SPANDIDOS PUBLICATIONS

Tumor classification based on DNA copy number aberrations determined using SNP arrays

YUHANG WANG[1], FILLIA MAKEDON[1] and JUSTIN PEARLMAN[2]

[1]Department of Computer Science, Dartmouth College, Hanover, NH 03755; [2]Departments of Medicine and Radiology, Dartmouth-Hitchcock Medical Center, Lebanon, NH 03756, USA

Received September 7, 2005; Accepted October 11, 2005

Abstract. High-density single nucleotide polymorphism (SNP) array is a recently introduced technology that genotypes more than 10,000 human SNPs on a single array. It has been shown that SNP arrays can be used to determine not only SNP genotype calls, but also DNA copy number (DCN) aberrations, which are common in solid tumors. In the past, effective cancer classification has been demonstrated using microarray gene expression data, or DCN data derived from comparative genomic hybridization (CGH) arrays. However, the feasibility of cancer classification based on DCN aberrations determined by SNP arrays has not been previously investigated. In this study, we address this issue by applying state-of-the-art classification algorithms and feature selection algorithms to the DCN aberration data derived from a public SNP array dataset. Performance was measured via leave-one-out cross-validation (LOOCV) classification accuracy. Experimental results showed that the maximum accuracy was 73.33%, which is comparable to the maximum accuracy of 76.5% based on CGH-derived DCN data reported previously in the literature. These results suggest that DCN aberration data derived from SNP arrays is useful for etiology-based tumor classification.

Introduction

High-density single-nucleotide polymorphism (SNP) array is a high-throughput technology that genotypes more than 10,000 human SNPs on a single array (1). Single nucleotide polymorphisms (SNPs) are the most common type of DNA polymorphisms, which occur when a single nucleotide in the genome sequence is altered. Because SNPs occur abundantly with even spacing along the human genome, they offer significantly greater potential to be used as biomarkers for diagnosing genetic diseases such as cancers, compared to other less common polymorphisms and microsatellite markers. It has been shown that SNP arrays can be used to determine not only SNP genotype calls (1), but also DNA copy number (DCN) aberrations that are common in solid tumors (2).

In the past, effective cancer classification was demonstrated using microarray gene expression data (3-5), or DNA copy numbers derived from comparative genomic hybridization (CGH) arrays (6-8). However, the feasibility of cancer classification based on DCN aberrations determined by SNP arrays has not been previously investigated.

Here, we address this issue by applying state-of-the-art classification algorithms and feature selection algorithms to the DCN aberration data derived from a public SNP array dataset. Performance was measured via leave-one-out cross-validation (LOOCV) classification accuracy. Experimental results showed that the maximum accuracy was 73.33%, which is comparable to the maximum accuracy of 76.50% based on CGH-derived DCN data reported previously in the literature.

Materials and methods

In the problem of cancer classification using DCN aberration data, we still encounter the typical curse-of-dimensionality problem similar to cancer classification based on gene expression data: i) the number of SNPs greatly exceeds the number of tissue samples; ii) most SNP loci do not show the DCN aberration, and are not related to the given cancer classification problem; and iii) to overcome this curse-of-dimensionality problem, we can use feature selection algorithms to select a small subset of SNPs as features for classification. After selecting the informative SNPs, we then applied a classification algorithm to the reduced data. We used the Relief-F feature selection algorithm and three classification algorithms, namely, *k*-nearest neighbor (*k*-NN), support vector machine, and Naive Bayes.

Relief-F feature selection algorithm. One of the most widely used feature filters is the Relief-F algorithm (9). The basic idea of Relief-F is to draw instances at random, compute their nearest neighbors, and adjust a feature weighting vector to give more weight to features that discriminate the instance from neighbors of different classes. Specifically, it tries to

Correspondence to: Dr Yuhang Wang, Computer Science and Engineering Department, Southern Methodist University, P.O. Box 750122, Dallas, TX 75275, USA
E-mail: wyh@cs.dartmouth.edu

Key words: tumor classification, DNA copy number, single-nucleotide polymorphism, microarray, Relief-F algorithm

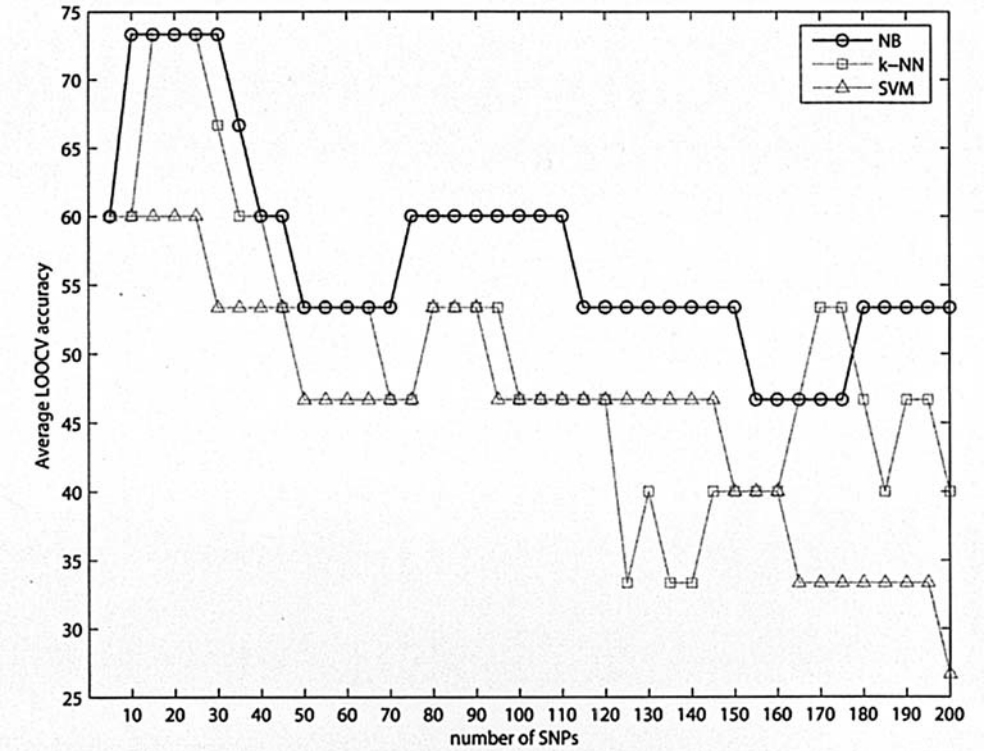

Figure 1. Comparison of the LOOCV accuracy of three classifiers combined with Relief-F.

find a good estimate of the following probability to assign as the weight for each feature f: $w_f = P$ (different value of f|different class) - P (different value of f|same class). This approach has shown good performance in various domains (10).

Nearest neighbor classifier. The *k*-nearest neighbor (*k*-NN) classifier is a well-known non-parametric classifier (11). To classify a new input x, the *k*-nearest neighbors are retrieved from the training data. The input x is then labeled with the majority class label corresponding to the *k*-nearest neighbors. For the *k*-NN classifier, we used the Euclidean distance as the distance metric in the experiments, and the best *k* between 1 and 10 was found by performing LOOCV on the training data.

Naive Bayes classifier. The Naive Bayes (NB) classifier is a probabilistic algorithm based on Bayes' rule and the simplifying assumption that feature values are conditionally independent given the class. Given a new sample observation, NB estimates the conditional probabilities of classes using the joint probabilities of training sample observations and classes.

Support vector machine. The support vector machine (SVM) belongs to a new generation of learning system based on advances in statistical learning theory (12). A linear SVM, which is used in our system, aims to find the separating hyperplane with the largest margin, defined as the sum of distances from a hyperplane (implied by a linear classifier) to the closest positive and negative exemplars. The expectation is that the larger the margin, the better the generalization of classifier. In a non-separable case, a linear SVM seeks a trade-off between maximizing the margin and minimizing the number of errors.

Results

Details about the data, preprocessing, experimental parameters, and results are provided in sections below using a public dataset.

Data

Data source. We used the SNP array dataset published by Zhao *et al* (2), which can be downloaded at http://research.dfci.harvard.edu/meyersonlab/snp/snp.htm. The original dataset contains raw data (CEL files) obtained from 43 tissue samples using Affymetrix *Xba*I mapping 130 array, which covers 10,043 SNP loci along all of the human chromosomes except the Y chromosome.

For cancer classification, we selected a subset from the original data. The selected subset contains data from 10 breast cancer patients and 5 small cell lung cancer (SCLC) patients.

Data processing. Raw data was processed following the same steps described previously (2). We re-analyzed the original raw dataset in its entirety using dChipSNP (13) to produce inferred DCN data for paired normal and tumor samples of the same individual. dChipSNP computes the raw DCN from the signal intensity, and employs a Hidden Markov model to infer DCN for each SNP, taking into account neighboring SNPs. The inferred DCN data are non-negative integers.

For each SNP in a pair of tumor/normal tissue samples, the DCN aberration was computed as the difference in DCN between the tumor and normal samples. For example, if the

n SNP locus is 2 for the normal sample, and 5 for sample, the corresponding DCN aberration is then 5-2=3. The DCN aberrations at all of the SNP loci were used as features.

Experimental settings. We consider the performance of the three machine learning models built by combining the Relief-F feature selection algorithm and the three classifiers discussed above. We implemented these models using Perl and the WEKA 3.4.3 (14), which is an open source collection of machine learning algorithms in Java.

In each fold of the LOOCV test, the DCN aberrations of 14 tissue pairs were used as training data, and the DCN aberrations of the one remaining tissue pair were used as test data. The feature selection algorithms were only applied to the training data, without prior knowledge of the test data. Therefore, in each LOOCV fold, the selected top-ranked SNPs may be different. In the LOOCV test, the classification accuracies of all 15 folds were averaged.

Results

Fig. 1 shows the LOOCV classification accuracies using *k*-NN, NB, and SVM combined with Relief-F. The x-axis is associated with the number of selected top-ranked SNPs, and the y-axis shows the average LOOCV accuracy. In the experiments, the top 5,10,15...200 SNPs were selected. From the results, we can observe that: i) the best LOOCV classification accuracy of 73.33% was achieved by *k*-NN and NB; and ii) selecting more SNPs does not necessarily increase the classification accuracy. In fact, all classifiers achieved the best performance when 5-30 SNPs were selected.

We also tried other feature selection algorithms, namely, information gain, gain ratio, and χ^2-statistic. Their performance in terms of LOOCV accuracy were comparable or worse than that of Relief-F (data not shown).

Discussion

This study presents some of the first results on applications of machine learning models in cancer classification using genome-wide DCN aberrations determined by SNP arrays. Using a public dataset, we found that the best LOOCV classification accuracy was 73.33%, which is comparable to the maximum accuracy of 76.50% based on CGH-derived DCN data reported previously (10). These results suggest that DCN aberration data derived from SNP arrays is useful for etiology-based tumor classification.

The informative SNPs selected by the feature selection algorithms may lead to the discovery of new tumor suppressor genes and oncogenes that are specific to a certain type of tumor. Although the selected top-ranked informative SNPs can lead to good LOOCV classification accuracy, their DCN properties still need to be confirmed by quantitative real-time PCR of the selected loci. The survey of additional cancer specimens will also help address their significance. We believe that the same machine learning models can also be applied to the classification of different subtypes of cancer, and the SNP arrays may be applied as a diagnostic tool in this area.

Acknowledgements

This work was supported in part by the National Science Foundation under grants ITR-0312629 and IDM-0083423, and also in part by FAMRI.

References

1. Matsuzaki H, Loi H, Dong S, Tsai YY, Fang J, Law J, Di X, Liu WM, Yang G, Liu G, Huang J, Kennedy GC, Ryder TB, Marcus GA, Walsh PS, Shriver MD, Puck JM, Jones KW and Mei R: Parallel genotyping of over 10,000 SNPs using a one-primer assay on a high-density oligonucleotide array. Genome Res 14: 414-425, 2004.
2. Zhao X, Li C, Paez JG, Chin K, Janne PA, Chen TH, Girard L, Minna J, Christiani D, Leo C, Gray JW, Sellers WR and Meyerson M: An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. Cancer Res 64: 3060–3071, 2004.
3. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD and Lander ES: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286: 531-537, 1999.
4. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick A, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Armitage JO, Weisenburger DD, Warnke R, Levy R, Wilson W, Grever MR, Byrd JC, Botstein D, Brown PO and Staudt LM: Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403: 503-511, 2000.
5. Gordon GJ, Jensen RV, Hsiao LL, Gullans SR, Blumenstock JE, Ramaswamy S, Richards WG, Sugarbaker DJ and Bueno R: Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. Cancer Res 62: 4963-4967, 2002.
6. Bastian BC, Olshen AB, LeBoit PE and Pinkel D: Classifying melanocytic tumors based on DNA copy number changes. Am J Pathol 163: 1765-1770, 2003.
7. Mattfeldt T, Gottfried HW, Wolter H, Schmidt V, Kestler HA and Mayer J: Classification of prostatic carcinoma with artificial neural networks using comparative genomic hybridization and quantitative stereological data. Pathol Res Pract 199: 773-784, 2003.
8. O'Hagan RC, Brennan CW, Strahs A, Zhang X, Kannan K, Donovan M, Cauwels C, Sharpless NE, Wong WH and Chin L: Array comparative genome hybridization for tumor classification and gene discovery in mouse models of malignant melanoma. Cancer Res 63: 5352-5356, 2003.
9. Kononenko I: Estimating attributes: analysis and extensions of RELIEF. In Proceedings of the European Conference on Machine Learning. Springer-Verlag Inc., New York, pp171-182, 1994.
10. Robnik-Sikonja M and Kononenko I: Theoretical and empirical analysis of ReliefF and RReliefF. Mach Learn 53: 23-69, 2003.
11. Dasarathy B: Nearest Neighbor Norms: NN Pattern Classification Techniques. IEEE Computer Society Press, 1991.
12. Vapnik VN: Statistical Learning Theory. Wiley-Interscience, 1998.
13. Lin M, Wei LJ, Sellers WR, Lieberfarb M, Wong WH and Li C: dChipSNP: significance curve and clustering of SNP-array-based loss-of-heterozygosity data. Bioinformatics 20: 1233-1240, 2004.
14. Witten IH and Frank E: Data mining: practical machine learning tools and techniques with Java implementations. Morgan Kaufmann, San Francisco, CA, 1999.