DETERMINACION DE LA EFICACIA DE LA BRAQUITERAPIA EN TRATAMIENTO DE CÁNCER BASADA EN MINERIA DE DATOS

Reparaz, D., Merlino, H., Rancan, C., Rodríguez, D., Britos, P., García-Martínez, R.

[1] Departamento de Ingeniería Industrial. Instituto Tecnológico de Buenos Aires
[2]Centro de Ingeniería del Software e Ingeniería del Conocimiento. Instituto Tecnológico de Buenos Aires
[3]Laboratorio de Sistemas Inteligentes. Facultad de Ingeniería. Universidad de Buenos Aires

{hmerlino, crancan, drodrigu, pbritos, rgm}@itba.edu.ar

1. INTRODUCCION

En las últimas décadas, hemos sido testigos de vertiginosos cambios en las tecnologías existentes. En este marco de rápida evolución, no podemos dejar afuera los avances producidos en las ciencias de salud. Frente a dichos cambios, aparecen en la medicina nuevas alternativas para el tratamiento de enfermedades. Al encontrarnos frente a estas nuevas alternativas, comienza a tomar mayor importancia el concepto de calidad de vida, ligada a cada uno de los tratamientos. Para el tratamiento del cáncer de próstata, existen tratamientos alternativos. Entre ellos se encuentran, la prostatectomía radical (abierta o laparoscópica), radioterapia conformada externa, terapia hormonal y braquiterapia. Para distintos pacientes con estadíos de enfermedad curables, la elección del tratamiento debe contrastarse contra los riesgos de la terapia, la edad del paciente, el comportamiento biológico del cáncer, la calidad de vida y otros factores. Esta situación, pone de manifiesto la necesidad de generar herramientas de toma de decisiones, para maximizar la eficacia a la hora de elegir el tratamiento adecuado para un paciente. En la actualidad, los especialistas cuentan con algunas herramientas que sirven de ayuda en el momento de tomar una decisión respecto al tratamiento adecuado para cada paciente. La herramienta por excelencia es la experiencia profesional, la cual confiere a los especialistas criterios consensuados a la hora de optar por un tratamiento u otro. Estos criterios se basan en una serie de indicadores, que varían en un rango de valores determinado, y que al interactuar generan un output, que es la decisión del profesional acerca del tratamiento "óptimo". Con el afán de brindar un marco teórico a las decisiones, tanto para optimizarlas como para dotar de un valor cuantitativo a la experiencia, es que la ingeniería industrial hace su aporte. Existen dos métodos de ayuda en la toma de decisiones. El primero, y el más extensamente estudiado y aceptado, es el conocido como Nomogramas (Partin). Los nomogramas son herramientas predictivas basadas en análisis de regresión multivariable. Son representaciones gráficas de modelos estáticos, que utilizan escalas, para calcular el "peso" del valor de cada variable, y luego predecir un determinado punto final (end point). Los puntos finales que se estudian entre otros pueden ser: estadío de la enfermedad, probabilidad de reaparición de la enfermedad [Shariat *et al*., 2005], predicción de retención urinaria aguda o intervención quirúrgica en pacientes con hyperplasia próstatica benigna [Slawin *et al*., 2006]. Las predicciones que se obtienen son resultado de los indicadores individuales de cada paciente. Los nomogramas están formados por una serie de ejes, cada uno de los cuales representa una variable. Las variables varían dentro de una escala, y a cada valor de la variable le corresponde una puntuación dependiendo del impacto que dicha variable tenga en la predicción. El eje final, concentra la puntuación final, que es transformada en la probabilidad de alcanzar el punto final. Este tipo de métodos, debe tener especial cuidado a la hora de definir como imputar el "peso" al valor de las variables, al descartar variables que puedan resultar importantes, al incorporar variables inadecuadas, entre otras. Los nomogramas son los métodos mas estudiados y por lo tanto existen varios estudios de validación de dichos

modelos. Entre otros podemos encontrar: Validación de nomograma para predecir resultados positivos de biopsia en cáncer de próstata [Yanke *et al*., 2005].

El otro método es el de la minería de datos [García-Martínez *et al*., 2003]. Éste último, es bastante novedoso para este tipo de aplicaciones, por lo cual existen desarrollos muy puntuales. Existe gran variedad de algoritmos (caracterización, inducción, etc), que tienen la capacidad de aprender de la experiencia. Están formados por nodos de ingreso, nodos ocultos y nodos de salida. Mediante este entrenamiento (Supervisado, No Supervisado) el modelo ajusta los pesos de las neuronas ocultas para optimizar la salida. La ventaja de minería frente a los nomogramas es que posee la capacidad de resolver relaciones no lineales complejas entre las variables, sin necesidad de hacer ninguna suposición previa respecto a dichas relaciones. La utilización de este método aún sigue siendo controversial, tanto por lo novedoso para estas aplicaciones y porque no resulta sencillo demostrar que sus resultados sean mejores a los arrojados por los nomogramas [Stephan *et al*., 2005].

En este contexto, el objetivo de este trabajo es estudiar la aplicación de algoritmos de caracterización e inducción [Fiszelew y García-Martínez, 2002] de forma tal de poder predecir la eficacia de la braquiterapia en el tratamiento del cáncer de próstata.

2. ESTADO DE LA CUESTIÓN

2.1. Caracterización de datos

La caracterización consiste en agrupar un conjunto de datos sin tener clases previamente definidas. Estos algoritmos operan basándose en la similitud de los valores de los atributos de los distintos datos. Este tipo de aprendizaje se realiza en forma no supervisada ya que no se saben de antemano las clases del set de datos de entrenamiento. La caracterización identifica regiones densamente pobladas, de acuerdo a alguna medida de distancia, en un gran conjunto de datos multidimensional [Chen & Han, 1996]. El análisis de clases se basa en maximizar la similitud de las instancias en cada cluster y minimizar la similitud entre clusters [Han & Lamber, 2001]. Se utiliza para reconocimiento de patrones, análisis de datos, procesamiento de imágenes entre otras. Como función de la *minería de datos*, el análisis de clases puede ser utilizado de forma independiente para obtener la distribución del set de datos, para caracterizar cada clase y dividir grupos para su análisis. Alternativamente, puede servir para el preprocesamiento de datos, antes de utilizar otros algoritmos.

2.2. Clasificación de Datos

Los algoritmos de clasificación, en delante de inducción, se utilizan para clasificar un conjunto de datos basado en los valores de sus variables. [Servente & García Martínez, 2002]. El objetivo de la inducción es analizar los datos de entrenamiento y a través de aprendizaje supervisado, desarrollar una descripción o un modelo para cada clase utilizando las características disponibles en los datos. Aún cuando existen varios enfoques para los algoritmos de inducción, se trabajará con aquellos que generan árboles de decisión conocida como la familia TDIT (*Top Down Induction Trees*). Entre otros importantes algoritmos de árboles de decisión, se destaca el ID3 [Quinlan, 1986] y su extensión C4.5 [Quinlan, 1993]. El J48 es una implementación mejorada del algoritmo C4.5, funcionando bien tanto con atributos nominales como numéricos.

2.3. Redes de Bayes

Una red bayesiana es un grafo acíclico dirigido en el que cada nodo representa una variable y cada arco una dependencia probabilística, en la cual se especifica la probabilidad condicional de cada variable dados sus padres, la variable a la que apunta el arco es dependiente (causa-efecto) de la que está en el origen de éste. La topología o estructura de la red nos da información sobre las

dependencias probabilísticas entre las variables pero también sobre las independencias condicionales de una variable (o conjunto de variables) dada otra u otras variables, independencias, simplifican la representación del conocimiento (menos parámetros) y el razonamiento (propagación de las probabilidades). Estas redes son utilizadas en diversas áreas aplicación como por ejemplo en medicina [Beinlinch et al., 1989; Hernández O.J. et al, 2004], ciencia [Breese & Blake, 1995; Hernández O.J. et al, 2004], y economía [Hernández O.J. et al, 2004]. Las mismas proveen una forma compacta de representar el conocimiento y métodos flexibles de razonamiento - basados en las teorías probabilísticas - capaces de predecir el valor de variables no observadas y explicar las observadas. Entre las características que poseen las redes bayesianas, se puede destacar que permiten aprender sobre relaciones de dependencia y causalidad, permiten combinar conocimiento con datos [Heckerman et al., 1995; Díaz & Corchado, 1999; Hernández O.J. et al, 2004] y pueden manejar bases de datos incompletas [Heckerman, 1995; Heckerman & Chickering, 1996; Ramoni & Sebastiani, 1996; Hernández O.J. et al, 2004].

3. DESCRIPCIÓN DEL PROBLEMA

En la actualidad, los especialistas médicos no cuentan con herramientas objetivas, que los ayuden a tomar una decisión respecto al tratamiento óptimo para un paciente. Utilizando la información que los especialistas consideran importante, se pretende encontrar patrones y relaciones entre las variables, de forma de poder predecir de antemano de eficacia de la braquiterapia para un paciente que padece cáncer de próstata. En este contexto, el objetivo del trabajo es caracterizar y clasificar la población de pacientes con cáncer de próstata mediante técnicas de minería de datos, esperando encontrar relaciones subyacentes en los datos que no pueden identificarse mediante un tratamiento estadístico clásico.

4. ABORDAJE DEL PROBLEMA

Con el objetivo de caracterizar a la población de pacientes y encontrar relaciones y patrones de comportamiento en los atributos considerados, se aborda la problemática de la siguiente manera:

1. Proceso de Caracterización utilizando atributos significativos de los pacientes.
2. Análisis y validación de las clases obtenidas con especialistas médicos.
3. Aplicación de algoritmos de inducción a cada clase para identificar reglas de decisión justifiquen la composición de cada grupo.

4.1. Estado de Avance

Hasta el momento se han definido las variables más importantes de la población de pacientes con cáncer de próstata. Se realizó la caracterización de datos y se definieron grupos, que han sido validados con los especialistas. Se ha procesado esta información y tratado bajo algoritmos de inducción. Se está trabajando en la interpretación de resultados y queda por delante el análisis y validación mediante redes bayesianas.

4.2. Descripción del Dataset

Se analizaron 206 registros de pacientes tratados con braquiterapia prostática. Una vez realizada la recolección, análisis y limpieza de los datos iniciales, se formo el set final de datos, en función de los atributos necesarios, formado por 116 registros. Las variables seleccionadas son:

- PSA Pre-implante.
- Nivel de Gleason.
- PSA diagnosticado.
- Edad.
- Volumen Ecográfico en gramos.
- Tiempo transcurrido desde el implante hasta el último seguimiento, Delta T.
- Resultado del tratamiento.

4.3. Resultados del Clustering de los Datos

A continuación se presentan los resultados una vez aplicado clustering, a traves de una red de mapas auto-organizados (tabla 1 y 2).

	Cluster Means		
	Cluster 1	**Cluster 2**	**Cluster 3**
Edad	65,0	64,0	70,0
Delta T	1,7	2,4	4,4
PSA diag.	9,9	8,6	9,6
Gleason	6,0	6,2	5,9
PSA preimp.	0,1	0,5	2,4
Volumen ecografico	33,4	37,0	41,5
Resultado	Fracaso	Fracaso	Éxito

Tabla 1. Centroides obtenidos mediante la Caracterización

	Cluster Variances		
	Cluster 1	Cluster 2	Cluster 3
Edad	24,7	14,3	26,5
Delta T	0,7	3,7	1,5
PSA diag.	3,2	30,4	21,5
Gleason	0,0	0,6	0,9
PSA preimp.	0,0	1,7	9,0
Volumen ecografico	112,2	176,1	294,6

Tabla 2. Varianzas obtenidas mediante la Caracterización

4.3.1. Primera Interpretación de los Clusters

A continuación se presenta el análisis de los resultados obtenidos:

- **Cluster 3 (83%):** Es el que posee mayor cantidad de registros, Está directamente asociado a los casos de tratamiento exitoso. Muestra un tiempo promedio de Delta T de 4,4 años, lo que lo que posiciona en un estadío estable de cura. El Gleason se encuentra en todos los casos por debajo de 7. El PSA preimplante es muy mayor de las otras clases, pero presenta una varianza muy grande.
- **Cluster 2 (13%):** Esta clase agrupa a la mayoría de los fracasos. La característica de los valores de sus atributos es que se encuentran distribuídos en un amplio rango de valores, es decir que sus variables no se encuentran sesgadas.
- **Cluster 1 (4%):** Agrupa una serie de fracasos, que representan un porcentaje minoritario y son los que a priori caracterizan al ruido del sistema.

4.4. Algoritmos de inducción

Una vez definidas las tres distintas clases, se ejecuto el algoritmo ID 3. El nodo objetivo no fue el Resultado, sino que se utilizó la clasificación propuesta por la caracterización y se buscó predecir la clase. El algoritmo generó reglas de decisión, que se dividen en reglas de éxito y reglas de fracaso:

Reglas de Éxito:

1. IF PSApreimp < 1.50 and Edad < 69.50 THEN Cluster = 3 con una confianza de 0.55
2. IF PSApreimp < 1.50 and Edad >=69.50 and PSAdiag < 18 and Gleason =< 7 THEN Cluster = 3 con una confianza de 1.0
3. IF PSApreimp >=1.50 and Volumen ecografico al imp >=45 THEN Cluster = 3 con una confianza de 1.0
4. IF PSApreimp >=1.50 and Volumen ecografico al imp < 45 and Edad >=61.50 and PSAdiag < 18 THEN Cluster = 3 con una confianza de 1.0

El cluster C denota la condición de Éxito. Debe tenerse en cuenta que los atributos: PSA preimplante, PSA diagnosticado, Volumen Ecográfico están categoriazadas, es decir que dichos valores son relativos a su categoría. Así pues, Si el PSA preimplante es menor que 1,5 y la edad es menor que 69,5 el tratamiento es exitoso. Sin embargo cuando la edad supera los 69,5 años, toma importancia que valores toman tanto el PSA diagnosticado como el Glesson. Para tener Éxito el PSA debe ser menor que 18 y el Gleason menor o igual que 7. En el caso que el PSA preimplante sea mayor que 1,5 y el tratamiento resulte exitoso toman importancia variables como Volumen ecográfico, Edad, PSA diagnosticado. Entonces, para PSA preimplante menor a 1,5 y volumen

ecografico mayor a 45 el tratamiento resulta exitoso. Pero si el volumen ecografico es menor a 45 la edad debe ser mayor a 61,5 y el PSA diagnosticado menor a 18 para hablar de Éxito.

Reglas de Fracaso:

1. IF PSApreimp < 1.50 and Edad >=69.50 and PSAdiag < 18 and Gleason >=8 THEN Cluster = 2 con una confianza de 1.0
2. IF PSApreimp >=1.50 and Volumen ecografico al imp < 45 and Edad < 61.50 THEN Cluster = 1 con una confianza de 1.0
3. IF PSApreimp >=1.50 and Volumen ecografico al imp < 45 and Edad >=61.50 and PSAdiag >=18 THEN Cluster = 1 con una confianza de 1.0
4. IF PSApreimp < 1.50 and Edad >=69.50 and PSAdiag >=18 THEN Cluster = 1 con una confianza de 1.0

La regla 1 de Fracaso debe ser mirada en contraste a la regla 2 de Éxito. Resultan análogas. La diferencia radica en el valor de Gleason. Cuando éste es inferior o igual a 7 se verifica Éxito, cuando el valor supera tal límite se convierte en Fracaso. El resto de la variables se mantienen en su rango de valores: PSA preimplante menor a 1,5, edad mayor a 69,5 y PSA diagnosticado menor a 18. Para PSA preimplante mayor a 1,5 y volumen ecografico menor a 45 aparecen dos casos. Si la edad es menor a 61,5 años el tratamiento fracasa, pero si la edad supera 61,5 para que fracase el PSA diagnosticado debe ser mayor a 18. Por ultimo, encontramos una última regla ligada también al fracaso. En el caso que el PSA preimplante sea menor que 1,5 y la edad supere los 69,5 años para que fracase el PSA diagnosticado debe ser mayor que 18. Si esto sucede no influye que valor tenga el Gleason.

5. FORMACIÓN DE RECURSOS HUMANOS

En la línea de investigación cuyos resultados parciales se reportan en esta comunicación, se encuentran trabajando: un tesista de doctorado, dos tesista de grado y tres investigadores formados.

6. CONCLUSIONES

En primer lugar se destaca la factibilidad de aplicar modelos de minería de datos para el tratamiento de información relativa a poblaciones de pacientes médicos. Se verifican relaciones e interacciones que no se ven a simple vista y se cuantificaron consensos médicos. Se continuará el proyecto de la siguiente manera: [a] aplicando Redes de Bayes para contrastar los resultados obtenidos, y [b] Estableciendo los criterios finales que sean de utilidad para los especialistas a la hora de tomar decisiones.

7. REFERENCIAS

Britos, P., Hossian, A., García-Martínez, R. y Sierra, E., 2005. Minería de Datos Basada en Sistemas Inteligentes. Editorial Nueva Librería. Buenos Aires. ISBN 987-1104-30-8.

Chen, H. y Han J., 1996. Data Mining: An overview from database perspective. IEEE Transactions on Knowledge and Data Eng.

Kantardzic, M., 2003. Data Mining: Concepts, Models, Methods, and Algorithms. John Wiley & Sons. ISBN 0471228524.

Hartigan, J.A., 1975. Clustering algorithms. John Wiley & Sons, New York.

Quinlan, J., 1993. Programs for Machine Learning. Morgan Kaufmann Publishers. Edición 1993.

Servente, M.; García-Martínez, R., 2002. Algoritmos TDIDT Aplicados a la Minería Inteligente. http://www.fi.uba.ar/laboratorios/lsi/R-ITBA-26-datamining.pdf Acceso Enero 2008.