Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Pattern Recognition 38 (2005) 2226–2228

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/patcog

Editorial

The fundamental role of pattern recognition for gene-expression/microarray data in bioinformatics

High-throughput measurement technologies, such as cDNA and oligonucleotide microarrays, are changing the practice of biology and medicine. Microarrays provide simultaneous expression (RNA abundance) measurements for thousands of genes and thereby facilitate analysis of the complex multivariate relations among genes. This new capability is being used to promote two major goals of functional genomics: (1) to use gene expression to classify disease on a molecular level; and (2) to discover genes that determine specific cellular phenotypes (diseases) and model their activity in a way that provides quantitative discrimination between normal and abnormal behavior. These goals correspond to diagnosing the presence or type of disease and to developing therapies based on the disruption or mitigation of aberrant gene function contributing to the pathology of a disease. Developing diagnostic tools at the RNA level involves designing expression-based classifiers to discriminate differences in cell state, such as one type of cancer or another. Engineering therapeutic tools involves synthesizing nonlinear dynamical networks to model gene regulation and deriving intervention strategies to modify network behavior. The classification methods of pattern recognition are clearly associated with diagnosis, but they also apply to therapy because prediction methods are used to identify gene–gene and gene–phenotype relations in network modeling. In discrete models, prediction of a target-gene value is given via a function of some predictor-gene values. This function is a multinomial classifier.

Expression-based microarrays phenotype classification requires designing a classifier that takes a vector of gene expression levels as input and outputs a class label to predict the class containing the input vector. Classification can be between different kinds of cancer, different stages of tumor development, different prognoses, or a host of such differences. Much early work in microarray phenotype classification has involved cancer, and many cancers have been considered. Classifier design involves assessing expression levels from RNA obtained from different tissues with microarrays, determining genes whose expression levels can be used as classifier variables, and then applying a rule to design the classifier from the sample microarray data. Expression values have randomness arising from both biological and experimental variability.

There has been an explosion of papers using a host of classification techniques; however, there has not been a concomitant effort at dealing with the daunting theoretical and procedural pattern-recognition issues raised by high-throughput classification. The problem: what is one to do when faced with thousands of potential features (gene expressions) and samples consisting of less than a hundred data points (microarrays)? This extreme disparity impacts the major aspects of classifier design: choosing a classification rule, error estimation, and feature selection. Owing to a lack of attention to this small-sample problem, the value of many results reported in the literature cannot be ascertained. What is one to make of using a complex classifier like CART or of using 50 features when the sample size is 30? The Vapnik–Chervonenkis theory should at least make scientists wary of using anything beyond very simple classification rules and a handful of features with the kinds of extremely small samples commonplace in microarray experiments [1]. The problem is not mitigated by obtaining a small error estimate, because in the vast majority of cases error estimation is achieved with some kind of cross-validation, which is not suitable for very small samples owing to its high variability in such circumstances. The imprecision of cross-validation is exacerbated by complex classification rules and large numbers of features.

The error of a classifier can be decomposed into the sum of the Bayes error for the feature-label distribution plus the error increase owing to constraining the form of the classifier plus the error increase owing to designing the classifier from sample data: $e_{classifier} = e_{Bayes} + e_{const} + e_{design}$. The Bayes error is outside the control of the designer; the choice of classification rule represents the engineering contribution. One would like to constrain classifier design to reduce e_{design} but at the same time not have e_{const} too high. Much outstanding work in pattern recognition has

0031-3203/$30.00 © 2005 Pattern Recognition Society. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2005.03.008

focused on bounding the expected design error in terms of the complexity of the classification rule and the sample size [2]. For the small samples confronting bioinformatics, the sample size is usually not sufficiently large to make the bound useful. Nonetheless, they tend to tell us that e_{design} is such an overriding problem that we should use very simple classifiers, even at the risk of increasing e_{const}. Basically, if the feature-label distribution is sufficiently complicated to make e_{const} unacceptably high, then the classification problem is intractable because the paucity of data will not allow good design. Even in the case of simple linear classifiers, regularization is advised. Ultimately the issue is how a classification rule behaves relative to distributional complexity for small samples. In this vein, there is evidence to support the heuristic that, if a complex classification rule provides good results, then with strong likelihood the task could have been accomplished with a very simple classifier [3]. The converse, however, is not true. Applying a complex classifier to an easily separable distribution will likely have much worse results than those obtained with a simple classifier, because in this case, e_{const} is small for both rules but e_{design} is large for the complex rule. We note that many practical problems of small-sample classification have long been known [4].

Owing to the huge numbers of variables in genomic problems, feature selection is a must. The problem is inherently combinatoric: to be assured of obtaining the best feature set of a certain size taken from a collection of potential features, one must test all features sets of the given size [5]. Except for very small feature sets, this is impossible. As the number of features grows, the error of a designed classifier commonly decreases and then increases with an increasing number of features. Owing to this “peaking phenomenon,” one cannot use all available features, or even more than a handful of them in some cases. Analysis of feature selection is usually based on simulation due to the difficulty of obtaining analytic results [6,7]. A central issue is to determine an optimal number of features for a classification rule and feature-label distribution relative to the sample size [8].

Perhaps, the area that has recently been the most overlooked in small-sample classification is error estimation. This is unfortunate because the scientific value of a classifier lies in the accuracy and precision of its error estimate. In the 1970s and 1980s, a decent amount of investigation went into error estimation [9], but thereafter it seems to have waned. Particularly troublesome is the extent to which cross-validation methods have been applied to small samples without justification, and often without caveat, even though their high variability is well known [2]. Recent studies demonstrate that cross-validation provides excessively imprecise error measurement [10] and poor feature-set ranking [11]. Significantly better performance is achieved by bootstrap [12] and the recently introduced bolstering [13], which has extensive roots within pattern recognition and is much more computationally efficient than re-sampling methods.

We next discuss clustering, which is being applied extensively in bioinformatics. Data clustering has historically lacked the two fundamental characteristics of pattern classification: (1) the error of a proposed classifier is estimated from data; and (2) given a family of classifiers from which to choose, a classification rule is used to obtain a classifier in the family. Of the two characteristics, the first is perhaps more basic, since without a decent error estimate, the worth of a classifier is unknown. Many validation techniques have been proposed for evaluating clustering results, but these are generally not set in the context of an encompassing probabilistic theory and therein based on an error criterion. The issue is serious. It goes to the epistemological foundations of clustering and therefore to the meaning of the conclusions based on clustering algorithms. Jain et al. [14] write. “Clustering is a subjective process; the same set of data items often needs to be partitioned differently for different applications. This subjectivity makes the process of clustering difficult.” Their warning should be heeded. Clustering is very difficult and should not be used without great prudence. But the problem is much deeper. Science is not subjective. It must involve a model leading to predictions that can be inter-subjectively tested. This is the role played by the error of a classifier, but that entails a probabilistic theory. To what is the output of a clustering algorithm to be compared, and how is the comparison to be measured so as to quantify predictability?

Whereas a classifier operates on a point to produce a label, a clustering algorithm operates on a set of points to produce a partition of the point set. The probabilistic theory of classification is based on a classifier being viewed as an operator on random points (vectors). A corresponding probabilistic theory of clustering would view a clustering algorithm as an operator on random point sets. Moreover, whereas the predictive capability of a classifier is measured by the decisions it yields regarding the labeling of random points, the predictive capability of a clustering algorithm must be measured by the decisions it yields regarding the partitioning of random point sets. Once this is recognized, the path to the development of error estimators for clustering accuracy and rules to learn clustering operators from data is open and the entire issue can be placed on firm epistemological ground [15]. This does not close the matter. While the manner in which clustering is used (operating on point sets) dictates a probabilistic theory in the context of random sets, new ideas and methods are required to employ the theory in scientific applications. This is not a trivial task. Whereas classification issues can be phrased in terms of the probability distribution function of a random vector, random point sets cannot be so easily modeled [16].

Given its key role in medical diagnosis and therapy, pattern recognition is poised to enter an exciting new phase, both in terms of application and theory. As briefly discussed, the huge imbalance between the numbers of features and the sample sizes requires new algorithms and error estimators whose small-sample properties are appreciated due to either mathematical theory or extensive simulation studies. One can confidently speculate that application advances will utilize biological knowledge in the design and choice of pattern-recognition procedures. The overall endeavor will provide challenging problems for decades and expand both theory and application immensely, perhaps in ways impossible to currently envision.

References

[1] E.R. Dougherty, Small sample issues for microarray-based classification, Comp. Functional Genomics 2 (2001) 28–34.

[2] L. Devroye, L. Gyorfi, G. Lugosi, A Probabilistic Theory of Pattern Recognition, Springer, New York, 1996.

[3] S.N. Attoor, E.R. Dougherty, Classifier performance as a function of distributional complexity, Pattern Recognition 37 (8) (2004) 1629–1640.

[4] S.J. Raudys, A.K. Jain, Small sample size effects in statistical pattern recognition: recommendations for practitioners, IEEE Trans. Pattern Anal. Mach. Intell. 13 (1991) 252–262.

[5] T. Cover, J. van Campenhout, On the possible orderings in the measurement selection problem, IEEE Trans. Systems Man Cybernet. 7 (1977) 657–661.

[6] M. Kudo, J. Sklansky, Comparison of Algorithms that Select Features for Pattern Classifiers, Pattern Recognition 33 (2000) 25–41.

[7] A.K. Jain, D. Zongker, Feature selection—evaluation, application, and small sample performance, IEEE Trans. Pattern Anal. Mach. Intell. 19 (1997) 153–158.

[8] J. Hua, Z. Xiong, J. Lowey, E. Suh, E.R. Dougherty, Optimal number of features as a function of sample size for various classification rules, Bioinformatics 21 (2005) 1509–1515.

[9] G.T. Toussaint, Bibliography on estimation of misclassification, IEEE Trans. Inform. Theory 20 (4) (1974) 472–479.

[10] U.M. Braga-Neto, E.R. Dougherty, Is cross-validation valid for small-sample microarray classification?, Bioinformatics 20 (2004) 374–380.

[11] C. Sima, U.M. Braga-Neto, E.R. Dougherty, Superior feature-set ranking for small samples using bolstered error estimation, Bioinformatics 21 (2005) 1046–1054.

[12] B. Efron, Estimating the error rate of a prediction rule: improvement on cross-validation, J. American Statistical Society 78 (1983) 316–331.

[13] U.M. Braga-Neto, E.R. Dougherty, Bolstered error estimation, Pattern Recognition 37 (2004) 1267–1281.

[14] A.K. Jain, M.N. Murty, P.J. Flynn, Data clustering: a review, ACM Comput. Surveys 31 (3) (1999) 264–323.

[15] E.R. Dougherty, M. Brun, A probabilistic theory of clustering, Pattern Recognition 37 (5) (2004) 917–925.

[16] D. Stoyan, W.S. Kendall, J. Macke, Stochastic Geometry and Its Applications, Wiley, Chichester, 1987.

Edward R. Dougherty
Department of Electrical Engineering,
Texas A&M University,
College Station, TX, USA
Division of Computational Biology,
Translational Genomics Research Institute,
Phoenix, AZ, USA