CLASIFICADOR BAYESIANO INGENUO EN RAPIDMINER

María de Lourdes Sandoval Solís[1], Imelda Hernández Baez[1]
[1]Benemérita Universidad Autónoma de Puebla,
maria.sandoval@correo.buap.mx, hebi.1305@gmail.com.mx

RESUMEN
En este trabajo se utiliza el Clasificador Bayesiano Ingenuo (CBI), incluido en el software libre de Minería de Datos llamado *RapidMiner*, para probar su eficiencia en el entrenamiento y predicción de bases de datos reales y académicas.

Se inicia mostrando la importancia de la clasificación en las distintas áreas de conocimiento, y se presenta el CBI, sus principales características, parámetros, así como su sencillez y eficiencia en la clasificación supervisada.
Se analiza también una variante del clasificador, llamado CBI-Kernel, que mantiene las ventajas del CBI pero además se puede usar en los casos en que los datos no sigan una distribución normal.

Se realizan pruebas con ambos operadores para evaluar su eficiencia, usando la base de datos PIMA-INDIANS-diabetes. El objetivo es entrenar al operador con la base de datos completa y con diferentes tamaños de muestra, para luego probar su desempeño al realizar la clasificación con nuevos datos. Se utilizan las opciones disponibles en *RapidMiner* para ambos operadores y se realiza una combinación con los diferentes parámetros para cada caso.

1. INTRODUCCIÓN

Clasificar cosas es parte de la vida desde que se es pequeño. Este proceso está implícito en muchas de nuestras actividades cotidianas, se clasifica la fruta como *verde* o *madura*, un auto como *último modelo* o *clásico*, el médico clasifica a los pacientes con base en ciertos estudios o valoraciones físicas como *apto* o *no apto* para realizar una cirugía, etc. Catalogar objetos en distintas clases, a partir de un criterio determinado, es sumamente común, y muchas veces necesario.

Hoy en día, con las tecnologías de la información relacionadas casi a todos los aspectos de la vida diaria, se puede tener acceso a grandes cantidades de información que guardan las características más comunes para clasificar objetos. Por ejemplo, una universidad puede saber a qué categoría pertenece un alumno con base en algunos atributos especiales que le solicita, y/o que va observando y almacenando a lo largo de su vida estudiantil. Un alumno puede ser apto para otorgarle una beca o puede ser candidato para estudiar un posgrado, o en definitiva se sabe que no estará en la universidad en el próximo periodo.

Debido a esto, se necesitan herramientas que faciliten el proceso de clasificación de grandes cantidades de información, y que sea relativamente fácil catalogar personas u objetos con base en ciertos criterios. Una propuesta es el Clasificador Bayesiano Ingenuo (CBI), conocido también como *Naive Bayes*, que toma las características de cada objeto y supone que todas ellas son independientes entre sí y no afectan en la clasificación, además sólo requiere una pequeña cantidad de datos de entrenamiento para lograr un resultado exitoso.

Existen algunas aplicaciones para utilizar el CBI, una alternativa es *RapidMiner*, una aplicación de software libre, que tiene una interfaz sencilla y ofrece una gran cantidad de operadores no sólo para clasificación, sino para otras técnicas de análisis y minería de datos.

2. TEORÍA

Desde un enfoque bayesiano, el problema de clasificación supervisada consiste en asignar a un objeto descrito por un conjunto de atributos o características, $X_1, X_2, \dots, X_n$, a una de *m* clases posibles, $c_1, c_2, \dots, c_m$, tal que la probabilidad de la clase dados los atributos se maximiza: [1]

$$Arg_C[MaxP(C|X_1, X_2, \dots, X_n)] \quad (1)$$

Si se denota al conjunto de atributos como: $X = \{X_1, X_2, \dots, X_n\}$, la ecuación (1) se puede escribir como: $ArgC[MaxP(C|X)]$. La formulación del clasificador bayesiano se basa en utilizar la regla de Bayes para calcular la probabilidad posterior de la clase, dados los atributos:

$$P(C|X_1, X_2, \dots, X_n) = \frac{P(C)P(X_1, X_2, \dots, X_n|C)}{P(X_1, X_2, \dots, X_n)} \quad (2)$$

Que se puede escribir de la forma:

$$P(C|X) = \frac{P(C)P(X|C)}{P(X)} \quad (3)$$

Entonces el problema de clasificación basado en la ecuación (3) se puede expresar como:

$$Arg_C\left[Max\left[P(C|X) = \frac{P(C)P(X|C)}{P(X)}\right]\right] \quad (4)$$

El denominador $P(X)$, no varía para las diferentes clases, por lo que se puede considerar como una constante si lo que interesa es maximizar la probabilidad de la clase:

$$Arg_C\big[Max[P(C|X) = \alpha P(C)P(X|C)]\big] \quad (5)$$

Para resolver un problema de clasificación bajo el enfoque bayesiano, se requiere la probabilidad *a priori* de cada clase, $P(C)$, y la probabilidad de los atributos dada la clase, $P(X|C)$, conocida como *verosimilitud*; para obtener la probabilidad *posterior* $P(C|A)$. En términos comunes, la ecuación se puede expresar como:

$$posterior = \frac{a\ priori * verosimilitud}{evidencia}$$

Entonces, para que este clasificador aprenda de un conjunto de datos, se requiere estimar estas probabilidades, a priori y verosimilitud, a partir de los datos, conocidos como los parámetros del clasificador.

La aplicación directa de la ecuación (5), resulta en un sistema muy complejo al implementarlo en una computadora, ya que el término $P(X_1, X_2, \dots, X_n|C)$, incrementa exponencialmente de tamaño en función del número de atributos; resultando en un requerimiento muy alto de memoria para almacenarlo, y también el número de operaciones para calcular la probabilidad crece significativamente. Una alternativa es considerar relaciones de independencia mediante lo que se conoce como el clasificador bayesiano simple, también conocido como Clasificador Bayesiano Ingenuo.

CLASIFICADOR BAYESIANO INGENUO

El Clasificador Bayesiano Ingenuo (CBI) se basa en la suposición de que todos los atributos son independientes dada la clase; esto es, cada atributo X_i es condicionalmente *independiente* de los demás atributos dada la clase: $P(X_1|X_j, C) = P(X_i|C), \forall j \neq i$. Considerando esto, la ecuación (2) se puede escribir como:

$$P(C \mid X\_1, X\_2, \dots, X\_N) = (P(C)P(X\_1 \mid C)P(X\_2 \mid C) \dots P(X\_n \mid C)/(P(X)) \quad (6)$$

Donde $P(X)$ se puede considerar como una constante de normalización.

Para que un CBI aprenda se requiere la probabilidad previa de cada clase, $P(C)$, y la probabilidad condicional de cada atributo dada la clase, $P(X_i|C)$. Estas probabilidades se pueden obtener mediante estimados subjetivos de expertos en el área, o a partir de datos mediante máxima verosimilitud (consiste en que las probabilidades se aproximan por las estadísticas de los datos).

CLASIFICADOR BAYESIANO INGENUO - KERNEL

Una variante del CBI es el Clasificador Bayesiano Ingenuo Kernel (CBI-Kernel), que mantiene las ventajas del CBI y además se puede aplicar en situaciones donde los datos no siguen una distribución normal.

Un **kernel** es una función de peso usada en técnicas de estimación no paramétrica. Los kernels son usados en Estimación de Densidad de Kernel (KDE, por sus siglas en inglés), para la estimación de la función de densidad de una variable aleatoria. [3]

Definición: Sea $(x_1, x_2, \dots, x_n)$ una muestra independiente e idénticamente distribuida trazada desde alguna distribución con una densidad no conocida f. Se está interesado en estimar la forma de esa función f. El **estimador de densidad de kernel** es: [4]

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x - x_i}{h}) \quad (7)$$

Donde $K(.)$ es el kernel – una función no negativa que se integra a uno y tiene media igual a cero– y $h > 0$ es un parámetro de suavizado llamado *ancho de banda*. Intuitivamente, se desea elegir h lo más pequeña posible como los datos lo permitan ($h \to 0$), para poder asegurar que $\hat{f}_h$ tiende a la verdadera densidad f de las variables x_i.
La elección correcta del parámetro h es el problema más difícil en la estimación no paramétrica. Si se elige demasiado pequeño, el estimador aparece "infrasuavizado", e incorpora demasiado "ruido", reflejado en la presencia de muchas modas, que no aparecen en la densidad que se desea estimar. Por el contrario, si h se elige demasiado grande, se da el fenómeno contrario de "sobresuavizado" y el estimador es casi insensible a los datos. [5]
Existen varias formas que permiten asignar h de manera óptima. Si se usan funciones Gaussianas para aproximar datos univariados, y la densidad subyacente es Gaussiana, la elección óptima de h es:

$$h = \left(\frac{4\hat{\sigma}^5}{3n}\right)^{\frac{1}{5}} \approx 1.06\hat{\sigma}n^{-\frac{1}{5}} \quad (8)$$

donde $\hat{\sigma}$ es la desviación estándar de la muestra. Esta se denomina *aproximación Gaussiana* o la *Regla de Thumb de Silverman*.

3. PARTE EXPERIMENTAL

La base de datos PIMA-INDIANS-DIABETES [2] consta de 768 entradas que representan pacientes de sexo femenino, de al menos 21 años de edad, a quienes se les realizaron pruebas para determinar si mostraban síntomas de diabetes mellitus, de acuerdo con el criterio de la OMS. Se definen 2 clases: 1, "positivo para diabetes" y 0, "negativo para diabetes". Los 8 atributos que se tomaron como referencia son:

1. Número de embarazos
2. Concentración de glucosa en sangre
3. Presión diastólica (mm Hg)
4. Espesor del pliegue cutáneo del tríceps (mm)
5. Insulina en suero (2 horas) (mu U/ml)
6. Índice de masa corporal
7. Función Pedigree Diabetes
8. Edad

Se probó el clasificador Bayesiano Ingenuo en RapidMiner [3], y se obtuvo lo siguiente:

Simple Distribution

```
Distribution model for label attribute clase
Class TRUE (0.349)
8 distributions
Class FALSE (0.651)
8 distributions
```

La tabla de distribuciones se muestra a continuación, en la figura 1:

ATRIBUTO	PARÁMETRO	TRUE	FALSE
No. Emb	mean	4.865671642	3.298
No. Emb	standard deviation	3.741239044	3.017184583
Plasma	mean	141.2574627	109.98
Plasma	standard deviation	31.93962206	26.14119976
Pre_Dias	mean	70.82462687	68.184
Pre_Dias	standard deviation	21.49181165	18.06307541
espesor_tric	mean	22.1641791	19.664
espesor_tric	standard deviation	17.6797114	14.88994711
2h-serum	mean	100.3358209	68.792
2h-serum	standard deviation	138.6891247	98.86528929
IMC	mean	35.14253731	30.3042
IMC	standard deviation	7.262967242	7.689855012
Diab_ped_func	mean	0.5505	0.429734
Diab_ped_func	standard deviation	0.372354484	0.299085304
edad	mean	37.06716418	31.19
edad	standard deviation	10.96825365	11.66765479

Figura 1. Tabla de distribución para los atributos de PIMA-INDIANS-Diabetes

Posteriormente se ejecutó el clasificador con algunos datos de entrenamiento de diferentes tamaños, y se probó su eficiencia en la predicción.

Es importante mencionar que la elección del conjunto de entrenamiento es muy importante para lograr una clasificación exitosa. Se debe considerar mantener la misma proporción de clases que en la base de datos original, pues de no ser así, el desempeño del clasificador se ve afectado, aumentando el número de errores en la predicción.
Para este caso, la proporción de clases en la base de datos original es de 35% "positivo" y 65% "negativo".
La figura 2 muestra la tabla resumen de las pruebas realizadas.

PRUEBA	TAMAÑO MUESTRA	PROPORCIÓN		DISTRIBUCIÓN		NÚMERO ERRORES	PORCENTAJE ERROR
		Positivo	Negativo	Positivo	Negativo		
1	50	19	31	0.38	0.62	206	26.82%
2	100	44	56	0.44	0.56	196	25.52%
3	200	68	132	0.34	0.66	179	23.31%
4	300	115	185	0.383	0.617	190	24.74%
5	300	105	195	0.35	0.65	181	23.57%
6	400	146	254	0.365	0.635	202	26.30%
7	400	140	260	0.35	0.65	200	26.04%
8	**50**	**25**	**25**	**0.5**	**0.5**	**226**	**29.43%**
9	**100**	**65**	**35**	**0.65**	**0.35**	**235**	**30.60%**

Figura 2. Resultados de las pruebas para PIMA

Se puede notar que, en general, el desempeño del clasificador es bueno. Las pruebas **3** y **5** dan los mejores resultados, cuando la muestra es de tamaño **200** y **300**, respectivamente; además la distribución de clases, se mantiene como la original. Aquí el número de errores es sólo de 179, teniendo una eficiencia de casi 77%.

Se observa en las pruebas **8** y **9**, que el porcentaje de error aumenta hasta el 30%, aquí no importa el tamaño de la muestra de entrenamiento, sino que no se mantuvo la proporción de clases que en la base original.

Se probó también el CBI-Kernel con el ancho de banda heurístico y óptimo para cada atributo. Se tomó como entrenamiento una parte de la base de datos completa, de diversos tamaños, y manteniendo la proporción de clases. La tabla resumen se muestra en la figura 3.

Prueba	Tamaño muestra	h	No. Errores	% de Error
1	100	heurística	211	27%
2	100	13.5	203	26%
3	100	128	261	34%
4	100	77.4	255	33%
5	100	64	241	31%
6	100	461	768	100%
7	100	31.5	198	26%
8	100	1.3	234	30%
9	100	47	220	29%
10	**200**	**heurística**	**176**	**23%**
11	200	64	255	33%
12	**300**	**heurística**	**177**	**23%**
13	300	77.4	254	33%
14	400	heurística	186	24%
15	400	13.5	193	25%

Figura 3. Tabla resumen con las pruebas usando CBI-Kernel

Observe que las pruebas 10 y 12 son las que tuvieron mejores resultados, con sólo un 23% de error. Elegir la opción heurística para el valor de h es la mejor opción; el usuario no debe preocuparse por el cálculo del ancho de banda, h. Mientras que en la prueba 6 la elección de $h = 461$, que es el valor óptimo para el atributo "insulina en suero", no tuvo éxito en la clasificación.

Se presentan a continuación las gráficas de los atributos más representativos en las que se observa la diferencia en la densidad de los datos. La línea azul representa la clase "positivo para diabetes", la línea en color rojo representa la clase "negativo para diabetes".

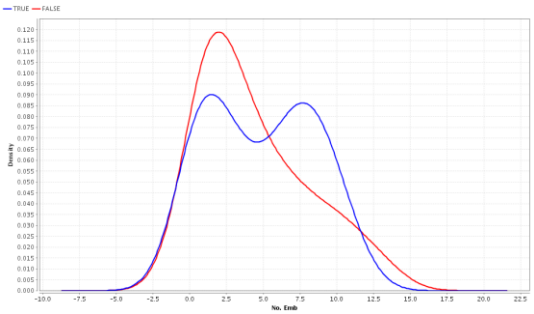

Figura 4. a) Atributo *No. Embarazos*

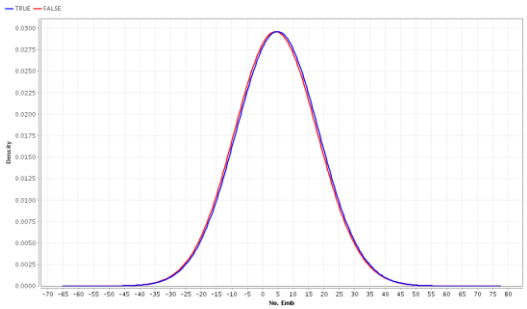

Figura 4. b) *No. Embarazos* con $h = 13.5$

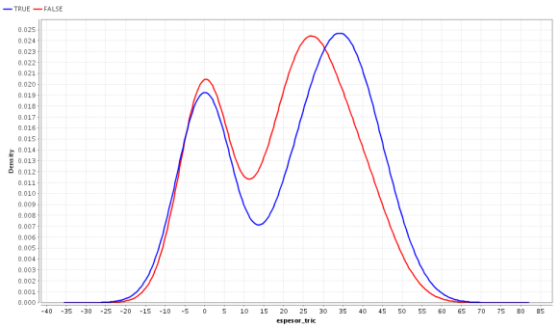

Figura 5. a) Atributo *espesor del tríceps*

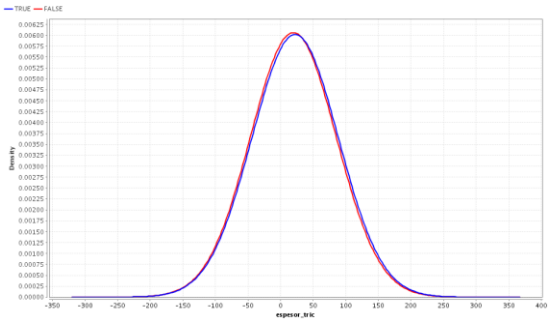

Figura 5. b) *espesor del tríceps* con $h = 64$

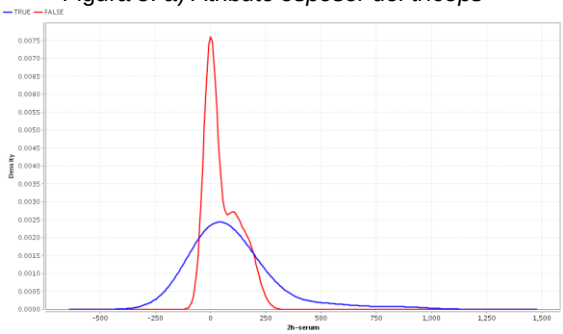

Figura 6. a) Atributo insulina en suero

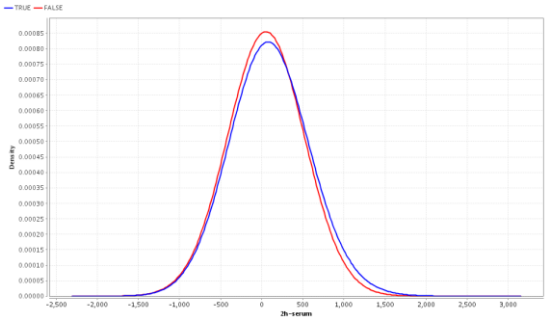

Figura 6. b) insulina en suero con $h = 461$

4. CONCLUSIONES

La clasificación está presente en muchos ámbitos de la vida diaria. Es importante contar con herramientas que faciliten este proceso cuando se cuenta con una gran cantidad de información que no es fácil de analizar.

El Clasificador Bayesiano Ingenuo y su variante Kernel, son dos algoritmos que funcionan adecuadamente para obtener una clasificación exitosa. La suposición de independencia de sus atributos los hace unos de los más sencillos clasificadores disponibles.

Estos clasificadores están presentes en algunas aplicaciones, una de ellas es el software *RapidMiner*, que ofrece estos operadores junto con una gran cantidad de herramientas para el análisis y minería de datos. Además de tener una interfaz muy sencilla, el resultado obtenido es fácil de interpretar, incluso para usuarios que no son expertos en el área.

El CBI clasifica los datos con una eficiencia considerable. No se necesitan conocimientos ni atributos especiales para su uso.
El Clasificador Bayesiano Ingenuo Kernel, funciona sin problema cuando se conoce el valor óptimo de la variable *h*, ancho de banda de los datos, y el número de kernels, que casi siempre está asociado con *h*.

En los casos en que no sea posible calcular el valor de h, este operador ofrece la ventaja de usarse en el modo de estimación *completo* y usar un ancho de banda *heurístico*, esto ofrece al usuario final una experiencia más agradable, ya que no se necesita calcular un ancho de banda ni saber cuántos kernels son apropiados para sus datos.

Cuando los datos de entrada no siguen una distribución normal, el CBI-Kernel aproxima estos datos con una función kernel que mejora la densidad y logra ajustarlos normalmente, por ende mejora la clasificación final.

BIBLIOGRAFÍA

1. Rich, Irina. "An Empirical Study of the naive Bayes Classifier". IBM Research Division. http://www.research.ibm.com/people/r/rish/papers/RC22230.pdf.
   Consultado el 25 de enero de 2016.

2. UCI Machine Learning Repository.
   http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes
   Consultado el 17 de febrero de 2016.

3. http://rapidminer.com/documentation/

4. Rosenblatt, M. (1956). "Remarks on Some Nonparametric Estimates of a Density Function". The Annals of Mathematical Statistics 27 (3): 832.
   doi:10.1214/aoms/1177728190
   Consultado el 29 de marzo de 2016.

5. Cuevas, Antonio. "El análisis estadístico de grandes masas de datos: algunas tendencias recientes". Departamento de Matemáticas. Universidad Autónoma de Madrid.
   http://www.mat.ucm.es/~rrdelrio/documentos/acuevas.pdf.
   Consultado el 29 de marzo de 2016