Data Fusion and Feature Selection for Alzheimer's Diagnosis

Blake Lemoine, Sara Rayburn, Ryan Benton, Alzheimer's Disease Neuroimaging Initiative[1]

Abstract. The exact cause of Alzheimer's disease is unknown; thus, ascertaining what information is vital for the purpose of diagnosis, whether human or automated, is difficult. When conducting a diagnosis, one approach is to collect as much potentially relevant information as possible in the hopes of capturing the important information; this is the Alzheimer's Disease Neuroimaging Initiative (ADNI) adopted approach. ADNI collects different clinical, image-based and genetic information related to Alzheimer's disease. This study proposes a methodology for using ADNI's data. First, a series of support vector machines is constructed upon nine data sets. Five are the results of clinical tests and the other four are features derived from positron emission tomography (PET) imagery. Next, the SVMs are fused together to determine the final clinical dementia rating of a patient: normal or abnormal. In addition, the utility of applying feature selection methods to the generated PET feature data is demonstrated.

NOTE:
This paper was published by Springer-Verlag in the proceedings of the International Conference on Brain Informatics, which took place in Toronto, ON, Canada from August 28, 2010 to August 30, 2010. The proceedings is Volume 6334 in the Lecture Notes in Computer Science series. The current link to the article is http://www.springerlink.com/content/ux22vjv745600573. In the event it changes, the original publication can be found through by searching the Springer-Verlag site at http://www.springerlink.com

1 Introduction

The Alzheimer's Disease Neuroimaging Initiative (ADNI) has made available a large quantity of data pertaining to Alzheimer's Disease (AD) [1]. This paper addresses the potential of this data to be used to train support vector machines (SVMs) for the purpose of automatic AD diagnosis. Support vector machines work by mapping data

[1] Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (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. ADNI investigators include (complete listing available athttp://www.loni.ucla.edu/ADNI/Collaboration/ADNI_Manuscript_Citations.pdf).

that may not be linearly separable into a high-dimensional space where it is separable. This mechanism explicitly requires that data be represented as a feature vector.

With the data sets utilized in this study, there are two distinct challenges that must be overcome. The first difficulty is the data sets containing an enormous number of features. The primary examples of this scenario are data sets containing several thousand features and only a few hundred patients. SVMs trained on these data sets are highly likely to over fit the data. Fortunately, it is likely that some features are more relevant to the diagnosis of AD than others. In this case, a classifier should use only those features that are relevant and disregard the remaining features.

With respect to the second difficulty, there are several dozen different tests, activities, and evaluations for any given patient. Each of these data sets is very different in nature. A medical doctor uses all of these variant sources of data to form a holistic picture of a patient's condition and renders an opinion based on that picture. Ideally, an automatic classifier would also take multiple sources of information about a patient into consideration simultaneously.

Considering these two challenges, it is clear that the problem of using the ADNI data for automatic classification of AD is one of information management. The task is to use as many relevant data sources as possible while ignoring as much irrelevant information within those sources as possible. Feature selection using the genetic algorithm and data fusion are the methods proposed here for addressing these challenges. This paper first presents the background of the problem and gives details of the data sets used. Next the method for performing feature selection using the genetic algorithm is presented, followed by the method for performing data fusion. Finally, the empirical results of classification experiments using this methodology are presented.

2 Background of Problem

In [2], the investigators assumed that each patient was represented by one or two features, each which described the metabolic activity within a region, along with a clinical dementia rating (CDR) value as described below. The CDR values were mapped into one of two categories, *normal or abnormal*. A threshold was assigned a priori and, then, each feature was compared to the threshold. If the feature's value was greater than the threshold, the feature was considered to be evidence for dementia. Otherwise, it was evidence against. If all the features were supporting dementia, then the patient was classified as abnormal. Otherwise, the patient was classified as being normal.

In [3], researchers report on efforts to build a classifier to distinguish between subjects with AD and Frontotemporal dementia (FTD) using features extracted from PET images. Calculated z-scores associated with locations within the cortical region were partitioned into groups that best distinguished AD and FTD. For each resulting region, the z-scores were used to generate a representative "z-score" value for the region. The representative values were then used to build a decision tree to distinguish between subjects with AD and FTD. A total of 48 subjects were used, 34 diagnosed with AD and 14 diagnosed with FTD. An accuracy of 94% was reported.

Results reported on in [4] applied voxel-based morphometry to a set of magnetic resonance imaging (MRI) scans in order to extract 20 features from each scan. The extracted features measured the amount of grey-matter found within a region. An artificial neural network was constructed to distinguish subjects diagnosed with AD and subjects diagnosed as normal condition (NC). The evaluated data consisted of 10 subjects diagnosed as AD and 12 subjects diagnosed as NC. The reported average accuracy was 100%.

Finally, results related to the classification of subjects with respect to mild cognitive impairment (MCI) have been reported. Specifically, in [5] researchers applied a nonlinear multivariate analysis technique to MRI images which resulted in classification accuracies of 81% and 74% with respect to MCI vs. NC and AD vs. MCI, respectively. The study included 66 subjects diagnosed as normal, 88 subjects diagnosed as MCI, and 56 diagnosed as AD. In an alternative study, which was based on extracted z-scores from PET images, it was determined that subjects diagnosed as normal could be distinguished from subjects diagnosed as MCI with an accuracy of 92% [6]. The latter study included 110 subjects diagnosed as NC and 114 subjects diagnosed as MCI.

3 Data

In this section, we will briefly describe the source of the data used in this study, explain what features were extracted from the PET imagery, provide an overview of the clinical data, and, finally, address some data cleaning issues.

3.1 Data Source

Data used in the preparation of this paper were obtained from the ADNI database (www.loni.ucla.edu/ADNI). The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether serial MRI, PET, and other biological markers could be used to detect the progression of MCI and early AD. Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials.

The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California - San Francisco. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 adults, ages 55 to 90, to participate in the research - approximately 200 cognitively normal older individuals to be followed for 3 years, 400 people with MCI to be followed for 3 years and 200 people with early AD to be followed for 2 years. For up-to-date information, see www.adni-info.org.

3.2 PET Data

As noted, we acquired the PET imagery from ADNI. However, the learning algorithms utilized assumes that the data is represented as feature vectors. Hence, we needed to convert each PET image into a usable representation.

First, we converted the PET image into a set of 15964 data points (voxels); each data point is a Z-Score representing the metabolic activity at that point. The data points were computed using a GE proprietary application known as Cortex ID. The regions are based on the Talairach-Tourneau atlas.

Next, as noted in [2], it is common practice to normalize each PET image set to a reference region. In this study, two different normalizations were considered. One reference region used was the Pons (PNS); in this case, each pixel value within the PET scan was divided by the average activity in the Pons. The other normalization considered was the average Global (GLB) activity of the brain. The resulting two versions of the 15,964 points are referred as the Full Feature (FF) Data Set.

In addition 31 regional Z-Score averages were also computed. The values for the left side and right side of the brain were calculated for the following regions: Parietal, Temporal, Frontal, Occipital, Post Cingulate, Ant Cingulate, Medial Frontal, Medial Parietal, Sensorimotor, Visual, Caudate, Cerebellum, Vermis, and the combined Parietal, Temporal and Frontal. In addition, a single value was computed for each of the following: Pons, Global (total brain) and Cortex. The averages are treated as an additional dataset, which, in this paper, is referred as PET 31. Thus, four data sets had been created: PET FF GLB, PET 31 GLB, PET FF PNS, and PET 31 PNS.

3.3 Clinical Data

There are several dozen clinical data sets provided by ADNI. Five were chosen as representative data sets spanning lab tests, questionnaires, and doctor interviews. These were: the ADAS-Cognitive behavior test (ADAS), the functional assessment questionnaire (FAQ), the family history questionnaire (FHQ), a homocysteine test (HCRES), and the baseline symptoms checklist (BLSCHECK). The result of each of these is a small number of "yes/no" values. These were attractive because they are easily representable as feature vectors and because feature selection would not be required for these data sets, due to the small number of features in each set.

3.4 Pre-processing the Data

For the described data sets to be useful for classification, a certain amount of pre-processing was performed. The ADNI data has a number of mismatches between CDR records, which provide the correct classification of a patient, and the PET imaging data and the clinical evaluation data. In order to generate usable data, one or the other of these must be chosen as constant. CDR records serve that purpose here. Each CDR was paired with exactly one PET scan, the one nearest to it in time. It was also paired with exactly one of each clinical measure to form a single problem instance. Any problem instance which did not include one or more of the diagnostic

measures was removed from the data set. Furthermore, many of the PET scans have measurements which are inconsistent with other PET scans. The feature values are Z-scores in comparison to a normal population. This sets an upper and lower bound for plausibility. A Z-score of +100 is simply so unlikely that it is more likely to be bad data. Any problem instances containing PET scans with values outside of a predefined acceptable range were removed from the data set. Finally, the Z-scores of PET scans were normalized to values between 0 and 1.

4 Feature Selection

Finding an optimal set of features to use for classification purposes requires exponential time in the worst case. Evolutionary algorithms have been found to efficiently solve such problems in the past [7]. Furthermore, there is some historic basis for the genetic algorithm being used for feature selection [8] and, as will be shown, it is useful in obtaining good results for the PET FF GLB and PET FF PNS sets.

The problem of feature selection in classification systems is, informally, a matter of deciding, with respect to the classification task, what information is relevant and what information is not. In order to translate this into an algorithm, however, the problem must be formalized. The task will be to find a set of features of size ***n***, where ***n*** is some fixed constant, that maximize some performance function when used to perform a classification task. Alternate definitions for the feature selection problem exist [9] but the one given above is the one adopted in this paper. For the procedure advocated here, ***n*** is chosen to be 30 and the evaluation function is chosen to be the area under the classifier's receiver operating characteristic (ROC) curve (AUC) [10].

1) Generate a random gene pool
2) Create an SVM for each chromosome; the SVM created will only use the features referenced by that chromosome.
3) Rank the SVMs by the area under ROC curve
4) Create a set of chromosomes that correspond to the top 10% of the SVMs as well as the chromosome that corresponds to the worst
5) Use the resultant set of chromosomes as the parents of a new gene pool
   a) Randomly select two chromosomes from the resultant set as parents.
   b) Perform crossover; this results in two new chromosomes.
   c) For each gene in each new chromosome, randomly change the value m% of the time.
   d) Insert the two new chromosomes into the new gene poll.
6) If the best SVM is better than the evaluation function's stopping metric then end, otherwise go to step 2

Figure 1. Outline of the feature selection algorithm

In adapting the genetic algorithm for this purpose, each feature will correspond to a single gene. A chromosome will be a set of ***n*** genes. The evaluation function will be

the AUC of the chromosome's SVM after 5-fold stratified cross validation. Crossover will consist of using a portion of one parent's features and a complementary portion of the other parent's features. Mutation will be incorporated by allowing features to be randomly changed to other features. The rate of mutation will be varied such that mutations will be more common in early generations than in later generations. An outline of this algorithm is given in Figure 1. Once the gene pool contains a member that meets the given stopping criterion then the best member of that generation will be returned as the best feature set. Hence, by the end of this processing, we will have two new data sets: PET FS GLB and PET FS PNS

5 Data Integration

The purpose of data integration is to combine the five clinical sets (ADAS, FAQ, FHQ, HCRES, and BLSCHECK) with four PET data sets (31 GLB, FS GLB, 31 PNS and FS PNS). Each of the sets are individually used to train and test SVMs. In order to create RBF SVMs, a Gaussian kernel was computed for each of the above data sets. Each training set consists of one feature set associated with a patient diagnosis (CDR value). These training sets are each used to perform 10-fold stratified cross validation of a soft margin SVM. The results for each data set are averaged across all 10 folds, yielding a single summary statistic, the AUC. This AUC is then used to apportion weights in the data integration step.

The combined kernel is created by a weighted linear summation of the individual kernel matrices. The weights are based on the individual ROC scores of the included data sources. A simple linear combination of kernels itself creates a valid kernel matrix because it maintains the positive, semi-definite nature of the kernel [11]. Let $\boldsymbol{K}$ be the set of n individual kernel matrices and $\boldsymbol{W}$ be the set of normalized relative performance of each kernel at the classification problem.

$$K_{summed} = \sum_{i=1}^{n} w_i k_i \tag{1}$$

Formula (1) is the linear combination of the individual kernel matrices. This combined kernel is then used as input to perform 10-fold stratified cross validation with an SVM classifier on the same problem.

6 Results

After data preprocessing there are 495 clinical dementia ratings (CDRs) in the ADNI data which have related records in each of the data sets considered. A CDR can have a value of 0, 0.5, or 1+, where 0 is no dementia, 0.5 is questionable and 1+ is dementia. These possible CDR values can be used to create a binary classification problem where the task is to distinguish normal from abnormal patients, where a CDR of 0 is normal, and a CDR of 0.5 or 1+ is considered abnormal [2]. The set of training samples contains 362 abnormal CDRs and 133 normal. For the purposes of evaluation, an abnormal CDR is taken as the positive class.

As stated earlier, all feature selection experiments thus far run with the ADNI Alzheimer's data generate chromosomes that contain 30 genes. This number was chosen somewhat arbitrarily, but fits the requirement that the number of features be much smaller than the number of examples. The stopping criteria is when an SVM is derived from the 30 features that has a minimum AUC of 0.98.

To determine how the combined data set performed it was compared to its constituent parts. Each individual data set was evaluated on the same training set and classification task, where the evaluation metric is the area under the ROC curve. After feature selection was performed on the full feature set for the PET data, this feature set was evaluated for the subset of CDRs chosen for the data integration experiments. The AUC for both the global and pons PET scans were 0.94, gaining 3% AUC over the PET feature set of 31 regional averages. The other non-imaging clinical data had AUCs ranging from 0.58 to 0.93, as shown in Table 1.

Table 1. AUC and weights for data sets evaluated

	AUC	Normalized Relative Performance (weight)
ADAS	0.93	0.16
BLSCHECK	0.62	0.11
FAQ	0.92	0.16
FHQ	0.72	0.13
HCRES	0.58	0.1
PET31 GLB	0.92	0.16
PET31 PNS	0.92	0.16
PET FS GLB	0.92	0.16
PET FS PNS	0.94	0.17
COMBINED	0.97	n/a

By combining the data sets with weights as shown in Table 1, we were able to improve the AUC of our classifier another 3% from 0.94 with the feature selected PET data to 0.97 with all of the data included.

7 Conclusions and Future Work

The results of this experiment demonstrate that both feature selection by the genetic algorithm and data integration can successfully increase the accuracy of SVM classification of AD using the available data from ADNI. Future work in this area will incorporate other clinical data to potentially increase accuracy even further. Additionally, the feature selection process could be used to identify which area(s) of the brain contain the most relevant information for the purposes of PET scan based Alzheimer's diagnosis. Determining whether or not any such areas exist is another area of future research.

Acknowledgements

First, we would like to thank Dr. Suresh Choubey at GE Healthcare for providing the 31 and 15964 PET feature data sets.

Second, data collection and sharing for this project was funded by the ADNI (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Abbott, AstraZeneca AB, Bayer Schering Pharma AG, Bristol-Myers Squibb, Eisai Global Clinical Development, Elan Corporation, Genentech, GE Healthcare, GlaxoSmithKline, Innogenetics, Johnson and Johnson, Eli Lilly and Co., Medpace, Inc., Merck and Co., Inc., Novartis AG, Pfizer Inc, F. Hoffman-La Roche, Schering-Plough, Synarc, Inc., as well as non-profit partners the Alzheimer's Association and Alzheimer's Drug Discovery Foundation, with participation from the U.S. Food and Drug Administration. Private sector contributions to ADNI are facilitated by the Foundation for the National Institutes of Health (www.fnih.org).

The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of California, Los Angeles. This research was also supported by NIH grants P30 AG010129, K01 AG030514, and the Dana Foundation.

References

1. *ADNI-Info*. (2009). Retrieved March 1, 2010, from Alzheimer's Disease Neuroimaging Initiative: http://www.adni-info.org/
2. Minoshima, S., Frey, K., Koeppe, R., Foster, N., & Kuhl, D. (1995). A Diagnostic Approach in Alzheimer's Disease Using Three-Dimensional Stereotactic Surface Projections of Fluorine-18-FDG PET. *Journal of Nuclear Medicine , 36* (7), 1238-1248.
3. Sadeghi, N., Foster, N., Wang, A., Minoshima, S., Lieberman, A., & Tasdizen, T. (2008). Automatic Classification of Alzheimer's Disease vs. Frontotemporal Dementia: A Spatial Decision Tree Approach with FDG-PET. *Proceedings of IEEE International Symposium on Biomedical Imaging: From Nano to Micro*, 408-411.
4. Huang, C., Yan, B., Jiang, H., & Wang, D. (2008). Combining voxel-based morphometry with Artificial Neural Network theory in the application research of diagnosing Alzheimer's disease. *International Conference on BioMedical Engineering and Informatics*, 250-254.
5. Fan, Y., Batmanghelich, N., Clark, C., & Davatzikos, C. (2008). Spatial patterns of brain atrophy in MCI patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline. *NeuroImage , 39* (4), 1731-1743.
6. Mosconi, L., Tsui, W., Herholz, K., Pupi, A., Drzezga, A., Lucignani, G., Reiman, E., Holthoff, V., Kalbe, E., Sorbi, S., Diehl-Schmid, J., Perneczky, R., Clerici, F., Caselli, R., Beuthien-Baumann, B., Kurz, A., Minoshima, S., and de Leon, M. (2008). Multicenter Standardized ^{18}F-FDG PET Diagnosis of Mild Cognitive Impairment, Alzheimer's Disease, and Other Dementias. *The Journal of Nuclear Medicine , 49* (3), 390-398.

7. DeJong, K., & Spears, W. (1989). Using Genetic Algorithms to Solve NP-Complete Problems. *Proceedings of the 3rd International Conference on Genetic Algorithms,* 124-132.
8. Fröhlich, H., Chapelle, O., & Schölkopf, B. (2003). Feature selection for support vector machines by means of genetic algorithms. *15th IEEE International Conference on Tools with AI (ICTAI 2003)*, 142-148.
9. Weston, J., Mukherjee, S., Chapelle, O., Pontil, M., Poggio, T., & Vapnik, V. (2000). Feature Selection for SVMs. *Advances in Neural Information Processing Systems*, 13.
10. Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27, 861-874.
11. Lanckriet, G., DeBie, T., Christianini, N., Jordan, M., & Noble, W. (2004). A statistical framework for genomic data fusion. *Bioinformatics*, 20.