www.roavs.com EISSN: 2223-0343
RESEARCH OPINIONS IN ANIMAL & VETERINARY SCIENCES

Research article

Genotype imputation using support vector machine in parent-offspring trios

Abbas Mikhchi[1], Mahmood Honarvar[2], Nasser Emam Jomeh Kashan[1*], Saeed Zerehdaran[3] and Mehdi Aminafshar[1]

[1]Department of Animal Science, Science and Research Branch, Islamic Azad University, Tehran, Iran; [2]Department of Animal Science, Shahr-e-Qods Branch, Islamic Azad University, Tehran, Iran
[3]Department of Animal Science, Ferdowsi University of Mashhad, Mashhad, Iran

Article history
Received: 14 Nov, 2015
Revised: 27 Nov, 2015
Accepted: 29 Nov, 2015

Abstract

An important problem in genomic selection in livestock is the cost of genotyping. Genotype imputation is a process of predicting unknown genotypes or un-typed Single nucleotide polymorphism (SNP), which uses reference population to predict missing genotypes for animal genetic variations. Support vector machines are algorithms based machine learning methods. We compared the Support Vector Machines (SVMs) and Beagle software for Genotype imputation in parent-offspring trios in term of imputation accuracy and computation of time. The methods employed uses simulated data (1000 trios with 10k SNPs) to impute the missing SNPs in parent-offspring trios. The genome consists of 5 chromosomes and each chromosome was set as 100 CM length. For simulated dataset five versions: NA10, NA30, NA50, NA70 and NA 90, were created (10, 30, 50, 70 and 90 percent of offspring genotypes are missing). Our results show that in all versions of simulated dataset Beagle outperformed SVM in term of imputation accuracy and computation of time. The Beagle requires almost no tuning and can easily handle missing predictor genotypes. We conclude to use of SVM in larger Sample size (*i.e* 10000) for imputation of parent-offspring trios.

Keywords: Genotype imputation; trios; support vector machine; machine learning methods

To cite this article: Mikhchi A, M Honarvar, NEJ Kashan, S Zerehdaran and M Aminafshar, 2015. Genotype imputation using support vector machine in parent-offspring trios. Res. Opin. Anim. Vet. Sci., 5(10): 416-419.

Introduction

Genotype imputation is a tool for increasing the performance of genomic selection. Imputation can be used for detection of additional associations, particularly when combining data from multiple studies genotypes on different platforms. Genotype imputation can be used for identifying an association between typed and un-typed SNP. In contrast, haplotype-based association testing is not limited to testing known genetic variants, but the interpretation of haplotype-based association analysis is typically more difficult. Imputation can be used for predicting genotypes at SNP that have not been genotyped. This is possible by using patterns of haplotype variation seen in another data set (the reference panel) that includes the largest set of markers (Browning and Browning, 2009). Imputation methods can be helpful to impute un-typed genotypes from cheaper SNP panels to higher density and follow decrease genotyping costs (Jafarikia, 2014).

Genotype imputation has been employed in the human genome for association study with gene expression and biomedical research. Genotype imputation can be used in inbred lines for identifying

***Corresponding author:** Nasser Emam Jomeh Kashan, Department of Animal Science, Science and Research Branch, Islamic Azad University, Tehran, Iran; E-mail: Nasser_ejk@yahoo.com

candidate genes for complex disease (Gualdrón Duarte et al., 2013). Imputation methods can be used for family (using linkage disequilibrium) and population methods (Sargolzaei et al., 2014). The "trio" included of genotypes from parent-offspring triplets and many phasing algorithms are adapted to be used in trio datasets (Lu and Cantor, 2014). There are many software programs for imputation, which are fast-PHASE (Scheet and Stephens, 2006), MACH (Willer et al., 2008), Beagle (Browning and Browning, 2007) and PLINK (Purcell et al., 2007). Some methods perform imputation using machine learning methods. Machine learning methods have been used in genomic studies to explore the predicting missing values of a marker (Goddard et al., 2013). Supervised machine learning methods can be used in building a model of learning such genetic patterns from a labelled set of training samples that will also provide accurate genetic predictions in new sample with similar genetic background (Ogutu et al., 2011). Support Vector Machines (SVMs) have been shown to have unique powers and the ability to create a binary classification based on multiple features (Goddard et al., 2013). The aim of the current study was to predict accuracy and running time of imputation at parent-offspring trio using support vector machine.

Materials and Methods

Data simulation

The genome structure could be clearly defined by the overall parameters and mutation rules applied in each current population. Generally, the number of chromosomes and the lengths of different chromosomes are assigned, e.g. 1 Morgan for each of five chromosomes. Data sets with 5 chromosomes, 100 cm each, were simulated to allow predicting of missing SNP in offspring using support vector machine, in term of accuracy of imputation. The population with effective size of 100 animals was simulated, of which 50% of the animals were female and the other half was male. This structure was kept fix for 50 generations using a mutation rate of 2.5 x 10^{-8} per site by selecting the parents of an animal randomly picked from animals of the last generation. Mating was performed by randomly selecting the parents of an animal randomly from the animals of the previous generation. Bi-allelic SNPs were defined on each of homologous chromosomes and used "0" and "1" to denote the two alleles at each SNP site. The allele with high frequency was defined as '0' and allele with low frequency as '1' and an unknown value as 'NaN'. The number of markers was 2000 for each chromosome. Finally, the population structure consisted of a dataset (Geno) with 1000 trios and each trios contained dam, sire and offspring. For this dataset five versions (NA10, NA30, NA50, NA70 and NA 90) were created (10, 30, 50, 70 and 90 percent of offspring genotypes is missing). All the imputations in this study were done using R software (Technow, 2015).

Imputation accuracy and running time

The imputation accuracy per SNPs in offspring were calculated as the correlation between imputed and observed SNPs, then mean of imputation accuracy were calculated across the 5 replicates. Computation time was measured based on running each program in the second on a PC with 7 core CPU Intel, 7 and a total of 8 GB RAM by R software.

Support Vector Machine (SVM)

Support vector machine is a supervised machine learning method which builds models based on 'training' and 'test' data. In supervised learning the goal is mostly to make a formula to classify samples in predefined process. The formula is created by learning from samples. The correct class assignments are known for the learning data and used in the making of the formula. The number of samples needed to make a classification formula is data set and classification method related. The training set is mostly a subset of all samples complete with all class and feature values and the resultant model is then applied to the extant test data (Goddard et al., 2013).

The aim of the SVM is to discriminate between two groups using a set of variables. It is particularly useful when the number of variables is greater than the number of individuals in the data set. SVM is based on a model with N ordered pairs where is a binary outcome with a vertex −1 assigned to one group and +1 to the other and, is a vector with M predictors (Lu and Cantor, 2014). The nature of SVM is a kernel function K (x_i, x) between any two SNP x and xi, that is the measure of similarity between two SNPs. Given a set of m SNPs (x_1,…, x_m), SVM estimates a scoring function for any new SNP x of the form using the following equation

$$f(x) = \sum_{i=1}^{m} \alpha i k(xi, x) + C \quad (1)$$

Where, "α_i" is the weights in the expression to be optimized by the SVM by maximizing the large positive scores for the SNP in the training set. The "C" is called complexity parameter that needs to be optimized for predication performance. It also controls possible over fitting of the training set. The Polynomial kernel function has two additional parameters, d (degree of freedom or order of the polynomial) and slope of alpha was used, the function of polynomial kernel was as follows:

$$k(x, y) = (\alpha x^{\dagger} T\ y + c)^{\dagger} d \quad (2)$$

The parameters of SVM are optimized by using a grid search before processing to the SVM classifier. The

R package of 'e1071' was used in this study (Chang and Lin, 2015).

Beagle software

Beagle 3.3.2 (Browning and Browning, 2009) programs were used to allow the reader to compare the change with SVM. Performances of beagle in terms of imputation accuracy and computational efficiency were also compared. Beagle uses a hidden Markov model (HMM) to infer haplotype phase with both typed and un-typed SNPs, and perform the association test with the inferred haplotypes. Beagle uses the localized haplotype cluster model to cluster haplotypes at each SNP marker and then determine an HMM to find the probable haplotype pairs based on the individual known genotypes (Weng et al., 2013).

Results

Imputation using SVM and Beagle was evaluated on the same dataset and using the same PC computer. The imputation accuracies in Geno datasets are shown in Table 1 for SVM and Beagle. For the simulated dataset, imputation accuracies ever decreased as the level of missing SNP increased. The imputation accuracies for SVM were high. On all of the level of Geno dataset has never been observed 100% imputation accuracy by SVM and Beagle. The imputation accuracy was lower for NA90 and NA70 versions compared to others. In general, for SVM and Beagle, the imputation accuracy decreased with the increase in the ratio of un-typed SNPs. Beagle performed foremost and yielded accuracies greater than 90% of all cases. When the ratio of un-type of SNPs was low, for example, NA70 or NA90, Beagle performed better compared with SVM. For each version of this dataset we made 5 replicates, and it is demonstrated that all imputation methods yield very small standard error in most conditions, showing that these methods perform robustly in genotype imputation.

The detailed runtime of both SVM and Beagle methods on Geno datasets at missing rate of 90% (NA90) is presented in Table 2. The number of SNP markers probably influences computing time. For Geno dataset, the Beagle was fastest algorithm. SVM required much more computing time than the Beagle, and its computing time increased when the number of un-typed SNP increased from 70% to 90%.

Table 1: Mean of imputation accuracy for each imputation method in various versions on the Geno dataset

Beagle	SVM	Version	Sample size	Density	Data set
0.9919(0.0011)	0.96032(0.012)	NA10	1000	10k	
0.9900(0.002)	0.9613(0.035)	NA30	1000	10k	
0.9899(0.002)	0.9326(0.044)	NA50	1000	10k	Geno
0.9844(0.003)	0.9381(0.036)	NA70	1000	10k	
0.9838(0.003)	0.9203(0.034)	NA90	1000	10k	
0.9879(0.0038)	0.9445(0.0206)	Mean	****	*****	

SVM=Support Vector Machine, NA10: 10% of genotype is missing per offspring, NA30: 30% of genotype is missing per offspring, NA50: 50% of genotype is missing per offspring, NA70: 70% of genotype is missing per offspring, NA90: 90% of genotype is missing per offspring.

Table 2: Average imputation runtime on Geno dataset, in seconds

Data Set	Sample size	Density	Version	SVM	Beagle
Geno	1000	10K	NA90	23421s	11830s

SVM=Support Vector Machine, NA10: 10% of genotype is missing per offspring, NA30: 30% of genotype is missing per offspring, NA50: 50% of genotype is missing per offspring, NA70: 70% of genotype is missing per offspring, NA90: 90% of genotype is missing per offspring

Discussion

We applied the performance of the SVM method using simulated data set for genotype imputation in parent-offspring trios. The abilities of the SVM method were compared with Beagle in terms of imputation accuracy and running time. There are various reports on the performance of the Beagle that show that imputation accuracy in beagle was slightly less than other methods (Pei et al., 2008; Li et al., 2009; Marchini and Howie, 2010). Weng et al. (2013) reported that Beagle was most robust in almost all conditions. Our results indicate that Beagle outperformed SVM and had the highest accuracy. A possible reason that Beagle was more accurate than SVM is that the HMM is better to adhere to the whole gamut of other algorithms, and the improvement in accuracy from pedigree information is small when LD is highly adequate (Li et al., 2009; Wang et al., 2013). Beagle also uses the haplotype clustering-based algorithm. Our result showed that the imputation accuracy for SVM was slightly less than beagle. It can be due to kernel function. SVM uses kernel function. In this study, we applied a Polynomial Kernel Function for SVM. There are some kernel functions for SVM: Radial Basis Function, Polynomial Kernel Function, linear and sigmoid (Gillani et al., 2014). Therefore, we concluded to use the other kernel function for imputation.

On the other hand, the datasets that we applied this research may have violated multivariate normality. In terms of computation time Wang et al. (2013) reported that in comparing with finhap, festivals and Beaglerun much slower, as they are both localized haplotype clustering-based HMM methods, and a large amount of time is required for Monte Carlo Markov Chains (MCMC) alterations. Between fastPHASE and Beagle, fastPHASE requires substantially most fast computation; as it takes into account all observed genotypes when imputing each missing genotype or un-typed SNP, whereas Beagle usually focuses on genotypes for a small

number of neighbour SNP when importing each missing genotype, which makes Beagle computationally more impressive (Li et al., 2009; Weng et al., 2013). Our results showed that Beagle was fastened and SVM needs more computation time for impute the Geno dataset. Because Beagle produces posterior genotype probabilities for imputed genotypes, when the number of un-genotyped SNPs is increased, the amount of calculation, including sampling haplotypes and producing posterior genotype probabilities, is correspondingly increased (Wang et al., 2013).

Conclusion

We compared SVM method with the Beagle software in performing genotype imputation based parent-offspring trios. Our comparisons comprised imputation accuracy and computation time. We found that Beagle outperformed SVM in term of imputation accuracy and computation time. The simulation study showed that SVM and Beagle performed very well for imputation of missing SNPs in offspring. The SVM appears to be accurate in the Geno dataset, but we would not recommend it over the HMM with such a small training set size (I. e. 500). The Beagle requires almost no tuning and can easily handle missing predictor genotypes. We recommend using SVM in larger Sample size (i.e. 10000).

Acknowledgements

We would like to express our gratitude to all those who helped to complete this paper, especially to Drs. Y Forghani, M Kamaei, Y Bernal Rubio, whose constructive suggestions and encouragements help us in all the time of this research.

References

Browning BL, Browning SR (2009) A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Amer J Human Gen 84: 210-223.

Browning BL, Browning SR (2007) Efficient multilocus association testing for whole genome association studies using localized haplotype clustering. Genet Epidemiol 31: 365-375.

Chang CC, Lin CC (2015) e1071: Misc. Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.6-7. Available at: http://CRAN.R-project.org/web/packages/e1071.

Gualdrón Duarte JL, Bates RO, Ernst CW, Raney NE, Cantet RJ Steibel JP (2013) Genotype imputation accuracy in a F2 pig population using high density and low density SNP panels. BMC Genetics 14: 38.

Gillani, Z, Akash, MSH, Rahaman, MMD, Chen, M, (2014) Compare SVM: supervised, Support Vector Machine (SVM) inference of gene regularity networks. BMC Bioinformatics 15: 395.

Goddard R, Eccles D, Ennis S, Rafiq S, Tapper W, Fliege J, Collins A (2013) Support vector machine classifier for estrogen receptor positive and negative early-onset breast cancer. PLoS ONE 8: e68606.

Jafarikia M (2014) Genomics Tools for Improving Health and Production Performance of Canadian Pigs, Proceedings of the 10th World Congress of Genetics Applied to Livestock Production.

Li Y, Willer C, Sanna S, Abecasis G (2009) Genotype Imputation. Annu Rev Genomics Hum Genet 10: 387-406.

Lu AT, Cantor RM (2014) Identifying rare-variant associations in parent-child trios using a Gaussian support vector machine. BMC Proceedings 8(Suppl 1): S98.

Marchini J, Howie B (2010) Genotype imputation for genome-wide association studies. Nat Rev Gen 11: 499-511.

Ogutu JO, Piepho HP, Streeck TS (2011) A comparison of random forests, boosting and support vector machines for genomic selection. BMC Proceedings 5(Suppl 3): S11.

Pei YF, Li J, Zhang L, Papasian CJ, Deng HW (2008) Analyses and comparison of accuracy of different genotype imputation methods. PLoS ONE 3: e3551.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559-575.

Sargolzaei M, Jansen GB, Schenkel FS (2014) A new approach for efficient genotype imputation using information from relatives. BMC Genomics 15: 478.

Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet 78: 629-644.

Wang Y, Cai Z, Stothard P, Moore S, Goebel R, Wang L, Lin G (2012) Fast accurate missing SNP genotype local imputation. BMC Res Notes 5: 404.

Willer CJ et al (2008) Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nat Genet 40: 161-169.