Universidad Central “Marta Abreu” de Las Villas.
Facultad Matemática, Física y Computación
Licenciatura en Ciencia de la Computación

TRABAJO DE DIPLOMA

Modificaciones al método de Grimson para la detección de conglomerados

AUTORA:

Leidys Cabrera Hernández

TUTORES:

Dra. Gladys Casas Cardoso

M.Sc Laureano Rodríguez Corvea

“Año del 50 Aniversario del triunfo de la Revolución”

3 de julio del 2009

Declaración de autoría

Hago constar que el presente trabajo fue realizado en la Universidad Central “Marta Abreu” de Las Villas como parte de la culminación de los estudios de la especialidad de Ciencia de la Computación, autorizando a que el mismo sea utilizado por la institución, para los fines que estime conveniente, tanto de forma parcial como total y que además no podrá ser presentado en eventos ni publicado sin la autorización de la Universidad.

_____

Firma del autor

Los abajo firmantes, certificamos que el presente trabajo ha sido realizado según acuerdos de la dirección de nuestro centro y el mismo cumple con los requisitos que debe tener un trabajo de esta envergadura referido a la temática señalada.

_____ _____

Firma del tutor

Firma del jefe del Laboratorio

Dedicatoria

A mis familiares

A la memoria de Antonio Florencio Herrera

Agradecimientos

Quiero agradecer a todas las personas que de una forma u otra han contribuido con su ayuda y sin la cual no hubiese sido posible la realización de este trabajo, en especial:

- *A Gladys Casas Cardoso tutora de este trabajo por su entera disposición y paciencia.*

- *A Laureano Rodríguez Corvea, también tutor de este trabajo por dedicarme momentos importantes de su tiempo.*

- *A mi papá Noel Cabrera Otero y mi mamá Magalys Hernández Herrera por estar siempre a mi lado y por ser los mejores padres del mundo.*

- *A mi novio Alejandro Capdevila Rey por toda su atención y dedicación en el tiempo que llevamos juntos, por estar a mi lado en los momentos malos y buenos, gracias por tu amor y entrega.*

- *A todos mis otros familiares por todo su apoyo, confianza y dedicación, en especial a mi abuela Carmen Herrera Benavides.*

- *A la familia de mi novio por tratarme como una hija más, a Alejandro Capdevila y Tania Rey.*

- *A Mabelyn Batule Domínguez y Grether Cabrera Peña, por demostrar ser verdaderas amigas en el tiempo que llevamos juntas y por toda su ayuda.*

- *A Dagnier Antonio Curra Sosa por su ayuda, sin la cual hubiese sido imposible avanzar.*

- *A todos mis compañeros de carrera con los cuales he disfrutado mis mejores 5 años.*

- *A todos, muchísimas gracias, ustedes forman parte de este logro. Solo ustedes saben el sacrificio que he hecho por alcanzarlo. Gracias por confiar en mí.*

Pensamiento

La mayor parte de las ideas fundamentales de la ciencia son esencialmente sencillas y, por regla general, pueden ser expresadas en un lenguaje comprensible para todos.

Albert Einstein

Resumen

La detección de conglomerados es un aspecto de gran importancia y vigencia en nuestros días. En epidemiología por ejemplo, se utiliza para detectar focos de enfermedades de manera temprana, antes que se desencadenen epidemias. En bioinformática y otras ciencias, también resuelven problemas de gran utilidad.

En este trabajo se muestra el método Grimson de detección de conglomerados. Con el objetivo de ampliar su aplicación a problemas bioinformáticos se le realiza una modificación que consiste en transformar la secuencia de datos de entrada (fechas inicialmente), en otra binaria, donde la categoría de interés se representa por uno y las demás categorías por cero. Se incluye además otra modificación utilizando la lógica borrosa para obtener mejores resultados.

Finalmente se muestran ejecuciones de los métodos con datos simulados y se concluye presentando y resolviendo un problema bioinformático.

Abstract

Cluster detection is a very important aspect and validity in our days. In Epidemiology for example, it can be used to detect clusters of sick persons in an early way, before epidemics are unchained. In Bioinformatics and other sciences, they also solve problems of great utility.

In this work, Grimson method for cluster detection is shown. In order to obtain a wider field of application, a modification is proposed. The input data (date of onset) is transformed into a binary sequence, where the value "one" represents the interest category. The other categories are represented by "zero". It is also included another modification using the fuzzy logic in order to improve the results.

Finally, an experiment is carried out with simulated data and you a real bioinformatic problem is also presented and successfully solved.

Índice

INTRODUCCIÓN 9

1. DETECCIÓN DE CONGLOMERADOS UTILIZANDO EL MÉTODO DE GRIMSON 13

1.1 TÉCNICAS DE DETECCIÓN DE CONGLOMERADOS 13
1.1.1 Conglomerados espaciales 13
1.1.2 Conglomerados temporales 14
1.1.3 Conglomerados espacio – temporales 14
1.2 EL MÉTODO DE GRIMSON CLÁSICO 15
1.3 MÉTODO DE GRIMSON PONDERADO 16
1.3.1 Análisis de un caso particular 17
1.3.2 Comparación de ambos métodos con un caso particular 18
1.3.3 Algunas consideraciones del método Grimson Ponderado 19
1.4 EL MÉTODO DE GRIMSON GENERALIZADO 20
1.5 ELEMENTOS DE LÓGICA BORROSA 21
1.5.1 Tipos de funciones de pertenencias 22
Representación explícita: 22
Representación analítica 22
Representación gráfica: 23
Función de pertenencia triangular 23
Función de pertenencia trapezoidal 23
Función de pertenencia Gaussiana 24
Función de pertenencia S 24
Función de pertenencia Gamma 25
1.5.2 Operaciones con conjuntos borrosos 26
1.5.3 Desfuzzyficación 26
1.6 MÉTODO GRIMSON BORROSO 27
1.7 DISEÑO DEL EXPERIMENTO 28
1.7.1 Análisis bifactorial no paramétrico 29
1.7.2 Algoritmo para un análisis bifactorial no paramétrico 29
CONSIDERACIONES FINALES 31

2. DETALLES SOBRE LA IMPLEMENTACIÓN DEL SISTEMA 33

2.1 ESPECIFICACIÓN DE REQUISITOS 33
2.1.1 Estructura de sus ficheros 34
2.1.2 Diagrama de Casos de Uso 38
2.1.3 Especificaciones de los Casos de Uso 40
2.2 DIAGRAMA DE ESTRUCTURA DE CLASES 46
CONSIDERACIONES FINALES 48

3 VALIDACIÓN CON DATOS SIMULADOS Y APLICACIONES. 50

3.1 VALIDACIÓN CON DATOS SIMULADOS 50
3.1.1 Bases de la simulación realizada 51
3.1.2 Validación de las variantes del método Grimson 52
3.1.3 Resultados del diseño de Experimentos 54
3.2 UNA APLICACIÓN BIOINFORMÁTICA 56
CONSIDERACIONES FINALES 60

CONCLUSIONES *61*
RECOMENDACIONES *62*
REFERENCIAS BIBLIOGRÁFICAS *63*
ANEXO *66*

Introducción

Resulta increíblemente grande el número de víctimas que las enfermedades transmisibles han causado a la humanidad. Existen evidencias irrefutables de la existencia de la tuberculosis en el Nuevo Mundo desde antes del descubrimiento de América. Se han encontrado lesiones compatibles con tuberculosis en Canadá(S 1984), Estados Unidos (Roentgenol 1925; Anthropol 1952; WA 1952; Surg 1957; MY 1979; Perzigian A 1979), Colombia (JV 1988; Correal G 1992; Sotomayor H 2004). Sin embargo, la prueba reina fue la identificación de segmentos de ADN de *tuberculosis* en lesiones de restos humanos encontrados en el norte de Chile y Perú (JV 1988). En Colombia, los hallazgos de lesiones compatibles con tuberculosis en un individuo proveniente de La Mesa de los Santos, Santander, con imágenes radiológicas de dos lesiones calcificadas en la región superior de la cavidad torácica izquierda (Correal G 1992), en la que recientemente se identificó ADN de tuberculosis (Sotomayor H 2004), los encontrados en siete sujetos de un cementerio prehispánico muisca en Soacha (JV 1988) y los observados en los restos óseos de individuos del departamento del Cauca sugestivos de tuberculosis (JV 1988) ;(Idrovo 2004).

A pesar de los notables avances en la Medicina Moderna y de los continuos esfuerzos del Ministerio de Salud Pública de Cuba por erradicar las enfermedades transmisibles, nuestro país no ha quedado fuera de su azote. Ejemplo de ello fue en 1981 los 344 203 casos de dengue reportados, 10 312 de Dengue Hemorrágico(DH) y 158 muertes (Guzmán MG 1990)-(Kouri GP 1989).

De manera general se calcula que 320 000 trabajadores de todo el mundo fallecen anualmente debido a enfermedades transmisibles causadas por riesgos biológicos representados por virus, bacterias, insectos u otras especies de animales (Trab 2007).

Debe mencionarse además, la existencia de enfermedades no transmisibles, como las malformaciones congénitas y los diferentes tipos de cánceres, que han sido imposibles de curar y que hoy en día constituyen una de las primeras causas de muerte en los países desarrollados y en Cuba.

La epidemiología por sí sola no puede dar todas las respuestas que se esperan de ella, por tanto se auxilia de ciencias como la matemática y la computación. Dentro de la matemática la estadística es la que desempeña esta labor con gran influencia. Una de las tareas más importante de los epidemiólogos es detectar a tiempo pequeños focos de

enfermedades para evitar que se conviertan en epidemias de serias complejidades y de difícil control.

Todos los países, sin importar su nivel de desarrollo en su sistema de salud, ponen énfasis en la medicina preventiva y perfeccionan constantemente su sistema de vigilancia, teniendo en cuenta tanto enfermedades transmisibles como no transmisibles. En esto, juega un papel fundamental, la detección temprana de conglomerados de enfermos. El objetivo fundamental de estas técnicas es detectar la presencia de un exceso de casos diagnosticados de una determinada enfermedad en espacio, tiempo o considerando ambos escenarios a la vez.

Las técnicas estadísticas clásicas de detección de conglomerados, (métodos jerárquicos, o de las k medias), no resuelven el problema de manera correcta, por lo que se hizo necesario desarrollar e implementar métodos matemáticos más específicos como son: el método de Scan y el de Grimson (Casas 2003).

Por otra parte, en un ámbito completamente diferente, la reciente secuenciación del genoma ha generado un amplio catálogo de miles de millones de secuencias de pares de bases. Para ayudar a los investigadores a determinar el sentido de este aluvión de datos han hecho falta numerosos instrumentos matemáticos e informáticos. Los investigadores de bioinformática aplican constantemente métodos utilizados en otros escenarios con el objetivo de extraer información relevante del código genético.

En investigaciones recientes se han desarrollado aplicaciones que ejecutan varias variantes de un método de detección de conglomerados: el método Scan (lineal y circular). (Rodríguez 2008) En este trabajo se quiere desarrollar una aplicación que ejecute variantes de otro método de detección de conglomerados: el método Grimson, con el propósito de ampliar su campo de aplicación así como facilitar el trabajo de los epidemiólogos. Se persigue además realizar validaciones del método propuesto con datos simulados y finalmente mostrar aplicaciones reales, en el campo de la epidemiología y de la bioinformática.

Para ello se formulan las siguientes preguntas de investigación:

1. ¿Cómo crear, de manera eficiente, una aplicación que contenga varias variantes del método Grimson?
2. ¿Cómo validar los métodos incorporados al software?

3. ¿Resolverán los nuevos métodos incorporados, de manera eficiente algunos problemas epidemiológicos y de bioinformática?

Para ellas se redacta el siguiente Objetivo General:

Obtener una aplicación donde se incluyan variantes del método Grimson (generalizada y borrosa), validarlos y mostrar su utilidad en la solución de problemas de bioinformática.

De manera muy resumida los objetivos específicos pudieran expresarse como:

- Diseñar un sistema computacional que implemente todas las variantes de los métodos Grimson.
- Implementar el sistema.
- Validar los métodos Grimson utilizando datos simulados.
- Demostrar de que forma los parámetros de los métodos influyen en los resultados de los métodos.
- Mostrar ejemplos bioinformáticos.

La tesis está estructurada en tres capítulos:

Capítulo 1. Detección de conglomerados utilizando el método Grimson: Se encuentran aspectos generales, definición, clasificación en tiempo, espacio, espacio-tiempo sobre las técnicas de detección de conglomerados. Se hace una explicación exhaustiva del método Grimson Clásico una primera modificación donde se generaliza el método para ser usado en la solución de problemas de bioinformática y una segunda modificación donde se emplean algunos elementos de la lógica borrosa para optimizar los resultados del método. Incluye además epígrafes donde se tratan elementos de lógica borrosa.

Capítulo 2. Detalles sobre la implementación del Sistema: Se muestran los detalles del diseño e implementación del software, así como los pasos a seguir para que el usuario tenga conocimientos de cómo utilizar el sistema.

Capítulo 3. Validación con datos simulados y Aplicaciones: Se hace un análisis de los resultados sobre varias secuencias simuladas que difieren en composición y tamaño. Se presenta y resuelve un problema actual de bioinformática: la detección de conglomerados de “gaps” sobre secuencias alineadas del virus de la influenza A H1N1.

Capítulo 1

1. Detección de conglomerados utilizando el método de Grimson

Según la literatura especializada se denomina conglomerado o *cluster* a un exceso de casos en un área geográfica determinada (conglomerado espacial), en un período de tiempo limitado (conglomerado temporal), o considerando dominios espacio temporales (conglomerado espacio-temporal).

1.1 Técnicas de detección de conglomerados

En los últimos años han surgido numerosos métodos capaces de detectar *clusters* en espacio, en tiempo o considerando ambos escenarios a la vez. No existen técnicas globales que puedan aplicarse a todas las situaciones, por eso hay gran diversidad de métodos con la misma finalidad. (Casas 2002).

1.1.1 Conglomerados espaciales

En ocasiones surgen preguntas sobre cuando una enfermedad resulta ser o no más prevalente en ciertas localidades o áreas específicas.

Como ejemplos concretos se pueden mencionar: el método de Cuzick and Edwards, (Cuzick 1990) (Jacquez 1994) y el de Pearson entre muchos otros. En (Marshall 1991) se puede consultar una revisión detallada de muchas de estas técnicas.

Algunos de estos tests trabajan con coordenadas espaciales *(X, Y)* que representan el lugar de residencia o de trabajo de la persona enferma, otros dividen la región que se estudia en pequeñas áreas y determinan allí la tasa de incidencia de la enfermedad en cuestión.

Cada uno de ellos, según sus particularidades es capaz de detectar si la distribución de enfermos en una determinada zona es superior a lo esperado. Si esto ocurre puede afirmarse que nos encontramos ante el origen incipiente de una posible epidemia(Casas 2003).

1.1.2 Conglomerados temporales

En el caso temporal se han desarrollado técnicas que detectan conglomerados de enfermos en series de tiempo cortas (Jacquez (1996 a), en las que no pueden aplicarse las técnicas tradicionales de modelación desarrolladas para series cronológicas,(Jenkins 1994). Estos tests trabajan mayormente con fechas de diagnóstico, de primeros síntomas o de muerte en aquellos casos en que se estudie la mortalidad. En ocasiones no se necesitan fechas propiamente dichas sino, por ejemplo, incidencia de casos diarios, semanales o mensuales.

Al igual que en el espacio, existen métodos específicos que trabajan sobre unos u otros datos y que permiten determinar la presencia de una agrupación de casos superior a lo esperado. Entre los más comunes se pueden mencionar los métodos Scan, (Nauss 1982); (Nagarwilla 1996) y (Larsen 1973).

1.1.3 Conglomerados espacio – temporales

La interacción espacio temporal no es la simple presencia de conglomerados de enfermos en espacio y en tiempo. Para que exista un *cluster* en espacio-tiempo tiene que ocurrir que los casos cercanos geográficamente coincidan con los casos cercanos en tiempo.

Este tipo de patrón es útil cuando se desean investigar enfermedades transmisibles, pues para que ocurra la transmisión de una persona enferma a una sana, se necesita del contacto personal directo o indirecto (a través de vectores). También se pone de manifiesto la interacción, cuando la causa de la enfermedad es la exposición a un agente geográficamente determinado: una sustancia tóxica, ciertos tipos de radiaciones, entre otras causas. Los individuos que están a la misma distancia del agente en cuestión, reciben dosis similares en el mismo período de tiempo, y por tanto manifiestan síntomas similares. (Casas 2003).

Entre los métodos más conocidos para detectar interacciones se pueden mencionar el de Knox,(Knox 1964 b), el de Mantel,(Mantel 1967) y el del k vecino más cercano, (Jacquez (1996 a) combinando de alguna forma lo anteriormente mencionado para los otros métodos.

1.2 El método de Grimson Clásico

El test de Grimson se considera uno de los métodos más generales y versátiles en la detección de conglomerados de enfermos porque puede utilizarse en la detección de conglomerados de enfermos en espacio, tiempo y en espacio tiempo.

Este método parte de dividir la región de estudio (espacial, temporal o espacio-temporal) en regiones más pequeñas llamadas celdas. Se dice que una celda está "marcada" si contiene una cantidad de enfermos superior a un cierto umbral definido por un especialista, matemáticamente este umbral pudiera determinarse por el valor esperado de una distribución de Poisson, (Grimson 1991).

El estadígrafo de Grimson es la cantidad "**A**" de celdas marcadas que son adyacentes. Dos celdas son adyacentes si ambas comparten uno de sus bordes. En el caso espacial puede entenderse por bordes las fronteras comunes que sean mayores que un único punto, en el caso temporal los bordes son límites entre dos intervalos de tiempo consecutivos y en el caso espacio – temporal los bordes estarán dados por la unión de las dos definiciones anteriores.

La hipótesis fundamental del test enuncia que bajo el supuesto de la no existencia de clusters, "**A**" debe ser pequeña, pues las celdas marcadas deben distribuirse con cierta uniformidad por toda la región. Existen fórmulas específicas para el cálculo del valor esperado de A (E(A)) y de su varianza (V(A)) que fueron publicadas en (Grimson 1991) y que aparecen detalladas más adelante en este documento. La hipótesis alternativa por su parte, enuncia que existe una gran número de celdas marcadas adyacentes, es decir, "**A**" tiene un valor demasiado grande sugiriendo así la presencia de al menos un conglomerado.

Sean :

c: número total de celdas,

m: número total de celdas marcadas, ($m < c$)

Yi: número de celdas adyacentes a la celda i, i = 1,2,…, c, o equivalentemente el número de bordes de la celda yi.

Para efectuar los cálculos necesarios, el método de Grimson necesita conocer el promedio de bordes de las celdas $(\bar{y})$ y su varianza $(V(y))$; ambas se calculan a partir de las fórmulas clásicas.

$$\bar{y} = \frac{\sum_{i=1}^{c} y_i}{c} \quad \text{y} \tag{1.1}$$

$$V(y) = \frac{c\sum_{i=1}^{c} y_i^2 - \left(\sum_{i=1}^{c} y_i\right)^2}{c(c-1)} \tag{1.2}$$

Como ya se había mencionado, el estadístico A de Grimson representa el número observado de pares de celdas marcadas adyacentes. Bajo la hipótesis nula, A tiene distribución normal asintótica con valor esperado y varianza que se calculan según fórmulas específicas desarrolladas en (Grimson 1991).

$$E(A) = \frac{\bar{y}\, m(m-1)}{2(c-1)} \tag{1.3}$$

$$V(A) = E(A)\left[1 + \frac{2(\bar{y}-1)(m-2)}{c-2} + \frac{(c\bar{y} - 4\bar{y} + 2)(m-2)(m-3)}{2(c-2)(c-3)} - E(A)\right] + V(y)\left[\frac{m(m-1)(m-2)}{(c-1)(c-2)(c-3)} - \frac{m(m-1)(m-2)(m-3)}{(c-1)(c-2)(c-3)(c-4)}\right] \tag{1.4}$$

Como caso particular, si c es grande y m/c pequeño, A sigue una distribución de Poisson con parámetro $E(A)$, (Jacquez (1996 a).

1.3 Método de Grimson Ponderado

El método Grimson Ponderado o Pesado es una variante que utiliza fórmulas distintas a Grimson Clásico pero que debe obtener iguales resultados. Su formulación se encuentra publicada en el mismo artículo (Grimson 1991).

En él se utiliza una matriz W que en principio representa una relación cualquiera de distancias entre las celdas. W es una matriz cuadrada de dimensión c x c. A partir de ella se calculan los siguientes parámetros:

$$S_0 = \sum_{i \neq j}^{c} W_{i\,j} \tag{1.5}$$

$$S_1 = \frac{1}{2}\sum_{i \neq j}^{c} 2\left(W_{i\,j} + W_{j\,i}\right) \tag{1.6}$$

$$S_2 = \sum_{i=1}^{c}\left(W_{i\,j} + \sum_{j=1}^{c} W_{j\,i}\right) \tag{1.7}$$

El estadígrafo del método de Grimson Ponderado **A'** representa la suma de los valores numéricos sobre la relación de pares de celdas marcadas.

Para obtener la esperanza matemática y la varianza de **A'** se utilizan las fórmulas que se describen a continuación, y las enunciadas desde (1.5) hasta (1.7)

$$E(A') = \frac{S_0\, m(m-1)}{2c(c-1)} \tag{1.8}$$

$$V(A') = \frac{1}{4} - \left[\frac{S_1\, m(m-1)}{c(c-1)} + \frac{(S_2 - 2S_1)m(m-1)(m-2)}{c(c-1)(c-2)} + \frac{\left(S_0^2 + S_1 - S_2\right)m(m-1)(m-2)(m-3)}{c(c-1)(c-2)(c-3)}\right] - E(A')^2 \tag{1.9}$$

Bajo la hipótesis nula, **A´** tiene distribución normal asintótica con valor esperado y varianza que se calculan según fórmulas (1.8) y (1.9). (Grimson 1991).

1.3.1 Análisis de un caso particular

Cuando la matriz utilizada W se utiliza para calcular los parámetros S_0, S_1 y S_2 se define de la manera siguiente:

$$W_{i\,j} = \begin{cases} 1 & \text{celda i adyacente a la celda j} \\ 0 & \text{en otros casos} \end{cases}$$

se cumple que:

$$E(A) = E(A')$$

y

$$V(A) = V(A')$$

Analicemos el caso del valor esperado:

$$E(A') = \frac{S_0\, m(m-1)}{2c(c-1)}$$

sustituyendo (1.5) se tiene que:

$$E(A') = \frac{S_0\, m(m-1)}{2c(c-1)} = \frac{\sum_{i\neq j}^{c} W_{ij}\, m(m-1)}{2c(c-1)} = \frac{\sum_{i\neq j}^{c} W_{ij}}{c}\,\frac{m(m-1)}{2(c-1)} \qquad (1.10)$$

Pero $\bar{y}$ es el promedio de los bordes de las celdas. Luego no es difícil concluir que:

$$\frac{\sum_{i\neq j}^{c} W_{ij}}{c} = \bar{y} \qquad (1.11)$$

De donde se llega a la expresión:

$$E(A') = \bar{y}\,\frac{m(m-1)}{2(c-1)} = \frac{\bar{y}\, m(m-1)}{2(c-1)} = E(A) \qquad (1.12)$$

Con la varianza ocurre lo mismo, sólo que los cálculos son mucho más complicados

1.3.2 Comparación de ambos métodos con un caso particular

Supongamos que se tiene un problema que conduce a la siguiente configuración de celdas:

Celdas: 10

Umbral: 5

Celdas marcadas: 5

Parámetros	Valores
Estadígrafo A	4
Media Y	1.8
Varianza Y	0.16
Esperanza(A)	2
Varianza(A)	0.66666
Significación	0.00715

Tabla 1.1. Resultados Grimson Clásico

Parámetros	Valores
Estadígrafo A’	4
Valor de S0	18
Valor de S1	36
Valor de S2	136
Esperanza (A’)	2
Varianza (A’)	0.66666
Significación	0.00715

Tabla 1.2. Resultados Grimson Pesado

Como puede observarse, los valores de los estadígrafos en ambos métodos coinciden. Lo mismo ocurre con los valores esperados y con las varianzas y consecuentemente con las significaciones.

1.3.3 Algunas consideraciones del método Grimson Ponderado

El método Grimson Pesado es computacionalmente más costoso que su variante Clásica, pues necesita para sus cálculos una matriz W que será mayor en la medida en la que aumente la cantidad de celdas a considerar.

1.4 El método de Grimson Generalizado

Como se explicó con anterioridad, el método de Grimson surge para darle respuestas a un problema epidemiológico: la detección temprana de conglomerados de enfermos.

En este epígrafe se muestra una generalización, para aumentar el dominio de aplicación del test. Para ello son necesarias algunas transformaciones.

En los métodos originales de detección de conglomerados temporales, la variable de interés es el tiempo en que ocurre el evento. Dicho evento pudiera ser la fecha de diagnóstico de la enfermedad o incluso, la fecha en la que aparecieron los primeros síntomas si esta es suficientemente precisa.

El primer paso del algoritmo consiste en ordenar cronológicamente los datos obtenidos. Posteriormente se divide el eje que representa el tiempo total considerado en intervalos fijos que puede ser años, meses o días.

A partir de este punto cada método sigue sus propios pasos para determinar si existen conglomerados. Todos estos algoritmos pueden transformarse para ampliar su campo de aplicación. La idea que se defiende es generalizarlos de manera que ellos puedan utilizarse para detectar conglomerados en un sentido más universal.

Para lograrlo se propone ordenar los datos por algún criterio determinado que depende del campo de aplicación. Si se trabaja, en problemas de bioinformática, con secuencias de bases que representan algún gen completo, o una porción de este, sería correcto asumir que tal juego de datos ya está ordenado.

El segundo paso consiste en transformar dicha secuencia en una secuencia análoga, pero dicotómica. El valor “uno” se colocara cada vez que aparezca la categoría de interés: una base, un aminoácido o una subsecuencia determinada dentro de una secuencia del ADN, ARN o de alguna proteína específica, una fecha (en problemas médicos) u otro evento que se considere. El valor “cero” se asociará a todas las demás categorías. Los datos transformados se representan en una línea. El nuevo problema que surge es el de determinar si en la secuencia dicotómica existen conglomerados de unos.

Por ejemplo, supóngase que se tiene una determinada secuencia de un gen y que dentro de ella resulta de interés determinar si existen conglomerados de la subsecuencia GCG. La transformación de la secuencia original en una dicotómica se realiza como se muestra en la siguiente figura:

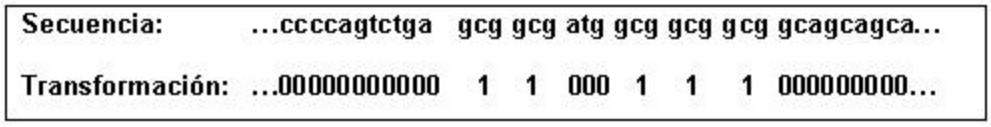

Figura 1.1: Ejemplo de conversión de una porción de la secuencia de un gen.

Obsérvese que la categoría de interés: subsecuencia GCG, se sustituyó por un uno, mientras que el resto de los casos considerados se sustituyó por el valor cero (Casas 2008).

Obsérvese además que el largo de la secuencia no se conserva, sino que disminuye.

Se pretende entonces, aplicar el método de Grimson ya explicado a esta secuencia de datos binarios. A este método se le llamará Grimson Generalizado.

1.5 Elementos de Lógica Borrosa

La lógica borrosa (o difusa) es algo que ha venido desarrollándose desde siglos atrás, hace 2500 años Aristóteles consideraba que existían ciertos grados de veracidad y falsedad y Platón había trabajado con grados de pertenencia.(Buckley 2006).

Un **conjunto borroso** es aquel que no está formado por números sino por etiquetas lingüísticas.

Una **etiqueta lingüística** es una palabra o conjunto de palabras, que representan los nombres de los conjuntos borrosos.

En los conjuntos clásicos se sabe si un elemento de un universo de discurso pertenece o no a él acudiendo a la lógica booleana. Es decir, estos conjuntos se pueden definir con un predicado que asigne a cada elemento del conjunto el valor 0 ó 1, en función de su pertenencia al conjunto. En los conjuntos borrosos esto no es posible. Así, cada elemento tendrá un valor asociado dentro del conjunto que indicará en qué “cantidad” pertenece a dicho conjunto. Esto es lo que se define como grado de pertenencia. Por ello, un conjunto borroso es la unión de los grados de pertenencia de todos aquellos elementos que forman parte de su universo de discurso.

El **universo de discurso** de un conjunto borroso es el intervalo en el que se incluyen los posibles valores que pueden tomar los elementos del conjunto. Con independencia de los valores que formen este universo, debe indicarse que siempre estará normalizado al intervalo [0,1]. Entonces un conjunto borroso A definido sobre un universo X es un par de

la forma (x, φA(x)). Donde φA(x) es llamada la función de pertenencia o membresía (MF) para el conjunto A. La MF asigna a cada elemento de X un grado de pertenencia en el intervalo [0,1]. A X seria el universo de discurso y puede ser un espacio discreto o continuo.

La existencia del grado de pertenencia para saber si un elemento pertenece a un conjunto o no, puede utilizarse para tratar problemas de imprecisión o incertidumbre en bases de datos, reconocimiento de patrones, clasificación, entre otras (Buckley 2006).

1.5.1 Tipos de funciones de pertenencias

Las funciones de pertenencia se pueden representar de tres formas distintas: representación explícita, analíticamente y gráficamente.

Representación explícita:

A= {(1, **0**), (2,**0**),..., (5, **0.1**),..., (10, **0.5**),..., (12, **0.8**),..., (16,**1**),..., (19, **0.9**),..., (35, **0.07**)}

A= **0**/1 + **0**/2+...+ **0.1**/5 +...+ **0.5**/10 +... + **0.8**/12 +... + **1**/16 +...+ **0.9**/19+... + **0.07**/35

Representación analítica

Triangular.

$$triangle(x;a,b,c)=\begin{cases}0, & x\le a,\\ \dfrac{x-a}{b-a}, & a\le x\le b,\\ \dfrac{c-x}{c-b}, & b\le x\le c,\\ 0, & c\le x\end{cases}$$

Figura 1.12 Representación analítica de la función triangular

Representación gráfica:

Seria realizar el gráfico correspondiente con los valores de los pares de la representación explícita

A continuación se mostrarán algunas funciones de pertenencia típicas:

Función de pertenencia triangular

Se define por sus límites inferior a y superior b, y el valor modal m, tal que a < m < b.

$$A(x)=\begin{cases}0 & \text{si } x \le a \\ (x-a)/(m-a) & \text{si } x \in (a,m] \\ (b-x)/(b-m) & \text{si } x \in (m,b) \\ 0 & \text{si } x \ge b\end{cases}$$

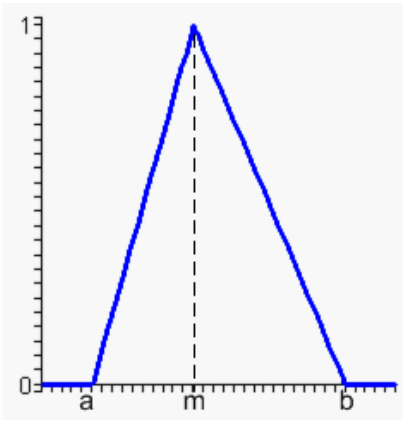

Figura 1.13 Función de pertenencia triangular

También puede representarse así: A(x; a, m, b) = máx {mín {(x-a)/ (m-a),

(b-x)/ (b-m)}, 0}

Función de pertenencia trapezoidal

Definida por sus límites inferior a y superior d, y los límites de su soporte, b y c, inferior y superior respectivamente.

$$A(x)=\begin{cases}0 & \text{si } (x \le a) \text{ o } (x \ge d) \\ (x-a)/(b-a) & \text{si } x \in (a,b] \\ 1 & \text{si } x \in (b,c) \\ (d-x)/(d-c) & \text{si } x \in (b,d)\end{cases}$$

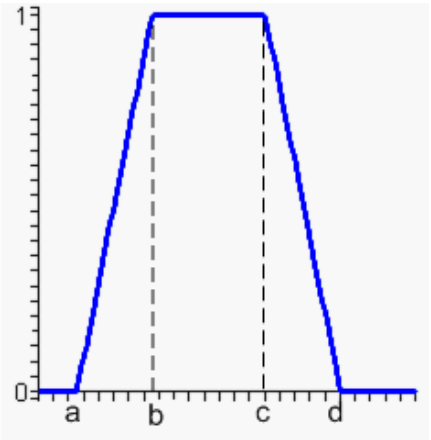

Figura 1.14 Función de pertenencia trapezoidal

Función de pertenencia Gaussiana

Definida por su valor medio m y el valor k>0.

Es la típica campana de Gauss. Cuanto mayor es el valor de k, más estrecha es la campana

$$A(x) = e^{-k(x-m)^2}$$

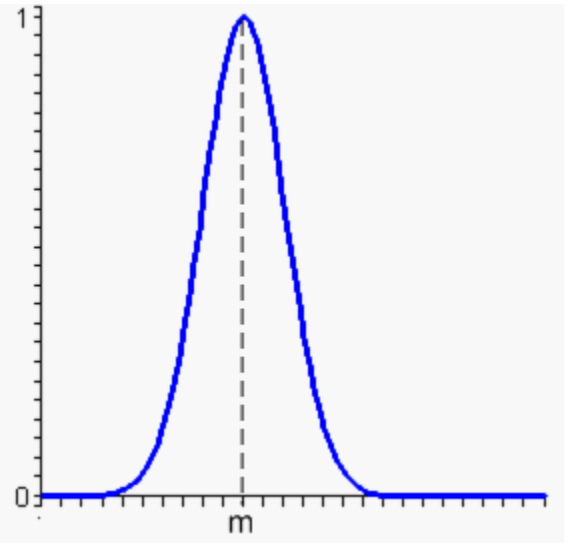

Figura 1.15 Función de pertenencia Gaussiana

Función de pertenencia S

La Función S está definida por sus límites inferior a y superior b, y el valor m, o punto de inflexión tal que a < m < b.

Un valor típico es: m=(a+b) / 2. El crecimiento es más lento cuanto mayor sea la distancia a-b.

$$A(x)=\begin{cases}0 & \text{si} \quad x \le a \\ 2\,\{(x-a)/(b-a)\}^2 & \text{si} \quad x \in (a,m] \\ 1-2\,\{(x-b)/(b-a)\}^2 & \text{si} \quad x \in (m,b) \\ 1 & \text{si} \quad x \ge b \end{cases}$$

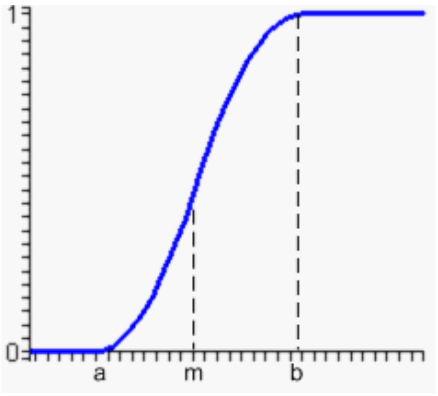

Figura 1.16 Función de pertenencia S

Función de pertenencia Gamma

Está definida por su límite inferior a y el valor k>0.

$$A(x)=\begin{cases}0 & \text{si} \quad x \le a \\ 1-e^{-k(x-a)^2} & \text{si} \quad x > a \end{cases}$$

$$A(x)=\begin{cases}0 & \text{si} \quad x \le a \\ \dfrac{k(x-a)^2}{1+k(x-a)^2} & \text{si} \quad x > a \end{cases}$$

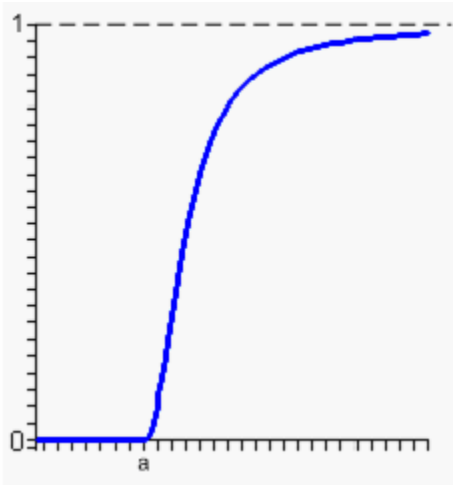

Figura 1.17 Función de pertenencia Gamma

Esta función se caracteriza por un rápido crecimiento a partir de a.

Mientras más grande sea el valor de k, el crecimiento es más rápido. La primera definición tiene un crecimiento más rápido. No llegan a tomar el valor 1, aunque tienen una asíntota horizontal en él.

1.5.2 Operaciones con conjuntos borrosos

Subconjunto:

El conjunto borroso A está contenido en el conjunto borroso B (es un subconjunto de B) o A es menor o igual que B, si y solo si, $\varphi B(x) \geq \varphi A(x)$ para todo x.

Unión (disyunción):

El resultado es un conjunto en el que se encuentren todos aquellos elementos de ambos conjuntos, cuya función de membresía se define como:

$\varphi C(x)$= máx ($\varphi A(x)$, $\varphi B(x)$). Es decir, la unión es el conjunto borroso más pequeño que contiene A y a B.

Intersección (conjunción):

El resultado de esta operación entre dos conjuntos borrosos, será un conjunto borroso en el que se encuentren aquellos elementos que están en ambos conjuntos. Si lo pensamos según sus funciones de pertenencia, se puede afirmar que: si se superponen ambas gráficas, la intersección es la zona en la que coinciden ambas funciones, cuya función de membresía se define como:

$\varphi C(x)$= mín ($\varphi A(x)$, $\varphi B(x)$)

Negación:

Esta es una operación válida sobre un único conjunto. La negación se define como todos aquellos elementos si forman parte de su universo de discurso, pero no forman parte de él, $\varphi \neg A(x) = 1 - \varphi A(x)$

1.5.3 Desfuzzyficación

Desfuzzyficar no es más que el proceso de encontrar el valor del universo que mejor "represente" al conjunto borroso. Para ello se traza una recta perpendicular al eje de las abscisas (donde se representa el universo del conjunto borroso), el punto donde se interceptan esta recta y ese eje es el valor que se asume. Los diferentes métodos de defuzzyficación determinan el lugar por donde se traza esa recta. Se pueden emplear varios métodos:

- Centroide del área: esta técnica del centroide o centro de gravedad encuentra el punto "balance" de la solución borrosa calculando la media ponderada de esa región borrosa. Dada una región borrosa A, la expresión de cálculo es:

$$z = \frac{\sum_{i=0}^{n} d_i * \varphi_A(d_i)}{\sum_{i=0}^{n} \varphi_A(d_i)}$$

Donde d es el i-ésimo valor del universo y φ (d) es el valor de pertenencia para ese punto. Este método es uno de los más usados por varias razones: los valores defuzzyficados tienden a moverse suavemente alrededor de la superficie borrosa, es decir, cambios en la topología del conjunto borroso provocan cambios suaves del valor defuzzyficado. Además es relativamente fácil de calcular.

- Medio de máximos: es el promedio de los valores del universo de discurso donde se alcanza el máximo grado de pertenencia.

- Menor de máximos: es el valor menor del universo de discurso con el cual se alcanza el máximo grado de pertenencia.

- Mayor de máximos: es el valor mayor del universo de discurso con el cual se alcanza el máximo grado de pertenencia.

1.6 Método Grimson Borroso

Producto de un experimento realizado con datos simulados para determinar las limitaciones del método Grimson generalizado, se observó que su mayor limitación es la cantidad de falsos positivos que detecta, es decir detecta conglomerados sobre secuencias que no los tienen.

Además se observan que los resultados en general dependen del ancho de la ventana disjunta y del umbral por tanto son de naturaleza borrosa, luego se puede aplicar la teoría

de la Lógica Borrosa para mejorar la clasificación en las secuencias de falsos conglomerados.

Por la naturaleza del método Grimson el parámetro umbral es el más factible a suavizar su borde inferior, esto implica alteraciones en el cálculo del estadístico, la cantidad de términos que se suavizan en el extremo inferior se conoce como el grado de incertidumbre o suavizamiento empleado.

Los conceptos anteriormente planteados pueden se utilizados en cualquiera de las variantes del método Grimson ya analizadas solo hay que modificar el cálculo del estadígrafo del la siguiente forma:

Se construye un arreglo que contiene en cada posición el peso correspondiente a cada celda según la siguiente función de pertenencia, donde valor es igual a la frecuencia de unos de cada celda:

Función de Pertenencia

0.8 si valor = umbral
0.9 si valor = umbral +1
1 si valor > umbral+1
0 en otro caso

Luego A = A + (Arreglo[i] + Arreglo[j]) / 2 *Wij.

Posteriormente se continúa el método igual a la forma tradicional.

1.7 Diseño del Experimento

Si se conocen qué efectos producen sobre la respuesta los parámetros: umbral y tamaño de la ventana, al mismo tiempo que se obtengan las interacciones entre ellos, se podrán orientar valores adecuados de estos parámetros para optimizar el método Grimson. Esto puede lograrse con un experimento factorial que permite estudiar simultáneamente varios factores y sus interacciones, de modo que los tratamientos se forman por todas las posibles combinaciones de los niveles de los factores.(Montgomery 2008).

Nuestros parámetros no tienen distribución normal y es conocido que, en las versiones del SPSS y de otros paquetes estadística avanzada, no están implementadas las posibilidades de hacer análisis de varianza multifactorial no paramétrico. Las alternativas

clásicas para resolver problemas de este tipo se basan en el uso de Técnicas de Análisis de Datos Cualitativos (Categóricos). Si en particular la variable dependiente puede ser dicotomizada, o discretizada en categorías nominales, puede utilizarse la Regresión Logística Binaria o Multinomial que permite detectar efectos de factores principales o de interacciones. Cuando todas las variables predictivas son categóricas, se puede pensar en el uso de procedimientos Loglinear o Probit del SPSS.

1.7.1 Análisis bifactorial no paramétrico

Existe un fundamento teórico de cómo puede realizarse tal análisis en el caso de diseños equilibrados. La idea esencial fundamentada por R.R. Sokal and F. J. Rohlf, 1995 (Sokal and Rohlf 1995) fue elaborar un Análisis de Varianza Bifactorial No Paramétrico ranqueando la variable dependiente, como lo hace el test de Kruskal-Wallis. Se utilizan las sumas de cuadrados de la variable dependiente ranqueada y se recalculan los grados de libertad de cada factor y su interacción para ofrecer finalmente una significación de cada efecto. Si algún factor tiene más de dos niveles, se pueden utilizar tests de comparaciones múltiples clásicos que se basan fundamentalmente en rangos para obtener subconjuntos homogéneos, por ejemplo, el test de Dunnet C, válido incluso ante falta de homogeneidad de varianzas.

1.7.2 Algoritmo para un análisis bifactorial no paramétrico

1. Ranquear la variable dependiente.
2. Aplicar el Análisis de Varianza sobre la variable dependiente ranqueada, para obtener la suma de cuadrados (SC) por cada factor y su interacción, así como sus grados de libertad.
3. Calcular el CMT (Cuadrado Medio Total)

   $$CMT = \frac{abn(abn+1)}{Total-de-datos}$$

   Donde,

   a: es el número de niveles del primer factor

   b: es el número de niveles del segundo factor

   n: es el número de réplicas de cada combinación
4. Calcular el estadígrafo H para cada factor y la interacción

$$H = \frac{SC(correspondiente)}{CMT}$$

5. Calcular la significación de cada H utilizando la distribución de Chi-cuadrado, teniendo presente los grados de libertad del factor o de la interacción analizada. (La variable H tiene distribución Chi-cuadrado).

Para facilitar el trabajo con el algoritmo anterior se han programado tres funciones simples en el paquete *Mathematica* 6.0 una de ellas utiliza el contexto de ANOVA dentro del paquete de Análisis de Varianza, para realizar el análisis paramétrico a la variable ranqueada.(Pavel 2008).

Consideraciones finales

En este capítulo se ha realizado una revisión bibliográfica sobre el método Grimson. Esta técnica surge para dar respuesta al problema epidemiológico de detección de conglomerados de enfermos. Se explicó el método clásico en sus tres variantes: temporal, espacial y espacio-temporal.

Bajo el nombre de Grimson Ponderado, se presentó una generalización del test, que utiliza una matriz de distancias W para realizar los cálculos de la significación. Esta variante es más general que la anterior, pero es también más costosa computacionalmente debido a que hay que almacenar la información de la matriz W.

Por su parte, el método de Grimson Generalizado es en esencia, el mismo método clásico (o ponderado). La transformación fundamental que se propone es sobre los datos de entrada, de manera que su aplicabilidad se pueda "generalizar" y extender más allá de los problemas epidemiológicos que trabajan con datos tipo fechas o coordenadas (X,Y) de residencia de los enfermos. El campo de la bioinformática es uno de los más beneficiados con esta generalización.

Finalmente en el capítulo se muestra el aporte fundamental de esta tesis: el método de Grimson Borroso. Esta variante flexibiliza el test considerando el umbral como un parámetro borroso.

La implementación de modificaciones que mejoran los resultados del método Grimson Clásico posibilita a los epidemiólogos y los investigadores bioinformáticas y de otras áreas, un trabajo más amplio en aplicación las técnicas de detección de conglomerados.

Capítulo 2

2. Detalles sobre la implementación del sistema

El software elaborado se utiliza para la detección de conglomerados (posibles epidemias en un ambiente epidemiológico, pero en general "conglomerados" de sucesos). Se ejecuta sobre *Windows* y brinda al usuario un ambiente cómodo. El sistema consta de un fichero: Grimson.exe. Lee los datos a partir de un fichero texto con una estructura explicada en el epígrafe 2.1.1.

El sistema está implementado en Borland Delphi. Usa las facilidades de las componentes visuales del lenguaje en aras de brindar un ambiente cómodo al usuario. Se elaboró según el paradigma de la programación orientado a objetos. A la hora del diseño se tuvo en cuenta el diseño de un ambiente sencillo.

Como primer paso fue necesario crear una estructura de datos que facilitara la llamada a todos los procedimientos y funciones que serían implementados, con parámetros similares.

2.1 Especificación de requisitos

En este epígrafe se realiza una breve descripción de los principales requisitos del sistema que se implementa.
El sistema debe ser capaz de:

- Capturar los datos de entrada para las ejecuciones del método deseado desde un archivo texto, con valores 0 y 1. Los valores 1 significan categoría de interés y 0 el resto de las categorías.
- Todos los métodos están sujetos a dos parámetros generales: ancho de la ventana y valor del umbral.
- Los parámetros mencionados anteriormente tienen las siguientes características:
    - Ancho de la ventana: no puede ser un valor vacío.
    - Ancho de la ventana: valor entero mayor que uno.
    - Valor del umbral: valor entero que no puede ser mayor que el ancho de la ventana.

- El usuario debe especificar la dirección de los dos ficheros textos que se utilizan antes de mandar a ejecutar cualquier método sino se muestra un mensaje de error.
- Los resultados deben ser mostrados en una forma de la aplicación y son guardados en el fichero texto especificado.

2.1.1 Estructura de sus ficheros

El sistema ayuda a los epidemiólogos en la detección temprana de focos epidémicos mediante la utilización del método Grimson y en algunas aplicaciones bioinformáticas donde es necesario saber la presencia de conglomerados.

Se utilizan solo dos ficheros textos, uno en el cual se entrará la(s) secuencia(s) de datos y otro donde se devolverán los resultados, la dirección de estos ficheros se especifica por el usuario. Debido a las transformaciones aplicadas y explicadas en el epígrafe 1.3, el fichero de datos se mostraría de la siguiente forma:

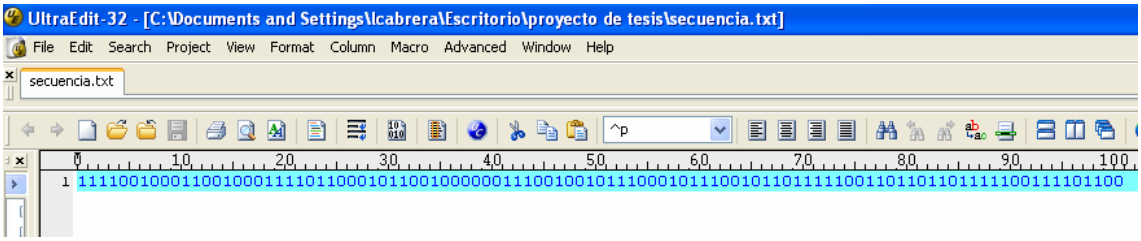

Figura 2.1.Procesamiento de una secuencia.

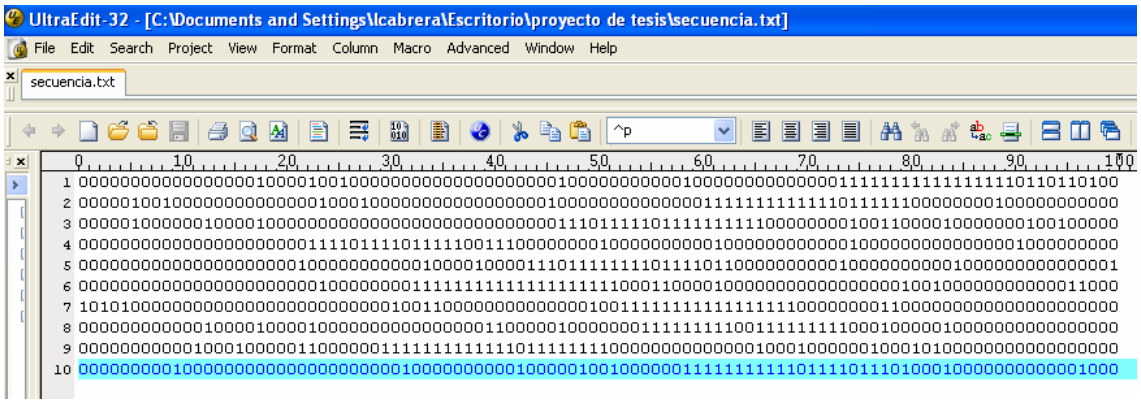

Figura 2.2. Procesamiento en lote con 10 secuencias.

En el fichero se guardan los datos de forma distinta según el tipo de procesamiento escogido pero esto no quiere decir que la información de ejecuciones anteriores se pierda, en él se van agregando los resultados consecutivamente y especificándose la hora exacta de cada ejecución.

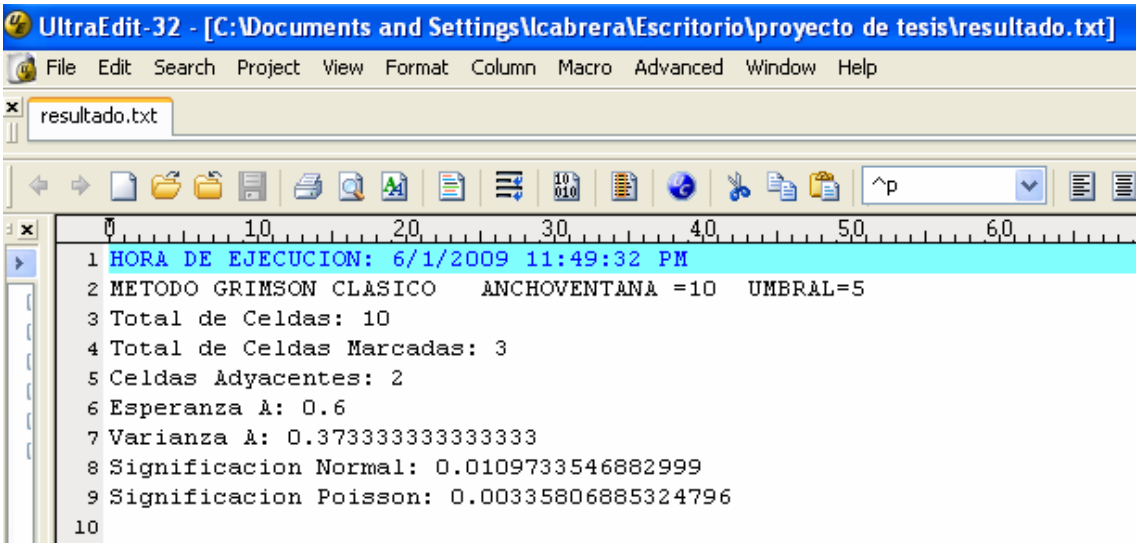

Figura 2.3. Resultado del procesamiento de una secuencia.

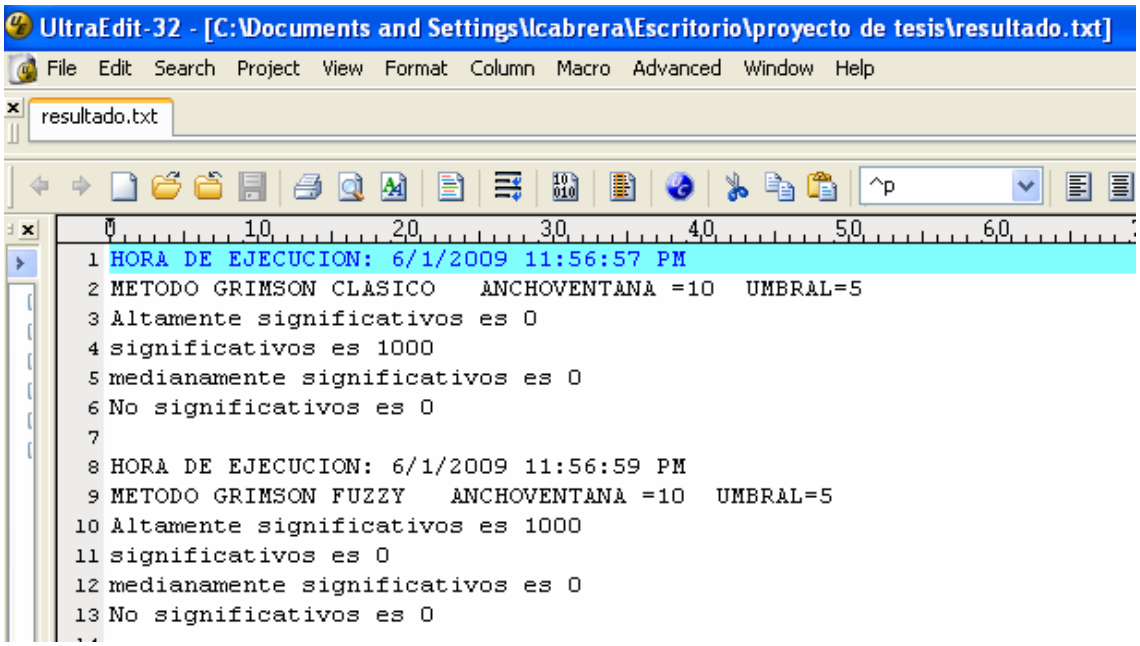

Figura 2.4. Resultado del procesamiento en lote

El sistema se presenta al usuario mediante la interfase siguiente con el nombre 'Variantes_Grimson', donde se entran los datos necesarios, se selecciona el tipo de procesamiento y el método que se desea ejecutar, además el usuario debe especificar el fichero de donde se van a leer los datos y donde se guardan los resultados a través del botón 'Configurar', para salir de la aplicación con el botón 'Salir':

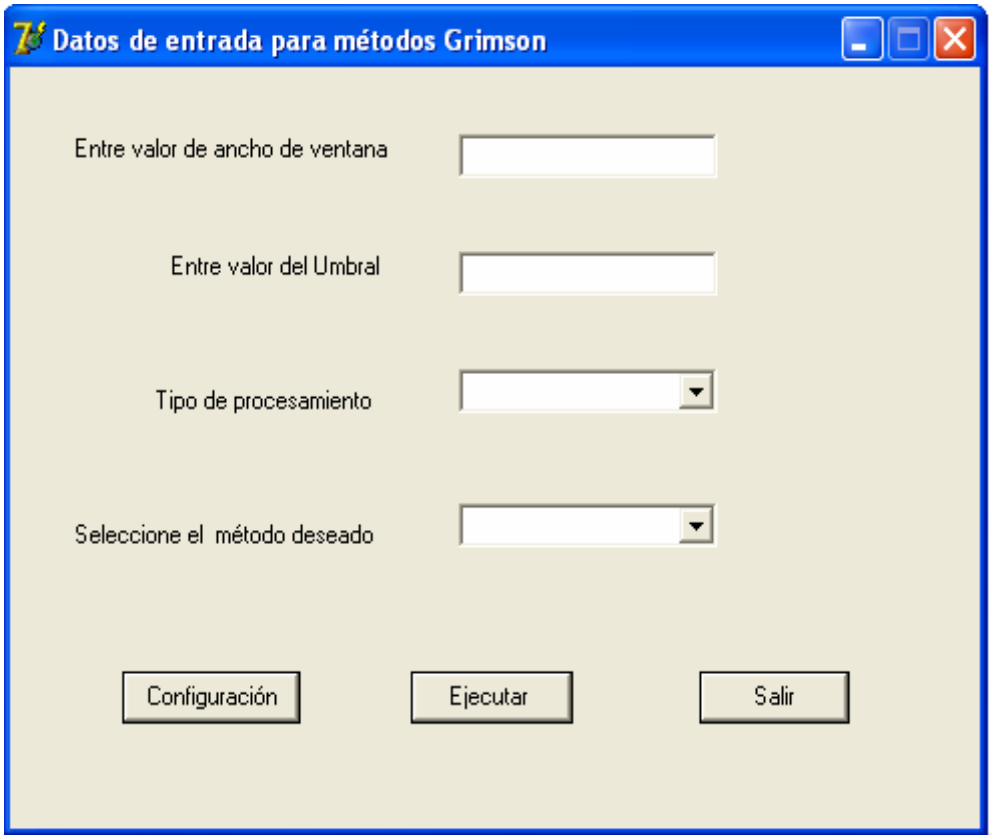

Figura 2.5 Interfase del Sistema 'Variantes_Grimson'.

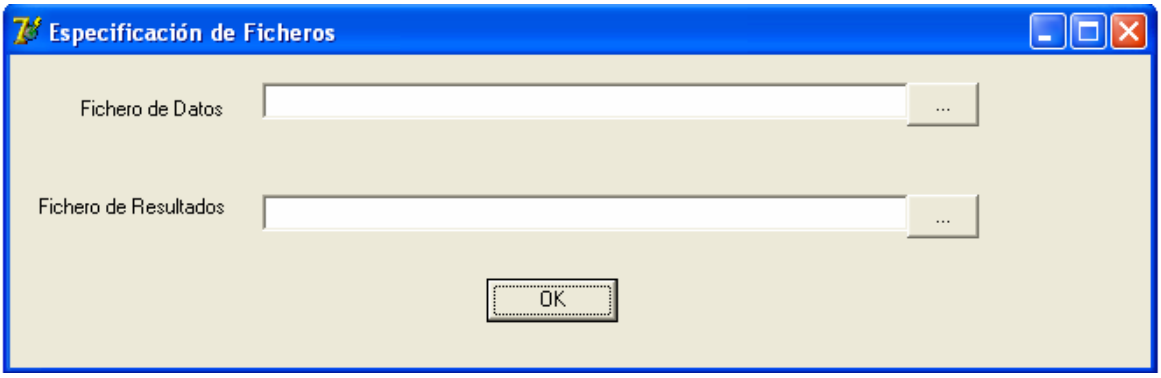

Figura 2.6 Especificación de ficheros

Cuando el usuario ejecuta cualquiera de los métodos, y el tipo de procesamiento es una secuencia se muestra una forma que contiene todos los resultados, la cual permite al usuario apreciarlos en el instante.

Sin embargo si se escogiera el procesamiento por lotes el interés estaría en saber la cantidad de casos significativos que detecta el método seleccionado, lo cual también se puede apreciar en otra forma:

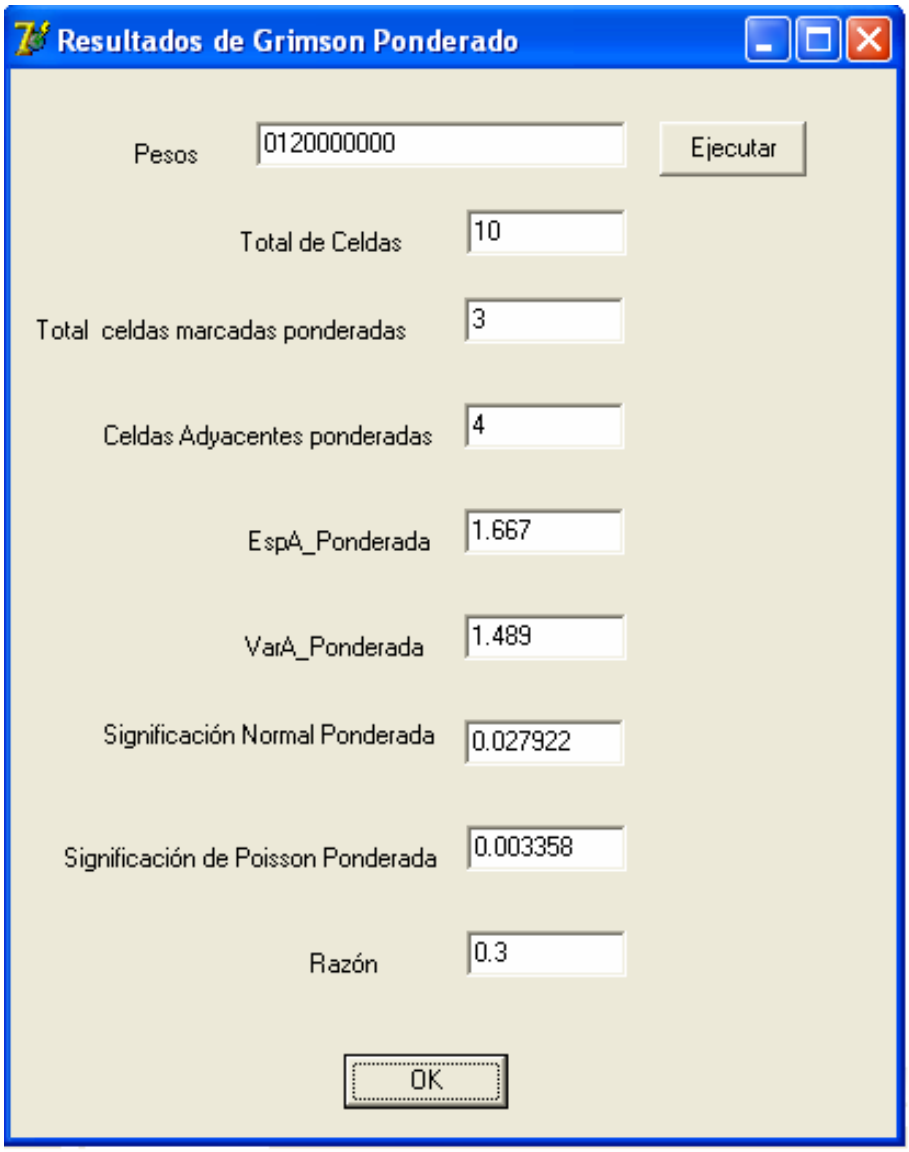

Figura 2.7 Resultados cuando el procesamiento es una secuencia.

Las formas correspondientes con los otros métodos son análogas a la anterior, solo que la mostrada da la posibilidad al usuario de cambiar la matriz de peso por tratarse de la variante ponderada, entrándole los números que se deseen probar en la matriz y utilizando el botón “Ejecutar” se actualizarían los resultados con los nuevos pesos.

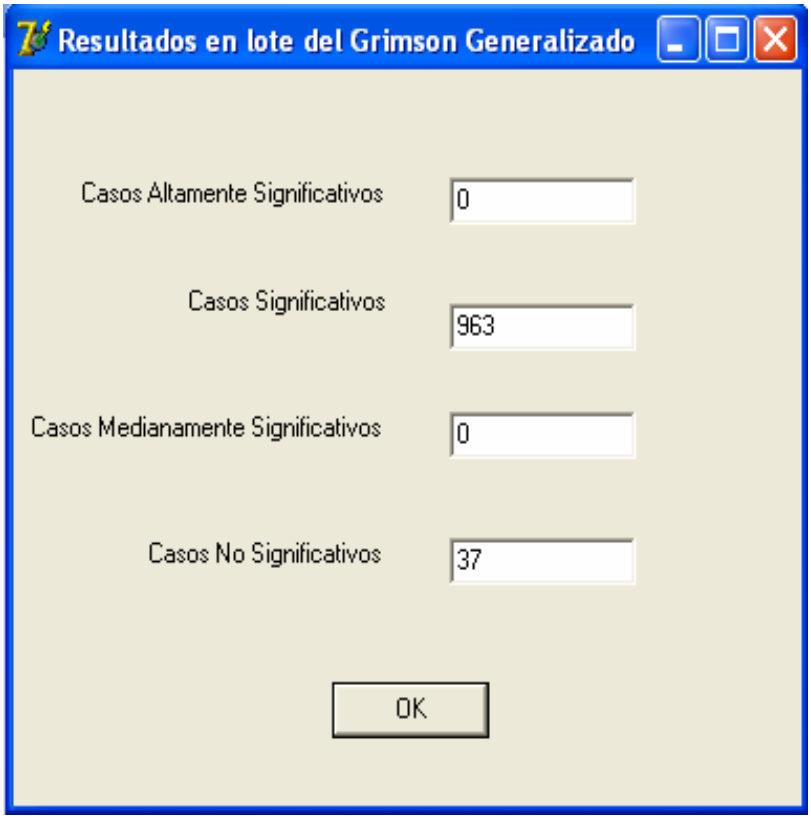

Figura 2.8 Resultados cuando el procesamiento es por lote.

Las formas correspondientes con los otros métodos son análogas a la anterior.

2.1.2 Diagrama de Casos de Uso

En el diagrama de casos de usos de la Figura 2.8 se hace un bosquejo de todos los casos de uso del sistema, que serán explicados detalladamente, más adelante.

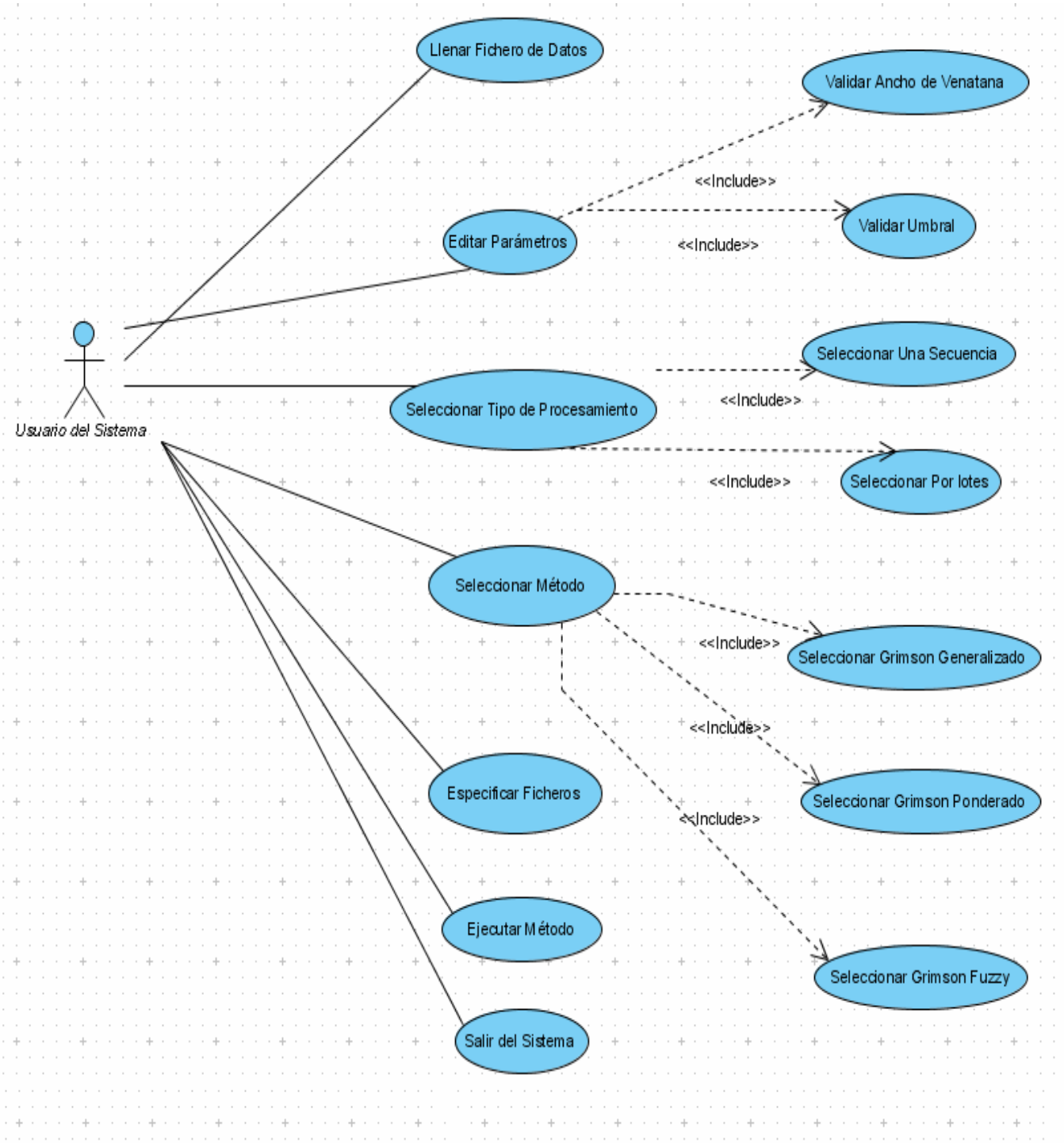

Figura 2.9 Diagrama de Casos de Uso del Usuario del Sistema

2.1.3 Especificaciones de los Casos de Uso

- **Primer Caso de Uso:** Llenar fichero de datos.

  Actor: Usuario del sistema.

  Propósito: Introducir los datos en el fichero destinado para esto.

Curso normal de los eventos

Acción del actor	Respuesta del sistema
1. Este caso de uso comienza cuando el usuario hace doble clic sobre el fichero de datos para abrirlo.	2. Muestra el fichero abierto
3. El usuario copia los datos para el fichero.	
4. El usuario guarda los datos utilizando la opción Guardar del fichero.	5. Guarda los datos copiados en el fichero texto.
6. El usuario hace clic en el botón cerrar del fichero.	7. Cierra el fichero.

Flujo alternativo

- Línea 3: El usuario decide no copiar los datos cancelando la acción.

- **Segundo Caso de Uso:** Editar Parámetros.

  Actor: Usuario del sistema.

  Propósito: Editar los valores de parámetros necesarios para la ejecución.

Sección Principal

Curso normal de los eventos

Acción del actor	Respuesta del sistema
1. Este caso de uso comienza cuando el usuario hace clic en alguno de los siguientes cuadros de textos: a) Ancho_de_Ventana b) Valor_del_Umbral	
2. El usuario escribe el valor en el cuadro de texto **Ancho_de_Ventana**.	3. El sistema valida el valor escrito y muestra en el cuadro de texto **Valor _ del_ Umbral**, el número correspondiente a la mitad del valor del ancho de la ventana.
4. El usuario puede reeditar los valores o editar otro campo.	

Flujo alternativo

- Línea 2: El usuario decide no editar los datos cancelando la acción y mostrándose un mensaje de error.

Sección Ancho de Ventana

Curso normal de los eventos

Acción del actor	Respuesta del sistema
1. El usuario escribe valor en el cuadro de texto **Ancho_de_Ventana**.	2. El sistema reconoce el valor como numérico.

Flujo alternativo

- Línea 2: El sistema detecta que el carácter escrito por el usuario no es numérico y se muestra mensaje de error.

- Línea 2: El sistema detecta que el valor escrito es igual a 1 y muestra mensaje de error.
- Línea 2: El usuario abandona el cuadro de texto dejándolo vacío y se muestra mensaje de error.

Sección Valor del Umbral

Curso normal de los eventos

Acción del actor	Respuesta del sistema
1. El usuario escribe valor en el cuadro de texto **Valor_del_umbral**.	2. El sistema reconoce el valor como numérico.

Flujo alternativo

- Línea 2: El sistema detecta que el carácter escrito por el usuario no es numérico y se muestra mensaje de error.
- Línea 2: El sistema detecta que el valor escrito es mayor que el ancho de la ventana y se muestra mensaje de error.
- Línea 2: El usuario abandona el cuadro de texto dejándolo vacío y se muestra mensaje de error.

- **Tercer Caso de Uso:** Seleccionar tipo de procesamiento.

Actor: Usuario del sistema.

Propósito: Seleccionar el tipo de procesamiento que se desea.

Curso normal de los eventos

Acción del actor	Respuesta del sistema
1. El usuario selecciona el tipo de procesamiento que desea en el Combobox1.	2. El sistema guarda en el campo texto del combobox1 el tipo de procesamiento seleccionado.

Flujo alternativo

- Línea 1: El usuario abandona el combobox1 sin seleccionar nada.

- **Cuarto Caso de Uso:** Seleccionar método que se desea ejecutar.

  Actor: Usuario del sistema.

  Propósito: Seleccionar el método que se desea ejecutar.

  Curso normal de los eventos

Acción del actor	Respuesta del sistema
1. El usuario selecciona el método que desea en el Combobox2.	2. El sistema guarda en el campo texto del combobox2 el método seleccionado.

Flujo alternativo

- Línea 1: El usuario abandona el combobox2 sin seleccionar nada.

- **Quinto Caso de Uso:** Especificar Ficheros.

  Actor: Usuario del sistema.

  Propósito: Especificar la dirección de fichero de datos y fichero de resultados.

Curso normal de los eventos

Acción del actor	Respuesta del sistema
1. Este caso de uso comienza cuando el usuario hace clic en el botón **Configuración** para buscar los siguientes ficheros : a) Fichero de datos inicialmente llenado b) Fichero donde guardar resultados	2. El sistema muesta la interfaz **Especificación de ficheros** donde será posible buscar todos los ficheros textos existentes.
3. El usuario selecciona los ficheros necesarios utilizando botones que se encuentran en la forma mostrada.	
4. El usuario cierra la interfaz **Especificación de ficheros** a través del botón **OK**.	6. El sistema cierra la interfaz.

Flujo alternativo

- Línea 3: No selecciona ningún fichero en específico y se muestra un mensaje de error.

Sección seleccionar fichero de datos

Curso normal de los eventos

Acción del actor	Respuesta del sistema
1. El usuario hace clic en el botón que le permite iniciar la búsqueda.	2. Se muestra el diálogo **Abrir** donde el usuario puede seleccionar el directorio que desea y ver todos los ficheros textos existentes en él.
3. El usuario especifica el fichero a utilizar.	4. El sistema muestra y guarda la dirección del fichero en el campo texto del Edit correspondiente en la interfaz

	Especificación de ficheros.

Flujo alternativo

- Línea 1: No busca ningún fichero en específico y se muestra un mensaje de error.

La sección correspondiente a la selección del fichero de resultados es análoga a la anterior.

- **Sexto Caso de Uso:** Ejecutar Método.

Actor: Usuario del sistema.

Propósito: Ejecutar método seleccionado, visualizar y guardar los datos.

Prerrequisitos: El sistema debe tener el camino completo del fichero de datos y todos los parámetros que se necesitan.

Curso normal de los eventos

Acción del actor	Respuesta del sistema
1. Este caso de uso comienza cuando el usuario hace clic en el botón **Ejecutar** de la interfaz del sistema.	2. El sistema carga los datos del fichero especificado anteriormente, ejecuta el método seleccionado, y muestra otra interfaz donde aparecen los resultados y los guarda en el fichero especificado. Si el método seleccionado fue Grimson Generalizado la interfaz mostrada será **Resultados de Grimson Generalizado**. Análogamente sucederá si se selecciona otro método.

Flujo alternativo

- Línea 1: El usuario decide no ejecutar nada y desiste de la acción.

- **Séptimo Caso de Uso:** Salir del sistema.

  Actor: Usuario del sistema.

  Propósito: Salir del sistema.

  Curso normal de los eventos

Acción del actor	Respuesta del sistema
1. Este caso de uso comienza cuando el usuario hace clic en el botón **Salir**.	2. Se cierra la interfaz del sistema.

2.2 Diagrama de Estructura de Clases

El sistema se elaboró en Borland Delphi 7.0. Para su desarrollo fue necesario crear tres clases que heredan de una donde están los procedimientos y funciones comunes, además de otra clase donde se encuentran todas las funciones matemáticas que se necesitan. También se crearon varias formas necesarias para la entrada y salida de datos.

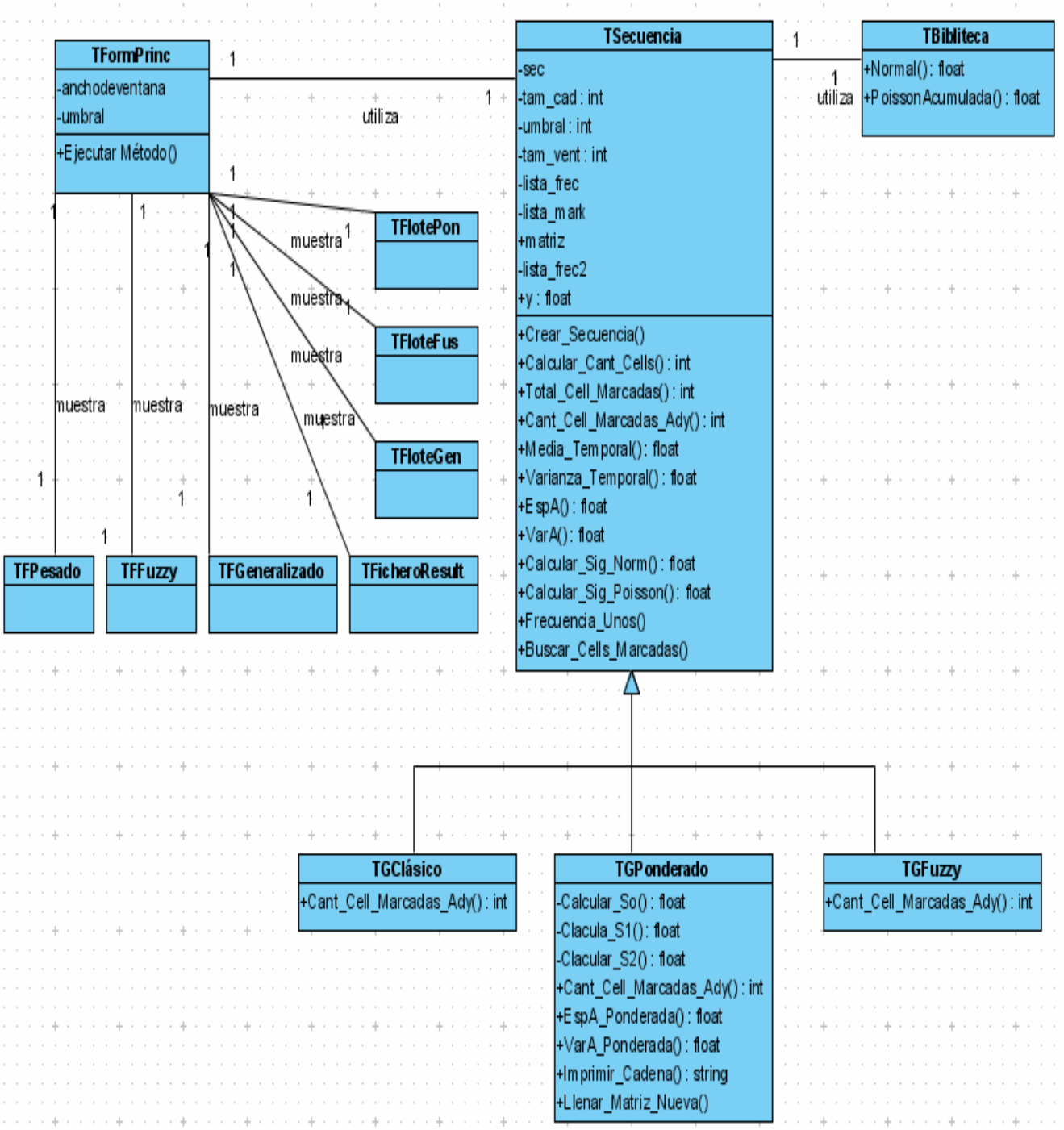

Figura 2.10 Diagrama de Estructura de Clases.

Consideraciones finales

En este capítulo aparecen detalles de la implementación del sistema así como la especificación de requisitos que hay que tener en cuenta para lograr una edición de los parámetros de forma segura y lograr una ejecución más eficiente, el diagrama de los casos de uso y el diagrama de clases utilizado. Además se muestra el manual de usuario.

Capítulo 3

3 Validación con datos simulados y Aplicaciones.

En este capítulo se muestra el uso del método Grimson y sus modificaciones con datos reales de una aplicación bioinformática.

3.1 Validación con datos simulados

Los avances de la Computación en las últimas décadas han contribuido notablemente al fortalecimiento de la matemática y esta a su vez ha sido utilizada como una poderosa herramienta de trabajo por otras ciencias entre las que se encuentra la medicina y la bioinformática. Varios modelos de epidemias han mostrado su utilidad en la prevención y en la aplicación de diversos procedimientos de control de enfermedades; sin embargo ninguno de ellos es infalible, todos tienen sus ventajas y limitaciones y ambas deben conocerse a fin de evitar aplicaciones incorrectas que pueden ser sumamente perjudiciales en un escenario epidemiológico, (Casas 2003).

Por razones elementales no es posible efectuar experimentos con enfermedades en humanos. Los datos que se tienen son los recopilados de las epidemias naturales que han ocurrido en años anteriores, pero muy frecuentemente estos son imprecisos o están incompletos. Es por ello que los modelos matemáticos y la simulación computarizada de datos juegan un papel fundamental en la realización de experimentos teóricos, en la estimación de sus parámetros y en la validación de pruebas estadísticas relacionadas con tales afecciones, (Ríos 2000).

Definitivamente los métodos de detección de conglomerados no pueden aplicarse en todas las situaciones en que se sospeche el origen de un foco epidémico, su campo de aplicación debe conocerse bien a fin de no utilizar técnicas incorrectas que conduzcan a falsas conclusiones. Es necesario entonces, desarrollar técnicas de simulación que permitan validar cualquier nuevo método antes de confrontarlo con datos reales, (Casas 2003).

Por su parte, en los problemas de Bioinformática generalmente el conocimiento es escaso. Resulta entonces muy importante, conocer el comportamiento de los métodos estadísticos de detección de conglomerados, para fin de evitar arribar a conclusiones falsas o poco certeras.

Es por ello que se decidió realizar un estudio experimental para determinar el comportamiento de la familia de los métodos Grimson, ante situaciones específicas.

3.1.1 Bases de la simulación realizada

La simulación se realizó con ayuda del paquete *Mathematica*® por razones de velocidad, facilidad de generación y precisión de los cálculos. Se generaron secuencias de datos de diversos tamaños para analizar su influencia en la capacidad de detección de las diferentes variantes.

Se generaron secuencias que contienen conglomerados (a ellas se les conglomerados verdaderos) y otras que no lo tienen (conglomerados falsos). Con todas ellas se ejecutaron los métodos implementados. Más adelante en este capítulo aparece un resumen de los resultados que se obtuvieron.

Para generar los conglomerados verdaderos se utilizó una distribución de Bernoulli (0.2) en el primer y último cuarto de la población, representando esto un 50% del total mientras que para el porcentaje restante se utilizó Bernoulli (0.8).

VERDADEROS CONGLOMERADOS (Sensibilidad)

0 1 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 0 0 1 0 0 1

Bernoulli (0.2) *Bernoulli (0.8)* *Bernoulli (0.2)*

Para generar los conglomerados falsos se utilizó una distribución de Bernoulli (0.2) en el total de los casos.

FALSOS CONGLOMERADOS (Especificidad)

0 1 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 1 0 0 0 0

Bernoulli (0.2)

3.1.2 Validación de las variantes del método Grimson

Se eligieron cuatro valores para el tamaño de las secuencias, que van desde las más pequeñas *(100, 300)* hasta otras un poco más grandes, *(500, 1000)* y en todas las simulaciones la cantidad de casos fue igual a 1000.

Las tablas 3.1 y 3.2 muestran los resultados de los diferentes juegos de datos, observe que en la variante borrosa mantiene aproximadamente los resultados iguales que la clásica para las secuencias con verdaderos conglomerados, pero mejora la clasificación con las secuencias de falsos conglomerados.

En este sentido, se utilizó la siguiente clasificación:

- Altamente significativo: $p < 0.01$
- Significativo: $0.01 \leq p < 0.05$
- Medianamente significativo: $0.05 \leq p < 0.1$
- No significativo: $p \geq 0.1$

Los datos se muestran con la nomenclatura siguiente: secuencia (*Sec*), tamaño de ventana (*TVe*), umbral (*Um*), altamente significativos (*AS*), significativos (*Sig*), medianamente significativos (*MS*) y no significativos (*No*).

Tamaño Sec	TVen	Um	Conglomerados								
			Verdaderos				Falsos				Exactitud
			AS	Sig	MS	No	AS	Sig	MS	No	
100	10	3	57	614	0	329	1	82	0	917	79.4
		5	0	963	0	37	25	22	0	953	95.80
		7	0	610	0	390	6	72	0	922	92.8
300	30	7	29	832	0	139	0	4	0	996	92.8
		15	0	982	0	18	23	19	0	958	97.00
		22	0	446	0	554	0	0	0	1000	72.3
500	50	12	4	983	0	13	0	1	0	999	99.3
		30	0	978	0	22	19	52	0	959	96.85
		37	0	437	0	563	0	0	0	1000	71.8
1000	100	25	0	1000	0	0	0	0	0	1000	100
		60	0	988	0	12	23	23	0	954	97.105
		75	0	355	0	645	0	0	0	1000	67.7

Tabla 3.1 Resultados del método Clásico.

Como puede apreciarse en la tabla 3.1, la columna *NO* correspondiente a los verdaderos conglomerados no tiene valores pequeños, obsérvese por ejemplo que en secuencias de tamaño 500, con un umbral de 37, este valor es de 563, es decir más del 50% de las secuencias están incorrectamente clasificadas, mientras que con los falsos los resultados fueron correctos en un 100%. Este es el problema fundamental del método Grimson: la detección excesiva de falsos positivos.

Tamaño Sec	*TVen*	*Um*	*Conglomerados*								
			Verdaderos				*Falsos*				*Exactitud*
			AS	*Sig*	*MS*	*No*	*AS*	*Sig*	*MS*	*No*	
100	*10*	*3*	*5*	*666*	*0*	*329*	*0*	*18*	*0*	*917*	*79.4*
		5	*0*	*963*	*0*	*37*	*0*	*20*	*7*	*973*	*97.15*
		7	*0*	*531*	*79*	*390*	*0*	*24*	*54*	*922*	*75.3*
300	*30*	*7*	*9*	*852*	*0*	*139*	*0*	*4*	*0*	*996*	*92.8*
		15	*0*	*982*	*0*	*18*	*3*	*23*	*8*	*966*	*97.80*
		22	*0*	*444*	*2*	*554*	*0*	*0*	*0*	*1000*	*72.2*
500	*50*	*12*	*0*	*987*	*0*	*13*	*0*	*1*	*0*	*999*	*99.3*
		30	*0*	*978*	*0*	*22*	*3*	*29*	*3*	*965*	*97.30*
		37	*0*	*437*	*0*	*563*	*0*	*0*	*0*	*1000*	*71.8*
1000	*100*	*25*	*0*	*1000*	*0*	*0*	*0*	*0*	*0*	*1000*	*100*
		60	*0*	*988*	*0*	*12*	*5*	*37*	*1*	*957*	*97.30*
		75	*0*	*355*	*0*	*645*	*0*	*0*	*0*	*1000*	*67.7*

Tabla 3.2 Resultados del método Borroso.

La tabla 3.2 muestra los resultados de realizar el mismo experimento (con las mismas secuencias), pero aplicando el método de Grimson Borroso.

3.1.3 Resultados del diseño de Experimentos

Dada un tamaño de ventana y un umbral, el método Grimson clasifica si en una secuencia existe al menos un conglomerado de la categoría de interés, por lo que nos interesa medir la influencia que producen dichos parámetros en su desempeño. Por tal razón la información analizada es la Exactitud (Accuracy) obtenida utilizando el conjunto de verdaderos y falsos conglomerados de cada caso donde hay secuencias de diferentes tamaños. Con el objetivo de generalizar los resultados en las distintas secuencias el tamaño de la ventana se trabaja en porciento.

Se conoce que en el Método Grimson por su naturaleza el tamaño de la ventana no puede ser mayor al tamaño total de la secuencia dividido 4. Además el desempeño de este método es nulo para ventanas mayores al 15 % del tamaño de la secuencia. Para el análisis de los resultados se fija el tamaño de la ventana alrededor de los

niveles: 5%, 9% y 13% y el umbral alrededor del 25%, 50% y 75% con respecto al tamaño de la ventana. Se controla el parámetro suavizado (fuzzy) repitiendo los experimentos para Grimson Generalizado y Grimson Borroso. Cada uno de estos experimentos tiene tres réplicas, cada una con probabilidades diferentes de presencia de la categoría de interés en el conglomerado (probabilidad de 0.9, 0.7 y 0.5)

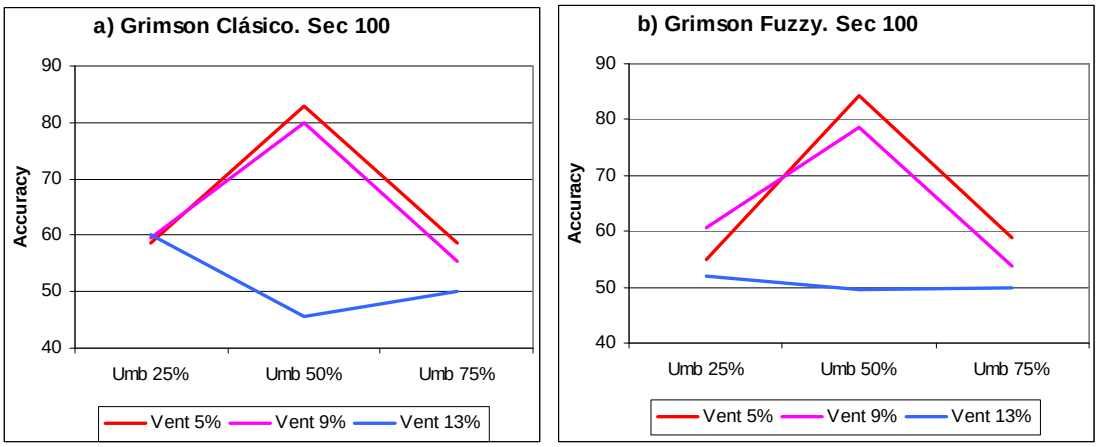

Tabla 3.3: Gráfico del factor ventana contra el factor umbral el método Grimson.

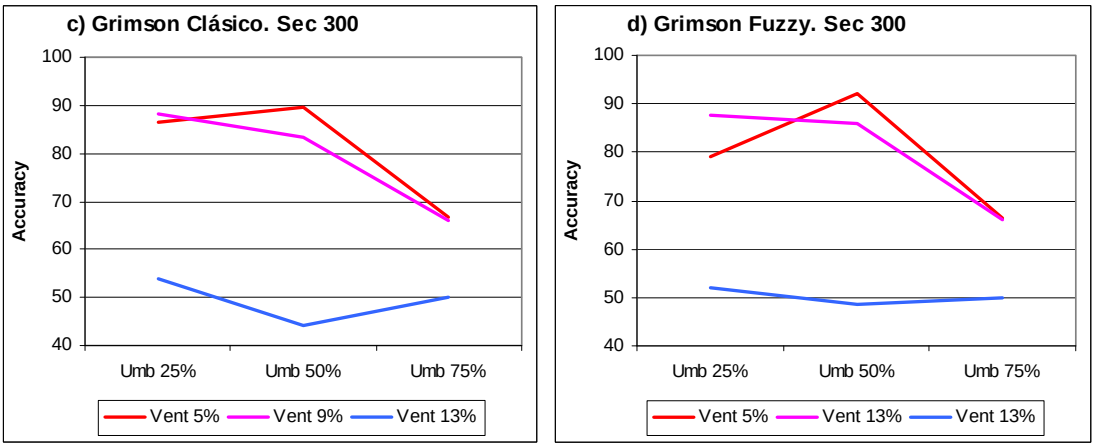

Tabla 3.4: Gráfico del factor ventana contra el factor umbral el método Grimson.

En las figuras 3.3 y 3.4 se ilustra que el método Grimson Clásico y su variante borrosa, tienen el siguiente comportamiento.

- El factor ventana para valores pequeños (5%) aumenta su respuesta para umbrales alrededor del 25 % con respecto al tamaño de la ventana.
- El factor ventana para valores intermedio (9%) disminuye su respuesta al aumentar los valores del umbral, cuando aumenta el tamaño de la secuencia, para secuencias pequeñas se comporta de forma similar a la anterior.

- El factor ventana para valores alrededor del 13% disminuye su respuesta para valores intermedios del umbral.

Ventanas alrededor del 5, 9 y 13%						
Umbrales alrededor del 25, 50 y 75% con respecto al tamaño de la ventana						
Tamaño Secuencia	Grimson Clásico			Grimson Fuzzy		
	Tamaño Ventana	Umbral	VxP	Ventana	Umbral	VxP
100	0.042	0.237	0.070	0.003	0.271	0.206
300	0.002	0.368	0.512	0.002	0.380	0.454

Tabla 3.5: Significación del análisis bifactorial no paramétrico.

En la tabla 3.5 se presenta la significación de los factores ventana, umbral y la interacción de ellos en cada uno de los experimentos del Grimson Generalizado y Borroso con las diferentes secuencias, se concluye que el total de casos bien clasificados es afectado en todos sus experimentos significativamente por el factor tamaño de ventana con una confiabilidad de 95%. Mientras que el factor umbral y la interacción entre ambos no afecta significativamente el desempeño de ninguno de los métodos en los experimentos realizados.

3.2 Una aplicación bioinformática

Un alineamiento de secuencias en bioinformática es una forma de representar y comparar dos o más secuencias o cadenas de ADN, ARN, o estructuras primarias proteicas para resaltar sus zonas de similitud, que podrían indicar relaciones funcionales o evolutivas entre los genes o proteínas consultados. Las secuencias alineadas se escriben con las letras (representando aminoácidos o nucleótidos) en filas de una matriz en las que, si es necesario, se insertan espacios para que las zonas con idéntica o similar estructura se alineen. (es.wikipedia.org/wiki-enciclopedia 2009).

El alineamiento múltiple de secuencias es una extensión del alineamiento de pares que incorpora más de dos secuencias al mismo tiempo. Los métodos de alineamiento múltiple intentan alinear todas las secuencias de un conjunto dado. Los alineamientos múltiples son usados a menudos en la identificación de regiones conservadas en un grupo de secuencias que hipotéticamente están relacionadas evolutivamente. Los

alineamientos son también utilizados para ayudar al establecimiento de relaciones evolutivas mediante la construcción de árboles filogenéticos. Los alineamientos múltiples de secuencias son computacionalmente difíciles de producir y la mayoría de las formulaciones del problema conducen a problemas de optimización combinatorial NP-completos. Sin embargo, la utilidad de estos alineamientos en la bioinformática ha dado lugar al desarrollo de una variedad de métodos adecuados para la alineación de tres o más secuencias.

Secuencias muy cortas o muy similares pueden alinearse manualmente. Aun así, los problemas más interesantes necesitan alinear secuencias largas, muy variables y extremadamente numerosas que no pueden ser alineadas por humanos. Existen diferentes *softwares* en internet que realizan el alineamiento de secuencias, como el Mega4 y el ClustalW.

Producto de la alineación de varias secuencias de ADN aparecen ciertos desplazamientos. A estos “espacio vacíos” se les denomina “gaps”. La distribución de los “gaps”, dentro de las secuencias alineadas, no sigue la misma distribución de las bases nucleotídicas.

Como resultados de investigaciones del grupo de Bioinformática de la Universidad Central de Las Villas, se obtuvo un nuevo modelo evolutivo basado en cinco bases: las cuatro del ADN y el “gap”.

Como ya se había mencionado, el comportamiento de los “gaps” dentro de la secuencia no es el mismo que el de las bases nucleotídicas. Debido a ello se hizo necesario probar la existencia de conglomerados de gaps dentro de secuencias alineadas.

Se alinearon 167 genomas completos de los virus de la influenza A H1N1, obtenidos de internet. Las bases nucleotídicas se sustituyeron por 0, mientras que los “gaps” se sustituyeron por 1. Con esas secuencias binarias se ejecutó el método Grimson Generalizado y Grimson Fussy, en ambos se obtuvieron los resultados siguientes:

# sec	Tamaño por sec.	TVen	Um	AS	Sig	MS	No
167	14158	10	5	167	0	0	0
167	14158	30	15	167	0	0	0
167	14158	50	25	167	0	0	0
167	14158	100	50	18	0	0	149
167	14158	100	40	18	0	38	111
167	14158	100	30	49	117	0	1

Tabla 3.6 Resultados de los genomas del virus de la influenza A H1N1.

Como puede apreciarse, los resultados fueron altamente significativos en la mayoría de los casos. Para tamaños de ventana pequeños y los valores de umbral recomendados (la mitad del tamaño de la ventana), los resultados fueron altamente significativos. Esto demuestra que efectivamente existen conglomerados de “gaps” en las secuencias alineadas, lo que desde el punto de vista bioinformática era lo que se quería demostrar.

Para complementar el estudio y a manera de ejemplo, se muestran los resultados de uno de los genomas utilizados en el procesamiento de datos de la tabla anterior, el método Grimson Borroso arrojó los mismos resultados.

METODO GRIMSON GENERALIZADO ANCHOVENTANA =30 UMBRAL=15	
Parámetros	**Valores**
Total de Celdas	471
Total de Celdas Marcadas	20
Celdas Adyacentes	11

Esperanza A	0.8067
Varianza A	0.7429
Significación Normal	9.999E-8

Tabla 3.7 Procesamiento de una secuencia

Consideraciones finales

La simulación de datos aquí mostrada tiene una gran importancia, pues permite determinar las limitaciones de las técnicas de detección de conglomerados y evita así aplicaciones incorrectas que pueden conducir a detectar falsas aglomeraciones o peor aún, a no detectar las verdaderas.

En este capítulo se ha realizado una simulación del método Grimson Generalizado y Grimson Borroso trabajando en su variante temporal solamente, considerando diversos valores del tamaño de la población y del umbral respectivamente.

También se muestra la facilidad del software de trabajar por lotes.

En todas las ocasiones los resultados correspondientes a los casos bien clasificados han alcanzado valores superiores al 95.0%, los que se catalogan como muy buenos. Esto muestra su importancia y confiabilidad.

Conclusiones

Este trabajo recoge de manera ordenada el surgimiento, evolución y desarrollo hasta la fecha de los métodos Grimson, enunciando sus ventajas y limitaciones. Con él se arriban a las siguientes conclusiones:

- Se diseñó e implementó un sistema computacional que tiene todas las variantes de los métodos Grimson: la clásica, la pesada y la borrosa. El sistema ofrece un ambiente cómodo que puede ser utilizado con facilidad por personal no especialista en computación. Además:
  - Brinda facilidades para procesar una sola secuencia, mostrando con detalles todos los valores estadísticos asociados a los métodos.
  - Brinda facilidades para realizar procesamientos en lotes, es decir, procesar muchas secuencias a la vez. Esta última opción sólo da resultados de manera agrupada.
- Los métodos propuestos se validaron con datos simulados. Los resultados obtenidos fueron satisfactorios.
- En el diseño de experimentos se concluye que el total de casos bien clasificados es afectado en todos sus experimentos significativamente por el factor tamaño de ventana con una confiabilidad de 95%. Mientras que el factor umbral y la interacción entre ambos no afecta significativamente el desempeño de ninguno de los métodos en los experimentos realizados.
- El método Grimson implementado se utilizó para resolver un problema de bioinformática: la detección de conglomerados de “gaps” en secuencias alineadas del virus de la influenza A H1N1. Los resultados obtenidos demuestran la existencia de conglomerados.

Recomendaciones

- Realizar un estudio experimental riguroso para determinar con mayor precisión los valores de los parámetros del método Grimson, los cuales influyen en la calidad de sus resultados, y para determinar los parámetros adecuados en dependencia del dominio de aplicación del problema a resolver.

Referencias Bibliográficas

Anthropol, A. J. P. (1952). "Ritchie WA. Paleopathological evidence suggesting Precolumbian tuberculosis in New York State." **10**: 305-18.

Buckley, J. J. (2006). " Fuzzy Probability and Statistics."

Casas, G. (2003). "Técnicas de detección de conglomerados incluyendo factores adicionales". FACULTAD DE MATEMÁTICA, FÍSICA Y COMPUTACIÓN-DEPARTAMENTO DE COMPUTACIÓN. Villa Clara, UNIVERSIDAD CENTRAL "MARTA ABREU" DE LAS VILLAS**:** 113.

Casas, G. C. R., Vivian Guerra Morales y Luis Felipe Herrera Jiménez (2002). "Técnicas de detección de clustes aplicadas a la investigación psicológica." Revista Cubana de Psicología **Vol 19 No1**

Casas, R. G. A., Mario Pupo Meriño y Laureano Rodríguez Corvea (2008). "generalización de dos métodos de detección de conglomerados. aplicaciones en bioinformática." Revista de Matemática: Teoría y Aplicaciones 2008 **15(1) 27–40**.

Correal G, F. I. (1992). "Estudio de las momias guanes de La Mesa de los Santos (Santander, Colombia)." Rev Acad Colomb Cienc **18**: 183-90.

Cuzick, J. a. E. (1990). ""Spatial clustering for inhomogeneous populations" (with discussion)." Journal of the Royal Statistics Society, Series B **52**: 73-104.

es.wikipedia.org/wiki-enciclopedia. (2009). " Alineamiento de secuencias."

Grimson, R. a. R. R. (1991). ""A versatile test for clustering and a Proximity Analysis of Neurons"." Methods of Information in Medicine **30**: , 299-303.

Guzmán MG, K. G., Bravo J, Soler M, Vázquez S, Morier L (1990). "Dengue hemorrhagic fever in Cuba , 1981: a retrospective seroepidemiologic study." Am J Trop Med Hyg **42**: 179-184.

Idrovo, A. J. (2004). "Raíces históricas, sociales y epidemiológicas de la tuberculosis en Bogotá, Colombia." Biomédica-Instituto Nacional de Salud Pública, Cuernavaca, México. **24-no.4**

Jacquez, G. (1994). ""Cuzick and Edwards test when exact locations are unknown"." American Journal of Epidemiology **140 (1)**: 58-64.

Jacquez, G. L. W. R. G. a. D. W. ((1996 a). ""The analysis of Disease Clusters, Part I: Stat of the Art"." Infection Control and Hospital Epidemiology **17 (5)**: 319-27.

Jenkins, B. a. (1994). "Time Series Analysis, Forecasting and Control, Prentice Hall.", Englewood Cliffs, 1994 **1994**.

JV, R. (1988). "Acerca de la supuesta debilidad mental y física de los muiscas como posible causa de conquista y posterior extinción" Arqueología **1**: 42-6.

Knox, E. (1964 b). ""Epidemiology of childhood leukemia in Northemberland and Durham"." British Journal of Preventative and Social Medicine **18, 17 24.**

Kouri GP, G. M., Bravo JR, Triana C (1989). "Dengue hemorrhagic fever/dengue shock syndrome: lessons from the Cuban epidemic, 1981." Bull World Health Organ **67:**: 375-380.

Larsen, R. y. c. (1973). ""A statistical test for measuring unimodal clustering: a description of the test and of its application of cases of acute leukemia in metropolitan Atlanta, Georgia", ." Biometrics **29**: 301-9.

Mantel, N. (1967). ""The detection of disease clustering and a generalized regression approach"." Cancer Research **27(2)** 209-20.

Marshall, R. (1991). ""A review of methods for the statistical analysis of spatial patterns of disease"." Journal of the Royal Statistical Society Association **154 (3)** 421-441.

Montgomery (2008). " Diseño y Análisis de Experimentos." D.C, , México: Limusa.

MY, E.-N. (1979). "Human treponematosis and tuberculosis: evidence from the New World. ." Am J Phys Anthropol **51**: 599-618.

Nagarwilla, N. (1996). ""A Scan statistic with a variable window"." Statistics in Medicine **15**: 845-50.

Nauss, J. (1982). ""Approximations for distributions of Scan statistics"." Journal of the American Statistics Association **77**: 377.

Pavel ,Silveira D.(2008).Análisis comparativo de secuencias genómicas mediante la ap licación de la Transformada de Fourier sobre campos finitos.Matemática-Física Computación, Universidad Central de Las Villas "Marta Abreu"

Perzigian A, W. L. (1979). "Evidence for tuberculosis in a prehistoric population." JAMA **241**: 2643-6.

Ríos , D., Ríos Insua, S. y Martín Jiménez, J. (2000). "Simulación: métodos y aplicaciones. ." ALFAOMEGA, S.A. **ISBN: 970-15-0509-3, 2000**.

Rodríguez, L., G. Casas, and R. Grau. . Cartagena de Indias. Colombia. (2008). Linear Fuzzy Scan Method to Detect Clusters. A Bioinformatic Application. in XIV Latin Ibero-American Congress on Operations Research (CLAIO 2008). .

Roentgenol, A. J. (1925). "Means H.Roentgenological study of the skeletal remains of the prehistoric Mound Builder Indians of Ohio." **13**: 359-67.

S, P. (1984). "Pfeiffer S. Paleopathology in an Iroquoian Ossuary, with special reference to tuberculosis." **65**: 181-9.

Sokal, R. R. and F. J. Rohlf (1995). The principles and practice of statistics in biological research. New York, W. H. Freeman and Company.

Sotomayor H, B. J., Arango M (2004). "Demostración de tuberculosis en una momia prehispánica colombiana por la ribotipificación del ADN de Mycobacterium tuberculosis." Biomédica **24(Supl.)**.

Surg, J. B. J. (1957). "Lichtor J, Lichtor A. Paleopathological evidence suggesting Precolumbian tuberculosis of the spine." **39**: 1398-9.

Trab, M. S. (2007). "Las enfermedades transmisibles amenazan a los trabajadores." Med Segur Trab 2007 **Vol LIII Nº 209**: 67-68.

WA, R. (1952). "Ritchie WA. Paleopathological evidence suggesting Precolumbian tuberculosis in New York State." **10**: 305-18.

Anexo

-Método Grimson-

```
Var
  f: textfile;
  Umbral, anchoventana: integer;
Begin
   If (anchoventana = 1) then write ('Mensaje de Error'); Salir;
   If (anchoventana = '') then write ('Mensaje de Error'); Salir;
   If (umbral > anchoventana)  then write ('Mensaje de Error'); Salir;

  Crear _ secuencia (f, umbral, anchoventana);   //Constructor donde se  inicializan los
                                                  parámetros pasados por
                                                  datos y además se lee el
                                                  fichero de datos.
Tamaño _ cadena=length (secuencia);
Cant_Cells = Tamaño _ cadena / anchoventana;
Calcular_Frecuencia_Unos ();                    //Cuenta la cantidad de unos de
                                                  cada celda.
Buscar_Cells_Marcadas ();                         //Celdas marcadas serán
                                                  aquellas donde la
                                                  frecuencia de unos es
                                                  mayor o igual al umbral.
Total_Cells_Marcadas ();
Cant_Cells_Marcadas_Ady ();                     //Contiene la cantidad de celdas
                                                  marcadas que son
                                                  adyacentes.
Media _ temporal = 2*(Cant_Cells-1) / Cant_Cells;
Varianza_temporal = 2*(Cant_Cells-2) / sqr (Cant_Cells);

Esperanza_A = Media _ temporal *Total_Cells_Marcadas*
                (Total_Cells_Marcadas-1) / 2*(Cant_Cells-1);
Temp0 = 2*(Media _ temporal-1) (Total_Cells_marcadas-2)/ (Cant_Cells-2);
Temp1 = (Cant_Cells*Media _ temporal-4*Media _ temporal+2)*
          (Total_Cells_Marcadas-2) (Total_Cells_Marcadas-3);
```

```
Temp2 = 2*(Cant_Cells-2) (Cant_Cells-3);
Temp3 = Total_Cells_Marcadas-*(Total_Cells_Marcadas-1)
            (Total_Cells_Marcadas-2);
Temp4 = (Cant_Cells-1) (Cant_Cells-2);
Temp5 = Total_Cells_Marcadas*(Total_Cells_Marcadas-1)*
            (Total_Cells_Marcadas-2)(Total_Cells_Marcadas-3);
Temp6 = (Cant_Cells-1)*(Cant_Cells-2) (Cant_Cells-3);
Varianza_A = Esperanza_A*(1+Temp0+Temp1/Temp2)+
                Varianza _ temporal*(Temp3/Temp4-Temp5/Temp6);
Razón = Total_Cells_Marcadas / Cant_Cells;
Nnobs = (Cant_Cells_Marcadas_Ady-Esperanza_A) /sqrt (Varianza_A);
Significación _ normal = 1-Normal (Nnobs);
  If  Significación _ normal < 0.05 then
        Inc (SIG);
  Else If Significación_normal <=0.1 then
        Inc (MED_SIG);
  Else
        Inc (NO);
Write ('SIG = ', SIG,'NO = ', NO,'MED_SIG' =, MED_SIG);
```