Feature Selection for the Classification of Longitudinal Human Ageing Data

Tossapol Pomsuwan
School of Computing,
University of Kent, UK
tp321@kent.ac.uk

Alex A. Freitas
School of Computing,
University of Kent, UK
A.A.Freitas@kent.ac.uk

***Abstract* — We propose a new variant of the Correlation-based Feature Selection (CFS) method for coping with longitudinal data – where variables are repeatedly measured across different time points. The proposed CFS variant is evaluated on ten datasets created using data from the English Longitudinal Study of Ageing (ELSA), with different age-related diseases used as the class variables to be predicted. The results show that, overall, the proposed CFS variant leads to better predictive performance than the standard CFS and the baseline approach of no feature selection, when using Naïve Bayes and J48 decision tree induction as classification algorithms (although the difference in performance is very small in the results for J4.8). We also report the most relevant features selected by J48 across the datasets.**

Keywords — classification, feature selection, longitudinal data, age-related diseases

I. INTRODUCTION

In machine learning, a classification algorithm aims to find a predictive relationship between features and the class variable. This is done by building a classification model from pre-classified instances. Afterwards, this model is used to predict the class label of previously unseen instances.

In classification datasets with a large number of features, feature selection methods are often applied in a data pre-processing step [1]–[3] in order to remove irrelevant or redundant features. This can lead to higher predictive accuracy and reduce the training time of classification algorithms.

The vast majority of works on the classification task focus on analysing the standard type of classification data, where each variable is measured at a single time point, so that there is no explicit temporal structure in the data. However, many important data sources – particularly in the biomedical domain – contain longitudinal data, where the values of a variable are repeatedly measured across several time points (often called waves) [4]. For instance, many hospital databases contain records with blood test results measured for the same patient across many time points.

In this work, we address the feature selection task, in the special context of longitudinal data. When analysing longitudinal data, a standard feature selection method would typically ignore the temporal nature of the features and treat each feature value at a given time point as a separate feature. That is, a standard algorithm would ignore the important difference between values of the same feature (measuring the same property of an instance) across different time points and values of fundamentally different features (measuring different properties of an instance) at the same time point.

In order to mitigate the above limitation of standard feature selection methods, we propose an adaptation of the well-known Correlation-based Feature Selection (CFS) method [5] to the context of longitudinal classification. The proposed adaptation of CFS works in two phases. First, it explicitly treats different values of the same feature across all time points as the same group of temporally related features, performing feature selection separately within each group of such related features. Second, it merges the selected features across all the groups in order to produce a single set of selected features which is then used as input by classification algorithms.

The proposed adaptation of CFS was evaluated on 10 longitudinal classification datasets created in this work using data from the English Longitudinal Study of Ageing (ELSA) [6]. Each dataset was involved in a classification task where the goal was to predict whether or not an individual would have an age-related disease, based mainly on the values of biomedical features measured for that individual in previous time points.

The experimental results showed that the proposed adaptation of CFS obtained, overall, higher predictive accuracy than the standard CFS (which ignores the temporal nature of the features) and the natural baseline approach of not performing feature selection in the created datasets.

This paper is organised as follows. Section II presents background and related work. Section III describes how the longitudinal datasets were created. Section IV introduces the proposed extension to the correlation-based feature selection method. Section V reports the computational results. Section VI presents the conclusion.

II. BACKGROUND

A. Feature Selection

In the classification task, feature selection is often performed in a data preprocessing step to select a subset of relevant features out of all original features. There are several motivations for feature selection [1], [2]. The main one is to remove irrelevant, noisy, or redundant features, which can reduce the predictive accuracy of the classification model [2]. In addition, identifying the most relevant features is a form of discovered knowledge by itself. Moreover, feature selection can improve the interpretability of the classification model due to the smaller number of features used to build the model. Finally,

2375-9259/17 $31.00 © 2017 IEEE
DOI 10.1109/ICDMW.2017.102

reducing the number of features can substantially speed up the execution of the classification algorithm.

In general feature selection methods have two components: a search method which decides how to generate new subsets of features to be evaluated, and an evaluation function which assigns a numerical quality value to each candidate feature subset. There are three types of feature selection approaches. The first one is the filter approach [3], which evaluates a feature subset without running the target classification algorithm – i.e. the algorithm that will use the selected features to build a classification model. Typically, the filter method uses simple statistical tests as an evaluation function. Clearly, the main advantage of this approach is that it is relatively fast.

In contrast, the wrapper and embedded approaches require running the target classification algorithm. The former evaluates the quality of a candidate feature subset by measuring the predictive accuracy (on training data) of the classification model built with that feature subset. This approach is very time consuming, since it requires many runs of a classification algorithm. The embedded approach builds a classification model and carries out feature selection at the same time. For example, when building a decision tree, the relevant features are automatically selected by the algorithm. This approach can also be very time consuming, depending on the classification algorithm used. In this work we use the filter approach, which is faster and more scalable to a large number of features.

B. Correlation-based Feature Selection

Correlation-based Feature Selection (CFS) is a filter method which evaluates candidate feature subsets based on the following principle: good feature subsets contain features highly correlated with the class variable, but uncorrelated with each other, i.e., with little or no redundancy among features. To implement this principle, the standard CFS method [5] tries to (a) maximize the average correlation between each feature in a candidate subset and the class variable and (b) minimize the average correlation between each pair of features in a candidate subset.

C. Longitudinal Classification

Unlike standard (non-longitudinal) datasets, longitudinal datasets consist of features whose values are assigned at multiple time points for each instance in a dataset. For example, a health-survey dataset, where instances represent patients, could contain features representing the results of different blood sample tests across several successive years. From a machine learning perspective, this type of datasets has temporal information about the features: how each feature's values change across time. In general, conventional classification algorithms do not explicitly exploit this temporal information, since they treat all occurrences of a feature in the same way regardless of how recent the feature values are.

In addition, the different values of a feature across time can exhibit some temporal redundancy in the sense that the value of a feature at a given time point may be correlated with values of the same feature in other time points (particularly closer time points). This is generally known as autocorrelation in the area of time series. Again, this kind of temporal redundancy is not explicitly detected by non-longitudinal classification or feature selection algorithms, which would not distinguish between measuring the correlation between two values of the same feature in two different time points (temporal redundancy) and measuring correlation between the values of two very different features at the same time point (non-temporal redundancy). By identifying these two types of redundancy, one can develop a feature selection algorithm that exploits the difference between them in order to try to improve the effectiveness of the feature selection procedure as will be seen in the later section.

In general, there are two approaches for longitudinal classification. The first one is the problem transformation approach, which transforms a longitudinal dataset into a non-longitudinal dataset before applying a conventional classification algorithm. The second approach is the algorithm-adaptation approach which adapts a non-longitudinal classification algorithm for longitudinal datasets. In this paper, we focus on the problem transformation approach, which is more generic (algorithm-independent), so that we can apply different classification algorithms and analyse different types of classification models.

As mentioned earlier, CFS can eliminate redundant and irrelevant features, but standard CFS ignores the temporal relation among the features so that it does not explicitly address the above mentioned temporal redundancy as a specific issue in longitudinal datasets. In the next Section, we briefly review related work on longitudinal feature selection methods, which were explicitly designed for longitudinal classification data.

D. Related Work on Longitudinal Feature Selection

Although there is a huge literature on conventional (non-longitudinal) feature selection [1]–[3], there are relatively few published studies on longitudinal feature selection for classification tasks. Here we briefly discuss the longitudinal feature selection methods most related to our work.

In [7], a longitudinal feature selection method was proposed for temporal gene expression data. They used the Minimum Redundancy Maximum Relevance (mRMR) method, whose evaluation function is conceptually similar to the CFS' one, based on maximising the candidate features' relevance with respect to the class variable and minimising redundancy among candidate features. A feature's degree of relevance is computed by the mean of the F-statistic over all the time points. A drawback is that the degree of relevance is averaged across all time points, ignoring that feature values at recent time points are intuitively more relevant for class prediction than older feature values. Also, the F-statistic makes the strong assumption that the data are normally distributed. The degree of redundancy among features is measured by using Dynamic Time Warping (DTW), also used in [8].

Another related work is [9], which proposed a margin-based feature selection method which transforms a feature space into a weighted feature space. A temporal margin is defined based on a measure of distance between two-time points, and then it selects the features with large weights that maximise each temporal margin. This method only considers a feature's relevance with respect to the class. In other words, the redundancy among features is ignored.

III. DATA PREPARATION

The classification datasets created in this work were derived from the English Longitudinal Study of Ageing (ELSA) [6] – https://www.elsa-project.ac.uk/. The ELSA study is a longitudinal survey of ageing and quality of life among older people that explores the dynamic relationships between health and functioning, social networks and participation, and economic position as people plan for, move into and progress beyond retirement. In this work, however, we focus only on the biomedical data in ELSA, such as the results of blood tests and other data collected by nurses, and the relationship between that data and the health status of patients, as will be described in more detail later. The ELSA subjects were recruited from a representative sample of the English population, who lived in private households, aged 50 and over [6]. The data was collected every two years: each data collection period was referred to as a 'wave', so that we can observe the variation of each feature's values for each individual across those waves. In total, seven waves of data were collected and have well-documented data.

It should be noted that the data in the ELSA database was not collected specifically for machine learning purposes. Hence, we had to spend a large amount of time with data preparation for the classification task. The first step was to define the instances (objects to be classified), the classes and the predictive features used for classification. In essence, the instances represent individuals in the ELSA database, the class variables represent age-related diseases and the features represent biomedical information collected by nurses or other relevant characteristics of an individual (age and gender). We next describe data preparation in detail.

A. Creating class variables representing age-related diseases

We aim at building classification models which help us understand what health factors play an important role in predicting whether or not a patient will have a certain age-related disease in the future. Therefore, we looked into the ELSA core data, and then identified ten age-related diseases, each used as a class variable in this work. These diseases were angina, arthritis, cataract, dementia, diabetes, high blood pressure, heart attack, osteoporosis, Parkinson's and stroke. Hence, we created ten datasets, each one with a different disease as the class variable to be predicted. More precisely, in each dataset, the binary class variable indicated the presence or absence of the corresponding disease in wave 7 (the most recent wave in ELSA).

Note that, for each disease, there was no variable in the ELSA database that directly indicated whether or not an individual had that disease in a given wave. This kind of information was rather represented indirectly by several related variables whose values depend on both whether or not the individual (patient) had the target disease in the past and whether or not the patient still has the disease or whether the disease was first diagnosed in the current wave. Therefore, we needed to create a well-defined class variable for each disease separately, combining information from the several related variables associated with that disease. In order to create such class variables, in general, the following rule was used for each disease, combining information about that disease's variables in wave 7:

IF ("whether confirms the disease diagnosis" = "yes")
OR ("whether still has the disease" = "yes")
OR ("the disease diagnosis newly reported" = "yes")
THEN Disease = "yes"
OTHERWISE Disease = "no".

In this rule, the terms between double quotes just before each "=" sign in the "IF" condition refer to original variables in ELSA's wave 7 core data. Note that, although each dataset had a different class variable, all datasets contained instances representing the same individuals and the same set of predictive features (described next).

B. Creating predictive features based mainly on Nurse data

In the created datasets, most features were created from raw variables available in the Nurse Visit data, part of the previously discussed ELSA database [6]. Those raw variables represent several types of biomedical information collected by a nurse, including for instance many types of blood sample tests. In addition, the nurse took several physical performance measurements that involved asking a patient to move his/her body in different ways. If a particular movement could not be done by the participant or he/she felt that it was unsafe, the attempt was marked as 'Not attempted' or 'Test not completed'. The Nurse variables were only available at ELSA waves 2, 4, and 6, so our created datasets contained only features for these waves. These features were then used to predict age-related diseases (classes) at the later wave 7, whose data were collected about two years later than the data in wave 6.

As mentioned earlier, the raw biomedical variables collected by the nurses were not collected specifically for machine learning and they also contained a large amount of obviously redundant or irrelevant information. Hence, we have created features for classification by extracting and combining information from the raw variables in the Nurse data files, as follows. First of all, we kept potentially predictive variables from the Nurse data, whilst many other variables which seemed intuitively useless for predicting age-related diseases were removed because such variables were collected mainly to record problems in data collection for other variables. For example, several variables capturing information such as the reasons why taking a blood sample test was refused by a patient, and information about several types of problems in some physical performance measurements were discarded.

In addition, many variables in the Nurse data represented clearly redundant information in cases where the same variable (e.g. the result of a blood test) was measured three different time-points in the same wave in order to represent the variability in test results. This resulted in duplication of variables representing the same biomedical property in each wave, and none of those three measures can be considered 'better' than the other two. Hence, instead of using any of the three underlying variables, we created a feature defined as the mean value over those three measures, for each individual (instance), for each wave.

Another point to consider was the occurrence of different types of missing values in many raw variables in the Nurse data,

which were originally labelled as different negative values as follows (using as an example a blood test result variable):

- -1 = Not applicable
- -6 = Time from collection to receipt in the lab > 5 days
- -8 = Don't know
- -9 = Refusal
- -11 = Blood sample not taken

Considering all these types of missing values separately would considerably complicate the task of the classification algorithms. Hence, to simplify, all these different negative values were assumed to have the same meaning of "missing value", so that we treated them in the same way by replacing all of them with the missing value symbol "?" (used in WEKA).

In addition to features created from Nurse data, we also included in our datasets two features directly extracted from the Core files in ELSA which intuitively represent potentially very relevant information for predicting age-related diseases, namely the features "w7indager" (age) and "indsex" (gender).

Finally, the most important point when creating the instances used in our datasets was that only the data from "core" members were used, that is, the ELSA records of their partners were ignored. The ELSA variable *"*idauniq*"*, which was a unique id for each individual, was added to our datasets to match up data about the same core member in different dataset files (across different waves). This variable was not used for classification purposes as it had no predictive power. Note that an instance was created for an individual only if that individual participated in wave 7, so the class variable values were available for all individuals in all datasets. However, some individuals in our datasets may not have participated in all waves used to create features (waves 2, 4 and 6). If an individual did not participate in a given wave, the corresponding features in that wave would have a missing value for that individual, and the feature selection and classification algorithms cope with those missing values in their own ways.

C. Constructing Longitudinal Features

Recall that the features created from variables in the Nurse data (the vast majority of features in the created datasets) were measured across three different time-points (waves), namely waves 2, 4 and 6 of the ELSA database. We used the term "conceptual feature" to refer to the abstract concept of such a feature regardless of its observed value in any given wave. For instance, "*chol*" (Blood total cholesterol level) was a conceptual (abstract) feature which was associated with three actual features, *w2chol*, *w4chol* and *w6chol*, which represented the observed value of that variable in waves 2, 4, and 6. For each conceptual feature, we created new features trying to capture temporal trends in the variation of that feature's values across the three waves, as follows.

First of all, we created *m* groups of temporally related features, thus one group for each of the *m* conceptual features. Each group considered all temporal variations of a conceptual feature across waves 2, 4 and 6, which were the waves before the wave with the class to be predicted (wave 7). Thus, each group contained observed features that were the variations of a conceptual feature across different waves. In the next step, these observed features were used to create six different types of Constructed Longitudinal Features (CLFs). Note that these CLFs only work for continuous (real-valued) observed features.

The first CLF was *mono_w246*, indicating whether the value of a base feature monotonically increased or decreased across waves 2, 4 and 6; as follows. Let $f_{(2)}, f_{(4)}, f_{(6)}$ be numeric values of feature *f* in waves 2, 4, 6. Then, *f_mono_w246* (*mono_w246* for feature *f*) has the value 1 (monotonic increase) if $f_{(2)} < f_{(4)} < f_{(6)}$, value -1 (monotonic decrease) if $f_{(2)} > f_{(4)} > f_{(6)}$, or value 0 (no monotonic property) otherwise. However, a few features had their values observed in only two waves, so that a *mono_w246* variable for such features cannot be created using the rule mentioned above. For such features, we created instead the CLF $up\_wt_1t_2$, indicating whether the values of *f* in the two time-indices (wave numbers) t_1 and t_2 go up or not. For instance, *f_up_w24 has the value* 1 if $f_{(2)} < f_{(4)}$, or value 0 otherwise. Note that if the value of the feature is missing in any of the waves, either of these CLFs has a missing value (denoted by "?").

Each of the other CLFs represents the difference between the values of a pair of features referring to the same conceptual feature in two different waves. Let $f\_diff\_wt_1t_2$ denote the difference between the values of feature *f* in the two time-indices (wave numbers) t_1 and t_2, for each of the three pairs of waves where $t_2 > t_1$. Then, these CLFs are defined as follows:

- $f\_diff\_w24 = f_{(4)} - f_{(2)}$
- $f\_diff\_w46 = f_{(6)} - f_{(4)}$
- $f\_diff\_w26 = f_{(6)} - f_{(2)}$

Hence, positive (negative) values of these constructed features denote an increase (decrease) in the value of feature f with time.

Table 1 shows the full set of 44 conceptual features used in all the datasets created in this work. This table shows, for each conceptual feature, its name and its description or definition in the ELSA database [6], the data source used to create the features. Note that the first 2 features, name age and gender, had one value for each individual, whereas the other 42 rows represent features from the Nurse data in ELSA which, in general, were longitudinal features with different values across waves (time points) for each individual. 36 of these 42 longitudinal features had values in three waves, whereas the other 6 were only available in some waves. This could be explained as follows: one feature (apoe) occurred only in wave 2, three features (hipval, whval, htpf) occurred only in waves 2 and 4, and two features (wbc, mch) occurred only in waves 4 and 6. Since 5 conceptual features had values in only two waves, each of these generated 3 features in our datasets (one feature for each of the two waves plus two CLFs). Furthermore, out of the 36 conceptual features having values in 3 waves, there were 22 conceptual features whose values were continuous (real-valued). Therefore, each of those 22 conceptual features generated 7 features in our datasets (one feature for each wave plus four CLFs). Table 2 shows the six types of CLFs, as explained earlier. The total number of features is 219.

Regarding missing values, a common approach in standard non-longitudinal classification is to replace a missing value by a default value, typically the mean of the known values of the feature across the dataset, in the case of numerical features; or the mode, in the case of nominal features.

Table 1: All conceptual features used in the created data sets

Feature (Variable)	Description in the ELSA database, or definition	Available in			Numeric
		wave 2	wave 4	wave 6	
indsex	Sex - Priority: DiSex, DhSex	Not Applicable			
w6indager	Definitive age variable collapsed at 90+ to avoid disclosure	Not Applicable			✓
clotb	Blood sample: Whether has clotting disorder	✓	✓	✓	
fit	Blood sample: Whether ever had a fit	✓	✓	✓	
apoe	Blood APOE level (mmol/l)	✓			✓
hasurg	Lung function: Whether respondent had abdominal or chest surgery in last 3 weeks	✓	✓	✓	
eyesurg	Lung function: Whether respondent has had eye surgery in the last 4 weeks	✓	✓	✓	
hastro	Lung function: Whether admitted to hospital for heart complaint in last 6 weeks	✓	✓	✓	
chestin	Lung function: Whether respondent had any respiratory infection in last 3 weeks	✓	✓	✓	
inhaler	Lung function: Whether used an inhaler/puffer in last 24 hours	✓	✓	✓	
mmssre	Side-by-side stand: Outcome	✓	✓	✓	
mmstre	Semi-tandem stand: Outcome	✓	✓	✓	
mmftre2	(D) Outcome of full tandem stand according to age	✓	✓	✓	
mmlore	Leg raise (eyes open): Outcome	✓	✓	✓	
mmlsre	Leg raise (eyes shut): Outcome	✓	✓	✓	
mmcrre	Chair rise: Single chair rise outcome	✓	✓	✓	
mmrroc	(D) Chair rise: Outcome of multiple chair rises, split by age	✓	✓	✓	
hipval	(D) Valid Mean Hip (cm)	✓	✓		✓
whval	(D) Valid Mean Waist/Hip ratio	✓	✓		✓
htpf	Lung function: Highest technically satisfactory PF reading (litres per minute)	✓	✓		✓
wbc	White blood cell count (x 10^9 cells/litre)		✓	✓	✓
mch	Blood mean corpuscular haemoglobin level (pg/cell)		✓	✓	✓
sysval	(D) Valid Mean Systolic BP	✓	✓	✓	✓
diaval	(D) Valid Mean Diastolic BP	✓	✓	✓	✓
pulval	(D) Valid Pulse Pressure	✓	✓	✓	✓
mapval	(D) Valid Mean Arterial Pressure	✓	✓	✓	✓
cfib	Blood fibrinogen level (g/l)	✓	✓	✓	✓
chol	Blood total cholesterol level (mmol/l)	✓	✓	✓	✓
hdl	Blood HDL level (mmol/l)	✓	✓	✓	✓
trig	Blood triglyceride level (mmol/l)	✓	✓	✓	✓
ldl	Blood LDL level (mmol/l)	✓	✓	✓	✓
fglu	Blood glucose level (mmol/L) - fasting samples only	✓	✓	✓	✓
rtin	Blood ferritin level (ng/ml)	✓	✓	✓	✓
hscrp	Blood CRP level (mg/l)	✓	✓	✓	✓
hgb	Blood haemoglobin level (g/dl)	✓	✓	✓	✓
hba1c	Blood glycated haemoglobin level (%)	✓	✓	✓	✓
htval	(D) Valid height (cm)	✓	✓	✓	✓
wtval	(D) Valid weight (Kg) inc. estimated>130kg	✓	✓	✓	✓
bmival	(D) Valid BMI - inc estimated>130kg	✓	✓	✓	✓
wstval	(D) Valid Mean Waist (cm)	✓	✓	✓	✓
htfvc	Lung function: Highest technically satisfactory FVC reading (litres)	✓	✓	✓	✓
htfev	Lung function: Highest technically satisfactory FEV reading (litres)	✓	✓	✓	✓
mmgsd_me	Created variable: grip strength: dominant hand (Kg), mean of 3 measures (mmgsd1, mmgsd2, mmgsd3)	✓	✓	✓	✓
mmgsn_me	Created variable: grip strength: non-dominant hand (Kg), mean of 3 measures (mmgsn1, mmgsn2, mmgsn3)	✓	✓	✓	✓

Table 2: The six types of proposed Constructed Longitudinal Features (CLFs)

Feature (Variable)	Description in the ELSA database, or definition	Numeric
f_mono_w246	CLF: whether the value of VAR monotonically increases (1), decrease (-1), or otherwise (0)	
f_up_w24	CLF: whether the value of f increases (1), or not (0), from the wave 2 to wave 4	
f_up_w46	CLF: whether the value of f increases (1), or not (0), from the wave 4 to wave 6	
f_diff_w24	CLF: f value in wave 4 - f value in wave 2	✓
f_diff_w46	CLF: f value in wave 6 - f value in wave 4	✓
f_diff_w26	CLF: f value in wave 6 - f value in wave 2	✓

However, in our context of the constructed temporal difference features for longitudinal classification, we can exploit additional temporal information about feature values when calculating the value that will replace the missing value (instead of using a pre-defined default value), as follows.

Let *i* and *j* be the indices of two waves associated with a temporal difference feature based on a given feature *f*, denoted by (*f_diff_wij*). If the value of the base feature *f* is missing for a given individual (instance) *x* in one of those two waves (say wave *i*), and the value of *f* is known in the other two waves (*j* and *k*), then the missing value of the constructed *f_diff_wij* feature for *x* will be replaced by a value calculated by equation (1), where wave index *k* denotes the "third" wave (i.e. nor wave *i* nor wave *j*) available in the dataset, so that data from all three waves are used to estimate the missing value.

$$f\_diff\_wij_x = \frac{f\_diff\_wkj_x \times mean\_f\_diff\_wij}{mean\_f\_diff\_wkj} \qquad (1)$$

In equation (1), *mean_f_diff_wij* and *mean_f_diff_wkj* are the mean values of all known values of the constructed *f_diff features* for the corresponding waves. For example, if the value of *f* is missing in wave 4 for a given individual *x*, the value of the constructed feature *f_diff_w24* for *x* is computed as:

$f\_diff\_w26_x \times (mean\_f\_diff\_w24_x / mean\_f\_diff\_w26_x)$.

The motivation for this approach is that it considers not only the known values of *f* for other individuals in wave *i*, but also the known values of *f* for both the same individual and other individuals in waves *j* and *k*. In other words, the ratio *mean_f_diff_wij* to *mean_f_diff_wkj* acts as a normalization factor, correcting for the different scales of *f_diff* values in different time periods. Note that this method only copes with the missing values for the constructed features, i.e., it does not attempt to fill in the missing values for the base feature. This latter possibility is left for future research.

IV. The Proposed Variant of Correlation-Based Feature Selection for Longitudinal Data

The proposed variant of the CFS method is based on the idea of first dividing the set of features into groups of temporally related features, with one group for each conceptual feature (see Section III-C). Each group contains two types of features: (a) all features representing different values of a conceptual feature across the different waves (time points); and (b) Constructed Longitudinal Features (CLFs) for the corresponding conceptual feature. For instance, the group of features for the conceptual feature "chol" (cholesterol level) contains seven features: w2chol, w4chol, w6chol, chol_mono_w246, chol_diff_w24, chol_diff_w46 and chol_diff_w26; where the first 3 features are the chol values at waves 2, 4 and 6, and the last four features are CLFs.

In general, exhaustive search evaluates all possible feature subsets and selects the best candidate feature subset based on the CFS Merit function. For a given set of *n* candidate features, the time complexity of this method is 2^n. The exhaustive search method is computationally feasible only if the number of candidate features is relatively small. This is the case for each feature group in this work, where the number of features in each group is at most 7 (three observed features and four CLFs). Therefore, in order to address the temporal redundancy problem mentioned in Section III, the exhaustive search is applied to each feature group separately. We call this the Exhaustive CFS per Group (Exh-CFS-Gr) method. Afterwards, we merge all groups of selected features, so a single feature subset is obtained and output as the result of the feature selection process. The basic idea of the proposed Exh-CFS-Gr method is summarized in graphical form in Figure 1.

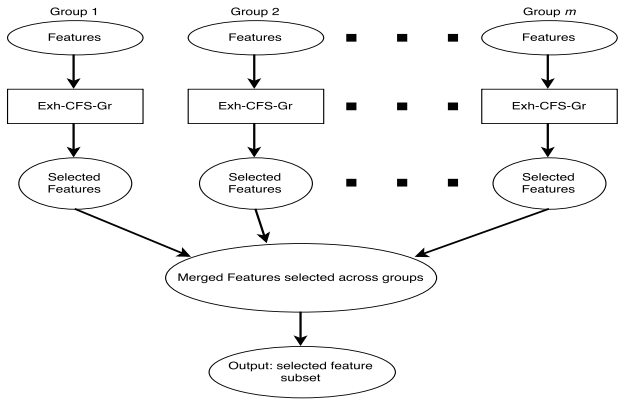

Figure 1: The basic idea of the proposed Exh-CFS-Gr method

V. COMPUTATIONAL RESULTS

A. Experimental Methodology

We report results for 10 datasets created from the ELSA data, as described earlier. Recall that each dataset had a different age-related disease in wave 7 as the class variable to be predicted, whilst all datasets had the same predictive features (derived in general from waves 2, 4, and 6). Predictive accuracy was measured by the F-measure, the harmonic mean between Precision and Recall [10], given by equation (2),

$$F\text{-}measure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (2)$$

where Precision is the proportion of instances predicted as positive which are really positive and Recall is the proportion of positive instances that were correctly predicted as positive. To compute these measures, each class label (presence or absence of the disease) was considered in turn as the positive class and the reported F-measure is the arithmetic (unweighted) mean of the F-measures for the two class labels. We report results for three feature selection approaches: the proposed Exh-CFS-Gr method, standard CFS, and no feature selection. Each of these approaches was evaluated using two classification algorithms, namely Naïve Bayes and the decision tree algorithm J48. All experiments were performed using the WEKA tool [11] using 10-fold cross-validation.

B. Predictive Accuracy Results

Table 3 compares the F-measure values obtained by Naïve Bayes (NB), NB using features selected by standard CFS (ignoring temporal information), and NB using the proposed Exh-CFS-Gr method (exploiting temporal information).

As shown in Table 3, NB using the proposed Exh-CFS-Gr obtained the best result in 6 out of the 10 datasets, whilst NB using standard CFS obtained the best result in 4 datasets. In all 10 datasets, NB without feature selection obtained the worst result (joint with standard CFS in the Dementia dataset). Also, NB with Exh-CFS-Gr obtained the best (lowest) average rank.

The Wilcoxon signed-ranks test [12] was used to compare the performances of two different methods. The main advantages of this test are its robustness against outliers and its non-parametric nature, making no assumption of normal distribution [10]. We are trying to reject the null hypothesis that Naïve Bayes (NB) with a given feature selection method obtains an F-measure value that is not significantly different from NB with another feature selection method (or just NB, without feature selection). We used the test with a significance level of $\alpha = 0.05$, and $N = 10$ (10 datasets) in our experiments.

First, comparing NB with Exh-CFS-Gr against NB without feature selection, the null hypothesis is rejected with a p-value of 0.005. Next, comparing NB with standard CFS against NB with no feature selection, the null hypothesis is rejected with a p-value of 0.008. Lastly, comparing NB with Exh-CFS-Gr against NB with standard CFS, the p-value is 0.959, so the null hypothesis cannot be rejected. To summarize, although there was no statistical evidence supporting that NB with Exh-CFS-Gr performed better than NB with the standard CFS. Both Exh-CFS-Gr+NB and CFS+NB performed significantly better than Naïve Bayes with no feature selection.

Table 3: F-measure values obtained by Naïve Bayes, after applying different CFS methods. The best F-measure value for each dataset (across all feature selection methods) is shown in boldface. The last row shows the average ranks.

Dataset	NB (No Feature Selection)	standard CFS + NB	Exh-CFS-Gr + NB
Angina	0.559	0.562	**0.576**
Arthritis	0.614	0.625	**0.629**
Cataract	0.641	**0.677**	0.658
Dementia	0.589	0.589	**0.603**
Diabetes	0.732	**0.760**	0.733
HBP	0.672	**0.693**	0.676
HeartAtt	0.606	**0.620**	0.619
Osteoporosis	0.610	0.613	**0.618**
Parkinsons	0.553	0.560	**0.570**
Stroke	0.590	0.602	**0.610**
Average Rank	2.95	1.65	**1.40**

Table 4 compares the F-measure values obtained by the decision tree algorithm J48 using all features (no feature selection in a pre-processing phase), by J48 using as input the features selected by standard CFS, and by J48 using as input the features selected by the proposed Exh-CFS-Gr. J48 using Exh-CFS-Gr obtained the best result in 5 out of the 10 datasets. J48 using standard CFS obtained the best result in three datasets, and J48 with no feature selection in a pre-processing phase was the winner in just two datasets.

Table 4: F-measure values obtained by J48 after applying different CFS methods. The best F-measure value for each dataset (across all feature selection methods) is shown in boldface. The last row shows the average ranks.

Dataset	J48 (No Feature Selection)	standard CFS + J48	Exh-CFS-Gr + J48
Angina	0.550	0.540	**0.550**
Arthritis	0.610	**0.620**	0.610
Cataract	0.670	0.670	**0.670**
Dementia	0.580	**0.590**	0.580
Diabetes	**0.770**	0.760	0.750
HBP	0.660	0.660	**0.670**
HeartAtt	0.610	**0.610**	0.600
Osteoporosis	0.610	0.610	**0.620**
Parkinsons	**0.590**	0.580	0.580
Stroke	0.600	0.590	**0.600**
Average Rank	2.05	2.00	**1.95**

Unlike the Naïve Bayes algorithm, the J48 algorithm obtained very similar average ranks for all three approaches (with two CFS versions and no CFS). Hence, J48 benefited less from feature selection in a pre-processing phase than NB. This

can be explained by the fact that, unlike NB, J48 performs embedded feature selection [13].

Again, we used the Wilcoxon signed-ranks test with significance level $\alpha = 0.05$ and $N = 10$ to evaluate if there was a significant difference in predictive performance between two methods for each of the three pairs of methods: J48 with Exh-CFS-Gr against J48 only, J48 with CFS against J48 only, and J48 with Exh-CFS-Gr against J48 with CFS. None of the three null hypotheses could be rejected, with p-values 0.953, 0.443 and, 0.959 respectively. In other words, there was no statistical evidence supporting that Exh-CFS-Gr+J48 performed better than CFS+J48 or J48 without feature selection in a preprocessing phase.

C. Discussion on the most relevant features selected by J48, using as input the features selected by Exh-CFS-Gr

For each dataset (each with a different age-related class variable), we looked at the decision tree built by J48 from the full dataset using as input the features selected by Exh-CFS-Gr. In each decision tree, we observed which feature was selected at the root node – and so it was used to classify all instances.

First, "*age*" was selected as the root node in four datasets: Stroke, Dementia, Cataract, Parkinson's. This was not surprising, since in our datasets the classes were age-related diseases. In the Parkinson's dataset, "*age*" was the only feature selected by J48.

For other datasets, in the following list of root features, the prefixes "w6" and "w2" at the start of a feature name denote that they were features observed in waves 6 and 2, respectively.

The six other root features were: "*w6LDL*" in the Heart Attack dataset, "*w2mmstre*" for Angina, "*w6hba1c*" for Diabetes, "*w2sysval*" for High Blood Pressure, "*w6mmgsd_me*" for Arthritis, and "*gender*" for Osteoporosis.

A brief description of these features can be found in Table 1, whilst a more detailed explanation can be found in the ELSA documentation. In essence, *LDL* (*Low Density Lipoprotein*) is known as the "bad" cholesterol (having a large amount of it is unhealthy), *Mmstre* refers to the patient's ability to keep their balance whilst standing for 10 seconds in a semi-tandem position, *Hba1c* is a measure of average plasma glucose concentration often used for testing if a patient has diabetes, *Sysval* means systolic blood pressure, and *Mmgsd_me* is a measure of grip strength. The choice of "*gender*" as the root node in the Osteoporosis dataset was natural, given that osteoporosis is more common in women than in men.

VI. CONCLUSION

In conclusion, the results of our experiments showed that there was a statistically significant improvement in the predictive accuracy when the proposed Exh-CFS-Gr was used as a feature selection method before running Naïve Bayes (NB) by comparison with the baseline approach of running Naïve Bayes with no feature selection. Moreover, overall this CFS variant obtained somewhat higher predictive accuracy than NB with standard CFS, which suggests some progress. In contrast with the results for NB, even though J48 with Exh-CFS-Gr achieved the best average rank, it showed no statistically significant difference in predictive accuracy when compared against J48 only. In fact, standard CFS did not improve the predictive accuracy of J48 either. The J48 algorithm constructs decision trees as classification models, following an embedded feature selection approach. Hence, irrelevant and/or redundant features have a smaller effect on J48 models compared with NB models.

REFERENCES

[1] J. Li *et al.*, "Feature Selection: A Data Perspective," *arXiv Prepr. arXiv*, 2016.

[2] H. and H. M. Liu, *Feature Selection for Knowledge Discovery and Data Mining*. Springer US, 1998.

[3] L. Wang, Y. Wang, and Q. Chang, "Feature selection methods for big data bioinformatics: A survey from the search perspective," *Methods*, vol. 111. pp. 21–31, 2016.

[4] C. E. Ribeiro, L. H. S. Brito, C. N. Nobre, A. A. Freitas, and L. E. Zárate, "A revision and analysis of the comprehensiveness of the main longitudinal studies of human aging for data mining research," *WIREs Data Min. Knowl. Discov.*, vol. 7, no. 3, p. e1202, 2017.

[5] M. A. Hall, "Feature Selection for Discrete and Numeric Class Machine Learning 1 Introduction," *Mach. Learn. Proc Seventeenth Int. Conf. Mach. Learn.*, pp. 1–16, 2000.

[6] NatCen Social Research, "English Longitudinal Study of Ageing," *Encyclopedia of Geropsychology*, 2016. .

[7] M. Radovic, M. Ghalwash, N. Filipovic, and Z. Obradovic, "Minimum redundancy maximum relevance feature selection approach for temporal gene expression data," *BMC Bioinformatics*, vol. 18, no. 1, p. 9, Jan. 2017.

[8] C. Furlanello, S. Merler, and G. Jurman, "Combining feature selection and DTW for time-varying functional genomics," *IEEE Trans. Signal Process.*, vol. 54, no. 6 II, pp. 2436–2443, 2006.

[9] Q. Lou and Z. Obradovic, "Analysis of temporal high-dimensional gene expression data for identifying informative biomarker candidates," in *Proceedings - IEEE International Conference on Data Mining, ICDM*, 2012, pp. 996–1001.

[10] N. Japkowicz and M. Shah, "Evaluating Learning Algorithms," p. 423, Cambridge University Press, 2011.

[11] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software," *ACM SIGKDD Explor. Newsl.*, vol. 11, no. 1, p. 10, Nov. 2009.

[12] F. Wilcoxon, "Individual Comparisons by Ranking Methods," *Biometrics Bull.*, vol. 1, no. 6, p. 80, 1945.

[13] V. Sugumaran, V. Muralidharan, and K. I. Ramachandran, "Feature selection using Decision Tree and classification through Proximal Support Vector Machine for fault diagnostics of roller bearing," *Mech. Syst. Signal Process.*, vol. 21, no. 2, pp. 930–942, 2007.