Estudio comparativo de medidas de calidad para biclusters en microarrays

Antonio Chaparro[1], Beatriz Pontes[1], Raúl Giráldez[2]

[1] Dpto. Lenguajes y Sistemas Informáticos, Universidad de Sevilla
antcharod@alum.us.es, bepontes@lsi.us.es
[2] Escuela Politécnica Superior, Universidad Pablo de Olavide
giraldez@upo.es

Resumen La evaluación de biclusters se ha convertido en una de las tareas más importantes en el análisis de microarrays mediante técnicas de biclustering. En la actualidad existen diversas medidas de evaluación para biclusters de datos génicos. En este trabajo se realiza un análisis experimental de tres de las medidas más utilizadas actualmente. Las conclusiones obtenidas permiten establecer un ranking de prioridades a la hora de utilizar dichas medidas como elemento de evaluación.

1. Introducción

Con fin de extraer información a partir de microarrays, se han utilizado diversas técnicas de clustering [2], generalmente agrupando genes teniendo en cuenta sus relaciones funcionales sobre todas las condiciones experimentales. Sin embargo, los genes que guardan una potencial relación entre sí no tienen por qué hacerlo con respecto a todas las condiciones [9].

Las técnicas de biclustering son una variante de las técnicas de clustering, donde la búsqueda se realiza simultáneamente sobre las filas y columnas en la matriz. En el caso del análisis de microarrays, dichas técnicas son utilizadas para poder identificar grupos de genes relacionados entre sí frente a subconjuntos de condiciones experimentales [4]. Estas técnicas se basan en la idea de que no todos los genes de un microarray son relevantes para todas las condiciones, aplicando así los conceptos de clustering sobre las dos dimensiones a la vez. Dado su gran interés, los métodos de biclustering para el análisis de datos biológicos han sido ampliamente abordados en los últimos años [3,7].

Un conjunto de genes puede ser agrupado en un bicluster si éstos presentan un comportamiento similar, sin que dichos genes exhiban valores de expresión iguales o muy similares. En otras palabras, un bicluster contendrá a aquellos genes que presenten un mismo patrón de comportamiento, independientemente de la escala en la que se encuentren sus valores de expresión [1]. De esta manera, podemos hablar de dos tipos de patrones diferentes en un bicluster: patrones de desplazamiento y patrones de escalado. Estos patrones son utilizados para describir el comportamiento común de los genes en un bicluster.

La evaluación de biclusters de datos de expresión genómica es una tarea muy importante, ya que permite la identificación de buenos biclusters obtenidos a

partir de microarrays. Dicha evaluación puede realizarse una vez obtenidos los biclusters, es decir, una vez aplicado el algoritmo de biclustering y obtenidos los resultados, o de forma intrínseca al algoritmo de búsqueda, de forma que permita discriminar las distintas soluciones conforme se van generando.

En este trabajo se presenta un estudio experimental de algunas de las medidas de evaluación más utilizadas en biclustering, generando para ello distintas submatrices basadas en los conceptos de patrones de desplazamiento y escalado, y utilizando además un porcentaje variable de error.

2. Evaluación de Biclusters

En este apartado se explican las distintas medidas de evaluación de biclusters utilizadas en este trabajo.

2.1. Residuo Cuadrático Medio

La mayoría de las técnicas de biclustering basadas en la evaluación se basan en la utilización del residuo, *Residuo Cuadrático Medio* (MSR), [4]. MSR trata de cuantificar la coherencia numérica presentada por los genes y condiciones de un bicluster $\mathcal{B}$, compuesto por I filas y J columnas. MSR se define como sigue:

$$MSR(\mathcal{B}) = \frac{1}{I \cdot J} \sum_{i=1}^{i=I} \sum_{j=1}^{j=J} (b_{ij} - b_{iJ} - b_{Ij} + b_{IJ})^2 \quad (1)$$

donde b_{ij}, b_{iJ}, b_{Ij} y b_{IJ} representan el elemento de la fila i y la columna j, la media de la fila i-ésima y la columna j-ésima, y la media de la submatriz, respectivamente. Cuando los niveles de expresión de los distintos genes siguen una evolución idéntica a través de las condiciones contenidas en el bicluster $\mathcal{B}$, entonces, el valor el residuo será nulo (MSR($\mathcal{B}$)= 0). De forma general, cuanto menor sea el valor de MSR, mejor será la calidad del bicluster. Sin embargo, cuando se trate de biclusters donde los genes no presenten variación alguna, o de biclusters triviales (un sólo gen o condición), el valor del residuo será también muy bajo. Para evitar evaluar este tipo de submatrices como buenas, se hace uso de otras medidas en combinación con el residuo, como pueden ser la varianza de gen y el volumen [4,5].

2.2. Medidas basadas en la estandarización

La estandarización es el proceso mediante el cual los genes pertenecientes a un biclusters son modificados, de manera que se permita una mejor comparación entre ellos. La forma en que se realiza dicha estandarización se define a continuación:

Definición: Estandarización. *Sea $\mathcal{B}$ un bicluster con J genes e I condiciones. Sean b_{ij} cada uno de los elementos de $\mathcal{B}$, con $1 \leq i \leq I$ y $1 \leq j \leq J$.*

Se define el bicluster estandarizado de $\mathcal{B}$ como un nuevo bicluster $\mathcal{B}'$, cuyos elementos b'_{ij} cumplen que $b'_{ij} = \frac{b_{ij}-b_{Ij}}{\sigma_{g_j}}, 1 \leq i \leq I, 1 \leq j \leq J$, donde σ_{g_j} es la desviación estándar de todos los valores de expresión del gen j.

La estandarización tiene dos objetivos principales: el primero de ellos es llevar el valor de expresión de todos los genes a un mismo rango (alrededor de 0 en este caso), para poder realizar una comparación más sencilla. El segundo de ellos es homogeneizar los valores de expresión de cada gen, modificando de esta forma sus valores bajo todas las condiciones, y suavizando su representación gráfica. Hay dos medidas de evaluación que se basan en el concepto de estandarización: Error Virtual (VE) y Área de Estandarización Máxima (MSA).

Error Virtual. La principal idea de *Virtual Error* (VE) [8] es crear un patrón para cada bicluster que represente la tendencia general de todos los genes contenidos en él. Dicho patrón debe ser creado de forma que sea un buen representante del comportamiento de los genes frente a las condiciones experimentales, cuando todos ellos varíen de forma similar a través de las condiciones, con independencia de los valores numéricos concretos. VE se basa en la creación de un patrón de comportamiento para cada bicluster, por lo tanto, la calidad de dicho patrón dependerá de la forma en que éste sea creado.

En las siguiente definición se especifica cómo el patrón utilizado para el cálculo de VE es creado, a partir de un bicluster $\mathcal{B}$.

Definición 1: Patrón virtual *Dado un bicluster $\mathcal{B}$ que contenga I condiciones y J genes, se define el patrón virtual como una colección de I elementos P_i, donde cada uno de ellos viene dado por:*

$$P_i = \frac{\sum_{j \in J} b_{ij}}{J}, b_{ij} \in \mathcal{B}, 1 \leq i \leq I, 1 \leq j \leq J$$

De esta forma, cada uno de los puntos del patrón representa el valor medio de todos los genes frente a una condición determinada.

Una vez que el patrón ha sido creado, el objetivo es cuantificar en qué medida los distintos genes del bicluster se ajustan a él. Es importante tener en cuenta que para poder comparar los valores de expresión génica a los valores contenidos en el patrón de comportamiento creado, todos ellos deben pertenecer al mismo rango de valores. Por lo tanto, el patrón de comportamiento debe ser también estandarizado, creando de esta forma un nuevo patrón llamado patrón virtual estandarizado. Este proceso se muestra en la ecuación 2, donde P_i denota el valor del patrón para la condición i, y donde $\overline{P}$, σ_P denotan la media y la desviación de los valores del patrón, respectivamente.

$$P'_i = \frac{P_i - \overline{P}}{\sigma_P} \tag{2}$$

Definición 2: Error Virtual. *Dado un bicluster $\mathcal{B}$ con I condiciones y J genes, y un patrón P que contiene I valores, se define VE como la media de las diferencias numéricas entre cada gen estandarizado y los valores del patrón estandarizado para cada condición:*

$$VE(\mathcal{B}) = \frac{1}{I \cdot J} \sum_{i=1}^{i=I} \sum_{j=1}^{j=J} (b'_{ij} - P'_i)$$

Los biclusters con valores más bajos de VE son considerados de mejor calidad que aquellos que tengan un valor más alto. Esto se debe al hecho de que VE calcula la media de las diferencias entre los genes estandarizados y el patrón estandarizado. Por lo tanto, cuando más parecidos sean los genes, menor será el valor de la medida VE.

Área de Estandarización Máxima. Esta medida recibe el nombre de *Área de Estandarización Máxima* (MSA), [6] siendo la idea principal poder cuantificar el area encerrada entre los máximos y mínimos valores de expresión presentados por todos los genes estandarizados en el bicluster. De esta forma, se evalúa el area que represente la máxima diferencia entre los niveles de expresión de los distintos genes para cada condición experimental. Para cada condición, se escogen los valores máximos y mínimos de cada gen estandarizado, de forma que esos valores definan una banda a través de todas las condiciones, que permita establecer el valor MSA. A continuación se presenta una descripción formal de dicha medida.

Definición 1: Límites de un bicluster $\mathcal{B}$. *Se define el límite superior de un bicluster $\mathcal{B}$ para la condición i como*

$$M_i(\mathcal{B}) = max_j \ b_{ij}, \forall j$$

y de manera similar se define el límite inferior de $\mathcal{B}$ para la condición i como

$$m_i(\mathcal{B}) = min_j \ b_{ij}, \forall j$$

Definición 2: MSA *Se define MSA (Maximal Standard Area), MSA($\mathcal{B}'$), como el área delimitada por los límites del bicluster estandarizado para cada condición:*

$$MSA(\mathcal{B}') = \sum_{i=1}^{I-1} \left| \frac{M_i(\mathcal{B}') - m_i(\mathcal{B}') + M_{i+1}(\mathcal{B}') - m_{i+1}(\mathcal{B}')}{2} \right|$$

donde $\mathcal{B}'$ es el bicluster estandarizado.

Cuando los genes de un bicluster $\mathcal{B}$ sigan un comportamiento coherente perfecto, el valor de la medida MSA será cero. Por el contrario, MSA será más alta para aquellos biclusters que contengan genes menos correlados, ya que los límites $M(\mathcal{B}')$ y $m(\mathcal{B}')$ estarán más distantes el uno del otro. Por lo tanto, un bicluster será considerado mejor que otro si su valor de MSA es más pequeño.

3. Patrones de comportamiento

Los genes que forman un bicluster pueden seguir un comportamiento similar, al que llamamos patrón. Dichos patrones fueron introducidos en [4], aunque se encuentran formalmente descritos en [1], dónde se distingue entre dos tipos de patrones que se definen a continuación.

Sea $\mathcal{B}$ un bicluster compuesto por I condiciones y J genes, de manera que cada elemento perteneciente al bicluster venga representado por $b_{ij} \in \mathcal{B}$, podemos definir los patrones de desplazamiento y escalado como sigue [1]:

- **Patrón de desplazamiento.** Un bicluster $\mathcal{B}$ sigue un patrón de desplazamiento cuando sus valores w_{ij} se pueden obtener sumando un cierto valor β_i, que será constante para la condición i-ésima, a un valor típico (π_j) para el gen j-ésimo. Formalmente, un bicluster presenta un patrón de desplazamiento cuando sus valores se rigen por la siguiente expresión:

$$b_{ij} = \pi_j + \beta_i + \xi_{ij} \tag{3}$$

  donde ξ_{ij} representa el error cometido por el patrón para el valor b_{ij}.
- **Patrón de escalado.** La definición de patrón de escalado es análoga a la del de desplazamiento, sustituyendo el valor aditivo β_i por un factor multiplicativo α_i, tal y como se muestra en la expresión:

$$b_{ij} = \pi_j \times \alpha_i + \xi_{ij} \tag{4}$$

  donde ξ_{ij} representa el error cometido por el patrón para el valor b_{ij}.
- **Patrón combinación de los patrones anteriores.** Este patrón es la suma de las dos expresiones de los patrones de desplazamiento y de escalado, aunando el error en una sola variable:

$$b_{ij} = \pi_j \times \alpha_i + \beta_i + \xi_{ij} \tag{5}$$

En ambos casos, cuando el error cometido ξ_{ij} es 0 para todos los valores del bicluster, se dice que se trata de un *bicluster perfecto*.

4. Diseño de los experimentos

En este apartado se definen los experimentos que se ha llevado a cabo. Por una lado, se han generado biclusters basados en los patrones de comportamiento que se han definido en el apartado anterior. Posteriormente, dichos biclusters han sido evaluados utilizando las medidas expuestas en la sección 2.

4.1. Descripción de los experimentos

La finalidad de los experimentos es la de evaluar un alto número de biclusters construidos con un mismo patrón, haciéndose uso para dicha evaluación de las medidas explicadas en la sección 2. De esta forma será posible comparar cuál de las medidas mide con más certeza la calidad de los biclusters, realizando un análisis de los distintos comportamientos de dichas medidas. El patrón que se ha utilizado para generar los distintos biclusters es una combinación entre un patrón de desplazamiento y un patrón de escalado.

Para generar el tamaño de los biclusters así como como los valores de cada uno de los niveles de expresión de los genes frente a las condiciones, se han teniendo en cuenta datos estadísticos extraídos de experimentos realizados por Cheng y Church [4] para el conjunto de datos de la levadura. El número medio de genes contenidos en 100 biclusters obtenidos a partir de dicho conjunto de datos es de 84, siendo su desviación de 300; de la misma forma, el número medio de condiciones en los biclusters es 6, mientras que su desviación es 3,3. Los valores de los niveles de expresión nunca superan el valor de 600, siendo su valor mínimo el 0.

Para la realización de estos experimentos se ha decidido construir un alto número de biclusters de modo que la población de biclusters sea significativa. Concretamente se han generado dos grupos de biclusters diferenciadas por su tamaño, de forma que en el primero de los grupos, cada bicluster generado tendrá un tamaño aleatorio, obtenido a partir de los valores medios y desviaciones de genes y condiciones. Este grupo de biclusters permitirá analizar el comportamiento de las medidas frente al tamaño de los biclusters.

El segundo grupo de biclusters está compuesto por biclusters de valores aleatorios, pero de forma que todos ellos tengan el mismo volumen (mismo número de genes y condiciones). De esta manera se elimina la variación del tamaño entre los biclusters, permitiendo estudiar el comportamiento de las tres medidas de forma independiente al volumen. El tamaño común para todos los biclusters ha sido, a su vez, obtenido de forma aleatoria a partir de las medias y desviaciones del número de genes y condiciones ya comentados.

Para los dos grupos de biclusters, sus valores se han obtenido de forma aleatoria, mediante la generación de los valores π_j, α_i y β_i que produzcan un nivel de expresión en el rango adecuado. El número de biclusters obtenidos para cada uno de los dos grupos ha sido de 1100, ya que se han construido 100 biclusters para cada uno de los porcentajes de error considerados. De esta manera los 100 primeros biclusters presentarán un patrón de desplazamiento y escalado perfecto, siendo su término ξ_{ij} igual a 0. Para los 100 siguientes biclusters, el valor ξ_{ij} será calculado de forma que se le añada al valor original de expresión un 1 %. Para cada uno de los errores (desde 0 % hasta 10 %) se han construido cien biclusters por lo que existen 1100 biclusters en cada grupo de biclusters.

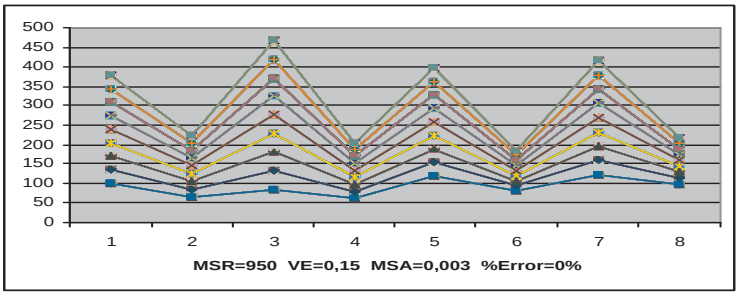

Figura 1. Bicluster bueno con MSR alto.

5. Resultados obtenidos

En este apartado se presentan los principales resultados obtenidos a partir del análisis de los experimentos ya descritos, en los que se han generado un total de 2200 biclusters (1100 del mismo volumen y otros 1100 de volumen aleatorio variable), y posteriormente han sido evaluados utilizando el error virtual (VE), residuo cuadrático medio (MSR) y area de estandarización máxima (MSA).

5.1. Biclusters de tamaño variable

Un primer resultado obtenido a partir de este grupo de biclusters es la contradicción que presenta el uso de la medida MSR. En la mayoría de estos casos se observa que MSR indica que el bicluster no es bueno a pesar de que las otras medidas apuntan a lo contrario. Véanse las figuras 1 y 2, y los valores de las tres medidas. Visualmente, estos biclusters son fácilmente reconocibles como buenos por un observador, pero MSR indica que no es así, de lo que podemos deducir que MSR no es siempre una buena medida para medir la bondad de un bicluster. Como caso particular, se puede comprobar también a partir de estas dos figuras que los valores de VE y MSA coinciden a la hora de decidir el mejor de dos biclusters, mientras que el MSR apunta en el sentido opuesto.

Por otra parte, los autores del residuo MSR establecen en su trabajo [4] un límite para esta medida, a partir del cual un bicluster no es considerado como bueno. Este límite depende del conjunto de datos que se esté utilizando, ya que esto influye en el rango numérico de los datos así como el volumen aproximado de las distintas soluciones. Para los biclusters aquí utilizados, el límite de referencia equivale a un valor alrededor de 300, ya que es el utilizado para el conjunto de datos a partir del cual se han obtenido los datos para generar estos resultados.

Como se puede apreciar en las figuras 1 y 2, los genes en ambos biclusters muestran un comportamiento común, siendo, además, el primero de ellos un bicluster perfecto ya que su error es de un 0 %. Sin embargo, el valor del residuo MSR no indica que dichos biclusters sean buenos, lo que supone una contradicción en dicha medida.

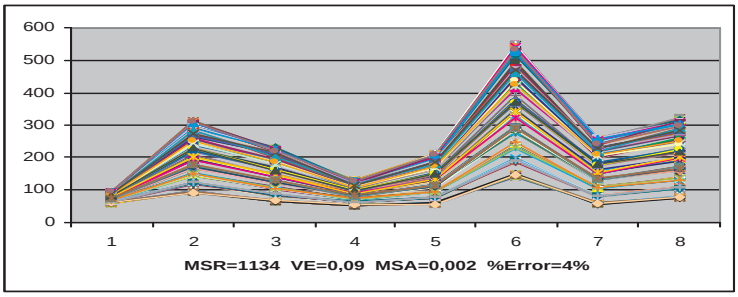

Figura 2. Bicluster bueno con MSR alto.

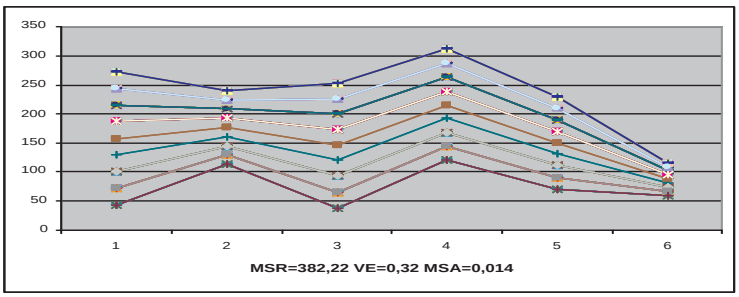

Figura 3. Bicluster con MSR aceptable.

Otro ejemplo que muestra la dificultad del residuo MSR para discernir entre buenos y malos bicluster puede verse en la figura 3, donde se muestra un conjunto de genes en los que el comportamiento no es del todo homogéneo bajo las tres primeras condiciones, ya que algunos genes tienen una tendencia al alza de sus niveles de expresión y otros a la baja. Sin embargo, aún ocurriendo esta situación, el MSR indica que este bicluster tiene mejor calidad que los ejemplos anteriores, en los que el comportamiento es homogéneo para todos los genes bajo todas las condiciones.

Analizando los valores de las medidas VE y MSA, se puede comprobar que éstas aumentan su valor para este último ejemplo, lo que indica que ambas han sido capaces de reconocer la disminución de la calidad de este bicluster con respecto a los anteriores.

En los resultados ha destacado la medida MSA frente a VE y MSR debido a que para algunos biclusters esta medida era mucho mayor que para el resto de biclusters. Como ya se ha comentado en el apartado de descripción de los experimentos, el tamaño de los biclusters se ha generado a partir de unos valores aleatorios comprendidos entre un valor medio del número de genes (84) y de condiciones (6) y su desviación (300 para el número de genes y 3.3 para el de condiciones). El número de genes es un valor mucho más alto tanto en valor medio como en desviación, de ahí que sea el factor determinante en el tamaño

del bicluster y, por tanto, de la diferencia de valores de MSA, ya que ésta se encuentra divida por el volumen total del bicluster.

5.2. Biclusters de tamaño fijo

Para poder estudiar el comportamiento de las distintas medidas de una manera independiente del tamaño, y así poder eliminar la dependencia de MSA con respecto al volumen de los biclusters, se ha realizado el mismo experimento anterior, pero manteniendo fijo el volumen de los biclusters generados. En concreto, el tamaño de biclusters utilizado ha sido de 195 genes y 8 condiciones.

Para ratificar la diferencia existente entre los valores de MSA para los biclusters de igual tamaño y los biclusters de diferente tamaño, podemos decir que la media y la desviación de estos valores es la siguiente:

	Tamaño variable	**Tamaño fijo**
Media MSA	0,008856483	0,004444377
Desviación MSA	0,014138219	0,001388656

Cuadro 1. Media y desviación de MSA para biclusters de tamaño variable y fijo.

En la tabla 1 se muestran el valor medio de la medida MSA y su desviación, calculado para los 1100 biclusters de tamaño variable (en la primera columna), y para los 1100 biclusters del mismo volumen (en la segunda columna). En dicha tabla puede apreciarse cómo la media de los valores es superior para el caso del tamaño variable, pero sobre todo lo es la desviación, que al ser superior al valor medio, hace que dicha información no sea significativa. Sin embargo, en el caso de los biclusters con tamaño constante, existe un valor medio para el MSA más fiable, con una desviación que dependerá del error cometido por cada bicluster, como se muestra a continuación.

En la figura 4 se muestra la evolución de las medias de los valores de MSA para los 100 biclusters pertenecientes a cada uno de los errores, donde se observa un claro incremento de la medida MSA conforme se aplica a biclusters con mayor porcentaje de error. Este crecimiento no es totalmente lineal pero sí mucho más constante que para las otras medidas.

Las figuras 5 y 6 muestran la evolución de las medidas MSR y VE respectivamente, para los distintos errores. Ambas muestran un comportamiento ligeramente ascendente, algo más pronunciado en el caso de VE, pero que no puede considerarse muy significativo ya que hay que tener en cuenta que se trata de los valores medios obtenidos.

Como punto común a estas tres gráficas, es posible destacar el comportamiento similar de la desviación para las tres medidas. Dicho comportamiento es homogéneo para todas las medidas, existiendo una equivalente diferencia en el rango de las medias y las desviaciones (inferior en éstas), y que indican la

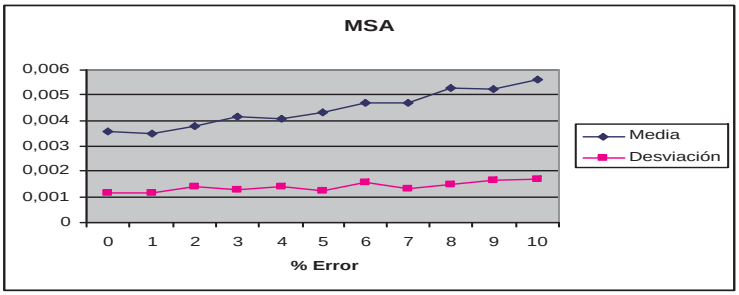

Figura 4. Media y desviación de MSA.

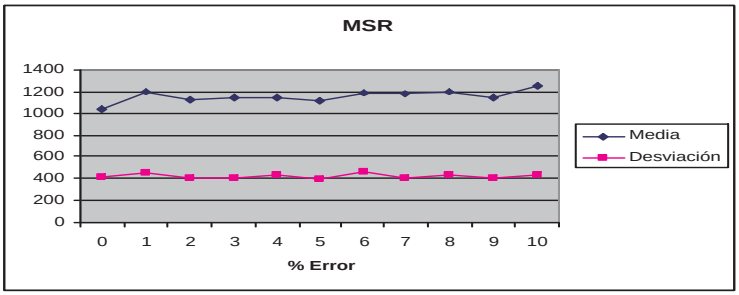

Figura 5. Media y desviación de MSR.

fiabilidad de los datos obtenidos, ya que si las desviaciones fuesen superiores no se podrían sacar conclusiones de estos resultados.

A partir de este análisis se puede destacar el comportamiento de la medida MSA frente al de VE o MSR como el más favorable a la hora de analizar biclusters, ya que presenta un mejor comportamiento a la hora de reconocer el error contenido en los biclusters.

6. Conclusiones

En este trabajo se ha presentado un estudio del comportamiento de tres medidas diferentes de evaluación de biclusters. Para realizar dicho estudio, se han generado un conjunto de biclusters obtenidos a partir de la definición de patrones de desplazamiento y escalado, y con un volumen y rango de valores obtenidos a partir de trabajos anteriores. Además, a los biclusters generados se le han añadido un cierto porcentaje de error, de forma que permita analizar el comportamiento de las diferentes medidas frente a los distintos errores.

Los resultados obtenidos muestran, por un lado que el residuo MSR no es una buena medida de evaluación de biclusters, ya que presenta un valor muy alto para ciertos biclusters clasificados como buenos o muy buenos por los patrones de comportamiento (error muy bajo o nulo) y las medidas VE y MSA. Por

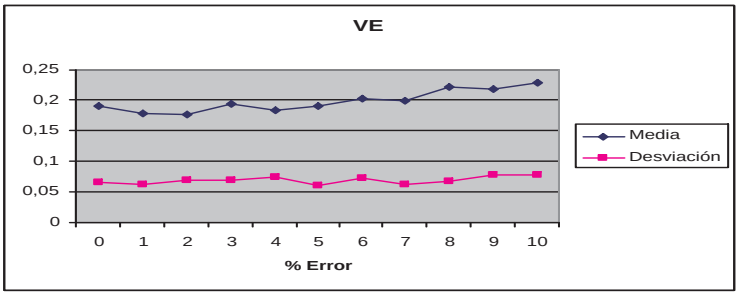

Figura 6. Media y desviación de VE.

otra parte, de las tres medidas analizadas, el área de estandarización máxima (MSA) es la que muestra un mejor comportamiento frente al error existente en los biclusters, ya que presenta una evolución ascendente más clara de los valores conforme se va aumentando el porcentaje de error.

Referencias

1. J. S. Aguilar-Ruiz. Shifting and scaling patterns from gene expression data. *Bioinformatics*, 21:3840–3845, 2005.
2. A. Ben-Dor, R. Shamir, and Z. Yakhini. Clustering gene expression patterns. *Journal of Computational Biology*, 6(3-4):281–297, 1999.
3. K. Bryan, P. Cunningham, and N. Bolshakova. Biclustering of expression data using simulated annealing. In *18th IEEE Symposium on Computer-Based Medical Systems*, pages 383–388, Dublin, Ireland, 2005.
4. Y. Cheng and G. M. Church. Biclustering of expression data. In *Proceedings of the 8th International Conference on Intellingent Systemns for Molecular Biology*, pages 93–103, La Jolla, CA, 2000.
5. F. Divina and J. Aguilar-Ruiz. Biclustering of expression data with evolutionary computation. *IEEE Transactions on knowledge & Data Engineering*, 18(5):590–602, 2006.
6. R. Giráldez, F. Divina, B. Pontes, and J. S. Aguilar-Ruiz. Evolutionary search of biclusters by minimal intrafluctuation. *2007 IEEE International Conference on Fuzzy Systems*, pages 1751–1756, 2007.
7. S. C. Madeira and A. L. Oliveira. Biclustering algorithms for biological data analysis: A survey. *IEEE Transactions on Computational Biology and Bioinformatics*, 1:24–25, 2004.
8. B. Pontes, F. Divina, R. Giráldez, and J. S. Aguilar-Ruiz. Virtual error: A new measure for evolutionary biclustering. *Fifth European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics (EvoBio 2007)*, 4447:217–222, 2007.
9. J. Wang, J. Delabie, H. C. Aasheim, E. Smeland, and O. Myklebost. Clustering of the som easily reveals distinct gene expression patterns: results of a reanalysis of lymphoma study. *BMC Bioinformatics*, 3(36):doi: 10.1186/1471–2105–3–36, 2002.