

BENCHMARKING INFORMATION DISCOVERY METHODS FOR MICROARRAY DATA IN BREAST CANCER

N. Lama
Milan University, Italy

P. Boracchi
Milan University, Italy

E. Biganzoli
INT, Italy

ABSTRACT

Feature selection & extraction refers to the task of dimensionality reduction and discriminant capacity control (1, 2).

The problem of selecting features related to a given outcome occurs in various machine learning applications. It is one of the most important issues in classification problems, because the choice of features can have a large impact on the performance of the produced classifier. This is a particularly relevant issue in the context of microarray datasets with thousands of features, most of which are likely to be uninformative or irrelevant for classification purposes (3).

Whilst the development of high throughput techniques such as microarray and proteomics has enormous potential to increase our understanding of biological system (4), cautionary tales against possible “noise discovery” has been published (5).

The need to suppress surrounding noise while extracting information has become imperative in order to face the challenge of relevant discovery in the molecular era.

Keywords: benchmarking, feature selection and extraction, microarray, breast cancer.

INTRODUCTION

The objective of the present study is to provide a starting point for a methodological approach to the design of benchmark studies aimed at the evaluation of performances of feature selection & extraction methods (from now on, simply FSE methods) applied to publicly available microarray datasets in Breast Cancer studies.

A set of measures will be proposed with the objective of studying the behaviour of these methods concerning the key characteristics of:

- *efficiency*; it can be measured directly by its training, tuning and running time
- *effectiveness*; it can be evaluated directly only over synthetic data, showing how the selected subset are similar to the optimal one. For real-world microarray data, only the predictive accuracy of the resulting model on the selected features can be used as an indirect measure.
- *reliability*; the repeatability of features selected measured within a re-sampling scheme (e.g. bootstrap or K-fold cross-validation).
- *descriptive capacity*; it might be evaluated by means

of information theory based measures and visualization techniques which could focus on the features ability to explain the underlying class structure.

This study should be viewed in the context of the benchmark and data-warehouse in microarray breast cancer research, anticipated in (6).

MATERIALS AND METHODS

Study Design

Such a kind of benchmark studies should be designed as retrospective (observational) study.

According to Altman (7) and Pepe (8) terminology, these studies might be considered a phase I-II diagnostic study. In fact, their purpose is basically exploratory and they are aimed to evaluate FSE methods performances on extracting the information value provided by gene expression profiling in breast cancer.

Study Subject

Study subjects are to be enrolled from breast cancer prognosis microarray studies on survival-related outcomes as overall or event-free survival (e.g. subjects who developed distant metastases within 5 years or subjects who continued to be disease free after a period of at least 5 years).

Currently available case-control studies with convenience sampling have different eligibility criteria for subject enrolment. Benchmarking FSE techniques on such a set of different operating settings would enable a more robust performance estimate.

Study Dataset

A “test bed” benchmark dataset repository is to be set up with Breast Cancer microarray datasets.

The datasets eligible for enrolment will be identified through the review process foreseen in (6).

The datasets should be publicly available in order to enable the reproducibility of results.

All available microarray datasets of breast cancer prognosis should be enrolled provided that they come from studies which earned scientific recognition and published in the last five years.

Quoting Russo et al. (9), the study of van't Veer et al. (10) has been one of the most extensive and informative studies performed to date. The dataset published in (10) will be therefore included in the benchmark dataset repository, taking into account of a great number of

studies which have re-analyzed it (11–16).

All the datasets originated from studies which compared their results to those published in (10) are to be carefully evaluated for the enrolment.

When applicable, original measures should be preferred to pre-processed dataset in order to prevent data transforming techniques from discarding relevant information and to avoid relying on possibly unrealistic assumptions.

Considering the high costs associated with microarray studies, the limited number of samples measured is a notorious challenge in microarray data analysis. For this reason, an additional eligibility criterion for the minimum number of patients that should be enrolled in the study is still an open issue.

FEATURE SELECTION AND EXTRACTION METHODS

For many classifiers, it is important to perform some type of feature selection; otherwise performance could degrade substantially with a large number of irrelevant features (3).

It can moreover be necessary to adopt different FSE techniques in succession in order to cope with practical limitation (memory or computation) of FSE algorithm implementations.

Langley (17) grouped different feature selection algorithms into two broad categories: the filter and the wrapper methods.

Filter methods perform FSE explicitly, independently of the inductive algorithm hence prior to the building of the classifier, in contrast to wrapper methods which perform FSE implicitly as an inherent part of the classifier building process hence using the inductive algorithm as the evaluation function.

In filters, the characteristics in the feature selection are less correlated to that of the subsequent learning method applied; therefore they have better generalization property.

Since in wrapper methods the usefulness of a feature is directly judged by the estimated accuracy of the learning method, Furnarello et al. (34) stressed the need to control the “selection bias” related to the optimization of a classification rule typical of wrapper algorithms.

Yu and Liu (49) claimed that when the number of features becomes very large, the filter model is usually a choice due to its computational efficiency.

Promising filter FSE techniques, which should be run over the benchmark repository datasets, are in particular: (Kernel) Principal Component Analysis (13, 14, 18), Factor Analysis (19), (Kernel) Canonical Correlation Analysis (18, 20) and Independent Component Analysis (21, 22).

Visualization techniques like Neuroscale (Lowe and Tipping (47)) and Generative Topographic Mapping

(Bishop et al. (48)) could also provide useful insight of the data structure as well as describing the intrinsic power of the selected features through a projection onto a latent space.

Performance evaluation techniques

An optimal subset of features is always relative to a certain evaluation function i.e. an optimal subset chosen using one evaluation function may not be the same as that which uses another evaluation function (23).

There are many possible measures for evaluating feature selection algorithms and classification models (23, 24). Three different types of performance evaluation studies should be conducted:

- Information Content Analysis
- Classification Model Behaviour Study
- Cost of FSE.

To the aim of features stability evaluation and classification performance results validation, bootstrap re-sampling methods have to be adopted (25, 26).

For each dataset in the repository at least 500 bootstrap samples of training sets should be constructed, where, all the subjects are randomly re-sampled with replacement.

The number of bootstrap samples has to be increased for confidence interval estimations.

The competing FSE algorithms will be run over the same bootstrap samples.

The significance of a feature is correlated with the repeatability of selection according to the probabilistic analysis given in Fu and Fu-Liu (36). They considered the validity of a selected gene by its reliability in the sense that more often a gene is selected; the less likely chance is a factor.

Feature stability can be measured by the average level of agreement between set of relevant features chosen in all the bootstrap samples and can be evaluated by means of Jaccard index, as proposed by Yeung and Bumgarner (12).

Information Content Analysis

The goal of training classification model is to reduce the uncertainty about predictions on class labels C for the known observations X as much as possible.

Among non-linear correlation measures, many measures are based on the information-theoretical concept of entropy. The entropy of a variable X is defined as:

$$H(X) = -\sum_i P(x_i) \log P(x_i)$$

It is a measure of uncertainty of a random variable X .

The uncertainty of the random variable C , the class label, measured by its entropy $H(C)$ can be decomposed in $H(C) = H(C|X) + I(X,C)$, where $I(X,C)$ measures the certainty about C that is resolved by X (also known as the information gain), and $H(C|X)$ measures the residual uncertainty about C . In this term, the goal of feature selection is to achieve the higher $I(X,C)$ with the fewer features.

As claimed by Liu et al. (33), a serious deficiency of currently used multivariate approaches for feature set selection is that they are based on selecting genes which are maximally relevant with respect to the classes in study. The problem with this approach is that there might still be genes among the selected set that are heavily correlated with each other and thus leading to a redundancy in the selected feature set.

In this kind of benchmark studies, relevance and redundancy of the selected features should be analyzed through investigation of information based measures proposed in literature.

Attention have to be paid to the relevance and redundancy measures: 1) based on the mutual information described in (27–29) 2) based on the Markov-blanket approach (30) and its approximation through the symmetrical uncertainty measure (31) and adopted in (32) 3) based on the normalized mutual information described in (33) 4) entropy based SVM weight distribution (34).

More details about the above mentioned approaches are available in appendix A.

When appropriate, the test published in (35) could be adopted in order to assess whether the features selected by each FSE algorithm are significantly related to the clinical outcome of interest.

Relevance of features repeatedly selected from different combination of data instances composing bootstrap samples could also be analyzed as described in (36).

If applicable, a between-datasets comparison of the different selected features (e.g. subsets of selected genes) should be performed. To this aim, integrating the data with Gene Ontology function classifications would allow the identification of certain clusters of genes with good correlation. Resources containing annotation data for several gene expression platforms and functions to match genes across platforms using either Unigene-ids or the actual base sequence are cited in (37).

Descriptive model behaviour study

An optimal validation procedure should check whether the FSE algorithm output (selected subset of features) is the same as the actual subset of relevant features. Unfortunately the actual subset is unknown in the case of real-world datasets chosen for benchmarking.

The selected subset of features will be therefore tested for its accuracy with the help of a committee of

classifiers suitable for this task.

As different selection methods have a different bias in selecting features, similar to that of different classifiers, it is not fair to use only certain combinations of FSE methods and classifiers and try to generalize from the results that some feature selection methods are better than others without considering the classifiers (23). It is therefore necessary that the features extracted from each FSE methods are used in the following classifiers: Fisher Linear Discriminant Analysis (14, 38), logistic classifier (13) and (least squares) support vector machine (14).

For each classifier, typical ROC measures should be ascertained: True Positive Rate (TPR), False Positive Rate (FPR) and the Area under the Curve (AUC).

The prognostic/diagnostic value of the classifier should be also described by means of positive/negative Diagnostic Likelihood Ratios (DLR), since they quantify the increase in knowledge that is gained through the classifier.

Point and interval estimates will be computed using the formula given in Pepe (8).

Overall performances comparison could be investigated by adopting the test statistic proposed by Pesarin (39), and specifically advised in Hothorn et al. (40) for benchmark studies. Being K the number of algorithm being benchmarked over B learning samples and p_{kb} the performance of the k th algorithm of the committee provided by the b th learning sample, the statistic:

$$t_{BK} = \frac{\sum_k \left(\frac{1}{B} \sum_b p_{kb} - \frac{1}{BK} \sum_{k,b} p_{kb} \right)^2}{\sum_{k,b} \left(p_{kb} - \frac{1}{B} \sum_b p_{kb} - \frac{1}{K} \sum_k p_{kb} + \frac{1}{BK} \sum_{k,b} p_{kb} \right)^2}$$

can be used to test whether the K algorithms perform equally well against the alternative that at least one of them outperforms all other candidates. To this aim, it can be constructed a permutation test, obtaining the null distribution by permuting the labels of the K algorithms for each of the B samples, independently.

For the purpose of pair-wise comparison of promising algorithms various metrics might be used: absolute differences, odds ratios and relative probabilities. The preference is for the latter, hence point and interval estimates of: relative false positive rate (rFPR), relative true positive rate (rTPR) and relative diagnostic likelihood ratio (rDLR) have to be computed as showed in (8).

In the case that large-scale comprehensive population-based study is made available, the analysis for the determination of which covariates affect ROC measures might be performed as advised in (8).

In the case of standard clinical predictor could be

compared to predictors of disease outcome derived from gene expression levels, this will be done in an unbiased way, e.g. pre-validation (11).

Cost of FSE analysis

In this kind of studies the total costs incurred by the “number-crunching machines” due to FSE analysis have to be investigated.

Both direct and indirect costs are included in this form of cost analysis.

Costs are here conceived as computing time directly elapsed during FSE analysis or indirectly to the extent it affects the time required by the subsequent classifier building process.

Information about the different computing facilities exploited by each FSE algorithm should be listed for a critical appraisal of the costs incurred.

It shall be however noted that this comparison of computational efficiency is purely indicative and should not be considered an exhaustive cost analysis study because of the prototype nature of the algorithms involved and the runtime interpreted software environments.

The information collected for cost-of-FSE could be subsequently used for the conduction of the more informative cost-effectiveness/utility analyses considering the impact of costs on the performances in terms of information content and the subsequent classification accuracy.

E-SCIENCE INFRASTRUCTURE

The participants to these studies should be able to cooperate exploiting a set of tools that will be made available within the dedicated directory, as foreseen in the Biopattern web portal. In particular, a working group forum and a document repository for the delivery of reports should be available.

Such a directory could also provide a Concurrent Versioning System (CVS) service for the maintenance of a common software repository of data handling routines.

The analysis could be performed using different software environments; the choice of the latest version of the open source R (41) and Matlab (42) is preferred. The software repository should initially consist of a set of import procedures for each enrolled dataset and for all the analysis environments considered. This software repository will be shared among the participants and collected into this Project directory within the Biopattern web site.

It should be investigated the feasibility to get all the different algorithms run on a same machine. If

necessary, an application for hiring time on a high performance number crunching machine could be made.

With the objective of moving towards the Virtual Research Group foreseen in (43), it should be considered to exploit grid facilities made available to the Biopattern project members. Parallel version of statistical procedures could be considered, in particular for bootstrap model evaluation. The Simple Network of Workstation (SNOW) package (44) is a promising effort in R environment, since with suitable support software, such as the Globus toolkit environment (45), it should also be possible to use a computational grid as the engine. Eventually, focus shall be made towards collaborative applications Grid (46), enabling and enhancing human-to-human interactions with a shared use of data archives and simulations.

Such a Virtual Research Group could also help to address community-wide issues like the skewness of the results. Similarly to what happen for reviews of publications, if researchers work separately, only those who get significant results will publish. To understand the possible consequence of this, one can imagine that, if a great number of researchers working separately try to study the effects of two particular FSA algorithms to determine which one is better, some of them might get significant result only due to chance.

The possibility to discuss, develop and share a common methodological framework for comparative studies is the ultimate goal of this study and its associated sub-project (6).

Acknowledgements: This work was partially supported by the EU BIOPATTERN Network of Excellence, under contract 508803. We would like to thank Prof. David Lowe and his team for their cooperation and hospitality in the start-up phase of this study.

REFERENCES

1. Jolliffe I.T. 1972. *J. R. Statist. Soc. B.* 21(2):160–173.
2. Jolliffe I.T. 1973. *J. R. Statist. Soc. B* 22:21–31.
3. Speed T. 2003. “Statistical Analysis of Gene Expression Microarray Data. Interdisciplinary Statistics”, Chapman & Hall/CRC
4. Hack C.J., Lopez J.A. 2004. In Proceedings of the Symposium KELSI 2004, LNAI 3003,:9–19.
5. Ioannidis J.P.A. 2005. *Lancet* 365(9458):454–5,
6. Lama N., Boracchi P., Biganzoli E. 2005. In Proceedings of MEDASP, 12-20 January
7. Altman D.G., Lyman G.H. 1998. *Breast Cancer Res Treat* 52:379–393.
8. Pepe M.S. 2003. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford

University Press

9. Russo G., Zegar C., Giordano A. 2003 Oncogene, 22:6497–6507.
10. van 't Veer L.J., Dai H., van de Vijver M.J., He Y.D., Hart A.A.M., Mao M., Peterse H.L., van der Kooy K., Marton M.J., Witteveen A.T., Schreiber G.J., Kerkhoven R.M., Roberts C., Linsley P.S., Bernards R., Friend S.H. 2002. Nature 415(6871):530–6
11. Tibshirani R.J., Efron B. 2002. Stat. Appl. Gen. Mol. Biol. 1:1–18
12. Yeung K.Y., Bumgarner R.E. 2003 Genome Biol. 4(12):R83.
13. Lama N., Ambrogi F., Antolini L., Boracchi P., Biganzoli E. 2004 In Proceedings First International Workshop on Evaluation and Assessment of Diagnostic Performance.
14. Pochet N., DeSmet F., Suykens J.A.K., DeMoor B.L.R. 2004. Bioinformatics 20(17):3185–3195.
15. Ein-Dor L., Kela I., Getz G., Givol D., Domany E. 2005. Bioinformatics 21(2):171–8
16. Michiels S., Koscielny S., Hill C. 2005. Lancet 365(9458):488–92.
17. Langley P. 1994 In Proceedings of the AAAI Fall Symposium on Relevance 1–5.
18. Shawe-Taylor J., Cristianini N. 2004 “Kernel Methods for Pattern Analysis.” Cambridge University Press.
19. Goeman J. 2004 In 25th Annual Conference of the ISCB, Leiden, the Netherlands, August 15-19.
20. Kuss M., Graepel T. 2003 Tech. Rep. TR-108, Max Planck Institute for Biological Cybernetics.
21. Hyvarinen A., Oja E. 2000 Neural Netw 13(4-5):411–30.
22. Kreil D.P., MacKay D.J.C. 2003 Comparative and Functional Genomics 4(3):300–317.
23. Dash M., Liu H.: 1997 Intelligent Data Analysis, 1(3):131–156.
24. Molina L.C., Belanche L., Nebot A. 2002 In ICDM, IEEE Computer Society:306–313.
25. Efron B. 1983 JASA, 78:316–331.
26. Efron B., Tibshirani R. 1997 JASA, 92(438):548–560
27. Ding C.H.Q., Peng H.: In CSB, IEEE Computer Society 2003:523–529.
28. Huang D., Chow T.W. 2005 Neurocomputing, 63:325–343.
29. Chow T.W.S., Huang D. 2005 IEEE Trans. on Neural Networks, 16.
30. Koller D., Sahami M. 1996 In ICML:284–292.
31. Press W.H., Flannery B.P., Teukolsky S.A., Vetterling W.T.: “Numerical recipes in C: the art of scientific computing”. New York, NY, USA: Cambridge University Press 1988.
32. Yu L., Liu H. 2004, JMLR, 5:1205–1224.
33. Liu X., Krishnan A., Mondry A. 2005 Bioinformatics, 6:76.
34. Furlanello C., Serafini M., Merler S., Jurman G. 2003 Bioinformatics, 4:54,
35. Goeman J.J., van de Geer S.A., de Kort F., van Houwelingen H.C. 2004 Bioinformatics, 20:93–99.
36. Fu L., Fu-Liu C. 2005 Bioinformatics, 6:67.
37. Cope L., Zhong X., Garrett E., Parmigiani G. 2004 Statistical Applications in Genetics and Molecular Biology, 3:Article 29.
38. Dudoit S., Fridlyand J., Speed T.P. 2002. JASA, 97:77–87.
39. Pesarin F. 2001 : “Multivariate Permutation Tests: With Applications to Biostatistics”. Chicester: Jhon Wiley & Sons.
40. Hothorn T., Leisch F., Zeileis A., Hornik K. 2004, Journal of Computational & Graphical Statistics:1–22.
41. R Development Core Team: 2004 “R: A language and environment for statistical computing. R Foundation for Statistical Computing”, Vienna, Austria.
42. Matlab: The MathWorks, Inc. [<http://www.mathworks.com/products/matlab>].
43. Biopattern: Virtual Research Insitute [<http://www.biopattern.org/objective2.html>].
44. Rossini A., Tierney L., Li N. 2003. UW Biostatistics Tech. Rep. 193
45. Globus: The Globus Project. 2002. [<http://www.globus.org>].
46. Foster I.T., Kesselman C. 2000 “Computational Grids”. In VECPAR:3–37
47. D. Lowe and M. E. Tipping, 1997, "Advances in Neural Information Processing Systems", 9, 543–549, London, UK.
48. C. M. Bishop, M. Svensen and C. K. Williams, 1997, Neural Computation, 10, 215–234.
49. L. Yu, H. Liu: ICML Washington DC, 2003.

APPENDIX A

This appendix examines feature subset selection methods based on Information Theory. The importance of minimum redundancy and maximum relevancy in gene selection is pointed out in Ding and Peng (27). Their argument was that if a gene has expressions randomly or uniformly distributed in the

different classes, its mutual information with these classes is zero. If a gene is strongly differentially expressed for different classes, it should have large mutual information. The mutual information between two random discrete variables x, y is defined by:

$$I(x, y) = \sum_{i,j} p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i)p(y_j)}$$

with $p(x, y)$, $p(x)$ and $p(y)$ the joint probabilistic distribution and marginal distributions of the two variables.

Ding and Peng proposed mutual information as a measure of relevance of genes. Their idea of minimum redundancy is to select genes such that they are maximally dissimilar, using the mutual information to measure the level of “similarity” between genes. Their redundancy criterion is defined by:

$$W = \frac{1}{|S|^2} \sum_{i,j \in S} I(g_i, g_j)$$

where g_i is the i th gene expression and $|S|$ the number of the selected features in the S subset.

To measure the level of discriminant power of genes when they are differentially expressed for different classes in study, the following relevance criterion was proposed:

$$V = \frac{1}{|S|^2} \sum_{i \in S} I(C, g_i)$$

being C the target class.

Ding and Peng proposed also the two different optimization measures to be maximized during a feature selection procedure:

$$\max(V - W)$$

$$\text{and } \max(V / W)$$

Since mutual information tends to favor features with more values, it should be normalized with their corresponding entropy. For this reason, Liu et al. (33), Yu and Liu (32) choose the symmetrical uncertainty measure (Press et al. 1988 (31)), defined as:

$$\begin{aligned} U(X, Y) &= 2 \frac{I(X, Y)}{H(X) + H(Y)} \\ &= 2 \frac{H(X) + H(Y) - H(X, Y)}{H(X) + H(Y)} \end{aligned}$$

This measure is symmetrical and ranges from 0 (low mutual relevance) to 1 (high mutual relevance). A value closed to 1 indicates that the knowledge of one variable predicts the values of the other, whilst a value closed to 0 indicates that X and Y are almost independent.

It is to be noted however that such a pair-wise feature comparisons approach could not determine exactly feature redundancy. To this aim, an optimal measure is provided by the Markov blanket criteria defined by Koller and Sahami (30), as follows:

Let F be the full set of feature and C the class label.

Given a feature F_i , let $M_i \subset F$ ($F_i \notin M_i$), M_i is said to be a Markov blanket for F_i iff:

$$P(S_i | F_i, M_i) = P(S_i | M_i)$$

where $S_i = \{F \cup C\} - M_i - \{F_i\}$

The Markov blanket condition requires that F_i is conditionally independent of S_i given M_i ; that is F_i gives no information about S_i beyond what is already in M_i .

A feature F_i is therefore said to be redundant iff has a Markov blanket within the set of features F .

Unfortunately, finding a Markov blanket might be hard; searching for an optimal subset is combinatorial in nature and an exhaustive search is prohibitive with a large number of features. Searching for an optimal subset on a finite sample of data should be controlled for the implicit multiple-test Type I error. For these reasons, Koller and Sahami (30), Yu and Liu (32) developed and proposed iterative approaches which approximate the Markov blanket criterion.

It shall however be underlined that the estimation of Markov blanket measures or entropy-based measures poses a great challenge when put in practice, since they need probability density functions (pdf). The point here is principally on continuous features. One possibility could be to discretise them and to use their histograms to estimate their pdf, simplifying substantially the integration operation for mutual information computations.

As claimed by Huang and Chow (29), the accuracy of most histogram estimators is substantially degraded because of the sparse distribution of data in high-dimensional space.

Continuous kernel-based pdf estimators could be considered as a good alternative for performing feature selection. With continuous pdf estimators, the integral operation in the mutual information poses great computational difficulty.

Huang and Chow (29) proposed a method to estimate mutual information by means of Gaussian function as kernel function and employing the quadratic mutual information estimator:

$$I_{CS}(X, Y) = \log \frac{\iint p(x, y)^2 dx dy \iint p(y)^2 p(x)^2 dx dy}{\left(\iint p(x, y) p(y) p(x) dx dy \right)^2}$$

Using the properties of Gaussian pdf estimator the integrals can be substantially simplified (see Huang and Chow (29)).