Aggregating Abstaining and Delegating Classifiers For Improving Classification performance: An application to lung cancer survival prediction

Mohamed-Ramzi TEMANNI[1,2,*], Sajjad Ahmed NADEEM[1,2,*], Daniel P. BERRAR[3], Jean-Daniel ZUCKER[1,2,4]

[1] Laboratoire d'Informatique Médicale et de Bioinformatique, LIM&BIO UFR SMBH, Université Paris 13, Bobigny, France

[2] Equipe INSERM U872 Nutriomique, Service de Nutrition, Hôpital Hôtel-Dieu, Paris, France

[3] School of Biomedical Sciences, University of Ulster at Coleraine, BT52 1ST, Northern Ireland

[4] IRD UR079, Centre IRD Ile de France 93143 Bondy Cedex, France

*both authors contributed equally to this paper

The incidence of lung cancer is increasing, particularly among elderly patients with approximately 40% arising in patients over 70 years [1]. In 2002, approximately 6,700,000 death were due to cancer (3,796,000 men and 2,928,000 women), corresponding to the third cause of death [2]. Lung cancer has the highest overall mortality in the Western World with over 1 million deaths annually. Data extracted from Microarrays chips is considered to be an important source for providing insight about cancer. Particularly, when it concerns survival prediction outcome, several studies have reported on the successful application of supervised machine learning approaches to prediction of cancer [3-5].

These models are, in general, built using microarray data and are known to provide more accurate results than models built using classical clinical parameter[6]. Nevertheless, the latter still provide good results and thus should not be discarded when trying to build survival outcome prediction methods. Different approaches show the efficiency of combining microarray and clinical data to provide a more accurate prediction model[7, 8]. These results show that heterogeneous data may provide complementary information that combined may provide better description about patients or at least about a subset of the patients. Although combining microarray and clinical data improves prediction accuracy this improvement is still limited.

For such a problem where it's hard to achieve high accuracy some new trend of approaches try to focus on giving a prediction only if it assumes that the sample has a high confidence. Friedel propose an abstaining classification model that abstains on samples from the dataset where we have doubt about the outcome and predict only sample with high confidence[9]. Such algorithms provide an improved accuracy and also provide the abstention rate of the suggested model. The empirical results shows that abstaining classifiers improve prediction accuracy but abstention rate could be high. Another approach aiming to improve the robustness of the result is the one proposed by Ferri[10]. This method delegates samples from the dataset that are more likely to be badly predicted to another algorithm. This model improves prediction accuracy but in some case the improvement could be limited as prediction methods tend to provide similar results on some dataset.

In our case we deal with a dataset where each patient is described by his gene expression profile and his clinical profile. We propose an abstaining/delegating method where a first model constructed using microarray data predict the outcome on a subset with high confidence and delegate the subset with low confidence to the model built using clinical data. This second model

does the same as the first model by predicting the outcome of high confidence patient and abstains on the remaining low confidence subset.

We applied our abstaining/delegating approach to five algorithms: support vector machines[11], decision trees[12], rule-based learning[13], naïve Bayes[14] and random forests[15]. Empirical results show that our method provide a higher accuracy compared to the use of standard algorithm. Compared to the abstaining model the delegation reduces the abstention rate.

Our general prediction model can be summarized as following: First, we train a classifier in 10-fold cross-validation using only the gene expression data. The task is to predict a patient's 5-year survival outcome. If the model makes a prediction with a confidence below the threshold X, then we do not accept this prediction. The corresponding case is delegated to the second model, which relies only on the clinical information. Although the predictions are provided by two different sub models, the results are treated as if only one model was applied to one separate validation set. This is reasonable as the training set size is only slightly smaller than the size of the original set and hence the different models are assumed to agree to a large extent. For investigating the effects of interchanging the use of clinical and microarray dataset, we also computed the prediction accuracies and abstention rates by using clinical data and delegating the abstained data to the second model where microarray data were used.

In this research work we used two publicly available microarray datasets: the Harvard lung cancer dataset [5] and the Michigan lung cancer dataset [3]. These datasets have a different number of genes and clinical attributes, but we kept only the attributes and genes common to both dataset as described in Berrar et al[16]. Microarray data used here were generated on an Affymetrix platform U133A.

Table 1: Datasets used for the 5-years survival prediction of cancer patients

	Lung Cancer (Harvard43)	**Lung Cancer (Michigan93)**
Class distribution	22 (survival) /21(no survival)	26(survival)/ 67(no survival)
Clinical Data	6 attributes (Age, Sex, T, N, M, Stage)	
Microarray data	3 588 genes	

From the Harvard lung cancer dataset, we used a subset of 43 patients. We selected only those patients having clinical and microarray data available and we excluded all patients that were censored and whose survival time was shorter than 5 years. In this selection, 22 patients survived after 5 years and 21 died before 5 years. Clinical data are age, sex, the TNM stage and the summary stage. T describes the primary tumor according to its size and location. N applies to the lymph nodes that drain fluid from the area of the tumor and whether the cancer has spread to them. M explains whether the cancer has spread to distant areas in the body. TNM stage was coded using three attributes, for example the following stage T2M1N0 is coded (2,1,0) in the case where we have a stage M1 the corresponding code will be (-1,-1,1) and we converted the summary stage in number(IA=1, IB=2, IIA=3, IIB=4, IIIA=5, IIIB=6, IV=7). The expression data table contains the transcriptional profile of 3588 genes. In the selection of patients from the Michigan dataset, 26 patients survived after 5 years and 67 died before 5 years. Clinical parameters selected for the Michigan dataset include age, sex, smoking history, TNM classification, tumor stage, survival time in months, and censor index and others. Expression data for 3588 genes is available. Table 1 resumes the data sets used in the present study.

Table 2 : Results using Havard Data set

Harvard Dataset		SMO	Random Forests	naïve Bayes	PART	J4.8
Without Abstention	*Accuracy (Clinical)*	75.0%	77.5%	72.5%	70.0%	67.5%
	Accuracy (Microarray)	72.5%	55.0%	67.5%	52.5%	50.0%
Abstention Clinical	*Accuracy*	78.6%	79.5%	93.1%	70.0%	64.9%
	Abstention Rate	**30.0%**	**2.5%**	**27.5%**	**0.0%**	**7.5%**
Abstention Microarray	*Accuracy*	72.5%	55.0%	67.5%	51.3%	50.0%
	Abstention Rate	**0.0%**	**0.0%**	**0.0%**	**2.5%**	**0.0%**
Delegation (Cli to Micro)	*Accuracy*	77.5%	81.6%	90.0%	75.5%	67.3%
	Abstention Rate	**0.0%**	**0.0%**	**0.0%**	**0.0%**	**0.0%**
Delegation (Micro to Cli)	*Accuracy*	72.5%	55.0%	67.5%	52.5%	50.0%
	Abstention Rate	**0.0%**	**0.0%**	**0.0%**	**0.0%**	**0.0%**

In the results obtained from the Harvard data set in the case of naïve Bayes, on abstaining at the rate of 27.50% we obtain 93.10% accuracy instead of 72.50% accuracy without abstention. While delegating although the accuracy decreases to 90.00% but abstention decreases from 27.50% to nil. In case of Random forests, we obtained 79.49% of accuracy and 2.50% abstention rate. On delegation we see that this abstention even vanishes and we obtain 81.63% accuracy. We observe a slight increase in accuracy using all algorithms in the Harvard data sets with absolutely no abstention after delegation. Table 2 summarizes the results obtained with Harvard data set.

Table 3 : Results for Michigan Data set

Michigan Dataset		SMO	Random Forests	naïve Bayes	PART	J4.8
Without Abstention	*Accuracy (Clinical)*	62.2%	64.4%	58.9%	64.4%	62.2%
	Accuracy (Microarray)	42.2%	63.3%	67.8%	55.6%	60.0%
Abstention Clinical	*Accuracy*	86.4%	62.2%	70.1%	64.5%	63.2%
	Abstention Rate	**51.1%**	**8.9%**	**25.6%**	**15.6%**	**15.6%**
Abstention Microarray	*Accuracy*	41.6%	62.7%	69.4%	55.6%	60.2%
	Abstention Rate	**1.1%**	**7.8%**	**5.6%**	**0.0%**	**2.2%**
Delegation (Cli to Micro)	*Accuracy*	63.5%	61.1%	67.4%	63.8%	61.4%
	Abstention Rate	**1.2%**	**1.1%**	**1.1%**	**1.2%**	**2.4%**
Delegation (Micro to Cli)	*Accuracy*	42.2%	63.5%	69.8%	56.0%	59.3%
	Abstention Rate	**0.0%**	**5.6%**	**1.1%**	**0.0%**	**0.0%**

It is also obvious from the results of Michigan data set that using SMO and clinical data in abstention model we have 86.36% accuracy and 51.11% abstention rate. Applying naïve Bayes and abstaining on clinical dataset we have 70.15% with 25.56% abstention rate. But when we use delegation abstention rate decreases to 1.15%. The abstention rates in other cases are also very low.

Our results reflect that when using abstention there is a significant increase in prediction accuracy. For instance, while using naïve Bayes in our proposed model with Harvard dataset (table 2), prediction accuracy increased from 76.50% to 93.10% with abstention rate of 27.50%.

On delegation, although accuracy decreases to 90.00% but abstention rate is reduced to nil. With Michigan dataset, we observe higher accuracies on abstaining especially while using SMO algorithm. For example prediction accuracy improved from 62.22% to 86.36%.

Experimental results demonstrate that as compared to classical machine learning algorithms, the use of microarray and clinical data in abstaining-delegating model produces good prediction accuracy and reduces the abstention rate. Comparing classical classifiers with delegating and abstaining classifiers is still an open issue. For instance we can define a cost matrix as below:

Table 4 : Cost Matrix

True Class	Predicted Class		
	Pp	Pn	Abstention
P	0	FN=1	0.5
N	FP=1	0	0.5 (1-0.5)

In the above cost matrix, we define the cost for misclassification (false negative and false positive) equal to 1 and costs for abstention equal to 0.5. Now assume that we have 30 patients in our test set and our non-abstaining classifier (i.e., any classical machine learning algorithm) makes 7 wrong predictions. Assume that the abstaining classifier abstained on 6 cases and misclassifies 3.

Table 5: Cost Calculation Example

No of Patients	Non Abstaining Classifier	Abstaining Classifier	Abstaining- Delegating Classifier
30	7 misclassified out of 30	6 abstained 3 misclassified	2 abstained 3 misclassified
Cost	7x1 = 7	6x0.5 + 3x1 = 6	2x0.5 + 3x1 = 4

Using the cost scenario presented in table 10 and the example described above, the costs calculated in table 5, show that the costs of abstaining classifiers are less than that of non-abstaining classifiers. If we use delegating-abstaining model, the cost are even less. This implies that each time our proposed method reduces cost of prediction. This concept is supported by our results because although prediction accuracy does not increase considerably when delegating, in every dataset but each time abstention rate decreases considerably

Table 6 : Result Summary of Model Proposed in our work: Delegating on Clinical and Abstaining to Microarray

	Harvard Dataset	Michigan Dataset
Accuracy Range	67.4% to 90.0%	61.1% to 67.4%
Abstention Range	00.0%	1.10% to 2.3%

Table 7 : Result Summary of Model Proposed in our work: Delegating on Microarray and Abstaining to Clinical

	Harvard Dataset	Michigan Dataset
Accuracy Range	50.0% to 72.5%	44.2% to 69.8%
Abstention Range	00.0%	00.0% to 5.56%

Analysis of tables 5, 6, 7 shows that as our abstention rates are very low with good prediction accuracies and our proposed model produces substantial/encouraging results.

In this research work, we propose a new delegating/abstaining approach that takes advantage of multiple sources in order to provide more accurate and more confident system. We analyzed the interest of using both delegation and abstention approach to investigate a new way for enhancing the performances of data mining algorithms. Empirical results show the importance of this new approach in terms of increased prediction accuracy and minimized abstention rates and raise questions about the evaluation of this class of methods. More attention should be drawn toward the evaluation of these methods. Future works will focus on the analysis of more algorithms and evaluate the approach on other dataset. We will also focus on constructing an optimization approach for automatically selecting the model that gives the optimal trade-off between high accuracy and low abstention rate based on a cost matrix provided by experts.

References

1. **Sophie, B., et al., *Lung cancer in elderly patients: A retrospective analysis of practice in a single institution.* Critical Reviews in Oncology/Hematology, October 2007. 64(1): p. 43-48**
2. **Ferlay, J., et al., *GLOBOCAN 2002 cancer incidence, mortality and prevalence worldwide, IARC Cancer.* Lyon: IARC Press, 2004. Base No. 5(Version 2.0).**
3. **Beer D.G, et al., *Gene-expression profiles perdict survival of patients with lung adenocarcinoma.* Nature Medicine, 2002. 8(8): p. 816-24.**
4. **Lee, Y. and C.-K. Lee, *Classification of multiple cancer types by multicategory support vector machines using gene expression data.* Bioinformatics, 2003. 19(9): p. 1132-1139.**
5. **Bhattacharjee, A., et al., *Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses.* Proc. Natl. Acad. Sci. USA 2001. 98(24): p. 13790-13795.**
6. **Eden, P., et al., *"Good Old" clinical markers have similar power in breast cancer prognosis as microarray gene expression profilers.* European Journal of Cancer, 2004. In Press, Corrected Proof.**
7. **Gevaert, O., et al., *Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks*. 2006. p. e184-190.**
8. **Nevins, J.R., et al., *Towards integrated clinico-genomic models for personalized medicine: combining gene expression signatures and clinical factors in breast cancer outcomes prediction.* Hum. Mol. Genet., 2003. 12(90002): p. 153R-157.**
9. **Friedel, C.C., R. Ulrich, and K. Stefan, *Cost Curves for Abstaining Classifiers.* Proceedings of the ICML 2006 workshop on ROC Analysis in Machine Learning Pittsburgh, PA, (2006).**
10. **Pietraszek, T., *Optimizing Abstaining Classifiers using ROC Analysis.* ACM Press, New York, NY, USA, (2005). 119: p. 665 - 672.**
11. **Burges, C.J.C., *A Tutorial on Support Vector Machines for Pattern Recognition.* Data Mining and Knowledge Discovery, 1998. 2(2): p. 121-167.**
12. **Quinlan, J.R., *C4.5: Programs for Machine Learning. Morgan Kaufmann, SanMateo, CA.* (1993).**
13. **Freund, Y. and R.E. Schapire, *A short introduction to boosting.* Journal of Japanese Society for Artificial Intelligence,, 1999. 14(5): p. 771-780.**
14. **Mitchell, T.M., *Machine Learning*. 1997, Boston: McGraw-Hill.**
15. **Breiman, L., *Random forests.* Machine Learning, 2001. 45: p. 5 - 32.**
16. **Berrar, D.P., et al. *Integration of microarray data for a comparative study of classifiers and identification of marker genes*. in *The 4th International Conference on Critical Assessment of Microarray Data Analysis 2003(CAMDA03)*. 2003. Durham, North Carolina, USA.**