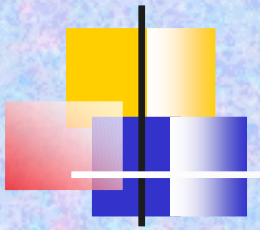


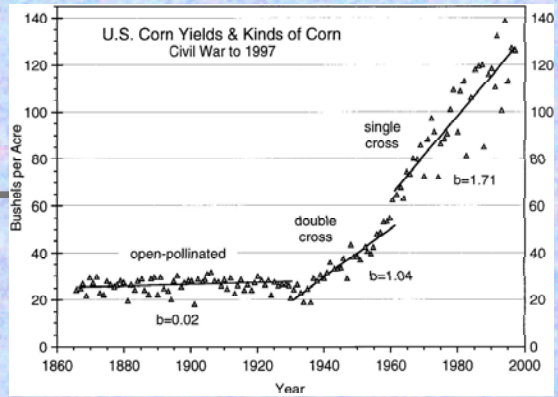
A classification approach for heterotic performance prediction based on molecular marker data

Leonardo Ornella and Elizabeth Tapia

Area de Comunicaciones, Facultad de Ingeniería, Ciencias Exactas y Agrimensura, UNR

etapia@eie.fceia.unr.edu.ar

Introducción



Crow, J. Genetics **148:** 923–928 (March, 1998)

Mejoramiento Genético Moderno de Maíz: (*Zea mays L*) 1900-

basado en el fenómeno de heterosis o vigor híbrido: incremento en el tamaño o tasa de crecimiento de la progenie respecto a los padres.

Metodología tradicional: agrupar las posibles líneas parentales en grupos heteróticos.

Grupo heterótico:colección de germoplasma que presenta mayor grado de heterosis cuando es cruzado con material de otro grupo heterótico a cuando es cruzado dentro del mismo grupo

Problemas

Las Estrategias utilizadas en mejoramiento (selección a partir de pocas líneas parentales)
frecuentemente generan una disminución de la diversidad genética en el acervo genético del germoplasma base)

Consecuencias en los nuevos materiales:

Menor resistencia a calor o sequía

Disminución de la resistencia a nuevas enfermedades y/o insectos

Mal de Río Cuarto: perdidas por u$120 millones(1997)

Solución

Introgresión de germplasma exótico en las poblaciones de mejoramiento : variedades no adaptadas, razas nativas,parientes no domesticados, etc.

Metodología Tradicional : Asignar los nuevos materiales a los grupos heteróticos previamente establecidos mediante cruzamientos de prueba

Problemas:el numero y costo de los ensayos de campo restringen la cantidad de posibles candidatos a evaluar

Alternativas:

Asignar las nuevas líneas evitando los ensayos de campo.

Métodos no moleculares: origen geográfico, fenotipo, relación de parentesco (bajo poder discriminativo)

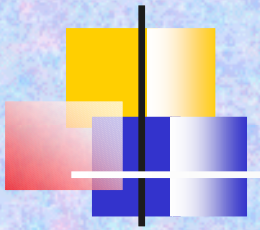
Métodos moleculares: Isoenzimas, RAPDs RFLPs, microsatélites, etc.

Ventajas: no dependen del ambiente, bajo costo,se pueden obtener en poco tiempo.

Problema:

los modelos estadísticos tradicionales no capturan completamente la relación entre genotipo y fenotipo (epistasis, etc)

Necesidad de nuevas metodologías que permitan analizar los datos

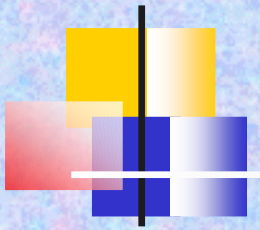


Machine Learning

Dominio de investigación asociado a la inferencia estadística, teoría de información y teoría de optimización

Permite construir sistemas capaces de resolver tareas dado un conjunto de ejemplos con distribución desconocida

Aprendizaje supervisado: seleccionar una función que permita relacionar atributos e hipótesis



Clasificación Multiclase

Objetivo: descomponer el problema de clasificación múltiple de orden M en *n* problemas de clasificación binaria

n *longitud de la palabra código*

One against All

clasificador

Clase	1	2	...	m
1	1	0	0	0
2	0	1	0	0
...	0	0	0	0
...	0	0	1	0
m	0	0	0	1

Código corrector de Error

clasificador

Clase	1	2	...	m	...	n
1	1	0	1	0	0	1
2	0	1	1	0	0	0
...	1	1	0	0	0	1
...	0	0	1	0	1	0
m	0	1	0	1	0	1

Decodificación

Distancia Mínima de Hamming

Decodificación iterativa

Materiales y Métodos

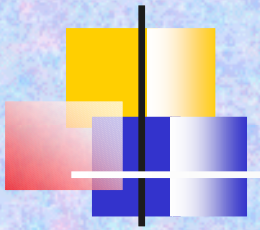
Cruzamientos de prueba entre 26 líneas derivadas de germoplasma argentino y poblaciones testers. Los cruzamientos permitieron agrupar a las líneas en 4 grupos heteróticos.

Datos moleculares: analisis mediante microsatélites de las líneas y dos de las poblaciones tester.

2 Datasets

Het6 :42 atributos: 21 loci de microsatelites (2 atributos por locus), 47 instancias, 6 clases (H1-H6): H1=4, H2=8, H3=6, H4=8, H5=12 y H6=9.

Het4 (solo líneas): 26 instancias y 4 clases (H1-H4)



Clasificadores

Naive Bayes

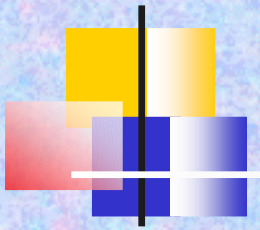
Arboles de decisión (Decision tree)

AdaBoost Decision Stump, 150 boosting steps

OneAgainstAll -Maquinas de vectores soporte con función de base radial (SVM-RBF)
constante de complejidad $C=1.0$, parámetro $\gamma = 0.01$

Clasificadores RECOC-LDPC codes :
+ SVM-RBF $C=1.0$, $\gamma = 0.01$

Entorno Weka : Java Machine Learning



Evaluación

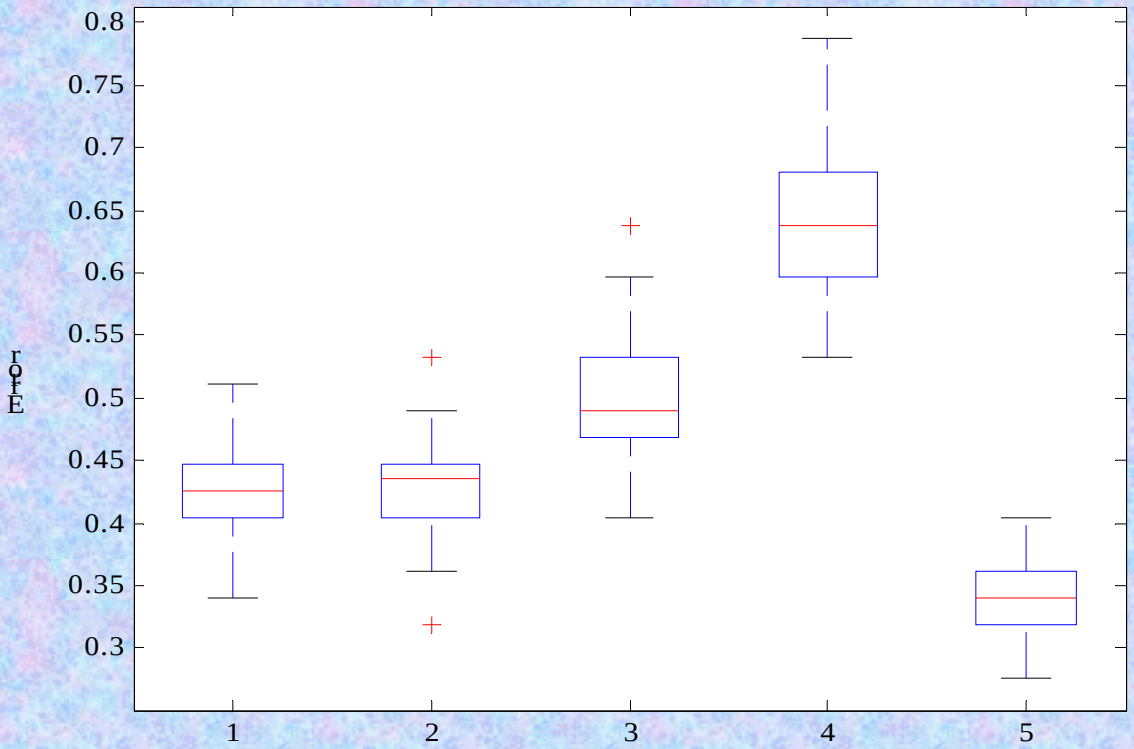
50 corridas Montecarlo 3 Fold Cross Validation

RESULTADOS

Multiclassifier	Number of Binary Classifiers	3 Fold CV
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

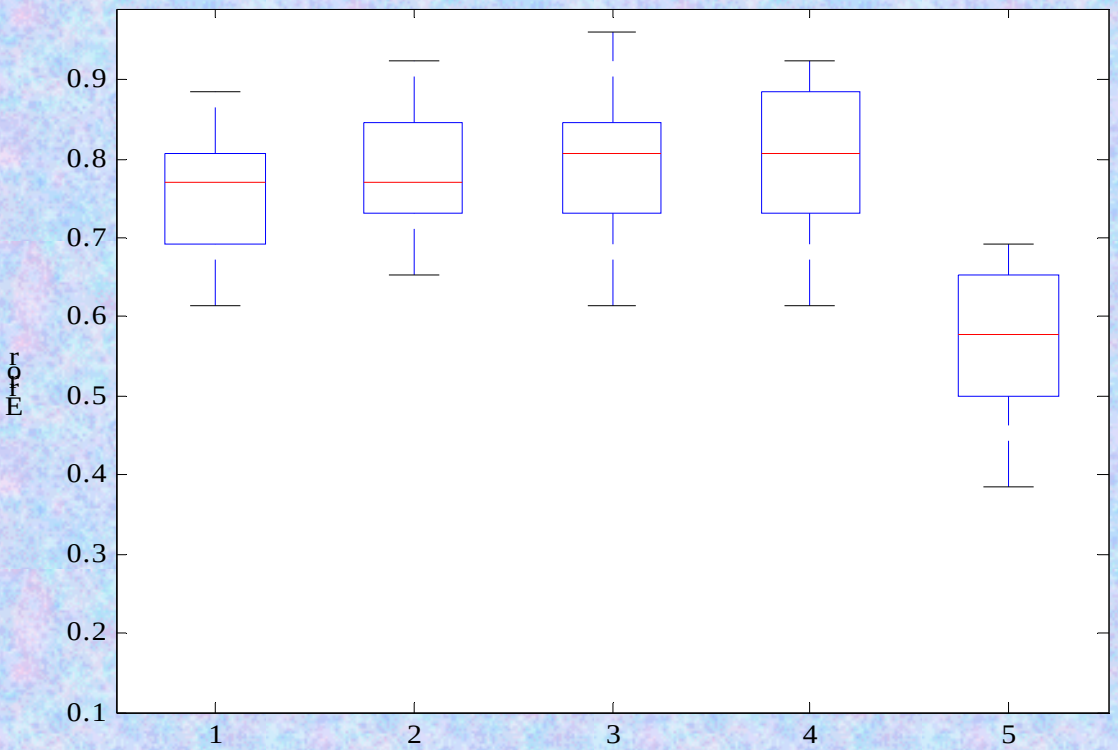
Error 3CV 50 corridas montecarlo - conjunto Het6 (6 clases)

Figura 1

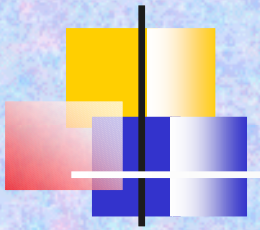


Boxplot -Het6: Error 3-CV, 50 corridas Montecarlo
1)SVM-OAA(C=1.0,γ=0.01), 2)Naive Bayes, 3) Decision Tree
4)AdaBoost-Decision Stump, 5)RECOC-LPDC

Figura 2



Boxplot -Het4: Error 3-CV, 50 corridas Montecarlo
1)SVM-OAA(C=1.0,γ=0.01), 2=Naive Bayes
3) Decision Tree 4)AdaBoost-Decision Stump 5)RECOC-LPDC



Conclusión

En ambos datasets los clasificadores basados en códigos correctores de error presentaron mejor comportamiento respecto de los clasificadores estandares:

Het6:
Multiclasificadores estandard a ~40 % Error 3Fold CV
RECOC-LDPC: ~34 % Error 3Fold CV

Het4:
Multiclasificadores estandard a ~78 % Error 3Fold CV
RECOC-LDPC: ~60 % Error 3Fold CV

Trabajo futuro

FILTRADO DE ATRIBUTOS:

Eliminar los datos moleculares menos informativos

Filtrado S2N (signal to noise)

Ejemplo para dos clases:

$$S2N(x_j) = \frac{\mu(x_j, Cod) - \mu(x_j, NoCod)}{\sigma(x_j, Cod) + \sigma(x_j, NoCod)}$$

Se selecciona la fracción Q de atributos con mayor S2N

Multiclase: media de todos los pares de clases posibles

Otras alternativas disponibles en weka

Agradecimientos:

Graciela Nestares and Guillermo Eyherabide por los datos de campo.

Agencia Nacional de Promoción Científica y Tecnológica.

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

Leonardo Ornella and Elizabeth Tapia

Muchas Gracias...