Aplicación del aprendizaje automático con árboles de decisión en el diagnóstico médico

Reinel Arias Montoya[1], Jhon Jairo Santa Chávez[2], Juan de Jesús Veloza Mora[3]

Resumen

Objetivo

Presentar la forma como la minería de datos es aplicable en la medicina como una técnica de clasificación que se puede utilizar para diagnosticar la existencia o no de enfermedades, con base la exploración sistemática de la información histórica disponible de casos previamente diagnosticados y documentados.

Metodología

Se enmarcan algunos problemas específicos de diagnóstico médico dentro del proceso general de descubrimiento de conocimiento en bases de datos. Se aborda con una introducción a la minería de datos y a la explicación de una de sus técnicas: La de inducción por árbol de decisión como herramienta seleccionada y desarrollada en particular. Se ilustra con dos ejemplos del campo de la medicina para mostrar la aplicabilidad de esta clase de herramientas en el área del diagnóstico médico y se analizan los resultados obtenidos en los casos abordados.

Resultados

Se evidencia la efectividad de utilizar el método de aprendizaje por árbol de decisión, para la exploración información multivariada en el apoyo a la toma de decisiones, en el área del diagnóstico de enfermedades. Los árboles de decisión obtenidos son de muy fácil entendimiento y utilización, ya que limita la labor de priorización de las variables críticas que más influyen en la respuesta.

Conclusiones

Si se dispone de una buena base de datos para soportar este proceso, la metodología de árbol de decisión puede ser una excelente herramienta como apoyo al diagnóstico temprano de enfermedades. La utilización de arboles de decisión para el diagnóstico de enfermedades lleva a la obtención de modelos fácilmente interiorizables ya que automatiza la labor de priorización de variables críticas que más influyen en la respuesta, permitiendo tener excelentes niveles de predicción con árboles poco profundos.

Palabras clave: *diagnóstico, minería de datos, neoplasias de la mama, enfermedades cardiovasculares (Fuente: DeCS, Bireme)*

Artículo recibido: *junio 19 de 2013* ***aprobado:*** *septiembre 10 de 2013*

1 Ingeniero Electricista, Magister en Investigación de Operaciones. Docente Universidad Tecnológica de Pereira. Correo electrónico: rarias@utp.edu.co.

2 Ingeniero Electricista, Magister en Ingeniería Eléctrica, Magister en Instrumentación Física, candidato a Doctor en Ingenierías. Docente Universidad Tecnológica de Pereira y Universidad Libre Seccional Pereira. Correo electrónico: jjsanta@utp.edu.co.

3 Ingeniero Electricista, Magister en Instrumentación Física. Miembro del Grupo de Investigación de Informática y docente Universidad Tecnológica de Pereira. Correo electrónico: veloza@utp.edu.co.

Application of machine learning with decision trees in medical diagnosis

Abstract

Objective
To present how data mining is applicable in medicine as a classification technique that can be used to diagnose the presence or absence of disease based on systematic exploration of historical information available from previously diagnosed and documented cases.

Methodology
The aim is to frame some specific problems of medical diagnosis within the overall process of knowledge discovery in databases. It deals with an introduction to data mining and an explanation of one of its techniques: The induction decision trees as a particularly selected and developed tool. Illustrated with two examples from the medical field to show the applicability of such tools in the area of medical diagnosis and the results obtained in both cases are discussed are analyzed

Results
The effectiveness of using the method of decision tree learning is evidenced, for exploration of multivariate information in supporting the decision making in the area of disease diagnosis. Decision trees obtained are very easy to understand and use, as it limits the work of prioritizing the critical variables that influence the response..

Conclusions
If you have a good database to support this process, the decision tree methodology can be an excellent tool to support the early diagnosis of diseases. The use of decision trees for disease diagnosis leads to obtaining simple models as easily as it automates the task of prioritizing critical variables that influence the response allowing us have excellent levels of prediction with small trees.

Key Words: *diagnosis, data mining, breast neoplasms, cardiovascular diseases (Source: MeSH)*

Introducción

El descubrimiento de conocimiento en las bases de datos, por sus siglas en inglés (knowledge discovery in databases o KDD), combina métodos de la teoría de bases de datos, de la estadística, del reconocimiento de patrones, inteligencia artificial, computación de alto desempeño y otros, describiendo el proceso de encontrar estructuras interesantes y útiles en los datos.

Por estructura se entiende patrones o modelos. Un patrón se define clásicamente como una descripción incluyente de un conjunto de datos. El proceso a recorrer a partir de los datos en bruto hasta encontrar "conocimiento" en los mismos, es un proceso largo y arduo. Los datos en bruto generados por cualquier proceso por ejemplo uno de manufactura, compras de los clientes o un grupo de datos estadísticos de personas con alguna enfermedad particular son difíciles de obtener.

El primer paso es recolectar y organizar estos datos en una bodega que suministre una visión lógica unificada de diversos aspectos del tema a considerar. Al construir una bodega de datos para tratarla específicamente con la técnica de aprendizaje por árbol de decisión se deben afrontar problemas básicos como la limpieza y consolidación de los mismos, etapa que en este caso es la discretización o categorización de las variables de entrada (e incluso de la variable de salida).

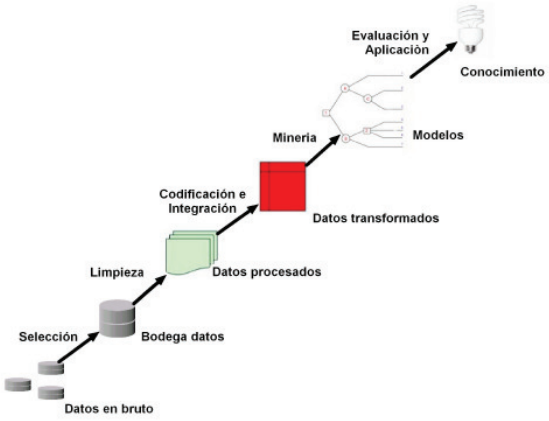
Esta discretización se puede hacer manualmente cuando los datos tienen un conjunto finito (y relativamente pequeño) de valores (dicotómicos, ordinales o nominales) o de manera manual o automatizada cuando el valor del dato es de naturaleza numérica continua. Cuando es manual, se logra definiendo rangos de valores para estas variables continuas y cuando es automático, esta automatización debe ir incluida dentro del algoritmo clasificador y es una modificación al algoritmo ID3 básico. El algoritmo C4.5 contiene estas y otras modificaciones tales como técnicas de poda y soporte a datos con valores faltantes (1).

Una ventaja de la técnica de árbol de decisión es que al modelo se le pueden introducir cualquier cantidad de variables o atributos de entrada y el algoritmo elegirá de manera automatizada los atributos más importantes, es decir, aquellos que mas impacten la variable de salida quedarán más cerca a la raíz.

A pesar de todas estas ventajas surge la pregunta: Como determinar las variables candidatas de entrada al problema? Esta es la parte realmente difícil y aquí aparecen dos opciones: una es la técnica estadística del diseño de experimentos para explorar el amplio espacio de las posibles variables de entrada y una segunda opción (la más efectiva) es la de recurrir a un experto en el área (p.e. un médico especialista) que exprese sus intuiciones y aserciones respecto a aquellas cosas que pueden influir en la variable de salida y plasmarlas en la definición de un conjunto de variables de entrada con las que se hará el muestreo, esperando que dentro de estas se encuentren aquellas que explican en mayor medida la variable de salida y que aparecerán en la raíz del árbol y sus descendientes cercanos.

El proceso KDD (Knowledge discovery in databases) involucra varios pasos tales como la selección, limpieza y transformación de los datos. (2) (Figura 1).

Figura 1. El proceso KDD (Knowledge discovery in databases)



Las herramientas de minería de datos

Existe un paso crítico en este proceso que es la minería, es decir, la búsqueda de patrones en esos datos. Para un conjunto arbitrario de datos, existen más patrones posibles que datos en sí; el objetivo de la minería de datos es saber que patrones buscar y enumerar.

Dado que la mayoría de patrones serian completamente superfluos y quizás hasta serian producto de la aleatoriedad o a veces describirían sentencias completamente obvias, es por este motivo que se hace atractiva la minería de datos al dedicarse a la búsqueda de patrones interesantes y útiles en los mismos. Así como en el apoyo en la definición del conjunto de variables de entrada, en la obtención del conocimiento, el análisis de los resultados por parte de un experto, toma importancia relevante previa su evaluación y aplicación.

Los arboles de decisión

De las herramientas disponibles en minería de datos tenemos las técnicas supervisadas, donde se encuentran los algoritmos de árboles de decisión como una de las principales herramientas predictivas que ayudan a seleccionar los atributos de mayor incidencia en una decisión de variable categórica (que usualmente pero no siempre, es dicotómica) basada en un árbol y generando una disyunción de conjunciones presentadas como un conjunto de reglas para tomar la decisión. (3)

Es un algoritmo TDIDT (top down, induction decision tree) cuya heurística principal es buscar el mejor atributo para ubicarlo en la raíz del árbol y para esto utiliza un estadístico llamado mayor ganancia de información el cual esta expresado como una diferencia de entropías según la teoría de información y el teorema de Shannon. (4)

La entropía de un conjunto *S* con *c* posibles clases dicotómicas (positivo, negativo) está dada por la siguiente expresión:

$$H(S) = -p_p \log_2 p_p - p_n \log_2 p_n$$

Donde Pi es la probabilidad de que la clase Ci tenga valor positivo

La entropía ponderada se define como el valor esperado de la entropía del conjunto

S cuando se particiona de acuerdo a un atributo *A* en particular. Esta entropía se calcula de acuerdo a la siguiente expresión:

$$H(S,A) \equiv \sum_{v \in Valores(A)} \frac{|S_v|}{|S|} H(S_v)$$

Con las dos expresiones anteriores se puede ahora obtener el estadístico de ganancia de información para cada atributo A de la información mediante la siguiente expresión:

$$H(S,A) \equiv \sum_{v \in Valores(A)} \frac{|S_v|}{|S|} H(S_v)$$

Es utilizando el mayor valor de esta ganancia que se selecciona el atributo que debe ir ubicado en la raíz del árbol.

Como segundo elemento de la heurística básica tenemos la redefinición del conjunto base para cada rama del árbol que corresponde a cada valor del atributo seleccionado en el paso anterior.

Este proceso debe repetirse recursivamente para cada nuevo nodo generado en el árbol. Cuando un nodo no tiene descendientes distintos entonces de convierte en una rama que contiene la decisión a tomar.

El modelo planteado

El modelo planteado incluye los pasos explicados en el proceso KDD desarrollados desde la selección hasta la codificación pasando por la limpieza y entregándolos a la herramienta de minería de árboles de decisión, utilizando en este caso una serie de scripts programados en lenguaje PHP con SQL para acceder a la base de datos de entrenamiento y de esta manera generar un modelo de conocimiento representado en forma de un conjunto de reglas de decisión que explican la salida en función de la entrada.

Resultados

Se analizan dos ejemplos de diagnóstico médico para ilustrar cómo a partir de un conjunto de datos tomados de muestras realizadas en pacientes, con atributos que se sospecha son influyentes en alguna medida en la enfermedad en estudio, es posible encontrar hipótesis que explican la existencia de la enfermedad con un alto porcentaje de efectividad.

En el primer ejemplo se tratará de predecir la existencia o no de enfermedad coronaria a partir de variables cardiovasculares, mientras que en el segundo caso se tratará de predecir la posibilidad de desarrollar cáncer de seno a partir de imágenes digitalizadas de muestras de tejido tomadas por el método de biopsia por aspiración con aguja fina (FNA). La técnica consiste en usar los datos disponibles para "entrenar" un modelo y luego aplicar el modelo obtenido a los datos disponibles y ver la efectividad en la predicción.

El Machine Learning Repository (5)(6)

La Universidad de California en Irvine publica desde 1987 una amplia colección de problemas en estas y otras áreas. La mayoría de ellos con datos reales aportados por diversos investigadores e instituciones y que son ampliamente referenciados en publicaciones científicas, llamada el Machine Learning Repository y que actualmente alberga más de 200 conjuntos de datos para la investigación y prueba de métodos y algoritmos de aprendizaje automático. En el análisis de la eficiencia del algoritmo utilizado en este trabajo, se hace uso de dos de los conjuntos de datos publicados en este repositorio aplicable a la medicina. En particular, los datos para

el problema de cáncer de seno fueron aportados por el Centro de Ciencias Clínicas de la Universidad de Winsconsin.

Resultados para un problema de clasificación de presencia o ausencia de enfermedades del corazón

En el primer ejemplo se toma una base de datos de 270 instancias o registros, cuyos atributos corresponden a variables cardiovasculares y que son el insumo al algoritmo de clasificación utilizado para observar su funcionalidad en un caso específico. (7)

Esta base de datos contiene 13 atributos o variables de entrada (que a su vez se han extraído de un conjunto mayor de 75 atributos):

1. Edad
2. Sexo
3. Tipo de dolor torácico (4 valores)
4. Presión arterial en reposo
5. Suero coleteral en mg/dl
6. Azúcar en sangre en ayunas > 120 mg/dl
7. Descanso resultados electrocardiográficos (valores 0,1,2)
8. Frecuencia cardiaca máxima alcanzada
9. Angina inducida por el ejercicio
10. Cldpeak: Depresión del segmento ST inducida por el ejercicio relativo al reposo.
11. La pendiente del segmento ST en ejercicio pico
12. Número de vasos mayores (0-3) coloreado por flouroscopia.
13. Thal: 3 = normal; 6 = defecto fijo; 7 = defecto reversible

Tipo de atributos de entrada:
Real: 1,4,5,8,10,12
Ordinal: 11
Binario: 2,6,9
Nominal: 7,3,13

Variable a predecir:
Fail

1. Ausencia de enfermedad cardiaca
2. Presencia de enfermedad cardiaca

Estos atributos están asociados a las siguientes variables en el algoritmo:
atributos[1]=”edad_c”
atributos[2]=”sexo”
atributos[3]=”tddt”
atributos[4]=”paer_c”
atributos[5]=”clts_c”
atributos[6]=”azsa”
atributos[7]=”drec”
atributos[8]=”fcma_c”
atributos[9]=”aipe”
atributos[10]=”dipe_c”
atributos[11]=”psst”
atributos[12]=”ndvm”
atributos[13]=”thal_c”

Se muestran cuatro registros de muestra: Dos de ellos con ausencia (1) y dos de ellos con presencia (2) de enfermedad del corazón.

70.0 1.0 4.0 130.0 322.0 0.0 2.0 109.0 0.0 2.4 2.0 3.0 3.0 **2**

67.0 0.0 3.0 115.0 564.0 0.0 2.0 160.0 0.0 1.6 2.0 0.0 7.0 **1**

57.0 1.0 2.0 124.0 261.0 0.0 0.0 111.0 0.0 0.3 1.0 0.0 3.0 **2**

64.0 1.0 4.0 128.0 263.0 0.0 0.0 105.0 1.0 0.2 2.0 1.0 7.0 **1**

A continuación se muestra la salida arrojada como resultado de una corrida del algoritmo: (8)

Total casos: 270
Casos acertados: 247
Porcentaje error: 8.52%
Tiempo ejecución=10s

Árbol resultante:

```
[] (120/270) thal_c
      [<= 4.5] (33/152) dipe_c
            [<= 2.5] (26/142) fcma_c
                  [<= 112.0] (6/7) 2
                  [> 112.0] (20/135) 1
            [> 2.5] (7/10) edad_c
       [> 45.5] (7/8) 2
       [<= 45.5] (0/2) 1
       [> 4.5] (87/118) paer_c
              [> 109.0] (86/112) tddt
       [3] (12/21) 2
            [2] (5/7) 2
                  [4] (67/76) 2
                   [1] (2/8) 1
             [<= 109.0] (1/6) 1
```

Reglas de decisión generadas

Regla 1:
Si thal_c es <= 4.5
y dipe_c es <= 2.5
y fcma_c es <= 112.0
entonces fail -> 2

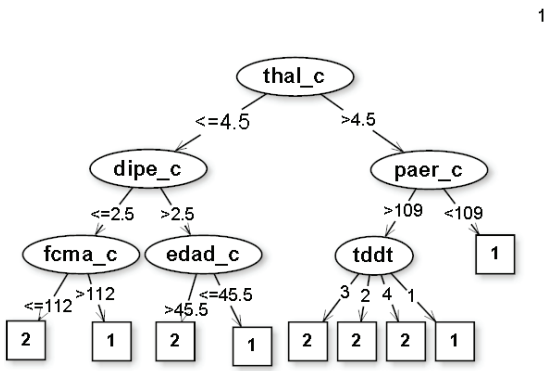
Regla 2:
Si thal_c es <= 4.5
y dipe_c es > 2.5
y edad_c es > 45.5
entonces fail -> 2

Regla 3:
Si thal_c es > 4.5
y paer_c es > 109.0
y tddt es 3
entonces fail -> 2

Regla 4:
Si thal_c es > 4.5
y paer_c es > 109.0
y tddt es 2
entonces fail -> 2

Regla 5:
Si thal_c es > 4.5
y paer_c es > 109.0
y tddt es 4
entonces fail -> 2

Representando gráficamente el árbol obtenido tenemos:



Se explica a continuación el caso de la Regla 1 entregada como resultado por el algoritmo: Esta regla se puede leer de la siguiente manera: Si el valor *tal_c* es menor o igual a 4.5, la presión arterial en reposo es menor o igual a 2.5 y la frecuencia cardiaca máxima alcanzada es menor o igual que 112.0, entonces el resultado será la presencia de enfermedad cardiaca. Observe los registros 1 y 3 de muestra que cumplen con estas condiciones expuestas por la regla 1.

De esta forma pueden ser analizados los casos para las reglas restantes en este nivel de ejecución de prueba del algoritmo. Se dice que el conjunto de reglas generadas es una disyunción de conjunciones, ya que cada regla explica una combinación diferente que arroja la presencia de la enfermedad. Para este caso de estudio el modelo presenta una tasa de error que no supera el 10,0%.

Como se mencionó previamente, en la obtención de conocimiento a partir de los datos, el análisis de resultados por parte del médico experto, es de importancia relevante previa su evaluación y aplicación. Aquí solo se pretende mostrar la capacidad de predicción del modelo generado por el algoritmo basado en el entrenamiento a través de unos datos específicos utilizados como insumo de entrada.

En la etapa de evaluación y aplicación se observa la facilidad de entendimiento, por parte del usuario destinatario, tanto del árbol como de las reglas de decisión, lo cual es tomado como referente para seleccionar el patrón hallado.

Resultados para un problema de detección previa de cáncer de seno

En el segundo ejemplo se trata de explicar la posibilidad de desarrollar cáncer de seno con base en una imagen digitalizada de muestras de tejido tomadas por aspiración con aguja fina y que describen las características de las células en la muestra. Se dispone de 685 registros de quistes (después de eliminar los registros incompletos) de las cuales 456 son benignos y 239 son malignos. (9)(10)(11)

Los atributos de cada muestra son los siguientes:

1. Grosor de la masa

2. Uniformidad del tamaño celular

3. Uniformidad de la forma celular

4. Adhesión marginal

5. Tamaño de célula de epitelio

6. Núcleo desnudo

7. Cromatinas suaves

8. Nucléolos normales

9. Mitosis

10. Clase: (2 benigno; 4: maligno)

El último atributo es la variable de salida que como se observa corresponde a la variable clasificadora entre maligno y benigno. Los atributos de entrada están asociados las siguientes variables en el algoritmo utilizado:

atributos[1]=”clump_thick”
atributos[2]=”unif_cell_size”
atributos[3]=”unif_cell_shape”
atributos[4]=”marg_adhesion”
atributos[5]=”epit_cell_size”
atributos[6]=”bare_nuclei”
atributos[7]=”bland_chroma”
atributos[8]=”normal_nucleoi”
atributos[9]=”mitoses”

Se muestran los 4 primeros registros de las muestras donde los tres primeros son negativos y el cuarto caso corresponde a un caso positivo.

5,4,4,5,7,10,3,2,1,2

3,1,1,1,2,2,3,1,1,2

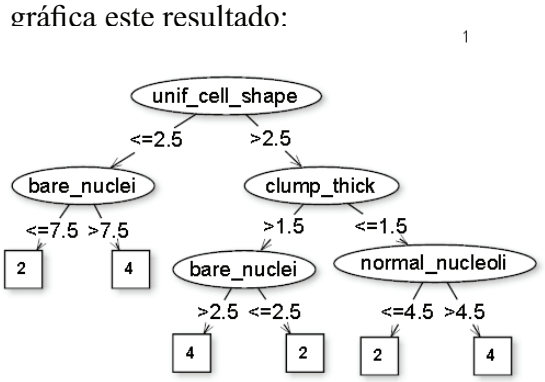
4,1,1,3,2,1,3,1,1,2

8,10,10,8,7,10,9,7,1,4

Para probar la efectividad del modelo de clasificación que arroja el algoritmo, se separan 227 registros del conjunto total de de muestras. Este conjunto de 227 datos consta de 83 positivos y 144 negativos. Los 456 registros restantes (156 positivos y 300 negativos), se usan como datos de entrenamiento para generar el árbol clasificador. La idea es probar el árbol obtenido presentándole los 227 casos separados inicialmente y determinar la efectividad de predicción del mismo ante casos nuevos, es decir, realizando lo que se conoce como una prueba no sesgada. El resultado que entrega el algoritmo es el siguiente:

```
[] (163/456) unif_cell_shape
  [<= 2.5] (4/264) bare_nuclei
      [<= 7.5] (0/260) 2
      [> 7.5] (4/4) 4
  [> 2.5] (159/192) clump_thick
      [> 1.5] (158/181) bare_nuclei
          [> 2.5] (145/153) 4
          [<= 2.5] (13/28) 2
      [<= 1.5] (1/11) normal_nucleoli
          [<= 4.5] (0/10) 2
          [> 4.5] (1/1) 4
```

A continuación se muestra de manera gráfica este resultado:



Evaluando este árbol contra los datos usados para el entrenamiento, se acierta en 445 de 456 casos, es decir, un error de predicción del 2.41%. Al hacer la prueba no sesgada probando el árbol contra los 227 casos que el algoritmo no ha visto, se acierta en 216 casos de 227, es decir, que la efectividad esperada en la predicción para datos nuevos será del 95.15%.

Discusión

Se ha enfatizado en la utilización de una técnica de minería de datos para extraer conocimiento de manera automatizada de un conjunto de datos médicos específicos y representarlo en la forma de un árbol, el cual que servirá como asistente a la hora de emitir un diagnóstico médico como un caso de clasificación. Es imprescindible el apoyo del experto a la hora de la definición de las variables de entrada al modelo asi como en la etapa de evaluación la efectividad del conocimiento extractado.

Dado que esta es una técnica supervisada, es decir, una técnica que va refinando el modelo de clasificación obtenido a medida que se le presentan más datos de entrenamiento; al utilizarla, lo mas importante es establecer políticas de retención de información histórica de calidad, insumo que servirá para evaluar y mejorar la efectividad de la predicción.

Referencias

1. Quinlan JR C4.5: Programs for Machine Learning. San Francisco. Morgan Kaufmann Publishers;1993

2. Han J, Kamber M. (2001). Data Mining: Concepts and Techniques. San Francisco: Morgan Kaufmann Publishers; 2001

3. Arias R. Detección temprana de fallas en la red de internet banda ancha aplicando minería de datos. [Proyecto de Grado] Maestría en Investigación de Operaciones. Universidad Tecnológica de Pereira, 2010.

4. Shannon CE. A Mathematical Theory of Communication. The Bell Technical Journal. 1948, 27: 379-423

5. Machine learning Repository. Centro de Aprendizaje automático y sistemas Inteligentes[Internet] [Consultado noviembre 7 de 2013] Disponible en: http://archive.ics.uci.edu/ml/

6. Jian-Per Z, Zhang-Er L, Jing Y. A parallel SVM training algorithm on large – scale classification problems. Machine Learning and Cybernetics; 2005

7. Machine learning Repository. Statlog datos (del corazón) [Internet] [Consultado noviembre 7 de 2013] Disponible en: http://archive.ics.uci.edu/ml/datasets/Statlog+(Heart)

8. Santa JJ, Veloza JDJ. Aplicación del aprendizaje automático con árboles de decisión al estudio de las variables del modelo de indicadores de gestión de las universidades públicas. [Proyecto de Grado] Maestría en Instrumentación Física. Universidad Tecnológica de Pereira; 2013.

9. Machine learning Repository. Breast Cancer Wisconsin datos (Diagnostic). [Internet] [Consultado noviembre 7 de 2013] Disponible en: http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

10. Street WN, Wolberg WH, Mangasarian OL. Nuclear feature extraction for breast tumor diagnosis. International Symposium on Electronic Imaging: Science and Technology. 1993, 1905: 861-870

11. Mangasarian OL, Street WN. Breast cancer diagnosis and prognosis via linear programming. Wisconsin. Operations Research. 1995, 43(4): 570-577