SADIO Electronic Journal of Informatics and Operations Research

http://www.dc.uba.ar/sadio/ejs

vol. 7, no. 1, pp. 24 - 30 (2007)

A classification approach for heterotic performance prediction based on molecular marker data

Leonardo Ornella [1] Elizabeth Tapia [1]

[1] Area Comunicaciones
Escuela de Ingeniería Electrónica
Facultad de Ciencias Exactas, Ingeniería y Agrimensura,
Universidad Nacional de Rosario
Riobamba 245 bis
2000 Rosario, Argentina
e-mail: etapia@eie.fceia.unr.edu.ar
No. tel. (+54) 341 4808543 (+ 47)

Abstract

A number of statistical methods based on molecular data are currently available for assigning new inbreds to heterotic groups in maize (*Zea mays L*), with variable results. We conjecture that the main flaw of such models is that they do not capture the non-linear relation between parental data and progeny performance. In this paper, we propose the use of supervised learning methods for handling such non-linearity. Standard and novel multiclassification methods are evaluated. Best results are obtained with the recently introduced class of multiclass, binary based, Recursive ECOC (RECOC) classifiers. RECOC classifiers are inspired in state of art Coding Theory solutions for the problem of transmitting symbols over noisy channels. For molecular marker data the noisy channel abstraction embeds the hardness of learning a classification function from noisy and scarce samples. Field data (top crosses between 26 inbreed lines and four tester populations), processed by cluster analysis in a previous work, was integrated with molecular marker data and used for training RECOC – AdaBoost Support Vector Machines RBF classifiers. A 34.10 % 3-CV error was achieved, clearly improving previously reported results on this task.

Keywords: Machine learning, maize, heterotic group

1 Introduction

Since the first maize hybrid was bred and produced in USA, hybrid breeding has become one of the primary goals in any maize breeding programs (Crow 1998). However, varietal development has become more competitive and costly; in the United States, for example, development of one variety of maize or soybean requires 6-8 years and $0.5-7.0 million dollars. The lifetime of a variety is usually 3-6 years before it succumbs to the challenges of the production environment, demands of consumers, and competition from new varieties. Consequently, breeding programs devote most of their resources to manipulation of a core of elite germplasm known to provide genetic gain and to satisfy objectives for the near term (Lee 1998b). It should be noted, however, that this type of elite breeding programs are often accompanied by disturbing trends e.g., decreased genetic diversity within the elite gene pool and increased genetic uniformity of the crop in production. Such disturbing trends may be exacerbated in the short term as a result of consolidation within the seed industry and deployment of transgenic crop varieties (Lee 1998b). This could result in greater problems, particularly less resistance to new diseases and insects or less tolerance to heat and drought (Salhuana et al. 1998). One way to solve the abovementioned problems is the introgression the exotic germplasm (e.g., unadapted varieties, plant introductions, land races, and undomesticated relatives) into heterotic groups previously established in order to improve the existing gene pool (Lee 1998b). An heterotic group is a collection of germplasm (related or unrelated genotypes) that, when crossed to germplasm external to this group (usually another heterotic group) tends to exhibit a higher degree of heterosis than when crossed to germplasm derived from their own group (Lee 1998a). By heterosis, or hybrid vigor, we mean an increase in size or rate of growth of offspring over parents e.g., hybrid vigor in crop plants can be observed as an increased grain yield (Lee 1998a).

Several non-molecular methods (phenotype, geographic origin, parentage, etc) have been used to avoid highly time consuming and expensive field tests regarding heterotic group assignment. However, these methods suffer from a lack of supportive genetic information and weak discriminatory power (Lee 1998b). As and alternative, variations in DNA sequences have been explored as molecular markers in plants and animals during the last two decades (Ferreira and Grattapaglia, 1998); with the advent of the polymerase chain reaction (PCR), new classes of markers emerged that combined the desired characteristics of being highly polymorphic and cost effective, such as RAPD (random amplified polymorphic DNA), AFLP (Amplified fragment length polymorphism), and more recently, microsatellites or SSR (single sequence repeats) (Ferreira and Grattapaglia, 1998). The latter consist of 2 to 6 base pairs nucleotide motifs repeated 10 to 50 times in tandem. Tandems can be amplified by PCR, provided that the non-repetitive sequences that flank them become known. They are present in various plant species (Kubis et al. 1998) and are useful as molecular markers in maize because they can be found in all chromosomes and contain a high level of polymorphism (Chin et al. 1996). Moreover, they are inexpensive, very easy to handle, require small amounts of template DNA (Ferreira and Grattapaglia, 1998), and can be included with previous results from breeding programs to avoid expensive and time consuming field-tests. Our proposal is to integrate this kind of data with supervised machine learning methods for heterotic group prediction.

1.1 Multiclassifiers and Molecular Marker Data for Heterotic Group Assignation

Decision Trees (Quinlan 1986), AdaBoost Decision Stumps (Schapire and Singer, 1998), Naive Bayes (Domingos and Pazzani, 1996) and OAA (One Against All) (Allwein et al. 2000) multiclassifiers were evaluated with maize microsatellite data regarding heterotic group assignation. In addition, the recently introduced class of ECOC (Error Correcting Output Coding) classifiers based on recursive error correcting codes (Tanner 1981, MacKay 1999), the so called class of RECOC (Recursive ECOC) classifiers, was also evaluated. RECOC classifiers remarkably improved current multiclassification approaches when used with microarray data (Tapia et al. 2005). However, their usefulness in small datasets (like the one we present in this work) remained to be explored. Both OAA and ECOC classifiers divide a multiclassification problem into a number of binary classification tasks. However, while OAA classifiers use a number of binary classifiers equalling the number of classes, ECOC classifiers use a number of binary classifiers equalling the size of an error correcting code. By assuming an error correcting code, ECOC classifiers aim the systematic correction of

binary classifiers' errors and, in this way, the improvement of multiclass decisions for OAA classifiers. For RECOC classifiers, binary base classifiers can be weakly boosted Support Vector Machines with Radial Basis functions kernels (SVM-RBF) (Schölkopf et al. 1997) or strongly boosted Decision Stumps (Schapire and Singer 1998).

For the sake of brevity, we only report experimental results OAA and RECOC results with SVM-RBF binary based classifiers. However, we note that closely approximated OAA and RECOC classification results were observed with strongly boosted Decision Stumps binary classifiers. Regarding the assembling of binary classifiers' decisions, both OAA and RECOC classifiers need the execution of corresponding decoding algorithms. For OAA classifiers, decoding algorithms are limited to loss variants of blockwise Minimum Hamming distance decoding. For RECOC classifiers, powerful bitwise iterative decoding algorithms can be used, allowing the potential solution of hard biological data multiclassification problems.

2 Materials and Methods

The field data analyzed in this study was taken from experiments that have been described in detail in (Nestares 1996). Briefly, our investigation involved 26 lines from a total of 48 evaluated for their combining ability with the testers: sB73 and sMo17 from the Reid x Lancaster pattern and with the flint testers HP3 and P5L2 from the local orange flint pattern during the 1991/92 season. The 48 lines were grouped according their combining ability with the testers populations into for heterotic groups (H1-H4) using fastclus procedure (Nestares 1996). Twenty six of the 48 inbred lines plus 2 testers populations were characterized using 21 evenly distributed in the genome (Morales Yokobori et al. 2002); all lines but one were derived from orange flint (native) populations (Nestares 1996).

A dataset comprising 42 attributes corresponding to 21 SSR (2 alleles per locus) were generated. This dataset, hereafter called Het6, contains 47 instances and 6 classes (C=6) defined by four heterotic groups and 2 tester populations (H1-H6): H1=4, H2=8, H3=6, H4=8, H5=12 and H6=9. In addition, a trimmed dataset comprising 26 instances of four classes of heterotic groups (H1-H4), hereafter called Het4 was also considered. Four standard multiclassifiers, provided by the Java WEKA library (Witten and Eibe 2000), were considered: i) Decision Trees, ii) Naïve Bayes, iii) OAA SVM-RBF with default constant complexity C_{SVM}=1.0 and the gamma parameter γ =0.01 and iv) AdaBoost Decision Stumps with a number of boosting steps T=150. In addition, RECOC classifiers constructed from LDPC codes (Gallager 1963) and based on AdaBoost SVM- RBF binary classifiers with default parameters C_{SVM}=1.0, γ =0.01, were also evaluated. In particular, RECOC LDPC classifiers involving a number binary classifiers n in the interval [k, 1.5* k], k = $\log_2$ C, were considered i.e., for C = 6, n $\in$ [4, 15] and for C = 4, n $\in$ [3, 10]. As in previous work, the performance of RECOC LDPC AdaBoost SVM-RBF classifiers was evaluated at a number of boosting steps T=15 and a number of iterative decoding iterations I =150 (at the classification stage performed by iterative decoding algorithms). Finally, due the absence of a public train-test partition, classification performance was evaluated by 50 runs of Montecarlo 3 Fold Cross Validation (3-CV). We note in passing that RECOC LDPC software was implemented as an extension of the Java WEKA library (Witten and Eibe 2000).

3 Experimental Results

In this section, we present experimental results (see Table 1) on the Het6 data set. The following multiclassifiers were evaluated: i) Naive Bayes, ii) Decision Trees, iii) AdaBoost Decision Stumps (T=150), iv) OAA SVMs with RBF kernels and v) RECOC LDPC AdaBoost (T=15) SVM-RBF. Best results were obtained with RECOC LDPC AdaBoost SVM-RBF multiclassifiers with parameter n=14: 3-CVerror=0.341. Among conventional classifiers, OAA SVM-RBF yielded the best results: 3-CVerror=0.422.

Table 1: 3 Fold CV(50 Montecarlo Runs) error on the Het6 data set involving six classes (C=6) The number of binary classifiers n is shown when applicable. Otherwise a NA label (not available) is shown.

Multiclassifier	**Number of Binary Classifiers**	**3-CV error**
Naïve Bayes	NA	0.434
Decision Trees	NA	0.508
AdaBoost Decision Stumps (**T**=150)	NA	0.639
OAA SVM-RBF (C_{SVM}=1.0, γ=0.01)	6	0.422
RECOC LDPC AdaBoost (**T**=15) SVM-RBF	4	0.393
	5	0.393
	6	0.393
	7	0.421
	8	0.386
	9	0.391
	10	0.362
	11	0.362
	12	0.362
	13	0.351
	14	**0.341**
	15	0.424

4 Discussion

A number of statistical methods based on molecular data are currently available for assigning new inbreds to heterotic groups in maize (*Zea mays L*), with variable results (dos Santos Diaz et al. 2004). In particular with this data, several "traditional" statistical methods were applied: Morales Yokobori et al. (2002) used Jaccard's coefficient to construct a similarity matrix and performed cluster analysis by unweighted pair-group method analysis (UPGMA) of the lines based on this matrix. Although grouping agreed with the pedigree several discrepancies were observed between dendogram and the grouping based on heterotic data. Ornella et al. (2002) applied several genetic distance estimators reported in the bibliography and evaluated their relationship with heterotic data using Pearson' correlation coefficient, with variable results. Ornella et al. (2004) used mixed models theory to predict the perfomance of hybrids derived from these lines under different variance-covariance structures. In this case, Pearson correlation between observed and predicted yield was 0.32 (at a statistical significance level $p<0.01$). We conjecture that the main flaw of mixed models and/or other traditional statistical theory is that they do not capture the non-linear relation between parental molecular data and progeny performance (Toolenar et al. 2004). Alternatively, our experimental results show that such type of non linearity can be easily captured by supervised Machine Learning models i.e., by multiclassifiers.

Microsatellite maize data was tested with four standard multiclassification methods and the recently introduced class of binary based RECOC classifiers (Tapia et al. 2005). Standard multiclassifiers yield a 3-CV error around the 40 % on Het6 while RECOC ones achieved a 3-CV error around the 34% (fig1). On Het4 data standard and RECOC multiclassifiers yield 78% and 60% 3 Fold -error respectively (fig2). These results outperform previous reports; however, and due the scarcity of Het4 and Het6 datasets (number of instances comparable or less than the number of features), feature selection preprocessing is being considered (Ornella et al. 2005). We expect that feature selection will considerably reduce prediction error. We remark that the dataset's size used here is typical in most molecular breeding programs, specially in third world countries (dos Santos Diaz et al. 2004).

Considering that our classification results are preliminary, they show the usefulness of a molecular based, classification approach for solving general heterotic group assignation problems. Furthermore, while in traditional genetic breeding programs the time scale for obtaining an heterotic categorization is in the order of years, in our proposed framework the time scale is in the order of weeks: two weeks for growing an small plant plus a week to obtain molecular data and a couple of days for computational analysis.

5 Conclusions

A molecular based, Machine Learning approach, for heterotic group prediction has been presented. Among all classifiers, RECOC ones showed the best classification performance. Based on previous work, we hypothesize that further application of feature selection methods i.e., the selection of highly discriminant molecular markers, might improve heterotic group assignation. This hypothesis is supported in the observed similarity between classification problems involving microsatellite marker and those involving microarray data (Dettling and Bühlmann 2003). In both cases, missing and noisy features might be present in scarce data samples. This type of classification noise can be properly limited by feature selection methods so that resulting data sets can be safely managed by binary based, Coding Theory inspired multiclassifiers.

Acknowledgment

The authors wish to Graciela Nestares and Guillermo Eyherabide for field data. Molecular data was obtained under the project PICT-08-03153: "Análisis de poblaciones heteróticas de maíz mediante marcadores moleculares", Agencia Nacional de Promoción Científica y Tecnológica, Argentina. Elizabeth Tapia's work

was supported the project PICT 11-15132, Agencia Nacional de Promoción Científica y Tecnológica, Argentina, which also provided a doctoral fellowship to Leonardo Ornella under the project PAV2003-00127.

References

Allwein, E., Schapire R., Singer Y.: Reducing Multiclass to Binary: A Unifying Approach for Margin Classifiers. J. Machine Learning Res. 1 (2000) 113-141

Chin, E.C.L., Senior, M.L., Shu, H., Smith, J.S.C.: Maize simple repetitive DNA sequences: abundance and allelic variation. Genome 39 (1996) 866-873

Crow, J.F.: 90 Years Ago: The Beginning of Hybrid Maize Genetics 148 (1998) 923-928

dos Santos Dias, L.A., de Toledo Picoli, E.A., Barros Rocha1, R., Couto Alfenas, A.: A priori choice of hybrid parents in plants Genet. Mol. Res. 3 (2004) 356-368

Dettling, M., Bühlmann, P.: Boosting for Tumor Classification with Gene Expression Data. Bioinformatics 19 (2003) 1061-1069

Domingos, P., Pazzani, M.: Beyond independence: Conditions for the optimality of the simple Bayesian classifier, in: Proceedings of the 13th International Conference on Machine Learning, Bari, Italy, (1996) 105-112.

Ferreira, M., Grattapaglia, D.: Introducción al uso de Marcadores Moleculares en el análisis genético. Brasilia, BR, EMBRAPA. (1998).

Gallager, R.: Low Density Parity-Check Codes. Ph.D. thesis. Cambridge, Massachusetts, M.I.T. Press, 1963.

Kubis, S., Schmidtj, T., Heslop-Harrison, J.S.: Repetitive DNA Elements as a Major Component of Plant Genomes Annals of Botany 82 (1998) 45-55

Lee, M.: DNA markers for Detecting Genetic Relationships among germplasm for Establishing Heterotic Groups. Maize Trainning Course, CIMMYT, Texcoco, Mexico, 1998.

Lee, M.: Genome projects and gene pools: New germplasm for plant breeding? Proc. Natl. Acad. Sci. USA 95 (1998) 2001-2004

MacKay, D.J.: Good Error Correcting Codes based on Very Sparse Matrices. IEEE Trans. Inf. Theory 45 (1999) 399-431
Morales Yokobori, M., Decker, V., Ornella, L., Nestares, G., Eyherabide, G.: Análisis de poblaciones heteróticas de maíz mediante marcadores moleculares XXXI Congreso Argentino de Genética. La Plata, Argentina (2002)

Nestares, G.: Caracterización de germoplasma elite de maíz (*Zea Mays L.)* por grupos de heterosis. Tesis de Maestría en Mejoramiento Genético Vegetal. UNR-INTA (1996)

Ornella, L., Morales Yokobori L., Decker V., Balzarini, M.: Estimación de distancias genéticas entre poblaciones de maíz utilizando microsatélites. 5to. Congreso Latinoamericano de Sociedades de Estadística (CLATSE V). Buenos Aires, Argentina (2002).

Ornella, L., Eyherabide G., DiRienzo J., Balzarini, M.: Predicción de Aptitud Combinatoria de Maíz (*Zea mays L.*) Utilizando Marcadores Moleculares y Teoría de Modelos Lineales Mixtos. IX reunión científica Grupo Argentino de

Biometría. La Rioja, Argentina (2004)

Ornella, L., Esteban, L., Serra, E., Tapia, E.: A Classification Approach for the Detection of Coding and non Coding Regions in T. cruzi Genomic Sequences. 7º Simposio Argentino de Inteligencia Artificial. Rosario, Argentina (2005)
Quinlan, J. R.: Induction of decision trees. Machine Learning, (1986) 1:81-106

Salhuana, W., Pollak, L.M., Ferrer, M., Paratori, O., Vivo, G.: Breeding Potential of Maize Accessions from Argentina, Chile, USA, and Uruguay. Crop Sci. 38 (1998) 866–872

Schapire, R. E., Singer, Y.: Improved boosting algorithms using confidence-rated predictions. In Proceedings of the Eleventh Annual Conference on Computational Learning Theory, (1998) 80-91

Schölkopf, B., Sung, K., Burges, C., Girosi, F., Niyogi, P., Poggio, T., and Vapnik, V.: Comparing support vector machines with gaussian kernels to radial basis function classifiers. IEEE Trans. Sign. Processing,. AI Memo No. 1599, MIT, Cambridge 45 (1997) 2758-2765

Tanner, M.: A recursive approach to Low Complexity Error Correcting Codes. IEEE Trans. Inf. Theory 27 (1981) 533 - 547

Tapia, E., Serra E., González J.C.: Recursive ECOC for Microarray Data Classification. In Oza, N.C., Polikar, R., Kittler, J., Roli, F. (eds.): Multiple Classifier Systems. Lecture Notes in Computer Science, Vol. 3541. Springer-Verlag, Heidelberg (2005) 108-117

Tollenar, M., Ahmadzadeh, A., Lee, E.A.: Physiological Basis of Heterosis for Grain Yield in Maize. Crop Sci. 44 (2004) 2086 – 2094

Witten, I. H., Eibe, F.: "Data Mining: Practical machine learning tools with Java implementations". Morgan Kaufmann, San Francisco (2000)