

Cancer Classification Using Loss of Heterozygosity Data Derived from Single-Nucleotide Polymorphism Genotyping Arrays

Yuhang Wang

Abstract—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 genome-wide Loss of Heterozygosity (LOH) calls can be derived by analyzing the genotypes calls measured by SNP arrays using paired tumor and normal tissue samples. The goal of this study is to evaluate the possibility of cancer classification using LOH calls. As a proof of concept, we applied 16 different combinations of classification algorithms and feature selection algorithms to a public data set that contains LOH calls of 10,043 SNP loci obtained from 10 breast cancer patients and 5 small cell lung cancer (SCLC) patients. Performance was measured in terms of the leave-one-out cross-validation (LOOCV) classification accuracy. Experimental results suggest that LOH calls derived from SNP arrays can be an excellent indicator of cancer type.

I. INTRODUCTION

Single-Nucleotide Polymorphism (SNP) array is a recently introduced high-throughput technology that genotypes more than 10,000 human SNPs on a single array [15]. 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 significant greater potential to be used as bio-markers for diagnosing genetic diseases including cancers, compared to other less common polymorphisms and microsatellite markers. Recently, it has been shown that genome-wide Loss of Heterozygosity (LOH) calls can be accurately derived by analyzing the genotypes calls measured by SNP arrays using paired tumor and normal tissue samples [12], [13].

LOH is particularly relevant to tumorigenesis. Mutated alleles of tumor suppressor genes (TSGs) are recessive and tumor cells can arise after LOH at relevant TSG locus/loci [10]. LOH can result from a variety of genetic mutation events, including hemizygous deletion, point mutation, mitotic nondisjunction, mitotic recombination and gene conversion [6]. While significant work has been done in cancer classification based on microarray gene expression data [3], [8], [5], [2], [9], cancer classification based on differences in LOH patterns has not been previously investigated. Effective machine learning models for cancer classification based on LOH data would be very useful because they can not only assist in clinical cancer diagnosis, but also lead to discovery of novel tumor suppressor genes that are specific to a certain type of tumor.

In the problem of cancer classification using LOH data, we still encounter the typical curse-of-dimensionality problem as in cancer classification based on gene expression data:

- The number of SNPs greatly exceeds the number of tissue samples.
- Most SNP loci do not show LOH, and are not related to the given cancer classification problem.

To overcome this curse-of-dimensionality problem, we can use feature selection techniques to select a small subset of SNPs as features for classification. However, unlike gene expression data where the features (expression levels of genes) have real numeric values, the features in LOH data have categorical values (we will explain the data type in detail in the following text.) Therefore, classical methods for gene selection for gene expression data, such as t-test, cannot be used in this case.

On the other hand, feature selection for features with discrete values is a well-studied problem in the field of text information retrieval [23]. Many methods have been proposed, including Information Gain [16], Gain Ratio [17], Mutual Information, χ^2 -statistic, etc.

In this study, we evaluate the effectiveness of machine learning models on cancer classification using LOH data. We built 16 different machine learning models by combining each of four widely used feature selection algorithms (Information Gain, Gain Ratio, χ^2 -statistic and Relief-F) and each of four widely used classification algorithms (k -NN, C4.5 Decision Tree, Support Vector Machine, and Naive Bayes). We applied these models to a public SNP data set and compared their performance according to the leave-one-out cross-validation (LOOCV) classification accuracy.

II. METHODS

A. Feature Selection Algorithms

Feature selection is an important topic in machine learning. In this study, we applied feature filtering algorithms to find informative SNP loci. Feature filters are techniques to compute statistics over the features that indicate which features are better than others. Features (SNPs) are ranked according to such statistics. In this paper, we use the following four popular feature filtering methods: Information Gain, Gain Ratio, χ^2 -statistic and Relief-F.

1) *Information Gain*: Information Gain (IG) [16] is a well-known and empirically proven method for high-dimensional feature selection. It measures the number of bits of information obtained for class prediction by knowing the value of a feature. Let $\{c_i\}_{i=1}^m$ denote the set of classes. Let

V be the set of possible values for feature f . IG of a feature f is defined to be:

$$IG(f) = -\sum_{i=1}^m P(c_i) \log P(c_i) + \sum_{v \in V} \sum_{i=1}^m P(f=v) P(c_i|f=v) \log P(c_i|f=v)$$

2) *Gain Ratio*: Gain Ratio (GR) is a normalized version of IG. It measures the fraction of the number of bits of information obtained for class prediction by knowing the value of a feature. GR of a feature f is defined to be:

$$GR(f) = IG(f) / -\sum_{i=1}^m P(c_i) \log P(c_i)$$

3) χ^2 -*statistic*: The χ^2 -statistic [14] measures the lack of independence between f and c . It is defined as follows:

$$\chi^2(f) = \sum_{v \in V} \sum_{i=1}^m \frac{(A_i(f=v) - E_i(f=v))^2}{E_i(f=v)}$$

where V is the set of possible values for feature f , $A_i(f=v)$ is the number of instances in class c_i with $f=v$, $E_i(f=v)$ is the expected value of $A_i(f=v)$. $E_i(f=v)$ is computed with

$$E_i(f=v) = P(f=v)P(c_i)N$$

where N is the total number of instances.

4) *Relief-F*: One of the most widely used feature filters is the Relief-F algorithm [11]. 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 find a good estimate of the following probability to assign as the weight for each feature f .

$$w_f = \frac{P(\text{different value of } f | \text{different class})}{-P(\text{different value of } f | \text{same class})}$$

This approach has shown good performance in various domains [19].

B. Classification Algorithms

After selecting the informative SNPs, we applied one of the following classifiers: k -Nearest Neighbor, linear Support Vector Machine, C4.5 Decision Tree and Naive Bayes.

1) *k -Nearest Neighbor*: The k -Nearest Neighbor (k -NN) classifier [7] is a well-known nonparametric classifier. To classify a new input x , the k nearest neighbors are retrieved from the training data. The input x is then labelled with the majority class label corresponding to the k nearest neighbors.

The distance metric we used for the k -NN classifier was the overlap metric [1], which is one of the most popular metric for data points with discrete features. In the overlap metric, the distance between different values of a feature is 1, and 0 if the values are the same. This metric basically counts the number of mismatching feature values in both data points. k is an important parameter for the k -NN classifier. In this study, the best k between 1 and 10 was found by performing LOOCV on the training data.

2) *Support Vector Machine*: The Support Vector Machine (SVM) belongs to a new generation of learning system based on recent advances in statistical learning theory [20]. A linear SVM, which is used in our system, aims to find the separating hyperplane with the largest margin, defined as the sum of the 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 the classifier. In a non-separable case, a linear SVM seeks a trade-off between maximizing the margin and minimizing the number of errors.

3) *C4.5 Decision Tree*: C4.5 [18] is a well-known decision tree based classifier. A decision tree is a tree structure where non-leaf nodes represent tests on one or more features and leaf nodes reflect classification outcomes. An instance is classified by starting at the root node of the decision tree, testing the attribute specified by this node, then moving down the tree branch corresponding to the value of the attribute. This process is then repeated at the nodes on this branch until a leaf node is reached.

In our experiments, we used the J4.8 algorithm in Weka [22], which is a Java implementation of C4.5 Revision 8.

4) *Naive Bayes*: The Naive Bayes (NB) classifier [16] is a probabilistic algorithm based on Bayes' rule and the simplifying assumption that the 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.

III. RESULTS

In this section, we present experiment results on a public data set. Details about the data, preprocessing, experimental parameters, and results are provided in sections below.

A. Data

1) *Data Source*: In this study, we used the SNP array data set published in [24] by Zhao *et al.* It can be downloaded at the following URL: <http://research.dfci.harvard.edu/meyersonlab/snp/snp.htm>.

The original data set contains raw data (CEL files) obtained from 43 tissue samples using Affymetrix XbaI mapping 130 array, which covers 10,043 SNP loci along all of the human chromosomes except the Y chromosome.

For our cancer classification study, we used a subset of the original data consisting of data from 10 breast cancer patients and 5 small cell lung cancer (SCLC) patients.

2) *Data Processing*: We processed the raw data following the same steps as described in [12]. First, we re-analyzed the whole original raw data set using dChipSNP [13] to produce SNP genotype calls. Then we performed LOH analysis on the whole original data set using dChipSNP to produce LOH calls based on the SNP genotype calls of paired normal and tumor samples of the same individual.

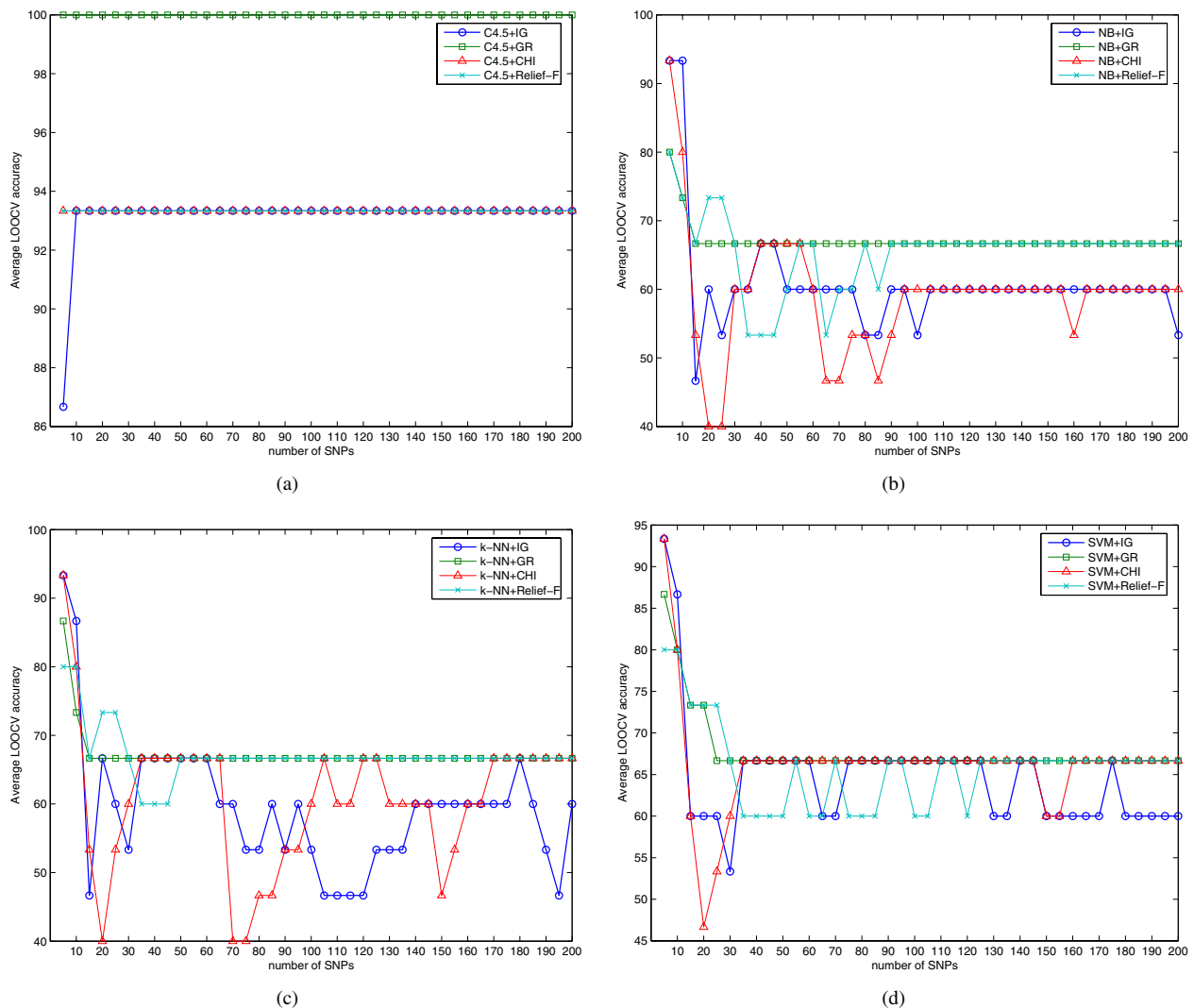


Fig. 1. Comparison of the LOOCV accuracy of (a) C4.5 (b) NB (c) k -NN (d) SVM using different feature selection algorithms.

3) *Data Types of LOH Calls:* For each SNP, possible assigned SNP genotype calls include A or B (homozygous for one allele), AB (heterozygous), and “No Call” (if signal is too poor to deduce a call). The possible LOH calls made by dChipSNP for each SNP are: Loss, Retention, Non-informative, and “No Call.” The relationship between the SNP calls of paired tissue samples and the corresponding LOH calls are described in [13]. Clearly, if we consider the LOH calls for SNPs as features, the features have categorical values.

B. Experimental Settings

We consider the performance of the 16 machine learning models built from all possible combinations of the four feature selection algorithms and the four classifiers discussed above. We implemented these models using Perl and the WEKA 3.4.3 [22], which is an open source collection of machine learning algorithms in Java.

In each fold of the LOOCV test, the LOH calls of 14 tissue pairs were used as training data, and the LOH calls

of the one tissue pair left was used as test data. To avoid the selection bias [4], the feature selection algorithms were applied to the training data only, without any 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 of the 15 folds were averaged.

C. Results

Figures 1 shows the LOOCV classification accuracy using different combinations of feature selection algorithms and classification algorithms. In each case, the top 5, 10, 15, ..., 200 SNPs were selected. We can observe from the results that:

- The best LOOCV classification accuracy of 100% was achieved by C4.5 combined with GR.
- Except for C4.5, all other classifiers achieved the best performance when 5 or 10 SNPs were selected.
- GR is the best among the four feature selection algorithms in this application.
- C4.5 is the best among the four classification algorithms

in this application. It is also robust to the addition of more selected SNPs. In the range of 10 to 200, the number of selected SNPs doesn't seem to affect its performance.

IV. DISCUSSION AND CONCLUSION

This study represents preliminary results on the application of machine learning models in cancer classification using genome-wide LOH data derived from SNP arrays. Using a public data set, we found that the best LOOCV classification accuracy of 100% was achieved by C4.5 classification tree with the Gain Ratio feature filtering algorithm. Experimental results suggest that LOH data as derived from SNP arrays are an excellent indicator of cancer type.

Because the 10k SNP array is relatively dense, nearby SNPs along the chromosomes often have the same LOH or retention status. The feature selection algorithms examined here do not take into account such dependence information. Therefore, the selected informative SNPs may be highly dependent. We could eliminate the "redundant" SNPs by applying our HykGene method [21]. However, since we are interested in the chromosome regions (possibly containing multiple SNPs) that are implicated in cancers, instead of a minimal number of SNPs for classification purpose only, we did not apply additional filtering steps to the selected SNPs.

Because the number of samples in the data set is small, the LOOCV classification accuracy of 100% may be an overestimate. The results should be interpreted as a proof of concept. Experiments on larger data sets are needed when such data sets are available.

The informative SNPs selected by the feature selection algorithms may lead to the discovery of new tumor suppressor genes that are specific to a certain type of tumor. Although the selected top-ranked informative SNPs can result in good LOOCV classification accuracy, their LOH properties still need to be confirmed by quantitative real-time PCR of the selected loci. The surveying of additional cancer specimens will also help to address their significance.

REFERENCES

- [1] D. W. Aha, D. Kibler, and M. K. Albert. Instance-based learning algorithms. *Machine Learning*, 6(1):37–66, 1991.
- [2] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson, J., L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503–511, 2000.
- [3] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *PNAS*, 96(12):6745–6750, 1999.
- [4] C. Ambrose and G. J. McLachlan. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc Natl Acad Sci U S A*, 99(10):6562–6566, 2002.
- [5] S. A. Armstrong, J. E. Staunton, L. B. Silverman, R. Pieters, M. L. den Boer, M. D. Minden, S. E. Sallan, E. S. Lander, T. R. Golub, and S. J. Korsmeyer. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat Genet*, 30(1):41–47, 2002.
- [6] W. K. Cavenee, T. P. Dryja, R. A. Phillips, W. F. Benedict, R. Godbout, B. L. Gallie, A. L. Murphree, L. C. Strong, and R. L. White. Expression of recessive alleles by chromosomal mechanisms in retinoblastoma. *Nature*, 305(5937):779–84, 1983. 0028-0836 Journal Article.
- [7] B. Dasarthy. *Nearest Neighbor Norms: NN Pattern Classification Techniques*. IEEE Computer Society Press, 1991.
- [8] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, 1999.
- [9] G. J. Gordon, R. V. Jensen, L. L. Hsiao, S. R. Gullans, J. E. Blumenstock, S. Ramaswamy, W. G. Richards, D. J. Sugarbaker, and R. Bueno. Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer Res*, 62(17):4963–4967, 2002.
- [10] J. Knudson, A. G. Mutation and cancer: statistical study of retinoblastoma. *Proc Natl Acad Sci U S A*, 68(4):820–3, 1971. 0027-8424 Journal Article.
- [11] I. Kononenko. Estimating attributes: analysis and extensions of relief. In *Proceedings of the European conference on machine learning on Machine Learning*, pages 171–182. Springer-Verlag New York, Inc., 1994.
- [12] M. E. Lieberfarb, M. Lin, M. Lechpammer, C. Li, D. M. Tanenbaum, P. G. Febbo, R. L. Wright, J. Shim, P. W. Kantoff, M. Loda, M. Meyerson, and W. R. Sellers. Genome-wide loss of heterozygosity analysis from laser capture microdissected prostate cancer using single nucleotide polymorphic allele (snp) arrays and a novel bioinformatics platform dchipsnp. *Cancer Research*, 63(16):4781–4785, 2003. 0008-5472 Journal Article.
- [13] M. Lin, L. J. Wei, W. R. Sellers, M. Lieberfarb, W. H. Wong, and C. Li. dchipsnp: significance curve and clustering of snp-array-based loss-of-heterozygosity data. *Bioinformatics*, 20(8):1233–1240, 2004. 1367-4803 Evaluation Studies Journal Article.
- [14] H. Liu and R. Setiono. Chi2: feature selection and discretization of numeric attributes. In *Proceedings of the Seventh International Conference on Tools with Artificial Intelligence*, pages 388–391, 1995.
- [15] H. Matsuzaki, H. Loi, S. Dong, Y. Y. Tsai, J. Fang, J. Law, X. Di, W. M. Liu, G. Yang, G. Liu, J. Huang, G. C. Kennedy, T. B. Ryder, G. A. Marcus, P. S. Walsh, M. D. Shriver, J. M. Puck, K. W. Jones, and R. Mei. Parallel genotyping of over 10,000 snps using a one-primer assay on a high-density oligonucleotide array. *Genome Res*, 14(3):414–25, 2004. 1088-9051 Journal Article.
- [16] T. M. Mitchell. *Machine Learning*. McGraw-Hill, New York, 1997.
- [17] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.
- [18] J. R. Quinlan. *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., 1993.
- [19] M. Robnik-Sikonja and I. Kononenko. Theoretical and empirical analysis of relief and rrelieff. *Mach. Learn.*, 53(1-2):23–69, 2003.
- [20] V. N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.
- [21] Y. Wang, F. S. Makedon, J. C. Ford, and J. Pearlman. HykGene: a hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data. *Bioinformatics*, 21(8):1530–1537, 2005.
- [22] I. H. Witten and E. Frank. *Data mining : practical machine learning tools and techniques with Java implementations*. Morgan Kaufmann, San Francisco, Calif., 1999.
- [23] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In D. H. Fisher, editor, *Proceedings of ICML-97, 14th International Conference on Machine Learning*, pages 412–420, Nashville, US, 1997. Morgan Kaufmann Publishers, San Francisco, US.
- [24] X. Zhao, C. Li, J. G. Paez, K. Chin, P. A. Janne, T. H. Chen, L. Girard, J. Minna, D. Christiani, C. Leo, J. W. Gray, W. R. Sellers, and M. Meyerson. An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. *Cancer Research*, 64(9):3060–3071, 2004. 0008-5472 Journal Article.