Una generalización del clasificador Naive Bayes para usarse en bases de datos con dependencia de variables

Ana E. Ruiz L.[1,2], Christopher R. Stephens S.[3,4], Hugo Flores [1,3]

[1] IIMAS, Universidad Nacional Autónoma de México, México D.F.
[2] Instituto Tecnológico de Minatitlán, Minatitlán, Ver., México
[3] C3 Centro de Ciencias de la Complejidad Universidad Nacional Autónoma de México, México D.F.
[4] Instituto de Ciencias Nucleares, Universidad Nacional Autónoma de México, México D.F.

aeruizlinares@hotmail.com, stephens@nucleares.unam.mx, hugo_fh@yahoo.com

Resumen. A pesar de la suposición que hace sobre la independencia de variables, el Clasificador Naive Bayes es muy utilizado en Minería de Datos y Aprendizaje Automático debido principalmente a su relativa simpleza y robustez mostrados frente a gran cantidad de problemas. Al suponer una independencia de variables, el modelo de NB proporciona un modelo no representativo cuando la base de datos tiene variables dependientes. Ante esta situación, se han propuesto varias aproximaciones que mejoran el desempeño del NB pero requieren mayores recursos y resultan complicados de implementar. Aquí se propone una nueva aproximación que puede ser usada cuando exista dependencia de variables conservando una sencillez de implementación. También se propone una métrica para determinar a priori si utilizar la aproximación más simple del clasificador NB o no. Los resultados obtenidos en cuatro bases de UCI mostraron que el modelo propuesto mejora el desempeño del NB cuando existe dependencia de variables.

Palabras clave: Clasificación · Naive Bayes · dependencia de variables.

1. Introducción

Un modelo muy utilizado en Clasificación debido a su robustez y sencillez de implementación es el Clasificador Naive Bayes (CNB), sin embargo, la suposición que hace de que todos los atributos son condicionalmente independientes dada una clase no siempre se cumple en aplicaciones del mundo real, produciéndose un decremento en su desempeño [1].

Con el fin de atenuar el impacto de la independencia de variables se han propuesto varias aproximaciones al CNB, entre ellos: *Semi-NB Classifiers* [2, 3, 4, 5]; *The Tree Augmented Naive Bayes (TAN)* [6]; *Super Parent TAN* [7,8]; *Improved Naive Bayes (INB)* [9]; *Weighted NB* [10-15]; Taheri et al. en [16] proponen un algoritmo llamado *Attribute Weighted NB (AWNB)* que asigna más de un peso a cada atributo. En [17] se muestra un panorama general del desempeño de varios de estos modelos.

Aunque estas aproximaciones, en términos generales, han presentado mejor desempeño que el NB, utilizan más recursos computacionales y resultan más complejas en el momento de implementarlas. Aquí se propone una aproximación llamada Naive Bayes Generalizado (NBG), con dos variantes: "simétrico" y "no simétrico", que puede ser usado cuando existen dependencias o correlaciones entre dos variables y conserva la sencillez de implementación del NB. Además, se presenta el uso de una herramienta de diagnóstico, denominada épsilon (ε), para averiguar si existen variables correlacionadas en la base de datos a examinar, permitiendo determinar a priori si la aproximación simple del NB es adecuada o es necesario invertir más recursos en la implementación de alguna otra aproximación más sofisticada del método.

2. Naive Bayes Generalizado NBG

El modelo de NBG - a diferencia del NB, que sólo analiza la acción de una variable sobre una clase- permite examinar el impacto sobre la clase de dos variables que actúan de forma sinérgica (fig. 1); cuando en una base de datos, la interacción de dos variables incrementa la probabilidad de clase, se dice que las variables son dependientes o están correlacionadas.

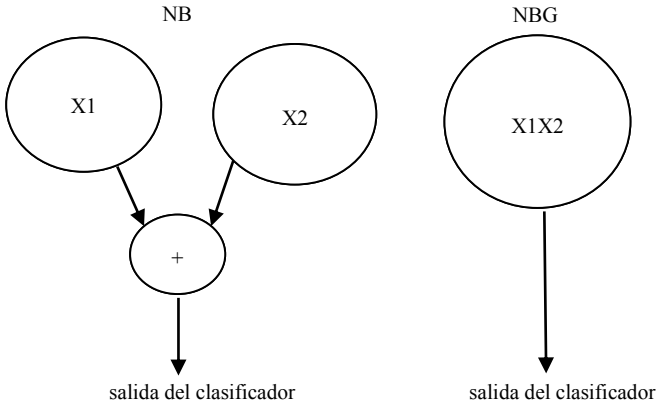


Fig. 1. Diferencia en el comportamiento de un clasificador NB y uno NBG. Se observa en el modelo de la derecha la acción independiente de cada variable, mientras que en el de la derecha se considera la acción conjunta de dos variables.

Al suponer una independencia de variables el modelo de NB proporciona un score de riesgo S_{NB} - función monotónica de la probabilidad condicional $P(C|\boldsymbol{X})$ siendo $\boldsymbol{X}$ un vector con n variables $\boldsymbol{X} = (x_1, x_2, \dots, x_{n)}$ y C la clase - determinado por la ec. 1, donde $P(C)$ es la probabilidad a priori de la clase y $P(\bar{C})$ la probabilidad a priori de la no clase. Con el fin de simplificar el análisis, supóngase que la base de datos sólo tiene dos variables x_1 y x_2, entonces el score de riesgo alcanzado con un CNB se re-

duce a la ec. 2; mientras que el modelo NBG, al contemplar la acción conjunta de dos variables sobre la clase, obtiene un score de riesgo S_{NBG} especificado por la ec.3.

$$S_{NB} = \ln\frac{P(C)}{P(\bar{C})} + \sum_{i=1}^{n} ln\frac{P(X_i|C)}{P(X_i|\bar{C})} \tag{1}$$

$$S_{NB} = \ln\frac{P(C)}{P(\bar{C})} + ln\frac{P(X_1|C)}{P(X_1|\bar{C})} + ln\frac{P(X_2|C)}{P(X_2|\bar{C})} \tag{2}$$

$$S_{NBG} = \ln\frac{P(C)}{P(\bar{C})} + ln\left(\frac{P(X_1X_2|C)}{P(X_1X_2|\bar{C})}\right) \tag{3}$$

2.1 Épsilon ε

Stephens en [18] plantea el uso de ε como una medida sobre el impacto que tiene cierto valor de variable sobre la clase, mencionando que valores ε>2 indican que ese valor de variable es predictiva de la clase, mientras que valores ε<-2 indican que es predictiva para no clase. Aquí se plantea el uso de ec. 4 para determinar si la combinación de $X_1 = i$ con $X_2 = j$ es predictiva de clase y ec. 5 para no clase.

N_C es el número de ocurrencia de clase; $N_{\bar{C}}$ es el número de ocurrencias para no clase; el número de ocurrencias de $X_1 = i$ y $X_2 = j$ con la clase es $P(X_{1i}X_{2j}|C)$ y con la no clase $P(X_{1i}X_{2j}|\bar{C})$; el número de coincidencias de $X_1 = i$ con la clase es $P(X_{1i}|C)$ y $P(X_{1i}|\bar{C})$ con la no clase; finalmente, el número de coincidencias de $X_2 = j$ con la clase es $P(X_{2j}|C)$y con la no clase es $P(X_{2j}|\bar{C})$.

$$\varepsilon_C = \frac{N_C\left(P(X_{1i}X_{2j}|C) - P(X_{1i}|C)P(X_{2j}|C)\right)}{\sqrt{N_C P(X_{1i}|C)P(X_{2j}|C)\left(1 - P(X_{1i}|C)P(X_{2j}|C)\right)}} \tag{4}$$

$$\varepsilon_{\bar{C}} = \frac{N_{\bar{C}}\left(P(X_{1i}X_{2j}|\bar{C}) - P(X_{1i}|\bar{C})P(X_{2j}|\bar{C})\right)}{\sqrt{N_{\bar{C}} P(X_{1i}|\bar{C})P(X_{2j}|\bar{C})\left(1 - P(X_{1i}|\bar{C})P(X_{2j}|\bar{C})\right)}} \tag{5}$$

El numerador de ec. 4 es la diferencia entre el número real de ocurrencias de $X_1 = i$ y $X_2 = j$ con la clase C y el número esperado de acuerdo con la hipótesis nula de que X_1 y X_2 son independientes de C, esto es, la hipótesis nula es la aproximación de Naive Bayes. El numerador de ec. 5 mide esta diferencia con respecto a no clase, $\bar{C}$. Los denominadores de las ec 4 y 5 están asociados con las desviaciones estándar de sus respectivos numeradores.

Ya que tanto ec. 4 como ec. 5 emplean una distribución binomial, valores $|\varepsilon_C| > 2$ y/o $|\varepsilon_{\bar{C}}| > 2$ representarían el intervalo de confianza del 95 % estándar y serían estadísticamente significativos para considerar $X_1 = i$ y $X_2 = j$ correlacionadas para clase y/o no clase. La idea es que si ε_C y/o $\varepsilon_{\bar{C}}$ son significativos: $X_1 = i$ y $X_2 = j$ son variables dependientes y el desempeño del modelo NB se verá reducida, en este caso se recomendaría el uso de alguna otra aproximación al modelo que considera las dependencias de variables, tal como lo hace el NBG.

En [19] utilizan ε como primer paso en la búsqueda de valores de variables que tienen mayor influencia en la predicción de clase. En [20] usan ε para medir la dependencia estadística de un taxón relativa a una hipótesis nula de independencia.

2.2 NBG "simétrico", NBG1

Se denomina simétrico porque se considera que una combinación $X_{1i}X_{2j}$ con un valor significativo de ε_C y/o $\varepsilon_{\bar{C}}$ tiene el mismo impacto en la predictibilidad de la clase y de la no clase. El clasificador que usa esta aproximación calcula primero el score de todas las combinaciones significativamente dependientes. Si la unión de esas dos variables no es estadísticamente significativa, se considera que cada variable actúa de forma independiente sobre la clase y el clasificador calcula el score de esas dos variables usando el modelo NB. Matemáticamente, el comportamiento del clasificador se muestra en la ec. 6.

$$S_{NBG1} = \sum ln\left(\frac{P(X_{\mathrm{lk}}X_{\mathrm{nm}}|C)}{P(X_{\mathrm{lk}}X_{\mathrm{nm}}|\bar{C})}\right) + \sum ln\frac{P(X_i|C)}{P(X_i|\bar{C})} + \ln\frac{P(C)}{P(\bar{C})} \tag{6}$$

2.3 NBG "no simétrico", NBG2

Esta otra variante se denomina no simétrico porque el clasificador considera que una combinación $X_{1i}X_{2j}$ no necesariamente tiene el mismo impacto en la predictibilidad de la clase que de la no clase. El clasificador calcula el score para clase (ec. 7) o el score para no clase (ec. 8) dependiendo si la combinación es considerada dependiente para clase o para no clase. Si la combinación de variables es estadísticamente independiente, utiliza el modelo de NB para calcular el score de esas dos variables por separado. Matemáticamente, el comportamiento del clasificador se muestra en la ec. 9.

$$S_C = \ln P(X_{\mathrm{lk}}X_{\mathrm{mn}}|C) \tag{7}$$

$$S_{\bar{C}} = \ln P(X_{\mathrm{lk}}X_{\mathrm{mn}}|\bar{C}) \tag{8}$$

$$S_{NBG2} = \sum S_C - \sum S_{\bar{C}} + \sum ln\frac{P(X_i|C)}{P(X_i|\bar{C})} + \ln\frac{P(C)}{P(\bar{C})} \tag{9}$$

2.4 Corrección de Laplace

En [21] Provest y Domingos muestran que la estimación de la probabilidad mejora si en vez de calcular la probabilidad basada en frecuencia, se utiliza la Corrección de Laplace para suavizar dicho cálculo (ec. 10).

$$P(X_i|C) = \frac{N_{X_iC}+1}{N_C+k} \tag{10}$$

donde N_{X_iC} es el número de ocurrencias en la base de datos donde aparece el valor de la variable X_i en el subconjunto de registros de la clase, N_C es el total de registros de la base de datos y k es el total de valores que puede tener la clase. En este trabajo se utilizó la corrección de Laplace para el cálculo de todas las probabilidades.

3. Pruebas en base de datos de la UCI

En la tabla 1, se mencionan las bases de datos del repositorio de la Universidad de Irvine (UCI Repository) [22] con los cuales se probaron los modelos anteriores. Antes de generar los modelos de entrenamiento de los tres clasificadores - NB, NBG1 y NBG2- cada una de las bases de datos fueron divididas en un conjunto de entrenamiento y un conjunto de prueba. Para conocer a priori la existencia de variables correlacionadas en las bases se utilizaron los conjuntos de entrenamiento para calcular los valores de ε_C y $\varepsilon_{\bar{C}}$ que se muestran en las figuras 2, 3, 4, 5, 6, 7, 8 y 9; en estas figuras, el eje X es el valor de ε y el eje Y es el número de combinaciones $X_1 = i$ y $X_2 = j$.

Tabla 1. Bases usadas para probar el modelo. Nombre original: nombre bajo el cual se puede encontrar en el repositorio de UCI; Nombre traducido: nombre con el cual se designará en el presente trabajo; No. Reg: número de registros de la base; No. Atrib: número de atributos; P(C): probabilidad de la clase.

Nombre original	Nombre traducido	No. Reg	No. Atrib	P (C)
Breast cancer Wisconsin	Cáncer de mama	699	10	0.34
Mushroom	Hongos	8124	22	0.48
Tic Tac Toe Endgame	Tic Tac Toe	958	9	0.65
Congressional Voting Reords	Votos	435	16	0.61

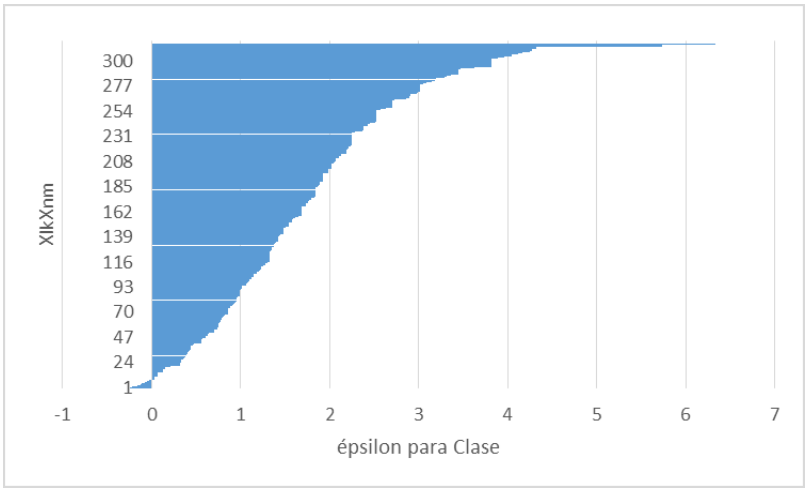


Fig. 2. Gráfico de valores de ε_C para la base de Cáncer.

Después de calcular ε_C y $\varepsilon_{\bar{C}}$ se supo que: en la base de hongos el 83% de las combinaciones de valores de variables son estadísticamente significativas para decir que actúan de forma correlacionada, mientras que en la base de cáncer sólo lo son el 38% de las combinaciones. Además, debido a que en la base de hongos se alcanzan valores más altos de ε_C y $\varepsilon_{\bar{C}}$ que en la de cáncer, se dice que en la base de hongos hay combinaciones de variables que son fuertemente predictivas para clase y para no clase, a diferencia de la base de cáncer, donde se observa en las fig. 2 y fig. 3 que hay más variables dependientes para clase que para no clase.

Lo anterior pudo ser la razón por la cual Pazzani en [23] encontrara mejores resultados con sus algoritmos que manejan dependencia de variables con respecto al NB en la base de hongos y no así en la de cáncer.

En la base de Tic tac toe alrededor de un 21 % de las combinaciones $X_{\mathrm{lk}}X_{\mathrm{nm}}$ indican correlación pero los valores de los indicadores no son muy altos, como se puede apreciar en las fig. 6 y fig. 7. Sucede algo similar en la base de Votos (fig. 8 y fig. 9).

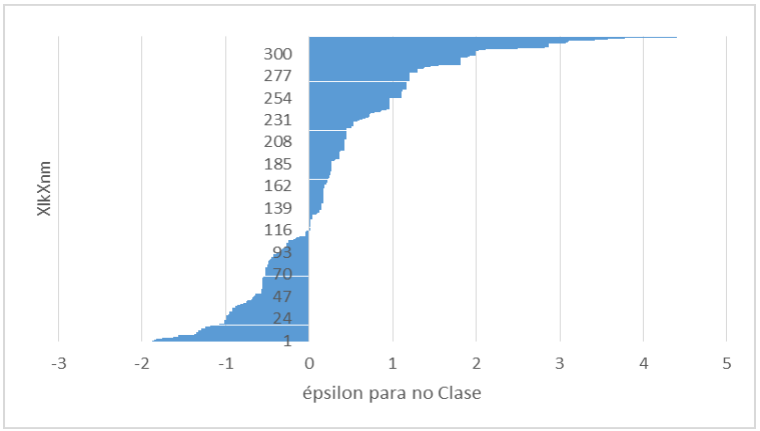


Fig. 3. Gráfico de valores de $\varepsilon_{\bar{C}}$ para la base de Cáncer.

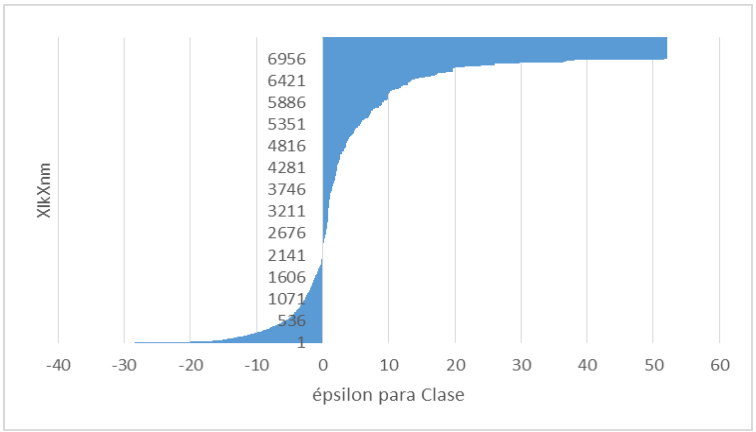


Fig. 4. Gráfico de valores de ε_C para la base de Hongos.

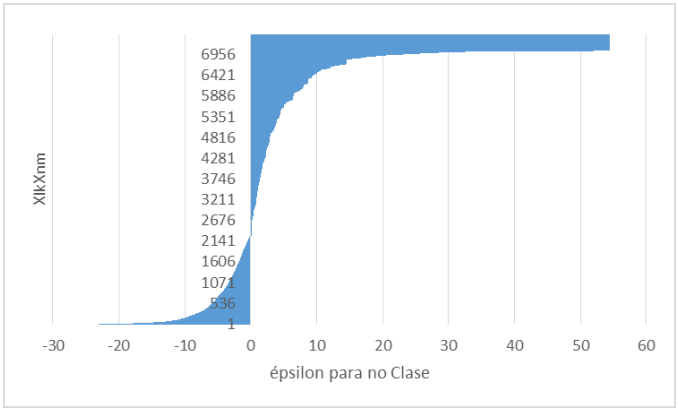


Fig. 5. Gráfico de valores de $\varepsilon_{\bar{C}}$ para la base de Hongos.

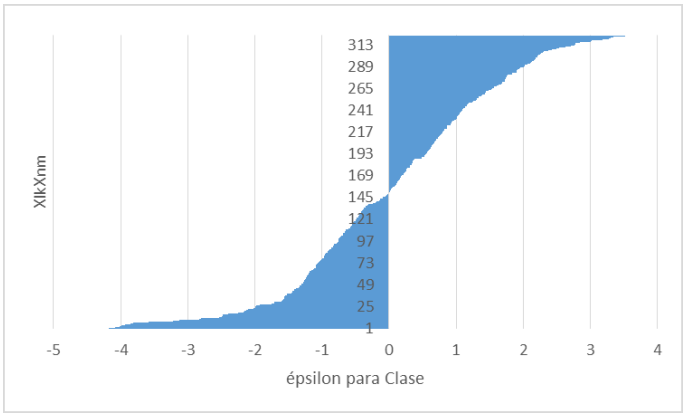


Fig. 6. Gráfico de valores de ε_{C} para la base de Tic tac toe.

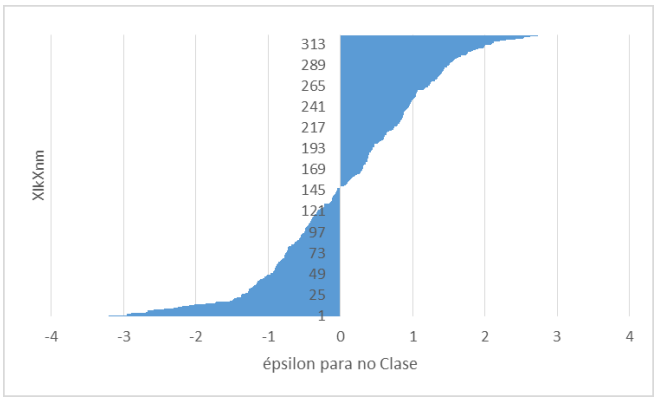


Fig. 7. Gráfico de valores de $\varepsilon_{\bar{C}}$ para la base de Tic tac toe.

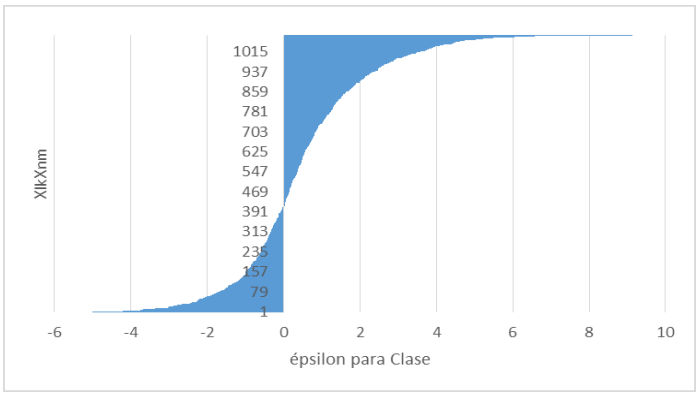


Fig. 8. Gráfico de valores de ε_C para la base de Votos.

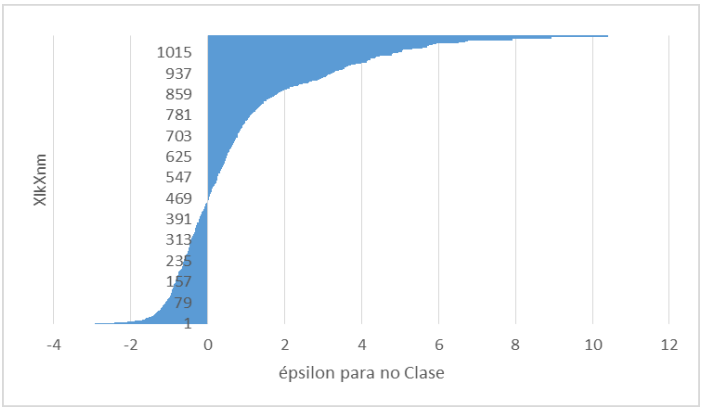


Fig. 9. Gráfico de valores de $\varepsilon_{\bar{C}}$ para la base de Votos.

Tabla 2. Matriz de confusión de la base de Cáncer de mama.

		salida del clasificador					
	CÁNCER	NB		NBG1		NBG2	
		clase	no clase	clase	no clase	clase	no clase
real	clase	75	1	75	1	75	1
	no clase	3	131	4	130	3	131

Tabla 3. Matriz de confusión de la base de Hongos.

		salida del clasificador					
	HONGOS	NB		NBG1		NBG2	
		clase	no clase	clase	no clase	clase	no clase
real	clase	1075	117	1175	17	1140	52
	no clase	5	1240	6	1239	4	1241

Después de correr los tres clasificadores en cada base de dato, se obtuvieron las matrices de confusión (tabla 2, 3, 4, 5) así como el desempeño de cada uno en términos de *sensibilidad*, *especificidad*, *eficiencia* y *error* (tabla 6).

Tabla 4. Matriz de confusión de la base de Tic tac toe.

		salida del clasificador					
	TIC TAC TOE	NB		NBG1		NBG2	
		clase	no clase	clase	no clase	clase	no clase
real	Clase	166	19	153	32	159	26
	no clase	62	41	52	51	48	55

Tabla 5. Matriz de confusión de la base de Votos.

		salida del clasificador					
		NB		NBG1		NBG2	
		clase	no clase	clase	no clase	clase	no clase
real	clase	64	12	64	12	64	12
	no clase	5	49	3	51	2	53

Tabla 6. Desempeño de los tres clasificadores aplicados a la bases de datos.

		SENSIBILIDAD	ESPEC	EFICIENCIA	ERROR
CÁNCER	NB	0.99	0.98	0.98	0.02
	NBG1	0.99	0.97	0.98	0.02
	NBG2	0.99	0.98	0.98	0.02
HONGOS	NB	0.9	1	0.95	0.05
	NBG1	0.99	1	0.99	0.01
	NBG2	0.96	1	0.98	0.02
TIC TAC TOE	NB	0.9	0.4	0.72	0.28
	NBG1	0.83	0.5	0.71	0.29
	NBG2	0.86	0.53	0.74	0.26
VOTOS	NB	0.84	0.91	0.87	0.13
	NBG1	0.84	0.94	0.88	0.12
	NBG2	0.84	0.96	0.89	0.11

Tabla 7. Desempeño promedio de 20 corridas de los tres clasificadores.

		SENSIBILIDAD	ESPEC	EFICIENCIA	ERROR
CÁNCER	NB	0.98	0.97	0.97	0.03
	NBG1	0.99	0.97	0.97	0.03
	NBG2	0.99	0.97	0.98	0.02
HONGOS	NB	0.91	0.99	0.95	0.05
	NBG1	0.99	0.99	0.99	0.01
	NBG2	0.97	1	0.98	0.02
TIC TAC TOE	NB	0.85	0.42	0.71	0.29
	NBG1	0.82	0.5	0.71	0.29
	NBG2	0.82	0.55	0.73	0.27
VOTOS	NB	0.88	0.92	0.9	0.1
	NBG1	0.89	0.93	0.9	0.1
	NBG2	0.89	0.95	0.91	0.09

En la tabla 7 se muestra el desempeño promedio de 20 corridas de los clasificadores. En la tabla 8 y tabla 9 se muestran las Curvas ROC (Receiver Operating Characteristics) y el área bajo la curva ROC conocido como AUC, ya que algunos trabajos [24,25] han mostrado que son una evaluación más discriminante que la razón de error en algoritmos de Aprendizaje Automático con estimaciones de probabilidad de clase. Se observa en dichas tablas que en la base de cáncer los tres clasificadores tienen el mismo valor de AUC, mientras que en las otras tres bases el valor de AUC es mayor en el modelo propuesto que el alcanzado con el CNB. Lo anterior es congruente con los valores obtenidos de ε.

4. Conclusiones y comentarios finales

Los resultados encontrados sugieren que el método propuesto, al tomar en cuenta la acción conjunta de dos valores de variables sobre la predictibilidad de la clase, mejora el desempeño del CNB en bases de datos donde existen variables correlacionadas. El grado de correlación o de dependencia se obtuvo a través de la herramienta de diagnóstico ε. Los valores que se logren con este indicador servirán para decidir si utilizar la aproximación más simple del clasificador NB o buscar otra más sofisticada.

Tabla 8. Curvas ROC y AUC de los tres clasificadores para la base de Cáncer (izquierda) y de Hongos (derecha). En el eje de las X está *(1-Especificidad)* y en el eje Y la *Sensibilidad*.

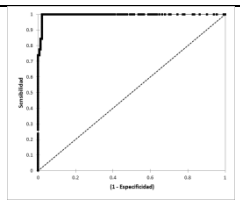
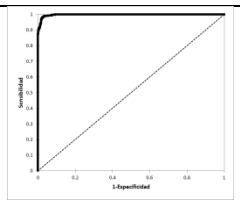
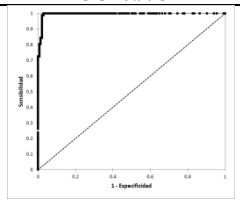
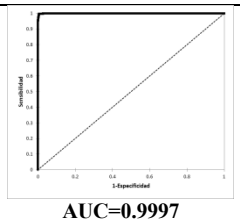
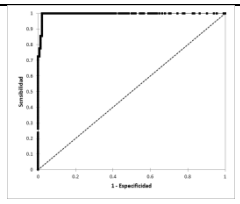
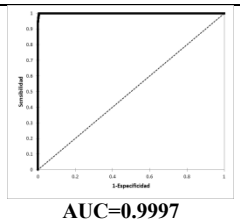
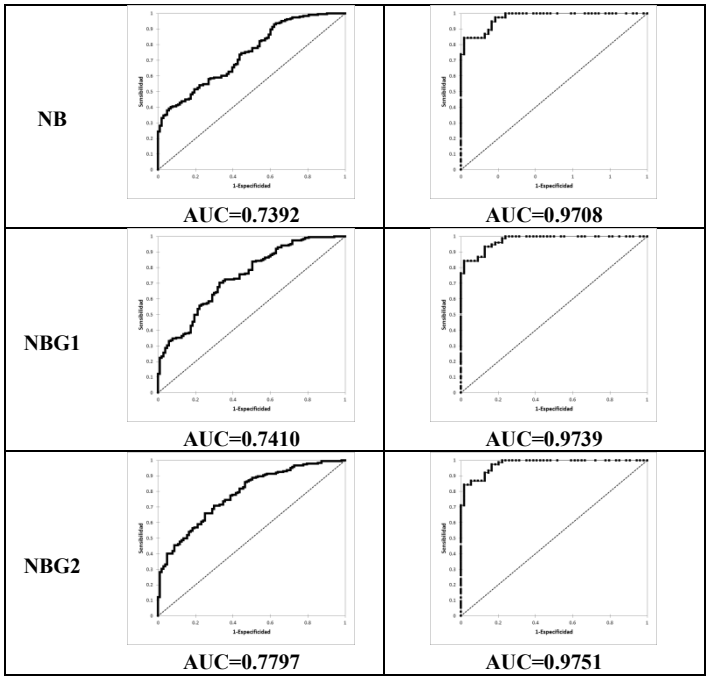
NB	 **AUC=0.9952**	 **AUC=0.9975**
NBG1	 **AUC=0.9952**	 **AUC=0.9997**
NBG2	 **AUC=0.9952**	 **AUC=0.9997**

Tabla 9. Curvas ROC y AUC de los tres clasificadores para la base de Tic tac toe (izquierda) y Votos (derecha). En el eje de las X está *(1-Especificidad)* y en el eje Y la *Sensibilidad*.



Aunque aquí sólo se consideran correlaciones entre un par de valores de variables, se puede buscar la correlación entre más de dos atributos haciendo múltiples pasos. Además, ya que los experimentos se hicieron sobre clasificaciones binarias, se tiene contemplado investigar el comportamiento del NBG con bases de datos multiclases. Resultaría también de gran interés, en un futuro, hacer una comparación en términos de complejidad computacional del método aquí propuesto con otras aproximaciones que también buscan atenuar la estricta restricción de independencia de variables que hace Naive Bayes.

Agradecimientos. Los autores agradecen el apoyo del proyecto PAPIIT-UNAM IN113414 para la realización de este trabajo.

Referencias

1. Jiang L, Wang D, Cai Z, Yan X.: Survey of Improving Naive Bayes for Classification. In: Berlin, Heidelberg: Springer Berlin Heidelberg; pp. 134–45 (2007)
2. Kohavi R: Scaling up the accuracy of naive-Bayes classifiers: a decision-tree hybrid. In: Proceedings of 2nd ACM SIGKDD International, Conference on Knowledge Discovery and Data Mining, pp 202–207 (1996)

3. Langley P, Saga S: Induction of selective Bayesian classifiers. In: Proceedings of tenth conference, uncertainty in artificial intelligence, Morgan Kaufmann, pp 399–406 (1994)
4. Pazzani MJ.: Constructive induction of Cartesian product attributes, ISIS: In-formation, Stat Induction Sci.,pp. 66–77 (1996)
5. Robles V, Larranaga P, Pena J M, Perez M S, Menasalvas E, Herves V.: Learning Semi Naive Bayes Structures by Estimation of Distribution Algorithms. Lecture Notes in Computer Science, vol. 292, pp. 244–58 (2003)
6. Friedman N, Geiger D, Goldszmidt M.: Bayesian network classifiers. Mach Learn, 29, pp. 131–163 (1997)
7. Keogh EJ, Pazzani MJ Learning augmented Bayesian classifiers: A comparison of distribution-based and classification based approaches. In: Proceedings of international workshop on artificial intelligence and statistics, pp. 225–230 (1999)
8. Webb GI, Boughton JR, Wang Z.: Not So Naive Bayes: Aggregating One-Dependence Estimators. Mach Learning, vol. 58, no. 1, pp. 5–24 (2005)
9. Taheri S, Mammadov M, Bagirov A.M.: Improving Naive Bayes classifier using conditional probabilities. In: Proceedings of ninth Australasian data mining conference (AusDM 2011), vol. 125, Ballarat, Australia (2011)
10. Zhang H, Sheng S.: Learning weighted Naive Bayes with accurate ranking. In: Proceedings of the 4th IEEE international conference on data mining, pp. 567–570 (2005)
11. Zhou Y, Huang T.S.: Weighted Bayesian network for visual tracking. In: Proceedings of the 18th international conference on pattern recognition (ICPR'O6) (2006)
12. Jiang L, Zhang H.: Weightily averaged one-dependence estimators. In: Proceedings of the 9th biennial pacific rim international conference on artificial intelligence, Guilin, China, pp. 970–974 (2006)
13. Hall M: A decision tree-based attribute weighting filter for Naive Bayes. Knowledge Based Systems , vol. 20, no. 2, pp.120–126 (2007)
14. Ozsen S, Gunecs S.: Attribute weighting via genetic algorithms for attribute weighted artificial immune system (AWAIS) and its application to heart disease and liver disorders problems. Expert Systems with Applications, vol. 36, no. 1, pp. 386–392 (2009)
15. Wu J, Cai Z.: Attribute weighting via differential evolution algorithm for attribute weighted Naive Bayes (WNB). Journal of Computational Information Systems, vol. 7, no. 5, pp.1672–1679 (2011)
16. Taheri S., Yearwood J., Mammadov M., Seifollahi S.: Attribute weighted Naive Bayes classifier using a local optimization. Neural Computing and Applications, vol. 24, no.5, pp. 995–1002 (2014)
17. Jiang L, Zhang H, Cai Z. A Novel Bayes Model: Hidden Naive Bayes. IEEE Trans Knowledge and Data Engineering, vol. 21, no.10, pp.1361–1371 (2009)
18. Stephens, C.: An introduction to data mining. In: Grover, R. and Vriens, M. (eds), The handbook of marketing research: uses, misuses and future advances. Sage Publ., pp. 445–486 (2006)
19. Stephens, C. R., Waelbroeck, H., and Talley, S.: Predicting healthcare costs using GAs. In: Proceedings of the 2005 workshops on Genetic and evolutionary computation, pp. 159–163, ACM (2005)
20. González-Salazar C., Stephens C.R., Marquet P.A.: Comparing the relative contributions of biotic and abiotic factors as mediators of species' distributions. Ecological Modelling, vol. 248, pp.57–70 (2013)
21. Provost F, Domingos P.: Tree Induction for Probability-Based Ranking. Machine Learning, vol. 52, no. 3, pp.199–215 (2003)
22. Murphy, P. M. , & Aha, D. W.: UCI Repository of machine learning databases. Irvine: University of California, Department of Information & Computer Science [Machine-readable data repository http://archive.ics.uci.edu/ml/datasets.html] (1995)

23. Pazzani, M. J.: Searching for dependencies in Bayesian classifiers. In: D. Fisher & H. Lenz (Eds.), Proceedings of the fifth International Workshop on Artificial Intelligence and Statistics, Ft. Lauderdale, FL (1995)
24. Bradley A.P.: The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition, vol.30, no. 7, pp.1145–59 (1997)
25. Huang J., Ling C.X.: Using AUC and accuracy in evaluating learning algorithms. Knowledge and Data Engineering, IEEE Transactions on, vol.17, no. 3, pp.299–310 (2005)