Gaceta Médica Espirituana
Univ. Ciencias Médicas. Sancti Spiritus

Vol.9, No.1 (2007)
ISSN 1608 - 8921

Trabajo Original

Facultad de Ciencias Médicas Dr. Faustino Pérez Hernández

Aplicación de métodos de selección de atributos para determinar factores relevantes en la evaluación nutricional de los niños.

Application of methods of selection of attributes to determine outstanding factors in the nutritional evaluation of children.

Lic. Roxana Martín Ramos[1], MSc. Rosa María Ramos Palmero[2], Dr. Ricardo Grau Ávalos[3], Dra. María Matilde García Lorenzo[3]

Lic. en Ciencias de la Computación. Profesor Asistente [1]
MSc. en Bioquímica. Profesor Auxiliar [2]
Dr (a). en Ciencias de la Computación. Profesor Titular [3]

RESUMEN

Las técnicas de aprendizaje automatizado constituyen un tema vigente en las investigaciones actuales, sobre todo por el amplio espectro en que pueden ser aplicadas. Existen diversos sistemas computacionales que permiten aplicar estas técnicas con muy buenos resultados. Una herramienta de gran utilidad y fácil acceso es el WEKA: sistema multiplataforma de extenso uso, probado bajo sistemas operativos Linux, Windows y Macintosh, que se encuentra disponible en Internet de forma gratuita y además es OpenSource. En el presente trabajo, se aplicaron los métodos de selección de atributos disponibles en el WEKA, versión 3.5.5, a una base de datos que contenía variables involucradas en el estado nutricional de niños de 6 a 11 años, con el propósito de precisar cuál de los métodos aplicados determinaba los factores que más aportaron a la evaluación nutricional. Pudo comprobarse que la selección de atributos brindó los factores más relevantes, cuyo comportamiento pudiera originar desviaciones del estado de nutrición, el que se encuentra bajo el sistema de vigilancia nutricional para detectar de forma temprana deficiencias nutrimentales y actuar en consecuencia.

DeCS: APRENDIZAJE, EVALUACIÓN NUTRICIONAL

SUMMARY

Techniques of automated learning are an effective topic in current investigations, mainly due to the wide spectrum in which they can be applied. There are several computational systems that allow to apply these techniques with very good results. A tool of great utility and easy access is the WEKA: a multiplatform system of extensive use, proven in Linux, Windows and Macintosh operating systems that is available in Internet in a free way and it is also OpenSource. In the present work, the methods of selection of available attributes were applied in the WEKA, version 3.5.5, to a database that contained variables involved in the nutritional state of children from 6 to 11 years, with the purpose of specifying which of the applied methods determined the factors that contributed the most to the nutritional evaluation. It could be proven that the selection of attributes provided the most outstanding factors whose behavior could originate deviations of the nutritional state, the one that is under the system of nutritional surveillance to detect in an early way deficiencies nutrimental deficiencies and to act accordingly.

MeSH: LEARNING, NUTRITION ASSESSMENT

INTRODUCCIÓN

El sistema de vigilancia nutricional materno-infantil permite la evaluación del crecimiento físico del niño y resulta una de las acciones más pertinentes a cumplir en la atención primaria.(1) Sin embargo, existen factores modificables que intervienen en la aparición de trastornos nutricionales que escapan al sistema establecido y constituyen factores de riesgo en el desarrollo de enfermedades crónicas en la edad adulta. La detección y prevención de conductas inadecuadas desde edades tempranas tendrá mayor impacto en el establecimiento de estilos de vida saludables en la adultez (2)(3).

En el empeño de lograr resultados coherentes en el campo de la salud, sus profesionales se han apoyado en diversas disciplinas que les proporcionen los instrumentos necesarios para el abordaje de los mecanismos etiológicos del proceso salud-enfermedad. Precisamente desde esa óptica ha resultado muy útil el empleo del Aprendizaje Automatizado, específicamente, de los métodos de selección de atributos, los cuales, como su nombre lo indica, reducen el número de variables, seleccionando aquellas que son “relevantes” dentro de un conjunto inicial de atributos (4). Existen diferentes métodos de selección que evalúan la calidad de los atributos basados en determinados criterios. Una de las herramientas de aprendizaje automatizado más utilizadas por la inmensa aplicabilidad que posee, y su fácil acceso, es el WEKA: sistema multiplataforma “OpenSource”, probado bajo sistemas operativos Linux, Windows y Macintosh y disponible gratuitamente en Internet, la cual ha resultado de preferencia por los especialistas del campo en los últimos tiempos.(5)

Ante la disyuntiva de determinar indicadores de riesgo metabólico en una población infantil relativamente amplia y diversa, se hizo necesario acudir a los sistemas computacionales de aprendizaje automatizado con el objetivo de precisar, utilizando varios métodos de selección de atributos, y un algoritmo de aprendizaje, los factores que más aportaron a la evaluación nutricional, así como determinar cuáles métodos de selección ofrecieron los mejores resultados.

MATERIAL Y MÉTODO

El estudio realizado respondió a un diseño observacional descriptivo.La base de datos procedía de un estudio nutricional realizado en una muestra de 278 niños de 6 a 11 años procedentes de escuelas primarias urbanas ubicadas en los 4 municipios de mayor población de la provincia espirituana. La muestra se obtuvo mediante un muestreo por conglomerado bietápico.La base de datos estuvo compuesta por variables generales como edad y sexo, además de otras específicas. Estas provenían de indicadores antropométricos, bioquímicos y dietéticos. Entre los primeros se obtuvo el índice de peso para la talla de acuerdo a las Normas Cubanas de Crecimiento y Desarrollo (6) y se clasificaron los niños de acuerdo a los puntos de corte establecidos según su ubicación en los percentiles.(7)En cuanto a los indicadores bioquímicos se obtuvo la concentración de hemoglobina en sangre y las concentraciones séricas de colesterol y triglicéridos, por los métodos vigentes aprobados por el Instituto de Nutrición e Higiene de los Alimentos (INHA).El estudio dietético se efectuó a través de una encuesta de frecuencia de consumo de alimentos, los cuales fueron clasificados en: leche, derivados lácteos, proteínas de origen animal, vísceras, embutidos, cereales, leguminosas, viandas, vegetales (con hojas y otros), frutas y grasas.

Para seleccionar las variables de mayor relevancia se utilizaron 8 métodos de selección de atributos disponibles en el WEKA, versión 3.5.5. La selección de atributos incluyó la combinación de una búsqueda, con la estimación de la utilidad del atributo, más su evaluación respecto a un esquema de aprendizaje específico.(8) En general, estos algoritmos pueden ser clasificados por varios criterios. Una categorización popular es aquella en la que los algoritmos se distinguen por su forma de evaluar atributos y se clasifican en: Filtros, donde se seleccionan y evalúan los atributos en forma independiente del algoritmo de aprendizaje y Wrappers (envoltorios), los cuales usan el desempeño de algún clasificador (algoritmo de aprendizaje) para determinar lo deseable de un subconjunto. (9) Otra taxonomía muy útil divide los algoritmos en: los que evalúan y ordenan cada atributo de forma individual y aquellos que evalúan subconjuntos de atributos. Este último grupo puede dividirse aún más atendiendo a la técnica de búsqueda comúnmente empleada con cada método para explorar el espacio del subconjunto de atributos.(8)

Se utilizaron los 4 algoritmos evaluadores de subconjuntos de atributos disponibles en el WEKA (10), los dos primeros clasificados como Filtros y los restantes como Wrappers. Se ejecutaron en combinación con el método de búsqueda Best First, el cual busca en el espacio de los subconjuntos de atributos utilizando la estrategia greedy hillclimbing con backtracking. La dirección de la búsqueda realizada por Best First fue hacia adelante partiendo del conjunto vacío de atributos.

1. CfsSubsetEval: Evalúa un subconjunto de atributos considerando la habilidad predictiva individual de cada variable, así como el grado de redundancia entre ellas. Se prefieren los subconjuntos de atributos que estén altamente correlacionados con la clase y tengan baja intercorrelación.(11)
2. ConsistencySubsetEval: Evalúa un subconjunto de atributos por el nivel de consistencia en los valores de la clase al proyectar las instancias de entrenamiento sobre el subconjunto de atributos.(12)
3. ClassifierSubsetEval: Evalúa los subconjuntos de atributos en los datos de entrenamiento o en un conjunto de prueba independiente, utilizando un clasificador. En este caso utilizamos el método J48 (árbol de decisión C4.5)
4. WrapperSubsetEval: Evalúa los subconjuntos de atributos utilizando un clasificador (también el J48). Emplea validación cruzada para estimar la exactitud del esquema de aprendizaje en cada conjunto.(13)

Los restantes 4 algoritmos empleados son evaluadores de atributos individuales y cada uno se aplicó unido al método Ranker, que devuelve una lista ordenada de los atributos según su calidad:

1. ChiSquaredAttributeEval: calcula el valor estadístico Chi-cuadrado de cada atributo con respecto a la clase y así obtiene el nivel de correlación entre la clase y cada atributo.
2. GainRatioAttributeEval: evalúa cada atributo midiendo su razón de beneficio con respecto a la clase.
3. InfoGainAttributeEval: evalúa los atributos midiendo la ganancia de información de cada uno con respecto a la clase. Anteriormente discretiza los atributos numéricos.(14)
4. OneRAttributeEval: evalúa la calidad de cada atributo utilizando el clasificador OneR, el cual usa el atributo de mínimo error para predecir, discretizando los atributos numéricos.

Con el objetivo de verificar la efectividad de la selección de atributos, se utilizó el árbol de decisión C4.5 (método J48 en el WEKA). Este algoritmo de aprendizaje fue aplicado a los datos antes y después de la selección de los atributos, para de esta forma comparar el porcentaje de individuos mal clasificados considerando todas las variables, con cada uno de los porcentajes aportados por la clasificación luego de la selección de los atributos. El árbol de decisión C4.5 fue escogido para este estudio, porque trata de enfocarse explícitamente hacia los atributos relevantes y de ignorar los irrelevantes, además es relativamente rápido, y representa uno de los algoritmos de aprendizaje más comúnmente usados en aplicaciones de minería de datos.(8)(15)

Por último, para precisar cuáles fueron los factores de mayor influencia en la evaluación nutricional, se tomaron los atributos resultantes de los dos métodos de selección que mejoraron la clasificación (tienen menor porcentaje de individuos mal clasificados que el J48 antes de la selección) y se ordenaron teniendo en cuenta la cantidad total de veces (frecuencia) con que fueron elegidos por los métodos. Para ordenar aquellos atributos que coincidían en la frecuencia, se consideró entonces, el orden con que aparecieron en la selección de los métodos evaluadores de atributos individuales.

RESULTADOS

La Tabla 1 muestra los atributos obtenidos por cada método de selección empleado. Los primeros cuatro métodos que aparecen, son evaluadores de subconjuntos de atributos (dos de ellos, Filtros y los otros dos, Wrappers) y los restantes evalúan individualmente los atributos y los ordenan por su relevancia en la clasificación.

En la Tabla 2 se observan los porcentajes de individuos mal clasificados que se obtuvieron con el algoritmo de aprendizaje J48, antes de la selección de atributos y después de realizada la misma, donde se tomaron solamente aquellos atributos seleccionados por cada uno de los métodos.

Finalmente, la Tabla 3 brinda la información referente a los atributos que tuvieron mayor influencia en la clasificación, los cuales aparecen ordenados por la frecuencia con que fueron elegidos por los métodos de selección.

DISCUSIÓN

Varios de los atributos que se muestran en la Tabla 1, coincidieron en la selección realizada por la mayoría de los métodos. En este sentido, pudo observarse que dos de los cuatro métodos evaluadores de atributos individuales (ChiSquaredAttributeEval e InfoGainAttributeEval), aportaron iguales selecciones y ordenamiento de las variables, así como, un tercer algoritmo (OneRAttributeEval) coincidió con los mencionados, en la selección de tres atributos y en el orden, de los dos primeros de ellos. El cuarto método de esta clasificación (GainRatioAttributeEval) seleccionó, en su totalidad, los mismos atributos de los dos algoritmos citados primeramente, aunque en este caso no los ordenó de la igual manera. Por otra parte, cada uno de los métodos evaluadores de subconjuntos de atributos, eligió entre 3 y 5 variables que se repitieron en las selecciones realizadas por el resto de los algoritmos de la misma y de diferente clasificación. Estos

resultados, aunque no fueron los determinantes, dieron una medida, de cuáles eran aquellos factores de mayor relevancia en la clasificación de los datos.

Los porcentajes de individuos mal clasificados que resultaron de aplicar el método J48 del WEKA y que aparecen en la Tabla 2, permitieron comprobar la efectividad de la selección de atributos. En la misma, se observa que el porcentaje obtenido al aplicar el clasificador antes de la selección de los atributos (14.75%) solo disminuyó en dos casos: al aplicar la selección con los métodos ClassifierSubsetEval (12.59%) y WrapperSubsetEval (13.67%). Es válido subrayar que ambos métodos son Wrappers, lo que trae consigo que generalmente brinden mejores resultados, al seleccionar los subconjuntos de atributos utilizando un algoritmo de aprendizaje como criterio de medida para la selección, combinado con el algoritmo de búsqueda (8)(10), por lo cual el hecho de que hayan aportado los menores porcentajes de mal clasificados, no constituye una sorpresa. Además, vale apuntar, que solamente con dos métodos de selección (ConsistencySubsetEval y OneRAttributeEval) empeoraron, de forma significativa desde el punto de vista estadístico (mayor que 1%) (8), los porcentajes de mal clasificados obtenidos; lo cual indica que de manera general podemos valorar de positiva la selección de atributos realizada.

La Tabla 3 ilustra los atributos que, finalmente, fueron seleccionados como los más significativos para la clasificación. Estos seis atributos que aparecen en la tabla, se tomaron de los subconjuntos seleccionados por los dos métodos de menores porcentajes de mal clasificados. Luego se ordenaron por la frecuencia con que aparecieron seleccionados y en los casos en que hubo coincidencia de este valor, se tomó en consideración el orden en que aparecieron en la selección de los métodos evaluadores de atributos individuales. Los atributos Leche, Lácteos y VegHoja, si bien aparecen como seleccionados por los métodos de mejores resultados, no se tomaron como atributos de mayor incidencia en la clasificación, por no aparecer seleccionados en ningún otro método. Resulta cierto que, tanto esta última decantación de los atributos, como el ordenamiento dado según la frecuencia de apariciones de los mismos en las selecciones hechas por los métodos, fueron realizados de forma manual, o sea, sin la intervención de método automático alguno, lo cual resultó acertado debido al profundo conocimiento que se tenía del problema en cuestión. (10)

CONCLUSIONES

1. A través de los métodos de selección y el algoritmo de aprendizaje utilizados, se precisó que los factores de mayor influencia en la evaluación nutricional de los niños estudiados fueron: las concentraciones séricas de triglicéridos y colesterol, el consumo de proteínas de origen animal, la edad, el consumo de vegetales sin hojas y la concentración de hemoglobina en sangre.
2. De acuerdo con la comparación que se realizó de las selecciones obtenidas por los métodos, utilizando el árbol de decisión C4.5, se determinó que los métodos de selección que ofrecieron los mejores resultados en cuanto a menores porcentajes de mal clasificados fueron: ClassifierSubsetEval y WrapperSubsetEval, ambos dentro de la categoría de Wrappers.

BIBLIOGRAFÍA

1. Jiménez S, Gay J. Vigilancia nutricional materno-infantil. Guías para la atención primaria en salud. INHA. La Habana : Editorial Caguayo SA; 2000. p. 2-13
2. González R, Baca A, Hormigo A, Villalba D, Ortega de la Cruz C, García C. Estudio de la relación entre los hábitos dietéticos y el rendimiento escolar. http://www.semg.es/revista/septiembre2001/599-602.pdf Consultado en diciembre 2006.
3. Polit D, Hungler B. Investigación Científica en Ciencias de la Salud. 6ta Ed. México : Editorial Interamericana; 2002.
4. Molina LC, Belanche Ll, Nebot Á. Evaluación de Algoritmos de Selección de Atributos. CCIA. octubre 2002. http://www.lsi.upc.es/~lcmolina/SC/html/paper/ccia02-fs.pdf Consultado en enero 2007.
5. WEKA HomePage. http://www.cs.waikato.ac.nz/~ml/weka/index.html Consultado en noviembre 2006.
6. Jordán J. Desarrollo humano en Cuba. La Habana : Editorial Científico Técnica; 1979. p. 232
7. Díaz ME. Manual de antropometría para trabajo en nutrición. INHA; 1998. p. 30.
8. Hall MA, Holmes G. Benchmarking Attribute Selection Techniques for Data Mining. Technical Report 00/10, University of Waikato, Department of Computer Science, Hamilton, New Zealand, Julio 2002. http://www.cs.waikato.ac.nz/~ml/publications/2000/00MH-GH-Benchmarking.pdf Consultado en febrero 2007.
9. Morales E. Cursos. Aprendizaje. Selección de atributos. http://ccc.inaoep.mx/~emorales/Cursos/Aprendizaje2/seleccion.pdf Consultado en enero 2007.
10. Witten IH, Frank E. Data Mining. Practical Machine Learning Tools and Techniques, 2th Ed. Morgan Kaufmann Publishers; 2005.
11. Hall MA. Correlation-based Feature Selection for Machine Learning. PhD Thesis. University of Waikato, Department of Computer Science, Hamilton, New Zealand; 1998.
12. Liu H, Setiono R. A probabilistic approach to feature selection - A filter solution. In: 13th International Conference on Machine Learning, 319-327. Morgan Kauffman; 1996.
13. Kohavi R, John GH. Wrappers for feature subset selection. Artificial Intelligence 1997; 97(1-2):273-324.
14. Lorenzo J. Selección de Atributos en Aprendizaje Automático basado en la Teoría de la Información. PhD thesis. Faculty of Computer Science, Univ. of Las Palmas. Gran Canaria; 2002.
15. Morales E, Sierra Araujo B. Aprendizaje Automático: conceptos básicos y avanzados. Aspectos prácticos utilizando el software WEKA. Pearson. Prentice Hall; 2006.

ANEXOS

Tabla 1: Atributos obtenidos por cada método de selección empleado.

Métodos de Selección de Atributos Utilizados							
Cfs	**Consistency**	**Classifier**	**Wrapper**	**ChiSquared**	**GainRatio**	**InfoGain**	**OneR**
Edad OtrosVeg Proteínas Triglic. Colest.	Edad OtrosVeg Proteínas Triglic.	Edad Lácteos Proteínas Triglic. Colest.	Leche VegHoja OtrosVeg Proteínas Triglic. Colest. Hb.	Triglic. Colest. Edad Proteínas OtrosVeg	OtrosVeg Proteínas Triglic. Colest. Edad	Triglic. Colest. Edad Proteínas OtrosVeg	Triglic. Colest. OtrosVeg. Hb. Vísceras

Tabla 2: Porcentajes de individuos mal clasificados por J48 antes y después de la selección de los atributos.

Porcentajes de individuos mal clasificados por J48								
Antes	**Después de la selección de los atributos**							
J48	**Cfs**	**Consistency**	**Classifier**	**Wrapper**	**ChiSquared**	**GainRatio**	**InfoGain**	**OneR**
14.75%	15.10%	16.18%	12.59%	13.67%	15.46%	15.46%	15.46%	16.90%

Tabla 3: Atributos de mayor influencia ordenados según la frecuencia con que se seleccionaron.

Atributos	**Frecuencia**
TRIGLICÉRIDOS	8
COLESTEROL	7
PROTEÍNAS	7
EDAD	6
OTROSVEG.	6
HB	2