Does Encoding Matter? A Novel View On The Quantitative Genetic Trait Prediction Problem

Dan He and Laxmi Parida
IBM T.J. Watson Research
Yorktown Heights, NY
{dhe, parida}@us.ibm.com

Abstract—Given a set of biallelic molecular markers, such as SNPs, with genotype values encoded numerically on a collection of plant, animal or human samples, the goal of genetic trait prediction is to predict the quantitative trait values by simultaneously modeling all marker effects. Genetic trait prediction is usually represented as linear regression models which require quantitative encodings for the genotypes. There are lots of work on the prediction algorithms, but none of the existing work investigated the effects of the encodings on the genetic trait prediction problem. In this work, we view the genetic trait prediction problem from a novel angle: we consider only the multiplication model for single marker model and epistasis model. It is known that the interactions of the markers may also contribute to the genetic traits under certain conditions, which is known as Epistasis [6]. The pairwise epistasis between two markers i and j is often modeled as the product of the two genotype values. Therefore, with the traditional representation, the linear regression model with pairwise epistasis effects is modified as the following:

\[ Y = \beta_0 + \sum_{i=1}^{m} \beta_i X_i + \sum_{i,j} \alpha_{i,j} X_i X_j + \epsilon_l \]  

where \( X_i \) is the product of the genotype values of the i-th and j-th genotype and it denotes the interaction of the two genotypes, \( \alpha_{i,j} \) represents the coefficient for the interaction. Thus in this epistasis model, the epistasis effects are considered as augmented genotypes besides the original genotype matrix \( X \). We call this model epistasis model. As multiplication is one of the most popular epistasis models, in this work, we consider only the multiplication model for epistasis.

For genetic trait prediction problem, in Equation 1, the genotypic values \( Y \) are quantitative. Thus the genotypes \( X_i \) needs to be quantitative as well. Researchers generally assign three distinct encodings to the three possible genotype values. A few common sets of encodings for genotypes are \{0, 1, 2\}, where 0 and 2 are for homozygous genotypes and 1 is for heterozygous genotype, and \{-1, 0, 1\}, where -1 and 1 are for homozygous genotypes and 0 is for heterozygous genotype.

We look at the problem from a novel angle: we consider the problem as a problem of multiple regression on categorical data, namely a regression on multiple categorical features. The genotype of each marker has three possible categories: homozygous with major allele, homozygous with minor allele and heterozygous. In order to conduct regression on categorical data, we need to first encode the categorical data. Many encoding methods have been proposed for categorical data, including dummy encoding, ordinal encoding, target-based encoding etc. The traditional coding of \{0, 1, 2\} is indeed the ordinal coding, which assumes that the categories follow a certain order. In this problem setting,
the three categories can be considered as following the order of the number of major or minor alleles.

In this work, we developed two hybrid encoding methods. The hybrid methods conduct target-based encoding for the two homozygous categories first, then encode the heterozygous category either as the mean of the trait for all samples, or the mean of the two homozygous categories. Thus they allow the flexibility of the encodings, where different markers can be encoded differently. We showed that the encoded value for the heterozygous category is always bounded by the two values of the homozygous categories. Therefore the hybrid encoding methods maintain the order of the categories. We also extended these hybrid encoding methods to epistasis model. We showed that our hybrid encoding methods are superior to both ordinal and target-based encodings for both single marker model as well as epistasis model. Due to space limit, we did not include the experimental results for the epistasis model, which will be included in the extended version of the work.

II. ENCODING MECHANISMS AND EVALUATION

Ordinal encoding assumes that the categories follow certain order and then encodes the categories with numerical values such as 0, 1, 2. This is indeed the case for our problem setting where the three categories can be considered as following the order of the number of major or minor alleles. When a combination of categories are considered, each category is encoded independently. For genotype encoding, a traditional way is to encode the three categories the same way across all markers. A different encoding mechanism, target-based encoding encodes each category as the mean of the target variable for that specific category. This encoding method allows each marker to be encoded differently. However, the order of the categories are not maintained. Thus for a combination of categorical variables, the order of the categories of each variable is relatively random.

A. Encoding on Single Marker Model

In this work, in order to address the drawbacks of the ordinal encoding and the target-based encoding while maintaining both of their advantages, we develop two hybrid encoding methods. Assuming 0 stands for the homozygous genotype with major allele, 2 stands for the homozygous genotype with minor allele, 1 stands for the heterozygous genotype, the first hybrid method computes new encodings of genotypes at marker i as the follows:

\[
E(i,0) = \text{Ave}(\text{trait}(i,0)) \quad (3)
\]

\[
E(i,2) = \text{Ave}(\text{trait}(i,2))
\]

\[
E(i,1) = \text{Ave}(\text{trait}(i,\{0,1,2\}))
\]

where \(E(i,0)\) is the new encoding for genotype of value 0 at marker \(i\), \(\text{trait}(i,0)\) is the set of traits for the samples whose genotypes are 0 at marker \(i\), \(\text{Ave}()\) is the function to compute the average value. We call this method Hybrid One.

The second hybrid method computes \(E(i,0)\) and \(E(i,2)\) the same as algorithm hybrid one does. However, instead of the average of the trait, \(E(i,1)\) is computed as the average of \(E(i,0)\) and \(E(i,2)\). We call this method Hybrid Two.

\[
E(i,0) = \text{Ave}(\text{trait}(i,0)) \quad (4)
\]

\[
E(i,2) = \text{Ave}(\text{trait}(i,2))
\]

\[
E(i,1) = \frac{E(i,0) + E(i,2)}{2}
\]

We can see that for both hybrid one and hybrid two, \(E(i,0)\) and \(E(i,2)\) are computed the same as those from target-based encoding. However, target-based encoding computes \(E(i,1)\) as \(\text{Ave}(\text{trait}(i,1))\) which then loses the order of the categories. For both hybrid one and hybrid two, it is guaranteed that \(E(i,1) = \text{Ave}(\text{trait}(i,\{0,1,2\}))\) and \(E(i,1) = \frac{E(i,0) + E(i,2)}{2}\) are in between of \(E(i,0)\) and \(E(i,2)\). Thus the order of categories is maintained. The difference is that \(E(i,1) = \text{Ave}(\text{trait}(i,\{0,1,2\}))\) is closer to \(E(i,0)\) as 0 stands for the heterozygous with major allele, where most of the samples have this genotype, thus the average of the whole trait values is close to \(E(i,0)\). On the contrary, \(E(i,1) = \frac{E(i,0) + E(i,2)}{2}\) requires its value as the mean of \(E(i,0)\) and \(E(i,2)\). From our experiments, the first strategy achieves slightly better performance.

B. Encoding on Epistasis Model

The hybrid encoding strategies can be naturally extended to pairwise epistasis effects or even higher dimensional epistasis effects. As shown in Figure 1, for pairwise epistasis effects, given the traditional encoding \(\{0,1,2\}\), we have 9 possible combinations for markers \(i\) and \(j\), organized in the \(3 \times 3\) grid matrix. Assuming 0 is the traditional encoding for homozygous genotype with major allele, 2 is the traditional encoding for homozygous genotype with minor allele, 1 is traditional encoding for heterozygous genotype, then the cell \((0,0)\) (from now on, for simplicity, we ignore the marker indices \(i, j\) for the cell) is the traditional encoding for a pair of homozygous genotypes, both with major allele, the cell \((2,2)\) is the traditional encoding for a pair of homozygous genotypes, both with minor allele, the cell \((1,2)\) is the traditional encoding for a pair of heterozygous genotype and homozygous genotype with minor allele. The meaning of the other cells can be inferred similarly.

Our goal is to encode each cell using the hybrid approach. We first compute the hybrid encoding for the four corner cells \((0,0), (0,2), (2,0), (2,2)\) as the average of their corresponding trait values, as shown in Figure 1. For example, \(E(i,j,0,0) = \text{Ave}(\text{trait}(i,j,0,0))\), where \(\text{trait}(i,j,0,0)\) is the set of traits for the samples whose traditional genotypes at marker \(i, j\) are 0 and 0, respectively. Then for the cells \((1,0), (0,1), (2,1), (1,2)\), we compute their hybrid encoding by extending the encoding
strategy for single markers. For example, $E(i, j, 1, 0) = Ave(trait(i, j, \{0, 1, 2\}, 0))$, where $trait(i, j, \{0, 1, 2\}, 0)$ is the set of traits for the samples whose traditional genotype at marker $i$ is 0 or 1 or 2, and at marker $j$ is 0, respectively. The intuition is that we consider the encoding for the three cells $(0, 0), (1, 0), (2, 0)$ for the marker pair $i, j$ as fixing the genotypes for marker $j$ as 0. Then the problem is converted to computing the encoding for a single marker $i$, whose genotype can be either 0, or 1, or 2. Similar encoding strategies are also applied on the cells $[(0, 2), (1, 2), (2, 2)]$, $[(0, 0), (0, 1), (0, 2)], [(2, 0), (2, 1), (2, 2)]$ to compute the encodings for cells $(1, 2), (0, 1), (2, 1)$, respectively. Finally for the cell in the center $(1, 1)$, we compute its hybrid encoding as the average of all the traits, namely $E(i, j, 1, 1) = Ave(trait(i, j, \{0, 1, 2\}, \{0, 1, 2\}))$.

The above is a straight-forward extension of Hybrid One for single markers. The extension of Hybrid Two is a similar procedure with the following differences: $E(i, j, 1, 0) = Ave(trait(i, j, E(i, j, 0, 0) + E(i, j, 2, 0), 0))$ and similar encodings for cells $(1, 2), (0, 1), (2, 1), E(i, j, 1, 1) = Ave(trait(i, j, E(i, j, 1, 1) + E(i, j, 1, 2) + E(i, j, 2, 1)))$.

The same hybrid encoding algorithm can be further extended to higher dimensional epistasis effects. In this work, we only focused on the application of the hybrid encoding algorithm on single marker and pairwise epistasis effects.

III. RESULTS

As rrBLUP [8] (Ridge-Regression BLUP, based on a linear regression model) is one of the most commonly used methods for genetic trait prediction, in our experiments, we evaluate the prediction accuracy for different encodings mainly using rrBLUP.

| Table I | SUMMARY OF THE DATA SETS |
|---------|--------------------------|
| Data Set | Num. of Markers | Num. of Samples |
| Maize: Flint (all three datasets) | 29094 | 261 |
| Maize: Dent (all three datasets) | 83027 | 261 |
| Rice (both datasets) | 36991 | 413 |
| Pig (both datasets) | 52842 | 3534 |
| QTLMAS (both datasets) | 10031 | 2326 |

We apply the new encoding strategy to four different data sets, summarized in Table I. We compare the performance of rrBLUP on both the traditional encoding and the two hybrid encodings and the target-based encoding. As $r^2$, the square of the Pearson’s correlation coefficient is the most common evaluation metric for genetic trait prediction problem, we show the average $r^2$ of 10-fold cross validation. Notice all the encodings are generated only from the training data and then applied to the test data accordingly.

The first data set is the Maize data set [9] which consists of two maize diversity panels with 300 Flint and 300 Dent lines developed for the European CornFed program. The two panels, Flint and Dent, were genotyped using a 50k SNP array, which after removing SNPs with high rate of missing markers and high average heterozygosity, yielded 29,094 and 30,027 SNPs respectively. Both of them contain 261 samples.

The second data set is the Asian rice, Oryza sativa, data set [11]. This data set was based on 44,100 SNP variants from 413 accessions of O. sativa, taken from 82 countries containing 34 phenotypes. We selected two phenotypes, one is polygenic (Protein.content), one is oligogenic (Peri-carp.color). The data sets have 36,901 markers and 413 samples.

The third data set is Pig data set, which is a collection data on male and female pigs born since 2000 and was taken from [5] and consists of 3,534 animals from a single PIC nucleus pig line yielding 52,842 SNPs with five measured traits (phenotypes). Only traits 2 and 4 were selected for study here. As described in [5], genotypes were sequenced from the Illumina PorcineSNP60 chip and full pedigree information is available, which we did not use in this study. In the original study, trait 2 was rescaled by a weighted mean of corrected progeny phenotypes. Whereas trait 4 was corrected for environmental factors such as year of birth and location. Genotypes were filtered for minor allele frequency less than 0.001 and with missing genotypes less than 10%. The original study used AlphaImpute to impute any missing data [14].

The fourth data set is QTLMAS data set, which was taken from the QTLMAS Workshop, which was held on May 17-18, 2010 in Poznan Poland [1]. The data set consists of 3,226 individuals over five generations (F0-F4) with 20 founders, five male and 15 females. There were two phenotype traits, the first a quantitative trait and the second a binary trait. Only the first four generations (2,326 individuals) have phenotype records. The genome is approximately 500 million bp with five chromosomes, each 100 million bp. In total, each individual was genotypes for 10,031 biallelic SNPs.

For genetic prediction, to our knowledge, there is no method can achieve consistently better performance than rrBLUP does with similar running time. Also compared with rrBLUP, even for cases where the performance can be improved, most of the other methods can not make an improvement over 5% [3]. Thus we consider an improvement of 5% as significant. As shown in Table II, in general the hybrid encodings are able to improve the prediction perfor-
mance and in many cases the improvement is significant. The target-based encoding is slightly better than the traditional encoding, but worse than both hybrid encodings. Thus for single marker model, the hybrid encodings are superior to the traditional encoding and the target-based encoding. The two hybrid encodings have similar performance.

We also applied the hybrid encoding strategies on the epistasis model shown in Formula 2. Due to space limit, we did not include the experimental results for the epistasis model, which will be included in the extended version of the work. However, our experiments indicate that the hybrid encoding strategies improved the prediction performance on the epistasis model as well.

IV. Conclusions

In this work, we showed that the quantitative genetic trait prediction problem heavily depends on the encoding of genotypes, for both single marker model and epistasis model. We developed two hybrid encoding methods which are simple but effective. Out experiments show that the hybrid encodings are able to improve the prediction accuracy for both the single marker model and the epistasis model. We also conducted a detailed analysis on the performance of the hybrid encodings. We would also like to investigate the effects of variation of allele frequency between train and test data and the effects of correlation of markers (linkage).

REFERENCES

[1] Matthew A Cleveland, John M Hickey, and Selma Forni. A common dataset for genomic analysis of livestock populations. *G3: Genes—Genomes—Genetics*, 2(4):429–435, 2012.

[2] BJ Hayes, PJ Bowman, AJ Chamberlain, and ME Goddard. Genomic selection in dairy cattle: Progress and challenges. *Journal of Dairy Science*, 92(2):433–443, 2009.

[3] Dan He, Irina Rish, David Haws, and Laxmi Parida. Mint: Mutual information based transductive feature selection for genetic trait prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, DOI 10.1109/TCBB.2015.2448071*, 2015.

[4] Elliot L Heffner, Mark E Sorrells, and Jean-Luc Jannink. Genomic selection for crop improvement. *Crop Science, 49*(1):1–12, 2009.

[5] Jean-Luc Jannink, Aaron J Lorenz, and Hiroyoshi Iwata. Genomic selection in plant breeding: from theory to practice. *Briefings in Functional Genomics, 9*(2):166–177, 2010.

[6] Jeffrey R Kilpatrick. Methods for detecting multi-locus genotype-phenotype association. PhD thesis, RICE UNIVERSITY, 2009.

[7] Russell Lande and Robin Thompson. Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics*, 124(3):743–756, 1990.

[8] T.H.E. Meuwissen, B.J. Hayes, and M.E. Goddard. Prediction of total genetic value using genome-wide dense marker maps. *Genetics, 157*:1819–1829, 2001.

[9] Renaud Rincent, Denis Laloë, Stéphane Nicolas, Thomas Altmann, Dominique Brunel, Pedro Revilla, Victor M Rodriguez, J Moreno-Gonzalez, Thomas Altmann, Dominique Brunel, Pedro Revilla, Victor M Rodriguez, J Moreno-Gonzalez, A Melchinger, Eva Bauer, et al. Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: Comparison of methods in two diverse groups of maize inbreds (zea mays l). *Genetics, 192*(2):715–728, 2012.

[10] Yunbi Xu and Jonathan H Crouch. Marker-assisted selection in plant breeding: from publications to practice. *Crop Science, 48*(2):391–407, 2008.

[11] Keyan Zhao, Chih-Wei Tung, Georgia C Eizenga, Mark H Wright, M Liakat Ali, Adam H Price, Gareth J Norton, M Rafiqul Islam, Andy Reynolds, Jason Mezey, et al. Genome-wide association mapping reveals a rich genetic architecture of complex traits in oryza sativa. *Nature communications, 2*:467, 2011.