Clasificación gaussiana de rasgos talasémicos y otras anemias

C. Vidal[1], J. Civera[2], J. Vicente[1], J.M. García-Gómez[1], A. Del Arco[3], A. Juan[2], M. Robles[1]

[1]Universidad Politécnica de Valencia. Grupo de Bioingeniería, Electrónica y Telemedicina. Área de Informática Médica.

[2]Universidad Politécnica de Valencia. Departamento de Sistemas Informáticos y Computación.

[3]Hospital Universitario Doctor Peset de Valencia. Servicio de Hematología.

Resumen

Con el fin de evitar el costoso análisis genético necesario para el correcto diagnóstico de algunos tipos de talasemia, en concreto la 'α-talasemia', se plantea la clasificación automática, utilizando datos de análisis clínicos, de esta hematopatología junto a otros tipos de anemias microcíticas, tanto talasemias ('β-talasemia' y 'β-δ-talasemia') como otras anemias de origen no genético ('anemia ferropénica' y 'ferropenia'), además de casos de control. También se ha discriminado por separado entre las clases 'α-talasemia' y 'anemia ferropénica', problemáticas en la práctica clínica. Se han utilizado clasificadores gaussianos, pero incluyendo registros con campos vacíos rellenados con los algoritmos Mean Imputation y Expectation-Maximization, aumentando el tamaño de la base de datos considerablemente. Se ha alcanzado un acierto del 84% al discriminar entre las seis clases y del 92% entre las dos clases utilizando todos los registros, demostrándose que la utilización de casos con datos perdidos mejora los resultados obtenidos.

1. Introducción

La talasemia es el trastorno genético más frecuente en el ser humano [1]. Los síndromes talasémicos de mayor interés clínico son los que afectan a la síntesis de las cadenas α y β, que forman parte estructural de la hemoglobina (Hb). La disminución de una de las dos cadenas rompe el equilibrio que existe entre ambas y conduce a la acumulación intracelular de la cadena no afectada, así como también a la formación de precipitados intracitoplasmáticos que producen la destrucción precoz de los hematíes y una reducción de su tiempo medio de vida.

Hoy en día, el método más fiable de diagnóstico del rasgo talasémico (especialmente la 'α-talasemia' en un paciente adulto) es la realización en laboratorios especializados de un estudio de la secuenciación del ADN talasémico.

Este estudio representa un elevado coste tanto temporal (porque debe llevarse a cabo en centros especializados con el consecuente retraso que ello supone), como económico.

Por ello, sería deseable disponer de herramientas que, utilizando información obtenida a través de los procedimientos normales (análisis clínicos), proporcionaran un diagnóstico fiable sin necesidad de recurrir a estudios genéticos posteriores.

Las técnicas de minería de datos y reconocimiento de patrones han demostrado sobradamente su utilidad en el campo de la medicina, en el apoyo al diagnóstico de multitud de patologías [2]. En el campo del diagnóstico de talasemias, en [3] y [4] se llevaron a cabo las primeras experiencias con éxito utilizando redes neuronales para discriminar entre pacientes sanos y pacientes con rasgo talasémico. En [5] se realizó un estudio sobre 1050 casos, pertenecientes a la misma base de datos utilizada en este trabajo, introduciendo la 'β-δ–talasemia' como clase, además de pacientes con 'anemia ferropénica', 'ferropenia' y casos de control. Los resultados obtenidos utilizando redes neuronales y k-vecinos superaron los calculados mediante fórmulas empíricas clásicas (índices de England-Frasier y Mentzer [6]).

Uno de los problemas que surgen al afrontar este estudio es la gran cantidad de datos que aparecen perdidos en la base de datos, debido a que no se han realizado todas las pruebas de manera sistemática a todos los pacientes, lo que no permite utilizar la totalidad de los registros disponibles. Las técnicas de reconocimiento de patrones necesitan gran cantidad de datos, por lo que sería interesante poder aprovechar toda la información.

En este trabajo se pretende estudiar, sobre el problema tratado en [5], si la utilización de los registros con datos perdidos rellenados proporciona robustez y mejora los modelos obtenidos sólo a partir de datos completos, comparando clasificadores gaussianos entrenados con datos rellenados con las técnicas Mean Imputation y Expectation-Maximization [7] con datos perdidos. Así mismo, se compararán los resultados con los obtenidos al aplicar clasificadores por los k-vecinos más próximos (K-NN) [7], utilizando sólo registros sin datos incompletos.

Así mismo, debido a la dificultad que entraña en la práctica clínica, se afronta también la discriminación específica entre 'α-talasemia' y 'anemia ferropénica'.

Por otro lado, se ha añadido al análisis un parámetro más, la *ferritina*, con poco más de la mitad de valores presentes, con el objetivo de comprobar si aporta información a los modelos.

2. Base de datos

La base de datos empleada para el desarrollo del estudio consta de 1.757 casos recopilados en el Hospital

Universitario Doctor Peset de Valencia, todos ellos pertenecientes a pacientes diagnosticados con alguna de las hematopatologías incluidas en el estudio (‘anemia ferropénica’, ‘ferropenia’, ‘α-talasemia’, ‘β-talasemia’ y ‘β-δ-talasemia’) consensuados por diferentes hematólogos del hospital, así como casos de pacientes etiquetados como normales, teniendo en cuenta que han sido personas que han acudido a la consulta por otros motivos. Los diagnósticos han sido emitidos a partir de los análisis clínicos realizados sobre muestras de sangre de los pacientes, obteniendo diferentes parámetros relativos a los índices eritrocitarios, cantidades de reserva, transporte y saturación del hierro y cantidad de hemoglobinas A_2 y fetal, así como la edad y el sexo. De entre las variables disponibles, debido a la ausencia de datos en algunas de ellas, y a partir del consejo de los expertos, se decidió inicialmente la utilización de ocho parámetros [5], *recuento de hematíes*, *hemoglobina*, *Volumen Corpuscular Medio (VCM)*, *Hemoglobina Corpuscular Media (HCM)*, *Amplitud de Distribución Eritrocitaria* (*RDW), Sideremia*, *hemoglobina A2* y *hemoglobina fetal*, reduciéndose la base de datos a 1050 casos. Para este estudio se ha añadido la variable *ferritina*, que se espera aporte información a la hora de discriminar las patologías del hierro y en un 46% de los casos vacía, y se han utilizado todos los registros disponibles. En la tabla 1 se observa la proporción de datos perdidos para cada parámetro.

Parámetro	Disponibilidad
Hematíes	96.4 %
Hemoglobina	96.3 %
VCM	96.3 %
HCM	91.0 %
RDW	72.4 %
Ferritina	54.3 %
Sideremia	83.3 %
HbA2	99.6 %
HbFetal	99.5 %

Tabla 1. *Disponibilidad de casos para cada parámetro.*

En cuanto a la distribución de los casos en las diferentes clases bajo estudio (Tabla 2), se dispone de numerosos casos de ‘β-talasemia’ y ‘β-δ-talasemia’, así como de pacientes normales, mientras que el resto de patologías están poco representadas, sobre todo las ‘α-talasemias’.

Clase	Nº Casos
α-talasemia	45
β-δ-talasemia	427
β -talasemia	787
Anemia ferropénica	74
Ferropenia	79
Normal	345

Tabla 2. *Distribución de los casos en las distintas clases.*

3. Técnicas y Metodología

En este trabajo se han utilizado clasificadores gaussianos (con una sola distribución gaussiana por clase), rellenando los campos vacíos mediante dos técnicas, Mean Imputation (MI) y el algoritmo Expectation-Maximization (EM), los cuales describimos a continuación.

Sea $\vec{x} = (\vec{x}_1,...,\vec{x}_N)^t$ una muestra extraída de una distribución gaussiana D-dimensional de media $\vec{\mu}$ y matriz de covarianzas Σ, ambas desconocidas. Los estimadores máximo-verosímiles de los parámetros desconocidos con respecto a $\vec{x}$ son bien conocidos:

$$\hat{\vec{\mu}} = \frac{1}{N}\sum_n \vec{x}_n \quad (1) \qquad \Sigma = \frac{1}{N}\sum_n (\vec{x}_n - \hat{\vec{\mu}})(x_n - \hat{\vec{\mu}})^t \quad (2)$$

Esto es, la media y matriz de covarianzas muestrales. Desafortunadamente, estas expresiones no son directamente aplicables cuando tenemos casos con campos vacíos. La técnica MI rellena los campos vacíos mediante la media de los casos presentes para la variable a la que pertenece el valor ausente. Esto es, si x_{nd} es un dato perdido, MI le asigna la media de la variable *d* sobre todos los casos en los que esta variable ha sido observada. En general, esta sencilla técnica ofrecerá una estimación precisa de x_{nd}. No obstante, es posible mejorar la precisión de esta estimación si, además de considerar la media observada de la variable *d*, también consideramos las posibles dependencias lineales entre variables. Por ejemplo, si $x_{nd'}$ es un dato observado y existe una dependencia casi-lineal entre las variables *d* y *d'*, esta dependencia permitirá mejorar la estimación de x_{nd} dada por MI.

Por su parte, el algoritmo EM rellena los datos perdidos teniendo en cuenta tanto medias observadas como dependencias lineales entre variables. En realidad, el rellenado de datos perdidos se lleva a cabo en el primero de sus dos pasos básicos, el paso E, para lo cual se necesita una estimación provisional de los parámetros desconocidos. El segundo paso básico, el paso M, actualiza dicha estimación provisional con base en (1) y (2) y los datos perdidos que se acaban de rellenar. El algoritmo EM aplica estos dos pasos iterativamente. Tras cada iteración se obtiene una nueva estimación de parámetros que será mejor o igual que la anterior en términos de verosimilitud respecto a los datos observados. Para más detalles sobre esta aplicación del algoritmo EM puede consultarse el artículo seminal de A. P. Dempster et al. [8], en cuya sección 4.1.3 se habla, precisamente, de esta aplicación.

En los casos de MI y EM, se ha implementado un suavizado sencillo de la matriz de covarianzas con la matriz identidad, probando diferentes proporciones de mezla.

Los diferentes algoritmos utilizados se han implementado en Octave bajo Linux.

En cuanto a la metodología de trabajo, con el fin de comparar los resultados obtenidos utilizando sólo registros completos y añadiendo registros con datos ausentes rellenados, se ha utilizado como método de evaluación la validación cruzada con los cuatro bloques utilizados en [5] para las pruebas sin datos perdidos, posibilitando así la comparación de los resultados al añadir la *ferritina* a esos mismos bloques. Al incluir los nuevos registros no utilizados anteriormente, se ha realizado una validación cruzada con seis bloques generados aleatoriamente para evitar el sesgo de distribuirlos en dos bloques nuevos, con muchos más campos perdidos que los cuatro bloques originales.

4. Resultados

A continuación se muestran (Tabla 3) los resultados obtenidos al discriminar entre las seis clases (6c) y entre 'anemia ferropénica' y 'α-talasemia' (2c), con los cuatros bloques de validación iniciales (4b), repitiendo los experimentos con y sin *ferritin*a, y añadiendo los nuevos casos generando seis bloques (6b).

	Con ferritina		**Sin ferritina**		
Prueba	**MI**	**EM**	**MI**	**EM**	**KNN**
6c/ 6b	17.7	17.9	16.6	16.6	-
2c/ 6b	9.2	9.2	7.6	7.6	-
6c/ 4b	22.6	22.6	21.1	21.1	19.3
2c/ 4b	10.6	10.6	8.6	8.6	12.5

Tabla 3. *Resultados (% error) obtenidos con las diferentes técnicas para los problemas de seis clases y dos clases, utilizando los cuatro bloques iniciales y con los dos nuevos bloques. Las pruebas con MI y EM se han repetido añadiendo el parámetro ferritina.*

En la tabla 4 se muestra la matriz de confusión para el problema de las seis clases y seis bloques de validación sin *ferritina*, obtenida por el mejor clasificador, en este caso utilizando datos completados mediante MI.

	A-F	**A**	**BD**	**B**	**F**	**N**
A-F	50	0	0	1	2	21
A	8	7	3	2	1	24
BD	0	1	392	15	8	11
B	7	3	38	720	2	17
F	18	0	4	1	9	47
N	42	3	6	2	5	287

Tabla 4. *Matriz de confusión para el problema de seis clases, utilizando estimación de datos perdidos mediante MI. Suavizado con un 6.5% de la matriz identidad. (A-F: Anemia Ferropénica, A: Alfa-talasemia, BD: Beta-Delta-talasemia, B: Beta-talasemia, F: Ferropenia, N: Normal).*

En la tabla 5 se observa la matriz de confusión para el problema de dos clases, utilizando los seis bloques disponibles, sin ferritina y con datos rellenados con MI.

	A-F	**A**
A-F	68	6
A	3	42

Tabla 5*. Matriz de confusión para el problema de dos clases, utilizando estimación de datos perdidos mediante MI. Suavizado con un 88.5% de la matriz identidad. (A-F: Anemia Ferropénica, A: Alfa-talasemia).*

5. Discusión

A la vista de los resultados, se observa que el hecho de utilizar el parámetro *ferritina* distorsiona los resultados en todas las pruebas, aunque el acierto con seis bloques, incluyendo la *ferritina* en la prueba, supera al obtenido sólo con los cuatro bloques sin *ferritina* y por lo tanto completos.

Este comportamiento se puede explicar debido a la falta de comportamiento gaussiano de esta variable, puesta de manifiesto en las diferentes pruebas de normalidad que se han realizado previamente sobre las variables por cada clase, ya que si se pretende utilizar clasificadores gaussianos se ha de partir de la hipótesis de que los datos se comportan de forma gaussiana, y la mayoría de las variables se pueden considerar normales en mayor o menor medida, excepto la *ferritina*.

Se deduce de los resultados que, en todos los casos, el aprovechamiento de los casos con campos vacíos proporciona una mejora en la eficacia, tanto en los clasificadores que utilizan la *ferritina* como en los que no lo hacen. Esta mejora es notable en el caso de la discriminación entre las seis clases, donde la mejora es de más de un 3% sin *ferritina* y de más de un 4% con esa variable. En el caso de la clasificación entre 'anemia ferropénica' y 'α-talasemia', las mejoras son menores y rondan el 1%.

Con respecto a la técnica K-NN, sus resultados (siempre sin casos perdidos) en el problema de las dos clases son inferiores a los obtenidos por los clasificadores gaussianos, incluso sin utilizar éstos registros con campos ausentes. Por el contrario, K-NN es superior en el problema de las seis clases sin registros con datos perdidos, necesitando los clasificadores gaussianos aumentar el tamaño de la base de datos utilizada mediante el rellenado de los datos perdidos para mejorar ampliamente los resultados.

Por otra parte, observando la matriz de confusión para las seis clases se aprecia que el error no se produce homogéneamente, sino que las clases 'α-talasemia' y 'ferropenia' no se clasifican bien, siguiendo la tendencia de la clase 'anemia ferropénica' y confundiéndose sobre todo con la clase 'normal'. A su vez, un gran número de casos 'normales' son clasificados como 'anemia ferropénica'. Los casos pertenecientes a 'β-talasemia' y 'β-δ-talasemia' son

los mejor clasificados, lo cual concuerda con la experiencia clínica. Todo esto nos lleva a la necesidad de afrontar el problema de la clasificación de estas hematopatologías desde otro punto de vista, como por ejemplo, implementando clasificadores por etapas. En el caso de la matriz de confusión para el problema de dos clases (Tabla 5), se observa que los errores son proporcionales en ambas clases.

En cuanto a la diferencia entre MI y EM como técnicas de rellenado de datos perdidos, los resultados demuestran que no existen diferencias entre ambos. La razón podría ser que las variables que en mayor número de ocasiones aparecen perdidas se encuentran poco correlacionadas con las que aparecen presentes en la mayoría de los casos, con lo que la utilidad del algoritmo EM es mímina, teniendo en cuenta que el rellenado de datos se realiza corrigiendo la media mediante la matriz de covarianzas.

6. Conclusiones y trabajo futuro

Como resultado de este trabajo se pueden extraer una serie de conclusiones importantes. En primer lugar, la clasificación automática de los diferentes tipos de talasemias y otras anemias microcíticas no genéticas es abordable mediante clasificadores gaussianos, con tasas de error del 16.6% para el problema complejo de discriminación entre seis clases, y del 7.6% para la discriminación entre 'anemia ferropénica' y 'ferropenia', problema a priori difícil de resolver.

La inclusión de la *ferritina* como variable de estudio tiene como resultado un empeoramiento en la eficacia de los mismos clasificadores sin *ferritina*, debido a la falta de normalidad de esta variable.

Por otro lado, sin tener en cuenta la *ferritina*, queda claro que la utilización de registros en principio desechados por contener campos incompletos, rellenados mediante técnicas como MI o EM, proporciona una mejora en todos los estudios realizados, tanto en la clasificación entre las seis clases bajo estudio, como en la discriminación entre 'anemia ferropénica' y 'α-talasemia'.

Finalmente, no se aprecia diferencia entre el rellenado de datos mediante MI y EM, debido a la baja correlación entre las variables ausentes con mayor frecuencia y las presentes, lo que reduce la aportación del algoritmo EM considerablemente.

Como trabajo pendiente, en el refinamiento de los clasificadores gaussianos se plantea como siguiente paso lógico la implementación del algoritmo EM con mixturas de gaussianas (es decir, más de una distribución gaussiana por clase).

En otro nivel, se está implementando un clasificador en cascada para las seis clases aquí estudiadas, realizando la discriminación en sucesivas fases: 'normal' vs. 'patológico', 'trastorno del hierro' vs. 'talasemia' y en el último paso la clasificación dentro de los dos grupos patológicos. Este clasificador se está empezando a utilizar en el servicio de hematología del hospital Doctor Peset de Valencia, integrado dentro una plataforma distribuida de ayuda a la decisión clínica [9].

7. Agradecimientos

Al Ministerio de Sanidad y al Instituto Carlos III por su financiación parcial de este trabajo mediante la red Inbiomed, así como al servicio de hematología del Hospital Dr. Peset de Valencia por su colaboración.

Referencias

[1] J. Sans-Sabrafen, C. Besses Raebel, and J. L. Vives Corrons, *Hematología clínica*, Harcourt, cuarta edition, 2001.

[2] J Bemmel, Mark A. Musen. *Handbook of Medical Informatics*. Cap. 15, 16, 18, 27 y 28. Springer, 1997.

[3] Brian S. Erler, Pamela Vitagliano, and Stephen Lee, "Superiority of neural networks over discriminant functions for thalassemia minor screening of red blood cell microcytosis," *Arch Pathol Lab Med*, vol. 119, pp. 350–354, 1995.

[4] S. R. Amendolia, A. Brunetti, P. Carta, G. Cossu, M. L. Ganadu, B. Golosio, G. M. Mura, andM. G. Pirastru, "A real time classification system of thalassemic pathologies based on artificial neural networks," *Medical Decision Making*, vol. 22, pp. 18–26, 2002.

[5] Javier Vicente, Aurora del Arco, Juan Miguel García–Gómez, César Vidal, Montserrat Robles. "Clasificación automática de rasgos talasémicos y otras anemias microcíticas e hipocromas". Congreso Anual de la Sociedad Española de Ingeniería Biomédica 2003. Libro de Actas, p.p. 69-72.

[6] Joseph J. Mazza, *Manual de hematología clínica*, Salvat, 1990.

[7] R. Duda, P. Hart, D. Stork. *Pattern Classification*. John Wiley & Sons, Inc., New York, 2001.

[8] A. P. Dempster, N- M. Laird and D. B. Rubin. "Maximum-likelihood from incomplete data via the EM algorithm". *J. Royal Statistical Society B,* 39:138, 1977.

[9] Juan M Garcia-Gomez, Cesar Vidal, Javier Vicente, Luis Marti-Bonmati, Montserrat Robles. "Medical Decision Support System for Diagnosis of Soft Tissue Tumors based on Distributed Architecture". *26th Annual International Conference IEEE Engineering in Medicine and Biology Societ.* San Francisco, EEUU, 1-5 de septiembre de 2004. (Libro de actas en prensa)

Contacto:

César Vidal Fernández

Telf. 0034 963879000 Ext. 75278

cevifer@eln.upv.es