Matemáticas aplicadas en la predicción en medicina

por

Dinora A. Morales Vega, Universidad Politécnica de Madrid

En este artículo revisaremos el paradigma naive Bayes aplicado en algoritmos computacionales cuya finalidad es aprender un modelo a partir de un conjunto de datos de entrenamiento, el cual contiene información del paciente y conocimiento del médico, para posteriormente predecir un caso nuevo en una de las posibles categorías del diagnóstico médico.

1. Introducción

Iniciaremos con un poco de historia de cómo surge un interés de estudio común entre las ciencias médicas y la ciencia de la computación. El campo de la ciencia en medicina es muy amplio, de suma importancia para la salud de la humanidad. El médico, como profesional de la salud, constantemente analiza los signos y síntomas, además de diferentes pruebas médicas (análisis clínicos, estudios por imágenes, etc.) para llevar a cabo el diagnóstico y posteriormente determinar el tratamiento y el seguimiento del paciente. Toda esta información es recogida y ordenada en la historia clínica, generándose grandes volúmenes de datos tanto en clínicas como en hospitales. Desde los años setenta, las comunidades científicas de la medicina y de la ciencia de la computación vieron crecer su interés hacia el campo científico conocido como inteligencia artificial en medicina, como la oportunidad de integrar nuevas herramientas computacionales para almacenar y organizar la información inicialmente recogida en la historia clínica del paciente, la extracción de conocimiento necesario para consultar en un caso difícil así como la sugerencia de un diagnóstico y pronóstico adecuados además de la sugerencia de decisiones

terapéuticas y también ayuda en la toma de decisiones [24].

Una de las aplicaciones más importantes de los sistemas de apoyo a la decisión en el campo médico es ayudar a encontrar respuestas a preguntas como las siguientes:

¿Cómo se puede aprender de la experiencia? Es decir, a partir del conocimiento que adquiere el doctor en medicina a través del tiempo y el número de pacientes aumenta.

En el diagnóstico, supuesto que un paciente presenta un conjunto de síntomas, ¿Cómo se decide? ¿Cuál es la enfermedad más probable?

¿Qué modelos pueden utilizarse para describir las relaciones entre los síntomas y las enfermedades?

¿Cuáles son las relaciones entre un conjunto de enfermedades (normalmente no observable) y un conjunto (observable) de síntomas?

¿Cuál es la contribución de cada uno de los síntomas o pruebas a la toma de decisión?

En la década de los 70 se desarrollaron varios sistemas basados en reglas, siendo la lógica clásica el motor de inferencia como la forma predominante de razonamiento, por ejemplo en 1972 se desarrollo un sistema basado en reglas, MYCIN, para el diagnóstico y tratamiento de infecciones bacterianas en sangre y posteriormente extendido a otras infecciones [25], en 1978 se desarrollo un sistema de red de asociación-causa llamado CASNET, para el diagnóstico y programa terapéutico del glaucoma y otras enfermedades relacionadas con los ojos [27,28], entre otros. Con el paso del tiempo se han aplicado diversas funciones matemáticas (e.j. regla de Bayes, regresión logística, lógica, etc.) en las distintas etapas de la construcción de un modelo capaz de predecir un nuevo caso en base a los casos aprendidos en forma de ejemplos. En Kononenko, 2001, presenta un estudio del estado del arte de la minería de datos para el diagnóstico médico [14]. Babita y Mishra, 2009, presentan una revisión de sistemas basados en el conocimiento, redes neuronales, modelos de computación evolutiva y sistemas difusos, los cuales requieren de heurísticas y lógica en el razonamiento aplicados al diagnóstico, planeación y tratamiento hasta nuestros días [21].

La pregunta ahora es: ¿Cómo se aprende un modelo matemático a partir de datos? Para ilustrar la respuesta partiremos de la minería de datos, la cual forma parte de la inteligencia artificial. Witten y Frank, 2000, definen la minería de datos como el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos [29]. Desde el punto de vista de la minería de datos es posible descubrir patrones de forma automática o semiautomática a partir de grandes volúmenes de datos, para ello se requiere una serie de pasos: 1) recopilación y almacenamiento de la información, 2) filtrado de la información, 3) extracción de la información para

llevar a cabo el estudio, 4) modelado matemático y 5) toma de decisiones.

La Figura 1 muestra el proceso de la minería de datos. Los pasos del 1 al 3 pertenecen al preproceso de los datos, el paso 4 se refiere a la construcción de la estructura del modelo y el aprendizaje de los parámetros en el caso de los modelos paramétricos y el paso 5 se refiere a la clasificación o predicción de un caso no visto, para ser catalogado en uno de los posibles diagnósticos.

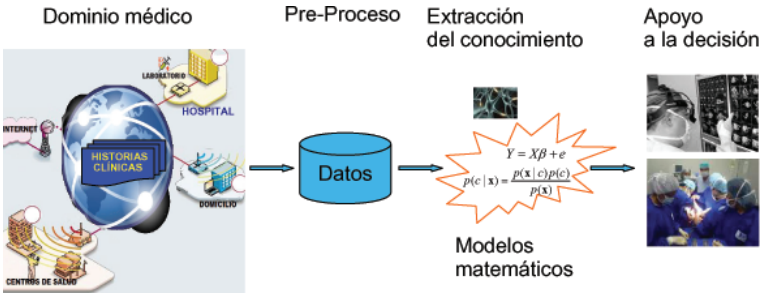

Figura 1: Proceso de la minería de datos. Preproceso de datos (pasos 1-3); Extracción del conocimento (paso 4); toma de decisiones (paso 5).

2. Aprendizaje automático

Para definir "la habilidad de aprendizaje" nos situaremos en el contexto del aprendizaje máquina definido por Mitchell, 1997, como: "Un programa de ordenador se dice que aprende desde la experiencia E con respecto a alguna clase de tareas T y medida de desempeño P, sí mejora su desempeño en T, con respecto a la medida P, basado en la experiencia E" [16]. Russell, 2002, menciona que "[...] cualquier sistema que se considere "inteligente" debería poseer la habilidad de aprender, es decir, mejorar automáticamente con la experiencia" [23], en este contexto la experiencia es el conocimiento que se extrae de los datos.

Nosotros nos centraremos en el aprendizaje automático cuyo objetivo es el desarrollo de paradigmas o modelos basados en un conjunto de relaciones matemáticas (relaciones lógicas, ecuaciones, etc.) que representen las relaciones existentes de la información suministrada en forma de ejemplos, representando de esta forma el conocimiento del experto. Este tipo de aprendizaje se conoce como aprendizaje estructural. Cuando se agregan datos o con la inclusión de una nueva regla en la base de conocimiento, se necesita aprender una nueva estructura del modelo que incluya este nuevo conocimiento. Existen diferentes tipos de paradigmas o modelos provenientes tanto de la estadística como del aprendizaje automático, como son los *modelos descriptivos* y los *modelos predictivos*.

Los modelos descriptivos identifican los patrones que explican los datos, ejemplo de ello son los modelos basados en reglas de asociación [1], los cuales encuentran las

relaciones entre variables, determinando patrones de comportamiento. Un segundo tipo de modelos son los basados en conglomerados (clusters), los cuales agrupan los casos homogéneos maximizando la similitud entre los casos de un conglomerado y minimizando la similitud entre los distintos conglomerados formados, ejemplo de ello el algoritmo EM [5].

El caso que nos ocupa, la predicción en medicina, conlleva un grado de incertidumbre relacionada con la información asociada a cada paciente (subjetividad, imprecisión, ausencia de información, errores, datos ausentes, etc.). Existen los modelos gráficos probabilísticos, como son las redes Bayesianas [22] que incorporan la incertidumbre inherente al problema.

Los modelos predictivos, como los modelos de clasificación supervisada, análisis discriminante y regresión logística, calculan los valores de las variables a predecir a partir de valores conocidos de otras variables predictoras. Estos modelos se caracterizan por ser modelos generativos o discriminativos. Modelos como el análisis discriminante [7] o el clasificador naive Bayes [15] son modelos generativos. Estos modelos aprenden el modelo a partir de la probabilidad conjunta de las variables predictoras $\mathbf{X} = (X_1, \ldots, X_n)$ y la variable clase C. A continuación el clasificador predice la clase de una nueva instancia $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ usando la regla de Bayes para calcular la probabilidad a posteriori de la variable clase dados los valores de las variables predictoras $P(c|\mathbf{x})$.

Los modelos de regresión logística [12] son modelos discriminativos, los cuales estiman los valores de las variables continuas a predecir en función de variables predictoras. Los modelos discriminativos obtienen sus parámetros modelando directamente la distribución de la clase dadas las variables predictoras, $P(c|\mathbf{x})$. En este caso los parámetros del modelo de clasificación pueden ser obtenidos maximizando el logaritmo de la función log-verosimilitud condicionada.

2.1. Clasificación supervisada

En este trabajo nos enfocaremos a revisar modelos predictivos de clasificación supervisada aplicados al diagnóstico en medicina. Primero formularemos el problema de clasificación supervisada y posteriormente recordaremos el teorema de Bayes.

En el caso de la clasificación supervisada se parte de un conjunto de casos N descritos por un vector de características $\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ etiquetado con un valor de la clase C. La clase real es denotada por $c_i \in C$ donde $\Omega_c = \{c_1, c_2, \ldots, c_r\}$ denota todos los posibles valores de la clase. A este conjunto de casos se le denomina conjunto de entrenamiento o conjunto de aprendizaje. La Tabla 1 muestra un ejemplo de un fichero de casos ordenados para la clasificación supervisada.

En base al conjunto de entrenamiento, en la clasificación supervisada se induce

	X_1	...	X_n	C
1	x_1^1	...	x_1^n	c_1
2	x_2^1	...	x_2^n	c_2
...		...		
N	x_n^1	...	x_N^n	c_N

Tabla 1: Problema de clasificación supervisada

un modelo o una función general que asigne un valor del conjunto de la variable clase C a una nueva instancia o caso $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$.

Un clasificador puede verse como una función, γ, la cual asigna la etiqueta correspondiente a cada una de las instancias:

$$\gamma : (x_1, x_2, \ldots, x_n) \to \{c_1, c_2, \ldots, c_r\}.$$

2.2. Teorema de Bayes

Thomas Bayes (Londres, 1702 - Tunbridge, Kent, 1761). El padre de Thomas, Josua Bayes fue unos de los seis primeros reverendos protestantes ordenados en Inglaterra. Thomas, recibió educación de forma privada, se desconoce el nombre de su tutor aunque algunas investigaciones apuntan a que este pudiera haber sido de Moivre. Se conoce que en 1719 estuvo matriculado en la Universidad de Edinburgo donde estudió Lógica y Teología.

En los registros de la Universidad de Edinburgo consta que el 14 de enero de 1721 pronunció la homilía "Mateo Capítulo 7 versos 24-27". Fue ordenado y comenzó como ayudante de su padre, párroco también, en Holborn y cerca de 1933 fue nombrado pastor en la capilla presbiteriana de Tunbridge Wells, Kent al sureste de Londres. Abandonó los hábitos en 1752 y continuó viviendo en la misma ciudad.

Publicó sus teorías en un artículo titulado Essay towards solving a problem in the doctrine of chances, publicado por The philosophical Transactions of the Royal Society of London en 1764 Tunbridge Wells.

Los trabajos que se sabe que Thomas Bayes publicó en vida son: *Divine Providence and Government Is the Happiness of His Creatures* (1731) y *An Introduction to the Doctrine of Fluxions*, and a Defence of The Analyst (1736), que fueron blanco de críticas por parte del obispo Berkeley, quien sustentaba sus ideas en los fundamentos lógicos del cálculo de Newton. En 1763 se publicó póstumamente Essay Towards Solving a Problem in the Doctrine of Chances, donde el reverendo Bayes abordó el problema de las causas a través de los efectos observados, y donde se enuncia el teorema que lleva su nombre. Este trabajo fue entregado a la Royal Society por Richard Price y publicado en Philosophical Transactions of the Royal Society, 53, 370-418.

Las conclusiones presentadas por Bayes fueron aprobadas por Laplace en una memoria de 1781, redescubiertas por Condorcet y permanecieron inalterables hasta que Boole las cuestionó en su libro Laws of Thought (1854). Desde aquel momento las técnicas de Bayes vienen siendo sometidas a controversia habiéndose creado una familia de estadísticos que las apoyan fielmente, los denominados Bayesianos.

Bayes fue miembro de la Royal Society desde 1742 y el primero en utilizar la probabilidad inductivamente y establecer una base matemática para la inferencia probabilística (la manera de calcular, a partir de la frecuencia con la que un acontecimiento ocurrió, la probabilidad de que ocurrirá en el futuro) [2].

A continuación se presenta el teorema de Bayes en su formulación de sucesos, para posteriormente formularlo para variables aleatorias.

Sean A y B dos sucesos aleatorios cuyas probabilidades se denotan por $p(A)$ y $p(B)$ respectivamente, dado que $p(B) > 0$. Supongamos conocidas las probabilidades a priori de los sucesos A y B, es decir, $p(A)$ y $p(B)$, así como la probabilidad condicionada del suceso B dado el suceso A, es decir $p(B|A)$. Con la probabilidad a posteriori del suceso A conocido se verifica el suceso B, es decir $p(A|B)$, puede calcularse a partir de la siguiente fórmula:

$$p(A|B) = \frac{p(A,B)}{p(B)} = \frac{p(A)p(B|A)}{p(B)} = \frac{p(A)p(B|A)}{\sum_{A'} p(A')p(B|A')}.$$

El teorema de Bayes puede formularse para variables aleatorias, tanto unidimensionales como multidimensionales. Para la formulación para dos variables aleatorias unidimensionales que denotamos por $\mathbf{X}$ e $\mathbf{Y}$, tenemos que:

$$p(\mathbf{Y} = \mathbf{y}|\mathbf{X} = \mathbf{x}) = \frac{p(\mathbf{Y} = \mathbf{y})p(\mathbf{X} = \mathbf{x}|\mathbf{Y} = \mathbf{y})}{\sum_{\mathbf{y}'} p(\mathbf{Y} = \mathbf{y}')p(\mathbf{X} = \mathbf{x}|\mathbf{Y} = \mathbf{y}')}.$$

Este resultado puede ser expresado en términos del número de componentes de cada una de las variables multidimensionales anteriores $\mathbf{X}$ e $\mathbf{Y}$, de la siguiente manera:

$$p(\mathbf{Y} = \mathbf{y}|\mathbf{X} = \mathbf{x}) = p(Y_1 = y_1, \ldots, Y_m = y_m|X_1 = x_1, \ldots, X_n = x_n) =$$
$$\frac{p(Y_1 = y_1, \ldots, Y_m = y_m)p(X_1 = x_1, \ldots, X_n = x_n|Y_1 = y_1, \ldots, Y_m = y_m)}{\sum_{y'_1,\ldots,y'_m} p(Y_1 = y'_1, \ldots, Y_m = y'_m)p(X_1 = x_1, \ldots, X_n = x_n|Y_1 = y'_1, \ldots, Y_m = y'_m)}.$$

en el caso de que la variable aleatoria $\mathbf{X}$ sea n-dimensional y la variable aleatoria $\mathbf{Y}$ sea m-dimensional.

2.3. Clasificadores bayesianos

A continuación, describiré los clasificadores naive Bayes y semi naive Bayes, los cuales son ampliamente empleados en la predicción en medicina.

Paradigma de clasificación naive Bayes [15]. El paradigma naive Bayes se basa en dos premisas sobre las variables predictoras (hallazgos, síntomas) y la variable a predecir (diagnóstico). La primer premisa es que los diagnósticos son excluyentes, es decir, la variable C a predecir toma uno de sus m posibles valores: $c^1, \ldots, c^m$.

La segunda premisa es que las variables predictoras son condicionalmente independientes dado el valor de la clase C, es decir, los síntomas, signos y pruebas médicas son condicionalmente independientes dado el diagnóstico. En el caso de conocer el valor de la variable diagnóstico, el conocimiento del valor de cualquiera de los hallazgos es irrelevante para el resto de los hallazgos. Esta condición se expresa como:

$$p(X_1 = x_1, \ldots, X_n = x_n | C = c) = \prod_{i=1}^{n} p(X_i = x_i | C = c).$$

Por tanto, la búsqueda del diagnóstico más probable, c^*, una vez conocidos los síntomas $(x_1, \ldots, x_n)$ de un determinado paciente, en el paradigma naive-Bayes se expresa como:

$$c^* = \text{argmax}_c p(C = c) \prod_{i=1}^{n} p(X_i = x_i | C = c).$$

El paradigma de clasificación semi naive Bayes [13], supera la asunción de independencia condicional del paradigma naive Bayes y detecta aquellas variables predictoras irrelevantes así como las variables dependientes entre si, creando una nueva variable a partir del producto cartesiano de las mismas. Considerando esta estructura, la instancia este modelo semi naive Bayes asigna la clase c de la siguiente manera:

$$c^* = \arg \max_c p(c) p(x_1|c) p(x_2, x_3|c) p(x_4|c)$$

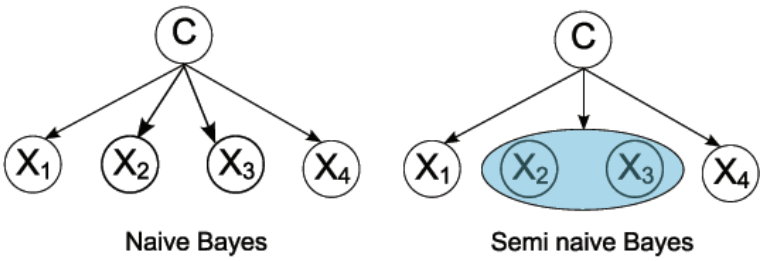

Figura 2: Ejemplo de las estructuras de clasificadores naive Bayes y semi naive Bayes. El clasificador naive Bayes, incluye todas las variables predictoras en el modelo, en este caso son cuatro variables predictora. La estructura del

clasificador semi naive Bayes, esta formada por tres nodos, uno de ellos formado por el producto cartesiano de las variables predictoras X_2 y X_3 y otros dos nodos que corresponden a las variables predictoras X_1 y X_4.

3. Validación de clasificadores

Existen diversos criterios para la evaluación de los algoritmos de clasificación. Su elección dependerá del dominio del problema. Por ejemplo, *el porcentaje de casos bien clasificados* mide la bondad (precisión) del clasificador. La bondad de un clasificador es una estimación de la probabilidad de la clasificación correcta de una instancia elegida al azar.

En clasificación, es aconsejable inducir un modelo a partir de un conjunto de datos llamado conjunto de entrenamiento y otro conjunto de datos llamado conjunto de prueba, el cual se aplica en la fase de clasificación. Con ello se evitan los resultados denominados optimistas. Es importante que para estimar la precisión de un clasificador se utilice un método con poca varianza.

3.1. Matriz de confusión

La matriz de confusión detalla el resultado de la clasificación. En la diagonal principal se reportan los casos correctamente clasificados y en la opuesta se detallan los errores de la predicción. Las columnas representan las clases presentes en los datos y las filas las clases en las que son predichas las instancias. En la tarea de clasificación con dos valores, dado un clasificador y una instancia se producen cuatro valores de salida como son: *verdadero positivo*, si la instancia es clasificada correctamente y su clase pertenece a la positiva; *verdadero negativo* se genera cuando la instancia es correctamente clasificada con la clase de valor negativo; *falso positivo* cuando la instancia es de la clase negativa y es clasificada como clase positiva y por último *falso negativo* se presenta cuando el clasificador clasifica erróneamente una instancia de la clase positiva como un caso de clase negativa. En base a estas cuatro salidas se puede construir una tabla de contingencia, o matriz de confusión, representando al conjunto de datos de prueba.

	clase real	*verdadero*	*falso*
clase predicha	*verdadero*	verdaderos positivos	falsos positivos
	falso	falsos negativos	verdaderos negativos

Tabla 2: Matriz de confusión.

A partir de la matriz de confusión se pueden extraer algunas medidas para comprender la distribución y naturaleza de los errores cometidos por el clasificador.

- La *sensibilidad* de un clasificador representa la fracción de verdaderos positivos y se calcula de la siguiente forma:

$$sensibibilidad = \frac{verdaderos\ positivos}{verdaderos\ positivos + falsos\ negativos}.$$

- La *especificidad* representa la proporción de verdaderos negativos y se calcula de la siguiente forma:

$$especificidad = \frac{verdaderos\ negativos}{verdaderos\ negativos + falsos\ positivos}.$$

- El *valor predictivo de la clase positiva* representa la precisión del clasificador en términos del porcentaje de casos positivos correctamente clasificados y es calculado de la siguiente forma:

$$valorpredictivopositivo = \frac{verdaderos\ positivos}{verdaderos\ positivos + falsos\ positivos}.$$

- El *valor predictivo de la clase negativa* se calcula de la siguiente forma:

$$valorpredictivonegativo = \frac{verdaderos\ negativos}{verdaderos\ negativos + falsos\ negativos}.$$

Otras dos medidas a tener en cuenta en la validación son el ratio de error verdadero y el ratio de error aparente.

El *ratio de error verdadero* de un clasificador es la tasa del error al clasificar mal los casos del conjunto de prueba que no han sido incluidos en el conjunto de entrenamiento utilizado para su inducción. Resulta ser una medida honesta de la estimación de la tasa de error.

Ratio de error aparente, es la tasa de error obtenido por un modelo al clasificar las instancias incluidas en su inducción. En particular esta medida tiende a ser muy optimista ya que los datos utilizados para su construcción se ajustan mejor al modelo que las instancias no incluidas en la inducción, generándose un fenómeno de sobre entrenamiento o también llamado *overfitting* y su error tiende a ser sesgado (*biased*). Para obtener un clasificador no sesgado, es recomendable probar el clasificador con un conjunto de datos no incluidos en su inducción. En el caso de que se cuente con pocos casos para la prueba, el error estimado tiende a tener una varianza grande.

4. Sistemas de apoyo a la decisión en el diagnóstico médico, basados en redes Bayesianas

Describiré dos ejemplos de la aplicación de los clasificadores Bayesianos en problemas médicos complejos con un alto índole social como son: la selección embrionaria en tratamientos in-vitro y el diagnóstico de la demencia en la enfermedad de Parkinson a partir de imágenes de resonancia magnética.

4.1. Selección embrionaria en tratamientos de reproducción asistida

Actualmente la infertilidad es considerada un problema social y es diagnosticada cuando una pareja después de un año no logra un primer embarazo de forma natural, la cual se conoce como la infertilidad primaria. Existe otro caso de infertilidad en la pareja, denominada infertilidad secundaria, caracterizándose por el hecho de que la pareja, después de un embarazo previo, no consigue en condiciones normales obtener un nuevo embarazo. Las causas de infertilidad pueden ser de origen femenino como masculino o, incluso, presentarse en ambos. La fertilización in-vitro (FIV) o la inyección introcitoplasmática del esperma (ICSI) son algunas de las técnicas que se emplean en la reproducción humana asistida a partir de las cuales es posible la fecundación, selección y transferencia de embriones al útero de la paciente, haciendo posible la implantación y el desarrollo del embrión. [1]

La tasa de lograr el embarazo de las personas que no tienen problemas reproductivos, se ubica en un rango de [20-30] % por cada intento. La tasa de implantación (éxito) en tratamientos de reproducción asistida, en algunos países Europeos y Estados Unidos, está en un rango del 29 % y 38 %. En el caso de tratamientos de reproducción asistida aplicados a mujeres menores de 36 años se obtiene una tasa de éxito del 38 %. La tasa de éxito del tratamiento de FIV para las mujeres de 36 a 39 años es del 29 %. Y, en las mujeres mayores de 40 años, el tratamiento de FIV tiene una tasa de éxito muy pequeña mejorando aproximadamente hasta un 13 % con donación de óvulos. Estos porcentajes de éxito son orientativos debido a la falta de homogeneidad en los tratamientos y procedimiento médicos, así como en los protocolos implementados en las unidades de fertilización in-vitro, todos ellos regidos por la legislación de cada país, lo cual dificulta la estandarización de la tasa de éxito entre las clínicas de reproducción asistida.

Los factores que más influyen en la variación de la tasa de embarazo son la edad de la mujer, la calidad de los óvulos extraídos, la calidad espermática, el número de embriones transferidos, el origen de la infertilidad así como los tratamientos médicos y la falta de criterios concluyentes de criterios en el proceso de selección

[1] En los tratamientos de reproducción asistida el término *transferencia* se refiere al procedimiento clínico utilizado para insertar los embriones seleccionados en el útero materno, mientas que el término *implantación* está relacionado con el éxito de lograr el embarazo (alguno de los embriones transferidos ha logrado adherirse a la pared del útero).

embrionaria. En algunos países como Estados Unidos, Suiza, Francia, Escandinavia y Países Bajos han llegado a obtener tasas de éxito del 40 % siendo este el resultado más optimista sin tener en cuenta el grupo de mujeres con más de 40 años de edad [9].

En las técnicas de reproducción asistida se busca maximizar la probabilidad de que nazca un niño sano y por otro lado se busca minimizar el riesgo de los embarazos múltiples. Siendo estos los motivos por los cuales el problema de la selección embrionaria en los tratamientos de fecundación in-vitro es tan importante.

La selección embrionaria forma parte de las rutinas de las unidades de reproducción asistida en todas las clínicas del mundo. La selección embrionaria se basa en la observación, evaluación y catalogación de los embriones realizada por el embriólogo y tiene como objetivo el determinar cuales, de entre los embriones en desarrollo, son los más viables para ser transferidos el segundo o tercer día después de la fecundación. La formación y capacitación del embriólogo resulta decisiva en la fecundación in-vitro, así como en la selección de los embriones ya que el factor tiempo influye en la evaluación y el biólogo no dispone de un periodo largo para realizarla, por lo que su criterio resulta muy importante para el éxito del tratamiento.

Los trabajos relacionados con la selección embrionaria aplicando clasificadores Bayesianos son descritos en [17] donde se propuso un sistema multi-clasificador basado en el esquema de pila, aplicando como meta-clasificador un clasificador Bayesiano adaptado al dominio continuo que predice el éxito del tratamiento de reproducción asistida a partir de las distribuciones de probabilidad de diferentes paradigmas de clasificación supervisada. En [18] se aplicaron los clasificadores a la selección de un único embrión a partir de datos de la imagen.

En Bayesian classification for the selection of in-vitro human embryos using morphological and clinical data [19] se presentó un nuevo planteamiento para el problema de clasificación supervisada. Este estudio consta de 63 tratamientos de fertilización in-vitro de pacientes entre 27-46 años del programa de fertilización in-vitro de la unidad de reproducción de la Clínica del Pilar en San Sebastián, Guipúzcoa durante el periodo de julio de 2003 a diciembre de 2005. El estudio consta de 18 casos exitosos donde se logró la implantación de al menos uno de los embriones transferidos y 45 casos no exitosos.

En todos los casos se transfirió un conjunto de tres embriones. El protocolo aplicado en la Clínica del Pilar para la selección embrionaria está basado en primer lugar en la catalogación del cigoto (formado durante las primeras 24h después de la fecundación) y posteriormente en la evaluación y catalogación morfológica del embrión (véase la Figura 3 para identificar algunas de las estructuras morfológicas). Además de las variables con información médica referentes al ciclo de fertilización in-vitro y datos de los pacientes. La Tabla 3 muestra el conjunto de

variables predictoras, correspondiendo cada una a una de las características mencionadas anteriormente. En este estudio se aplicaron ocho diferentes clasificadores Bayesianos.

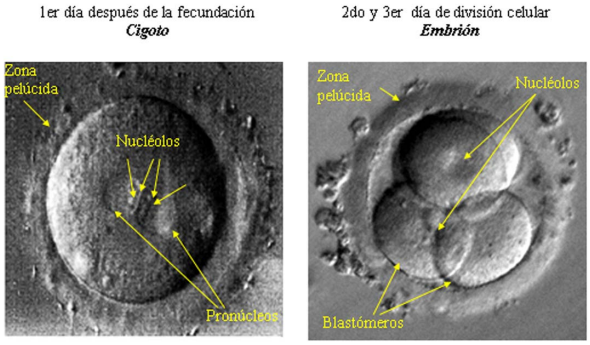

Figura 3: Se muestran algunas de las estrucutras mofológicas del cigoto y del embrión.

Variables clínicas	Variables del tratamiento	Características morfológicas
Edad	Núm. transferidos	Catalogación del cigoto
Ciclo actual	Núm. congelados	Catalogación del embrión
Tratamientos previos	Calidad en la transferencia	Número de células
Donante	Día de transferencia	Tamaño de los blastómeros
Tipo de esterilidadinfertilidad		Blastómeros multinucleados
Calidad del semen		Grado de fragmentación
Causa femenina		Grosor de la zona pelúcida
Causa masculina		

Tabla 3: Lista de variables registradas en la base de datos de la Unidad de reproducción de la Clínica del Pilar-Donostia.

Métodos

El problema de la selección embrionaria se ha planteado como un problema de clasificación supervisada creando un modelo que asigne a cualquier nuevo caso de fertilización in-vitro $\boldsymbol{x} = (x_1, \ldots, x_n)$ un valor de variable clase C. Formalmente denotaremos por $\boldsymbol{x} = (x_1, \ldots, x_n)$ el vector de características de un conjunto de embriones transferidos así como las variables clínicas de la paciente y del tratamiento de fertilización in-vitro. El valor de la variable clase está denotado por $c \in C = \{0, 1\}$, asignándose el valor 1 si ha ocurrido la implantación y el valor 0 en caso contrario.

Los datos se han discretizado en dos intervalos de igual frecuencia. Se aplicó una validación cruzada dejando un caso fuera con el propósito de evitar el sobre ajuste de los siguientes clasificadores: naive Bayes, selective naive Bayes, semi

naive Bayes, clasificador Bayesiano aumentado a árbol (TAN), El clasificador k-dependiente y los clasificadores filter selective naive Bayes, filter TAN y filter-k-DB. Los clasificadores semi naive Bayes y selective naive Bayes, tienen implícito en la construcción del modelo, una selección de envoltura, la cual selecciona un subconjunto de variables con las que se va evaluando el porcentaje de casos bien clasificados del modelo. En el caso de los filter, se aplica el criterio de evaluación a todas las variables predictoras y las variables que superan en este caso la prueba de Chi-cuadrada forman un subconjunto de variables con las que posteriormente se construye el modelo.

Resultados

Los resultados de este experimento se describen en términos del porcentaje de casos bien clasificados, los cuales se encuentran resumidos en la Tabla 4. El clasificador semi naive Bayes obtuvo el porcentaje de casos bien clasificados más alto con un 71.43 % con una sensibilidad del 22.22 % y una especificidad del 91.11 % presentando un valor de predicción para los casos implantados del 50.00 % y un 74.54 % en la predicción de los casos no-implantados. Los clasificadores naive Bayes, selective naive Bayes, FSNB, y FTAN obtuvieron una precisión del 68.25 %, mientras que el clasificador naive Bayes obtuvo un 38.89 % de sensibilidad, 80.00 % de especificidad y un valor de predicción de implantación del 43.75 %. El clasificador Bayesiano kDB fue el que obtuvo la menor precisión con un 60.32 % de casos bien clasificados.

Clasificador	Precisión	Sensibilidad	Especificidad	Valor predictivo Implantación	Valor predictivo No-implantación
naive Bayes	68.25	38.89	80.00	43.75	76.59
semi naive Bayes	71.43	22.22	91.11	50.00	74.54
selective naive Bayes	68.25	5.55	93.33	25.00	71.19
TAN	63.49	5.55	86.67	14.29	69.64
kDB	60.32	5.55	82.22	11.11	68.52
FSNB	68.25	11.11	91.11	33.33	71.92
FTAN	68.25	11.11	91.11	33.33	71.92
FkDB	63.49	0.00	88.89	0.00	68.96

Tabla 4: Resultados de la clasificación de un conjunto de tres embriones por medio de diferentes clasificadores Bayesianos en términos de la sensibilidad, la especificidad, el valor de predicción de implantación y del valor predictivo de no-implantación.

Al aplicar la prueba de McNemar para evaluar la diferencia estadística del rendimiento entre parejas de clasificadores Bayesianos no se encontró diferencia significativa entre el mejor clasificador, semi naive Bayes, y cada uno de los clasificadores empleados en este estudio (naive Bayes, selective naive Bayes, kDB, TAN, FSNB, FTAN y FkDB). McNemar $p - value = 0{,}6892$ con $\alpha = 0{,}05$

La Figura 4 muestra la estructura del clasificador semi naive Bayes, el cual obtuvo la mayor precisión en la selección de un conjunto de embriones en tratamien-

tos de fertilización in-vitro. El clasificador semi naive Bayes está formada por dos nodos: el primero incluye el producto cartesiano de las variables predictoras *Embrión1-Tamaño Blastómeros, Embrión1-Fragmentación Blastómeros, y Calidad del esperma*. El segundo nodo está formado por el producto cartesiano de las variables predictoras *Embrión1-Multinucleado, Embrión2-Grosor de la zona pelúcida y Embrión3-Tamaño de Blastómeros*. Las características morfológicas del embrión así como la información clínica del tratamiento resultaron ser de especial interés para el experto, las cuales son acorde con la literatira [8].

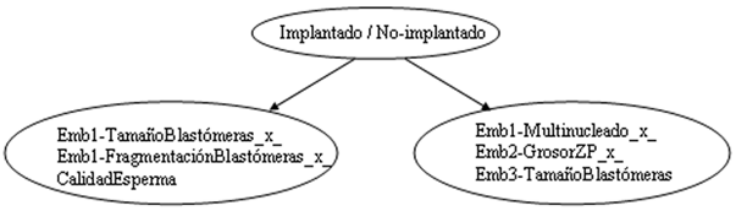

Figura 4: Modelo semi naive Bayes para la clasificación de un conjunto de embriones en tratamientos de fertilización in-vitro.

Conclusiones

Este estudio permitió la aplicación de los clasificadores Bayesianos al problema de la selección embrionaria en tratamientos de reproducción asistida. La selección de un subconjunto de variables por cada uno de los diferentes clasificadores Bayesianos nos permitió orientar la investigación hacía la extracción automática de características del embrión y en especial, la medición de la variación del grosor de la zona pelúcida, que como se ha visto, resulta de interés por su alto valor discriminativo.

4.2. Predicción de demencia en pacientes con Parkinson

En está sección les comentaré de un estudio presentado en [20] en cual se aplicaron técnicas de selección de variables para extraer caraterísticas altamente predictivas.

Los pacientes con Parkinson (PD) presentan alteraciones estructurales en el cerebro como el deterioro de la corteza cerebral y perdida de volumen en estructuras subcorticales que pueden servir como biomarcadores, los cuales pueden ser registrados por imágenes de resonancia magnética (IRM) y cuantificados por herramientas como FreeSurfer [4].

El 83 % de los pacientes con Parkinson desarrollan demencia a lo largo de la evolución de la enfermedad [11], sin embargo, las características extraídas de

las imágenes de RM no han sido estudiadas en profundidad. En este estudio se aplicaron diferentes clasificadores Bayesianos [22] con el propósito de identificar pacientes con demencia (PDD) de aquellos pacientes sin demencia (PDWD).

El estudio consta de 30 pacientes con Parkinson, 17 hombres and 13 mujeres, con un rango de edad de 74 años. Se formaron dos grupos, el primero con 14 PDD y el segundo con 16 PDWD. Todas las imágenes fueron adquiridas en un equipo de IRM de 3 Teslas (3D-FFE, Philips Achieva) las cuales fueron transferidas para su postprocesamiento al PIC (PIC, IFAE, UAB). Se extrajeron un total de 214 medidas de volumen de estructuras corticales y subcorticales y del grosor de la corteza cerebral. La Figura 5 muestra tres imágenes, la primera es una imagen anatómica de RM, la segunda es la imagen con neurodegeneración cerebral y la tercera se muestra un ejemplo de la parcelación automática realizada con el programa FreeSurfer.

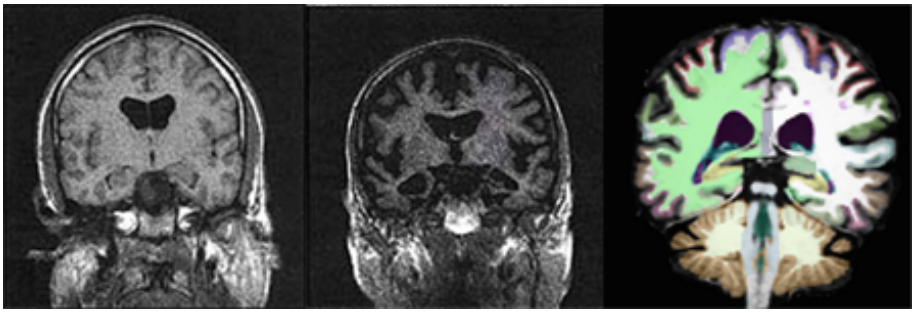

Figura 5: Se muestra en la primer figura la imagen de RM, posteriormente una imagen con neurodegeneración y finalmente la segmentación automática de las estructuras cerebrales hecha por FreeSurfer.

Métodos

Este estudio consta de dos etapas. La primera consta de la comparación de dos clasificadores Bayesianos de distinta complejidad con el propósito de construir un sistema de apoyo a la decisión en el diagnóstico de la demencia en pacientes con Parkinson. La segunda etapa consta de la identificación de las estructuras cerebrales más informativas que distingan los dos grupos, PDWD y PDD, aplicando el algoritmo basado en selección de variables midiendo su grado de correlación (CFS) [10]. Los datos fueron discretizados por entropía con el algoritmo de Fayat e Irani [6] y la validación de los clasificadores fue hecha aplicando la validación cruzada dejando un caso fuera.

Resultados

Los resultados están expresados en el porcentaje de casos bien clasificados, la sensibilidad, especificidad y la medida-F. El clasificador naive Bayes resulto ser el

mejor clasificador aplicando las 212 medidas obteniendo un porcentaje de casos bien clasificados del 93.33 % con las medidas-F para los grupos PDWD y PDD de 0.938 y 0.929 respectivamente, con una sensibilidad de 0.938 y 0.929. La Tabla 5 muestra los resultados obtenidos por el clasificador semi naive Bayes y CFS-naive- Bayes que mejoró el porcentaje de casos bien clasificados a 96.66 % con una sensibilidad de 0.929 y especificidad de 1.

Clasificador	% casos bien clasificados	F-measure	Sensibilidad	Especificidad
naive Bayes	93.33	0.929	0.929	0.937
semi NB	86.66	0.786	0.937	0.846
CFS-NB	96.66	0.963	0.929	1

Tabla 5: Resultados de los clasificadores Bayesianos en la predicción entre los grupos PDWD y PDD.

Lista de variables	semi NB	CFS-NB
Hipocampo derecho	✓	✓
Hipocampo izquierdo	✓	✓
Horno temporal del ventrículo derecho	✓	✓
Horno temporal del ventrículo izquierdo	✓	✓
Vol. de materia blanca hemisferio izquierdo		✓
Grosor promedio del núcleo cuneatus izquierdo		✓

Tabla 6: Lista de variables predictoras seleccionadas por los clasificadores semi naive Bayes y CFS- naive Bayes.

Al aplicar los modelos semi naive Bayes y CFS-naive Bayes, se han seleccionado una serie de variables predictoras para distinguir entre los dos grupos PDWD y PDD. En ambos modelos de clasificación se eligieron como variables relevantes los volúmenes del horno lateral de los ventrículos izquierdo y derecho así como los hipocampos de ambos hemisferios izquierdo y derecho. La Tabla 6 resume las estructuras seleccionadas por ambos modelos.

Conclusiones

Los resultados de aplicar los clasificadores Bayesianos en la predicción de demencia entre los pacientes con Parkinson sin demencia y los que sí la presentan. Resulta de gran interés para la predicción de la demencia en estados tempranos de la enfermedad. Después de aplicar los métodos de selección de variables, el algoritmo de envoltura en el semi naive Bayes y el algoritmo de selección de variables basados en la correlación, los resultados confirman que los volúmenes de los hornos temporales de los ventrículos laterales, así como los ambos hipocampos, son biomarcadores que pueden ayudar en el diagnóstico temprano de los pacientes

con demencia. Estos resultados son acorde con la literatura relacionada con los casos de demencia en pacientes con Parkinson [3,26].

5. Conclusiones generales

La comunicación entre médicos y la comunidad de minaría de datos es fundamental para la construcción de los sistemas de apoyo a la decisión. La reducción de tiempo en el diagnóstico médico puede suponer algunas ventajas como: la atención temprana y adecuada al paciente así como la reducción de costos del sistema sanitario.

Desde el punto de vista de la minería de datos y el aprendizaje automático, es deseable la construcción de paradigmas de clasificación aplicando diferentes métodos matemáticos con el propósito de que sean condiables para su aplicación en problemas reales como lo es la medicina.

Por último, me gustaría agradecer a Marta y Raúl la invitación a participar en el ciclo de charlas de Un Paseo por la Geometría y también decir que es admirable el trabajo que realizan para divulgar las matemáticas.

Bibliografía

[1] R. Agrawal, T. Imielinski, A. Swami, *Mining association rules between sets of items in large databases*, Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, 207-216, 1993.

[2] Biografía Bayes, http://www-history.mcs.st-and.ac.uk/Biographies/Bayes.html

[3] A. Bruck, T. Kurki, V. Kaasinen, T. Vahlberg, J.O. Rinne, *Hippocampal and prefrontal atrophy in patients with early non-demented Parkinson's disease is related to cognitive impairment*, The Journal of Neurology, Neurosurgery, and Psychiatry 75(10), 1467-1469, 2004.

[4] A.M. Dale, B. Fisch, M.I. Sereno, *Cortical surface-based analysis I. Segmentation and surface reconstruction*, Neuroimage 9(2), 179-194, 1999.

[5] A.P. Dempster, N.M. Laird, D.B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm (with discussion)*, J. Roy. Statist. Soc. (Ser. B) 39, 1-38, 1977.

[6] U.M. Fayyad, K.B. Irani, *Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning*, Proceedings of the International Joint Conference on Uncertainty in AI, 1022-1027, 1993.

[7] R.A. Fisher, *The use of multiple measurements in taxonomic problems*, Eugen. 7, 179-188, 1936.

[8] A. Gabrielsen, S. Lindenberg, K. Petersen*The impact of the zona pellucida thickness variation of human embryos on pregnancy outcome in relation to sub-*

optimal embryo development. A prospective randomized controlled study, Human Reproduction 16(10), 2166-2170, 2001.

[9] C. Giorgetti, E. Hans, P. Terriou, J. Salzmann, B. Barry, V. Chabert-Orsini, J.M. Chincole, J.P. Franquebalme, E. Glowaczower, M-C. Sitri, M-C. Thibault, R. Roulier, *Early cleavage: An additional predictor of high implantation rate following elective single embryo transfer*, Reproductive BioMedicine Online 14(1), 85-91, 2007.

[10] M. A. Hall, *Correlation-based Feature Selection for Machine Learning*, Ph.D dissertation. Dept. of Computer Science, Waikato University, 1998.

[11] G. Halliday, M. Hely, W. Reid, J. Morris, *The progression of pathology in longitudinally followed patients with Parkinson's disease*, Acta Neuropathology 115(4), 409-415, 2008.

[12] D. W. Hosmer, S. Lemeshow, *Applied Logistic Regression*, John Wiley, 1989.

[13] I. Kononenko, *Semi-naïve Bayesian classifiers*, Proceedings of the 6th European Working Session on Learning, 206-219, 1991.

[14] I. Kononenko, *Machine learning for medical diagnosis: History, state of the art and perspective*, Artificial Intelligence in Medicine 23(1), 89-109, 2001.

[15] M. Minsky, *Steps toward artificial intelligence*, Transactions on Institute of Radio Engineers 49, 8-30, 1961.

[16] T.M. Mitchell, *Machine learning*, McGraw-Hill, 1997.

[17] D.A. Morales, E. Bengoetxea, P. Larrañaga, *Combining multi-classifiers with Gaussian network for selection of in-vitro human embryos using morphological and clinical data*, Data Mining and Medical Knowledge Management: Cases and Applications, P. Berka, J. Rauch, D. Zighed (eds.), IGI Global Inc., 307-331, 2009.

[18] D.A. Morales, E. Bengoetxea, P. Larrañaga, *Selection of human embryos for transfer by Bayesian classifiers*, Computers in Biology and Medicine 38 (11-12), 1177-1186, 2008.

[19] D.A. Morales, E. Bengoetxea, P. Larrañaga, M. García, Y. Franco, M. Fresnada, M. Merino*Bayesian classification for the selection of in-vitro human embryos using morphological and clinical data*, Computer Methods and Programs in Biomedicine 90, 104-116, 2008.

[20] D.A. Morales, E. Bengoetxea, Y. Vives, B. Gomez, P. Larrañaga, J. Pagonabarraga, J. Kulisevsky, G. Llebaria, R. Rotger, I. Corcuera, *Predicting Dementia Development in Parkinson Disease using Bayesian Networks*, Proceedings Human Brain Mapping, Barcelona, 2010.

[21] B. Pandey, R.B. Mishra, *Knowledge and intelligent computing system in medicine*, Computers in Biology and Medicine 39, 215-230, 2009.

[22] J. Pearl, *Probabilistic Reasoning in Intelligence Systems*, Morgan Kaufmann, Los Altos CA., 1988.

[23] R.A. Russell, *An Environment for Robot Learning*, Proccedings 2002 Aus-

tralasian Conference on Robotics and Automation, Auckland, 27-29 November, 2002.
[24] W.B. Schwartz, *Medicine and the computer: The promise and problems of change*, New Engl. Journal Medicine 283, 1257-1264, 1970.
[25] E.H. Shortliffe, R. Davis, S.G. Axline, B.G. Buchanan, C.C. Green, S.N. Cohen, *Computer-based consultations in clinical therapeutics: Explanation and rule acquisition capabilities of the MYCIN system*, Computers and Biomedical Research 8(4), 303-320, 1975.
[26] Ch. Summerfield, C. Junqué, E. Tolosa, P. Salgado-Pineda, B. Gómez-Ansón, M. Martí, P. Pastor, B. Ramírez-Ruíz, J. Mercader, *Structural brain changes in Parkinson disease with dementia. A voxel-based morphometry study*, Archives of Neurology 62(2), 281-285, 2005.
[27] S. Weiss, C.A. Kulikowski, A. Safir, *Glaucoma consultation by computer*, Comp. Biol Med. 8, 24-40, 1978.
[28] S. Weiss, C. Kulikowski, S. Amarel, A. Safir, *A Model-based method for computer-aided medical decision-making*, Artificial Intelligence 11, 145-172, 1978.
[29] I.H. Witten, E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann Publishers, 2005.

Dinora A. Morales Vega
Universidad Politécnica de Madrid
Facultad de informática
Departamento de Inteligencia Artificial
Campus de Montegancedo, Boadilla del Monte
28660 (Madrid)
e-mail: ***dinora.morales@fi.upm.es***