Feature Selection Study on Separate Multi-modal Datasets: Application on Cutaneous Melanoma

Konstantinos Moutselos[1], Aristotelis Chatziioannou[2,*], and Ilias Maglogiannis[1]

[1] University of Central Greece
{kmouts,imaglo}@ucg.gr
[2] Institute of Biological Research and Biotechnology,
National Hellenic Research Foundation
achatzi@eie.gr

Abstract. In this work, we study the behavior of a feature selection algorithm (backwards selection) using random forests, by fusing multi-modal data from different subjects. Two separate datasets related to cutaneous melanoma, obtained from image (dermoscopy) and non-image (microarray) sources are used. Imputations are applied in order to acquire a unified dataset, prior the effect of machine learning algorithms. The results suggest that application of the normal random imputation method acts as an additional variation factor, helping towards stability of potential recommended biomarkers. In addition, microarray-derived features were favorably selected as best predictors compared to image-derived features.

Keywords: feature selection, random forest, biomarkers, cutaneous melanoma.

1 Introduction

Integration of multi-modal and multiscale data is of known importance in the context of personalized medicine and the electronic health records. Much effort is exerted for the assessment of the appropriate data fusion schemes which could utilize the most of the available information contained in these datasets. In the context of Virtual Physiological Human (VPH) vision, an integrated framework should make it possible to interconnect predictive models defined at different scales, with different methods, and at different levels of detail into functional networks that consolidate and test systemic hypotheses [1].

Information fusing algorithms can be categorized as being either combination of data (COD) or combination of interpretations (COI) [2]. COD methods aggregate features from each source into a single feature vector before classification, while COI methods classify the data from each source independently and then aggregate the results. Rohlfing et al compared the two methods to combine information sources in different biomedical image analysis applications, while Haapanen and Tuominen [3] followed a COD approach for the combination of satellite image and aerial photograph features for forest variable estimation. On the other hand, Jesneck et al

* Corresponding author.

L. Iliadis et al. (Eds.): AIAI 2012 Workshops, IFIP AICT 382, pp. 36–45, 2012.

[4], on a COI path, optimized a decision-fusion technique to combine heterogeneous breast cancer data. Lee et al [5], proposed a Generalized Fusion Framework (GFF) for homogenous data representation and subsequent fusion in the meta-space using dimensionality reduction techniques. The meta-space is created by projecting the heterogeneous data streams into a space where these scale and dimensionality differences are alleviated. Such meta-space representation approaches, which transform data into a homogeneous space allowing for direct combination of modalities, are embedding projections and kernel space projections [6].

GFF algorithms assume that we have raw data from sources $S_i(x_1, x_2, .., x_k)$, where $x_1, x_2, .., x_k$ represent the k observations in a study and i represents one of the N data sources, $i \in \{1,2, \dots, N\}$. While this could be the case for specific studies or electronic patient records, most available databases contain single modal data from different patients. The number of observations from each source S_i differs. Nonetheless, the modalities from each source reflect information regarding the same disease, and it is an open question how these interconnections could be exploited.

In this work, we study the behavior of a feature selection algorithm (random forests) as obtained by fusing multi-modal data from different subjects. We refer to these data as *separate* datasets. The integration of separate datasets referring though to the same disease, is an innovative approach which can contribute significantly towards the extraction of better biomarkers involved in various diseases.

1.1 Cutaneous Melanoma

Cutaneous melanoma (CM) is considered a complex multigenic and multifactorial disease that involves both environmental and genetic factors. It is the most life-threatening neoplasm of the skin, and its incidence and mortality have been increasing worldwide. CM tumorigenesis is often explained as a progressive transformation of normal melanocytes to nevi that subsequently develop into primary cutaneous melanomas (PCM). However, the molecular pathways involved have not been clearly elucidated, although considerable progress has been made [7]. Despite the success of genomics in defining genomic markers or gene signatures for other kinds of cancers (such as breast cancer), there has been no similar progress related to malignant melanoma.

The microarray studies that have been performed on CM by different groups, used different microarray platforms in highly heterogeneous patient cohorts and pathological sample collections [8]. These differences make comparisons quite difficult and result in a reduced total cohort size and diversity, since independent cohorts from different studies are hard to be summed [9].

Regarding the clinical diagnostic methods for diagnosis of melanoma, there are several standard approaches for analysis and diagnosis of lesions. For example the Menzies scale, the Seven-point scale, the Total Dermoscopy Score based on the ABCD rule, and the ABCDE rule (Asymmetry, Border, Color, Diameter, Evolution). In these methods, digital images can serve as a basis for the medical analysis and diagnosis of lesions under consideration. As there is a general lack of precision in human interpretation of image content, advanced computerized techniques can assist

doctors in the diagnostic process [10]. A review of image acquisition and feature extraction methods utilized in the literature regarding existing classification systems can be found in [11].

1.2 Feature Selection

Feature selection techniques do not alter the original representation of the variables, but merely select a subset of them, in contrast to other dimensionality reduction techniques like those based on projection (e.g. principal components analysis) or compression (e.g. using information theory). Thus, they preserve the original semantics of the variables, offering the advantage of interpretability by a domain expert [12].

The main objectives of feature selection are: a) to avoid overfitting and improve model performance, b) to provide faster and more cost-effective models and c) to gain a deeper insight into the underlying processes that generated the data.

Regarding the used feature selection procedure in this study, a wrapper type technique was applied (sequential backward elimination) using the random forest algorithm [13], which utilizes ensembles of decision trees. In addition, a multivariate filter was used as an option to reduce the co-linearity among features of the microarray dataset, prior to the application of the wrapper method. This filtering together with the imputation procedure, is a departure from a merely COD method, towards a GFF approach, although here no further transformation is applied to the feature vectors.

The random forest algorithm, among other ensemble learning methods, is reported to be successful in variance reduction which is associated with overfitting [14]. In addition, we utilized the option of stratifying the bootstrapped samples with equal number of cases per class [15]. This is compatible with the Balanced Random Forest (BRF) approach which is computationally more efficient with large imbalanced data, since each tree only uses a small portion of the training set to grow. Additionally is less vulnerable to noise (mis-labeled class) than the Weighted Random Forest (WRF) were a heavier penalty is placed on misclassifying the minority class [16]. BRF alleviated the class imbalance problem which is a common problem in disease diagnosis where the disease cases are rare as compared with normal populations. The recognition goal is to detect people with the disease, thus a favorable classification model in one that provides a higher identification rate on the disease category.

2 Multi-modal Data Fusion of Separate Datasets

All the programming of the workflow was implemented in R [17].

Image Data

The dataset derived from skin lesion images contained 972 instances of nevus skin lesions and 69 melanoma cases. Three types of features are analyzed: Border Features which cover the A and B parts of the ABCD-rule of dermatology, Color Features

which correspond to the C rules and Textural Features, which are based on D rules. The total number of features assessed was 31 from the initial set of 32 (one feature was removed as having zero variation across the samples). The relevant pre-processing which produced all the features is described in [18].

Microarray Data

The dataset regarding microarray data was taken from the Gene Expression Omnibus (GEO) [19], GDS1375. In that experiment, total RNA isolated from 45 primary melanoma, 18 benign skin nevi, and 7 normal skin tissue specimens were analyzed on an Affymetrix Hu133A microarray containing 22,000 probe sets [20]. The dataset contains the values of MAS5-calculated signal intensities after global scaling the average intensity to 600.

The data retrieval from GEO was performed using GEOquery [21] and processed with limma [22] R packages from the Bioconductor project [23], following the main steps as listed in the R script produced by the GEO2R tool [24]. The input contrast levels were differentially expressed genes between melanoma versus skin and melanoma versus nevus. 1701 genes from a linear model fit were extracted setting FDR for multiple testing adjustment, p-value 0.001 and 2-fold changes as thresholds. As a normalization step, after taking the logarithms of the values, the mean values of normal skin were subtracted from the rest of the data, and the normal skin columns were removed from the table.

Data Integration

The two tables containing the microarray and image data were merged to one block sparse matrix with dimensions 1104 rows x 1734 columns, marking the not available values as NA. The rows contain the microarray and image data samples, and the columns microarray and image features plus one binary response variable (0 for nevus and 1 for melanoma).

Missing Values Imputation

Although there are several software packages implementing advanced imputation methods [25], they could not be utilized in this unified dataset where the multi-modal data have only the class variable column as complete. In this study we considered two simple imputation methods: mean value imputation per class and random normal imputation per class. In the second case, after taking the mean value (m) and standard deviation (sd) of each feature (ignoring the NA values) per class, we randomly filled the missing values sampling from an assumed normal distribution having as parameters: (m, sd).

For the efficient execution of the imputations, the `plyr` R package was used [26].

3 Feature Selection

The setup of the in-silico experiment involved the examination of the reported selected feature subsets when: a) applying a co-linearity removal filter to the

microarray dataset prior to the execution of the selection algorithm (marked as: Filtered/Unfiltered), and b) setting a 95% tolerance threshold to the best obtained performance criterion (Tolerance/Best). The tolerance in the performance method allows the selection of a subset size that is small enough but without sacrificing too much performance, and can produce good results where there is a plateau of good performance for larger subset sizes. The combination of these two parameters (prior filtering and tolerance threshold) resulted in the examination of four distinct cases.

For each of the four cases, a 10-fold cross-validation procedure was performed with 50 repetitions on four different datasets: only the microarray data (marked as om), the unified dataset with mean imputations (m.i), the unified dataset with normal random imputations for the NA values (nr.i), and only the image data (oi). In all the repetitions, the nr.i dataset was imputed anew, thus providing more sampling variations. Prior the application of the repetitions, the datasets were centered and scaled as a pre-processing step on the predictors.

The feature selection workflow was setup using the R package `caret` (classification and regression training) [27]. The search algorithm employed in caret uses the recursive feature elimination method on predefined sets of predictors, and in this study the length of the variable subsets was defined as [1 to 10, 15, 20, 25, 30, 35, 40, 45, 50], except for the image only data where the subsets were [1 to 10, 15, 20, 25, 30, 31].

For each of the 50 repetitions, the optimum subset number of predictors was recorded, along with the names of the predictors, and the performance attained. As performance metric the area under the ROC curve (auc) was set. The auc of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance. This is equivalent to the Wilcoxon test of ranks, and also it is closely related to the Gini coefficient [28].

4 Preliminary Results and Discussion

The results of the trials are depicted at Fig. 1.

Regarding the median value of optimum performances, in all cases an almost perfect score was achieved in the case of the unified datasets. Only-image dataset (oi) exhibited the lower performance and the higher subset numbers.

The application of the co-linearity reduction filter reduced the dispersion of the optimum subset number. Furthermore, the execution time in the reduced dataset was 4 times faster, analogous with the remained feature number after the use of the filter (482 from the initial 1701 differentially expressed genes in the microarray dataset).

The results on the imputed datasets exhibited also a minimization of the dispersion of the subset numbers. In addition, the two imputed datasets presented almost the same distribution. Nevertheless this similarity ended when we compared the gene sets retrieved in each case. As shown at Tables 1 and 2, the normal random imputation dataset (nr.i) resulted in a considerably more stable selection of features comparing to the mean imputation unified dataset (m.i). The same pattern is observed for the

unfiltered cases too. In the unfiltered cases, the nr.i dataset exhibited far better stability in the predictors' selection even to the microarray-only dataset. The features resulted from the mean imputation unified dataset presented high instability, and though proved as the least preferable option to the imputation procedure.

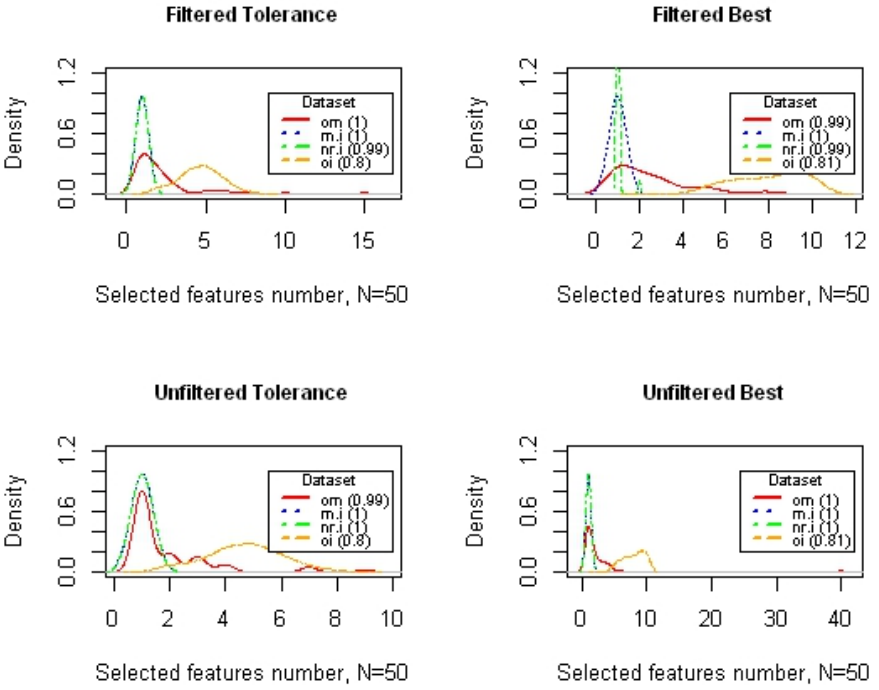

Fig. 1. Density plots of the optimum features number from 50 repetitions. The four datasets are: only microarray (om), mean imputation (m.i), normal random imputation (nr.i) and only image (oi). In parentheses are the medians of the obtained performances (auc) for each dataset.

The deficiency of using only performance indicators for marker discovery has been noted in the literature [29] and this is consisting with the findings of this study. The measure of stability of feature selection results with respect to sampling variations provides higher confidence in discovered biomarkers. But in this case, the imputations with the normal random method acted on the unified dataset as an additional variation factor. This additional variation resulted in the retrieval of smaller optimum subsets of features, consisting of fewer re-occurring genes as possible biomarkers.

Notably, none of the image-derived features were present to the top selected features of the unified datasets, as seen at Tables 1 and 2. In order to assess the importance ranking of image-features, 50 repetitions of the random forest algorithm

were run for the unified dataset imputed by the two methods (m.i and nr.i). Each of the resulted 50 lists of features was sorted by decreasing importance. Next, the positions of the image-features in the lists were collected and the density plots for the filtered/unfiltered cases are shown at Fig 2. Random forest avails four importance measures [30] and in this case the "MeanDecreaseGini" criterion was chosen. The results using the other three criteria were similar.

Table 1. Top features (genes) selected after 50 repetitions of the 10-fold cross-validation modeling for the Best-Filtered case in each of the three datasets

Feature (om)	Freq. (om)	Feature (m.i)	Freq. (m.i)	Feature (nr.i)	Freq. (nr.i)
CDC37L1	47	NEIL1	4	CDC37L1	49
RRAS2	34	IFI16	3	RRAS2	2
SLC7A8	18	CTDSPL	2		
HPCAL1.1	14	DLK2	2		
IFT81	8	NADK	2		
SSBP2	6	OR2A9P	2		
GIPC2	5	PIK3C2G	2		
CTDSPL	3				

Table 2. Top features selected at the Tolerance-Filtered case

Feature (om)	Freq. (om)	Feature (m.i)	Freq. (m.i)	Feature (nr.i)	Freq. (nr.i)
CDC37L1	45	PARD3	5	CDC37L1	40
RRAS2	25	ACOT9	3	RRAS2	6
SLC7A8	17	CYP4F3	3	HPCAL1.1	2
HPCAL1.1	10	FZD10	3	SSBP2	2
IFT81	6	NEIL1	3		
GIPC2	5	ACADL	2		
CTDSPL	4	MTUS1	2		
NEIL1	4	PER3	2		
SSBP2	3	PPP2R3A	2		
SMAD5OS	2	SMAD5OS	2		

The majority of the image-features ranked as less important when compared to the microarray features. This implies their lower informative power with respect to the total observed variation in the integrated dataset, probably due to the fact that technical covariance but also size, leave their fingerprint in the integration process, despite the application of normalization techniques, thus inflicting their effect on the response vector of the disease. When using the nr.i method however, a better performance of the image features is observed, which is captured as their more frequent presence in higher positions of the classifier's vector, in discord with the results of the m.i method. Mean imputation process resulted in scoring all image features in the lowest positions of the complete feature set, considering them less informative compared to the microarray features. In this sense, it is obvious that the

normal random imputation method yields a more impartial effect, as can be surmised from the improved score of the image related features, providing practical value to its application in the integration process, as the simulated dataset thus derived, is a more realistic representation of the real one.

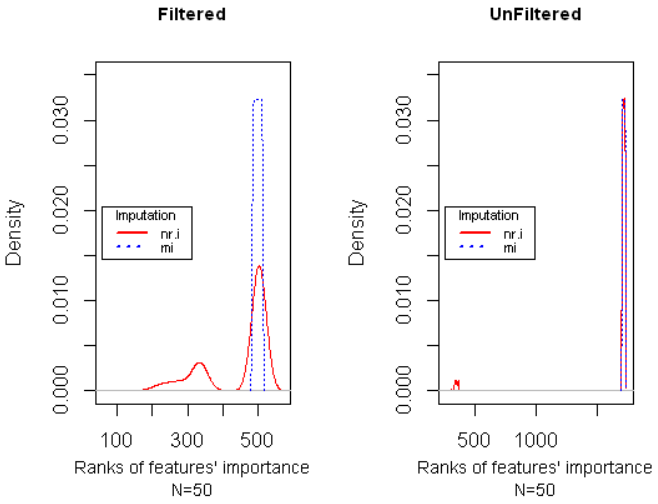

Fig. 2. Density plots of importance ranks for image-derived features (ranking in the x-axis is in decreasing order of importance)

This is the first attempt, to the best of our knowledge, to assess feature selection algorithms on integrative datasets retrieved from separate sources (modalities) dissecting the same pathological mechanism.

As future work we intend to examine a broader and more versatile in its mathematical origin, set of imputation methods for heterogeneous data integration from separate sources (data perturbation), as well as the application of different data-mining algorithms (function perturbation).

Acknowledgments. The authors would like to thank Dr. Max Kuhn creator of the `caret` R package for his kind assistance for the exploitation of this package.

References

1. Fenner, J.W., Brook, B., Clapworthy, G., Coveney, P.V., Feipel, V., Gregersen, H., Hose, D.R., Kohl, P., Lawford, P., McCormack, K.M., Pinney, D., Thomas, S.R., Van Sint Jan, S., Waters, S., Viceconti, M.: The EuroPhysiome, STEP and a roadmap for the virtual physiological human. Philos. Transact. A Math. Phys. Eng. Sci. 366, 2979–2999 (2008)
2. Rohlfing, T., Pfefferbaum, A., Sullivan, E.V., Maurer, C.R.: Information fusion in biomedical image analysis: combination of data vs. combination of interpretations. Inf. Process. Med. Imaging 19, 150–161 (2005)

3. Haapanen, R., Tuominen, S.: Data combination and feature selection for multi-source forest inventory. Photogrammetric Engineering and Remote Sensing 74, 869–880 (2008)
4. Jesneck, J.L., Nolte, L.W., Baker, J.A., Floyd, C.E., Lo, J.Y.: Optimized approach to decision fusion of heterogeneous data for breast cancer diagnosis. Med. Phys. 33, 2945–2954 (2006)
5. Lee, G., Doyle, S., Monaco, J., Madabhushi, A., Feldman, M.D., Master, S.R., Tomaszewski, J.E.: A knowledge representation framework for integration, classification of multi-scale imaging and non-imaging data: preliminary results in predicting prostate cancer recurrence by fusing mass spectrometry and histology. In: Proceedings of the Sixth IEEE International Conference on Symposium on Biomedical Imaging: From Nano to Macro, pp. 77–80. IEEE Press, Boston (2009)
6. Tiwari, P., Viswanath, S., Lee, G., Madabhushi, A.: Multi-Modal Data Fusion Schemes for Integrated Classification of Imaging and Non-Imaging Biomedical Data. In: IEEE International Symposium on Biomedical Imaging (ISBI), pp. 165–168 (2011)
7. Balazs, M., Ecsedi, S., Vizkeleti, L., Begany, A.: Genomics of Human Malignant Melanoma. In: Tanaka, Y. (ed.) Breakthroughs in Melanoma Research. InTech (2011)
8. Timar, J., Gyorffy, B., Raso, E.: Gene signature of the metastatic potential of cutaneous melanoma: too much for too little? Clin. Exp. Metastasis 27, 371–387 (2010)
9. Martins, W.K., Esteves, G.H., Almeida, O.M., Rezze, G.G., Landman, G., Marques, S.M., Carvalho, A.F., Reis, L.F.L., Duprat, J.P., Stolf, B.S.: Gene network analyses point to the importance of human tissue kallikreins in melanoma progression. BMC Med. Genomics 4, 76 (2011)
10. Ogorzalek, M., Nowak, L., Surowka, G., Alekseenko, A.: Modern Techniques for Computer-Aided Melanoma Diagnosis. In: Murph, M. (ed.) Melanoma in the Clinic - Diagnosis, Management and Complications of Malignancy. InTech (2011)
11. Maglogiannis, I., Doukas, C.N.: Overview of advanced computer vision systems for skin lesions characterization. IEEE Trans. Inf. Technol. Biomed. 13, 721–733 (2009)
12. Saeys, Y., Inza, I., Larranaga, P.: A review of feature selection techniques in bioinformatics. Bioinformatics 23, 2507–2517 (2007)
13. Breiman, L.: Random Forests. Machine Learning, 5–32 (2001)
14. Sun, Y., Wong, A.K.C., Kamel, M.S.: Classification of imbalanced data: A review. International Journal of Pattern Recognition and Artificial Intelligence 23, 687–719 (2009)
15. Liaw, A., Wiener, M.: Classification and Regression by randomForest. R News 2, 18–22 (2002)
16. Chen, C., Liaw, A., Breiman, L.: Using Random Forest to Learn Imbalanced Data (2004), http://www.stat.berkeley.edu/users/chenchao/666.pdf
17. R Foundation for Statistical Computing, http://www.R-project.org
18. Maragoudakis, M., Maglogiannis, I.: Skin lesion diagnosis from images using novel ensemble classification techniques. In: 10th IEEE EMBS International Conference on Information Technology Applications in Biomedicine, Corfu, Greece (2010)
19. Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Muertter, R.N., Holko, M., Ayanbule, O., Yefanov, A., Soboleva, A.: NCBI GEO: archive for functional genomics data sets–10 years on. Nucleic Acids Res. 39, D1005–D1010 (2011)
20. Talantov, D., Mazumder, A., Yu, J.X., Briggs, T., Jiang, Y., Backus, J., Atkins, D., Wang, Y.: Novel genes associated with malignant melanoma but not benign melanocytic lesions. Clin. Cancer Res. 11, 7234–7242 (2005)
21. Davis, S., Meltzer, P.: GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinformatics 14, 1846–1847 (2007)

22. Smyth, G.K.: Limma: linear models for microarray data. In: Bioinformatics and Computational Biology Solutions using R and Bioconductor, pp. 397–420. Springer, New York (2005)
23. Gentleman, R.C., Carey, V.J., Bates, D.M., et al.: Bioconductor: Open software development for computational biology and bioinformatics. Genome Biology 5, R80 (2004)
24. NCBI GEO, http://www.ncbi.nlm.nih.gov/geo/info/geo2r.html
25. Horton, N.J., Kleinman, K.P.: Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. Am. Stat. 61, 79–90 (2007)
26. Wickham, H.: The Split-Apply-Combine Strategy for Data Analysis. Journal of Statistical Software 40, 1–29 (2011)
27. Kuhn, M., Weston, S., Williams, A., Keefer, C., Engelhardt, A.: caret: Classification and Regression Training (2012), http://CRAN.R-project.org/package=caret
28. Fawcett, T.: An introduction to ROC analysis. Pattern Recogn. Lett. 27, 861–874 (2006)
29. He, Z., Yu, W.: Stable feature selection for biomarker discovery. Comput. Biol. Chem. 34, 215–225 (2010)
30. Breiman, L.: Manual on setting up, using, and understanding random forests v3.1. p. 10, 11 (2002), http://oz.berkeley.edu/users/breiman/Using_random_forests_V3.1.pdf