Multi-Source Feature Selection to Improve Multi-class Brain Tumor Typing

V. Metsis[1], D. Mintzopoulos[2,3], H. Huang[1], M. N. Mindrinos[4], P. M. Black[5], F. Makedon[1], and A. A. Tzika[2,3]

[1]Computer Science, University of Texas, Arlington, TX, United States, [2]NMR Surgical Laboratory, MGH & Shriners Hospitals, Harvard Medical School, Boston, MA, United States, [3]Radiology, Athinoula A. Martinos Center for Biomedical Imaging, Boston, MA, United States, [4]Biochemistry, Stanford University School of Medicine, Palo Alto, CA, United States, [5]Neurosurgery, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, United States

Introduction

Brain tumors are one of the leading causes of death in adults [1]. The potential value of combining high resolution magic angle spinning (HRMAS) proton (^{1}H) Magnetic Resonance Spectroscopy (MRS) and gene expression data for brain tumor typing has been previously proposed [2]. Also the molecular classification of brain tumor biopsies using ^{1}H HRMAS MRS and robust classifiers has been recently reported [3]. However, this classification was limited to the binary classification problem of discriminating between tumor types using the one-versus-all classification methodology. Here, we use machine learning algorithms to create a novel framework to perform the heterogeneous data fusion on MRS and gene expression data coming from the same brain tumor biopsies, to identify different profiles of brain tumors. We concentrate on the data fusion for the problem of assigning each sample to one of the multiple possible tumor type classes. Therefore, we select features (biomarkers) from multi-source simultaneously and those selected features are discriminative to all brain tumor types used in our study, not just to individual ones.

Methods

We used 46 samples of normal (control) and brain tumor biopsies from which we obtained ex vivo HRMAS 1H MRS and gene expression data respectively. The samples came from tissue biopsies taken from 16 different people. Out of the forty-six biopsies that were analyzed, 9 of them were control biopsies from epileptic surgeries and the rest 37 were brain tumor biopsies. The tumor biopsies belonged to 5 different categories: 11 glioblastoma multiforme (GBM); 8 anaplastic astrocytoma (AA); 7 meningioma; 7 schwanoma; and 5 from adenocarcinoma. From the MRS data we extracted and used as features 15 significant metabolites: choline (Cho), phosphocholine (PC), glycerophosphocholine (GPC), phosphoethanolamine (PE), ethanolamine (Etn), γ-aminobutyric acid (GABA), n-acetyl aspartate (NAA), aspartate (Asp), alanine (Ala), polyunsaturated fatty acids (PUFA), glutamine (Gnl), glutamate (Glu), lactate (Lac), taurine (Tau) and lipids (Lip). For the gene expression profiles the original feature space comprised 54,675 genes. We experimented with feature selection from both dataset types in order to reduce redundancy and noise before using them for classification. Because the main goal of this work was to examine the potential gain from combined heterogeneous data for tumor typing we only performed experiments with the most well studied feature selection and classification methods. The feature selection methods we applied include the filter methods Relief-F (RF) [4], Information Gain (IG) [5] and χ^2-statistic [6], and a Wrapper feature selection method for each classification algorithm. As for the classification methods, we used Naïve Bayes [7] and Support Vector Machines (SVMs) [8]. The methodology of our classification framework is summarized in figure 1 and is comprised of 3 main steps: feature selection from each dataset separately, merging of the top features from both datasets, and classification based on the combined feature set.

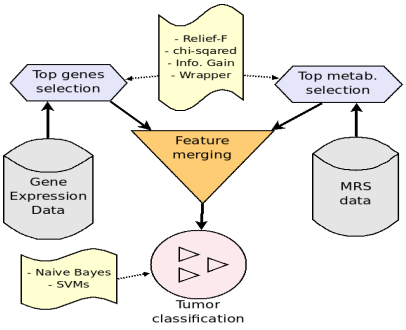

Figure 1. Data fusion and classification framework.

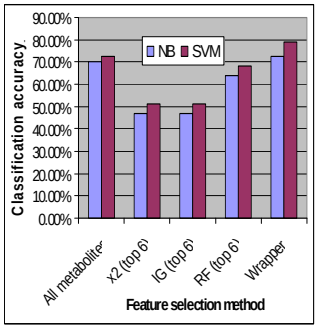

Figure 2: Classification accuracy for the using MRS data only.

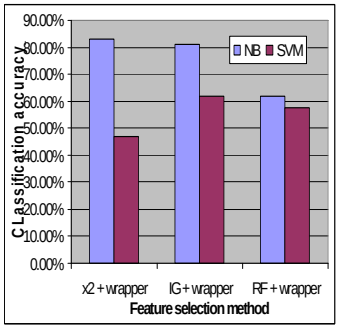

Figure 3: Classification accuracy using gene expression data only.

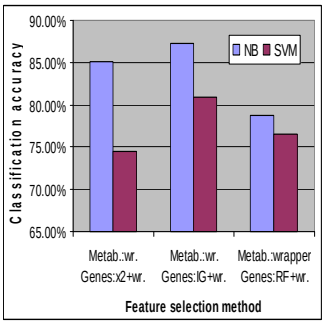

Figure 4: Classification accuracy using a combination of features from multi-source.

Results

We performed experiments to evaluate the classification accuracy when using each of the datasets separately and in combination. For the *MRS data* we tested our classifiers for the case of using all available metabolites and for the cases of applying each of the 4 feature selection methods. The best accuracy of 78.72% was obtained by the SVM classifier by using the wrapper feature selection method. For the *gene expression data* we followed a hybrid feature selection approach selecting the top 100 genes by using a filter feature selection method and then further reducing the feature number by using wrapper feature selection. The best accuracy we could get for this dataset came from the Naïve Bayes classifier and it reached 82.98%. Finally, we experimented with the *combination* of the best features drawn from each of the above dataset. In this case both classifiers outperformed the respective best accuracies for the individual datasets. The best result of 87.23% was given by the NB classifier when using wrapper feature selection for the MRS dataset and a combination of IG and wrapper feature selection for the gene expression dataset.

Discussion

The results of the proposed machine learning based, data fusion framework, to type different brain tumors and detect biomarkers from both MRS and genomic gene expression data, show that the combination of heterogeneous data sources can significantly improve the classification performance when considering the multi-class classification problem. The wrapper and feature filtering methods are used to remove redundant features and select the most significant features to classify five different brain tumors. In our experiments, we successfully applied the proposed methodology on real life MRS and gene expression data extracted from the same biopsy samples. Since our framework is a general method, it can also be applied to any other biomedical and biological data fusion for sample classification and biomarker detection.

References

1. Legler JM, Gloeckler Ries LA, Smith MA, et al., J Natl Cancer Inst 92: 77A-78, 2000.
2. Tzika AA, Astrakas L, Cao HH, et al., International Journal of Molecular Medicine 20: 199-208, 2007.
3. Andronesi OC, Blekas KD, Mintzopoulos D, et al., *International Journal of Oncology* 33: 1017-1025, 2008
4. H. Liu and R. Setiono, *IEEE Computer Society Washington*, DC, USA, 1995.
5. T.M. Mitchell. *Machine Learning. 1997*. Burr Ridge, IL: McGraw Hill.
6. I. Kononenko, *Lecture notes in Computer Science*, pages 171–171, 1994.
7. V. Metsis et al., *Third Conference on Email and Anti-Spam (CEAS)*, 2006
8. V. Vapnik. *Statistical Learning Theory*. 1998. NY Wiley.