Diseño Balanceado de Clasificadores para Estudios de Asociación poligenética

Marcel Brun[†*], Virginia Ballarín[†], Inti Anabela Pagnuco[±]

[†]Dto. de Electrónica. Facultad de Ingeniería. UNMdP , Bs As, Argentina
[±]FI-UNER, Universidad de Entre Rios, Argentina
[*]Contacto: Marcel Brun - mbrun@fi.mdp.edu.ar

Resumen

Un área de investigación relativamente reciente en el área de procesamiento de señales genómicas es el de la clasificación de fenotipos basado en información genotípica, en especial SNPs (single nucleotide polymorfisms), usando combinaciones de SNPs para predecir una característica fenotípica. En su mayor parte los datos para un estudio son recolectados sin considerar la probabilidad a priori de las clases a clasificar. Por ejemplo en el contexto de casos vs. controles, donde los casos pueden representar individuos expresando alguna enfermedad y los controles representan individuos sanos, es común que haya más muestras de control que de casos. En este trabajo analizamos las razones para utilizar las técnicas de balance de datos para los estudios de predicción de fenotipos basado en genotipos, mostrando como estas técnicas pueden mejorar los resultados, generando clasificadores más robustos.

Palabras Clave
SNP, Clasificación, Genotipo, Fenotipo, Balanceo

Introducción

Los polimorfismos de un solo nucleótido, o SNPs (del inglés "Single nucleotide polymorphisms"), son una variación de una sola base en la secuencia del ADN, se están convirtiendo rápidamente en una buena opción para definir marcadores para estudios de asociación [1,2,3]. Los SNPs ubicados dentro de una secuencia codificante pueden modificar o no la respectiva cadena de aminoácidos; si no lo hacen, se denominan SNP sinónimos (o mutación silenciosa), mientras que la alternativa es el SNP no-sinónimo. Los SNP que se encuentran en regiones no codificantes no tendrían efecto sobre la secuencia de las proteínas, pero podrían afectar procesos regulatorios tales como el splicing, la unión de factores de transcripción o la modificación de la secuencia de RNA (Ribonucleic Acid) no codificante. Por estas razones, se considera a los SNPs como una fuente importante de variabilidad fenotípica. Otras fuentes de variación de genotipo, como el caso de los QTL (Quantitative trait locus) pueden estar asociados a variaciones en fenotipos [4,5].

Además de las técnicas clásicas de asociación para estudiar la relación entre genotipos y fenotipos, el uso de las técnicas de reconocimiento de patrones permite determinar genotipos asociados a un fenotipo, por medio de reglas de predicción (clasificadores) y estimaciones de sus tasas de error. En el caso específico de genotipos, los datos finales a procesar son de naturaleza completamente discreta, como por ejemplo el tipo de alelo AA, AB o BB en SNPs, y técnicas clásicas de diseño de clasificadores se aplican [6,7]. En todos estos casos, el diseño de clasificadores está orientado a optimizar el error sobre la distribución empírica, la cual estima la distribución poblacional a través de los datos disponibles.

Un problema existente usualmente en el diseño de clasificadores a través de ejemplos es que los datos disponibles representan, usualmente, las distribuciones condicionales de ambas clases, pero no sus probabilidades a priori. El balanceo de datos es una técnica muy usada en el diseño de clasificadores continuos [8-11], pero poco estudiado en el caso de clasificadores discretos.

Elementos del Trabajo y Metodología

Clasificación Discreta

En el caso de la clasificación de fenotipos basado en genotipos, el espacio de observaciones consiste en vectores de la forma $x = (x_1, x_2, x_3, ..., x_n)$ donde x_i toma valores en un conjunto finito $K = \{0, 1, ..., k\}$. Se supone que los vectores están generados por una variable aleatoria $Z = (X,Y)$ tomando valores en $K^n \times \{0,1\}$, donde X corresponde a las observaciones (vectores x) e Y corresponde a la etiqueta de la clase a la que corresponde la observación. Un clasificador binario discreto es una función $\psi : K^n \rightarrow \{0,1\}$ que asigna a cada vector x una etiqueta 0 o 1 dependiendo de la clase a la que es asignado el vector.

Como ejemplo, en el caso de estudios de asociación de caracteres poligenéticos, o QTLs (del inglés *quantitative trait loci*), que son regiones del ADN asociados a un fenotipo, los estudios de asociación se realizan por medio de mediciones fenotípicas y genotípicas de los mismos individuos. Las mediciones genotípicas se realizan sobre *marcadores*, usualmente SNPs o microsatélites [4,5]. Para el caso de los SNPs, el espacio de valores posibles para x_i es $K = \{0, 1, 2\}$ con una posible asignación siendo AA=0, AB=1, BB=2. El caso de los microsatélites es similar al de SNPs, con $K = \{0, 1, 2\}$, donde los casos posibles son definidos por combinaciones de alelos AA=0, AB=1, BB=2.

La calidad de un clasificador está dado por su probabilidad de cometer errores bajo la distribución P de Z: $error(\psi) = P(\psi(X) \neq Y)$. Otras medidas de calidad son las tasas de falsos positivos (FPR) y falsos negativos (FNR), definidos por $FPR = P(\psi(X) = 1 \mid Y = 0)$ y $FNR = P(\psi(X) = 0 \mid Y = 1)$. El error puede escribirse como una combinación de FPR y FNR de la siguiente manera: $error(\psi) = P_0 \cdot FPR + P_1 \cdot FNR$, donde P_0 y P_1 son las probabilidades marginales de las dos clases: $P_0 = P(Y = 0)$ y $P_1 = P(Y = 1)$.

Las técnicas de diseño estadístico de clasificadores se basan en encontrar un clasificador ψ_{opt} que minimiza el error, de tal forma que $error(\psi_{opt}) \leq error(\psi)$ para todo ψ tal que $\psi: K^n \rightarrow \{0,1\}$. Esta búsqueda se realiza mediante el método de minimización de riesgo empírico, o ERM (del inglés *Empirical Risk Minimization*), que consiste en minimizar el error en un conjunto de datos de ejemplo (o entrenamiento) $(\mathbf{x}_i, y_i)$, $i = 1,...,m$:

$$err\_emp(\psi) = \sum_{i=1}^{m} | y_i - \psi(\mathbf{x}_i) |$$

En el caso discreto, la regla plug-in permite determinar el clasificador óptimo en forma simple, definiendo $\psi_{opt}(\mathbf{x}) = 1$ si $n_1(\mathbf{x}) > n_0(\mathbf{x})$ y $\psi_{opt}(\mathbf{x}) = 0$ si $n_1(\mathbf{x}) \leq n_0(\mathbf{x})$, donde $n_1(\boldsymbol{x})$ y $n_0(\boldsymbol{x})$ son la cantidad de observaciones $(\mathbf{x},y)$ en el conjunto de datos de

entrenamiento de la forma ($\boldsymbol{x}$,1) y ($\boldsymbol{x}$,0), respectivamente. Diferentes técnicas [3,7] se pueden utilizar para extender la decisión a configuraciones (vectores x) no observados en los datos de entrenamiento.

Por ejemplo en el caso de predicción de genotipos a partir de los fenotipos, los datos de *entrenamiento* son pares ($\mathbf{x}_i$,y_i), i=1,...,m, donde cada par está asociado a una muestra, $\mathbf{x}_i$ es un vector definido por los valores de n genotipos observados en la muestra i, y el valor y es 0 o 1 dependiendo del fenotipo observado para la muestra i. Por ejemplo si para una muestra se observan los genotipos AA, AB, AA para 3 marcadores seleccionados, y el fenotipo observado es 0, entonces la observación se puede codificar como $\mathbf{x}_i$=(0,1,0) e y=0.

El problema de desbalance

Un problema existente usualmente en el diseño de clasificadores a través de ejemplos es que los datos disponibles representan, usualmente, las distribuciones condicionales de ambas clases, pero no sus probabilidades a priori. Esto significa que el clasificador va a ser óptimo sobre los datos de entrenamiento, pero puede no ser óptimo con respecto a la distribución P de la población. Adicionalmente, inclusive si los datos de entrenamiento reflejan correctamente la distribución de la población, a veces el clasificador que minimiza el error no es el más interesante, ya que puede hacerlo en función de clasificar todas las muestras como 0 o como 1, lo que genera un clasificador óptimo de cero utilidad práctica. Por lo tanto, podemos ver dos situaciones desventajosas del ERM:

- Los datos de ejemplo reflejan las distribuciones condicionales de las clases P(X=$\mathbf{x}$|Y=y) pero no las probabilidades a priori P_y=P(Y=y).
- El clasificador que minimiza el error lo hace a costas de una tasa FPR o FNR muy alta, inaceptable para la aplicación específica.

Para analizar estos casos realizamos una simulación, donde se define una distribución P para la población, con probabilidades a priori P_1=0.4 y P_0=0.6. A partir de esta población se generaron muestras de tamaño grande, para evitar errores debido al muestreo, variando la cantidad de muestras de cada clase para obtener proporciones que van desde 0%-100%, hasta 100%-0%, con incrementos de 1%, para las clases 0 y 1. De esta forma se obtienen 101 conjuntos de entrenamientos, con probabilidades empíricas para P_1 que varían en 0, 0.01, 0.02, ..., 0.99, y 1.

La figura 1 muestra los resultados de diseñar clasificadores con muestras cuyas proporciones no son las dadas por las probabilidades a priori P_0 y P_1. El eje horizontal muestra la proporción de muestras de la clase 0 usadas en los datos de entrenamiento. En el eje vertical se muestran distintas mediciones sobre los clasificadores diseñados sobre esas muestras. La línea azul (oscura) muestra el error empírico del clasificador. Como podemos ver, si todas las muestras son de clase 1 (extremo izquierdo del gráfico), el clasificador óptimo es el que clasifica todo como 1, y su error en los datos de entrenamiento es cero. Pero al aplicarlo a la población donde 60% de las muestras son de clase 0, este clasificador cometerá un error de 60%. En el otro extremo se encuentra el clasificador diseñado con muestras que son de clase 0 en su totalidad. En este caso el clasificador óptimo es el que clasifica cualquier observación como 0, y su error empírico es 0%, pero su error en la población es del 40%. Podemos ver que cuanto más desbalanceados los datos de

entrenamiento, mayor es la diferencia entre el error empírico (estimado sobre las muestras de entrenamiento) y el error verdadero del clasificador estimado. Esta es la situación que se quiere evitar.

El análisis de la figura 1 nos permite ver también el efecto de los datos desbalanceados en las tasas de falsos positivos y falsos negativos, como líneas punteadas. En cada extremo del gráfico una de las dos medidas crece rápidamente. Esto significa que el clasificador diseñado tendrá un rendimiento muy pobre para una de las dos clases. Por ejemplo, para una proporción de 0.1 de clase 0 en los datos (lado izquierdo del gráfico), el clasificador clasificará la mayor parte de las observaciones como 1, por lo que la tasa de falsos negativos (FNR) va a ser baja (línea punteada violeta), pero la tasa de falsos positivos (FPR) va a ser alta, ya que la mayor parte de las observaciones que deberían pertenecer a la clase cero serán clasificadas como 1 también.

También puede observarse que en el caso donde el muestreo se realiza con la proporción correcta de 60% de clase cero, el error estimado del clasificador es igual a su error real (en la población), pero la tasa de falsos negativos es alta, cercana al 80%. En cambio, cuando se utilizan la misma cantidad de muestras para cada clase (50%), se obtiene un decremento de la tasa de falsos negativos, a cambio de un incremento en los falsos positivos y la tasa de error.

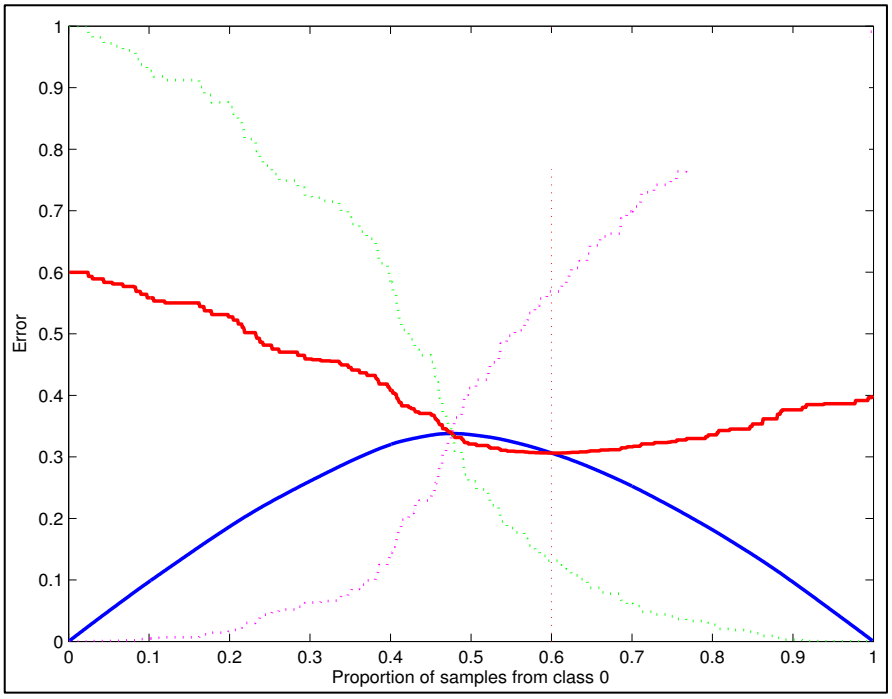

Figura 1: Resultados del error de los clasificadores obtenidos con diferentes proporciones en el muestreo. Las líneas sólidas representan el error verdadero (rojo) y aparente (azul) para el clasificador óptimo. Las líneas punteadas muestras los valores de FPR (verde) y FNR (rojo). La línea vertical muestra la proporción poblacional de verdaderos-falsos.

Balanceo

Una solución clásica al problema del desbalance de datos de entrenamiento consiste en modificar las muestras, con técnicas de sub-muestreo o sobre-muestreo [11]. En estos casos se eliminan muestras de la clase más abundante (sub-muestreo) o se replican muestras de la población menos abundante (sobre-muestreo), para obtener una nueva muestra con cantidades similares para ambas clases. Otras técnicas operan sobre el diseño del clasificador para obtener un diseño donde las tasas de falsos positivos y negativos son similares [9,10]. En muchos casos, modificar el diseño de los clasificadores es similar al re-muestreo de los datos [8].

En este trabajo presentamos un modelo de diseño *balanceado* de clasificadores discretos, que elimina la desproporción de muestras positivas y negativas, evitando el sub-muestreo o sobre-muestreo. La técnica está basada en la regla de decisión óptima, definida previamente por $\psi_{opt}(\mathbf{x}) = 1$ si $n_1(\mathbf{x}) > n_0(\mathbf{x})$ y $\psi_{opt}(\mathbf{x}) = 0$ si $n_1(\mathbf{x}) \leq n_0(\mathbf{x})$, donde $n_1(\boldsymbol{x})$ y $n_0(\boldsymbol{x})$ son la cantidad de observaciones ($\mathbf{x}$,y). Esta regla está originalmente basada en la comparación de las probabilidades condicionales de la distribución empírica.

Sea $P^*(X,Y)$ la *distribución empírica* definida por $P^*(\mathbf{x},y) = P^*(X=\mathbf{x},Y=y) = n_y(\mathbf{x})/m$, donde $n_y(\mathbf{x})$ es la cantidad de observaciones del par ($\mathbf{x}$,y) en los datos de entrenamiento, y *m* es la cantidad total de muestras de entrenamiento. P* es un estimador de la distribución poblacional P(X,Y), y el clasificador óptimo es el clasificador de Bayes sobre esta distribución empírica: $\psi_{opt}(\mathbf{x}) = 1$ si $P^*(\mathbf{x},1) > P^*(\mathbf{x},0)$ y $\psi_{opt}(\mathbf{x}) = 0$ si $P^*(\mathbf{x},1) \leq P^*(\mathbf{x},0)$. Si los datos de entrenamiento no están balanceados, entonces las probabilidades marginales de las clases son diferentes: $P^*(1) \neq P^*(0)$. El objetivo de cualquier técnica de balanceo es obtener una nueva distribución empírica $P^b(X,Y)$ tal que las probabilidades de las clases son iguales (50%), y se mantienen las distribuciones condicionales:

a) $P^b(1) = P^*(0) = 0.5$
b) $P^b(\mathbf{x}|0) = P^*(\mathbf{x}|0)$, $\forall \mathbf{x}$
c) $P^b(\mathbf{x}|1) = P^*(\mathbf{x}|1)$, $\forall \mathbf{x}$

Estas condiciones usualmente no se cumplen cuando se modifica el muestreo, pero si se pueden obtener modificando directamente la distribución empírica, definiendo:

1) $P^b(\mathbf{x},0) = P^*(\mathbf{x},0) \cdot 0.5 / P^*(0)$
2) $P^b(\mathbf{x},1) = P^*(\mathbf{x},1) \cdot 0.5 / P^*(1)$

Puede demostrase que las condiciones (a), (b) y (c) se cumplen para P^b asi definida. Una vez definida esta distribución empírica balanceada, se puede definir el clasificador *balanceado* por $\psi^b_{opt}(\mathbf{x}) = 1$ si $P^b(\mathbf{x},1) > P^b(\mathbf{x},0)$ y $\psi^b_{opt}(\mathbf{x}) = 0$ si $P^b(\mathbf{x},1) \leq P^b(\mathbf{x},0)$.

Resultados

Balanceo

La figura 2 muestra los resultados del diseño del clasificador balanceado sobre los mismos datos sintéticos que los usados para la figura 1. En este caso podemos sacar varias conclusiones:

- Aunque no se percibe en la figura, debido a la gran cantidad de datos de entrenamiento usados, en todos los casos se obtiene el mismo clasificador balanceado, ya que las distribuciones condicionales no cambian, y después del balanceo las probabilidades de las clases son siempre 0.5.
- Aunque el error estimado para clasificadores balanceados sigue siendo diferente al error real, la diferencia es mucho menor para todo el rango de proporciones.
- Para la proporción correcta de muestras (60%-40%), el clasificador balanceado no es el óptimo (respecto a la población), pero su diferencia en error se compensa por los mejores valores (en promedio) de FNR y FPR obtenidos.
- Los valores obtenidos de error, FPR y FNR son independientes de la cantidad de muestras usadas, a diferencia del ejemplo de la figura 1.

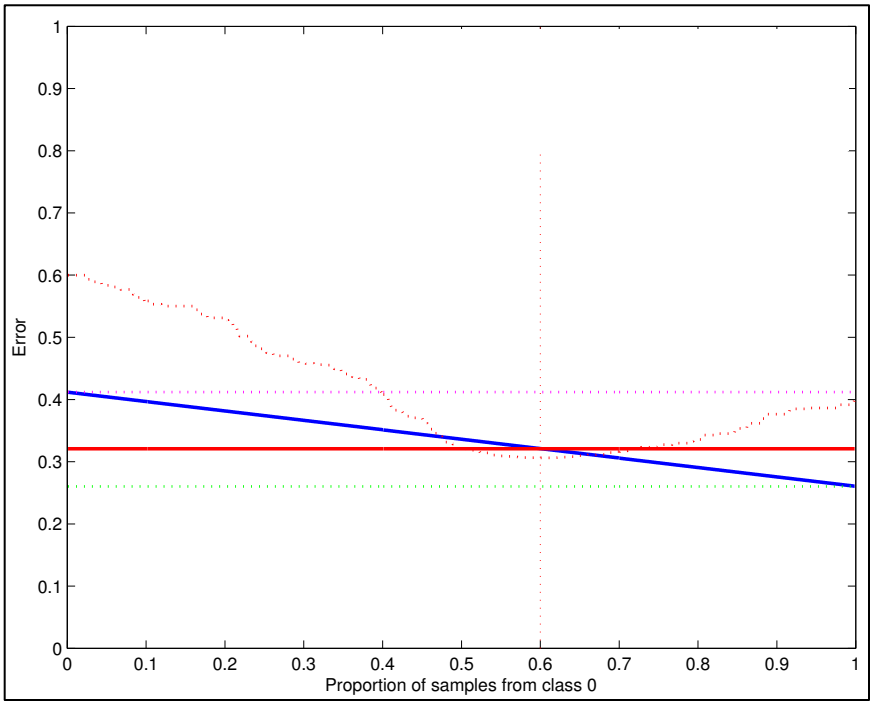

Figura 2: Resultados del error de los clasificadores obtenidos con diferentes proporciones en el muestreo, usando diseño balanceado. Las líneas sólidas representan el error verdadero (rojo) y aparente (azul) para el clasificador óptimo. Las líneas punteadas muestras los valores de FPR (verde) y FNR (rojo). La línea vertical muestra la proporción poblacional de verdaderos-falsos.

Predicción de fenotipo basado en genotipos

Como ejemplo de aplicación de balanceo de datos utilizamos datos de susceptibilidad de ratones a *listeria monocytogenes* [12]. Este estudio consiste en el análisis de la relación entre QTLs y la supervivencia de 120 ratones después de infectarlos con listeria. Qua*titative trait loci* (o QTLs) son partes de un gen que afecta un carácter cuantitativo. Animales con supervivencia de más de 240 horas fueron considerados *supervivientes*. La proporción final fue de 35 supervivientes y 85 no supervivientes. Un total de 133 QTLs fueron analizados para cada ratón, donde el genotipo presente para cada *loci* se puede categorizar como AA (A), AB (H), o BB (B). El carácter discreto de estos datos permite utilizar las técnicas de clasificación discreta para predecir el fenotipo (supervivencia) a partir del genotipo.

QTLs que no poseen información genotípica para todas las muestras fueron removidos. Aunque existen técnicas para estimación de datos faltantes, e incluso el diseño de operadores con datos incompletos, pero está fuera del alcance de este trabajo. El filtro removió 92 QTLs con datos incompletos, resultando en 28 QTLs completamente caracterizados en las 133 muestras. La figura 3 muestra una imagen de los genotipos y fenotipos para las 120 muestras y los 28 QTLs que poseen información para todas las muestras, donde los genotipos están representados por colores verde (A), negro (H) y rojo (B), y los fenotipos están indicados por los colores rojo (no supervivencia) y azul (supervivencia). Para el análisis discreto, se mapearon los valores al rango 0 (A), 1 (H) y 2 (B) para las variables predictoras (genotipos) y 0 (no supervivencia) y 1 (supervivencia).

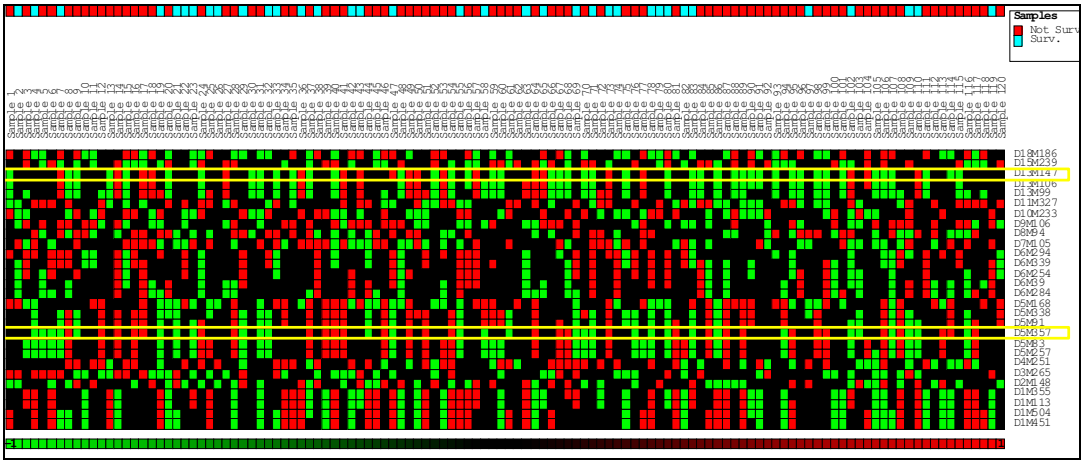

Figura 3. Genotipos y Fenotipos de los 28 QTLs seleccionados sobre las 133 muestras. Los recuadros en amarillo muestran las dos regiones seleccionadas para análisis.

Las regiones elegidas como predictoras son D13Mit147 y D5Mit357, seleccionados en el artículo original por su alta relación con el fenotipo [13], y marcadas en la figura 3. Dos clasificadores discretos fueron diseñados, usando la regla de plug-in, en base a los datos disponibles para el genotipo de ambas regiones, uno utilizando balanceo y el otro no. Para estimar el error de los clasificadores utilizamos 1000 iteraciones de validación cruzada de Monte Carlo [14], separando los datos en 80 muestras para el diseño del clasificador y 40 muestras para la evaluación del error, en cada iteración.

La tabla 1 muestra los resultados de entrenar un clasificador a partir de los datos con y sin balanceo. Al aplicar el diseño balanceado obtenemos un pequeño incremento en la tasa de error y en la tasa de falsos positivos, pero con un decremento de más del 50% de la tasa de falsos negativos. Resultados similares, donde una de las tasas de falsos positivos o negativos mejora considerablemente, a cambio de un bajo crecimiento de los otros valores, se observa consistentemente para otros pares de regiones predictoras.

Tabla 1 – Resultados del diseño de operadores con y sin balanceo

	Tasa de Error	**FPR**	**FNR**
Diseño Clásico	29.7%	50.9%	19.8%
Diseño Balanceado	30.3%	51.4%	9.7%

Discusión

El diseño balanceado usualmente incrementa la tasa de error del clasificador diseñado, pero a cambio de una mejora sustancial en las tasas de falsos positivos o falsos negativos. Para el caso particular de datos discretos, como en datos genómicos, una ventaja adicional es que el balance artificial de los datos se puede obtener sin necesidad de re-muestreo.

Agradecimientos

Marcel Brun recibió fondos de la Agencia Nacional de Promoción Científica y Tecnológica (PICT-2006-02313) para el desarrollo de este trabajo.

Referencias

[1] Hirschhorn Joel N. and Daly Mark J., "Genome-wide association studies for common diseases and complex traits" Nature Reviews Genetics , 6 , pp.95-108 , 2005

[2] Emahazion Tesfai , Feuk Lars , Jobs Magnus , Sawyer Sarah L. , Fredman David , Clair David St , Prince Jonathan A. , Brookes Anthony J., "SNP association studies in Alzheimerÿs disease highlight problems for complex disease analysis", in Trends in Genetics , 17 (7) , pp.407-413 , 2001

[3] M. Gonzalez, M.; Brun, Marcel; Corva, Pablo y V. L. Ballarin, "Clasificación de razas bovinas por un número reducido de SNPs", XVII Congreso Argentino de Bioingeniería- VI Jornadas de Ingeniería Clínica (SABI), ISBN: 978-950-605-505-9, Polo Tecnológico Rosario, 2009.

[4] K. W. Broman. "Review of statistical methods for qtl mapping in experimental crosses". Laboratory Animals, 30(7):44-52, 2001.

[5] Karl W. Broman, Saunak Sen, "A Guide to QTL Mapping with R/qtl", Statistics for Biology and Health, Springer, 2009

[6] Dougherty, E. R., Random Processes for Image and Signal Processing, Series on Imaging Science and Engineering, SPIE Press and IEEE Presses, Bellingham, 1999

[7] Dougherty, E. R., Bittner, M., Chen, Y., Kim, S., Sivakumar, K., Barrera, J., Meltzer, P., and Trent, J., "Nonlinear Filters in Genomic Control" in Proceedings of the IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing, Antalya, Turkey.

[8] Maloof Marcus A., "Learning When Data Sets are Imbalanced and When Costs are Unequal and Unknown", Workshop on Learning from Imbalanced Data Sets II, ICML, Washington DC, 2003.

[9] TAO Xiao-yan, JI Hong-bing, XIE Yu-xin, "A Modified PSVM and its Application to Unbalanced Data Classification", Third International Conference on Natural Computation (ICNC 2007), 24-27 Aug. 2007, Haikou, China

[10] Muhlbaier M., Topalis A., Polikar R., "Incremental learning from unbalanced data," Proc. of Int. Joint Conference on Neural Networks (IJCNN 2004), pp. 1057-1062, Budapest, Hungary, July 2004.
[11] Jianping Zhang, Inderjeet Mani, "kNN Approach to Unbalanced Data Distribution: A Case Study Involving Information Extraction", Workshop on Learning from Imbalanced Datasets II, ICML, Washington DC, 2003.
[12] Victor L. Boyartchuk, Karl W. Broman, Rebecca E. Mosher, Sarah E.F. D'Orazio, Michael N., "Multigenic control of Listeria monocytogenes susceptibility in mice", Nature Genetics, Vol 27, pp. 259-260, 2001
[13] Karl W.Broman, Victor L. Boyartchuk and William F. Dietrich, "Mapping time-to-death quantitative trait loci in a mouse cross with high survival rates", Technical Report MS00-04 Department of Biostatistics, Johns Hopkins University, 5 May 2000
[14] Annette M. Molinaro, Richard Simon, Ruth M. Pfeiffer, "Prediction error estimation: a comparison of resampling methods", Bioinformatics 2005 21(15):3301-3307