FEATURE SELECTION AND TIME REGRESSION SOFTWARE: APPLICATION ON PREDICTING ALZHEIMER'S DISEASE PROGRESS

Dimitrios Ververidis, Mark Van Gils, Juha Koikkalainen and Jyrki Lötjönen

VTT Technical Research Center of Finland, P.O. Box 1300, 33101 Tampere, Finland,
e-mail: Ext-Dimitrios.Ververidis@vtt.fi; Mark.vanGils@vtt.fi; Juha.Koikkalainen@vtt.fi; Jyrki.Lotjonen@vtt.fi

ABSTRACT

In this paper, the Bayes classifier is used to predict Alzheimer's disease progress. The classifier is trained on a subset of the Alzheimer's Disease Neuroimaging Initiative database. Subjects are diagnosed by doctors as belonging to healthy, mild-cognitive impaired, and Alzheimer's disease class. A software tool for features selection and time regression is developed. The tool utilizes a variant of the Sequential Forward Selection (SFS) algorithm for feature selection, where the criterion used for selecting features is the correct classification rate of the Bayes classifier. The tool also employs linear regression to predict future values of selected biomarkers, such as the hippocampus volume, from past measurements, so that future class of the subject can be predicted.

1. INTRODUCTION

Great effort has been made to find a drug that slows down Alzheimer's disease (AD) progress. Unfortunately, until now no such drug has been found, mainly because there is no biomarker to track faithfully the progress of the disease [1]. Clinicians use neuropsychological features (tests) that track dementia level in order to decide if a subject is healthy, early onset AD patient, or AD patient. An intermediate group is formed by MCI (mild-cognitive impairment) cases; persons who are not yet diagnosed with Alzheimer. These can be subdivided further into persons who either may develop into AD (so called progressive MCI), or may stay at the current level (so called stable MCI). Early prediction of in which groups a certain person falls is of high importance from a health-care point of view.

Biomarkers related to AD are divided into several categories. One category of these biomarkers is the concentration of β-amyloid peptides in the cerebrospinal fluid (CSF). In AD patients, most of β-amyloid peptides are accumulated in the brain, preventing inter-neuron communication. Therefore, low concentration of these peptides is expected in CSF [1]. The metabolic activity of the brain calculated with positron emission tomography (PET) scans is another biomarker category. Low metabolic activity in neurons of AD patients has been reported because neurons that do no communicate to each other do not metabolize any substance [2]. Furthermore, shapes and sizes of brain parts estimated with anatomical magnetic resonance imaging (MRI) is also a category of biomarkers. The brain shrinks because neurons that do not operate eventually die [3]. High cholesterol levels in blood [4], and certain genes expressions [**?**] have been related also to AD.

In order to select the features, that discriminate AD, MCI, and Healthy subjects, either neuro-psychological or biomarkers, a wrapper scheme was selected. The wrapper employs the cross-validated correct classification rate (CCR) of the Bayes classifier to find a feature subset that maximizes CCR when classifying subjects into AD, MCI, and Healthy classes. In a previous investigation, a statistical variant of the Sequential Forward Selection algorithm, denoted here as StatSFS, was proposed [6]. It is faster and more accurate than the standard SFS due to statistical tests for preliminary rejection of a feature and comparisons of CCRs with confidence limits that depend on cross-validation variance. Later, in another investigation, the loss of classification information due to the curse of dimensionality was calculated [7]. This loss of information was employed to find a lower bound of CCR to guarantee the performance of the selected feature set. Both methods are incorporated into the wrapper used throughout this paper called as *InfoStatSFS*. The class-conditional probability density function (pdf) is modeled as a multivariate Gaussian in order to maintain low execution time [6]. In order to find the progress of the AD, a method to estimate future measurement values of the selected features based on linear regression with least squares training method is proposed.

The outline of this paper is as follows. In Section 2, the feature selection, the linear regression on features, the data, and the software tool used in the experiments are described. Experimental results are reported in Section 3. Finally, conclusions are drawn in Section 4.

2. METHODS

Let us denote the set of subjects $\mathscr{U} = \{u_i\}_{i=1}^{N}$ where N is the total number of subjects. Each subject is treated as a pattern

$$u_i = \left\{\underline{x}_i^{\mathscr{W}}(t), c_i(t)\right\}_{t=0}^{T_{iE}} \quad (1)$$

where T_{iE} is the expected life period in months of u_i; $\underline{x}_i^{\mathscr{W}}(t) = [x_{i1}(t) \;\; x_{i2}(t) \;\; \ldots \;\; x_{iD}(t)]$ is the D measurements vector on time stamp t taken from u_i; $\mathscr{W} = \{w_d\}_{d=1}^{D}$ is the whole feature set of D features that it is measured on a subject; and $c_i(t) \in \{\Omega_1, \Omega_2, \Omega_3\}$ is the class that u_i belongs at t, where Ω_1=AD, Ω_2=MCI, Ω_3=Healthy.

D.Ververidis work was carried out during the tenure of an ERCIM fellowship.

M. Van Gils, J. Koikkalainen, and J. Lötjönen's work is partially funded under the 7th Framework Programme by the European Commission (http.//cordis.europa.eu/ist; EU-Grant-224328-PredictAD.)

 ISSN 2076-1465

2.1 Feature Selection

The InfoStatSFS consists of an internal step and an external step. In the internal step several feature sets of the same cardinality are compared. Let d be the instance counter of the internal step, that is also the dimensionality of the selected feature set. Initially ($d = 0$), the subset of selected features $\mathscr{Z}_d = \emptyset$.

We seek the feature $w^+ \in \mathscr{W} - \mathscr{Z}_d$ to include in $\mathscr{Z}_d$ such that

$$w^+ = \underset{w \in \mathscr{W} - \mathscr{Z}_d}{\operatorname{argmax}} \operatorname{MCCR}_B(\mathscr{Z}_d \cup \{w\} \mid \mathscr{U}), \tag{2}$$

where $\operatorname{MCCR}_B(\mathscr{Z}_d \cup \{w\} \mid \mathscr{U})$ is the average correct classification rate over B cross-validation repetitions using $\mathscr{Z}_d \cup \{w\}$ estimated on $\mathscr{U}$, i.e.

$$\operatorname{MCCR}_B(\mathscr{Z}_d \cup \{w\} \mid \mathscr{U}) = \frac{1}{B} \sum_{b=1}^{B} \operatorname{CCR}_b(\mathscr{Z}_d \cup \{w\} \mid \mathscr{U}). \tag{3}$$

B is estimated in our previous investigation [6]. In a cross-validation repetition b, the patterns set $\mathscr{U}$ is split into a design set $\mathscr{U}_{\mathscr{D}b}$ containing $0.9N$ patterns, and a remaining test set $\mathscr{U}_{\mathscr{T}b}$ that contains $N_{\mathscr{T}} = 0.1N$ patterns.

The estimate of correct classification rate (CCR) in repetition b using feature set $\mathscr{Z}_d \cup \{w\}$ is

$$\operatorname{CCR}_b(\mathscr{Z}_d \cup \{w\} \mid \mathscr{U}) = \frac{\mathrm{y}_b^{\mathscr{Z}_d \cup \{w\}}}{N_{\mathscr{T}}}, \tag{4}$$

where $\mathrm{y}_b^{\mathscr{Z}_d \cup \{w\}}$ is the number of subjects in the test set that are correctly classified in repetition b, when using feature set $\mathscr{Z}_d \cup \{w\}$. Then,

$$\mathrm{y}_b^{\mathscr{Z}_d \cup \{w\}} = \sum_{u_i \in \mathscr{U}_{\mathscr{T}b}} \mathscr{L}[c_i, \hat{c}_i], \tag{5}$$

where $\mathscr{L}[c_i, \hat{c}_i]$ denotes the *zero-one loss function* between the label c_i and the predicted class label $\hat{c}_i$ returned by the Bayes classifier for u_i.

Instead of using $\underset{w \in \mathscr{W} - \mathscr{Z}_d}{\operatorname{argmax}} \operatorname{MCCR}_B(\mathscr{Z}_d \cup \{w\} \mid \mathscr{U})$ operator, statistical comparisons of MCCRs with a t-test have been employed for accurate results. More details can be found in our previous investigation [6].

In the external step, feature sets of different cardinality, $d = 1, 2, \ldots, D'$, are compared. Each feature set is the same as the previous feature set plus one feature, the one found in the internal step. D' is found as follows. The selected feature set is increasing until the classification information loss due to the curse of dimensionality exceeds 50% [7]. For example, for the set used that consists of 270 patterns per class, classification information loss exceeds 50% when more than D'=34 features are selected. The feature set $\mathscr{Z}_{D_{\text{opt}}}$ that achieves the maximum lower limit of CCR is the optimum one [7].

2.2 Linear Regression

Some features are measured after 6 and 12 months from the first measuring time. The first time of measuring is called as time 0 or screening time for a patient. However, distant-future measurements that are particularly interesting are unknown, and therefore, a classification of the subject into the three classes in distant-future is not feasible. Regression is used for obtaining distant-future measurements that can be used to predict when an MCI patient will become an AD one, which is important information when a drug is tested whether it delays AD progress or not. Additionally, past measurements can be obtained in the same manner. The past measurements will be used in order to estimate the time stamp that the subject became AD patient, i.e. how far the AD has gone.

For a certain subject u_i and a certain feature w_d, some measurements $x_{id}(t)$ for $t = T_{1i}, T_{2i}, \ldots, T_{\Lambda i}$ are available, where Λ is the number of measurement values in the certain time frame. Regression estimates $x_{id}(t)$ for $t < T_{1i}$ or $t > T_{\Lambda i}$. In order to estimate future or past values of a certain feature w_d on a certain subject u_i outside the known time frame, the linear model

$$\hat{x}_{id}(t) = \hat{a}_{id} t + \hat{b}_{id} \tag{6}$$

is employed, where the uknown parameters $\hat{a}_{id}$ and $\hat{b}_{id}$ are found with least squares method, i.e.

$$\hat{a}_{id} = \frac{\sum t x_{id}(t) - \Lambda^{-1} \sum t \sum x_{id}(t)}{\sum t^2 - \Lambda^{-1} (\sum t)^2}, \quad \text{and} \tag{7}$$

$$\hat{b}_{id} = \Lambda^{-1} (\sum x_{id}(t) - \hat{a}_{id} \sum t) \tag{8}$$

where $\sum$ stands for $\sum_{t=T_{1i}}^{T_{\Lambda i}}$.

2.3 Data

The biomarker measurements are obtained from Alzheimer's disease Neuroimage Initiative (ADNI) database which is publicly available [8]. The subset of ADNI used here consists of 2712 neuropsychological and biomarker features measured over 819 subjects (patterns). 800 subjects were used in the experiments as 19 out of 819 subjects had no label. The distribution of subjects at time 0 into classes is: 185 subjects are AD patients, 389 are MCI patients, and 226 subjects are healthy. The ground truth of the pattern is the clinician's diagnosis, which may not be always correct, but it is assumed in our experiments as the 'ultimate' truth. The features are divided into categories. Some of the 80 categories of ADNI subset used are 'Demographic', 'Vital signs', 'MRI', 'PET', 'CSF', 'Mini-mental exams', etc. Category information is important because experiments should be contacted separately for neuro-psychological and biomarker feature sets.

An accurate decision about the discrimination information of the feature can not be taken when many measurements are missing. If less than 10% of the feature is missing then the values missing are replaced with the feature mean, otherwise feature is discarded. Discrete measurements can cause singularities during the estimation of the covariance matrix in Gaussian pdf estimation. When the unique measurements are less than 50, then a small variance (0.01) noise is added to the whole feature measurements. For certain subjects, certain features are measured again after a period of 6 and 12 months. These measurements allow us to use regression in order to predict distant-future feature values.

2.4 Software tool

A software in Matlab encompassing all the modalities explained in the previous section is presented in Figure 1.

9. Visualization of the patterns × selected features matrix during the external step. Patterns are sorted according to targets so as the discrimination information of a feature can be seen, i.e. feature 857 differs significantly above AD, MCI, and Healthy parts of the targets ribbon.

6. The class information is plotted as a color ribon.

7. The internal step. It can be viewed in real-time.

12. Button with pattern name leads to class-conditional histograms per selected feature and certain pattern values overlaid as in Figure 2.

1. The user can define optionally categories of features that should not be used in feature selection process.

13. Upper multi-color bar visualizes the class-conditional probability of the certain pattern. Bar's length represents probability.

2. A report about the features discarded due to missing values.

14. Prediction according to class of highest probability.

15. Lower single color bar contains the ground truth.

3. A list of the features transformed from discrete to continuous.

16. Button for pattern future classification using linear regression as in Figure 3.

17. A new pattern measurement vector can be provided 'live'.

4. The features available for feature selection process.

18. The confusion matrix for the classified patterns.

5. Visualization of patterns × features matrix in an image. Rows are the features and columns are the patterns.

8. The external step. It can be viewed in real-time.

10. Indices of selected features during the external step.

11. The confusion matrix found with averaging confusion matrices over cross-validation repetitions for the selected feature set.

Figure 1: Feature selection and linear regression software. Abbreviations: Use: Features available for feature selection; UnqVal: unique valued features that are discarded; Low: low presence features that are discarded; Out: excluded by the user features; Contin.: continuous features; Discr.: discrete features.

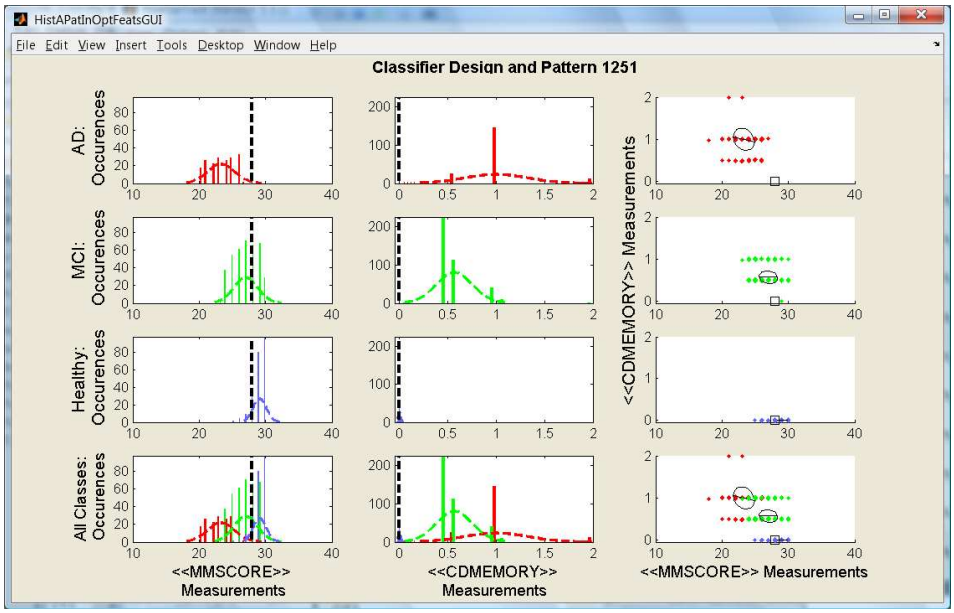

Figure 2: Features MMSCORE and CDMEMORY for all patterns against pattern 1251.

3. EXPERIMENTS

3.1 Feature selection results

Feature selection by InfoStatSFS is performed separately for neuro-psychological features and biomarkers. In Table 1, the selection results for neuro-psychological features are presented. InfoStatSFS achieved 94.2±1.5% CCR when the selected features are the clinical dementia rating (857, CDMEMORY) and mini mental state exam (215, MMSCORE). First, CDMEMORY is selected as it achieves 90.5% CCR. Second, MMSCORE is added as it improves CCR by 4%. The lower CCR is the lower limit of CCR due to curse of dimensionality. The dimensionality curse did not affect strongly CCR because only two features were selected. It was found that 44 cross-validation repetitions were enough to produce a confidence interval of ± 1.5% at 95% confidence level. The pdfs of the two features can be seen in Figure 2. First column contains histograms that present distribution of MMSCORE over the classes AD, MCI, and Healthy. Similarly, second column corresponds to feature CDMEMORY. In third column both features are plotted in 2 dimensional scatter plots. Rows correspond to classes. In last row all classes are plotted. The measurements of pattern 1251 are indicated by a dashed line in histogram plots, and by a square in 2D scatter plots. It can be inferred that CDMEMORY and MMSCORE are inversely proportional. So, CCR was not improved greatly by the second feature.

The optimum biomarkers selected by InfoStatSFS are reported in Table 2. The maximum CCR achieved is 57.9%. The curse of dimensionality for 9 features slightly affected the result by -0.8% as it is seen in the last row. Features CEREB8L, PRECUNEUSL, HIPPR, and CALCARINEL belong to UA (Gene Alexander) MRI SPM voxel based morphometry (VBM) analysis. HMT16 belongs to the category Laboratory Data. VSTMPSRC belongs to Vital Signs. APGEN2 belongs to ApoE genotyping. AVGJACOB is the Average Jacobian - Temporal (Paul Thompson's Lab). LONISID belongs to MRI MPRAGE Ranking. The pdf of feature HIPPR (hippocampus volume-right part) which is plotted in the upper-right part of Figure 3. It is seen that hippocampus volume is smaller in MCI than it is in Healthy subjects, and smaller in AD than it is in MCI subjects.

Table 1: InfoStatSFS cross-validated CCR results in % for selecting neuropsychological tests. Conf. interv. stands for confidence interval.

Step	Feature (ADNI Index)	CCR	Lower CCR	Conf. interv.
1	CDMEMORY (857)	90.5	90.5	±1.5
2	MMSCORE (215)	**94.2**	**93.9**	±**1.5**

3.2 Time regression results on biomarkers

Classification results over time for the subject 1382 are plotted in Figure 3. The hippocampus volume (HIPPR) measurements for time 0 and time +12 months were used for linear regression, and the predicted values are shown with asterisk in the upper-left part of the figure. It is inferred that the hip-

Table 2: InfoStatSFS cross-validated CCR results in % for selected biomarkers.

Step	Feature (ADNI Index)		CCR	Lower CCR	Conf. interv.
1	HMT16	(1028)	49.9	49.9	±1.5
2	VSTMPSRC	(8)	50.2	50.2	±1.5
3	CEREB8L	(2125)	51.5	51.4	±1.5
4	APGEN2	(1070)	52.9	52.6	±1.5
5	PRECUNEUSL	(2089)	53.9	53.6	±1.5
6	AVGJACOB	(2010)	55.0	54.5	±1.5
7	LONISID	(1439)	55.1	54.6	±1.5
8	HIPPR	(2060)	55.9	55.2	±1.5
9	CALCARINEL	(2065)	**57.9**	**57.1**	±**1.5**

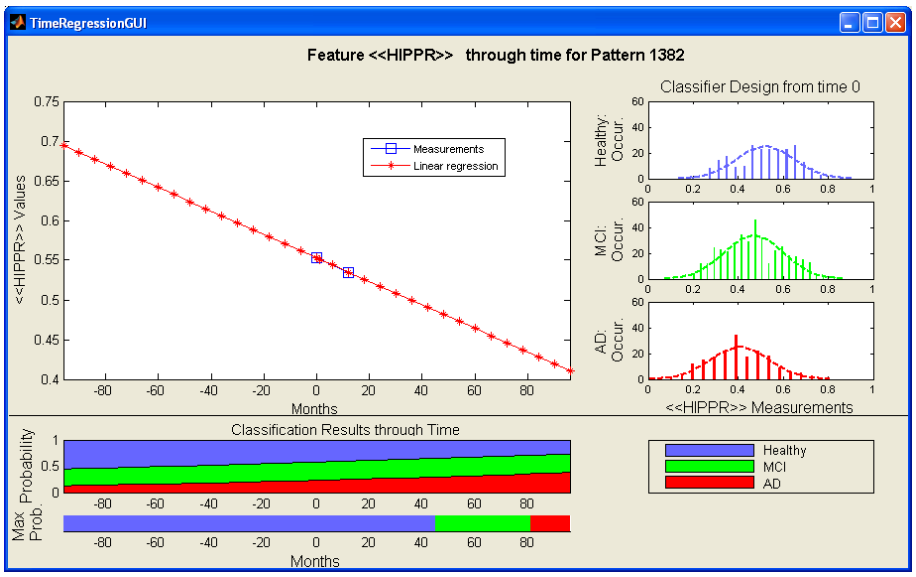

Figure 3: Future and past classification of subjects according to hippocampus volume values predicted by linear regression for plus 96 months, minus 96 months from the first recording of the certain feature. Class-conditional is estimated from Gaussian modeled pdfs fitted on screening dataset at time 0.

pocampus will become smaller as time lapses. The Bayes classifier employs the class-conditional pdfs for all subjects at time 0 plotted in the upper-right part of figure. The class-conditional probabilities over time are plotted as areas in the left-down part of the figure. It is seen that AD probability increases as time lapses, whereas probability of the subject being healthy becomes smaller. Below the areas of probability, the classification result of the classifier is plotted. It is observed that healthy patient is expected to become an MCI one after 42 months, and an AD one after 84 months from 0 time point.

4. CONCLUSIONS

From the feature selection results, it is inferred that neuropsychological tests outscored biomarkers with 94% against 58% correct classification rate, with random classification being at the level of 33%. Actually, it was expected that neuropsychological tests MMSCORE and CDMEMORY will be selected as best features and will achieve high classification score because clinicians rely on them for AD diagnosis. The most useful biomarkers are the ones related to the brain volume. It is observed that brain is smaller in MCI and AD subjects than in healthy subjects, as it is also observed in [1, 3]. Among the brain parts, hippocampus was the most informative. Its volume distribution fits well in the Gaussian model employed in the classifier, which was not the case for features HMT16 or APGEN2 that present discrete distribution with 2 values. Therefore, a pdf modeled by a Gaussian mixture will be tested in future experiments.

The linear regression method proposed can be used by clinicians to predict when a healthy subject will become MCI or AD patient. Based on two measurements a future value of a feature is estimated for a certain subject. However, currently the ADNI database does not contain enough measurements for accurate predictions of features. ADNI is expected to contain more feature measurements for more subjects over a greater time frame, so that confidence limits about linear regression estimates can be employed.

REFERENCES

[1] G. Miller, "Alzheimer's biomarker initiative hits its stride," *Els. Neurosc. meth.*, vol. 326, pp. 386–389, 2009.

[2] J. Ramírez, J. Górriz, D. Salas-Gonzalez, A. Romero, M. López, I. Álvarez, and M. Gómez-Río, "Computer-aided diagnosis of Alzheimer's type dementia combining support vector machines and discriminant set of features," *Els. Inf. Sciences*, vol. In Press, Corr. Proof, 2009.

[3] J. Lötjönen, R. Wolz, J. Koikkalainen, L. Thurfjell, G. Waldemar, H. Soininen, D. Rueckert, and The Alzheimer's Disease Neuroimaging Initiative, "Fast and robust multi-atlas segmentation of brain magnetic resonance images," *Els. Neuroimage*, vol. 49, no. 3, pp. 2352–2365, 2009.

[4] S. Vasto, G. Candore, F. Listí, C. Balistreri, G. Colonna-Romano, M. Malavolta, D. Lio, D. Nuzzo, E. Mocchegiani, D. Di Bona, and C. Caruso, "Inflammation, genes and zinc in Alzheimer's disease," *Els. Brain research reviews*, vol. 58, pp. 96–105, 2008.

[5] W. Liang, T. Dunckley, and T. B. et al., "Neuronal gene expression in non-demented individuals with intermediate Alzheimer's disease neuropathology," *Els. Neurobiology of Aging*, vol. 31, no. 4, pp. 549–566, 2010.

[6] D. Ververidis and C. Kotropoulos, "Fast and accurate feature subset selection applied to speech emotion recognition," *Els. Signal Processing*, vol. 88, no. 12, pp. 2956–2970, 2008.

[7] D. Ververidis and C. Kotropoulos, "Information loss of the Mahalanobis distance in high dimensions: Application to feature selection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 12, pp. 2275–2281, 2009.

[8] The Alzheimer's Disease Neuroimaging Initiative, "Database home page," www.loni.ucla.edu/ADNI.