Schwender, Holger; Ickstadt, Katja

Working Paper

Identification of SNP interactions using logic regression

Technical Report, No. 2006,31

Provided in Cooperation with:
Collaborative Research Center 'Reduction of Complexity in Multivariate Data Structures' (SFB 475), University of Dortmund

Suggested Citation: Schwender, Holger; Ickstadt, Katja (2006) : Identification of SNP interactions using logic regression, Technical Report, No. 2006,31, Universität Dortmund, Sonderforschungsbereich 475 - Komplexitätsreduktion in Multivariaten Datenstrukturen, Dortmund

This Version is available at:
https://hdl.handle.net/10419/22675

Identification of SNP Interactions Using Logic Regression

Holger Schwender and Katja Ickstadt

Collaborative Research Center 475

Department of Statistics

University of Dortmund

holger.schwender@udo.edu

Abstract

Interactions of single nucleotide polymorphisms (SNPs) are assumed to be responsible for complex diseases such as sporadic breast cancer. Important goals of studies concerned with such genetic data are thus to identify combinations of SNPs that lead to a higher risk of developing a disease and to measure the importance of these interactions.

There are many approaches based on classification methods such as CART and Random Forests that allow measuring the importance of single variables. But with none of these methods the importance of combinations of variables can be quantified directly.

In this paper, we show how logic regression can be employed to identify SNP interactions explanatory for the disease status in a case-control study and propose two measures for quantifying the importance of these interactions for classification. These approaches are

then applied, on the one hand, to simulated data sets, and on the other hand, to the SNP data of the GENICA study, a study dedicated to the identification of genetic and gene-environment interactions associated with sporadic breast cancer.

Keywords: Single Nucleotide Polymorphism, Feature Selection, Variable Importance Measure, GENICA

1 Introduction

Even though humans share far more than 99% of their DNA, there are still millions of differences between the DNA of two individuals. The most common – and so far the best investigated – type of such variations are single nucleotide polymorphisms (SNPs). A SNP occurs when a single nucleotide is altered, i.e. when (usually two) different sequence alternatives exist at a single base pair position. To distinguish a SNP from a mutation, the less frequent variant has to occur in at least 1% of the population. Since the human genome is diploid, i.e. consists of two pairs of chromosomes, each SNP is explained by two bases. Therefore, each SNP can take three realizations:

- *Homozygous reference genotype*: Both bases explaining the SNP are the more frequent variant.
- *Heterozygous variant genotype*: One of the bases is the more frequent and the other is the less frequent variant.
- *Homozygous variant genotype*: Both bases are the less frequent variant.

SNPs are assumed to alter the risk for developing a particular disease. It is, however, very unlikely that individual SNPs play an important role

in the development of complex diseases such as sporadic breast cancer. Instead, high-order interactions of SNPs are supposed to explain the differences between low and high risk groups (Garte, 2001).

In an association study concerned with SNP data, it is thus of interest to construct classification rules of the following type:

"If SNP A is of the heterozygous variant genotype *AND* SNP B is of the homozygous variant genotype *OR* both SNP C *AND* D are *NOT* of the homozygous reference genotype,
then a person has a higher risk to develop the disease of interest."

A procedure developed for solving exactly this type of problems is logic regression (Ruczinski et al., 2003) which attempts to identify Boolean combinations of binary variables for the prediction of, e.g., the case-control status of an observation.

Other classification methods such as CART (Breiman et al., 1984), Bagging (Breiman, 1996), Random Forests (Breiman, 2001) and Support Vector Machines (Vapnik, 1995) can also be applied to SNP data (Schwender et al., 2003). But in comparisons with, on the one hand, CART and Random Forests (Ruczinski et al., 2004), and on the other hand, with other regression procedures (Kooperberg et al., 2001, Witte and Fijal, 2001), logic regression has shown a good performance when applied to SNP data.

Another goal when analyzing SNP data is to quantify the importance of the identified SNPs and SNP interactions for classification. Many classification methods provide approaches to measure the importances of variables. Examples are the variable importance measures of Random Forests or CART, and the squared weights used in RFE-SVM (Guyon et al., 2002) for recursive feature elimination with support vector machines. These methods, however,

do not allow to compute the importance of interactions of variables directly unless these interactions are included as variables into the procedure. This, however, is impractical since analyzing only 50 variables would lead to more than 250,000 input variables if we were interested in interactions up to four-way interactions.

Thus, methods are needed that just use the variables themselves as inputs into the model but enable us to identify combinations of variables and to quantify the importance of these interactions. In this paper, we propose approaches based on logic regression that exactly fulfill these needs.

The paper is organized as follows: In Section 2, a brief introduction to Boolean algebra and logic regression is given. While we describe a method based on logic regression and bootstrapping for identifying potentially interesting interactions in Section 3, two measures for quantifying the importance of the interactions are proposed in Section 4. In Section 5, these approaches are applied to simulated data sets and to the SNP data of the GENICA study, a study dedicated to the identification of genetic and gene-environment interactions associated with sporadic breast cancer.

2 Logic Regression

Logic regression is an adaptive regression methodology for predicting the outcome in classification and regression problems based on Boolean combinations of variables such as

S_1 : "SNP S is not of the homozygous reference genotype,"

or

S_2 : "SNP S is of the homozygous variant genotype,"

i.e. of binary variables that are either true or false. These variables can be negated by the operator ^{C} (e.g., S_2^C means "SNP S is NOT of the homozygous variant genotype.") and combined to a logic expression by the operators $\wedge$ (AND) and $\vee$ (OR).

In logic regression, these logic expression are represented by logic trees (for an example, see Figure 2.1). Logic trees, however, cannot only be employed as nice graphical representations of logic expressions, but also to generate new logic trees in the search for the best model. Permissible moves in this tree-growing process are alternating an operator or a variable, respectively, pruning or growing a branch, and adding or removing variables (for details, see Ruczinski et al., 2003).

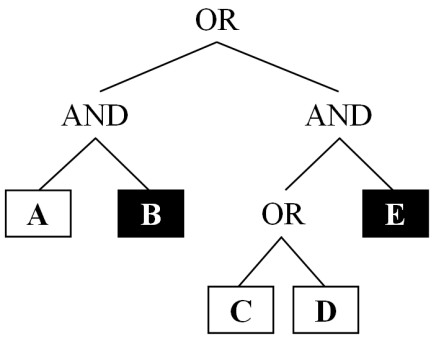

FIGURE 2.1. Logic tree representing the logic expression $L = (A \wedge B^C) \vee ((C \vee D) \wedge E^C)$.

In a case-control study, e.g., logic regression searches for the logic expression L that best explains the cases. If L is true for a new observation, this observation will be classified as case.

Logic regression, however, not only allows to grow a single tree but also provides the possibility to adaptively construct several logic expressions $L_i, i = 1, \ldots, p$, and to combine them by a generalized linear model

$$g\left(\mathrm{E}(Y)\right) = \beta_0 + \sum_{i=1}^{p} \beta_i L_i$$

with response Y, parameters $\beta_i, i = 0, \ldots, p$, and link function g. Since our interest centers on case-control studies we assume g to be the logit function.

Even though the logic expression displayed in Figure 2.1 is still relatively easy to interpret, it becomes more complicated to interpret such expressions the more variables they contain. Therefore, we propose to convert each logic expression into a disjunctive normal form (DNF), i.e. an OR-combination of AND-combinations. The DNF of the logic tree $L = A \wedge B^C \vee (C \vee D) \wedge E^C$ displayed in Figure 2.1 is, e.g., given by

$$L = (A \wedge B^C) \vee (C \wedge E^C) \vee (D \wedge E^C).$$

The advantage of the DNF is that interactions are directly identifiable since they are given by the AND-combinations. The above logic expression, e.g., consists of the three interactions $A \wedge B^C$, $C \wedge E^C$ and $D \wedge E^C$ and is true if at least one of these conjunctions is true.

To avoid redundancy the DNF should only consist of prime implicants, i.e. minimal AND-combinations. If, e.g., $A \wedge B \wedge C$ and $A \wedge B \wedge C^C$ are part of the DNF, then C will be redundant and only the prime implicant $A \wedge B$ is needed.

Our goal is to identify all interactions that might have an influence on the risk of developing a disease. Therefore, we are not interested in obtaining a minimal DNF, i.e. a DNF consisting of a minimum number of prime implicants, but a DNF containing all prime implicants. Schwender (2006) presents a fast algorithm based on matrix algebra for generating such a DNF of a logic expression.

3 Identification of Interesting Interactions

One of the search algorithms used in logic regression is based on the Markov Chain Monte Carlo approach. Kooperberg and Ruczinski (2005) run this algorithm on the whole data set not to find a single best logic regression model but to obtain a large collection of models that fit almost as well as the best one. This set is then used to identify combinations of variables occurring frequently in these models, and these interactions are assumed to be the most important ones.

Contrary to Kooperberg and Ruczinski (2005), we propose a subset-based approach in which the default search algorithm of logic regression, i.e. simulated annealing, is applied to different subsets of the data. More precisely, we suggest the following procedure called *logicFS* for the identification of potentially interesting variables and interactions that might be explanatory for the case-control status of an observation.

Algorithm 1 (logicFS – Identification of Interesting Interactions)

1. Draw a bootstrap sample of size n from the n observations of the data set of interest.
2. Construct a logic regression model based on the bootstrap sample.
3. Convert each of the logic expressions into a disjunctive normal form consisting of prime implicants.
4. Repeat steps 1.-3. B times.

Some of the interactions identified by logicFS are very important for the prediction. Others are not important at all, or might actually be obstructive

for a good classification. It is, hence, necessary to quantify the importance of each of these potentially interesting interactions.

4 Measuring the Importance of Identified Interactions

For a first impression of which variables or interactions might be important or not, the proportion of models generated by logicFS that contain a specific interaction can be computed for each identified interaction. This is similar to the approach used by Kooperberg and Ruczinski (2005) to quantify the importance of the variables and combinations of variables.

It is, however, assumed that some of the SNP interactions are explanatory for only a small subset of patients. Such interactions will hardly be found, and it is likely that they appear only in very few of the models. They would thus be called unimportant by the above measure even though they are actually very important for the correct prediction of some of the patients. Moreover, a suitable measure should quantify how much a particular interaction improves the classification. This improvement should not be computed on the same data set on which the classification rule has been trained but on an independent data set containing new observations.

Since in logicFS a logic regression model is constructed based on a subset of the data the out-of-bag (oob) observations, i.e. the observations not contained in the bootstrap sample, can be employed to estimate the importance of the interactions.

As mentioned in Section 2, there exists both a single and a multiple trees approach of logic regression. While logicFS can handle either of these methods, different importance measures are employed for the two approaches.

In the single tree case, the importance of a prime implicant, i.e. a variable or an interaction, P for classification is computed by

$$\text{VIM}_{\text{Single}} = \frac{1}{B}\left(\sum_{b:\, P\in L_b}\left(N_b - N_b^-\right) + \sum_{b:\, P\notin L_b}\left(N_b^+ - N_b\right)\right), \tag{4.1}$$

where

L_b is the set of prime implicants identified in the b^{th} iteration of logicFS, $b = 1, \ldots, B$,

N_b is the number of oob observations in the b^{th} iteration that are correctly classified by the logic regression model constructed in the b^{th} iteration,

N_b^-/N_b^+ is the number of oob observations correctly classified by the b^{th} model after P has been removed from / added to the model.

We thus compare how well the logic regression models perform when P is part of the logic expressions or not to get a measurement of the influence of P on the correct classification.

In the multiple tree case, it is not possible to unambiguously add an interaction to one of the logic trees since it is not clear to which of the logic expressions it should be appended. The prime implicant P is, therefore, only removed from (and not added to) the models, and the multiple tree measure is determined by

(a) calculating the number N_b of correctly classified oob observations for each of the B iterations,

(b) removing P from all models,

(c) recalculating the number of correctly classified oob observations – now denoted by $N_b^\star$ – for each of the B iterations,

(d) and computing

$$\begin{aligned} \text{VIM}_{\text{Multiple}} &= \frac{1}{B} \sum_{b=1}^{B} (N_b - N_b^\star) \\ &= \frac{1}{B} \sum_{b:\, P \in L_b} (N_b - N_b^\star)\,. \end{aligned} \tag{4.2}$$

The multiple tree measure is similar to the variable importance measure of Random Forests. The only difference is that Breiman (2001) does not remove the variable from the CART trees but permutes the outcome of the variable once and computes $N_b^\star$ based on the permuted outcomes.

For a particular interaction, a large value of both (4.1) and (4.2) corresponds to a high importance of this interaction, whereas a value of about zero leads to the assumption that the interaction has no importance for classification. A prime implicant showing a negative importance is obstructive for a good classification since the number of misclassifications will increase if this interaction is added to the the model.

5 Application to SNP Data

In this section, we apply logicFS and the two variable importance measures (4.1) and (4.2) to simulated and real SNP data. Since the input variables of logic regression and hence of logicFS have to be binary, each SNP $S_i, i = 1, \ldots, m$, is split into the two variables

> S_{i1} : "At least one of the bases explaining S_i is of the homozygous variant genotype,"

and

S_{i2} : "Both bases are of the homozygous variant genotype."

These dummy variables are used instead of the SNPs themselves, where S_{i1} codes for a dominant variation, and S_{i2} for a recessive effect.

5.1 Simulated SNP Data

To investigate if our procedures are able to identify the influential interactions in case-control studies, we employ two simulations: In the first simulation, we are particularly interested in the stability of the approaches, i.e. whether logicFS always identifies the interactions intended to be influential, and whether the importance of an interaction provided by either (4.1) or (4.2) is always about the same, when the approaches are applied to the same data set. The goal of the second simulation is to determine if our procedure can cope with real association studies in which single interactions might have moderate effects and a high percentage of the cases cannot be classified by the measured SNPs.

To examine the former issue, data of 1,000 observations (500 cases and 500 controls) and 50 SNPs are simulated, where an observation is classified as a case if one of the following four logic expressions is true:

- S_{12} (explaining 100 cases),
- $S_{21}^C \wedge S_{32}$ (150),
- $S_{42} \wedge S_{52} \wedge S_{62}$ (100),
- $S_{72} \wedge S_{82}$ (150).

Apart from the SNP values explaining the cases, the values of each of the 50 SNPs are randomly drawn such that the minor allele frequency, i.e. the frequency of the less frequent variant, of each SNP lies between 0.2 and 0.4, and the Hardy-Weinberg equilibrium is fulfilled.

Using 100 Bootstrap samples and allowing a maximum of 20 variables in each of the logic expression models, logicFS is applied to this data set twice – once with the single tree approach and once with the multiple tree approach allowing three logic trees to grow. Afterwards, the single and the multiple tree measure are computed for each of the interactions in the respective approaches. This procedure is repeated 50 times.

TABLE 5.1. Numbers of variables and interactions appearing in a particular number of iterations when logicFS is applied 50 times to the simulated SNP data set using both the single and the multiple tree approach.

Iterations	1	2	3-10	11-47	48	49	50
Single Tree	72	15	21	17	3	0	9
Multiple Tree	4,649	739	1,022	154	1	4	16

Table 5.1 shows how many of the identified interactions appear in how many of the 50 iterations. In the single tree approach, e.g., 72 interactions appear only once, and 15 in two of 50 iterations. Only 9 of the interactions found in the single tree approach and 16 of the interactions in the multiple tree approach are identified in all 50 iterations. Figure 5.1 displays the median and the 25% and 75% quantiles of the 50 values of the importance measures for each of these 9 respectively 16 interactions.

In the single tree case, only the four interactions explanatory for the cases and the three two-way interactions contained in the explanatory three-way

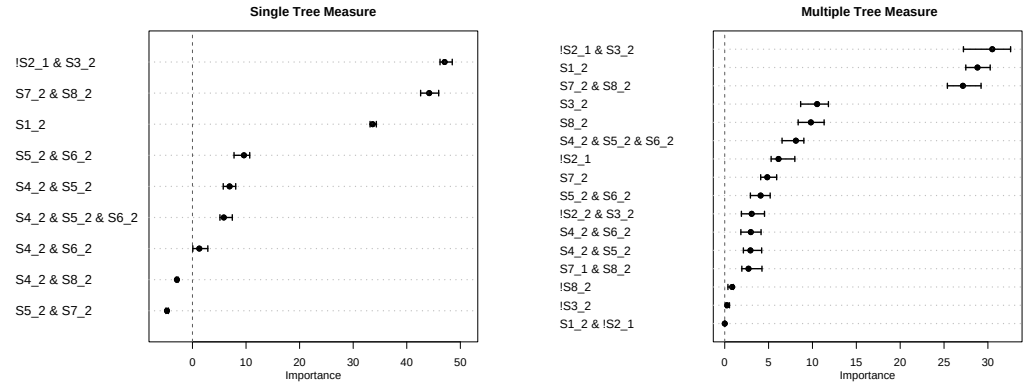

FIGURE 5.1. Single tree (left panel) and multiple tree (right panel) measures of the interactions identified in all 50 iterations of the application of logicFS to the simulated SNP data set. For each of these 9 or 16 interactions, the solid dot represents the median and the bold line the interquartile range of the 50 values of the single or multiple tree measure, respectively. "!" denotes the complement of a variable, and "&" is synonymous to "$\wedge$".

interaction are identified with a positive importance in all iterations. As expected from the fact that typically about 37% of the observations are out-of-bag, the two-way interactions have an importance a little smaller than $0.37 \cdot 150 = 55.5$ and S_{12} has an importance slightly smaller than 37. Figure 5.1 also reveals that the single tree estimates of the importances are very stable since they do not differ much between the 50 iterations.

As in the single tree case, the top 3 logic expressions are the two explanatory two-way interactions and S_{12}. The three-way interaction shows the 6^{th} highest importance and is surrounded by the binary variables belonging to the two two-way interactions. The latter also explains why the importance of the two-way interactions is smaller in the multiple tree case compared to the single tree case: Even though the importances of both the variables themselves and the corresponding two-way interactions are computed sepa-

rately, they are considered jointly in the computation. This might also be a reason for the larger variances of the estimates – compared to the single tree measure.

As a second simulation, SNP data are considered that are more realistic for a genetic association study. Data of 1,000 observations and 50 SNPs are generated, where each SNP exhibits a minor allele frequency of 0.25. The case-control status y of each observation is randomly drawn from a Bernoulli distribution with mean $\text{Prob}(Y = 1)$, where

$$\text{logit}\left(\text{Prob}(Y = 1)\right) = -0.5 + 1.5L_1 + 1.5L_2$$

with $L_1 = S_{61} \wedge S_{71}^C$ and $L_2 = S_{31}^C \wedge S_{91}^C \wedge S_{10,1}^C$.

Thus, the probability of being a case in this association study is 0.378 even if an observation exhibits none of the two interactions intended to be explanatory for the case-control status. A reason for this might be that there are other genetic or environmental factors that have not been surveyed in this study but have an influence on the disease risk.

This procedure is repeated 50 times such that 50 data sets are generated. The mean number of cases and controls over these data sets for the different

TABLE 5.2. Probabilities for being a case when showing none, one or both of the influential interactions, and the mean number of cases and controls over the 50 simulated data sets.

Interactions	Probability	Cases	Controls
0	0.378	388	232
1	0.731	91	245
2	0.924	3	40

probabilities of being a case are summarized in Table 5.2.

Both the single tree approach with a maximum of six variables and the multiple tree approach with two trees and a maximum of eight variables are applied to each of these data sets using $B = 50$ iterations.

Table 5.3 reveals that the two SNP interactions intended to be influential for the disease risk are detected in all of the 50 data sets. Moreover, they are identified as the two most important expressions in almost any of these data sets, where $S_{61} \wedge S_{71}^C$ mostly ranks first with a mean importance of 18.88 in the single and 15.19 in the multiple tree approach, and $S_{31}^C \wedge S_{91}^C \wedge S_{10,1}^C$ ranks second with a mean importance of 12.21 or 6.44, respectively. If one of these interactions ranks third (or lower), then the expressions identified to be more important typically contain this or the other influential interaction plus another variable.

TABLE 5.3. Ranks of the two SNP interactions intended to be influential for the case-control status in the applications of both the single and the multiple tree approach to each of the 50 simulated data sets.

	$S_{61} \wedge S_{71}^C$		$S_{31}^C \wedge S_{91}^C \wedge S_{10,1}^C$	
Rank	Single	Multiple	Single	Multiple
1	45	42	5	6
2	4	6	42	32
3	1	2	3	10
4	0	0	0	2

5.2 The GENICA Study

The GENICA study (http://www.genica.de) is carried out by the Interdisciplinary Study Group on *G*ene *EN*vironment *I*nteraction and Breast *CA*ncer in Germany, a joint initiative of researchers dedicated to the identification of genetic and environmental risk factors associated with sporadic breast cancer. This age-matched and population-based case-control study has been initially launched within the activities of the German Human Genome Project (DHGP) and continues until present.

Even though exogenous risk factors such as reproduction variables, hormone variables and life style factors have also been assessed, we here focus our interest on a subset of the genotype data of the GENICA study. More precisely, data of 1,258 women (609 cases and 649 controls) and 40 SNPs belonging to the DNA repair or the xenobiotic and drug metabolism pathway are available for our analysis.

All observations having more than five missing values as well as SNPs having more than 10% missing values or fewer than 30 women showing not the homozygous reference genotype are removed from the analysis leading to a total of 35 SNPs and 1,191 women (561 cases and 630 controls). The remaining missing values are replaced SNP-wise by random draws from the marginal distribution.

For the application of logicFS to the GENICA SNP data, each of the SNPs is again coded by two dummy variables. Only 59 of these 70 binary variables are used in the analysis since for each of the other 11 variables there are less than ten women for which this variable is true.

Using $B = 200$ iterations logicFS is then applied to this data set twice – once with a single tree and a maximum of 10 variables contained in this tree, and once allowing three trees to grow with a maximum of 16 variables in all

three trees combined.

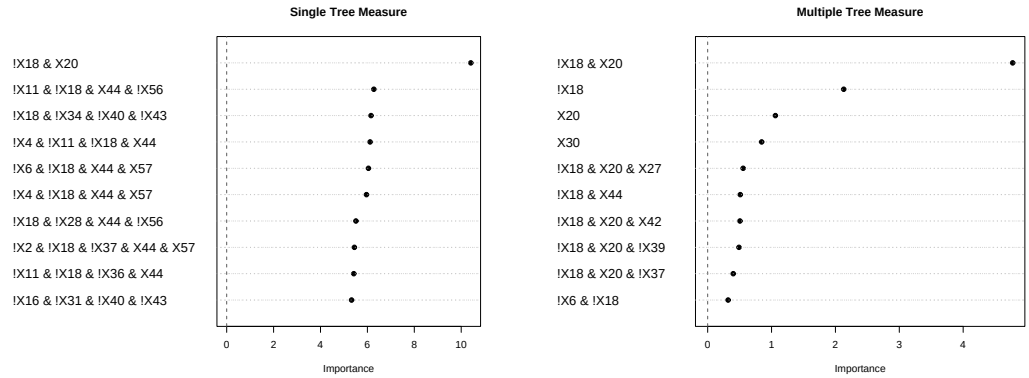

FIGURE 5.2. Single tree (left panel) and multiple tree (right panel) importances of the interactions identified in the analysis of the GENICA SNP data set. Since the SNP names are too long for graphical representation, they are coded.

In the single tree case, this leads to the detection of 1,052 potentially interesting SNPs and SNP interaction, whereas in the multiple tree case 1,589 potentially interesting SNPs and SNP interactions are identified. Figure 5.2, however, reveals that just one interaction, namely $!X18$ & $X20$ or decoded

$$\text{ERCC2\_6540}_1^C \wedge \text{ERCC2\_18880}_1,$$

consisting of two SNPs from the gene ERCC2 (Excision Repair Cross-Complementing group 2; formerly XPD) seems to be associated with the case-control status. If thus ERCC2_6540 (refSNP ID: rs1799793) is of the homozygous reference genotype and ERCC2_18880 (rs1052559) is not of this genotype, then a women will have a little higher risk of developing breast cancer.

The moderate importances of the other interactions in the single tree case are mostly due to the inclusion of ERCC2_6540_1^C in each of these expressions.

6 Discussion

A common and important task in genetic association studies is the identification of SNPs and SNP interactions associated with a covariate of interest, e.g., a disease. Since SNP interactions are assumed to be more influential than individual SNPs an appropriate method needs to be able to identify such interactions. For a good prediction of the covariate of interest, this method should, in addition, provide a possibility to quantify the importance of interactions.

In this paper, we have introduced a procedure called logicFS based on a combination of bootstrap and logic regression for the identification of potentially interesting logic expressions that, e.g., represent SNP interactions, and two measures for quantifying the importance of these features for classification in case-control studies.

In the applications to simulated SNP data, all logic expressions intended to be explanatory for the case-control status of the observations are identified in any of the repetitions always having the highest importances. In the analysis of the GENICA SNP data set, only one interaction between two SNPs of the ERCC2 gene could be detected that slightly increases the risk of developing breast cancer. This supports the findings of Justenhoven et al. (2004).

Since the goal of a case-control study is the construction of a classification rule based on as few variables as possible, the identification of SNP interactions associated with the case-control status is just the first but a very important step. In a next step, one could, e.g., take the k most important features, or all interactions exceeding a specific importance, and use them as binary variables in logic regression or in any other classification procedure.

The variable importance measures are currently restricted to analyses

of data with a binary outcome. They can, however, be extended to, e.g., QTL (Quantitative Trait Loci) studies in which the covariate of interest is quantitative. In this case, the sums of squares would replace the numbers of correctly classified observations in (4.1) and in (4.2), and the signs of the differences in (4.1) and (4.2) would have to be changed. Since logic regression already comprises linear regression (Ruczinski et al., 2003), logicFS can be used as is to identify interactions associated with the quantitative trait.

In this paper, we have employed simulated annealing for model search since this is the standard search algorithm in logic regression. But our method is not restricted to this algorithm. Neither it is restricted to logic regression. logicFS and – at least – the single tree measure can be applied to any procedure whose output is a logic expression.

Moreover, logicFS itself can be employed as a classification procedure since it can actually be viewed as a bagging (Breiman, 1996) version of logic regression. Using the output of logicFS, the case-control status of a new observation can be predicted by majority voting, i.e. by assigning the observation to the class predicted by the majority of the B logic regression models, or by averaging over the class probabilities. Since logic trees and CART trees are related – each logic tree can be transformed into a CART tree, and vice versa – logic trees might also be instable classifiers. It is, therefore, likely that the bagging version of logic regression might improve the classification.

All the approaches presented in this paper have been implemented in the `R` package `logicFS` that can be downloaded from http://www.bioconductor.org, the web page of the Bioconductor project (Gentleman et al., 2004). This package also contains a version of logicFS enabling to perform bagging on logic regression models.

Acknowledgements

Financial support of the Deutsche Forschungsgemeinschaft (SFB 475, "Reduction of Complexity in Multivariate Data Structures") is gratefully acknowledged. The authors would also like to thank all partners within the GENICA research network for their cooperation.

References

Breiman, L. (1996). Bagging Predictors. *Machine Learning*, **26**, 123–140.

Breiman, L. (2001). Random Forests. *Machine Learning*, **45**, 5–32.

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984). *Classification and Regression Trees*, Wadsworth, Belmont.

Garte, S. (2001). Metabolic Susceptibility Genes as Cancer Risk Factors: Time for a Reassessment? *Cancer Epidemiology, Biomarkers and Prevention*, *10*, 1233–1237.

Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A.J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J.Y.H. and Zhang, J. (2004). Bioconductor: Open Software Development for Computational Biology and Bioinformatics. *Genome Biology*, **5**, R80.

Guyon, I., Weston, J., Barnhill, S. and Vapnik, V. (2002). Gene Selection for Cancer Classification Using Support Vector Machines. *Machine Learning*, **46**, 389–422.

Justenhoven, C., Hamann, U., Pesch, B., Harth, V., Rabstein, S., Baisch, C., Vollmert, C., Illig, T., Ko, Y., Brüning, T. and Brauch, H. (2004). ERCC2 Genotypes and a Corresponding Haplotype are Linked with Breast Cancer Risk in a German Population. *Cancer Epidemiology, Biomarkers and Prevention*, **13**, 2059–2064.

Kooperberg, C. and Ruczinski, I. (2005). Identifying Interacting SNPs Using Monte Carlo Logic Regression. *Genetic Epidemiology*, **28**, 157-170.

Kooperberg, C., Ruczinski, I., LeBlanc, M. and Hsu, L. (2001). Sequence Analysis Using Logic Regression. *Genetic Epidemiology*, **21**, 626–631.

Ruczinski, I., Kooperberg, C. and LeBlanc, M. (2003). Logic Regression. *Journal of Computational and Graphical Statistics*, **12**, 475–511.

Ruczinski, I., Kooperberg, C. and LeBlanc, M. (2004). Exploring Interactions in High-Dimensional Genomic Data: An Overview of Logic Regression, with Applications. *Journal of Multivariate Analysis*, **90**, 178–195.

Schwender, H. (2006). Minimization of Boolean Expressions Using Matrix Algebra. *Technical Report*, SFB 475, Department of Statistics, University of Dortmund, Germany.

Schwender, H., Zucknick, M., Ickstadt, K. and Bolt, H.M. (2004). A Pilot Study on the Application of Statistical Classification Procedures to Molecular Epidemiological Data. *Toxicology Letters*, **151**, 291–299.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer, New York.

Witte, J.S. and Fijal, B. A. (2001). Introduction: Analysis of Sequence Data and Population Structure. *Genetic Epidemiology*, **21**, 600–601.