretic difference.

5. Results

In this section we evaluate the combination of BoW and multi-label classification algorithms. In the Portavita dataset, we perform our evaluation using a leave-one-patient-out cross validation (LOPO CV). LOPO CV proved to be suitable for the medical domain [37], as it avoids situations that happen with leave-one-out (LOO), where records of the same patient are both in the training and testing set. Furthermore, LOPO CV presents an advantage with respect to N-folds CV, in which the selection of the random splits may lead to choose suboptimal parameters. Given the fact that we have 525 patients and 3528 health records, the computational cost of LOPO CV is affordable for the Portavita dataset. For model selection, we split our dataset into a training/validation set and a testing set, applying a LOPO CV on the training/validation set to select the best model for the testing phase. We withheld 375 patients for the training/validation and 150 patients for the testing.

Concerning the MIMIC II dataset extraction, we used a 10-fold CV approach for the grid search, splitting the dataset and keeping 70% of the patients (1844) for training and validation and 30% (791) patients for testing, while keeping the same distribution of labels in the test dataset. 10-folds CV was chosen as this dataset counts 2635 patients for a total of 159 127 health records, and LOPO CV was computationally infeasible to run a grid search.

5.1. Parameters Selection with CV

The combination of BoW and dimensionality reduction techniques involves many parameters: size of the codebook, neighbours for the affinity matrix, the regularization coefficient, and the number of components to retain in the dimensionality reduction. Given the large amount of parameters to evaluate, we decided to run a grid search with a step of 100 for the size of the codebook, identifying $cb = 600$ as the best size for the codebook for all the considered algorithms in the Portavita dataset and $cb = 800$ for the MIMIC II dataset.

After the dimensionality reduction applied by LDA, KDA, and KLFDA, we always retain components that can explain at least 99% of the variance of the model. Fig. 2 shows the grid search on average precision and hamming

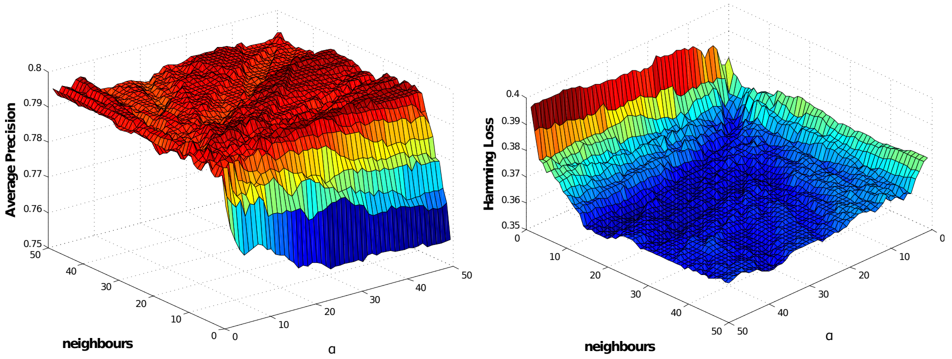


Figure 2: Grid Search on Hamming Loss and Average Precision for KLFDA.

loss, run on the training set, for the other parameters of KLFDA in the Portavita dataset.

Algorithm	Parameters Portavita	Parameters MIMIC II	Interval	Step
KDA-BR	λ=40,cb=600	λ=10, cb=800	[1:50],[10^2:10^3]	lin,lin
KDA	λ=49,cb=600	λ=20, cb=800	[1:50],[10^2:10^3]	lin,lin
KLFDA-BR	N=30, λ=5,cb=600	N=40,λ=3,cb=800	[1:50],[1:50],[10^2:10^3]	lin,lin,lin
KLFDA	N=34, λ=9,cb=600	N=47,λ=3,cb=800	[1:50],[1:50],[10^2:10^3]	lin,lin,lin
MLkNN	N=29, σ=2,cb=600	N=19,σ=3,cb=800	[1:50],[1:50],[10^2:10^3]	lin,lin,lin
DMLkNN	N=29, σ=12,cb=600	N=19,σ=5,cb=800	[1:50],[1:50],[10^2:10^3]	lin,lin,lin
BPMLL	λ=10^{-6},HN=5,cb=600	λ=10^{-7}, HN=7,cb=800	[10^{-8}:1],[1:20],[10^2:10^3]	log,lin,lin
BR-SVM	C=10,γ=10^{-1},cb=600	C=5, γ=1, cb=800	[1:10^2],[10^{-3}:10^3],[10^2:10^3]	lin,log,lin

Table 5: Selected Parameters.

Table 5 summarizes the parameters selection performed with LOPO CV and 10-fold cross validation concerning the algorithms studied for the Portavita and MIMIC II datasets. The parameter N identifies the number of neighbours, while λ identifies the regularization factor, σ the smoothing parameter for MLkNN and DMLkNN, γ the exponent of the RBF kernel, and HN the number of hidden nodes in the BPMLL algorithm. In particular, the most difficult algorithm to train has been BPMLL as it requires more parameters than the other algorithms. To simplify the search, we decided to keep the learning rate constant to the default value $\alpha = 0.05$.

5.2. Results on the Portavita Dataset

After training and validation of the model, we perform our testing using 150 additional patients from the Portavita dataset, with respect to the performance measures discussed in Section 4. Table 6 shows the results for the selected algorithms on the Portavita dataset, with a confidence interval of 95%. The BoW Cosine approach is taken as a baseline for the comparison with the other algorithms.

Algorithm/Metric	Average Precision	Hamming Loss	Ranking Loss	Coverage	One-Error
BoW	68% ± 4%	51.6% ± 4.2%	50% ± 5.6%	2.2 ± 0.14	50% ± 7.8%
LDA	67.8% ± 4%	51.1% ± 4.3%	51.9% ± 5.7%	2.3 ± 0.14	48% ± 8%
LDA-BR	67% ± 4%	52.8% ± 4.3%	52.2% ± 5.8%	2.3 ± 0.14	48% ± 8%
KDA-BR	67.8% ± 3.8%	57.5% ± 3.2%	58% ± 3.9%	2.25 ± 0.15	48% ± 8%
KDA	78.3% ± 3.8%	39.1% ± 4.1%	33.4% ± 5.5%	1.94 ± 0.16	32.1% ± 7.5%
KLFDA-BR	73.5% ± 3.8%	41.6% ± 3.5%	38.9% ± 5.2%	2.07 ± 0.16	35.7% ± 7.7%
KLFDA	**78.8%** ± 3.7%	**37.3%** ± 4%	**32.2%** ± 5.2%	**1.87** ± 0.16	32.1% ± 7.5%
MLkNN	78.2% ± 3.7%	44% ± 3.8%	36 ± 5.5%	2 ± 0.16	**30%** ± 7.5%
DMLkNN	76.1% ± 3.8%	44.8% ± 3.8%	37% ± 5.4%	2.1 ± 0.15	33.3% ± 7.5%
BPMLL	75.7% ± 3.8%	42.6% ± 4.1%	38.3% ± 5.6%	2 ± 0.16	36% ± 7.7%
BR-SVM	78.2% ± 3.7%	39.3% ± 4.2%	34% ± 5.3%	1.98 ± 0.16	33.3% ± 7.5%

Table 6: Results for the Portavita Dataset.

The classes of patients of the Portavita dataset do not appear to be linearly separable and linear techniques such as LDA and LDA-BR do not seem to improve the results with respect to a BoW classifier. We think that this is due to the tendency of non-regularized LDA to overfit when the ratio between the classes and the features of the training elements is small [38, 39].

KLFDA and KDA with feature concatenation achieve a better hamming and ranking loss than the other considered algorithms. KLFDA also achieves a better average precision. This suggests that in the case of classifying patients affected by DT2, the possibility of using supervised kernel methods and LPP brings an advantage in terms of classification. Another advantage of both KDA and KLFDA when compared to the other considered algorithms is the possibility of concatenating the projected features calculated by each of the classifiers. The KLFDA-BR and KDA-BR algorithms, on the contrary, do not perform much better than the standard BoW approach. In particular KDA-BR performs exactly the same as the BoW case. This is probably due to the fact that the relationship between the labels is not taken into consideration, degrading performances. KLFDA-BR shows an improvement with respect to BoW. This suggests that algorithms considering the locality of the data are likely to perform better than algorithms considering only the labels.

MLkNN performs similarly to KDA and KLFDA, with the best one-error score, confirming that making use of neighbourhood properties of the dataset is quite important in the case of DT2 patients. In contrast, BPMLL does not generalize too well with respect to new data after the training. The main issue of BPMLL is the large number of parameters to train, which makes it difficult to tune properly. Furthermore, BPMLL seems to be affected more by the large rate of missing values in the Portavita dataset than the other considered algorithms.

BR-SVM performs well in both training and testing, despite not considering the interaction between the labels, which seems to explain the difference in performance with KLFDA concerning hamming loss and ranking loss.

5.3. Results on the MIMIC II dataset

Table 7 shows the results for the selected algorithms on the MIMIC II dataset, with a confidence interval of 95%. As it is clear from Table 7, BoW

Algorithm/Metric	Average Precision	Hamming Loss	Ranking Loss	Coverage	One-Error
BoW	47.5% ± 1.6%	44.5% ± 1.1%	39.5% ± 1.6%	5.6 ± 0.14	69% ± 3.4%
LDA	50.9% ± 1.6%	42.7% ± 1.1%	38.2% ± 1.5%	5.4 ± 0.17	63.2% ± 3.3%
LDA-BR	43.3% ± 1.4%	40.2% ± 0.8%	39.2% ± 1.4%	5.14 ± 0.15	82.93% ± 2.5%
KDA-BR	64.74% ± 1.8%	23.7% ± 1.2%	26% ± 1.4%	4.67 ± 0.19	39.7% ± 3.3%
KDA	66% ± 1.8%	23.4% ± 0.9%	23.2% ± 1.4%	4.2 ± 0.18	40.5% ± 3.3%
KLFDA-BR	64.79% ± 1.8%	23.8% ± 1.1%	25.9% ± 1.4%	4.59 ± 0.18	40% ± 3.4%
KLFDA	65.5% ± 1.8%	23.3% ± 0.9%	23.7% ± 1.4%	4.25 ± 0.16	41.1% ± 3.3%
MLkNN	**68.4%** ± 1.8%	21.7% ± 1%	**20%** ± 1.5%	**4** ± 0.17	35.6% ± 3%
DMLkNN	68.1% ± 3.8%	**21.6%** ± 0.8%	21.5% ± 1.5%	4 ± 0.17	**34.1%** ± 3.2%
BPMLL	67.8% ± 1%	26% ± 0.8%	37.1% ± 3.3%	4 ± 0.18	37% ± 3.3%
BR-SVM	57.7% ± 1.8%	22.2% ± 0.9%	37% ± 1%	5.8 ± 0.19	38.6% ± 3.2%

Table 7: Results for the MIMIC II Dataset.

without any transformation is affected by the curse of dimensionality and LDA-BR does not really give meaningful results. LDA manages to improve the results with respect to BoW, but as in the Portavita dataset, the algorithm does not perform well.

First, Table 7 shows that for the MIMIC II dataset, the two best performing algorithms are MLkNN and DMLkNN, while the KDA and KLFDA algorithms perform similarly to MLkNN and DMLkNN. The fact that MIMIC II dataset has less missing values than Portavita, seems to favour the BPMLL algorithm, which performs well from the perspective of the average precision. BPMLL still does not perform well for the hamming loss and the ranking loss, which we believe related to the difficulty in training the algorithm.

Second, binary relevance approaches seem to perform well on MIMIC II, except for BR-SVM. KDA-BR performs similarly to KLFDA-BR: this may happen because MIMIC II has 194 different symbols, and thus the interaction between the illnesses is quite complex, limiting the advantage of LPP projections. Furthermore, KDA-BR and KLFDA-BR seem to have comparable results to KLFDA and KDA, where the calculated features are concatenated.

This may be related to the fact that MIMIC II is imbalanced. Concatenating features moves the centroid of a label depending on the features calculated for the other labels, but majority labels may have more impact in defining the centroids, degrading the performance. The difference in performance between BR-SVM, KDA-BR and KLFDA-BR is of more difficult interpretation. This could be caused by the use of regularization or of the nearest centroid classifier in KDA-BR and KLFDA-BR algorithms.

5.4. Discussion

The fact that KLFDA and KDA perform better than the other algorithms for the Portavita dataset in respect to hamming loss and ranking loss is quite important in medical applications such as classification of diabetic patients complications. Hamming loss discriminates the capability of the algorithm to identify the presence of a complication, while ranking loss discriminates how well the algorithm ranks the labels. These metrics allow a caregiver to understand which patient illnesses have a strong expression, giving an indication on where to act more promptly.

The performed evaluation illustrates the strengths and weaknesses of KDA and KLFDA for multi-label classification tasks: the behaviour of KDA and KLFDA is comparable with that of state-of-the-art multi-label classification algorithms, but they seem to present an advantage with respect to datasets with a large number of missing values and with a high label density such as the Portavita dataset. We can have a better idea of the behaviour of KDA and KLFDA by looking at Table 8, comparing the hamming loss per symbol of KLFDA, KDA, BR-SVM and MLkNN, in the Portavita dataset (confidence intervals are omitted as we only have 10 elements per symbol).

Symbol	H	D	Mi	Ma	KLFDA	KDA	BR-SVM	MLkNN
1	no	no	no	yes	40%	42.5%	60%	60%
2	no	no	yes	no	40%	35%	47.5%	50%
3	no	no	yes	yes	40%	45%	42.5%	42.5%
4	no	yes	no	no	25%	32.5%	27.5%	40%
5	no	yes	no	yes	42.5%	45%	40%	42.5%
6	no	yes	yes	no	20%	15%	15%	30%
7	no	yes	yes	yes	40%	40%	42.5%	42.5%
8	yes	no	no	no	35%	32.5%	40%	45%
9	yes	no	no	yes	40%	40%	45%	57.5%
10	yes	no	yes	no	47.5%	57.5%	47.5%	42.5%
11	yes	no	yes	yes	45%	40%	62.5%	52.5%
12	yes	yes	no	no	27.5%	40%	20%	40%
13	yes	yes	no	yes	40%	40%	40%	42.5%
14	yes	yes	yes	no	35%	37.5%	30%	30%
15	yes	yes	yes	yes	42.5%	40%	30%	42.5%

Table 8: Hamming Loss Per Symbol in the Portavita Dataset. H = Hypertension, D = Dyslipidemia, Mi= Microvascular, Ma = Macrovascular.

In these results, we see that KDA has an advantage where the patients have only one label, which are also those patients presenting many missing values in Portavita dataset. For the other classes, KDA performs similarly to BR-SVM, with some exceptions, probably caused by the fact that BR-SVM finds support vectors, whereas KDA is a variance based method. KLFDA seems to combine the behaviour of KDA, BR-SVM and MLkNN: the eigenvectors explaining little variance are discarded just like in KDA; the use of kernel methods allows KLFDA to deal with non-linearity in the data, similarly to BR-SVM; the LPP transformation allows KLFDA to consider the neighbourhood of the elements, similarly to MLkNN, but in addition if there are enough elements per symbol, with a high label density, the retained eigenvectors would be able to characterise those symbols expressing most variance. In this sense, when dealing with datasets presenting the three aspects of missing data, high label density and non-linearity, KLFDA may have an advantage with respect to other techniques.

In MIMIC II, KLFDA and KDA perform slightly worse on the average precision than MLkNN and DMLkNN, but they are comparable for hamming loss and ranking loss. A possible reason for this is that MIMIC II dataset is imbalanced. KLFDA and KDA are variance based methods, so an imbalanced estimation of the classes variance impacts the calculated model and its performance. Looking at the hamming loss per symbol in MIMIC II, we found that the absence of missing values in MIMIC II, cancels the advantage of KLFDA and KDA, as they behave similarly to MLkNN for patients with only one complication. Additionally, MIMIC II has a low label density, with 194 symbols and few patients for most of the symbols, which may be difficult to characterise for the variance based model calculated by KLFDA and KDA. If this is the case, only the effect of the LPP projection of KLFDA, and of the nearest centroid classifier for KLFDA and KDA would be present and that would explain the similar behaviour of KLFDA and KDA with MlKNN.

Finally, an advantage of KLFDA and KDA it that they compute a model based on eigenvectors, which allows to include new patients' records by projecting their BoW representation and then recalculating the centroids for each class, while MLkNN and DMLkNN have to store the new instances in memory, that is infeasible with big datasets.

6. Conclusions

In this paper we studied the combination of the BoW model in medical time series with dimensionality reduction approaches for multi-label patient classification. When taking the Portavita dataset into consideration, the

KLFDA algorithm with a nearest centroid classifier achieves the best results. In the MIMIC II dataset, dimensionality reduction algorithms are comparable to state-of-the-art multi-label classification algorithms, but suffer from the fact that the dataset is imbalanced.

There are several possible extensions to this work. At the moment we are using a single kernel mapping, but extensions of KLFDA and KDA that work with multiple kernel learning have already been defined [40]. Multiple kernels could achieve a better mapping for our data and improve the precision of KLFDA and KDA.

Another promising approach could be to develop a multi-label version of KLFDA and KDA, similarly to what is proposed in [41]. This would require modifying the definition of the scatter matrices in KLFDA and KDA to consider multiple labels, which is quite a challenging problem.

In Section 3, we identified the issue of dealing with values sampled at different frequencies. Quantizing patient data with different sampling frequencies or considering descriptive statistics rather than a codebook, could be suitable approaches. Finally, we could apply a different substitution to LOCF and generate physiological values with a maximum likelihood model, provided that enough patients' records are available.

Acknowledgment

This work was partially supported by the FP7 287841 COMMODITY12 project. We thank the anonymous reviewers for the useful comments that allowed us to improve the paper considerably.

References

[1] C. McAdam Marx, Economic implications of type 2 diabetes management, Am J Manag Care 19 (2013) 143–148.

[2] J. G. Ghably, B. J. Paterson, A. N. Peiris, Endocrinology in crisis?, South. Med. J. 106 (2013) 245.

[3] T. G. Kannampallil, A. Franklin, R. Mishra, K. F. Almoosa, T. Cohen, V. L. Patel, Understanding the nature of information seeking behavior in critical care: Implications for the design of health information technology, Artificial Intelligence in Medicine 57 (2013) 21–29.

[4] J. Sun, D. Sow, J. Hu, S. Ebadollahi, Localized Supervised Metric Learning on Temporal Physiological Data, in: Proceedings of the 2010 20th International Conference on Pattern Recognition, ICPR '10, IEEE Computer Society, Washington, DC, USA, 2010, pp. 4149–4152.

[5] J. Sun, F. Wang, J. Hu, S. Edabollahi, Supervised patient similarity measure of heterogeneous patient records, SIGKDD Explorations 14 (2012) 16–24.

[6] R. B. Marimont, M. B. Shapiro, Nearest Neighbour Searches and the Curse of Dimensionality, IMA Journal of Applied Mathematics 24 (1979) 59–70.

[7] J. Wang, P. Liu, M. F. She, S. Nahavandi, A. Z. Kouzani, Bag-of-words Representation for Biomedical Time Series Classification, CoRR abs/1212.2262 (2012).

[8] M. Jiu, C. Wolf, C. Garcia, A. Baskurt, Supervised learning and codebook optimization for bag of words models, Cognitive Computation 4 (2012) 409–419.

[9] P. Ordóñez, T. Armstrong, T. Oates, J. Fackler, Using modified multivariate bag-of-words models to classify physiological data, in: M. Spiliopoulou, H. Wang, D. J. Cook, J. Pei, W. Wang, O. R. Zaïane, X. Wu (Eds.), ICDM Workshops, IEEE, 2011, pp. 534–539.

[10] G. Madjarov, D. Kocev, D. Gjorgjevikj, S. Dzeroski, An extensive experimental comparison of methods for multi-label learning, Pattern Recognition 45 (2012) 3084–3104.

[11] M.-L. Zhang, Z.-H. Zhou, A Review On Multi-Label Learning Algorithms, IEEE Transactions on Knowledge and Data Engineering 99 (2013) 1.

[12] S. Ji, J. Ye, Linear dimensionality reduction for multi-label classification, in: Proceedings of the 21st international jont conference on Artifical intelligence, IJCAI'09, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2009, pp. 1077–1082.

[13] E. Pacharawongsakda, C. Nattee, T. Theeramunkong, Improving multilabel classification using semi-supervised learning and dimensionality reduction, in: P. Anthony, M. Ishizuka, D. Lukose (Eds.), PRICAI, volume 7458 of *Lecture Notes in Computer Science*, Springer, 2012, pp. 423–434.

[14] B. Qian, I. Davidson, Semi-Supervised Dimension Reduction for Multi-Label Classification, in: M. Fox, D. Poole (Eds.), AAAI, AAAI Press, 2010.

[15] J. P. Pestian, C. Brew, P. Matykiewicz, D. J. Hovermale, N. Johnson, K. B. Cohen, W. Duch, A Shared Task Involving Multi-label Classification of Clinical Free Text, in: Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing, BioNLP '07, Association for Computational Linguistics, Stroudsburg, PA, USA, 2007, pp. 97–104.

[16] W. T. Cade, Diabetes-related microvascular and macrovascular diseases in the physical therapy setting, Phys Ther 88 (2008) 1322–1335.

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

[18] M.-L. Zhang, Z.-H. Zhou, ML-KNN: A lazy learning approach to multi-label learning, Pattern Recognition 40 (2007) 2038–2048.

[19] J. Shawe-Taylor, N. Cristianini, Kernel Methods for Pattern Analysis, Cambridge University Press, 2004.

[20] A. Barla, F. Odone, A. Verri, Histogram intersection kernel for image classification, in: ICIP (3), 2003, pp. 513–516.

[21] X. He, P. Niyogi, Locality preserving projections, in: S. Thrun, L. Saul, B. Schölkopf (Eds.), Advances in Neural Information Processing Systems 16, MIT Press, Cambridge, MA, 2004.

[22] K. Fukunaga, Introduction to statistical pattern recognition (2nd ed.), Academic Press Professional, Inc., San Diego, CA, USA, 1990.

[23] M. Sugiyama, Dimensionality Reduction of Multimodal Labeled Data by Local Fisher Discriminant Analysis, Journal of Machine Learning Research 8 (2007) 1027–1061.

[24] R. J. Little, R. D'Agostino, M. L. Cohen, K. Dickersin, S. S. Emerson, J. T. Farrar, C. Frangakis, J. W. Hogan, G. Molenberghs, S. A. Murphy, J. D. Neaton, A. Rotnitzky, D. Scharfstein, W. J. Shih, J. P. Siegel, H. Stern, The prevention and treatment of missing data in clinical trials, N. Engl. J. Med. 367 (2012) 1355–1360.

[25] J. D. Dziura, L. A. Post, Q. Zhao, Z. Fu, P. Peduzzi, Strategies for dealing with missing data in clinical trials: from design to analysis, Yale J Biol Med 86 (2013) 343–358.

[26] D. Jackson, I. R. White, M. Leese, How much can we learn about missing data?: an exploration of a clinical trial in psychiatry, J R Stat Soc Ser A Stat Soc 173 (2010) 593–612.

[27] R. A. Fisher, The Use of Multiple Measurements in Taxonomic Problems, Annals of Eugenics 7 (1936) 179–188.

[28] Z. Younes, F. Abdallah, T. Denoeux, H. Snoussi, A dependent multilabel classification method derived from the k-nearest neighbor rule, EURASIP J. Adv. Sig. Proc. 2011 (2011).

[29] M.-L. Zhang, Z.-H. Zhou, Multilabel Neural Networks with Applications to Functional Genomics and Text Categorization, IEEE Transactions on Knowledge and Data Engineering 18 (2006) 13381351.

[30] C. Cortes, V. Vapnik, Support-vector networks, Machine Learning 20 (1995) 273–297.

[31] Y. Ko, A study of term weighting schemes using class information for text classification, in: Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, SIGIR '12, ACM, New York, NY, USA, 2012, pp. 1029–1030.

[32] J. Sivic, A. Zisserman, Efficient Visual Search of Videos Cast as Text Retrieval, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (2009) 591–606.

[33] J. Sivic, A. Zisserman, Efficient Visual Content Retrieval and Mining in Videos, in: K. Aizawa, Y. Nakamura, S. Satoh (Eds.), Advances in Multimedia Information Processing - PCM 2004, volume 3332 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, 2005, pp. 471–478.

[34] J. B. MacQueen, Some Methods for Classification and Analysis of Multivariate Observations, in: L. M. L. Cam, J. Neyman (Eds.), Proc. of the

fifth Berkeley Symposium on Mathematical Statistics and Probability, volume 1, University of California Press, 1967, pp. 281–297.

[35] G. Tsoumakas, I. Katakis, I. P. Vlahavas, Mining Multi-label Data, in: O. Maimon, L. Rokach (Eds.), Data Mining and Knowledge Discovery Handbook, Springer, 2010, pp. 667–685.

[36] R. E. Schapire, Y. Singer, Boostexter: A Boosting-based System for Text Categorization, Machine Learning 39 (2000) 135–168.

[37] M. Dundar, G. Fung, L. Bogoni, M. Macari, A. Megibow, B. Rao, A methodology for training and validating a CAD system and potential pitfalls, International Congress Series 1268 (2004) 1010–1014. CARS 2004 - Computer Assisted Radiology and Surgery. Proceedings of the 18th International Congress and Exhibition.

[38] A. Jain, B. Chandrasekaran, Dimensionality and Sample Size Considerations, in: P. Krishnaiah, L. Kanal (Eds.), Pattern Recognition in Practice, 1982, pp. 835–855.

[39] K. Linnet, On the sensitivity of linear discriminant analysis to sampling variation and analytical errors, Computers and Biomedical Research 21 (1988) 158–168.

[40] Z. Wang, X. Sun, Multiple kernel local Fisher discriminant analysis for face recognition, Signal Processing 93 (2013) 1496–1509. Special issue on Machine Learning in Intelligent Image Processing.

[41] C. H. Park, M. Lee, On applying linear discriminant analysis for multi-labeled problems, Pattern Recognition Letters 29 (2008) 878887.