Revista Cubana de Ciencias Informáticas
Vol. 11, No. 4, Octubre-Diciembre, 2017
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 67- 84

Tipo de artículo: Artículo original
Temática: Inteligencia artificial
Recibido: 09/02/2017 | Aceptado: 23/09/2017

Estudio experimental para la comparación del desempeño de Naïve Bayes con otros clasificadores bayesianos

Experimental Study for the Comparison of Naïve Bayes with other Bayesian Classifiers

Alain Pereira-Toledo[1] *, José D. López-Cabrera [2], Luis A. Quintero-Domínguez [1]

[1] Universidad de Sancti Spíritus "José Martí Pérez". Comandante Manuel Fajardo s/n, Olivos 2, Sancti Spíritus, Sancti Spíritus, Cuba. {alain, lqdominguez}@uniss.edu.cu
[2] Universidad Central "Marta Abreu" de Las Villas. Carretera a Camajuaní km 5½. Santa Clara, Villa Clara, Cuba. josedaniellc@uclv.cu

* Autor para correspondencia: alain@uniss.edu.cu

Resumen

En este artículo se realiza un estudio experimental para la comparación del desempeño de la clasificación de Naïve Bayes con otros métodos bayesianos. Otros experimentos reportan su competitividad, pero no utilizan la metodología mucho más apropiada que fue propuesta por Demšar. Por tanto, el propósito es volver sobre el estudio de Naïve Bayes frente a otros métodos bayesianos. Para ello se escogieron cinco clasificadores bayesianos más Naïve Bayes, todos implementados en WEKA; y también, tres conjuntos de bases de datos para realizar un experimento general, uno sobre atributos irrelevantes y un tercero sobre ejemplos ruidosos. Finalmente, siguiendo la metodología de Demšar, se mostró evidencia empírica que ubica aún a Naïve Bayes como una alternativa competitiva frente al resto de los clasificadores bayesianos seleccionados.

Palabras clave: Naïve Bayes, clasificadores bayesianos, estudio experimental, atributos irrelevantes, ejemplos ruidosos

Abstract

An experimental study is conducted for the comparison of Naïve Bayes classification performance with other Bayesian classifiers. There are some other similar reports of experiments, but they do not use the more suitable methodology proposed by Demšar. That is why the purpose is revisiting Naïve Bayes in comparison with other

Bayesian classifiers. In consequence, we chose five Bayesian classifiers and Naïve Bayes; all of them are implemented in WEKA. In addition, we chose three set of databases for the three set of experiments: a general experiment, an irrelevant feature experiment and a noisy values experiment. Finally, we showed by means of the Demšar methodology that there is empirical evidence for stating that Naïve Bayes still is a competitive alternative to the rest of the selected Bayesian classifiers.

Keywords: *Naïve Bayes, Bayesian classifiers, experimental study, irrelevant features, noisy values*

Introducción

Naïve Bayes es uno de los algoritmos de aprendizaje inductivo más eficientes y efectivos. Simplifica considerablemente el aprendizaje mediante el supuesto de independencia de los atributos y, no obstante, compite en la práctica con clasificadores más sofisticados. Ha sido considerado además como uno de los 10 algoritmos de minería de datos más influyentes por varias razones (Wu et al., 2008). Es fácil de construir y no necesita de esquemas iterativos de estimación de parámetros. Ello implica que puede ser aplicado a grandes bases de datos. Es también fácil de interpretar por usuarios no familiarizados con los clasificadores.

Sin conformarse, algunos han intentado mejorar el desempeño de Naïve Bayes. La manera de hacerlo generalmente es mediante la relajación de su supuesto fundamental, la independencia de los atributos. Para ello es necesario una representación adecuada de tales dependencias condicionales. Las redes bayesianas son una respuesta apropiada, pero ha sido probado que el problema de aprender la red óptima para una base de datos es de clase NP-hard (Chickering, 1996). Por tanto, una estrategia ha sido la de restringir la red bayesiana que se pretende aprender de los datos o, lo que es lo mismo, aumentar la red simple que es Naïve Bayes. Otras direcciones seguidas al respecto, han sido la selección de subconjuntos de atributos en los cuales se cumpla el supuesto de independencia, el empleo del principio de aprendizaje local y la expansión de los datos de entrenamiento (Jiang et al., 2007).

Pero tales modificaciones son complicaciones que atentan contra la simplicidad de Naïve Bayes, una de sus cualidades fundamentales. En consecuencia, es necesario volver hacia su estudio en el contexto de los clasificadores de su misma familia, con el propósito de comprobar su competitividad en cuanto a la precisión en la clasificación. Con ello no se pretende desestimular la investigación en el campo de los clasificadores bayesianos, sino mostrar si Naïve Bayes es una alternativa viable ante un problema donde la simplicidad sea una característica deseada.

En este mismo sentido, (Fernández-Delgado, Cernadas, Barro, 2014) han cuestionado la necesidad de crear nuevos algoritmos cuando muchos de los estudios comparativos realizados muestran deficiencias. Según (Demšar, 2006), las pruebas estadísticas usadas hasta ese entonces, no eran las más adecuadas para comparar el desempeño de la clasificación de múltiples algoritmos sobre múltiples bases de datos, por lo que propuso otra metodología para ello. Esta idea fue mejorada y extendida por (García, Herrera, 2008). Y como también apuntaba (Fernández-Delgado, Cernadas, Barro, 2014), existe un sesgo en cuanto a la selección de algoritmos para comparar.

Varios estudios experimentales en los que se comparan clasificadores bayesianos han sido llevados a cabo con diversas metodologías. Evidentemente todos aquellos conducidos antes de 2006 (Friedman, Geiger, Goldszmidt, 1997; Cheng, Greiner, 1999; Rish, 2001; Webb, Boughton, Wang, 2005; Zhang, Jiang, Su, 2005), no utilizan la metodología de Demšar, pues fue este el año en que la publicó. Pero lo más sorprendente es que estudios posteriores (Jiang et al., 2007; Jiang, Zhang, Cai, 2009), tampoco la hayan tomado en cuenta.

Debe notarse además que la mayoría de los experimentos conducidos con clasificadores bayesianos, se restringen a comparaciones generales del desempeño. Los problemas en los que se aplican tales clasificadores difieren en cuanto al origen, tipos de atributos, balance de clases o números de ejemplos. Pero también se diferencian en cuanto a niveles de ruido generados de forma natural a partir de, por ejemplo, la adquisición de datos (Nettleton, Orriols-Puig, Fornells, 2010), y a la presencia de atributos irrelevantes que se derivan, generalmente, del proceso de recolección de datos disponibles en el dominio (Güvenir, 1998).

Por tanto, se pretende conducir un estudio comparativo de Naïve Bayes con otros clasificadores bayesianos mediante la metodología propuesta por (Demšar, 2006) y ampliada por (García, Herrera, 2008), que incluya adicionalmente el desempeño ante la presencia de ejemplos ruidosos y atributos irrelevantes. Tres secciones se han escrito para ello. La primera introduce las definiciones y conceptos fundamentales, de Naïve Bayes y otros clasificadores bayesianos. Luego se describe la configuración del experimento y finalmente, se muestran y discuten los resultados obtenidos de las pruebas estadísticas utilizadas.

Materiales y métodos

El estudio experimental está dividido en tres. El primero es de tipo general, en el cual se seleccionan un conjunto de bases de datos con características diversas. El segundo está orientado a la presencia de atributos irrelevantes; mientras

que el tercero, a la contaminación de ejemplos con ruido. Para ello se han escogido un grupo de clasificadores bayesianos y las bases de datos, para realizar las pruebas según la metodología de evaluación estadística de Demšar.

El clasificador Naïve Bayes

El objetivo de la clasificación en el aprendizaje automático es entrenar un método determinado a partir de un conjunto de datos para construir un modelo que sea capaz de predecir uno de los valores nominales que pertenecen al dominio de un atributo llamado clase (Witten, Frank, Hall, 2011). Usualmente un ejemplo *E* es representado como una tupla de valores de atributos (x_1, x_2,..., x_n), donde x_i es el valor del atributo X_i. Sea también *C* la representación del atributo clase y *c* el valor de *C*. Por cuestiones de simplicidad y solo en el caso de la definición que a continuación se detalla, se asumen clases binarias con los valores c_0 y c_1. Zhang (2004), quien expone con claridad los formalismos de Naïve Bayes, define a un clasificador como una función que asigna el valor de una clase a un ejemplo:

$$f_{nb}(E) = \frac{p(C = c_1)}{p(C = c_0)} \prod_{i=1}^{n} \frac{p(x_i|C = c_1)}{p(x_i|C = c_0)}.$$

A la función $f_{nb}(E)$ se le conoce como *clasificador Naïve Bayes* o simplemente *Naïve Bayes*. La distribución de probabilidades aprendida puede ser visualizada en forma de árbol. Dado el supuesto de independencia de los atributos, este árbol siempre está formado por la raíz que es la clase padre e inmediatamente por las hojas, las cuales corresponden a los atributos de la base de datos.

Si se relajase el supuesto de independencia, entonces se formaría un grafo con arcos que pueden conectar a los nodos, lo que expresa las dependencias entre los atributos. A este tipo de estructura que generaliza a un clasificador Naïve Bayes se le conoce como red bayesiana. De manera que un clasificador Naïve Bayes es la forma más simple de red bayesiana. Un estudio más detallado de este clasificador se encuentra en (John, Langley, 1995). Asimismo, (Witten, Frank, Hall, 2011) explica con claridad su funcionamiento y características.

Redes bayesianas

Naïve Bayes aprende las estimaciones de la probabilidad de una clase. De hecho, lo que se aprende es la distribución de probabilidad condicional de los valores de una clase dados los valores de los atributos. No obstante, Naïve Bayes solo es capaz de aprender distribuciones de probabilidad condicional muy simples. Así, una alternativa teóricamente bien fundada de representar distribuciones de probabilidad condicional arbitrarias, de una manera concisa y comprensible es la de las redes bayesianas (Witten, Frank, Hall, 2011). Ellas pueden ser representadas como un grafo

Revista Cubana de Ciencias Informáticas
Vol. 11, No. 4, Octubre-Diciembre, 2017
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 67- 84

dirigido acíclico donde existe un nodo para cada atributo, y cada arco es una expresión de la dependencia entre atributos. Formalmente, una red bayesiana se define como sigue (Bouckaert et al., 2015a).

Sea $U = \{x_1, x_2,..., x_n\}, n \geq 1$ un conjunto de atributos. Una red bayesiana B sobre el conjunto U es una estructura de red B_s, la cual es un grafo acíclico sobre U y un conjunto de tablas de probabilidades $B_p=\{p(u|pa(u)|u \in U)\}$, donde $pa(u)$ es el conjunto de padres de u en B_s. Una red bayesiana representa entonces una distribución de probabilidad $P(U) = \Pi_{u \epsilon U}\, p(u|pa(u))$.

La manera principal de construir un algoritmo de aprendizaje para una red bayesiana es la definición de dos componentes: una función para evaluar una red aprendida de los datos y un método para la búsqueda a través de todas las posibles redes (Witten, Frank, Hall, 2011). La calidad de una red dada se mide mediante la probabilidad de los datos dada la red. Para ello se calcula la probabilidad que la red otorga a cada ejemplo y luego se combinan tales probabilidades.

Por su parte, la búsqueda en el espacio de todas las redes posibles implica estimar las tablas de probabilidad condicional y la medida de calidad para cada una de las redes. Esto, como se ha apuntado, es un problema de clase NP-hard que motiva la búsqueda de un método de complejidad computacional mucho menor mientras mantenga la capacidad de representación de las dependencias entre los atributos.

Clasificadores seleccionados

Los clasificadores seleccionados son una muestra de la familia bayesiana (ver Tabla 1). El primero de ellos, y que es objeto de estudio, es el Naïve Bayes de (John, Langley, 1995). Se seleccionaron también cuatro redes bayesianas. Dos de ellas usan una búsqueda heurística para determinar la red y las dos redes restantes, una metaheurística. Por último, se eligió A2DE, un multiclasificador que combina clasificadores bayesianos para obtener mayor precisión en la clasificación, el cual ha demostrado, según sus autores, un buen desempeño dentro de la familia bayesiana, aunque no ha sido evaluado con la metodología estadística de (Demšar, 2006).

Tabla 1. Clasificadores seleccionados para compararse con Naïve Bayes.

Clasificador	Referencia	Descripción
BNK2	(Cooper, Herskovits, 1992)	Red bayesiana que usa el algoritmo K2 para la búsqueda de una red que cumpla con cierto grado de calidad. Consiste en la utilización de *hill climbing* en la cual se asume cierto orden de los atributos.
BNHC	(Bouckaert et al., 2015b)	Variante de K2 en la que no se usa un orden prefijado, conocida como *hill climbing*.
BNT		Variante en la que se usa la metaheurística de búsqueda tabú. De la misma manera que *hill climbing* y K2, se utiliza para buscar la red que maximice la probabilidad dada una base de datos. Lo que le distingue, es su intento de salir de los óptimos locales.

TAN	(Friedman, Geiger, Goldszmidt, 1997)	Aumenta la capacidad de representación de Naïve Bayes mediante una estructura de árbol, sin incrementar demasiado el consumo de tiempo. En el nodo raíz, la clase, apunta a cada nodo; y cada uno de estos, en representación de los atributos, solo tiene un padre aparte de la raíz, el cual es otro nodo no raíz.
A2DE	(Webb et al., 2012)	Modelo que evita, a diferencia de TAN, aprender una estructura mientras mantiene la capacidad de representación de las dependencias. Esto es conseguido por el algoritmo AODE (del inglés *Avaraged One-dependence Estimators*) (Webb, Boughton, Wang, 2005). La implementación usada en este estudio es una versión mejorada de AODE que aparece descrita en el trabajo de (Webb et al. 2012).

Todos los algoritmos seleccionados están disponibles para WEKA, en su versión 3.7.13, la cual es una conocida herramienta para utilizar diferentes algoritmos de aprendizaje automático (Witten, Frank, Hall, 2011). Es además gratuita, extensible con nuevos algoritmos y ampliamente usada. Cuenta con el *Experimenter*, interfaz visual con la cual se diseñan, ejecutan y prueban los experimentos.

En todos los casos, con excepción de BNHC, se usó la configuración por defecto que ofrece WEKA. Con BNHC se optó por relajar la opción que restringe el número posible de padres, con lo cual se ha querido representar una red bayesiana sin límites en las dependencias que se puedan crear. Como se usa la configuración por defecto, en este caso Naïve Bayes estima los atributos numéricos mediante una distribución normal. Además, según se explica en el manual (Bouckaert et al., 2015a), todos los clasificadores basados en redes que están implementados asumen que los atributos son discretos y finitos, y que no existen valores perdidos. Para tratar tales casos, WEKA aplica de manera automática los filtros no supervisados *Discretize* y *ReplaceMissingValues*. El primero de ellos, en su configuración por defecto, transforma los atributos continuos en discretos mediante la subdivisión del dominio en intervalos de igual tamaño, método conocido como *equal-width binning* (Witten, Frank, Hall, 2011). El segundo, reemplaza los valores perdidos con la moda en caso de atributos nominales, y con la media en caso de atributos numéricos (Witten, Frank, Hall, 2011).

Bases de datos seleccionadas

Otro paso para el diseño experimental es la selección de las bases de datos. En el caso del experimento general, 22 de ellas fueron escogidas de manera que exhibieran características distintas según muestra la Tabla 2, con el objetivo de generalizar el estudio y no centrarlo en una situación particular. De la misma manera, se decidió incluir bases de datos con atributos numéricos dado que es esta una situación común cuando clasificadores bayesianos se enfrentan a problemas de la realidad, como se atestigua en (Frías-Blanco et al., 2016; John, Langley, 1995; Webb et al., 2012; Kononenko, 1993; Sebastiani, 2002; Lewis, 1998), a pesar de que están mejor preparados para trabajar sobre dominios discretos. Tres de las bases de datos se obtuvieron artificialmente mediante generadores artificiales

disponibles en WEKA. Para AGD1 y AGD2 se empleó el RDG1; y LED24 para el generador del mismo nombre. Otras dos, *saheart* y *vowel*, proceden del KEEL *Repository*; y el resto, del UCI *Machine Learning Repository* (Lichman, 2013). Debe notarse además que, si bien la mayoría utiliza algún subconjunto procedente de este último repositorio, los conjuntos de las bases de datos seleccionadas en estudios comparativos similares no son iguales ni en número ni en origen. Por tal razón, se ha considerado conformar un conjunto de bases de datos que difiere necesariamente del resto de los demás estudios, aunque existen coincidencias sobre todo en aquellas extraídas del UCI *Machine Learning Repository*. Otro criterio de selección reside en las características apropiadas para evaluar el desempeño de los algoritmos ante la presencia de atributos irrelevantes y ejemplos ruidosos.

Tabla 2. Características de las bases de datos para el experimento general.

Base de Datos	Ejemplos	Atributos	Numéricos	Nominales	Clases	Valores perdidos	Ruido	RD*
AGD1	100	9	0	9	2	No	No	1.9
AGD2	100	9	9	0	2	No	No	1.3
breast-cancer	286	9	0	9	2	Sí	-	2.4
wisconsin-breast-cancer	699	9	9	0	2	Sí	-	1.9
horse-colic	368	22	7	15	2	Sí	-	1.7
credit-rating	690	15	6	9	2	Sí	-	1.2
german_credit	1000	20	7	13	2	No	No	2.3
echocardiogram	132	12	3	9	2	Sí	-	2.1
Glass	214	9	9	0	7	No	-	8.4
haberman	306	3	3	0	2	No	No	2.8
hungarian-14-heart-diseas	294	13	6	7	5	Sí	-	1.8
heart-statlog	270	13	13	0	2	No	No	1.3
hepatitis	155	19	6	13	2	Sí	-	3.8
ionosphere	351	34	34	0	2	No	-	1.8
led24	200	24	24	0	10	No	Sí	1.9
pima	768	8	8	0	2	No	No	1.9
saheart	462	9	8	1	2	No	No	1.9
sonar	208	60	60	0	2	No	-	1.1
tic-tac-toe	768	9	0	9	2	No	No	4.4
vehicle	846	18	18	0	4	No	-	1.1
vote	435	16	0	16	2	Sí	-	1.6
vowel	990	13	10	3	11	No	-	1

* El ratio de desbalance (RD) se calcula como el número de la clase mayoritaria entre el de la clase minoritaria.

Ante la variedad de las bases de datos seleccionadas, el teorema *no free lunch* (Wolpert, 2002) advierte contra la tendencia a buscar el mejor clasificador, en lugar del mejor para su propósito. Ello se debe a que tal algoritmo será significativamente mejor en un dominio específico y peor en otro. Esto no significa en manera alguna que no sea

posible encontrar el de mejor comportamiento en el contexto específico de un conjunto determinado de bases de datos seleccionadas para evaluar la precisión en la clasificación de un grupo de algoritmos bayesianos. Por tanto, cualquier conclusión derivada de los experimentos, debe ser leída en este contexto.

Los dos experimentos restantes, se configuraron para tratar la presencia de atributos irrelevantes y de ejemplos ruidosos. Para el primer caso, se escogieron 10 bases de datos (ver Tabla 3), de las cuales 9 son del UCI *Machine Learning Repository* y una, del KEEL *Repository*. A cada una, basado en el esquema propuesto en (Güvenir, 1998), se le añadieron incrementalmente atributos nominales en cantidades de 5, 10 y 15 atributos, cuyos valores se generaron de forma aleatoria según la distribución uniforme. Para datos ruidosos se seleccionaron 9 bases de datos (ver Tabla 3), de las cuales 8 son del UCI *Machine Learning Repository* y 1 del KEEL *Repository*. A partir de cada una de las bases de datos originales, se obtuvieron dos versiones contaminadas, la primera con el 15% de los ejemplos y la segunda, con el 20%. Para ello, se sustituyó un valor por otro seleccionado aleatoriamente dentro del dominio de cada atributo del ejemplo seleccionado para contaminar. Tal procedimiento se basa en el esquema seguido por (Nettleton, Orriols-Puig, Fornells, 2010).

Tabla 3. Características de las bases de datos para experimentos con atributos irrelevantes y ejemplos ruidosos.

Base de Datos	No. ejemplos	No. atributos	No. Numérico	No. Nominal	No. clases	V. perdidos	Ruido	RD	I*	R*
balance-scale	625	4	4	0	3	No	No	5.9	x	
bupa	345	6	6	0	2	No	No	1.4	x	x
german-credit	1000	20	7	13	2	No	No	2.3		x
haberman	306	3	3	0	2	No	No	2.8	x	
hayes-roth	160	5	0	5	3	No	No	2.1	x	x
heart-statlog	270	13	13	0	2	No	No	1.3	x	x
ionosphere	351	34	34	0	2	No	-	1.8		x
iris	150	4	4	0	3	No	No	1	x	x
pima	768	8	8	0	2	No	No	1.9	x	x
sonar	208	60	60	0	2	No	-	1.1		x
saheart	462	9	8	1	2	No	No	1.9	x	
tic-tac-toe	768	9	0	9	2	No	No	4.4	x	
wdbc	569	30	30	0	2	No	No	1.7		x
wine	178	13	13	0	3	No	No	1.5	x	

* Las marcadas en la columna I son usadas en el experimento con atributos irrelevantes y las en R, en el de ejemplos ruidosos.

Las bases de datos seleccionadas para los experimentos con presencia de atributos irrelevantes y ejemplos ruidosos, se escogieron de manera que cumplieran con los siguientes requerimientos: a lo sumo tres clases; relación número de atributos a número de ejemplos menor que 1:10; ausencia de valores perdidos; no desbalance crítico, es decir, el ratio

de desbalance es menor que 9 (García, Herrera, 2009). Aunque es deseable el predominio de atributos nominales en este contexto, no fueron muchas las bases de datos encontradas con tal característica y que cumplieran con los requerimientos establecidos. Con tales restricciones se intenta descartar cualquier característica que pudiera influir en los resultados de los clasificadores y provocar un sesgo en el problema estudiado.

Evaluación experimental

La metodología para comparar múltiples algoritmos con múltiples bases de datos es la descrita en (Demšar, 2006), y extendida por (García, Herrera, 2008). En ella se propone un conjunto de pruebas estadísticas para escenarios en los que se comparan dos o más algoritmos. En el primer paso se prueba si existen diferencias significativas entre los algoritmos, en caso de lo cual se realizan otras pruebas para conocer entre cuáles específicamente existen tales diferencias. Para ello se usa el paquete SCMAMP del lenguaje R (Calvo, Santafé, 2016). Para correr los experimentos, se configuraron 10 corridas con validación cruzada de 10-fold para cada uno.

Existen varios tipos de estadísticos disponibles para la clasificación. Entre ellos, se destacan dos medidas de la calidad: *Classification Percentage Accuracy* (precisión del porciento de clasificación) y *Receiver Oparating Characteristics* (ROC), también conocida esta última como área bajo la curva ROC (AUC, del inglés *Area Under the Curve*). En (Demšar, 2006) se aboga por el uso de AUC y en (Provost, Fawcett, Kohavi, 1998) también. Además, (Chawla, 2005) afirma que el AUC es una medida de precisión en la clasificación adecuada para casos en que las bases de datos presentan desbalance. Por tales razones, en este estudio se usa AUC como medida del desempeño de la clasificación.

Resultados y discusión

Los resultados obtenidos se dividen para su análisis según los experimentos diseñados. Por lo que se reportan y discuten los resultados del experimento general, el de atributos irrelevantes y luego el de ejemplos ruidosos.

Experimento general

Como resultado se obtuvo una matriz con el promedio del desempeño de cada algoritmo en cada base de datos (ver Tabla 4). En ella se muestra además el promedio del desempeño de cada algoritmo a lo largo de todas las bases de datos, cuyo valor máximo es alcanzado por el clasificador A2DE. Ahora debe comprobarse que las diferencias de precisión en la clasificación basada en la AUC, es significativa.

Para ello se aplica el test de Friedman. Con este test no paramétrico se busca la existencia de al menos dos algoritmos que posean diferencias significativas. Construye además un ranking de los algoritmos en el cual entre mayor es el número, peor es la precisión en la clasificación.

Tabla 4. Promedio del desempeño de cada clasificador por cada base de datos en el experimento general.

Base de Datos	NaiveBayes	TAN	BNK2	BNHC	BNT	A2DE
AGD1	0.83	**0.89**	0.84	0.86	0.81	**0.89**
AGD2	0.83	**0.91**	**0.91**	0.9	**0.91**	**0.91**
breast-cancer	0.7	0.66	0.7	0.64	**0.71**	**0.71**
wisconsin-breast-cancer	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**	**0.99**
horse-colic	0.84	0.86	0.84	**0.87**	0.85	**0.87**
credit-rating	0.9	**0.92**	**0.92**	**0.92**	**0.92**	**0.92**
german_credit	**0.79**	0.77	0.78	0.75	0.76	**0.79**
echocardiogram	0.96	**0.97**	0.96	0.96	0.96	0.96
Glass	0.73	0.87	**0.89**	0.88	**0.89**	0.88
haberman	**0.64**	0.63	0.63	0.63	0.63	0.63
hungarian-14-heart-diseas	0.91	0.91	0.91	0.9	0.91	**0.92**
heart-statlog	0.9	**0.91**	**0.91**	0.9	**0.91**	**0.91**
hepatitis	0.86	0.86	**0.88**	0.81	**0.88**	**0.88**
ionosphere	0.94	**0.97**	0.95	0.96	0.95	**0.97**
led24	0.91	0.83	0.91	**0.92**	0.91	0.87
pima	**0.82**	**0.82**	0.81	**0.82**	0.81	**0.82**
saheart	**0.76**	0.73	0.72	0.73	0.72	0.73
sonar	0.8	**0.87**	0.86	0.84	0.86	**0.87**
tic-tac-toe	0.74	0.82	0.74	0.82	0.75	**0.97**
vehicle	0.7	**0.84**	0.75	0.83	0.75	**0.84**
vote	0.97	**0.99**	0.97	**0.99**	0.97	**0.99**
vowel	0.97	**1**	0.98	**1**	0.99	**1**
Promedio	**0.84**	**0.865**	**0.857**	**0.86**	**0.856**	**0.878**

El resultado calculado del *p-value* de Friedman es 0.006554, lo cual es mucho menor que el nivel de significación α=0.05. Por tanto, es posible afirmar que existen diferencias significativas entre al menos dos algoritmos. Como muestra la Tabla 5, A2DE se comporta como el mejor algoritmo, mientras que Naïve Bayes es el peor en este caso.

Tabla 5. Ranking promediado de los algoritmos para el experimento general.

NaiveBayes	TAN	BNK2	BNHC	BNT	A2DE
4.386	2.977	3.841	3.682	3.750	**2.364**

Dado el resultado de Friedman, se continúa con los test post hoc. Entre todas las posibilidades, se escoge el test de Bergmann y Hommel, considerado el más poderoso aunque computacionalmente costoso (Calvo, Santafé, 2016). Los resultados de este test se ilustran mediante la Figura 1, en la cual se dibuja una línea gruesa horizontal sobre todos

Revista Cubana de Ciencias Informáticas
Vol. 11, No. 4, Octubre-Diciembre, 2017
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 67- 84

aquellos algoritmos que no tienen diferencias significativas entre ellos, a partir de la matriz de valores ajustados de los *p-values* de este test, con referencia al nivel de significación α=0.05. Aunque el test de Friedman ubica a Naïve Bayes como el peor, la Figura 1 muestra que no es significativamente peor que el resto, con excepción de A2DE. Para el resto de los pares de algoritmos, no existen diferencias significativas.

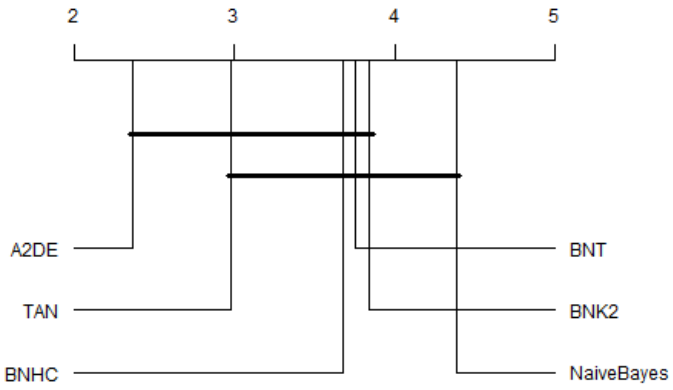

Figura 1. Diferencias significativas según test de Bergmann y Hommel.

Por su parte, los gráficos de caja proveen al investigador de una mejor idea de la distribución de los datos del promedio de las medidas de desempeño de un algoritmo respecto a todas las bases de datos seleccionadas. Por ejemplo, la Figura 2 muestra que, aunque la mediana del desempeño de A2DE es alta, como corresponde al mejor algoritmo, su largo bigote hacia abajo, induce a pensar que existe un sesgo a la izquierda. Naïve Bayes, en contraste, presenta una media menor, mayor variabilidad entre el primer y tercer cuartil, y largos bigotes a ambos extremos. Ello implica una mayor variabilidad en el comportamiento de su precisión en la clasificación.

Experimento con atributos irrelevantes

Para probar la tolerancia ante la presencia de atributos irrelevantes, se realizaron cuatro experimentos. Con cada uno de ellos se indaga por diferencias significativas ante el incremento gradual de los atributos adicionados con valores aleatorios; primero con ningún atributo irrelevante, luego con 5, 10 y por último, 15 atributos irrelevantes. Pero antes, para forjar una idea general del comportamiento de los distintos clasificadores ante el incremento gradual de los atributos irrelevantes, se promedió el AUC de cada uno a lo largo de todas las bases de datos.

Revista Cubana de Ciencias Informáticas
Vol. 11, No. 4, Octubre-Diciembre, 2017
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 67- 84

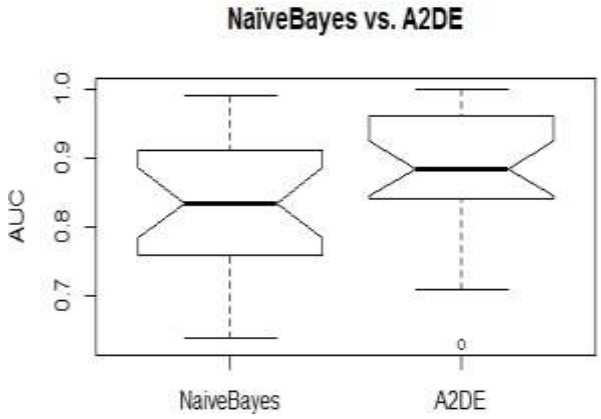

Figura 2. Gráfico de caja para Naïve Bayes y A2DE.

La Figura 3 muestra la degradación del comportamiento de todos los clasificadores en cuanto a la precisión en la clasificación, cuando solo se han añadido cinco atributos irrelevantes. También se aprecia un mejor desempeño medio de Naïve Bayes frente al resto de los clasificadores. Es interesante cómo Naïve Bayes y A2DE, a pesar de mostrar ambos el mejor desempeño medio, son los más sensibles a la presencia de atributos irrelevantes, mientras que el resto mantiene un comportamiento mucho más estable al añadirse los atributos irrelevantes. Con las bases de datos originales (con 0 atributos irrelevantes), se observa que Naïve Bayes tiene el mejor comportamiento, lo cual contrasta con los datos obtenidos en el experimento general. Ello se debe a las restricciones impuestas en las bases de datos seleccionadas, las cuales crean el ambiente propicio para tal resultado.

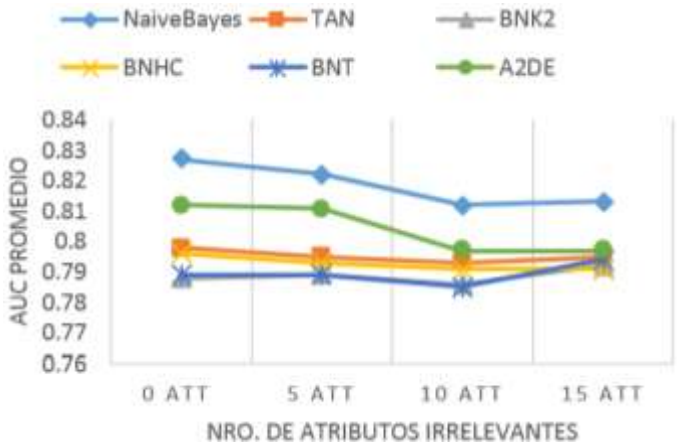

Figura 3. Comportamiento de los clasificadores ante el incremento de atributos irrelevantes.

En los cuatro experimentos el test de Friedman ofreció *p-values* mayores que el nivel de significación especificado (ver Tabla 6). Por tanto, las diferencias observadas en la Figura 3 no son significativas. De todas maneras, los

Revista Cubana de Ciencias Informáticas
Vol. 11, No. 4, Octubre-Diciembre, 2017
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 67- 84

rankings obtenidos mediante el test de Friedman en los cuatro experimentos (ver Tabla 6), corroboran que el mejor desempeño es el de Naïve Bayes en cada incremento de atributos irrelevantes.

Experimento con ejemplos ruidosos

En el experimento para probar la tolerancia a la presencia de ejemplos ruidosos, se escogieron bases de datos que son conocidas por no tener ruido, o por tener poco. Luego, se escogió incrementalmente un porciento de los ejemplos y se contaminó con valores aleatorios de sus dominios. En este caso, los incrementos fueron del 15 y el 20% de ejemplos ruidosos. De la misma manera que en el caso de los atributos irrelevantes, se graficó el comportamiento medio de la precisión en la clasificación para los distintos clasificadores a lo largo de todas las bases de datos, por cada incremento a partir del caso en que se conoce que no hay ruido o es muy poco (ver Figura 4).

Tabla 6. Ranking promediado de los algoritmos para los experimentos con atributos irrelevantes.

***P-values* para los cuatro experimentos con atributos irrelevantes.**						
	Con 0 atribs.	Con 5 atribs.		Con 10 atribs.	Con 15 atribs.	
*Friedman p-value**	0.4554	0.1602		0.5618	0.7088	
Ranking promedio para los experimentos con atributos irrelevantes.						
	NaiveBayes	TAN	BNK2	BNHC	BNT	A2DE
0 atributos irrelevantes	**2.550**	3.250	4.100	3.800	3.950	3.350
5 atributos irrelevantes	**2.500**	3.800	4.000	4.150	3.950	2.600
10 atributos irrelevantes	**2.550**	3.550	3.550	3.950	4.000	3.400
15 atributos irrelevantes	**2.750**	3.600	3.550	4.150	3.350	3.600
***P-values* para los tres experimentos con ejemplos ruidosos**						
	0%.		15%.		20%.	
*Friedman p-value**	0.538		0.2303		0.7729	
Ranking promedio de los algoritmos para los experimentos con ejemplos ruidosos						
	NaiveBayes	TAN	BNK2	BNHC	BNT	A2DE
0%	3.667	2.944	3.722	4.056	3.944	**2.667**
15%	4.167	3.167	3.778	3.889	3.833	**2.167**
20%	3.444	3.500	3.500	4.167	3.611	**2.778**

*El nivel de significación es α=0.05.

A diferencia del caso de los atributos irrelevantes, los clasificadores muestran una degradación similar tanto en el incremento del 15% como en el del 20% de ejemplos ruidosos. Naïve Bayes y A2DE vuelven a destacarse como los mejores en cuanto a la tolerancia ante la presencia de atributos irrelevantes y datos ruidosos, aunque discretamente. Pero de igual manera, Naïve Bayes también se destaca como el mejor cuando las bases de datos originales no han sido contaminadas aún (0% de ruido), efecto causado por idénticas razones. Los *p-values* obtenidos para el test de Friedman, en los tres experimentos que corresponden al caso inicial más los dos incrementos, demuestran que las diferencias no son significativas (ver Tabla 6).

Revista Cubana de Ciencias Informáticas
Vol. 11, No. 4, Octubre-Diciembre, 2017
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 67- 84

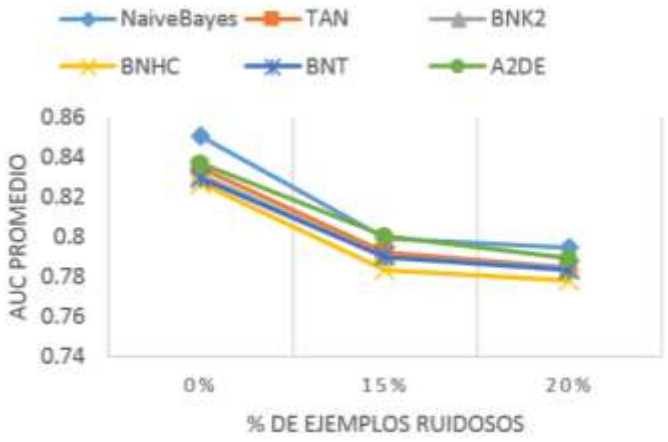

Figura 4. Comportamiento de los clasificadores ante el incremento de ejemplos ruidosos.

En cuanto a los *rankings* obtenidos por el test de Friedman, A2DE muestra el mejor desempeño ante la presencia ejemplos ruidosos en cada uno de los casos contemplados. En contraste, el desempeño de Naïve Bayes es peor, a pesar de mostrar el mejor AUC promedio, como lo ilustra la Figura 4. Esta aparente contradicción se explica por la característica variabilidad que Naïve Bayes mostró en el experimento general.

Conclusiones

Se realizó un estudio experimental sobre el desempeño en la clasificación de 6 clasificadores bayesianos. Para ello se dividió el estudio en un experimento general que vinculó bases de datos con características variadas, otro conjunto de experimentos para explorar la tolerancia al incremento de atributos irrelevantes y un último conjunto de experimentos para determinar el comportamiento de la precisión en la clasificación ante la contaminación incrementada de los ejemplos de las bases de datos.

Para las bases de datos seleccionadas, Naïve Bayes mostró ser competitivo en cuanto a la precisión en la clasificación con el resto de los algoritmos en el experimento general. A2DE, el de mejor desempeño, es la excepción, pues presentó diferencias significativas con Naïve Bayes. Tampoco se encontró evidencia para afirmar que Naïve Bayes es diferente en su desempeño al resto de los clasificadores ante la presencia de atributos irrelevantes. De hecho, muestra la mejor tolerancia. Mostró además el mejor promedio del AUC, pero no es el de mejor *ranking* en el caso de la contaminación con ruido según el test de Friedman. No obstante, las diferencias con el resto de los clasificadores no son significativas. Todo ello es evidencia en su favor, por lo que continúa siendo una opción viable y una alternativa

competitiva a otros clasificadores bayesianos, más aún en situaciones donde la simplicidad sea una característica necesaria.

Referencias

BOUCKAERT, Remco R., FRANK, Eibe, HALL, Mark, KIRKBY, Richard, REUTEMANN, Peter, SEEWALD, Alex y SCUSE, David, 2015a. *WEKA Manual for Version 3-7-13*. 9 octubre 2015. University of Waikato.

BOUCKAERT, Remco R., FRANK, Eibe, HALL, Mark, KIRKBY, Richard, REUTEMANN, Peter, SEEWALD, Alex y SCUSE, David, 2015b. *WEKA Manual for Version 3-7-13*. University of Waikato.

CALVO, Borja y SANTAFÉ, Guzmán, 2016. scmamp: Statistical Comparison of Multiple Algorithms in Multiple Problems. *The R Journal*. 2016. Vol. 8, no. 1, p. 248-256.

CHAWLA, Nitesh V., 2005. Data Mining for Imbalanced Datasets: An Overview. En: *Data Mining and Knowledge Discovery Handbook* [En línea]. Springer US. p. 853-867. [Consultado el 3 junio 2016]. ISBN 978-0-387-24435-8. Disponible en : http://link.springer.com/chapter/10.1007/0-387-25465-X_40

CHENG, Jie y GREINER, Russell, 1999. Comparing Bayesian Network Classifiers. En: *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence* [En línea]. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. 1999. p. 101–108. [Consultado el 23 junio 2016]. UAI'99. ISBN 978-1-55860-614-2. Disponible en : http://dl.acm.org/citation.cfm?id=2073796.2073808

CHICKERING, D. M., 1996. Learning Bayesian networks is NP-Complete. En: *Learning from Data: Artificial Intelligence and Statistics V*. Heidelberg: Springer. p. 121-130.

COOPER, Gregory F. y HERSKOVITS, Edward, 1992. A Bayesian method for the induction of probabilistic networks from data. *Machine learning*. 1992. Vol. 9, no. 4, p. 309–347.

DEMŠAR, Janez, 2006. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*. 2006. Vol. 7, p. 1–30.

FERNÁNDEZ-DELGADO, Manuel, CERNADAS, Eva y BARRO, Senén, 2014. Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? *Journal of Machine Learning Research*. 2014. Vol. 15, p. 3133-3181.

FRÍAS-BLANCO, Isvani, VERDECIA-CABRERA, Alberto, ORTIZ-DÍAZ, Agustín y CARVALHO, Andre, 2016. Fast Adaptive Stacking of Ensembles. En: *Proceedings of the 31st Annual ACM Symposium on Applied Computing* [En línea]. New York, NY, USA: ACM. 2016. p. 929–934. [Consultado el 5 abril 2017]. SAC '16. ISBN 978-1-4503-3739-7. Disponible en : http://doi.acm.org/10.1145/2851613.2851655

FRIEDMAN, Nir, GEIGER, Dan y GOLDSZMIDT, Moises, 1997. Bayesian Network Classifiers. *Machine Learning*. 1997. Vol. 29, no. 2-3, p. 131-163. DOI 10.1023/A:1007465528199.

GARCÍA, Salvador y HERRERA, Francisco, 2008. An extension on 'Statistical comparisons of classifiers over multiple data sets' for all pairwise comparisons. *Journal of Machine Learning Research*. 2008. Vol. 9, p. 2677–2694.

GARCÍA, Salvador y HERRERA, Francisco, 2009. Evolutionary Undersampling for Classification with Imbalanced Datasets: Proposals and Taxonomy. *Evolutionary Computation*. 2009. Vol. 17, no. 3, p. 275–306.

GÜVENIR, H. Altay, 1998. A classification learning algorithm robust to irrelevant features. En: *Artificial Intelligence: Methodology, Systems, and Applications* [En línea]. Springer. p. 281–290. [Consultado el 24 mayo 2016]. Disponible en : http://link.springer.com/10.1007%2FBFb0057452

JIANG, Liangxiao, WANG, Dianhong, CAI, Zhihua y YAN, Xuesong, 2007. Survey of Improving Naive Bayes for Classification. En: *Advanced Data Mining and Applications* [En línea]. Springer Berlin Heidelberg. p. 134-145. Lecture Notes in Computer Science, 4632. [Consultado el 23 junio 2016]. ISBN 978-3-540-73870-1. Disponible en : http://link.springer.com/chapter/10.1007/978-3-540-73871-8_14

JIANG, Liangxiao, ZHANG, Harry y CAI, Zhihua, 2009. A novel Bayes model: Hidden naive Bayes. *IEEE Transactions on Knowledge and Data Engineering*. 2009. Vol. 21, no. 10, p. 1361–1371.

JOHN, George H. y LANGLEY, Pat, 1995. Estimating continuous distributions in Bayesian classifiers. En : *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence* [En línea]. Morgan Kaufmann Publishers Inc. 1995. p. 338–345. [Consultado el 16 septiembre 2016]. Disponible en : http://dl.acm.org/citation.cfm?id=2074196

KONONENKO, Igor, 1993. Inductive and Bayesian learning in medical diagnosis. *Applied Artificial Intelligence an International Journal*. 1993. Vol. 7, no. 4, p. 317–337.

LEWIS, David D., 1998. Naive (Bayes) at forty: The independence assumption in information retrieval. En : *Machine Learning: ECML-98* [En línea]. Springer, Berlin, Heidelberg. 21 abril 1998. p. 4-15. [Consultado el 5 abril 2017]. Disponible en : https://link.springer.com/chapter/10.1007/BFb0026666

LICHMAN, M., 2013. *UCI Machine Learning Repository* [En línea]. 2013. University of California, Irvine, School of Information and Computer Sciences. Disponible en : http://archive.ics.uci.edu/ml

NETTLETON, David F., ORRIOLS-PUIG, Albert y FORNELLS, Albert, 2010. A study of the effect of different types of noise on the precision of supervised learning techniques. *Artificial Intelligence Review*. 1 abril 2010. Vol. 33, no. 4, p. 275-306. DOI 10.1007/s10462-010-9156-z.

PROVOST, Foster J, FAWCETT, Tom y KOHAVI, Ron, 1998. The case against accuracy estimation for comparing induction algorithms. En: *ICML*. 1998. p. 445–453.

RISH, Irina, 2001. An empirical study of the naive Bayes classifier. En: *IJCAI 2001 workshop on empirical methods in artificial intelligence* [En línea]. IBM New York. 2001. p. 41–46. [Consultado el 23 junio 2016]. Disponible en : https://www.researchgate.net/profile/Irina_Rish/publication/228845263_An_Empirical_Study_of_the_naive_Bayes_Classifier/links/00b7d52dc3ccd8d692000000.pdf

SEBASTIANI, Fabrizio, 2002. Machine Learning in Automated Text Categorization. *ACM Comput. Surv.* marzo 2002. Vol. 34, no. 1, p. 1–47. DOI 10.1145/505282.505283.

WEBB, Geoffrey I., BOUGHTON, Janice R. y WANG, Zhihai, 2005. Not so naive Bayes: aggregating one-dependence estimators. *Machine learning*. 2005. Vol. 58, no. 1, p. 5–24.

WEBB, Geoffrey I., BOUGHTON, Janice R., ZHENG, Fei, TING, Kai Ming y SALEM, Houssam, 2012. Learning by extrapolation from marginal to full-multivariate probability distributions: decreasingly naive Bayesian classification. *Machine Learning*. 2012. Vol. 86, no. 2, p. 233–272.

WITTEN, Ian H, FRANK, Eibe y HALL, Mark A, 2011. *Data Mining: Practical Machine Learning Tools and Techniques*. 3. USA: Elsevier.

WOLPERT, David H., 2002. The Supervised Learning No-Free-Lunch Theorems. En: *Soft Computing and Industry* [En línea]. Springer London. p. 25-42. [Consultado el 7 diciembre 2016]. ISBN 978-1-4471-1101-6. Disponible en : http://link.springer.com/chapter/10.1007/978-1-4471-0123-9_3

WU, Xindong, KUMAR, Vipin, QUINLAN, J. Ross, GHOSH, Joydeep, YANG, Qiang, MOTODA, Hiroshi, MCLACHLAN, Geoffrey J., NG, Angus, LIU, Bing, YU, Philip S., ZHOU, Zhi-Hua, STEINBACH, Michael, HAND, David J. y STEINBERG, Dan, 2008. Top 10 algorithms in data mining. *Knowledge and Information Systems*. 2008. Vol. 14, p. 1-37. DOI 10.1007/s10115-007-0114-2.

ZHANG, Harry, JIANG, Liangxiao y SU, Jiang, 2005. Hidden naive bayes. En: *AAAI* [En línea]. 2005. p. 919–924. [Consultado el 16 septiembre 2016]. Disponible en : http://www.aaai.org/Papers/AAAI/2005/AAAI05-145.pdf

ZHANG, Harry, 2004. The optimality of naive Bayes. *AA*. 2004. Vol. 1, no. 2, p. 6.