A Class Imbalance Ordinal Method for Alzheimer’s Disease Classification

Conference Paper  · June 2018
DOI: 10.1109/PRNI.2018.8423960

CITATIONS
0

READS
26

3 authors:

Ricardo Cruz
INESC TEC, Porto, Portugal
8 PUBLICATIONS  35 CITATIONS
SEE PROFILE

Margarida Silveira
Technical University of Lisbon
45 PUBLICATIONS  648 CITATIONS
SEE PROFILE

Jaime S. Cardoso
University of Porto
239 PUBLICATIONS  2,542 CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Project
Learning in evolving video streams View project

Project
ZATLAB View project
A Class Imbalance Ordinal Method for Alzheimer’s Disease Classification

Ricardo Cruz∗, Margarida Silveira†, Jaime S. Cardoso∗
for the Alzheimer’s Disease Neuroimaging Initiative
∗INESC TEC and
Faculty of Engineering, University of Porto
Email: rpcruz@inesctec.pt, jaime.cardoso@inesctec.pt
†ISR and
Instituto Superior Técnico, Universidade de Lisboa
Email: msilveira@isr.tecnico.ulisboa.pt

Abstract—The majority of computer-aided diagnosis methods for Alzheimer’s disease (AD) from brain images either address only two stages of the disease at a time (and reduce the problem to binary classification) or do not exploit the ordinal nature of the different classes. An exception is the work by Fan et al. [1], which proposed an ordinal method that obtained better performance than traditional multiclass classification. Still, special care should be taken when data is class imbalanced, i.e. when some classes are overly represented when compared to others. Building on top of [1], this work makes use of a recently published ordinal classifier, which transforms the problem into sets of pairwise ranking problems, in order to address the class imbalance in the data [2]. Several methods were experimented with, using a Support Vector Machine as the underlying estimator. The pairwise ranking approach has shown promising results, both for traditional and imbalance metrics.

I. INTRODUCTION

Alzheimer’s disease (AD) is the most frequent type of dementia in the elderly. It is characterized by memory loss along with the decline of other cognitive functions. Typically, the onset of AD is preceded by an intermediate stage known as mild cognitive impairment (MCI) where individuals experience cognitive impairments beyond what is expected, but that do not interfere significantly with their daily activities [3]. MCI patients are likely to convert to AD, particularly the amnestic MCI subtype, but many MCI patients remain stable or convert to a different type of dementia.

Since there is currently no cure for AD, there is a great interest in the development of treatments that can delay or even prevent the progression of the disease. For these treatments to be effective, it is key that AD is accurately diagnosed at the early MCI stages and that the distinction between stable MCI (s-MCI) and progressive MCI (p-MCI) is done reliably.

Brain imaging is a largely accepted tool for this purpose and several imaging biomarkers have been used including two main modalities: structural Magnetic Resonance Images (MRI) and Positron Emitting Tomography (PET). Together with machine learning techniques, these imaging biomarkers have been shown to provide accurate diagnosis.

978-1-5386-6859-7/18/$31.00 ©2018 IEEE
is skewed, which occurs frequently in AD data, that biases traditional classification models towards the larger classes. Pairwise ranking has been previously proposed for binary classification [14] so that learning compares pairs of observations from each class, and therefore both classes contribute equally to the decision boundary. This work has since been extended for ordinal classification [2], [15].

II. DATA

We used FDG-PET and 1.5T MR image volumes from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.ucla.edu/), a global research effort committed to the search for biomarkers that can be used in Alzheimer’s disease clinical treatment trials. Data from participants with FDG-PET scans available at baseline were retrieved: 75 from Normal Controls (NC), 135 from MCI subjects and 58 from AD subjects. In order to test our methods on 4-class data we further divided the 135 MCI subjects into non-converters or stable (s-MCI) and converters or progressive (p-MCI). Our criterion to identify MCI converters was based on the score of two neuropsychological scores, the Mini Mental Scale Exam (MMSE) and the Clinical Dementia Rating (CDR). MCI patients with a CDR increase of 0.5 to 1 or an MMSE reduction to values smaller than 24 in a follow-up period of 24 months were labeled as p-MCI converters. The remaining MCI subjects were labeled as s-MCI. We thus obtained 51 p-MCI converters and 84 s-MCI non-converters.

Some clinical and Neuropsychological characteristics of each group are summarized in Table I. T-tests were performed for all the clinical group pairs confirming that age of the different groups does not vary significantly (p-value > 0.05).

All images were warped into the MNI152 standard space. The warping procedure used PET and MRI images, both of which had already been preprocessed by ADNI researchers. Firstly, skull-stripping and segmentation into white-matter (WM) and gray-matter (GM) was performed in all MR images, with FreeSurfer and SPM8, respectively. Secondly, each PET image was co-registered with the corresponding skull-stripped MR image with SPM8, using Rigid-body transformations and an objective function based on the mutual information between the two images. Thirdly, all MR images were non-linearly registered to an inter-subject template using the DARTEL toolbox from SPM8 [16]. These templates were then mapped to the MNI-ICBM 152 nonlinear symmetric atlas (version 2009a) [17] using an affine transformation. Finally, the original PET images were resampled into the MNI-152 standard space with a 3 x 3 x 3 mm resolution using the appropriate composition of transformations from the above steps.

PCA was applied to choose those components that explained 90% of the variance. The number of features was originally 557,780 and was then reduced to 147.

Furthermore, the data were prepared for two experiments: #1) all four classes were used, and #2) the intermediate classes, representing the MCI stage, are used as a single class. The classes and frequencies are represented in Table II.

<table>
<thead>
<tr>
<th>Group</th>
<th>Class 1 (NC)</th>
<th>Class 2 (s-MCI)</th>
<th>Class 3 (p-MCI)</th>
<th>Class 4 (AD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of subjects</td>
<td>75</td>
<td>84</td>
<td>51</td>
<td>58</td>
</tr>
<tr>
<td>Age</td>
<td>76.0</td>
<td>75.5</td>
<td>74.7</td>
<td>76.1</td>
</tr>
<tr>
<td>Sex (M/F)</td>
<td>49/26</td>
<td>58/26</td>
<td>30/21</td>
<td>34/24</td>
</tr>
<tr>
<td>MMSE</td>
<td>29.1</td>
<td>27.0</td>
<td>27.4</td>
<td>23.5</td>
</tr>
<tr>
<td>CDR</td>
<td>0.0</td>
<td>0.5</td>
<td>0.5</td>
<td>0.8</td>
</tr>
</tbody>
</table>

III. MODELS

The underlying model used throughout this work was a Support Vector Machine (SVM) with a linear kernel. Since SVM is a binary classifier, in order to use it on K classes, the problem must first be converted into one or more binary problem(s). The five approaches used were:

1) **One vs Rest**: the number of models required, M, is the same as the number of classes, K = M, such that the labels k given to each model m are +1 for observation i when k_i = m, or -1 otherwise, and prediction is performed by finding the closest SVM decision hyperplane.

2) **One vs One** where M = K(K – 1) models are trained in pairs, using only a subset of the data corresponding to the two classes involved, and prediction is done by majority voting.

3) **Fan et al.** [1], as previously discussed, is an ordinal ensemble with M = K–1, originally proposed by Frank and Hall [18]. This ensemble takes the class lexicographical order into account by training the i-th model using classes {C_1, ..., C_i} against those to its right, {C_{i+1}, ..., C_K}. Prediction is then performed by majority voting.

4) **oSVM** uses a single SVM by first applying a pre-processing step which expands the feature space. This transformation is applied in such a way that the coefficients from the extra features can then be transformed into intercepts to create new decision hyperplanes [19].

5) **The balanced ranking** adaptation described in the next section.

IV. PROPOSAL

Pairwise scoring ranking is a family of models from the learning-to-rank literature. In pairwise ranking, observations are trained in pairs; in scoring ranking, predictions are produced individually in the form of a score, making it possible to convert them back into classes for classification.

The proposed method converts an SVM classifier into a pairwise scoring ranker, where data is balanced, and then back to a classifier using a threshold. It consists of a single SVM trained on a transformation of the original data domain (x_i, k_i) into
a domain of binary differences, \( \{(x_{ij}, +1), (-x_{ji}, -1) \mid x_{ij} \equiv x_i - x_j \} \) whenever \( x_i > x_j \), meaning \( x_i \) should be ranked higher than \( x_j \). The distance of \( x_i \) to the hyperplane is taken to be the ranking score \( s_i \). Note that since the binary differences \( x_{ij} \) are between observations from different classes, there is no imbalance during training.

This model was proposed for binary classification in [14] when class imbalance is present. Defining \( x_i > x_j \) whenever \( x_i \) belongs to the positive class and \( x_j \) to the negative class, then a new balanced pairwise representation of the data is composed, as illustrated in Fig. 1.

An SVM can then be trained in this space and the ranking score \( s_i \), for each observation \( i \) in the training data, is used to train a threshold to convert predictions back to classes. An extension to the ordinal case from [2], [15] determines the threshold by finding the breaking points where \( k \) should be incremented for each score \( s_i \), such that the cost function \( f \) is minimized,

\[
    f(s_i, k_i, \hat{k}) = \min \left\{ \varepsilon_{k_i, \hat{k}} + f(s_{i+1}, k_{i+1}, \hat{k}), \ f(s_i, k_i, \hat{k} + 1) \right\}.
\]

The process starts with \( i = 1 \) and finishes when \( i = N \) (number of observations). This expression is minimized by making binary decisions on whether to (a) increment \( i \) with a cost of \( \varepsilon_{k_i, \hat{k}} \) or (b) increment \( \hat{k} \) and set a threshold at score \( s_i \). Iterating all permutations is computationally inexpensive (\( O(KN) \)) if done using dynamic programming to avoid redundant computations. \( \varepsilon = \varepsilon_{k_i, \hat{k}} \) is a cost matrix. Several cost matrices have been proposed: uniform, absolute and inverse class frequency cost matrices [2].

Notice that this method would not work for nominal classes. Classes must be ordinal, so that a ranking “greater than” lexicographical order exists. However, the method does work even if classes are not evenly-spaced, i.e. it does not assume cardinality unlike regression methods.

V. Evaluation Metrics

Traditionally, ordinal classification problems use as metric the Mean Absolute Error, \( \text{MAE} = \frac{1}{N} \sum_i |y_i - \hat{y}_i| \). But when class imbalance is present, we might want to use such metrics as Average Mean Absolute Error, or Maximum Absolute Error, \( \text{MMAE} = \max \{ \text{MAE}_k \mid k = 1, \ldots, K \} \), where \( \text{MAE}_k \) is the MAE metric applied to a single class \( k \). These metrics benefit the most challenging and less abundant classes.

Another commonly used metric is Spearman ranking correlation \( \rho \) which evaluates the correlation between the predicted classes and the true classes. Also, Accuracy (Acc) is used as a measure of the fraction of correct predictions, but it completely ignores the ordinal nature of the data – and punishes equally a prediction of AD of a healthy NC person as it does for an MCI patient.

VI. Results

The aforementioned models were used in the FDG-PET data of 268 patients, and evaluated using the previously mentioned metrics: Spearman \( \rho \), Acc, MAE, AMAE and MMAE. The latter two metrics take imbalance into consideration.

The results for both experiments mentioned in section II are presented in Table III and Table IV for Experiment #1.

<table>
<thead>
<tr>
<th>Model</th>
<th>( \rho )</th>
<th>Acc</th>
<th>MAE</th>
<th>AMAE</th>
<th>MMAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Most Freq</td>
<td>0.000 ± 0.00</td>
<td>0.313 ± 0.03</td>
<td>0.903 ± 0.04</td>
<td>1.000 ± 0.00</td>
<td>2.000 ± 0.00</td>
</tr>
<tr>
<td>One vs One</td>
<td>0.496 ± 0.17</td>
<td>0.407 ± 0.13</td>
<td>0.791 ± 0.19</td>
<td>0.801 ± 0.20</td>
<td>1.203 ± 0.29</td>
</tr>
<tr>
<td>One vs One (w)</td>
<td>0.461 ± 0.21</td>
<td>0.399 ± 0.13</td>
<td>0.813 ± 0.20</td>
<td>0.829 ± 0.22</td>
<td>1.278 ± 0.30</td>
</tr>
<tr>
<td>One vs Rest</td>
<td>0.388 ± 0.32</td>
<td>0.362 ± 0.11</td>
<td>0.925 ± 0.23</td>
<td>0.931 ± 0.25</td>
<td>1.369 ± 0.32</td>
</tr>
<tr>
<td>Fan et al.</td>
<td>0.418 ± 0.27</td>
<td>0.377 ± 0.12</td>
<td>0.896 ± 0.18</td>
<td>0.923 ± 0.20</td>
<td>1.419 ± 0.36</td>
</tr>
<tr>
<td>Fan et al. (w)</td>
<td>0.499 ± 0.15</td>
<td>0.373 ± 0.12</td>
<td>0.806 ± 0.14</td>
<td>0.803 ± 0.14</td>
<td>1.193 ± 0.25</td>
</tr>
<tr>
<td>Fan et al. (w)</td>
<td>0.505 ± 0.22</td>
<td>0.396 ± 0.15</td>
<td>0.784 ± 0.23</td>
<td>0.782 ± 0.23</td>
<td>1.192 ± 0.29</td>
</tr>
<tr>
<td>oSVM</td>
<td>0.401 ± 0.30</td>
<td>0.392 ± 0.13</td>
<td>0.873 ± 0.28</td>
<td>0.891 ± 0.31</td>
<td>1.465 ± 0.56</td>
</tr>
<tr>
<td>Balanced ranking</td>
<td>0.568 ± 0.18</td>
<td>0.451 ± 0.10</td>
<td>0.720 ± 0.23</td>
<td>0.737 ± 0.23</td>
<td>1.183 ± 0.37</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>( \rho )</th>
<th>Acc</th>
<th>MAE</th>
<th>AMAE</th>
<th>MMAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Most Freq</td>
<td>0.000 ± 0.00</td>
<td>0.504 ± 0.01</td>
<td>0.496 ± 0.01</td>
<td>0.667 ± 0.00</td>
<td>1.000 ± 0.00</td>
</tr>
<tr>
<td>One vs One</td>
<td>0.423 ± 0.24</td>
<td>0.485 ± 0.11</td>
<td>0.590 ± 0.16</td>
<td>0.559 ± 0.19</td>
<td>0.832 ± 0.19</td>
</tr>
<tr>
<td>One vs One (w)</td>
<td>0.424 ± 0.23</td>
<td>0.470 ± 0.11</td>
<td>0.601 ± 0.15</td>
<td>0.563 ± 0.18</td>
<td>0.845 ± 0.19</td>
</tr>
<tr>
<td>One vs Rest</td>
<td>0.392 ± 0.21</td>
<td>0.466 ± 0.12</td>
<td>0.623 ± 0.15</td>
<td>0.593 ± 0.18</td>
<td>0.933 ± 0.29</td>
</tr>
<tr>
<td>One vs Rest (w)</td>
<td>0.379 ± 0.32</td>
<td>0.463 ± 0.10</td>
<td>0.619 ± 0.12</td>
<td>0.604 ± 0.13</td>
<td>0.935 ± 0.28</td>
</tr>
<tr>
<td>Fan et al.</td>
<td>0.432 ± 0.32</td>
<td>0.522 ± 0.13</td>
<td>0.522 ± 0.13</td>
<td>0.572 ± 0.14</td>
<td>0.902 ± 0.22</td>
</tr>
<tr>
<td>Fan et al. (w)</td>
<td>0.422 ± 0.17</td>
<td>0.526 ± 0.09</td>
<td>0.519 ± 0.10</td>
<td>0.523 ± 0.12</td>
<td>0.814 ± 0.19</td>
</tr>
<tr>
<td>oSVM</td>
<td>0.285 ± 0.29</td>
<td>0.474 ± 0.18</td>
<td>0.608 ± 0.20</td>
<td>0.646 ± 0.23</td>
<td>1.052 ± 0.39</td>
</tr>
<tr>
<td>Balanced ranking</td>
<td>0.492 ± 0.18</td>
<td>0.507 ± 0.13</td>
<td>0.552 ± 0.15</td>
<td>0.552 ± 0.15</td>
<td>0.876 ± 0.22</td>
</tr>
</tbody>
</table>
the naive multiclass approaches. The SVM models with the suffix (w) that are shown in Tables III and IV are weighted models, in which training used a cost matrix such that each entry $k$ has a weight inverse to the frequency of observations of that class $N_k$, $w_k = \frac{N}{N_k}$. This is another typical approach to maximize class imbalance metrics. The “Most Freq” model is a baseline which always predicts the most frequent class.

The proposed ranking-based method outperforms all other models for Experiment #1. When using less classes, the difference between the models is not as clear. Fan et al. approach [1] is more competitive in this case, when using weights (w). In any case, ordinal classifiers clearly outperform the naive multiclass approaches.

VII. CONCLUSION

Traditionally, automatic classification systems resort to either two-classes, ignoring some of the stages, or do not take the ordinal information into consideration. A pairwise ranking model [2] is here introduced to both (a) take the ordinal aspect in consideration, which had already been considered in [1], and (b) account for the class imbalance present in the data.

The ordinal classifiers clearly outperformed traditional multiclass classifiers which do not take ordinality into account. The pairwise ranking model that was tested was demonstrably competitive, especially in the experiment taking the full range of four classes into the prediction.

As future work, we intend to take the various modalities (PET and MRI images) into account in a single model to strengthen the predictions. Other pairwise scoring ranking methods could be used, together with neural networks, AdaBoost or other models; the balancing idea suggested is not restricted to the space of differences technique exploited in this paper.

ACKNOWLEDGMENT

This work was funded by the Project “NanoSTIMA: Macro-to-Nano Human Sensing: Towards Integrated Multimodal Health Monitoring and Analytics/NORTE-01-0145-FEDER-000016” financed by the North Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, and through the European Regional Development Fund (ERDF), and also by Fundação para a Ciência e a Tecnologia (FCT) within PhD grant number SFRH/BD/122248/2016 and FCT project UID/EEA/50009/2013. Data used in the preparation of this article were obtained from the ADNI database (http://www.loni.ucla.edu/ADNI). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report.

REFERENCES


