Árboles de Clasificación y Regresión basados en atributos funcionales y su utilización en el contexto de procesos epidémicos

23 de septiembre de 2009

Tesis de Maestría en Ingeniería Matemática
Facultad de Ingeniería, UDELAR

Juan Piccini

Tutores : Dr. Gonzalo Perera
Dr. Ing. Franco Robledo

Tribunal : Dr. Juan Cristina
Dr. Ing. Pablo Musé
MSc. Ing. Omar Viera

Agradecimientos

Son muchas las personas a quienes debo agradecer, y los intentos de nombrarlas a todas me hicieron tomar conciencia de lo imposible de la tarea. Es por ello que solamente nombraré algunos especialmente involucrados, que espero representen a todos.

Vaya pues mi gratitud hacia Eduardo Canale, Badih Ghattas, Gonzalo Perera, Franco Robledo, Ariel Roche y Marco Scavino.

"Que otros se jacten de
los libros que han escrito,
yo me enorgullezco
de los que he leído"
J.L. Borges

Índice

1. Introducción

1.1. Líneas generales

En este trabajo se construyen grafos aleatorios de 2000 nodos que en algún sentido imitan las redes de contactos entre individuos. Luego se simulan epidemias en dichos grafos siguiendo el clásico modelo SIR (Susceptible-Infectado-Removido).

La idea central de esta tesis es que cada simulación puede verse como la realización de un proceso estocástico al cual asociaremos determinadas funciones como descriptores. Aplicando herramientas de Machine Learning a dichos descriptores podremos clasificar las epidemias según las funciones utilizadas para describirlas (aquellas realizaciones o epidemias con descriptores similares serán clasificadas como del mismo tipo), permitiendo hacer una suerte de agrupamiento o *clustering* de las mismas.

Esto permite no solamente comparar procesos estocásticos (epidemias) entre sí, sino que también brinda la posibilidad de adelantar como sería el comportamiento de una epidemia "media"o "típica"sobre una red determinada, fijados ciertos parámetros.

Organización de este trabajo

- En este capítulo (secciones 1.2 y 1.3) se muestra la vinculación entre epidemias y redes y se introduce el modelo epidémico SIR en su formulación clásica.

- El capítulo 2 muestra una síntesis de las herramientas utilizadas para modelar epidemias sobre redes así como los distintos tipos de redes que emergen como adecuadas para modelar redes sociales. Se muestran definiciones de medidas para características globales y locales de redes.

- El capítulo 3 muestra una sítesis de los elementos de la teoría del **Machine Learning** (Statistical Learning, Aprendizaje Automático).

- Los capítulos 4 y 5 introducen las técnicas y algoritmos utilizados para la generación de grafos aleatorios así como para la simulación epidemias sobre los mismos.

- En el capítulo 6 simulamos epidemias con distintos grados de virulencia en tres grafos aleatorios distintos para luego usar árboles de clasificación y regresión con los cuales agrupar las distintas epidemias según sus

curvas de cantidad de casos en función del tiempo y de distribución de frecuencias de los grados de los nodos afectados. Asimismo probaremos distintas estrategias de inmunización o vacunación.

- El capítulo 7 contiene las conclusiones así como algunas líneas de trabajo a seguir.

1.2. Epidemias y Redes

Las redes ocupan hoy día un lugar predominante. Vivimos en un mundo de redes, y cualquier sistema complejo puede ser modelado o pensado como una red, donde los elementos del sistema representan vértices o nodos y las aristas representan las interacciones entre dichos elementos. Ejemplos provenientes de la biología, ciencias sociales, ingeniería, etc. atestiguan la versatilidad y potencia del concepto de red o grafo como sustrato matemático de los mismos.

De particular relevancia son las redes sociales, donde cada individuo se representa mediante un nodo y sus relaciones con otros individuos mediante aristas que conectan dicho nodo con los nodos que representan a los otros individuos.

Muchas de las enfermedades que se propagan en poblaciones, presuponen el contacto o cercanía entre individuos infectados (portadores del agente que provoca la enfermedad) e individuos susceptibles (quienes no han tenido todavía dicha enfermedad, pero podrían adquirirla). El patrón de estos contactos causantes de contagio puede modelarse mediante una red.

La mayoría de los modelos matemáticos clásicos suponen que la población está homogéneamente mezclada ("fully mixed"). Esto es, cualquier individuo infectado es igualmente capaz de contagiar a cualquier otro individuo de la misma población (asumiendo que el individuo contactado sea susceptible).

En poblaciones de gran tamaño (los modelos suponen poblaciones cuyo tamaño tiende a infinito), esto permite escribir sistemas de ecuaciones diferenciales no lineales para describir por ejemplo la evolución temporal del número de infectados, número de susceptibles, etc. Usualmente existe un umbral en el espacio de parámetros del modelo, que separa el estado donde la enfermedad no prospera (no logra infectar a una fracción apreciable de la población) del estado donde se declara una epidemia.

Cuando hablemos de **epidemia** nos estaremos refiriendo a brotes epidémicos que afectan a una fracción no nula de la población cuando el tamaño de ésta tiende a infinito (esto es, los individuos infectados son una parte consi-

derable de la población aún en poblaciones muy grandes). Usaremos indistintamente el término **brotes** o **brotes epidémicos** para aquellos que no llegan a afectar a una fracción apreciable de la población (o que afectan una fracción que tiende a cero cuando el tamaño de la población tiende a infinito).

Típicamente el parámetro R_0, llamado tasa reproductiva (la cantidad de contagios que un infectado medio puede lograr) es quien hace de umbral. Cuando $R_0 < 1$, no hay epidemia. Aunque con estos modelos que presuponen un mixing total se pueden hacer muchas cosas, no toman en cuenta la topología de la red: un individuo no tiene igual probabilidad de contactar al resto de la población, contacta con mayor frecuencia a unos que a otros.

Usualmente cada individuo tiene contactos solamente con una muy pequeña fracción de la población, y la cantidad de contactos "cotidianos "varía de individuo a individuo. Desde el advenimiento de Internet y los virus informáticos, estos modelos mostraron discrepancias entre sus predicciones y la realidad ([21]). Aparecen epidemias informáticas con valores de R_0 muy inferiores al valor crítico de uno, lo que hizo se prestara atención a la topología de la red subyacente.

En años recientes se ha acumulado cada vez más evidencia sobre la importancia de la topología de la red sobre la cual está discurriendo el proceso epidémico en cuestión. En particular se han estudiado los dos modelos más célebres: el modelo SIS (Susceptible-Infectado-Susceptible) y el modelo SIR (Susceptible-Infectado-Removido) ([22]).

1.3. El modelo SIR clásico

Probablemente la clase de modelos epidémicos más estudiada es la clase de modelos SIR (Susceptible-Infectados-Removidos). Bosquejaremos la versión original, formulada por Lowell Reed y Wade H. Frost en la década de 1920. En ésta, los individuos solo pueden existir en uno de tres estados: Susceptible (S), Infectado (I) y Removido (R). Esta última categoría representa a aquellos individuos que adquieren inmunidad luego de recuperarse de la enfermedad.

En este modelo los individuos infectados o infecciosos (dado que inmediantamente pueden contagiar a los susceptibles) tienen contactos al azar con otros individuos (de cualquiera de los tres estados) a una tasa media por unidad de tiempo β, y se recuperan y adquieren inmunidad a una tasa media por unidad de tiempo γ. Si el individuo contactado por un infectado es un susceptible, resulta contagiado y pasa a ser infectado. Describiremos el modelo a grandes rasgos, siguiendo [13].

Para una población suficientemente grande, el modelo puede describirse mediante el sistema de ecuaciones diferenciales no lineales:

$$\begin{aligned}\frac{ds}{dt} &= -\beta is\\ \frac{di}{dt} &= \beta is - \gamma i\\ \frac{dr}{dt} &= \gamma i\end{aligned}$$

donde $s(t) = S(t)/n, i(t) = I(t)/n, r(t) = R(t)/n$ son las proporciones de la población que se encuentra en cada estado en el instante t. La última ecuación puede omitirse por cuanto $s + i + r = 1$ ([13]).

El modelo supone la incidencia estándar y recuperación (salida de I) a razón de $\gamma\dfrac{I}{n}$ individuos por unidad de tiempo. Esto nos dará un tiempo de espera (o tiempo de permanencia en la clase I) de $e^{-\gamma t}$, con media $\dfrac{1}{\gamma}$. Puesto que dicho período de tiempo es corto, el modelo no tiene dinámica vital (nacimientos y muertes naturales), por lo que sólo resulta adecuado para describir enfermedades de rápido desarrollo y conclusión, que además confieren inmunidad a quien la ha padecido, como la influenza.

Asimismo, se supone que la población está totalmente mezclada (los individuos con los que un susceptible tiene contacto son elegidos al azar entre la población). También se supone que todos los individuos tienen aproximadamente el mismo número de contactos en el mismo tiempo, y que estos contactos transmiten la enfermedad con la misma probabilidad. Al quitar la última ecuación, obtenemos

$$\begin{aligned}\frac{ds}{dt} &= -\beta is \qquad s(0) = s_0 \geq 0\\ \frac{di}{dt} &= \beta is - \gamma i \quad i(0) = i_0 \geq 0\end{aligned} \tag{1.1}$$

Podemos adelantar algunas conclusiones de carácter cualitativo sobre $s(t)$ e $i(t)$ a partir del estudio del plano de fases ([13]) :

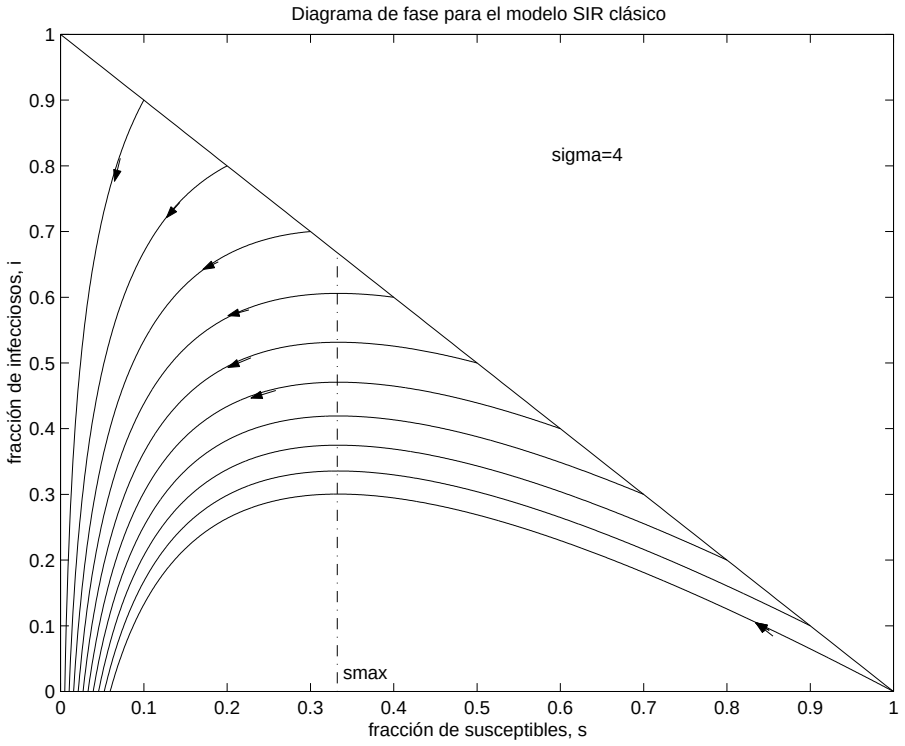
1. como $s(t)+i(t) \leq 1\ \forall t$, y $s(0)+i(0) = 1$, las curvas $(s(t), i(t))$ partirán en $t = 0$ de la recta $i = 1 - s$, y para tiempos posteriores estarán contenidas en el triángulo $T = \{(s, i) :\ s \geq 0;\ i \geq 0;\ s + i \leq 1\}$

2. $s'(t) < 0$, de modo que las abcisas se mueven hacia la izquierda.

3. $i' = 0$ sii $\beta si = \gamma i$, esto es, si $s = \dfrac{\gamma}{\beta}$. El signo de i' es negativo a la izquierda de $s = \dfrac{\gamma}{\beta}$ y positivo a la derecha, y como partimos desde la derecha (la recta $s + i = 1$), primero i crece a medida que s decrece, hasta que alcanzamos el valor $s = \dfrac{\gamma}{\beta}$. Luego, a medida que s sigue decreciendo, i también decrece, por lo que en el punto $s_{max} = \dfrac{\gamma}{\beta}$ tendremos un máximo para i.

La Figura 1 ilustra lo anterior.

Figura 1:



Notemos que $\dfrac{\beta}{\gamma}$ es la tasa de contacto β multiplicada por la duración media $\dfrac{1}{\gamma}$ del período de infección, de modo que podemos interpretarlo como

el número medio de contactos adecuados de un infeccioso típico durante su período infectivo.

Llamemos $\sigma = \dfrac{\beta}{\gamma}$, por lo que $s_{max} = \dfrac{1}{\sigma}$. Entonces, si llamamos $R(t)$ a la tasa media a la que un infectado típico contacta a los susceptibles (esto es, el número medio de contagios que logrará); al comienzo de la epidemia $R = \sigma s(0) = R_0$ que es la tasa media de contactos o contagios del primer infectado (el caso cero).

Si dividimos la segunda ecuación de (1.1) entre la primera, obtendremos

$$di/ds = -1 + 1/(\sigma s)$$

resolviendo dicha ecuación tenemos

$$i(s) + s - (log(s))/(\sigma) = i(0) + s(0) - (log(s(0)))/(\sigma)$$

$di/ds = 0$ cuando $s = \dfrac{1}{\sigma}$, y el signo de di/ds es negativo para $s < \dfrac{1}{\sigma}$ y positivo para $s > \dfrac{1}{\sigma}$. Esto muestra que i crece si $s > \dfrac{1}{\sigma}$, y decrece si $s < \dfrac{1}{\sigma}$.

Por tanto, el comportamiento de las soluciones $(s(t), i(t))$ de (1.1) puede resumirse como sigue:

1. Si $\sigma s(0) \leq 1$ entonces $i(t)$ decrece a cero cuando $t \longrightarrow +\infty$.

2. Si $\sigma s(0) > 1$, entonces $i(t)$ primero crece hasta alcanzar su máximo, $i_{max} = i(0) + s(0) - \dfrac{1}{\sigma} - \dfrac{log(\sigma s(0))}{\sigma}$, y luego decrece a cero cuando $t \longrightarrow +\infty$

3. La fracción final de susceptibles, s_∞ será la raíz de la ecuación

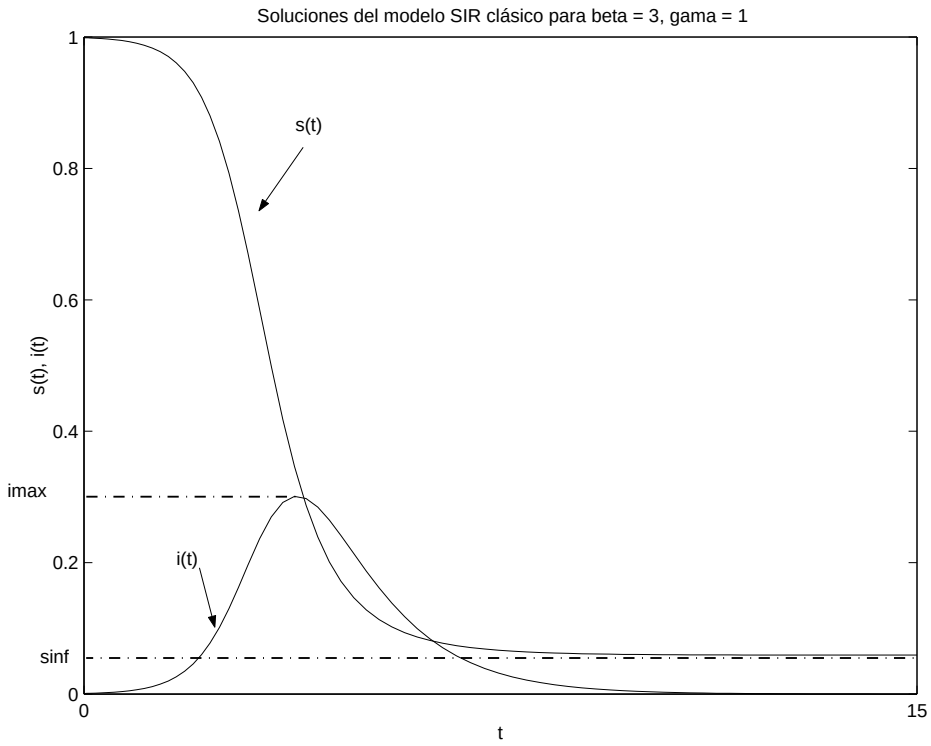
$$s_\infty + \frac{log(s_\infty)}{\sigma} = i(0) + s(0) - \frac{log(s(0))}{\sigma} \tag{1.2}$$

La Figura 2 muestra las soluciones del sistema 1.1 para $\beta = 3$ y $\gamma = 1$.

Observaciones

1. El modelo no toma en cuenta nacimientos o muertes, lo cual supone que describen la situación durante un período de tiempo corto. La notación $t \longrightarrow +\infty$ debe entenderse como un período de tiempo muy grande en

Figura 2:



comparación con $\frac{1}{\gamma}$. De modo que solamente explicará razonablemente aquellas enfermedades con tiempo medio de infección muy corto (del orden de días) en relación con el período de tiempo en el cual podemos despreciar los nacimientos y las muertes que ocurren en la población y que nada tienen que ver con la enfermedad.

2. Si el tiempo medio de infección no es muy pequeño comparado con el tiempo en el cual podemos despreciar nacimientos y muertes, entonces lo razonable es tomar los resultados hasta el tiempo

$$t_f : \ s(t_f) = s_{max}; \ i(t_f) = i_{max}$$

(si $\sigma s(0) > 1$).

3. La fracción de susceptibles $s(t)$ siempre decrece, pero su valor límite s_∞ es positivo. La epidemia se apaga puesto que cuando s cae por debajo de $\frac{1}{\sigma}$, entonces σs se hace menor que 1. Esto tiene sentido, dado que a medida que s decrece, llega un momento en que los infecciosos contactan casi enteramente a otros infecciosos, o a individuos recuperados, por lo que el número de reemplazo R se hace menor que 1. Esto es, en promedio cada infeccioso no consigue contagiar ni a un solo susceptible, por lo que al no haber nuevos contagios y recuperarse los existentes se termina la epidemia.

4. Aparece aquí un umbral, $R_0 = \sigma s(0)$, que separa el comportamiento del sistema según estemos a un lado o al otro de él.

1.4. Descripción del problema

El comportamiento de epidemias SIR sobre grafos se ha estudiado en varios trabajos ([18],[19],[20]). Mediante técnicas de percolación se obtienen expresiones explícitas para el umbral R_0, la distribución de probabilidades del tamaño final del brote epidémico, etc.

Simulaciones hechas sobre grafos del tipo *small-world* (grafos con diámetros pequeños extensamente estudiados en [18],[19],[20],[22]) avalan dichas predicciones. Dichos grafos tienen como característica distribuciones de grados del tipo potencial o exponencial (Sección 2.3). Sin embargo para poblaciones humanas, parece más apropiado una distribución unimodal similar a la gaussiana, pero asimétrica, con cola a la derecha.

Esto es, si pensamos una red donde los nodos representan individuos y las aristas entre nodos representan contactos entre individuos, la mayoría de los mismos tiene contactos con un número relativamente reducido de individuos (la gente con la que uno trata cotidianamente), y sólo unos pocos individuos tienen habitualmente contacto con muchos (p. ej. docentes, choferes de ómnibus, etc.).

Se pretende pues modelar una epidemia en grafos donde las distribuciones de grados, si bien tienen colas a la derecha, no son las típicas de las redes *small-world*.

Cada simulación puede verse como la realización de un proceso estocástico, al cual podemos asociar determinados descriptores que servirán para caracterizar al proceso. Por ejemplo: las gráficas del número de casos en función

del tiempo, la distribución de frecuencias de los grados de los nodos afectados, contribución de cada grafo en el tamaño de la epidemia, imágenes formadas con la distribución espacial del número de casos etc.

Esto permitirá aplicar a dichos descriptores las técnicas de **machine learning** para clasificación y regresión. En este trabajo se caracterizará cada epidemia mediante los dos primeros atributos mencionados, a saber: gráficas del número de casos en función del tiempo y la distribución de frecuencias de los grados de los nodos afectados. Se usarán árboles de clasificación y de regresión para datos con atributos funcionales.

Mediante los árboles de clasificación veremos si epidemias distintas en cuanto a duración y virulencia son efectivamente reconocidas como distintas (lo que avalará la elección de los descriptores elegidos). Con los árboles de regresión estudiaremos el tipo de realización más frecuente y tendremos una suerte de proceso epidémico "medio".

En el capítulo 2 veremos una síntesis de las herramientas matemáticas utilizadas para modelar epidemias sobre redes. También veremos los distintos tipos de redes que surgen como adecuadas para modelar redes sociales y definiremos medidas para características globales y locales de las mismas.

2. Enfoque clásico : Percolación y epidemias en redes

2.1. Introducción

En la vida real difícilmente se cumplan los presupuestos en los que se basa el modelo SIR anteriormente descripto. Los individuos en una población no hacen contactos al azar, y la mezcla homogénea debe reemplazarse por una red de conexiones entre individuos a través de los cuales se dan los contactos contagiantes. En este capítulo se sigue ([18],[19],[20]).

Las aristas de la red o grafo representan aquellas parejas de individuos predispuestas a tener contactos contagiantes, pero no los garantizan. Para cada nodo, las aristas de las que participa representan el conjunto de personas con las cuales el individuo asociado a dicho nodo tiene contacto durante el tiempo que dura la enfermedad (personas con las que convive, compañeros de trabajo o de estudio, desconocidos que viajan a su lado en el ómnibus, etc.)

Podemos variar la cantidad de contactos o aristas de cada nodo eligiendo cualquier distribución de grados para la red. Recordemos que el grado de un nodo es la cantidad de aristas de las cuales participa, y mide la cantidad de contactos habituales del individuo, la gente con las que se relaciona con mayor frecuencia.

Tampoco es realista suponer que la probabilidad de contagio entre parejas de individuos sea la misma para todos. Es posible que algunas parejas tengan mayor probabilidad de tener contactos contagiantes que otras.

Consideremos dos individuos o nodos i, j que están conectados. Supongamos que el individuo $i \in I$ (infectado) y que $j \in S$ (susceptible) y que la tasa media por unidad de tiempo a la que se producen contactos contagiantes entre ambos es r_{ij} y que el individuo i permanece infectado durante un período de tiempo τ_i.

Llamemos C_{ij} al evento "el individuo i contagia al individuo j ", y T_{ij} a la probabilidad de dicho evento. Dividamos el intervalo de tiempo de longitud τ_i en subintervalos de longitud Δt (tenemos $\frac{\tau_i}{\Delta t}$ subintervalos). La probabilidad de que no ocurra contagio en un subintervalo es $1 - r_{ij}\Delta t$.

Suponiendo que la probabilidad de contagio en una unidad de tiempo es independiente de lo ocurrido antes, tenemos que

$$1 - T_{ij} \cong (1 - r_{ij}\Delta t) \times \cdots \times (1 - r_{ij}\Delta t)$$

$\frac{\tau_i}{\Delta t}$ veces, o sea $1 - T_{ij} \cong (1 - r_{ij}\Delta t)^{\frac{\tau_i}{\Delta t}}$. Tomando límite cuando $\Delta t \to 0$, queda $1 - T_{ij} = e^{-r_{ij}\tau_i}$.

Entonces la probabilidad de que el individuo j sea contagiado por el i es $T_{i,j} = 1 - e^{-r_{ij}\tau_i}$. Para el caso de tiempo discreto (como en las simulaciones, donde $\Delta t = 1$), tenemos $T_{ij} = 1 - (1 - r_{ij})^{\tau_i}$, donde τ_i se mide en pasos de tiempo.

En general r_{ij} y τ_i varían de un individuo a otro. Supongamos inicialmente que son variables aleatorias iid, $r_{ij} \sim F_r,\ \tau_i \sim F_\tau$. Entonces T_{ij} también es una variable aleatoria. Como r_{ij} no tiene por qué ser simétrica (pensemos por ejemplo que i es el servidor de un sitio web, j un usuario, entonces la tasa a la que i envía paquetes de datos a j va a ser mayor que la tasa a la que j los envía a i), tampoco lo será T_{ij}. La probabilidad a priori de contagio entre dos individuos cualesquiera es la esperanza de T_{ij}.

Llamemos T a dicha esperanza, entonces

$$T = E(1 - e^{-r_{ij}\tau_i}) = 1 - E(e^{-r_{ij}\tau_i}) = 1 - \int_0^{+\infty} \int_0^{+\infty} f_r(r) f_\tau(\tau) e^{-r\tau} dr\ d\tau$$

Para el caso de tiempo discreto

$$T = 1 - \int_0^{+\infty} \sum_{\tau=0}^{+\infty} f_r(r) P_\tau(\tau) (1 - r)^\tau dr$$

Este parámetro $T \in [0, 1]$ (llamado ***transmisibilidad*** de la infección), puede interpretarse como la probabilidad de que un infectado típico contagie a un susceptible típico.

El que las probabilidades de contagio puedan variar entre individuos no supone diferencia alguna, dado que en la población considerada como un todo la enfermedad se propagará como si todas las probabilidades de transmisión fueran iguales a T. Esta hipótesis "mean-field"es la que permite tener soluciones explícitas.

Para el caso de períodos de latencia (el individuo infectado todavía no es infeccioso), como T_{ij} es la probabilidad integrada en el tiempo de transmisión de la enfermedad entre dos individuos, cualquier variación será suavizada.

Entonces el comportamiento temporal preciso de r_{ij} u otras variables no importa. Por tanto este modelo podrá tratar con cualquier variación temporal en la infectividad de los individuos infectados, y la transmisión podrá representarse mediante el parámetro T.

Imaginemos ahora que observamos un brote epidémico, el cual comienza en un individuo (nodo) y comienza a diseminarse por la red. Si marcamos u ocupamos cada arista del grafo a través de la cual se ha propagado la enfermedad (lo que ocurre con probabilidad T), el tamaño final del brote epidémico será la cantidad de nodos del subgrafo (que crece en forma arborescente) formado por las aristas ocupadas.

Esto equivale a un modelo de percolación sobre la red que representa los contactos de la población, donde la probabilidad de ocupación de cada arista es T. Este vínculo entre epidemias y percolación fue de hecho una de las motivaciones originales del modelo de percolación en sí ([6],[12]).

Por ello serán de interés las distribuciones de probabilidad de las variables aleatorias que describen el grado de un vértice elegido al azar (cuya función generatriz llamaremos G_0), el número de nodos vecinos de dicho vértice que se alcanzan al seguir una arista al azar de las que salen del mencionado nodo (cuya función generatriz denotaremos G_1), el tamaño del subgrafo formado por los nodos que pueden alcanzarse siguiendo una arista que sale de dicho vértice elegido al azar (cuya generatriz denotaremos H_1) y el tamaño total de todos los subgrafos como el anterior, o sea grafos formados por aquellos vértices alcanzables desde el vértice elegido al azar (cuya generatriz denotaremos H_0).

En lo que sigue supondremos un grafo aleatorio con una distribución de grados prefijada. Se ha observado que en las redes sociales dichas distribuciones no son simétricas, sino que exhiben colas hacia la derecha. La mayoría de los individuos tienen unos pocos contactos, pero existe un pequeño número con gran cantidad de contactos que pueden tener un gran efecto en la propagación de la epidemia.

2.2. Herramientas matemáticas

La distribución de los grados de los nodos puede describirse a través de la probabilidad p_k de que un nodo elegido al azar tenga grado k. Se pretende encontrar el comportamiento medio de un grafo con dicha distribución de grados bajo el proceso de percolación de enlaces con probabilidad de ocupación de enlaces T.
Llamando n_k a la cantidad de vértices de grado k, entonces $p_k = \frac{n_k}{\sum_{i=0}^{\infty} n_i}$.

Definición 2.1. *La función generatriz de la sucesión* $\{p_0, p_1, \ldots, p_n, \ldots\}$ *es* $G_0(x) = \sum_{k=0}^{\infty} p_k x^k$.

Para el caso en que $\{p_0, \ldots, p_n, \ldots\}$ es una distribución discreta de probabilidad, $G_0(1) = \sum_{k=0}^{\infty} p_k = 1$.

La función generatriz contiene toda la información sobre dicha distribución de probabilidad. Por ejemplo podemos recuperar la distribución a partir de la función generatriz derivando la misma tantas veces como haga falta $p_k = \frac{1}{k!}\frac{d^k G_0}{dx^k}|_{x=0}$ (por ello se dice que la función generatriz genera a la distribución).

Las funciones generatrices poseen varias propiedades que las vuelve atractivas para trabajar con ellas en lugar de hacerlo directamente con los objetos que ellas generan. En particular destacan dos de ellas:

1. Si la distribución de probabilidad de una cierta propiedad p de un objeto es generada por una función generatriz G_p, entonces la distribución de probabilidad de la suma de p sobre m realizaciones independientes del objeto estará generada por $(G_p)^m$.
   Por ejemplo, si elegimos m nodos al azar en un grafo cuya distribución de probabilidad para los grados de los nodos está generada por G_0, entonces la distribución de probabilidad para la suma de los grados de dichos vértices estará dada por $(G_0(x))^m$.

2. La esperanza de la distribución de probabilidad generada por una función generatriz está dada por la derivada primera de dicha función evaluada en $x = 1$.
   Así, el grado medio de un nodo en una red estará dado por $< k >= \sum_k k p_k = G_0'(1)$. Los momentos de orden mayor se obtienen derivando sucesivamente : $< k^n >= \sum_k^{\infty} k^n p_k = \left[\left(x\frac{d}{dx}\right)^n G_0(x)\right]_{x=1}$

Mientras que G_0 genera la distribución del grado de un nodo elegido al azar, precisamos otra función G_1 que genere la distribución del grado de los nodos alcanzados al seguir una arista elegida al azar.

Si elegimos una arista al azar y la seguimos hasta uno de los nodos que toca, entonces es más probable que dicho nodo tenga grado alto que si elegimos un nodo al azar, dado que los nodos de mayor grado son los que participan de más aristas.

Como la probabilidad de elegir una arista que toca a un nodo de grado k es $\frac{kp_k|G|}{\sum_k kp_k|G|} = \frac{1}{< k >} kp_k = c\widetilde{p}_k$, donde $c = \frac{1}{< k >}$, $\widetilde{p}_k = kp_k$, la función generatriz es $c\sum_k \widetilde{p}_k x^k = \frac{\sum_k kp_k x^k}{< k >} = x\frac{G_0'(x)}{G_0'(1)}$.

Como queremos contar la cantidad de maneras de salir de un vértice excluyendo la arista a través de la cual llegamos a él (el grado del vértice menos uno), entonces debemos dividir la función hallada entre x, lo que nos da

$$G_1(x) = \frac{G_0'(x)}{G_0'(1)} = \frac{\sum_k kp_k x^{k-1}}{< k >}$$

Ahora veamos la distribución de probabilidad para los tamaños de las componentes conexas del grafo aleatorio. Esto es, elegimos una arista al azar y la seguimos haciendo una búsqueda a lo ancho Esto es, seguimos la arista hasta un nodo, luego seguimos cada arista que sale de dicho nodo hasta los respectivos nodos en los que terminan, y repetimos el procedimiento (algoritmo BFS, [11]). En el contexto epidémico será la descendencia originada por un nodo infectado.

Llamemos $H_1(x)$ a la función generatriz de la distribución de los tamaños de las componentes conexas finitas (descendencias) así formadas. El caso de descendencias infinitas (esto es, que tienden a infinito cuando la cantidad de nodos tiende a infinito, componente gigante en la jerga de percolación) será tratado luego.

$$H_1(x) = \sum_{n=0}^{<\infty} P_c(n)x^n$$

donde $P_c(n)$ es la probabilidad de que la componente conexa tenga n nodos. Entonces $xH_1(x) = \sum_{n=0}^{\infty} P_c(n)x^{n+1} = \sum_{m=1}^{\infty} P_c(m-1)x^m$, donde $P_c(m-1)$ es la probabilidad de que la componente conexa generada a partir del nodo padre tenga $m-1$ nodos, con lo que el total contando al padre es $(m-1)+1 = m$ nodos. Cuando elegimos una arista al azar y la seguimos, el nodo al que lleguemos tendrá grado $n = 0, 1, 2, \ldots$ Sea q_k a la probabilidad de que dicho

nodo tenga k aristas a seguir (además de la empleada para llegar hasta él, de modo que q_k es la probabilidad de que dicho nodo tenga grado $k+1$).

Usando la regla de la suma, tenemos que dicha arista puede conducirnos a un nodo terminal (esto es, un nodo sin ninguna otra arista a excepción de la usada para llegar a él) con probabilidad q_0, a un nodo con una arista por la cual seguir (que producirá una componente cuya distribución de tamaños está generada por $H_1(x)$) con probabilidad q_1, a un nodo terminal con dos aristas por las cuales seguir, (cada una de las cuales producirá una componente con distribución generada por $H_1(x)$) con probabilidad q_2, etc. Asumiendo que para grafos con muchos nodos los tamaños de dichas componentes son independientes (hipótesis discutible en algunas redes), podemos aplicar la propiedad de potencias de las funciones generatrices.

En virtud de la primera de las dos propiedades destacadas de las funciones generatrices, la distribución de la suma de tamaños de cada uno de esos casos estará dada por las potencias de $H_1(x)$, por lo que

$$H_1(x) = xq_0 + xq_1H_1(x) + xq_2H_1^2(x) + \ldots$$

y como q_k es el coeficiente de x^k en la función generatriz $G_1(x)$, tenemos que

$$H_1(x) = xG_1(H_1(x))$$

Si partimos de un vértice infectado elegido al azar y vemos su descendencia, tendremos una componente al final de cada arista que sale de dicho nodo, por lo que con los argumentos utilizados para H_1 tenemos que la función generatriz para el tamaño total de la descendencia será

$$H_0(x) = xG_0(H_1(x))$$

Por tanto, si conocemos las funciones $G_0(x)$ y $G_1(x)$, podemos hallar $H_1(x)$ y con ella $H_0(x)$. Podemos por ejemplo calcular la probabilidad de que un vértice elegido al azar pertenezca a una componente de tamaño s calculando la derivada de orden s de H_0. Si P_s es la distribución de probabilidad para los tamaños s de los brotes epidémicos (descendencias), entonces $H_0(x) = \sum_{s=0}^{<\infty} P_s x^s$.

Supongamos que hay una componente gigante, de modo que $\sum_{s=0}^{<\infty} P_s x^s = 1 - S$, donde S es la fracción de la población de nodos perteneciente a la

componente gigante. Llamemos u a la fracción de nodos del grafo que no pertenecen a la componente gigante. Podemos interpretar a u como la probabilidad de que un nodo elegido al azar no pertenezca a la componente gigante, que es tambin la probabilidad de que ninguno de sus vecinos pertenezcan a la componente gigante, que es u^k si el nodo tiene grado k. Por tanto u debe cumplir (cuando la cantidad de nodos es muy grande)

$$u = \sum_{k=0}^{\infty} p_k u^k = G_0(u) \tag{2.1}$$

Ahora bien, en los cálculos anteriores se supuso que si un vértice infectado tenía una arista con otro susceptible, el contagio era seguro. Si tomamos en cuenta que existe una probabilidad de contagio u ocupación de la arista, debemos tomar en cuenta la transmisiblidad.

Llamemos $G_0(x, T)$ a la función generatriz de la distribución del número de aristas ocupadas que tocan a un nodo en función de la transmisibilidad T. De modo similar, sea $G_1(x, T)$ la función generatriz de la distribución de probabilidad de las aristas ocupadas que salen de un nodo al que se llegó siguiendo una arista elegida al azar, como función de T.

Para obtener $G_0(x, T)$, notemos que la probabilidad de que un nodo de grado k tenga exactamente m de sus aristas ocupadas está dada por la distribución binomial $\binom{k}{m}T^m(1-T)^{k-m}$ (alcanza con considerar la variable aleatoria X que cuenta la cantidad de aristas ocupadas para cada nodo, que tiene distribución $\sim Bin(k, T)$), y por tanto la distribución de probabilidad de X está generada por $G_0(x, T) = \sum_{k=0}^{\infty}\sum_{m=0}^{k} p_k\binom{k}{m}T^m(1-T)^{k-m}x^m =$

$$= \sum_{k=0}^{\infty} p_k \sum_{m=0}^{k}\binom{k}{m}(xT)^m(1-T)^{k-m} \text{ =(usando la fórmula del binomio)=}$$

$$\sum_{k=0}^{\infty} p_k(1-T+xT)^k = \sum_{k=0}^{\infty} p_k(1+(x-1)T)^k = G_0(1+(x-1)T)$$

De modo similar la distribución de probabilidad para la cantidad de aristas ocupadas que tocan a un nodo al que se llegó a través de una arista elegida al azar tiene como función generatriz a $G_1(x, T) = G_1(1+(x-1)T)$.

Esto se debe a que la variable aleatoria que cuenta la cantidad de aristas ocupadas en este caso es $X \sim Bin(k-1, T)$, y por tanto $G_1(x, T) =$

$$\sum_{k=0}^{\infty}\sum_{m=0}^{k-1}\frac{kp_k}{<k>}\binom{k-1}{m}T^m(1-T)^{k-1-m}x^m = \sum_{k=0}^{\infty}\frac{kp_k}{<k>}\sum_{m=0}^{k-1}\binom{k-1}{m}(xT)^m(1-T)^{k-1-m}$$

que es igual a $\sum_{k=0}^{\infty}\frac{kp_k}{<k>}(1-T+xT)^{k-1} = \sum_{k=0}^{\infty}\frac{kp_k}{<k>}(1+(x-1)T)^{k-1} = G_1(1+(x-1)T)$.

Observemos que $G_0(x,1) = G_0(x),\ G_0(1,T) = G_0(1),\ G_0'(1,T) = TG_0'(1)$, donde $G_0'(x,T)$ representa la derivada de G_0 respecto a su primer argumento, de modo similar para G_1.

Repitiendo argumentos anteriores, tenemos que

$$H_1(x,T) = xG_1(H_1(x,T),T)$$

y

$$H_0(x,T) = xG_0(H_1(x,T),T) = \sum_{s=0}^{<\infty} P_s(T)x^s$$

El tamaño medio $< s >$ de un brote epidémico (o descendencia de un infectado) está dado por $< s >= H_0'(1,T) = 1 + G_0'(1,T)H_1'(1,T)$.
Por otro lado $H_1'(1,T) = 1+G_1'(1,T)H_1'(1,T)$, de donde $H_1'(1,T) = \frac{1}{1-G_1'(1,T)}$.
Entonces

$$< s >= 1 + \frac{G_0'(1,T)}{1-G_1'(1,T)} = 1 + \frac{TG_0'(1)}{1-TG_1'(1)}$$

Notemos que $< s >\to \infty$ cuando $TG_1'(1) \nearrow 1$.Teniendo en cuenta que $G_1'(1) = \frac{G_0''(1)}{G_0'(1)}$ (número medio de segundos vecinos/grado medio o número medio de primeros vecinos, donde los segundos vecinos son los vecinos de los primeros vecinos), obtenemos que existe una transmisibilidad media crítica por encima de la cual el brote epidémico medio explota (los nodos infectados forman una componente gigante en la jerga de percolación). Este valor crítico es

$$T_c = \frac{G_0'(1)}{G_0''(1)}$$

Cuando $T \nearrow T_c$, tenemos una epidemia ($< s >\to +\infty$).

En este último caso aparece un brote o componente gigante (afecta a una fracción no nula de la población, el brote tiene un tamaño que tiende a infinito cuando el tamaño de la población total tiende a infinito), tenemos que

$\sum_{k=0}^{<\infty} P_s(T) = 1 - S(T)$, donde $S(T)$ es la fracción de la población de nodos perteneciente a la componente gigante.
Tenemos entonces $S(T) = 1 - H_0(1,T) = 1 - G_0(H_1(1,T),T)$.

Si ahora llamamos u a la probabilidad de que una arista elegida al azar nos lleve a un nodo que no forma parte de la componente gigante (o sea que permanezca susceptible durante una epidemia), por un argumento anlogo al que nos llevó a la ecuación (2.1), tenemos que u debe cumplir $u = G_1(u,T)$, o sea $u = H_1(1,T)$.

Observación : Aún cuando exista una epidemia (esto es, existe un brote que afecta a una fracción significativa de la población), pueden existir brotes que solamente afecten a una fracción despreciable de la población (esto es, dicha fracción tiende a cero cuando la población total tiende a infinito).

Esto contrasta con los modelos clásicos "fully-mixed"según los cuales cualquier brote provoca una epidemia cuando se supera el umbral crítico. En el modelo de percolación la probabilidad de que un brote devenga en una epidemia para una transmisibilidad T dada es $S(T)$.

La probabilidad de que un vértice no resulte infectado a través de una de sus aristas es $v = 1-T+Tu$ (la suma de la probabilidad $1-T$ de que la arista no esté ocupada y la probabilidad Tu de que esté ocupada pero conecte con un vértice no infectado). Si el vértice tiene grado k, la probabilidad total es v^k.

La probabilidad de que un vértice tenga grado k, dado que no ha sido infectado es $\frac{p_k v^k}{\sum_k p_k v^k} = \frac{p_k v^k}{G_0(v)}$, cuya función generatriz es $\frac{G_0(vx)}{G_0(v)}$.
Si derivamos y evaluamos en $x = 1$ tendremos el grado medio de los vértices no infectados, $z_{ni} = \frac{vG_0'(v)}{G_0(v)} = \frac{vG_1(v)}{G_0(v)}z = \frac{u[1-T+Tu]}{1-S}z$.

De modo similar, la distribución de los grados de los nodos infectados está generada por $\frac{G_0(x) - G_0(vx)}{1 - G_0(v)}$, de donde obtenemos el grado medio para dichos vértices, $z_{in} = \frac{1 - vG_1(v)}{1 - G_0(v)}z = \frac{1 - u[1-T+Tu]}{S}z$, donde $z =< k >$.

Observaciones:

1. Como todos los coeficientes de $G_0(x,T)$ son por definición positivos (forman una distribución de probabilidad), tenemos que
$$1 - S(T) = G_0(u,T) \leq u$$

2. Todas las derivadas de $G_0(x, T)$ son no negativas, por lo que es una función convexa.

3. $z_{ni} \leq z \leq z_{in}$

4. La probabilidad de que un vértice de grado k sea infectado es

$$1 - v^k = 1 - e^{-k \log(1/v)}$$

tendiendo exponencialmente a uno cuando el grado aumenta.

5. Pese a la potencia de estos resultados, esta teoría describe el estado o resultado al finalizar la epidemia. No nos brinda una descripción de la dinámica temporal del proceso : el tiempo está ausente.

2.3. Tipos de redes

Supongamos una población de n individuos que forman una red social con m contactos en total y un grafo $G = (V, E)$ donde $|V| = n$ y $|E| = m$. Existen en total $C^m_{n(n-1)/2}$ grafos posibles con n vértices y m aristas que podrían describir dicha red. Entre ellos destaca el grafo completo K_n con $m = n(n-1)/2$, donde todos contactan con todos (lo más cercano a la situación de fully mixed population), y puede ser usado como una suerte de referente, de hipótesis nula.

Cualquier grafo con menos aristas que K_n puede obtenerse mediante el algoritmo desarrollado por Erdös y Rëyni ([9]) para generar grafos aleatorios. Este algoritmo parte de los n vértices y elige (en forma aleatoria) aristas de un conjunto de posibles aristas. Esto permite modelar poblaciones donde los contactos son al azar (homogeneous random mixing).

Estos grafos aleatorios tienden a tener un diámetro pequeño, dado que la conexión al azar entre vértices conecta todas las partes de la red con igual frecuencia. También suelen tener pocos cúmulos de vértices, pues la probabilidad de que un vértice se conecte con un vecino de su vecino no es distinta a la de conectarse con cualquier otro vértice de la red.

En el otro extremo tenemos los reticulados o lattices, grafos altamente regulares que también pueden usarse para generar grafos aleatorios. Los reticulados tienen un alto grado de cumulación o clusterización: un vértice tiene más chance de estar conectado al vecino de su vecino que en un grafo aleatorio.

Esta estructura tan regular se traduce en grandes diámetros, tiene caminos más cortos promedialmente largos respecto de los grafos aleatorios con igual cantidad de vértices y aristas, dado que algunas regiones de la red serán topológicamente remotas respecto de otras.

La estructura de contactos en poblaciones reales son marcadamente no-homogéneas ([10]) y se ha visto que incluso pequeños niveles de heterogeneidad (tanto social como espacial) puede tener grandes efectos en por ejemplo la propagación de epidemias ([2]).

2.3.1. Redes "Small World"

En el mundo real existen restricciones que operan sobre las redes. Cada arista tiene un costo, lo que lleva a que progresivamente sea más difícil añadir aristas a un nodo dado. Cada vértice tiene así, una cierta capacidad máxima en cuanto a la cantidad de aristas de las que participa. Así, se originan redes que ni son totalmente regulares, ni son totalmente aleatorias.

También existe el envejecimiento de los nodos, que hace que su capacidad de interactuar con sus vecinos vaya decreciendo, pudiendo llegar a cero. En este caso, aunque el nodo sigue participando en las estadísticas de la red, no juega un papel activo en el proceso que discurre sobre la misma, ni aumenta su grado.

Estas restricciones operan para que las redes aleatorias observadas en el mundo real sean significativamente distintas de los grafos aleatorios o de los reticulados, produciendo redes llamadas "*small-world*", que están "a medio camino"de ambos extremos.

La experiencia sugiere ([1]) que la distribución de los grados de los vértices en este tipo de red puede agruparse en tres casos :

1. Redes "Scale-free", caracterizadas por una distribución de grados que decae potencialmente. Este tipo de redes emerge en contextos de redes que crecen y donde las nuevas aristas se conectan preferentemente a los vértices de mayor grado.

2. Redes "Single-scale", caracterizadas por una distribución de grados con colas que decaen rápido, exponencialmente.

3. Redes "Broad-scale". Estas últimas presentan un decaimiento potencial, seguido por un abrupto corte exponencial. Algo así como una mezcla de los dos tipos anteriores, con un neto predominio en los grados

bajos del tipo "Scale-free", pasando a ser "Single-scale "en los grados altos.

2.3.2. Eficiencia local y global

Los conceptos de eficiencia local y global ([15]) permiten describir características locales y globales de grafos. Seguiremos el desarrollo de ([15]). Si pensamos las aristas o links de la red como canales a través de los cuales circulan cosas (información, vehículos, contactos entre personas, etc.), cobra relevancia el poder medir la eficiencia de la red a ese respecto. Precisemos el alcance de dicho concepto.

Sea un grafo o red $G = (V, E)$ donde $|V| = n$ y $|E| = m$. Supongamos que G no es ponderado (todas las aristas tienen el mismo peso o importancia, no hay aristas preferidas), y que es disperso o ralo, esto es $m << \frac{n(n-1)}{2}$. También supondremos que G es conexo: entre cualquier par de nodos existe un camino que los conecta en un número finito de pasos.

Podemos representar a G mediante su ***matriz de adyacencia*** . Esta es una matriz $A_{n\times n}$ donde la entrada $a_{i,j}$ vale uno si existe una arista entre el nodo i y el nodo j, y cero sino. Así, el grado del nodo (cantidad de aristas de las cuales participa) i puede obtenerse sumando los unos de la fila i-sima. Como ya vimos en la sección 2.2, el grado medio de un nodo es

$$< k >= \sum_k k p_k = \frac{2m}{n}$$

A partir de la matriz de adyacencia se puede construír otra matriz D donde en cada entrada $d_{i,j}$ se encuentra la longitud del camino más corto entre los nodos i, j. Dicho camino puede hallarse usando por ejemplo el algoritmo de Dijkstra ([11]). Como G es conexo, las entradas de D serán no negativas y finitas. Para cuantificar las propiedades estructurales o topológicas de G, en ([23]) se proponen dos medidas, una de carácter global y la otra local

1. La longitud característica L, que es el promedio de las longitudes de los caminos más cortos entre nodos:

$$L = \frac{1}{n(n-1)} \sum_{i\neq j} d_{i,j}$$

con $d_{i,j}$ la longitud del camino más corto entre los nodos i, j. Es claro que L mide una característica global de G: si elegimos dos nodos i, j al

azar, el valor esperado de la longitud del camino más corto entre ellos es L.

2. El coeficiente de cumulación

$$C = \frac{1}{n}\sum_i C_i$$

donde $C_i = \dfrac{2|E_i|}{k_i(k_i-1)}$ es la cantidad de aristas que existen en el subgrafo G_i formado por los nodos conectados con el nodo i (los vecinos de i), dividido el máximo de aristas que tendría G_i si fuera un grafo completo. Esto es, C_i mide que tan lejos está G_i de ser K_{k_i}. Es claro que C_i es una característica local de G, siendo C el promedio de las mismas.

Para grafos regulares tales como reticulados, se observan valores grandes de L y de C;

$$L \sim \frac{n}{< k >},\ C = \frac{3}{4}\frac{< k > -2}{< k > -1}$$

([15]).
En el otro extremo, para grafos aleatorios se observan valores más bajos,

$$L \sim \frac{\ln(n)}{\ln(< k > -1)},\ C \sim \frac{< k >}{n}$$

([3],[5]).

Sin embargo entre medio se observa un comportamiento mixto: el grafo tiene un alto índice de cumulación similar al de los reticulados, y por otra parte tiene longitudes características pequeñas, como en los grafos aleatorios. Esta conducta es otra forma de caracterizar los grafos o redes llamados *small-world*. En [15] se propone un enfoque más general a través del concepto de *eficiencia*, mediante el cual puede describirse el comportamiento *small-world* y asignarle sentido físico. Bosquejaremos dicho enfoque.

Representaremos una red real mediante un grafo G, el cual puede ser no conexo y no necesariamente disperso. Precisaremos dos matrices: la matriz de adyacencia A ya vista, y una matriz $L = ((l_{i,j}))$ de distancias "físicas". Esto es, $l_{i,j}$ puede ser una distancia geográfica entre i, j, o una medida de la interacción entre los nodos i, j (el tiempo que demanda mandar un paquete

de datos entre routers, la probabilidad de contactar al otro nodo, etc.). Supondremos que $l_{i,j}$ es conocido aunque no exista una arista entre los nodos i, j (esto es, $a_{i,j} = 0$).
Si el grafo no es ponderado, tendremos $l_{i,j} = 1 \, \forall \, i, j = 1 \ldots, n$ si existe una arista que conecta los nodos i, j.

La longitud de un camino entre los nodos i, j será la suma de las distancias l de las aristas que forman el mismo, y la longitud $d_{i,j}$ del camino más corto será la del camino entre i, j que minimice dichas sumas.
Por tanto la matriz $D = ((d_{i,j}))$ de caminos más cortos se construye utilizando las matrices A y L. Es claro que $d_{i,j} \geq l_{i,j}$, y que la igualdad sólo se da cuando existe una arista entre el nodo i y el nodo j (esto es, son vecinos).

La eficiencia $e_{i,j}$ en la comunicación o interacción entre los nodos i, j se define como

$$e_{i,j} = \frac{1}{d_{i,j}}$$

Notemos que cuando no existe un camino entre dichos nodos (por ejemplo porque pertenecen a distintas componentes conexas de G), es $d_{i,j} = +\infty$, por lo que $e_{i,j} = 0$. Además $0 \leq e_{i,j} \leq 1$.

La ***eficiencia media*** de G se define como

$$Ef(G) = \frac{\sum_{i \neq j} e_{i,j}}{n(n-1)} = \frac{1}{n(n-1)} \sum_{i \neq j} \frac{1}{d_{i,j}}$$

Para normalizar esta cantidad la dividimos entre su máximo valor, que es el que se alcanza cuando $G = K_n$, donde $d_{i,j} = l_{i,j} \forall \, i, j = 1, \ldots, n$. Este máximo es

$$Ef_{\max} = \frac{1}{n(n-1)} \sum_{i \neq j} \frac{1}{l_{i,j}}$$

Definimos pues la eficiencia global de un grafo G

Definición 2.2. ***(Eficiencia global)***

$$Ef_{\text{glob}}(G) = \frac{Ef(G)}{Ef_{\max}} \text{ (se cumple } [0 \leq Ef_{\text{glob}}(G) \leq 1]) \qquad (2.2)$$

Así, tenemos una medida global definida aún cuando G es no conexo. Para cada nodo i, podemos usar esta medida restringida al subgrafo G_i formado por los vecinos de i y tener así una medida local de la conectividad del entorno o vecindario de i.

Definamos la *eficiencia local* (recordemos que $G = (V, E)$):

Definición 2.3.

(***Eficiencia local***)

$$Ef_{\text{loc}} = \frac{1}{n} \sum_{i \in V} Ef_{\text{glob}}(G) \tag{2.3}$$

Esta medida juega un rol muy similar al del índice de cumulación, y como $i \notin G_i$, Ef_{loc} mide la resistencia o tolerancia del grafo G a fallas (*fault tolerance*). Esto es, cuán eficiente es la comunicación entre los nodos vecinos al nodo i cuando éste es removido de la red. Al igual que antes, se tiene que $0 \leq Ef_{\text{loc}} \leq 1$.

Así, usando un único concepto (la eficiencia) podemos medir características globales y locales de una red o grafo, inclusive si es no conexa y no es dispersa.
Puede entonces redefinirse una red *small-world* en términos de su eficiencia para facilitar el flujo a través de sus aristas: Una red *small-world* es aquella que tiene alta eficiencia global y alta eficiencia local.

Para el cálculo de la eficiencia global se programó un algoritmo que utiliza Dijkstra para fabricar la matriz D, luego se utiliza (2.2). La eficiencia local se calcula mediante un algoritmo que, para cada nodo i primero halla G_i, luego usa Dijkstra para hallar la matriz D_i asociada a G_i. Por último se utiliza (2.3).

En el capítulo 3 daremos una síntesis de los elementos de la teoría de **Machine Learning** (Statistical Learning, Aprendizaje Automático) utilizada en este trabajo.

3. Elementos de Machine Learning

3.1. Definiciones y resultados

3.1.1. Clasificadores y error

Sean (X, Y) variables aleatorias, X a valores en un espacio $\mathbb{X}$ (habitualmente $\mathbb{X} = \mathbb{R}^d$ salvo que expresamente digamos lo contrario), Y a valores en $\{0, 1\}$. Las X son observaciones con un objeto Y (conocido) asociado. Cuando llegan nuevas observaciones X, deseamos predecir o asignarle el correspondiente Y.

Si bien luego trabajaremos con Y a valores en $\{1, \ldots, m\}$, el uso árboles binarios permite reducir todo al caso de valores $\{0, 1\}$.

Sean $\mu : \mu(A) = P(X \in A)$ y $\eta : \eta(x) = P(Y = 1|X = x) = E(Y|X = x)$. Entonces μ, η describen la distribución del par (X, Y).

En efecto, $\forall\, C \subset \mathbb{R}^d \times \{0, 1\}$ podemos escribir

$$C = (C \cap (\mathbb{R}^d \times \{0\})) \cup (C \cap (\mathbb{R}^d \times \{1\})) \stackrel{Def.}{=} (C_0 \times \{0\}) \cup (C_1 \times \{1\})$$

por lo que

$$P(\{(X, Y) \in C\}) = P(\{X \in C_0, Y = 0\}) + P(X \in C_1, Y = 1) =$$

(si C es Borel-medible)

$$= \int_{C_0} (1 - \eta(x))\mu(dx) + \int_{C_1} \eta(x)\mu(dx)$$

Entonces la pareja (μ, η) determina la distribución de (X, Y).

La función $\eta : \mathbb{R}^d \to [0, 1]$ suele llamarse ***probabilidad a posteriori***.

Definición 3.1. ***(Clasificador o función de decisión)***

Llamaremos así a cualquier función $g : \mathbb{R}^d \to \{0, 1\}$.

Definición 3.2. ***La probabilidad de error***

o simplemente ***error*** de g es

$$L(g) = P(\{g(X) \neq Y\})$$

De particular relevancia es el clasificador de Bayes,

$$g^*(x) = \begin{cases} 1 & \text{si } \eta(x) > \frac{1}{2} \\ 0 & \text{sino} \end{cases}$$

dado que es el que minimiza la probabilidad de error. Esto es:

Teorema 3.3.

Para todo clasificador $g : \mathbb{R}^d \to \{0, 1\}$ se tiene que

$$P(\{g^*(X) \neq Y\}) \leq P(\{g(X) \neq Y\})$$

O sea que g^* es la mejor función de decisión, es optimal. La demostración puede verse en [8].

De la prueba del enunciado anterior surge que para cualquier clasificador g:

$$L(g) = 1 - E(\mathbb{I}_{\{g(X)=1\}}\eta(X) + \mathbb{I}_{\{g(X)=0\}}(1 - \eta(X)))$$

En particular para el clasificador de Bayes,

$$L^* = P(\{g^*(X) \neq Y\}) = 1 - E(\mathbb{I}_{\{\eta(X)>\frac{1}{2}\}}\eta(X) + \mathbb{I}_{\{\eta(X)\leq\frac{1}{2}\}}(1 - \eta(X))$$

A la probabilidad de error L^* se le llama ***error de Bayes o riesgo de Bayes.***

La probabilidad a posteriori $\eta(x) = P(\{Y = 1|X = x\}) = E(Y|X = x)$ minimiza el error cuadrático cuando Y es predicho por alguna $f : \mathbb{R}^d \to \mathbb{R}$. Esto es,

$$E((\eta(X) - Y)^2) \leq E((f(X) - Y)^2)$$

La prueba puede encontrarse en [8].

3.1.2. Clasificadores Plug-in

Dada la observación X, la mejor predicción para Y está dada por la función de decisión de Bayes,

$$g^*(x) = \begin{cases} 1 & \text{si } \eta(x) > \frac{1}{2} \\ 0 & \text{sino} \end{cases} = g^*(x) = \begin{cases} 1 & \text{si } \eta(x) \geq 1 - \eta(x) \\ 0 & \text{sino} \end{cases}$$

La función η es en general desconocida. Supongamos que tenemos funciones no negativas $\widetilde{\eta}, 1-\widetilde{\eta}$ que aproximan a η y $1-\eta$ respectivamente. Entonces es natural usar los clasificadores plug-in:

$$g(x) = \begin{cases} 1 & \text{si } \widetilde{\eta}(x) > \frac{1}{2} \\ 0 & \text{sino} \end{cases}$$

como aproximación a g^*.

El siguiente teorema establece que si $\widetilde{\eta}(x)$ está cerca de $\eta(x)$ en la norma L_1, entonces $L(g)$ estará cercano a $L(g^*)$.

Teorema 3.4.

Sea g el clasificador plug-in definido arriba, entonces

$$L(g) - L(g^*) = 2\int_{\mathbb{R}^d} |\eta(x) - \frac{1}{2}\mathbb{I}_{\{g(X)\neq g^*(X)\}}|\mu(dx)$$

además,

$$L(g) - L(g^*) \leq 2\int_{\mathbb{R}^d} |\eta(x) - \widetilde{\eta}(x)|\mu(dx) = 2E(|\eta(X) - \widetilde{\eta}(X)|)$$

La prueba de este teorema puede consultarse en [8].

Cuando el clasificador g es de la forma

$$g(x) = \begin{cases} 1 & \text{si } \widetilde{\eta_1}(x) \geq \widetilde{\eta_0}(x) \\ 0 & \text{sino} \end{cases}$$

donde $\widetilde{\eta}_1(x)$ y $\widetilde{\eta}_0(x)$ son aproximaciones de $\eta(x)$ y $1-\eta(x)$ respectivamente, tenemos el siguiente resultado:

Teorema 3.5.

Para el clasificador g definido arriba, se tiene que

$$L(g) - L(g^*) \leq \int_{\mathbb{R}^d} |(1-\eta(x)) - \widetilde{\eta}_0(x)|\mu(dx) + \int_{\mathbb{R}^d} |\eta(x) - \widetilde{\eta}_1(x)|\mu(dx)$$

(ver [8])

3.1.3. Consistencia

Supongamos una secuencia de datos de entrenamiento

$$D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$$

donde las X forman una muestra aleatoria simple. Lo ideal en un clasificador es que alcance la probabilidad de error de Bayes, L^*. Aunque no podemos esperar dicho resultado ideal, sí podemos construír una secuencia de clasificadores $\{g_n\}$ (también llamada ***regla de decisión***) de modo tal que la probabilidad de error pueda estar arbitrariamente cerca del óptimo L^* con gran probabilidad (esto es, para la mayoría de los D_n).

Más precisamente, sea $L_n = L(g_n) = P(\{g_n(X, D_n) \neq Y | D_n\})$.

Definición 3.6. ***(Consistencia débil y fuerte)***

Una regla de decisión es ***consistente*** en sentido débil para una distribución del par (X, Y) si

$$E(L_n) = P(\{g_n(X, D_n) \neq Y\}) \underset{n}{\to} L^*$$

Una regla de decisión es ***fuertemente consistente*** si

$$\lim_n L_n = L^*$$

con probabilidad 1.

Observación:
La consistencia débil es pues la convergencia de $E(L_n)$ a L^*. Como L_n es una variable aleatoria a valores entre L^* y 1, esta convergencia equivale a la convergencia de L_n a L^* en probabilidad. Esto es,

$$\forall \epsilon > 0, \lim_n P(\{L_n - L^* > \epsilon\}) = 0$$

Como la convergencia casi segura (convergencia en un conjunto de probabilidad 1) implica la convergencia en probabilidad, tenemos que la consistencia fuerte implica la consistencia débil.

Una regla de decisión consistente garantiza que incrementando el número de datos, la probabilidad de que la probabilidad de error esté cerca del óptimo alcanzable (el riesgo de Bayes), puede acercarse a 1 tanto como uno quiera.

Intuitivamente, la regla puede eventualmente aprehender al clasificador óptimo a partir de un gran número de datos con alta probabilidad. La consistencia fuerte nos dice que si usamos más datos, la probabilidad de error tiende al riesgo de Bayes para cualquier secuencia de datos de entrenamiento, excepto un subconjunto de probabilidad 0.

Una regla de decisión puede ser consistente para una cierta clase de distribuciones del par (X, Y) pero no para otras. Claramente es deseable tener una regla que sea consistente para tantas distribuciones como se pueda. Como en muchas situaciones no tenemos información previa sobre la distribución, es fundamental una regla que tenga buena performance para cualquier distribución. Esto motiva la siguiente definición.

Definición 3.7. ***(Consistencia universal)***

Una regla se dice ***universalmente (fuertemente)consistente*** si es (fuertemente) consistente para cualquier distribución del par (X, Y).

La consistencia de las reglas de clasificación puede deducirse de la consistencia de la estimación en regresión. Es claro que

$$\eta(x) = P(\{Y = 1|X = x\} = E(Y|X = x))$$

es la regresión de Y en X. En muchos casos la probabilidad a posteriori $\eta(x)$ se estima a partir de los datos de entrenamiento D_n por una función $\eta_n(x) = \eta_n(x, D_n)$.

En este caso, la probabilidad de error $L(g_n) = P(\{g_n(X) \neq Y|D_n\})$ del clasificador plug-in

$$g_n(x) = \begin{cases} 0 & \text{si } \eta_n(x) \leq \frac{1}{2} \\ 1 & \text{sino} \end{cases}$$

es una variable aleatoria, y por el teorema 3.4 tenemos:

Corolario 3.8.

La probabilidad de error del clasificador $g_n(x)$ definido arriba satisface

$$L(g_n) - L^* \leq 2\int_{\mathbb{R}^d} |\eta(x) - \eta_n(x)|\mu(dx) = 2E(|\eta(X) - \eta_n(X)||D_n)$$

De la desigualdad de Cauchy-Schwarz surge este otro corolario:

Corolario 3.9.

Si g_n es el clasificador definido arriba, su probabilidad de error satisface

$$P(\{g_n(X) \neq Y|D_n\}) - L^* \leq 2\sqrt{\int_{\mathbb{R}^d} |\eta(x) - \eta_n(x)|^2 \mu(dx)}$$

La consecuencia más interesante del teorema 3.4 es que la mera existencia del estimador η_n de la función de regresión η, para el cual

$$\int_{\mathbb{R}^d} |\eta(x) - \eta_n(x)|^2 \mu(dx) \to 0$$

en probabilidad (o con probabilidad uno), implica que la regla de decisión plug-in g_n es consistente (o fuertemente consistente).

A partir del teorema 3.5 podemos obtener un corolario análogo al corolario 3.8 cuando las probabilidades $\eta_0(x) = P(\{Y = 0|X = x\})$ y $\eta_1(x) = P(\{Y = 1|X = x\})$ se estiman a partir de datos mediante algunas $\eta_{0,n}$ y $\eta_{1,n}$ respectivamente. La consistencia de la regla de clasificación habitualmente se prueba escribiendo la regla como plug-in y estableciendo la convergencia de $\eta_{0,n}$ y $\eta_{1,n}$ a η_0 y η_1 respectivamente en L_1.

3.1.4. Reglas de Partición o Split

Muchas reglas de clasificación (en particular los árboles) particionan a $\mathbb{R}^d$ en subconjuntos disjuntos $A_1, A_2, \ldots$ y clasifican en cada uno de ellos según el voto mayoritario entre las etiquetas de los X que habitan el mismo subconjunto. Esto es :

$$g_n(x) = \begin{cases} 0 & \text{si } \sum_{i=1}^{n} \mathbb{I}_{\{Y_i=1\}} \mathbb{I}_{\{X_i \in A(x)\}} \leq \sum_{i=1}^{n} \mathbb{I}_{\{Y_i=0\}} \mathbb{I}_{\{X_i \in A(x)\}} \\ 1 & \text{sino} \end{cases}$$

donde $A(x)$ es el subconjunto que contiene a x.

Las particiones pueden cambiar con n, también pueden depender de los puntos $X_1, \ldots, X_n$, pero supondremos que las etiquetas no juegan ningún rol en la construcción de la partición. El siguiente teorema es un resultado general de consistencia para tales reglas de partición. Las mismas deberán cumplir dos propiedades :

1. Los subconjuntos deben ser lo suficientemente pequeños como para poder detectar cambios locales de la distribución.

2. Los subconjuntos deben ser lo suficientemente grandes como para contener una cantidad de puntos tal que el promediado entre etiquetas sea efectivo.

Si llamamos diam(A) al diámetro de A (esto es, diam$(A) = \sup_{x,y \in A} ||x-y||$), sea

$$N(x) = n\mu_n(A(x)) = \sum_{i=1}^{n} \mathbb{I}_{\{X_i \in A(x)\}}$$

la cantidad de X que caen en el mismo subconjunto que x. Las condiciones para el teorema requieren que este subconjunto aleatorio (elegido según la distribución de X), tenga diámetro pequeño y contenga muchos puntos, con alta probabilidad.

Teorema 3.1.3

Consideremos una regla de clasificación que particiona a $\mathbb{R}^d$ como arriba. Entonces $E(L_n) \to L^*$ si:

1. diam$(A(X)) \to 0$ en probabilidad

2. $N(X) \to +\infty$ en probabilidad

La prueba puede verse en [8].

El siguiente teorema (Stone, 1977) nos permite deducir la consistencia universal de varias reglas de clasificación.

Teorema 3.10. ***(Teorema de Stone)***

Supongamos que para cualquier distribución de X los pesos satisfacen las siguientes tres condiciones:

1. Existe una constante c tal que para cualquier función no negativa medible f con $E(f(X)) < \infty$ se cumple $E\left(\sum_{i=1}^{n} W_{ni}(X) f(X_i)\right) \leq cE(f(X))$

2. Para todo $a > 0$, $\lim_{n\to\infty} E\left(\sum_{i=1}^{n} W_{ni}(X)\mathbb{I}_{\{||X_i - X|| > a\}}\right) = 0$

3. $\lim_{n\to\infty} E\left(\max_{1\leq i\leq n} W_{ni}(X)\right) = 0$

Entonces g_n es universalmente consistente. La prueba puede consultarse en [8].

Veamos un ejemplo : sea la regla

$$\eta_n(x) = \sum_{i=1}^{n} \mathbb{I}_{\{Y_i=1\}} W_{ni}(x) = \sum_{i=1}^{n} Y_i W_{ni}(x)$$

donde los pesos $W_{ni}(x) = W_{ni}(x, X_1, \ldots, X_n)$ son no negativos y además $\sum_{i=1}^{n} W_{ni}(x) = 1$.

La regla de clasificación es

$$g_n(x) = \begin{cases} 0 & \text{si } \sum_{i=1}^{n} \mathbb{I}_{\{Y_i=1\}} W_{ni}(x) \leq \sum_{i=1}^{n} \mathbb{I}_{\{Y_i=0\}} W_{ni}(x) \\ 1 & \text{sino} \end{cases}$$

o lo que es lo mismo

$$g_n(x) = \begin{cases} 0 & \text{si } \sum_{i=1}^{n} Y_i W_{ni}(x) \leq \frac{1}{2} \\ 1 & \text{sino} \end{cases}$$

El estimador η_n es un promedio ponderado. Intuitivamente parece razonable esperar que las parejas (X_i, Y_i) donde X_i está cerca de x provean mayor información sobre $\eta(x)$ que las lejanas. Por tanto los pesos serán típicamente mayores en las cercanías de X, por lo que podemos interpretar a $\eta_n(x)$ como la frecuencia relativa con que aparecen los X con etiqueta 1 entre los puntos cercanos a X, un estimador del promedio local. Por su parte g_n sería un voto mayoritario ponderado localmente.

3.2. Árboles

3.2.1. Generalidades

Los árboles son modelos estadísticos que predicen una respuesta (una etiqueta o clase en caso de clasificación, un número o vector Y en caso de

regresión) dado un conjunto de variables explicativas $X_1, \ldots, X_n$. A partir de un conjunto de observaciones D_n estos modelos particionan al espacio X y predicen la respuesta Y mediante un modelo que es una función constante en cada subconjunto de la partición u hojas.

Dado un nuevo dato (X) se le asigna un Y (o se predice Y) como el valor de la respuesta asociado a la hoja o miembro de la partición al que pertenece X. Este valor puede ser la etiqueta mayoritaria en la hoja a la que pertenece X (si estamos haciendo clasificación), o el promedio de los Y_i correspondientes a los X_i pertenecientes al elemento de la partición al que pertenece X (si estamos haciendo regresión).

Los árboles particionan al espacio $\mathbb{X}$ (por lo general $\mathbb{X} = \mathbb{R}^d$), usualmente en hiperrectángulos paralelos a los ejes. De particular relevancia son los árboles binarios, donde cada nodo da origen a dos subnodos o hijos. En un árbol de clasificación cada nodo representa un subconjunto de $\mathbb{R}^d$, el cual puede o no ser vuelto a partir (cada nodo tendrá 2 hijos o ninguno). Si un nodo u representa al conjunto A y sus hijos u', u'' representan a A', A'' respectivamente, entonces $A = A' \cup A''$ con $A' \cap A'' = \phi$. El nodo raíz representa a todo el espacio, mientras que los nodos terminales u hojas constituyen una partición del espacio; cada clase estará asociada a una hoja.

Dado $x = (x^1, \ldots, x^d) \in A$, las preguntas que performan dicha partición suelen ser del tipo:

1. ¿$x^i \leq \alpha$? Esto nos lleva a los árboles binarios, los cuales particionan a $\mathbb{R}^d$ en hiperrectángulos.

2. ¿$a_1 x^1 + \cdots + a_d x^d \leq \alpha$? Esto nos lleva a los árboles oblicuos. En este caso cada split insume más tiempo, pero la partición del espacio en poliedros convexos es más flexible (aunque más difícil de interpretar).

3. ¿$||x - z|| \leq \alpha$? (Aquí $z \in \mathbb{R}^d$ puede cambiar en cada nodo). Esto induce una partición en bolas, de donde a dichos árboles se les llamará árboles esféricos.

4. ¿$\psi(x) \geq 0$? Donde ψ es una función no lineal, distinta en cada nodo. De hecho todo clasificador puede escrbirse así, puesto que decidimos mandar al punto x a una clase u otra según $\psi(x) \geq 0$ o no para una cierta ψ. Sin embargo dichos árboles son de difícil interpretación a la hora de entender relevancia de las variables y posibles causalidades.

Definición 3.11. ***(Clasificador natural)***

Si una hoja representa una región A, ***diremos que el clasificador (árbol)*** g_n ***es natural*** si

$$g_n(x) = \begin{cases} 1 & \text{si } \sum_{i:X_i \in A} Y_i > \sum_{i:X_i \in A} (1 - Y_i),\ x \in A \\ 0 & \text{sino} \end{cases}$$

Esto es, en cada región (u hoja) tomamos el voto mayoritario sobre todas las parejas (X_i, Y_i) con X_i en dicha región. Los empates se fallan a favor de una clase (por lo general de forma pesimista, asignándolos a la clase menos conveniente para nuestros fines).

3.2.2. Impureza

Si A es la región a ser partida en A' y A'', definamos

$$A'_i(A, \alpha) = \{(X_j, Y_j) : X_j \in A,\ X_j^i \leq \alpha\}$$
$$A''_i(A, \alpha) = \{(X_j, Y_j) : X_j \in A, X_j^i > \alpha\}$$

(esto es, los conjuntos de parejas que caen en los hijos izquierdo y el derecho respectivamente cuando la decisión se toma mirando a la componente i-sima). Llamemos $N'_i(A, \alpha) = |A'_i(A, \alpha)|$ y $N''_i(A, \alpha) = |A''_i(A, \alpha)|$ a la cantidad de dichas parejas, y contemos entre ellas las que tienen etiquetas cero o uno mediante

$$N'_{i,0}(A, \alpha) = |\{(X, Y) \in A'_i(A, \alpha) : Y = 0\}|$$
$$N'_{i,1}(A, \alpha) = |\{(X, Y) \in A'_i(A, \alpha) : Y = 1\}|$$
$$N''_{i,0}(A, \alpha) = |\{(X, Y) \in A''_i(A, \alpha) : Y = 0\}|$$
$$N''_{i,1}(A, \alpha) = |\{(X, Y) \in A''_i(A, \alpha) : Y = 1\}|$$

Definición 3.12. ***(Función de impureza)***

Definamos una función de impureza para un posible split (i, α) como

$$I_i(A, \alpha) = \psi\left(\frac{N'_{i,0}}{N'_{i,0} + N'_{i,1}}, \frac{N'_{i,1}}{N'_{i,0} + N'_{i,1}}\right) N'_i + \psi\left(\frac{N''_{i,0}}{N''_{i,0} + N''_{i,1}}, \frac{N''_{i,0}}{N''_{i,0} + N''_{i,1}}\right) N''_i$$

Donde ψ es una función no negativa que satisface los siguientes requerimientos:

1. $\psi(\frac{1}{2}, \frac{1}{2}) \geq \psi(p, 1-p)\ \forall p \in [0,1]$

2. $\psi(0,1) = \psi(1,0) = 0$

3. $\psi(p, 1-p)$ es creciente en $[0, \frac{1}{2}]$ y decreciente en $[\frac{1}{2}, 1]$

De todas las particiones posibles de A en A', A'' elegiremos aquella que minimiza la impureza de A', A''. Esto es, la que hace mayor la diferencia entre la impureza del padre y las impurezas sumadas de los hijos. Pero esto de por sí no garantiza la consistencia del estimador, a menos que la regla de partición basada en la disminución de la impureza cumpla las hipótesis del teorema 3.8.

3.2.3. Consistencia de un árbol

Siguiendo la notación de 3.1.4, sea $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ un conjunto de datos, sean $A_1, \ldots, A_N$ (donde N puede ser aleatorio) las hojas del árbol construído a partir de D_n. Dichas hojas conforman una partición del espacio de las X. Sea N_j la cantidad de $X \in A_j$ (de donde $\sum_{j=1}^{N} N_j = n$). Recordemos que $\text{diam}(A_j)$ es el diámetro de A_j, esto es, la máxima distancia entre dos puntos de A_j. La clasificación se hace por voto mayoritario, de modo que si $x \in A_j$, la regla es

$$g_n(x) = \begin{cases} 1 & \text{si } \sum_{i:X_i \in A_j} Y_i > \sum_{i:X_i \in A_j} (1 - Y_i),\ x \in A_j \\ 0 & \text{sino} \end{cases}$$

Denotemos por $A(x)$ el A_j donde cae x, y $N(x)$ la cantidad de X que caen en $A(x)$. El teorema 3.8 nos dice que si

1. $\text{diam}(A(X)) \to 0$ en probabilidad

2. $N(X) \to +\infty$ en probabilidad

entonces $E(L_n^*) \to L^*$.

3.3. CART para datos funcionales

Daremos una descripción de la herramienta CART (acrónimo de Classification And Regression Trees). Los datos funcionales aparecen como forma dominante en virtualmente toda disciplina científica. En medicina por ejemplo, cada paciente tiene una historia clínica que contiene datos tanto de naturaleza cuantitativa como cualitativa, varios de los cuales aparecen bajo la forma de series de tiempo o gráficos.

Un dato funcional $\mathcal{A}$ consiste en una lista de valores $(a_1, \ldots, a_n)$ (generalmente ordenados cronológicamente), donde a_i (o $\mathcal{A}(t_i)$) denota la medida $(t_i, \mathcal{A}(t_i))$ en tiempo i. Como en general dichos datos son series de tiempo, de aquí en adelante los llamaremos TS.

Un conjunto de datos D_n consiste de n individuos o ejemplos $\{e_1, \ldots, e_n\}$ (las $X_1, \ldots, X_n$ de antes). Cada ejemplo e_i tiene K atributos $\mathcal{A}_{i,1}, \ldots, \mathcal{A}_{i,K}$ y una etiqueta de clase m_i que denota la clase a la que pertenece el ejemplo (hasta ahora sólo teníamos dos valores posibles, 0 y 1). De ahora en más llamaremos D_n al conjunto formado por las antiguas X's, sin las correspondientes Y's. Los atributos que pertenecen a una misma clase deben tener la misma longitud.

Construiremos un clasificador basado en D_n, el cual será usado luego para predecir la clase a la cual pertenece un nuevo ejemplo o individuo e (en el caso de clasificación), o asignarle un determinado valor u objeto (regresión).

Por ejemplo, D_n puede ser un archivo médico con registros de pacientes $e_1, \ldots, e_n$ cuya evolución es conocida. Entonces, cuando aparece un nuevo ejemplo e, parece natural buscar en D_n para hallar el caso más parecido para así realizar una predicción (¡ y un tratamiento !) basado en dicha información.

Dicho esto, el problema que surge es la comparación entre datos que corresponden a series de tiempo o a la evolución temporal de ciertos parámetros, para decidir si lucen similares o no.

El árbol clasificador construído a partir de la muestra de entrenamiento tiene un atributo en sus nodos internos, y produce la partición (split) basado en la similitud entre parejas de dicha muestra.

Para dicho split test elegimos al "mejor"individuo que existe en los datos (el individuo que produce la partición más homogénea entre quienes se le parecen y quienes no) mediante una búsqueda exhaustiva entre los individuos del nodo ([24]).

3.3.1. Introducción

La entrada (muestra de entrenamiento) es una matriz donde cada fila corresponde a un individuo. Las primeras cuatro columnas contienen el número y la clase o valor que caracterizan al individuo ("sano", "enfermo", probabilidad de padecer cierta enfermedad, tipo de epidemia, año lluvioso o seco, etc.). En este trabajo cada individuo es el resultado de simular una epidemia, y como etiqueta de clase usaremos los parámetros (μ, σ) de la distribución normal utilizada para sortear los tiempos de vida como infecciosos de los nodos (individuos) contagiados.

Las restantes columnas contienen los atributos que serán utilizados. Supongamos que el primer atributo es un vector de n_1 componentes, el segundo de n_2 componentes, etc. Las columnas desde la 5 hasta la $5 + n_1$ contienen las medidas que forman el primer atributo, las columnas desde la $5 + n_1 + 1$ hasta la $5 + n_1 + 1 + n_2$ contienen el segundo atributo, etc.

Usaremos dos atributos: la distribución de frecuencias de los nodos según su grado (cantidad de aristas que lo involucran, o cantidad de nodos con los cuales comparte una arista), que será el atributo 1. Como atributo 2 usaremos la curva cantidad de casos/tiempo (cantidad de infectados por unidad de tiempo).

Un split-test $st(e, \mathcal{A}, \theta)$ consiste de un ejemplo standard e, un atributo $\mathcal{A}$ y un umbral θ. Sea el valor $e(\mathcal{A})$ para el ejemplo e en el atributo $\mathcal{A}$, y la medida de disimilaridad (que no tiene por qué ser una distancia) $d((e(\mathcal{A})), (e'(\mathcal{A})))$ entre TS, entonces el split-test divide el nodo actual en dos conjuntos disjuntos $S_1(e, \mathcal{A}, \theta),\ S_2(e, \mathcal{A}, \theta)$. S_1 es el conjunto de ejemplos e' para los cuales $d((e(\mathcal{A})), (e'(\mathcal{A}))) < \theta$, y S_2 es el complemento.

A continuación una vista global del algoritmo:

1. Dado el conjunto de ejemplos actual (nodo), hallar su mejor partición. Para eso, repetir :

   a) Para cada ejemplo e en el nodo actual : para cada atributo $\mathcal{A}$, ordenar los otros ejemplos e' en el nodo usando $d(e(\mathcal{A}), e'(\mathcal{A}))$.

   b) Para cada e' tomar $\theta' = d(e(A), e'(\mathcal{A}))$ y construir el correspondiente split $S_1(e, \mathcal{A}, \theta'),\ S_2(e, A, \theta')$.

   c) Elegir la mejor partición S_1, S_2 del nodo actual (implica una medida para la bondad de la partición).

d) Si el split es aceptado, añadir los nuevos dos nodos a la lista de nodos que esperan inspección para split (decidir si se parten o no).

2. Repetir (a) con el siguiente nodo en la lista.

La salida es una lista con los ejemplos o individuos e, los umbrales θ, los atributos $\mathcal{A}$ y las etiquetas o valores de los individuos que forman cada hoja.

3.3.2. Medida de disimilaridad

Dadas dos muestras de entrenamiento A y B, una forma muy simple de compararlas es el uso de la distancia Euclídea entre parejas (a_i, b_i).

Sin embargo está el caso en que las dos secuencias tienen aproximadamente la misma forma general, pero no están verticalmente alineadas en el eje X (usualmente eje temporal).

Para medir la similaridad entre secuencias debemos permitir deformaciones en el eje X de una o ambas secuencias para tener el mejor alineamiento.

El Dynamic Time Warping (DTW)([14]) es una técnica para lograr esto, la cual intenta explicar la variabilidad en el eje Y deformando el eje X. Bosquejaremos la versión más simple de la DTW, que es la aplicada en este caso.

Sean $A = (a_1, \ldots, a_n)$, $B = (b_1, \ldots, b_m)$ y una distancia d. Para alinear dichas secuencias usando DTW, construímos una matriz $D_{A,B} = ((d_{i,j}))$, $i = 1 \ldots, n$; $j = 1 \ldots m$ de tamaño $n \times m$, donde $d_{i,j} = d(a_i, b_j)$ (tipicamente d es la distancia Euclídea en $\mathbb{R}^2$).

Cada $d_{i,j}$ corresponde al alineamiento entre los puntos a_i y b_j. Un *warping path* W es un conjunto de elementos de la matriz "diagonalmente contiguos"que definen un mapa entre A y B. El k-ésimo elemento de W se define como $w_k = (di, j)_k$, y por tanto tenemos :

$$W = w_1, w_2, \ldots, w_k, \ldots, w_K, \text{ con } \max(n, m) \leq K < n + m + 1$$

W está sujeto por lo general a varias restricciones.

1. **Condiciones de borde:** $w_1 = d_{1,1}$ y $w_K = d_{n,m}$. Esto significa que el warping path comienza y termina en celdas de las esquinas diagonalmente opuestas de $D_{A,B}$.

2. **Contigüidad :** Dado $w_k = d_{i,j}$, entonces $w_{k+1} = d_{i',j'}$, donde $i - i' \leq 1$ and $j - j' \leq 1$. Esto restringe los pasos factibles en el warping path a celdas adjacentes, incluyendo celdas diagonalmente adyacentes.

3. **Monotonía :** Dado $w_k = d_{i,j}$ entonces $w_{k-1} = d_{i',j'}$ donde $0 \leq i - i'$ y $0 \leq j - j'$. Esto fuerza a los puntos en W a estar monótonamente espaciados en el tiempo (eje X).

En cierto modo los warping paths juegan el rol de los Q-Q plots en estadística. Existen muchos warping paths que satisfacen las condiciones mencionadas, estaremos interesados en el camino que minimiza el costo

$$DTW(A, B) = \text{mín} \left\{ \frac{\sqrt{\sum_{k=1}^{K} w_k}}{K} \right\}$$

(El denominador K compensa el hecho de que los warping paths pueden tener distintas longitudes)

Este camino óptimo puede hallarse usando programación dinámica para evaluar la siguiente recurrencia, la que define la distancia acumulada $\gamma(i, j)$ como la distancia $d_{i,j}$ hallada en la celda actual y el mínimo de las distancias acumuladas de los elementos adyacentes.

$$\gamma(i, j) = d_{i,j} + \text{mín}\{\gamma(i-1, j-1), \gamma(i-1, j), \gamma(i, j-1)\}$$

3.3.3. Eligiendo el mejor split

Clasificación

Sea D_n un conjunto de ejemplos (individuos) que toman valores sobre m clases posibles $F_1, \ldots, F_m$ y $\{\mathcal{A}_1, \ldots, \mathcal{A}_K\}$ un conjunto de atributos.
Para cada atributo $\mathcal{A}_k$ definamos

$$P_j = \frac{|D_n \cap F_j|}{|D_n|}$$

$$I(D_n) = -\sum_{j=1}^{m} P_j \log_2 P_j$$

$$E(\mathcal{A}_k, D_n) = \frac{1}{|D_n|} \sum_{i=1}^{p} |D_{n,i}| I(D_{n,i})$$

donde

- P_j es la probabilidad de ocurrencia de cada clase F_j en el conjunto D_n de ejemplos, definida como la proporción de ejemplos en D_n que pertenecen a la clase F_j.
- $I(D_n)$ Es la entropía del conjunto (una medida de la aleatoriedad de la distribución de los ejemplos entre las m clases o etiquetas, su grado de desorden).
- p es el número de valores posibles del atributo $\mathcal{A}_k$.
- $|D_{n,i}|$ es el número de ejemplos en D_n que toman el valor V_i para el atributo $\mathcal{A}_k$.
- $|D_n|$ es el número de ejemplos en el nodo.

Los conjuntos $D_{n,1}, \ldots, D_{n,p}$ forman una partición de D_n generada por los p valores de $\mathcal{A}_k$ mientras que $I(D_{n,i})$ mide la aleatoriedad de la distribución de los ejemplos en el conjunto $D_{n,i}$ sobre las clases posibles y está dado por

$$I(D_{n,i}) = -\sum_{j=1}^{p} Pj|i \ \log_2(Pj|i)$$

donde $Pj|i = \dfrac{|D_{n,i} \cap F_j|}{|D_{n,i}|}$.

$E(\mathcal{A}_k, D_n)$ es por tanto la información esperada al partir el conjunto D_n usando $\mathcal{A}_k$. Esta información esperada es el promedio ponderado sobre los p valores del atributo $\mathcal{A}_k$ de las medidas $I(D_{n,i})$.

La siguiente fórmula (*ganancia de información de Quinlan*) mide la información ganada al partir usando los valores del atributo $\mathcal{A}_k$:

Definición 3.13. ***(ganancia de información de Quinlan)***

$$G(\mathcal{A}_k, D_n) = I(D_n) - E(\mathcal{A}_k, D_n)$$

esto es, mide la diferencia en la aleatoriedad de la distribución de ejemplos en D_n sobre las m clases posibles $F_1, \ldots, F_m$ y sobre las p clases provocadas al partir según los valores del atributo $\mathcal{A}_k$.

El atributo seleccionado es el que maximiza dicha ganancia (disminución en la aleatoriedad, aumento de la homogeneidad).

Sin embargo, esta medida está sesgada hacia los atributos con mayor número de valores. Para compensar esto, Quinlan introduce la siguiente modificación: Sea

$$IV(\mathcal{A}_k) = -\sum_{i=1}^{p} \frac{|D_{n,i}|}{|D_n|} \log_2 \left(\frac{|D_{n,i}|}{|D_n|} \right)$$

que es la entropía de la partición de D_n provocada por los p valores del atributo $\mathcal{A}_k$. Entonces $IV(\mathcal{A}_k)$ mide el contenido de información del atributo mismo, y define el *Cociente de Ganancia*

Definición 3.14. ***(Cociente de ganancia)***

$$G_R(\mathcal{A}_k, D_n) = \frac{G(\mathcal{A}_k, D_n)}{IV(\mathcal{A}_k)}$$

G_R es una medida mejor que G, pero tiene las siguientes limitaciones :

- Puede no estar bien definida (el denominador puede ser cero).
- Puede elegir atributos con un muy bajo $IV(\mathcal{A}_k)$ antes que otros con alto $G(\mathcal{A}_k, D_n)$.

Siguiendo [16], sea P_C la partición $\{C_1, \ldots, C_m\}$ del conjunto D_n en sus m clases, y sea P_D la partición $\{D_{n,1}, \ldots, D_{n,p}\}$ generada por los p posibles valores del atributo $\mathcal{A}_k$. La intersección de ambas produce una nueva partición de D_n, para cada etiqueta $i = 1, \ldots, m$ tenemos a los individuos particionados según el valor que presentan en el atributo $\mathcal{A}_k$.

Definamos

$$P_{ij} = P(C_i \cap D_{n,j}) = \frac{|C_i \cap D_{n,j}|}{|D_n|}$$

(proporción de ejemplos con la etiqueta i que toman el valor V_j del atributo $\mathcal{A}_k$). La *entropía* de dicha partición es

$$I(P_C \cap P_D) = -\sum_i \sum_j P_{ij} \log_2(P_{ij})$$

y la *entropía condicional* es

$$I(P_D|P_C) = I(P_D \cap P_C) - I(P_C)$$

Entonces la medida $d(P_C, P_D) = I(P_C|P_D) + I(P_D|P_C)$ es una distancia, y la normalización $d_N(P_C, P_D) = \dfrac{d(P_C, P_D)}{I(P_C \cap P_D)}$ es una distancia en el intervalo $[0, 1]$.

Puede probarse que

$$1 - d_N(P_C, P_D) = \frac{G(\mathcal{A}_k, D_n)}{I(P_D \cap P_C)}$$

de modo que trabajaremos con esta versión del gain-ratio en lugar de la de Quinlan. En este caso, el denominador $I(P_C \cap P_D)$ no puede ser cero si el numerador es diferente de cero.

Regresión

Para el caso de regresión, a cada individuo se le asigna ya sea un valor o una función, y en lugar del gain ratio se utiliza como medida de impureza de un nodo el grado de dispersión de sus individuos respecto de una medida de centramiento.

Si a cada individuo se le asigna un valor (por ejemplo el máximo pico epidémico), si usamos la media como centro y si el nodo contiene individuos

con valores $y_1, \ldots, y_n$, la dispersión se mide como $\sum_{i=1}^{n}(y_i - \overline{y})^2$, donde $\overline{y}$ es el promedio de los n valores y_i.

Si usamos la mediana como centro, la dispersión se medirá como

$$med\{|y_i - \widehat{y}|\}, i = 1 \ldots n$$

donde $\widehat{y}$ es la mediana de los n valores y_i y $med\{|y_i - \widehat{y}|\}$ es la mediana del conjunto de diferencias.

Cuando a cada individuo se le asigna una función en vez de un valor numérico, se puede usar como medida de dispersión $\sum_{i=1}^{n} ||y_i - \overline{y}||^2$, donde $||y||^2 = \int_a^b (y)^2$. Esto es, y es la observación de una respuesta funcional aleatoria Y perteneciente a un espacio de Hilbert con soporte $[a, b]$, $Y \in L^2[a, b]$ y la norma es la inducida por el producto interno habitual

$$< f, g >= \int_a^b f(t)g(t)dt$$

([17])

3.3.4. Bondad del split

Una vez que tenemos el mejor split, éste puede ser aceptado o no. Para decidirlo usamos la medida de impureza llamada *índice de Gini* (si estamos haciendo clasificación): Dado un conjunto de ejemplos (individuos, nodos) t, la impureza o diversidad de clases contenida en t se define como

$$I(t) = \sum_{i,j=1, i \neq j}^{m} Pi|tPj|t = 1 - \sum_{j=1}^{m}(Pj|t)^2$$

($Pi|t$ y $Pj|t$ como en la sección 3.3.3).

Sea s una partición del nodo t en dos nodos t_L, t_R , entonces la diferencia entre las impurezas de t y t_L, t_R se mide mediante

$$\Delta(s, t) = I(t) - p_L I(t_L) - p_R I(t_R)$$

donde p_L, p_R es la proporción de los individuos en t que van a t_L, t_R respectivamente.

Para el caso de regresión, en lugar del índice de Gini se utiliza la devianza, por lo que $I(t), I(t_L), I(t_R)$ seran las devianzas de los nodos t, t_L y t_R respectivamente.

Así, un split s del nodo t es aceptado si $\Delta(s,t)$ es satisfactoria (esto es, mayor que una cierta tolerancia) y las hojas resultantes contienen una cantidad de individuos mayor o igual que un mínimo prefijado.

En el capítulo 4 introduciremos las técnicas y algoritmos utilizados para la generación de grafos aleatorios así como para simular epidemias sobre los mismos.

4. Técnicas utilizadas, descripción

Se generan varios grafos aleatorios, cada uno de ellos con una distribución de grados prefijada (algoritmo grafoaleatoriosparse). Estos grafos serán luego partes de una red (o las ciudades que luego se interconectarán, si pensamos en poblaciones humanas). Nos referiremos a estos grafos o ciudades como componentes cuasi-conexas, por cuanto inicialmente la cantidad de aristas entre componentes es mucho menor que la existente dentro de cada una.

Pensando que cada nodo es un individuo, el mismo tendrá una suerte de legajo donde conste el estado en cada iteración de variables tales como en cual grafo o ciudad se encuentra en ese momento, su estado sanitario, la cantidad de nodos con los que se conecta, etc. Una vez generados los distintos grafos, se procede a modelar la epidemia como sigue :

1. El programa principal (Simulavarios) llama a la subrutina Estocasticoredes que es la que hace cada corrida o simulación. Simulavarios es meramente un ciclo para repetir n veces el programa Estocasticoredes y almacenar los resultados. La subrutina Estocasticoredes a su vez llama a las siguientes subrutinas principales: Iniciales-Cambiasitio-Contagioredes-Remocion, en ese orden. Estocasticoredes es el que hace para cada valor de $t = 1, \ldots, T$ (T es la duración total de la epidemia, fijado por el usuario), los ciclos Inicializar-Viaje-Contagio-Remoción.

2. Para cada t entre 1 y T, dentro del programa principal estocasticoredes hacemos Iniciales − Cambiasitio − Contagioredes − Remocion, y actualizamos las poblaciones de cada sitio, cantidades de infectados y proporciones infectados/susceptibles. La cantidad de ciclos o tiempo (T) es introducida por el usuario, aunque el algoritmo puede terminar antes si en un tiempo dado la cantidad de infectados es cero. A continuación una breve descripción de las subrutinas involucradas:

    a) Iniciales: Crea la configuración inicial: esta configuración es una lista con una fila por nodo (son 10000 nodos, 2000 por cada grafo o ciudad). Dicha lista tiene ocho columnas, las cuales se actualizan para cada t:

        1) La primera tiene el número del nodo en la población total (un número del 1 al 10000).
        2) La segunda su status sanitario (0=susceptible, -1=removido, k=tiempo de vida que le queda como infeccioso).

3) La tercera columna dice en cual red está actualmente (para cada t) el nodo (red 1,2,3,4 o 5).
4) La cuarta columna tiene el número de la red original a la que pertenece el nodo (si pensamos en personas, su ciudad de residencia habitual, el lugar donde vive).
5) La quinta columna tiene el número del nodo en su red natal (su documento de identidad local).
6) La sexta columna tiene el tiempo en que se contagia (lo que permite medir el tiempo transcurrido desde que el nodo se volvió infeccioso).
7) La séptima columna tiene el tiempo de duración de su peródo de contagio.
8) La octava columna el grado del nodo.

Se eligen al azar los individuos que serán los infectados iniciales. A cada uno se le sortea un tiempo de vida como infectado (durante el cual podrá contagiar) con una cierta distribución prefijada de antemano, así como un perfil de infectividad que controla la evolución temporal de la probabilidad de que un contacto en dicho período entre el nodo infectado y otro nodo susceptible resulte en un contagio. Para el tiempo de vida se eligió un valor sorteado de la distribución normal de parámetros μ, σ, por cuanto parece razonable suponer que la mayoría de los infectados debe tener un tiempo de vida similar, con colas simétricas. Para evitar tiempos negativos se toma el máximo entre dicho tiempo y cero.

b) Cambiasitio: En cada ciclo (o sea para cada t entre 1 y T) sortea para cada nodo si permanece en su red (ciudad) o se conecta con otra (viaja a otra ciudad). La probabilidad de dicho evento está dada por una matriz P, donde $P_{i,j}$ es la probabilidad que tiene un nodo de la red i de ser conectado con un nodo cualquiera de la red j. Si pensamos en ciudades e individuos, es la probabilidad de que un habitante de la ciudad i viaje a la ciudad j. Se elige al azar un porcentaje igual a $P_{i,j}$ de nodos de la red i y en la columna correspondiente (la tercera) se le asigna el valor correspondiente de la red a la cual será conectado (o de la ciudad a la cual viajará).

c) Contagioredes: Para cada nodo infectado sortea (usando un proceso de Poisson de intensidad λ, donde λ se elige uniformemente

entre cero y el grado del nodo) a cuantos nodos contactará durante su tiempo de vida como infeccioso. Sea n_i el resultado de dicho sorteo. Si el nodo está en su red original, se eligen n_i de sus vecinos, si está conectado a otra red que no es la suya original, el sorteo se hace entre todos los nodos de la red donde se encuentra. Si el nodo contactado es susceptible, se sortea si será contagiado o no. La probabilidad depende del perfil de infectividad del nodo infectado. Este perfil se le asigna a cada nodo cuando es contagiado. Esto es, junto con el tiempo que durará su estado como infectado se le asigna una función "bump"o meseta, de modo que al comienzo la probabilidad de contagiar es baja, luego va creciendo y sobre el final de su período como infectado decrece nuevamente (figura 5). Recordemos que en la lista de todos los nodos, la sexta columna tiene el tiempo transcurrido desde que el nodo se infectó, lo que nos permite saber en que parte del perfil de infectividad estamos y por tanto ajustar la probabilidad de que dicho nodo contagie a sus vecinos.

d) Remocion: Al final de cada ciclo actualiza (bajando en una unidad) el tiempo remanente de los nodos infectados, cambiando su status a removidos cuando dicho tiempo expira.

3. La salida del programa principal son gráficas mostrando la evolución temporal del número de nodos infectados en cada red, la contribución de cada red en la epidemia total de la mega-red, y la distribución de grados de los nodos afectados y de los no afectados.

Así, cada simulación de una epidemia puede verse como la realización de un proceso estocástico que origina una trayectoria o un vector de trayectorias por cada red. El algoritmo permite asimismo simular epidemias sobre un solo grafo sustituyendo la matriz de probabilidades de viaje por la matriz identidad. Esto es, la probabilidad de viajar es cero.

Para analizar el algoritmo, simulamos cien veces la misma epidemia: esto es, sobre el mismo grafo (si simulamos la epidemia en un grafo aislado) o conjunto de grafos (si simulamos la epidemia en un conjunto de grafos interconectados) y con los mismos parámetros corremos 100 veces la simulación.

Cuando tenemos un solo grafo, tenemos entonces un conjunto de datos con los resultados de las cien corridas. Cada dato contiene la evolución temporal

del número de casos, la distribución del grado de los nodos afectados y una etiqueta asociada al grado de virulencia de la epidemia.

Cuando tenemos un conjunto de grafos interconectados con cierta probabilidad, tendremos un conjunto de cien datos donde ahora aparece para cada grafo la información anteriormente citada, así como la evolución temporal del porcentaje que los infectados de cada grafo representan sobre el total de infectados en dichos grafos.

Utilizando algoritmos CART (Classification And Regression Trees) para datos con atributos funcionales (en este caso concreto, la evolución temporal de la cantidad de casos para cada sitio, así como las distribuciones de frecuencias de grados de los nodos infectados y los no infectados), se construyen árboles de clasificación y regresión.

Dichos algoritmos fueron programados en lenguaje R junto con Ariel Roche bajo supervisión de Badih Ghattas en el marco del curso de maestría "Machine Learning y sus Aplicaciones "por él dictado en la Facultad de Ingeniería, UDELAR en el segundo semestre de 2007.

Usando parte de los datos como muestra de entrenamiento y otra parte como muestra de testeo, se pueden construir árboles de clasificación para agrupar las epidemias ocurridas sobre el mismo sustrato (red o conjunto de redes) en clases (y según sea el sustrato y los parámetros investigar cuántas clases surgen). Asimismo se pueden construir árboles de regresión a partir de los cuales construír epidemias "típicas"o "epidemias medias".

En el caso de poblaciones humanas, esto plantea la posibilidad de ajustar el modelo a la red de ciudades y pueblos. Si se logra capturar la red social típica de nuestras ciudades y pueblos (lo que implica contar con datos que deben ser fruto de un trabajo interdisciplinario), se pueden generar grafos aleatorios con la estructura de dichas redes.

Usando parámetros conocidos sobre la enfermedad bajo estudio se podrá simular una epidemia típica, así como los escenarios o realizaciones de menor probabilidad, investigar estrategias de control, etc.

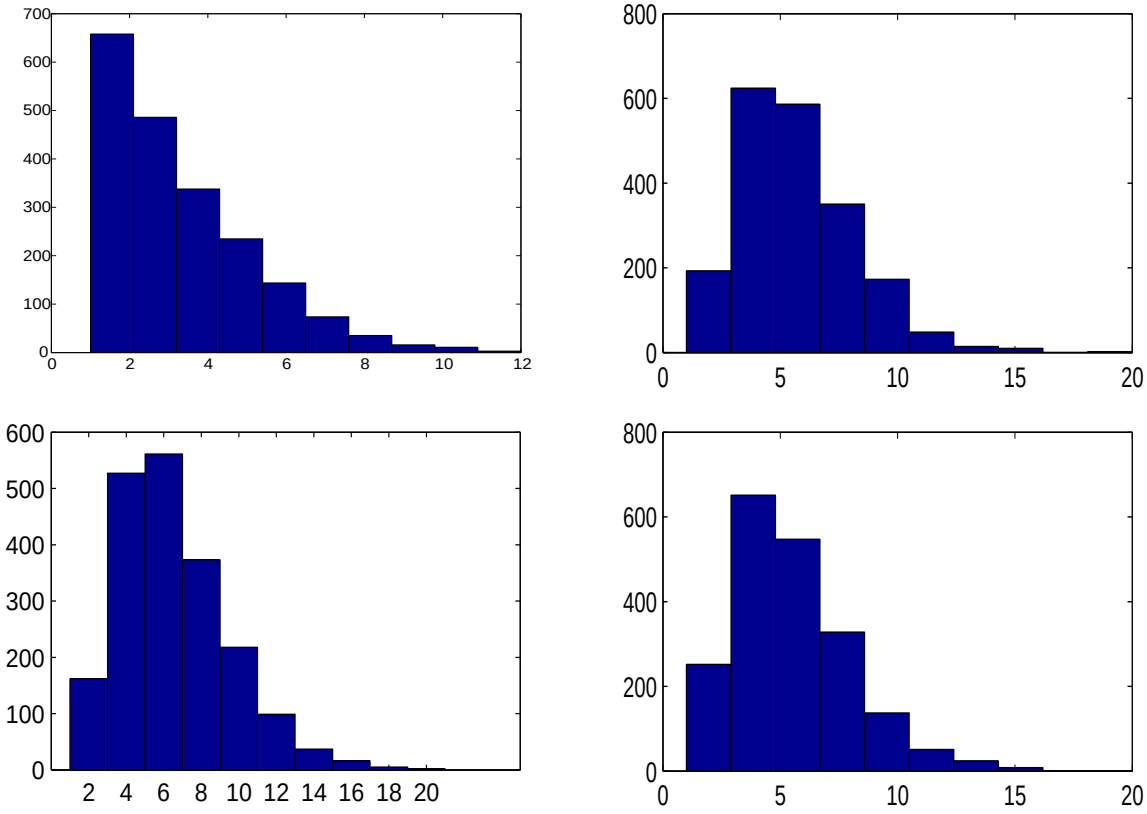
En el capítulo 5 daremos una descripción general de los algoritmos utilizados para generar los grafos aleatorios, así como la matemática subyacente.

5. Generación de casos de prueba

5.1. Comentarios generales

Se generaron varias secuencias de grados distintas de 2000 elementos. Para tener una distribución de frecuencias con unos pocos vértices de grado alto y que no sea puramente potencial o exponencial, se usó la distribución Gamma. En la Figura 3 se muestran histogramas para secuencias de 2000 nodos, con el grado (eje horizontal) versus la cantidad de nodos (eje vertical).

Figura 3: Histogramas para secuencias de 2000 nodos



5.2. Algoritmos

Definición 5.1. ***(Secuencia gráfica)***

Dada una secuencia $\{d_1, \ldots, d_p\}$ de enteros no negativos, diremos que la misma es gráfica cuando existe algún grafo $G = (V, E)$, con $|V| = p$ y nodos $\{v_1, \ldots, v_p\}$ tal que grado$(v_i) = p_i \; \forall i = 1 \ldots p$

Una condición necesaria y suficiente para que una tal secuencia sea gráfica es la de Havel-Hakimi:

Teorema 5.2. ***(Havel-Hakimi)***

Una secuencia $\{d_1, \ldots, d_p\}$ de enteros no negativos con $d_1 \geq d_2 \geq \cdots \geq d_p$, $p \geq 2$, $d_1 \geq 1$, es gráfica si y solo si la secuencia $s_1 = \{d_2 - 1, d_3 - 1, \ldots, d_{d_1+1} - 1, d_{d_1+2}, \ldots, d_p\}$ es gráfica. [7]

Basado en el teorema se programó un algoritmo (Dictamen) que permite determinar cuando una secuencia de enteros no negativos es gráfica. ([7]).

Algoritmo "Dictamen"

Dada una secuencia de enteros no negativos $d_1, \ldots, d_p$, $p \geq 1$:

1. Si alguno de los enteros de la secuencia es mayor a $p - 1$ o negativo, o si la secuencia suma un número impar, la secuencia no es gráfica. Sino, siga al paso 2.

2. Si todos los enteros de la secuencia son cero, la secuencia es gráfica. Sino, siga al paso 3.

3. Si hace falta, reordene la secuecia de mayor a menor y siga al paso 4.

4. Elimine el primer número (digamos n) de la secuencia y reste 1 a los siguientes n números de la secuencia. Vuelva al paso 2.

Una vez que disponemos de una secuencia de grados gráfica, el siguiente paso es crear un grafo aleatorio cuyos vértices tengan dicha secuencia de grados. Para ello nos basamos en ([4]).

Supongamos que tenemos una secuencia $d = \{d_1, d_2, \ldots, d_p\}$ de enteros no negativos que es gráfica, con $\sum_{i=1}^{p} d_i = 2m$ y un conjunto de vértices $\{v_1, \ldots, v_p\}$. Entonces existe al menos un grafo simple con esos grados.

Algoritmo "Grafoaleatoriosparse"

1. Sea E el conjunto de aristas, $\hat{d} = \{\hat{d}_1, \ldots, \hat{d}_p\}$ una secuencia de enteros. Inicializamos $E = \Phi$, $\hat{d} = d$

2. Elijamos dos vértices $v_i, v_j \in V$ con probabilidad proporcional a

$$\hat{d}_i \hat{d}_j (1 - \frac{\hat{d}_i \hat{d}_j}{4m})$$

   entre todos los pares i, j con $i \neq j$ y $(v_i, v_j) \notin E$. Reduzcamos $\hat{d}_i$ y $\hat{d}_j$ en una unidad.

3. Repitamos el paso anterior hasta que no se puedan añadir más aristas a E.

4. Si $|E| < m$ reportar error y comezar de nuevo. Sino la salida es el grafo $G = (V, E)$

Este algoritmo genera grafos aleatorios con probabilidad uniforme (esto es, de todos los grafos aleatorios que tienen dicha secuencia de grados, produce uno que es resultado de un sorteo donde todos tienen la misma probabilidad).

Con dichas secuencias se construyeron grafos aleatorios que las realizan. Si pensamos en los nodos como individuos de una población, entonces el grafo aleatorio representa la red de contactos de dicha población. Cada grafo puede pensarse como una entidad (una ciudad, si pensamos en poblaciones humanas). Generamos así varios grafos de 2000 nodos.

Estos grafos luego se conectan aleatoriamente. Esto es : para cada tiempo t, cada vértice o habitante tiene una cierta probabilidad de conectarse con vértices de otro grafo (modelando el viajar a otra ciudad).

Dicha conexión se hace al azar, según sea la probabilidad de viajar de una ciudad a otra. Los grafos originales pueden pensarse como componentes cuasi-conexas del mega-grafo o red de ciudades y pueblos así formado.

En el capítulo 6 simularemos epidemias con distintos grados de virulencia en tres grafos aleatorios distintos y luego pondremos a prueba la idea de usar árboles de clasificación y regresión para agrupar las distintas epidemias según sus curvas de cantidad de casos en función del tiempo y de distribución de frecuencias de los grados de los nodos afectados. Asimismo probaremos distintas estrategias de inmunización o vacunación.

6. Casos construídos

6.1. Repaso

Recordemos como trabajan los árboles, tanto en clasificación como en regresión (sección 3.3.1): Cada realización (corrida o simulación de una epidemia sobre un grafo) tiene dos atributos : la gráfica de la distribución de frecuencias de los grados de los nodos afectados (para cada grado nos dice que porcentaje del total de nodos afectados representan los nodos con dicho grado), y la curva con la evolución temporal del número de afectados o casos (atributos 1 y 2 respectivamente).

Si simulamos 100 veces una epidemia sobre el mismo grafo, tendremos 100 realizaciones que conformarán el nodo raíz del árbol. Luego se prueban distintas particiones de dicho nodo: para cada atributo de los dos ya mencionados se comparan estas realizaciones mediante la DTW (sección 3.3.2)). Aquellas realizaciones con gráficas similares del atributo en cuestión quedarán en el mismo subconjunto de la partición (irán para el mismo lado, dado que cada partición es binaria).

Se elige la mejor partición (la que parte al nodo padre en dos nodos hijos de modo tal que los hijos sean lo más homogéneos o puros posible). Para los árboles de clasificación se asigna a cada corrida o realización la etiqueta con la pareja (μ, σ) correspondiente (los parámetros de la distribución normal usada para sortear los tiempos de vida como infecciosos, que dan una idea de la virulencia de la epidemia). En este contexto, donde cada realización tiene una etiqueta o tipo, un nodo puro será aquel que solamente tenga realizaciones del mismo tipo o etiqueta.

En regresión no hay etiquetas, sino que a cada realización se le asocia un valor y (por ejemplo el máximo valor del atributo 2, que es la función n^o de casos vs. tiempo). Este máximo se usa como característica de cada corrida para implementar árboles de regresión sobre dichos datos. Puede decirse que y es una etiqueta, la diferencia con la clasificación es que no tenemos muchas realizaciones y pocas etiquetas, sino que cada uno tiene su propia etiqueta, es una clase en sí mismo. Por tanto la pureza de cada nodo se mide según el grado de dispersión de dichos valores respecto de la media para el nodo (ver sección 3.3.3).

Se repite el proceso de partición para cada nodo obtenido hasta que o bien se llega a un nodo puro (si estamos haciendo clasificación), o se llega a un nodo cuyo grado de pureza es aceptable (ya no ganamos gran cosa en

homogeneidad si lo seguimos partiendo), o se llega a un nodo con demasiado pocas realizaciones.

Una vez que tenemos el árbol hecho, cada nueva realización es asignada a la hoja donde se encuentran las realizaciones que más se le parecen. Si estamos haciendo clasificación, a dicha realización se le asignará la etiqueta o clase correspondiente a la hoja a la que fue asignada. Si se trata de regresión se le asigna como valor y el promedio de los valores de las realizaciones que forman dicha hoja.

6.2. Simulaciones en grafos aleatorios de 2000 nodos

Trabajaremos con tres grafos aleatorios de 2000 nodos en dos escenarios o paisajes :

1. Grafos aislados (sin interconexiones entre ellos).

2. Cinco grafos de 2000 nodos cada uno con interconexiones aleatorias entre ellos.

Sobre cada uno de estos escenarios se simularán epidemias variando el lugar de aparición de los infectados iniciales, así como el tiempo medio de duración del período durante el cual cada nodo infectado puede contagiar.

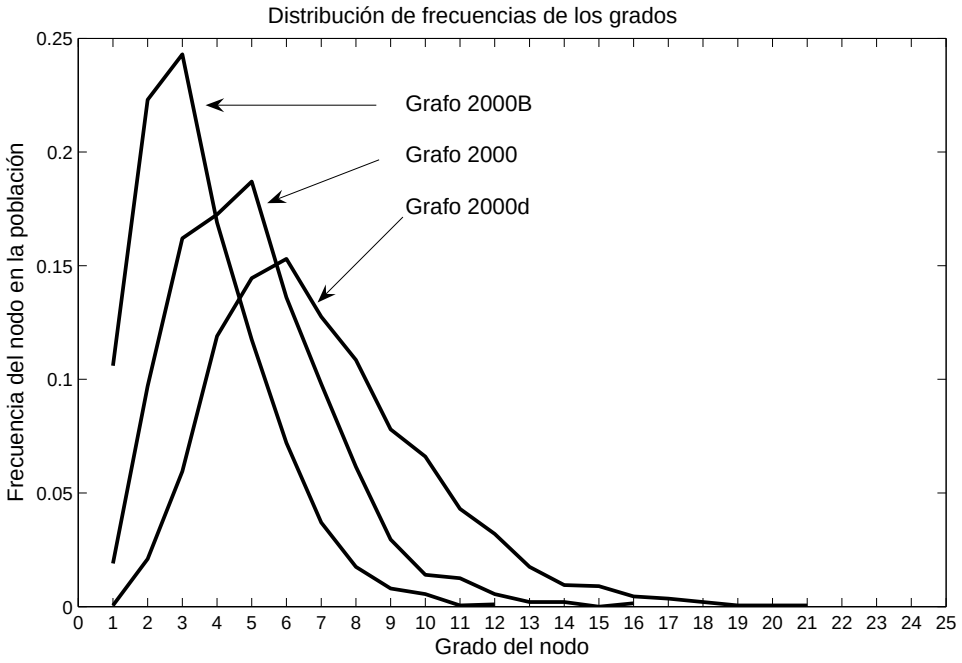
Se construyeron tres grafos aleatorios de 2000 nodos tomando las secuencias de grados cuya distribución de frecuencias se muestran en la Figura 4.

Luego se realizaron simulaciones de epidemias de distinto grado de virulencia sobre los mismos. Dicho grado de virulencia está dado por la duración del período de infectividad para cada individuo (nodo), la cual fue sorteada usando la distribución normal de parámetros (μ, σ), con $\mu = 2, 3, 5, 10, 20$ y tomando el máximo entre dicho valor y cero.

La cantidad de contactos que un individuo o nodo tendrá durante dicho lapso se sortea usando la distribución Poisson de media λ, donde λ se elige con probabiliad uniforme discreta en el número de sus vecinos. Una vez elegidos (μ, σ), se realizan 100 simulaciones de la epidemia, en cada una de ellas el tiempo durante el cual un individuo infectado es capaz de contagiar es una variable aleatoria cuya distribución es normal de parámetros (μ, σ).

La capacidad de contagiar durante dicho período no tiene porqué ser constante, por lo que a cada individuo que se contagia se le asigna un potencial infectivo. Si por ejemplo un individuo permanece infectado entre $t = a$ y

Figura 4:



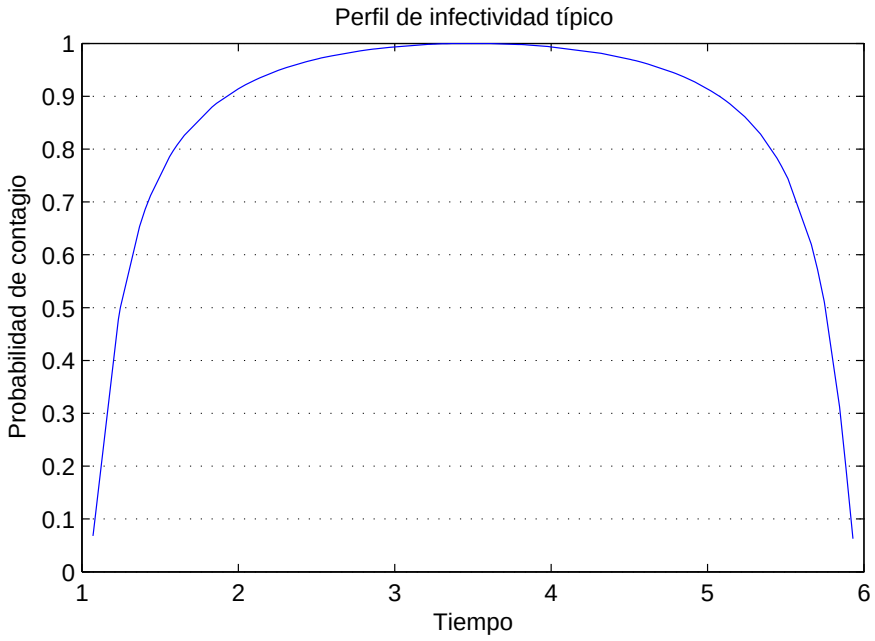
$t = b$, su potencial infectivo estará dado por $f(t) = e^{-\frac{1}{(b-t)(t-a)}}$. Esta es una función que crece, luego tiene una meseta y por último decrece. Se muestra un ejemplo en la Figura 5, con $a = 1,\ b = 6$.

Se eligió un conjunto de parámetros con $\mu = 2, 3, 5, 10, 20$ y $\sigma = 1$ ($\sigma = 5$ en un solo caso), generando las etiquetas $\{(2,1),(3,1),(5,1),(10,1),(10,5),(20,1)\}$. Esto es, a cada simulación de una epidemia sobre un grafo donde los lapsos de tiempo en que cada infectado puede contagiar se sortean de la distribución normal de parámetros (μ, σ), se le asigna la etiqueta (μ, σ). El tiempo total en cada corrida se fijó en 300 iteraciones y se tomaron 100 corridas de los tipos $(2,1),(3,1),(5,1),(10,1),(20,1)$.

Comentarios

En cada corrida se guarda la distribución de frecuencias de los grados de los nodos afectados, la cantidad de nodos infectados en función del tiempo, y la etiqueta $((\mu, \sigma)$ si vamos a hacer clasificación o el máximo pico epidémico

Figura 5:



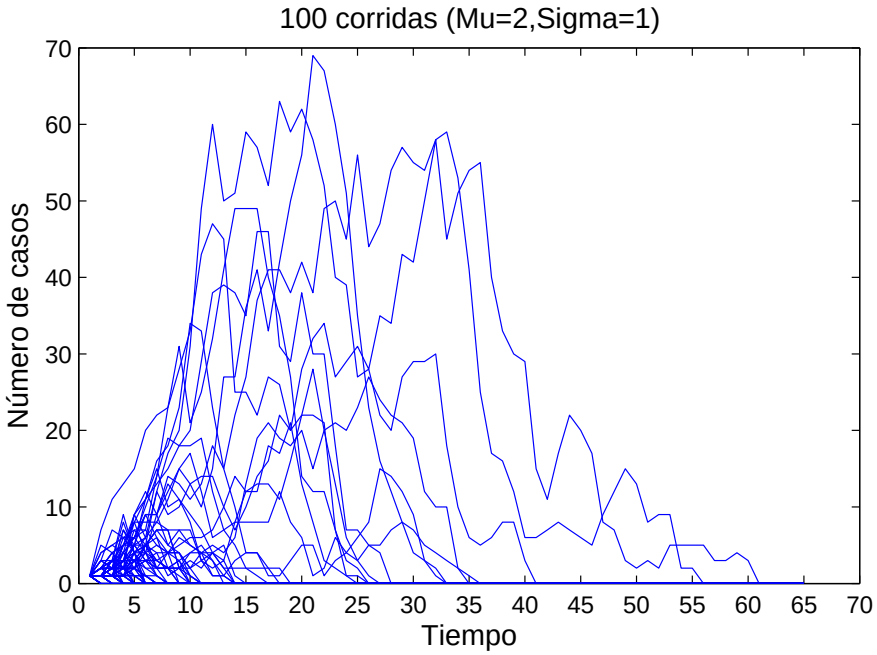
de la anterior función si vamos a hacer regresión). De este modo por ejemplo para el grafo 2000B cada corrida se guarda como un vector con 12+300+1 lugares (12 porque es el grado máximo del grafo, 300 por las iteraciones y 1 por la etiqueta).

A continuación mostramos los resultados para el grafo "2000B"

- $\mu = 2,\ \sigma = 1$. Se hicieron 100 corridas. El número de iteraciones (tiempo) es de 300, pero todas las trayectorias se anulan antes de t=65. Si por epidemia entendemos la afectación de una fracción apreciable de la población, podríamos decir que no llega a darse ninguna epidemia. Como puede apreciarse en la Figura 6, las trayectorias son altamente disímiles entre sí, y en la mayoría de las corridas se extinguen antes de t=15, como muestra la Figura 7.

- $\mu = 3,\ \sigma = 1$. El comportamiento es similar al caso anterior (Figura 8), aunque ahora la proporción de corridas donde ocurren epidemias es mayor. Se muestran la mediana y la media. Observamos que esta

Figura 6:



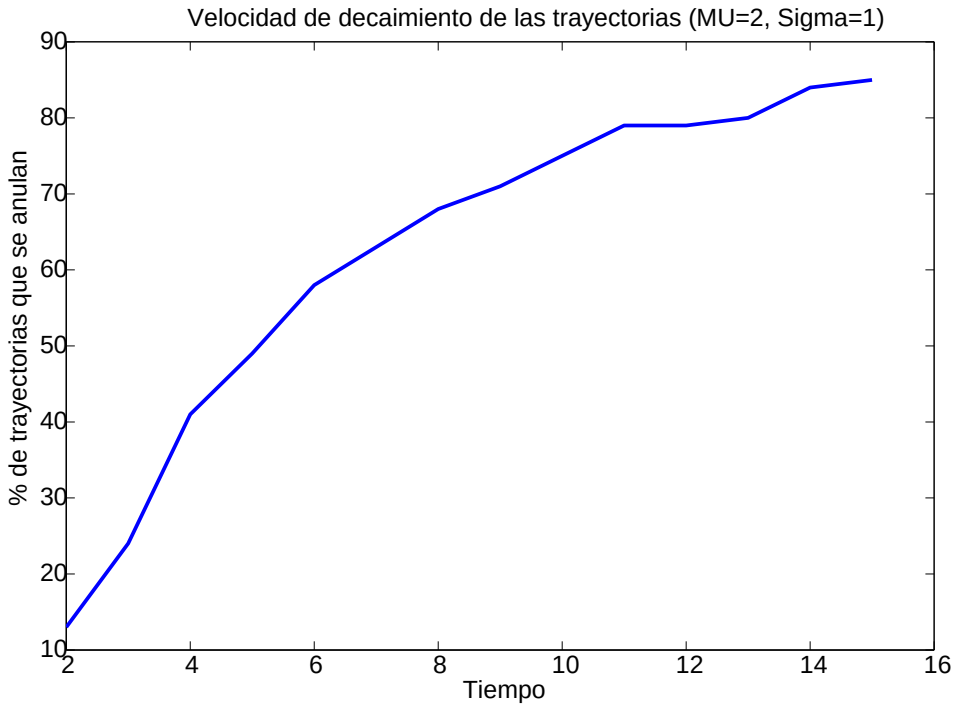
última sigue siendo fuertemente afectada por aquellas corridas donde no hubo epidemia.

- $\mu = 5,\ \sigma = 1$. El número de iteraciones (tiempo) se fijó en 50, dado que en cada corrida la epidemia concluye entre t=35 y t=40 (fig.9). Aunque para este caso se hicieron 200 corridas, luego se eligieron 100. Notemos que las trayectorias son mucho más regulares que en el caso anterior, aunque parecen agruparse en dos conjuntos. En la Figura10 mostramos la epidemia "media"junto con las bandas de confianza al 95 %.

- $\mu = 10,\ \sigma = 1$. También en este caso las trayectorias son muy similares entre sí, formando un único grupo. Todas las trayectorias se anulan antes de t=120 como puede verse en la Figura 11.

- $\mu = 10,\ \sigma = 5$. De nuevo las trayectorias son regulares, como puede apreciarse en la Figura 12. Al ser mayor la dispersión σ en los tiempos

Figura 7:

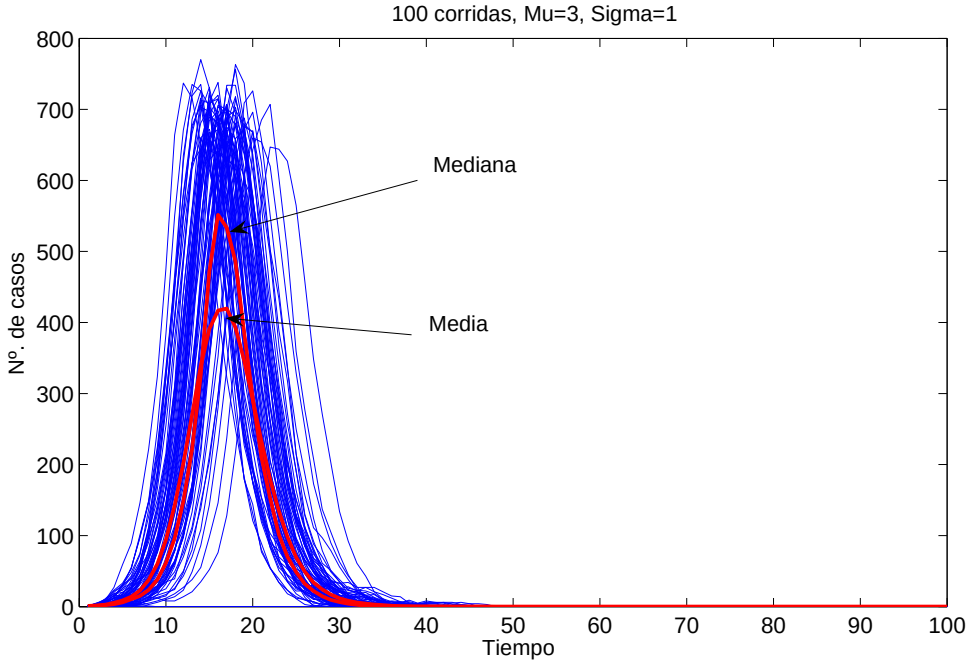


de vida sorteados a los nodos al infectarse, tenemos tiempos de vida más variados entre los nodos infectados, pero el resultado medio no presenta cambios sustanciales respecto del caso anterior.

- $\mu = 20,\ \sigma = 1$. Se observa un comportamiento similar al caso anterior (Figura 13). Los tiempos de extinción son mayores y comienzan a aparecer picos de menor cuantía sobre el final, de un modo similar a lo que ocurre en el modelo SIR con demografía ([13]).Se nota una tendencia creciente en el valor de los picos epidémicos al aumentar μ (Figura 14). Esto es razonable, por cuanto el infeccioso promedio tiene un período infectivo de duración cada vez mayor, lo que hace más probable que pueda contagiar una mayor cantidad de sus vecinos.

- $\mu = 20,\ \sigma = 5$ Similar al caso anterior, comentarios similares al caso

Figura 8:



(10,5).

- En la Figura 15 se muestran las epidemias medias para los casos anteriores con $\sigma = 1$.

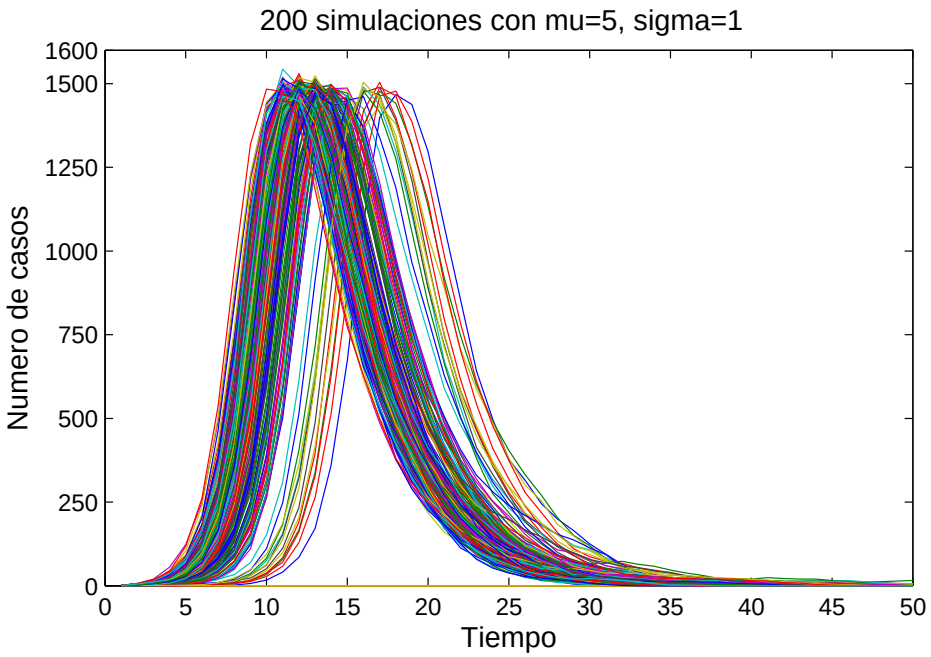
Observaciones

1. Al aumentar la duración del período durante el cual el individuo puede contagiar, las trayectorias se van haciendo más regulares, la variabilidad disminuye.

2. Los picos epidémicos de las epidemias medias van aumentando en tamaño.

3. En tiempos avanzados aparecen brotes secundarios. Al no haber demografía, una explicación es la siguiente: Luego del crecimiento explosivo

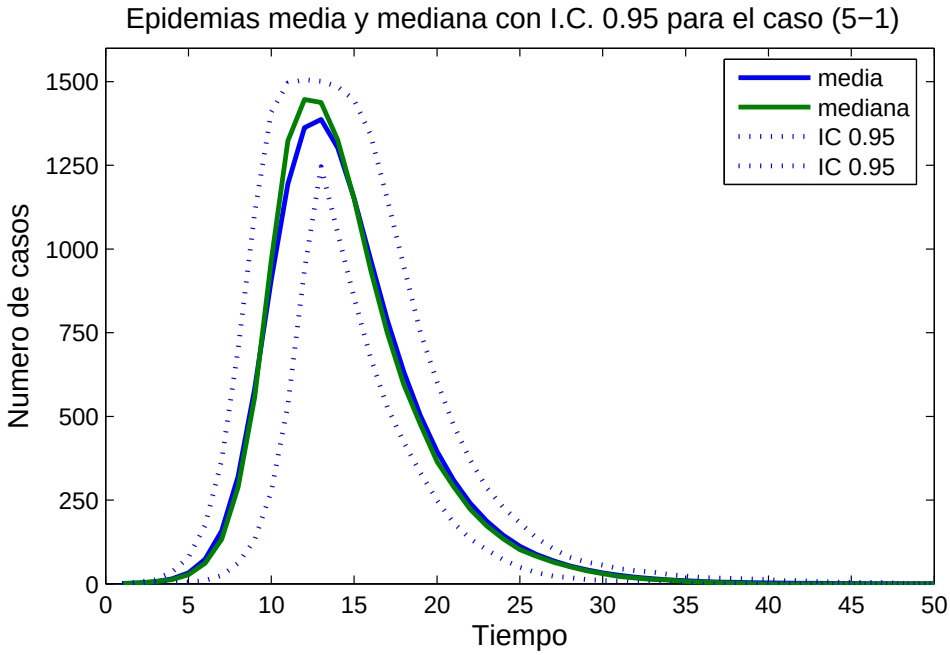
Figura 9:



inicial, se llega a una saturación, en el sentido que cada individuo infectado contactará mayormente realizaciones también infectados o removidos. Sólo unos pocos podrán contactar los pocos individuos susceptibles remanentes, dando lugar a un pequeño brote epidémico sobre los nodos restantes.

4. El caso $(\mu, \sigma) = (10, 5)$ se descartó por cuanto no presentaba diferencias relevantes respecto del caso $(10, 1)$. Asimismo, todas las gráficas de todas las corridas de todos los casos fueron recortadas a $t = 100$ para acortar los tiempos de cómputo. Puede sospecharse que esto vaya en detrimento de la capacidad de distinguir entre los casos $(10, 1)$ y $(20, 1)$, que son los que tienen tiempo de extinción superior a $t = 100$. Esto es, como vamos a comparar similitudes entre gráficas muchas de las cuales valen cero antes que otras, al remover las colas de aquellas gráficas que tardan más en anularse estamos dejando fuera algo que puede contribuír

Figura 10:



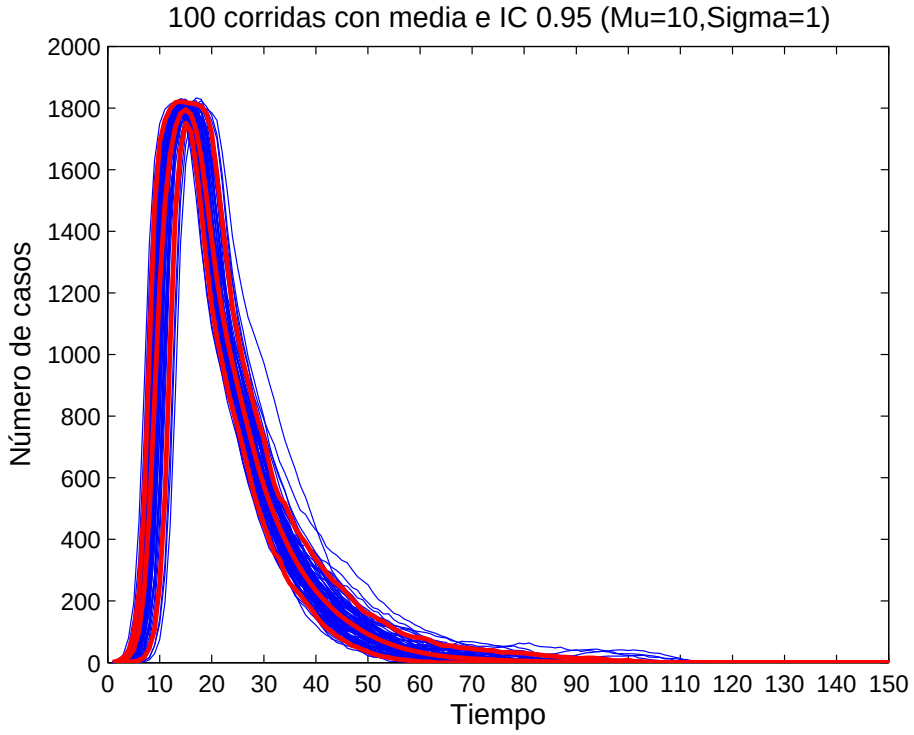
a diferenciarlas de las otras. Sin embargo veremos que ello no sucede.

6.2.1. Clasificación

La idea central de asociar a cada corrida o realización del proceso estocástico "epidemia"una salida susceptible de ser tratada con técnicas de machine learning, se somete a prueba usando árboles de clasificación (sección 6.1). Para ello se realizan 500 corridas (100 para cada una de las parejas de parámetros (μ, σ) citados) en dos grafos aleatorios distintos (2000B y 2000d).

A cada epidemia se la etiqueta según la pareja (μ, σ), teniendo cinco tipos: (2,1),(3,1),(5,1),(10,1) y (20,1). Podemos pensar a dicha epidemia como un punto de un cierto espacio S, al que asociamos dos funciones (la que describe la cantidad de infectados por unidad de tiempo, y la distribución de frecuencias de los grados de los nodos infectados), estamos asignando a

Figura 11:



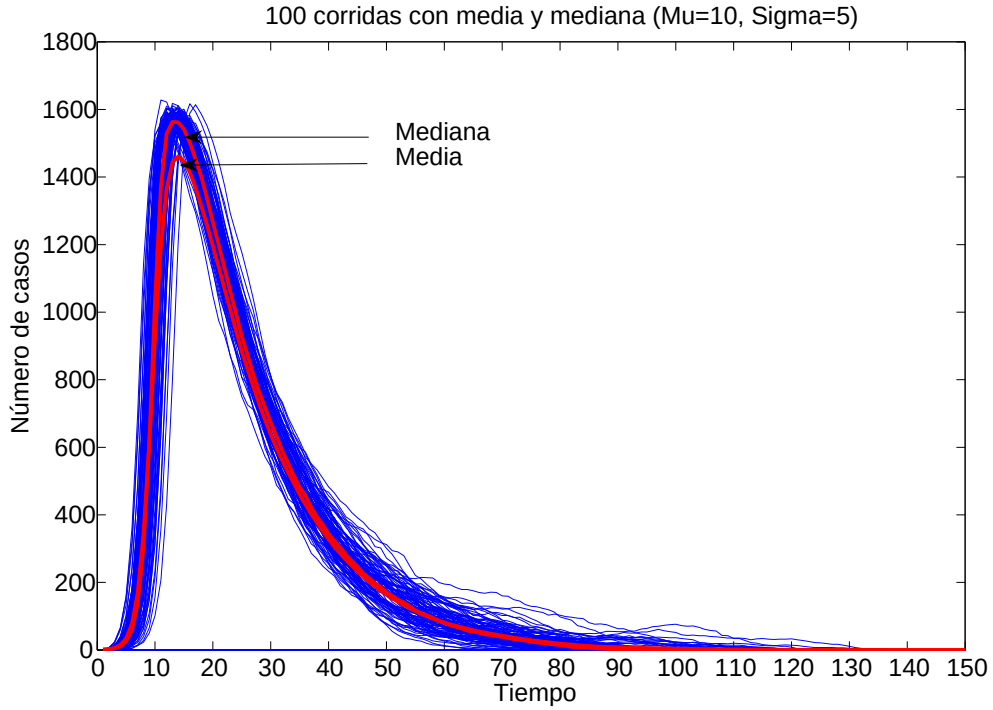
$s \in S$ un vector $(s_1, s_2) \in \mathbb{R}^d \times \mathbb{R}^g$, donde $d = 100$ es la longitud del atributo 2 (curva cantidad de infectados vs. tiempo) y $g = \{12, 16, 21\}$ (según sea el grafo en el cual simulamos la epidemia) es la longitud del atributo 1 (grados de los nodos afectados).

Entonces el clasificador (árbol) induce una partición en $\mathbb{R}^d \times \mathbb{R}^g$, la que a su vez induce una partición en S, de modo que dos realizaciones del proceso llamado epidemia irán a la misma clase o subconjunto de S si sus vectores (s_1, s_2) asociados van a la misma clase en $\mathbb{R}^d \times \mathbb{R}^g$.

Para el grafo 2000B, se usó la técnica de validación cruzada (ten-fold cross-validation): se parte la población en 10 muestras de aproximadamente el mismo tamaño, se construye un árbol de clasificación con 9/10 de la población y se testea con el décimo restante.

El error de clasificación cometido (esto es, se etiqueta erróneamente a la

Figura 12:



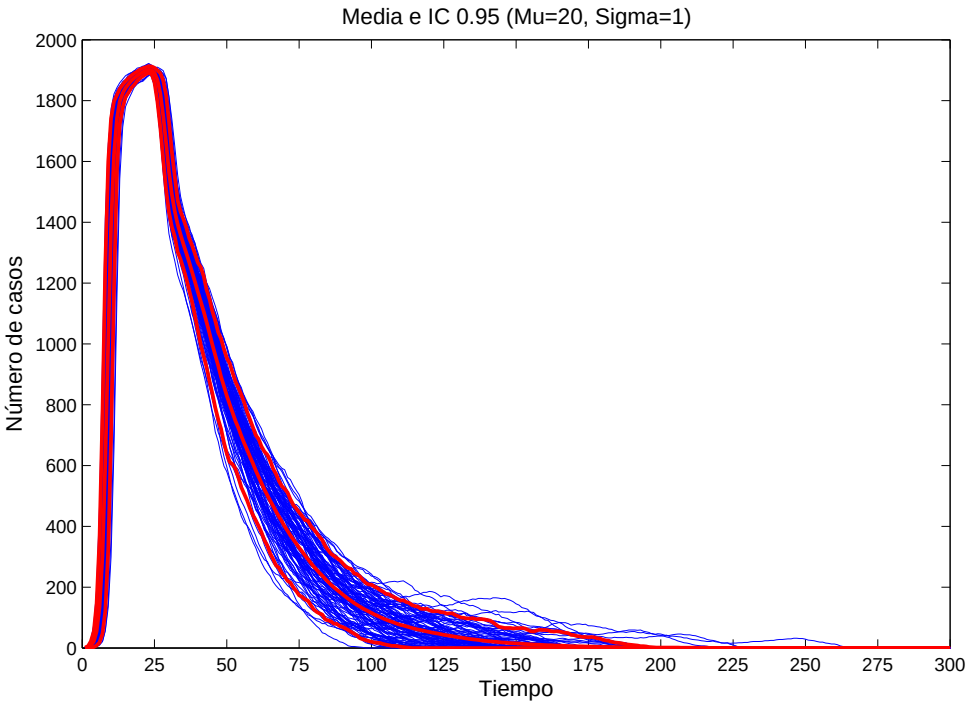
realización asignándole una etiqueta que no es la suya) se guarda. Se repite esto 10 veces variando el décimo usado para testear y se promedian los errores de clasificación cometidos cada vez.

Para el parámetro hojamin = 4 (recordemos que hojamin es la cantidad mínima de realizaciones que puede tener una hoja, por lo que un nodo cuya partición provoque hijos con menos realizaciones que hojamin, no será partido, sin importar cuán impuro pueda ser), el error de clasificación medio es de 0.076, con un desvío estándar de 0.029.

Como la muestra L es grande (aproximadamente 450 realizaciones), probamos con hojamin = 50 y hacemos nuevamente crossvalidation. En este caso el error de clasificación medio es de 0.07 y el desvío estándar es 0.034.

Para evaluar la performance con muestras de menor tamaño, se eligen al azar 85 realizaciones de entre la población de 500. Se extraen 75 de ellas al

Figura 13:



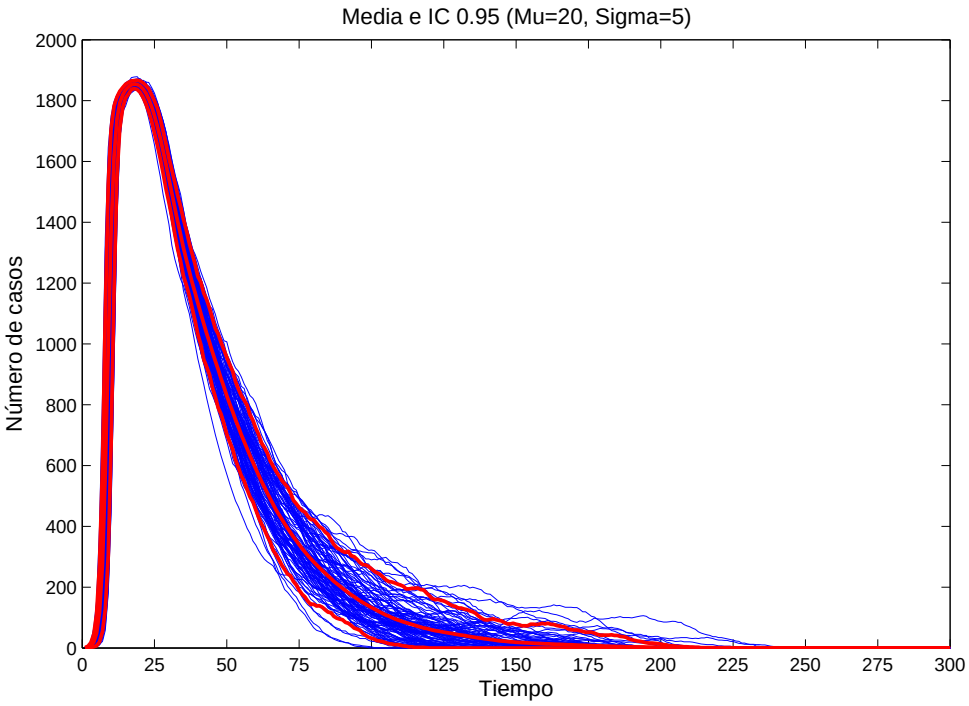
azar para conformar la muestra de entrenamiento L, las 10 restantes serán utilizados como muestra test T.

Al hacer un árbol de clasificación sobre L (la cantidad mínima de realizaciones de un nodo se fija en 4), obtenemos que el algoritmo agrupa correctamente las epidemias similares. Los casos más borrosos como era de esperar son las epidemias más suaves, de parámetros $(2, 1)$ y $(3, 1)$, cuyas descriptores son muy similares, con una proporción menor de epidemias de parámetros $(5, 1)$ (éstas forman una suerte de grupo de transición entre los tipos $(2, 1), (3, 1)$ y los tipos $(10, 1), (20, 1)$).

Para medir el error de clasificación se utilizaron tres criterios distintos:

1. "Error duro": Si la realización recibe una etiqueta distinta a la de su tipo real, se contabiliza un error.

Figura 14:



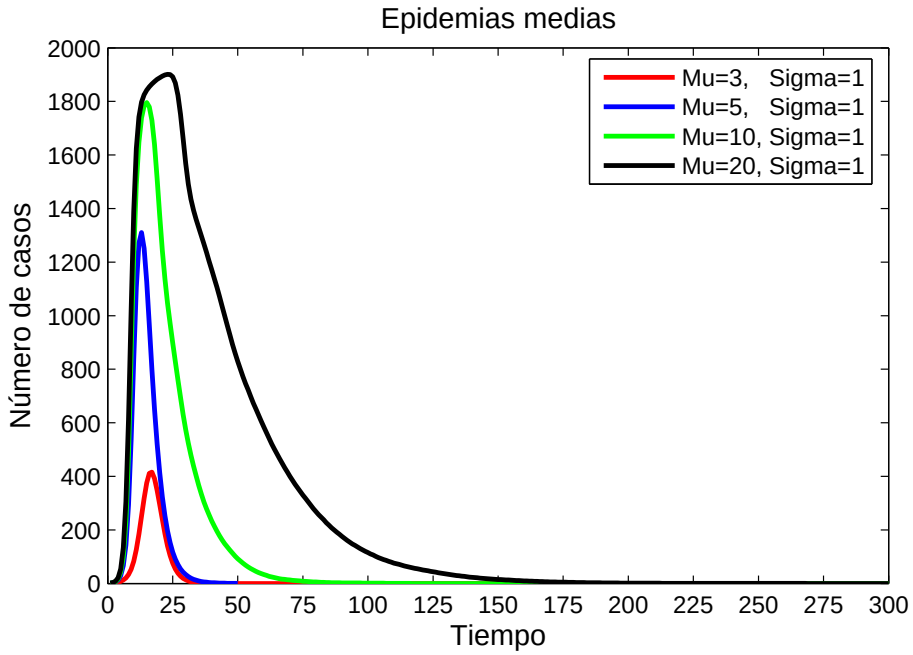
2. "Error ponderado": Cuando el árbol de clasificación tiene hojas que no son puras, por ejemplo un 49 % de realizaciones del tipo $(2, 1)$, el resto $(3, 1)$, puede suceder que un dato (realización) del tipo $(2, 1)$ sea asignado a dicha hoja, con lo cual será clasificado como $(3, 1)$. En vez de incrementar el contador de errores en una unidad, lo incrementamos en 0.51 unidades, por cuanto el algoritmo no estuvo tan errado.

3. "Error suave" En este caso se cuenta un error solamente cuando la realización va a dar a una hoja donde no hay nadie de su mismo tipo.

A continuación los resultados sobre los datos reservados para testear para los grafos 2000B y 2000d.

Grafo 2000B : Por ejemplo, el árbol construído con la muestra L (Figura 16, parámetro hojamin = 4) tiene 6 hojas, 5 de las cuales son puras. Excepto

Figura 15:



en un nodo, siempre se utiliza el atributo 1 para partir. La raíz se divide en los nodos 2 y 3.

El nodo 2 se divide en la hoja 4 (pura, 15 realizaciones del tipo $(5, 1)$ y el nodo 5, el cual tiene mezcla de realizaciones $(10, 1)$ y $(20, 1)$.

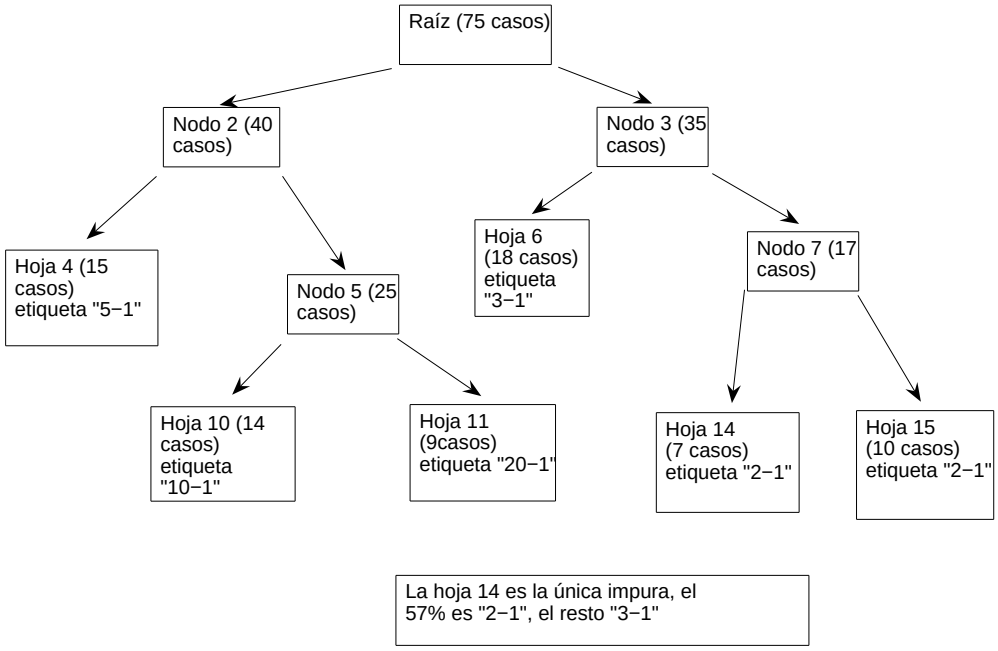
El nodo 3 se divide en la hoja 6 (pura, 18 realizaciones del tipo $(3, 1)$) y el nodo 7, el cual tiene mezcla de realizaciones $(2, 1)$ y $(3, 1)$.

El nodo 5 se divide en las hojas 10 (pura, 14 realizaciones del tipo $(10, 1)$) y 11 (pura, 9 realizaciones del tipo $(20, 1)$).

El nodo 7 se divide en las hojas 14 (7 realizaciones, 57.1 % tipo $(2, 1)$, el resto $(3, 1)$) y 15 (pura, 10 realizaciones del tipo $(2, 1)$).

De los 10 datos de la muestra de testeo T, 9 de ellos son correctamente clasificados. El único error es en un individuo del tipo $(2, 1)$, que es correctamente etiquetado en la hoja 14 como $(2, 1)$, pero como dicha hoja tiene un 42.9 % de casos $(3, 1)$, el contador de errores ponderados se incrementa en 0.429. El saldo final es: error duro 0, error ponderado 0.0429, error suave 0. Otras corridas muestran performances similares.

Figura 16:



Grafo 2000d: Construímos un árbol (Figura 17) con la correspondiente muestra L con 75 realizaciones y hojamin = 4. Dicho árbol tiene 6 hojas, tres de las cuales son puras, otra casi pura y las dos restantes con dos poblaciones en proporciones 60/40. El atributo elegido para partir siempre es el 2 (gráfica cantidad de casos vs. tiempo).

La raíz se divide en los nodos 2 (33 realizaciones, impureza=0.48) y 3 (42 realizaciones, impureza= 0.53).

El nodo 2 se divide en los nodos 4 (hoja con 20 realizaciones, 95 % tipo $(2, 1)$ y el resto $(3, 1)$) y 5 (hoja pura con 13 realizaciones tipo $(3, 1)$).

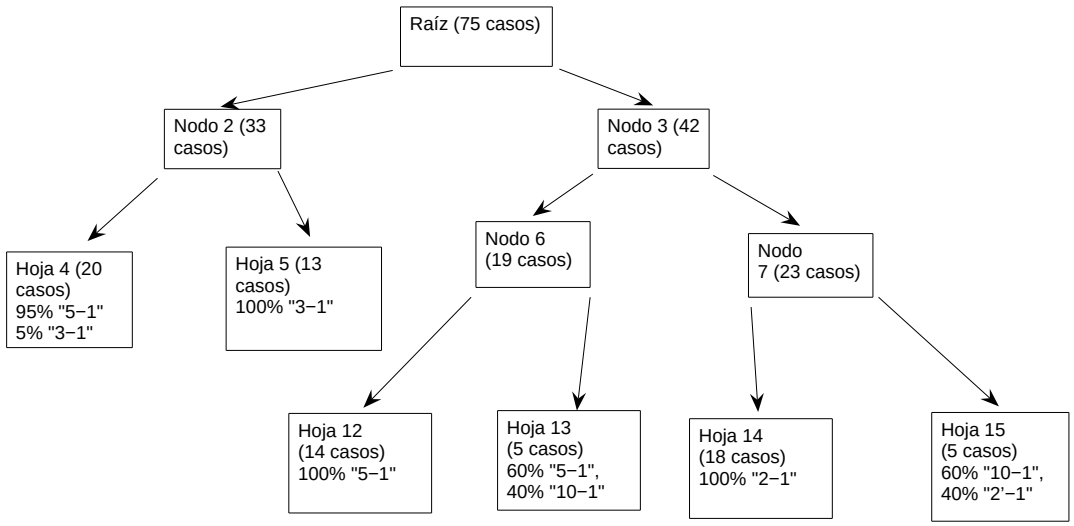
El nodo 3 se divide en los nodos 6 (19 realizaciones, impureza=0.18) y 7 (23 realizaciones, impureza=0.15).

El nodo 6 se divide en los nodos 12 (hoja pura con 14 realizaciones tipo $(5, 1)$) y 13 (hoja con 5 realizaciones, 2 del tipo $(10, 1)$ y 3 del tipo $(5, 1)$).

El nodo 7 se divide en los nodos 14 (hoja pura con 18 realizaciones del tipo $(10, 1)$) y 15 (hoja con 5 realizaciones, 3 del tipo $(10, 1)$ y 2 del tipo $(20, 1)$).

De los 10 datos de la muestra de testeo, 9 son correctamente clasificados. El único error es un $(3, 1)$ que es clasificado como $(5, 1)$. Los errores duro, ponderado y suave son respectivamente 0.1, 0.005 y 0.

Figura 17:



Agregación de predictores

Pequeños cambios en la muestra L pueden producir árboles muy distintos, los cuales darán predicciones distintas sobre a cual hoja debe ir un nuevo dato (fenḿeno conocido como *inestabilidad*). Dada la inestabilidad del árbol , se emplea el método *Bagging* (**B**ootsrap**a**gregating) para tener un predictor estable: Una vez elegidas L y T, fabricamos muestras bootstrap (remuestras con reposición) de L con las cuales hacemos los correspondientes árboles de clasificación. Se asignan los datos de T a las hojas correspondientes usando cada árbol y luego se da una asignación final usando voto mayoritario.

Grafo 2000B: Las remuestras se toman de L (75 realizaciones) y el testeo se hace con T (10 realizaciones). Los datos mal clasificados por voto mayoritario suelen ser aquellos correspondientes a epidemias $(2, 1)$, $(3, 1)$ (al ser muy parecidas entre sí es de esperar ese tipo de errores) y eventualmente $(5, 1)$, las que por ser una suerte de transición entre las epidemias "suaves"y las más "virulentas", puede llegar a clasificarse erróneamente como $(10, 1)$ o $(20, 1)$.

A modo de ejemplo, ejecutando 5 veces el método *Bagging* con 10 corridas cada vez, los errores de clasificación por voto mayoritario en c/u de las 5 ejecuciones son respectivamente 0.20, 0, 0, 0 y 0, dando una media de 0.04 . Esto muestra que 10 corridas ya producen un buen resultado.

Sin embargo observemos que cada vez que ejecutamos el *Bagging* cambian la muestra L de la cual luego se toman remuestras, así como la muestra T que se usa para testeo. Repetimos ahora las 5 ejecuciones del *Bagging* pero con 20 corridas cada uno, en cada una de las cuales las remuestras se hacen sobre la misma L, y los testeos con la misma T. El error de clasificación medio por voto mayoritario es de 0.06.

Grafo 2000d: Las remuestras se toman sobre una muestra L de 70 realizaciones, el testeo se hace sobre una muestra T de 30 realizaciones. El parámetro hojamin se fijó en 10. Esto provoca que muchos de los árboles tengan cuatro hojas, y como son cinco etiquetas esto implica que las realizaciones de la muestra T que posean una etiqueta para la cual no hay hoja con la misma etiqueta serán todas mal clasificadas.

Cuando hacemos 20 iteraciones, obtenemos una tasa de error de 0.1 : dos realizaciones del tipo $(2, 1)$ que son clasificadas como $(3, 1)$ y una realización $(20, 1)$ clasificada también como $(3, 1)$. Tanto el atributo 1 como el 2 son utilizados, algunos con mayor frecuencia según el árbol. En algunos árboles se usa casi exclusivamente el atributo1, en otros se usa casi exclusivamene el atributo 2.

Observaciones

1. El algoritmo reconoce los distintos tipos de epidemias y las agrupa razonablemente bien, a pesar de lo exigüo de la muestra de entrenamiento (75 casos de 500). Asimismo clasifica razonablemente bien los casos para testear.

2. Esto alienta la idea central de asociar a un proceso estocástico un vector de funciones como descriptor.

6.2.2. Regresión

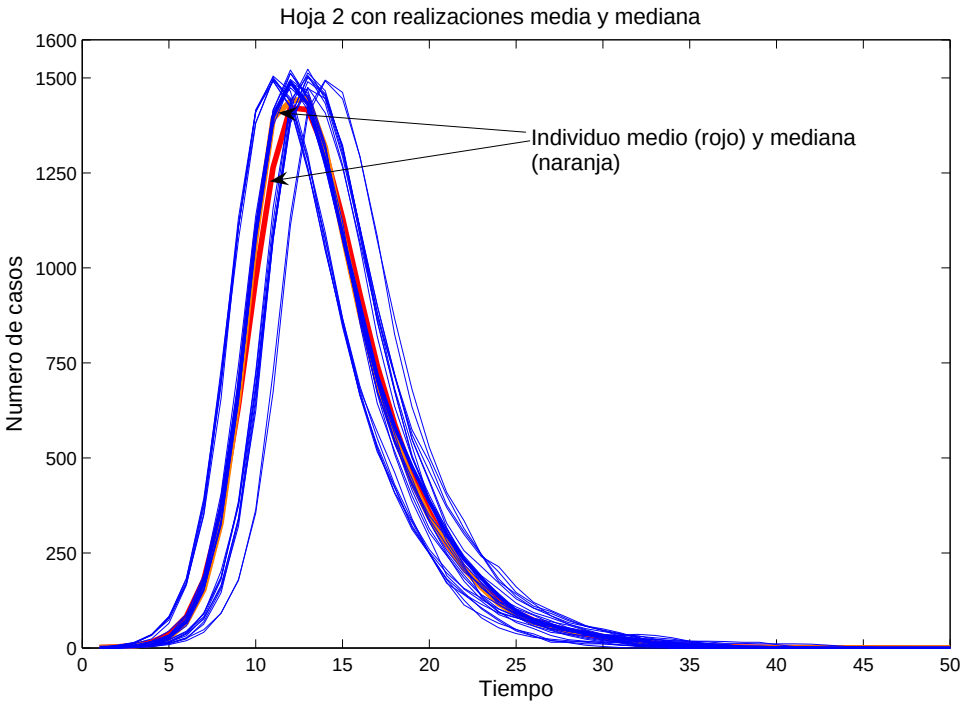
Ejemplo 1): Comentaremos el caso de parámetros $\mu = 5,\ \sigma = 1$, por ser una suerte de caso "intermedio"entre las epidemias suaves y las más virulentas. Los resultados para una muestra de entrenamiento de 100 realizaciones se ven a continuación (ocurrieron epidemias en 94 casos). Se fijó el tamaño mínimo de una hoja (hojamin) en 12 realizaciones, de modo que el algoritmo elige agrupar realizaciones minimizando la impureza pero se detiene si dicha disminución entraña subdividir un grupo en subgrupos de menos de 12 realizaciones, o si la ganancia en disminución de la impureza es menor a 0.05. Como medida de impureza se tomó $\sum_{i=1}^{n}(y_i - \overline{y})^2$.

De los 100 realizaciones de la raíz, 24 van a parar al nodo 2 (que es una hoja) y los restantes van al nodo 3. Se muestra el nodo 2 (Figura 18) con un valor promedio para el máximo de $y = 1497$
El nodo 3 se parte según el atributo 2 en los nodos 6 (hoja con 22 realizaciones e $y = 1486$, Figura 19) y el nodo 7.

El nodo 7 a su vez se subdivide en los nodos 14 y 15 (hoja con 19 realizaciones e $Y = 1459$). La Figura 20 muestra las realizaciones de la hoja 15. El nodo 14 se subdivide en las hojas 28 (14 realizaciones e $y = 1485$, Figura 21) y la 29 (18 realizaciones e $y = 1473$, Figura 22). Se muestran ambas hojas, siempre con las realizaciones media y mediana. Notemos que el algoritmo reconoce subgrupos entre los 100 casos "(5-1)".

Ejemplo 2): Ahora haremos una regresión con una muestra L conformada por 75 realizaciones de los 5 tipos $((2,1),(3,1),(5,1),(10,1)$ y $(20,1))$. La muesta T consta de otras 10 realizaciones. El tamaño mínimo de un nodo se fijó en 8 realizaciones. La devianza (dispersión) inicial del nodo raíz es de 44.980.125. Este nodo raíz se divide en los nodos 2 y 3, con devianzas de

Figura 18: Hoja 2



1.402.055 y 3.097.483 respectivamente. El atributo según el cual se efectúa el split es el 1 en casi todos los casos.

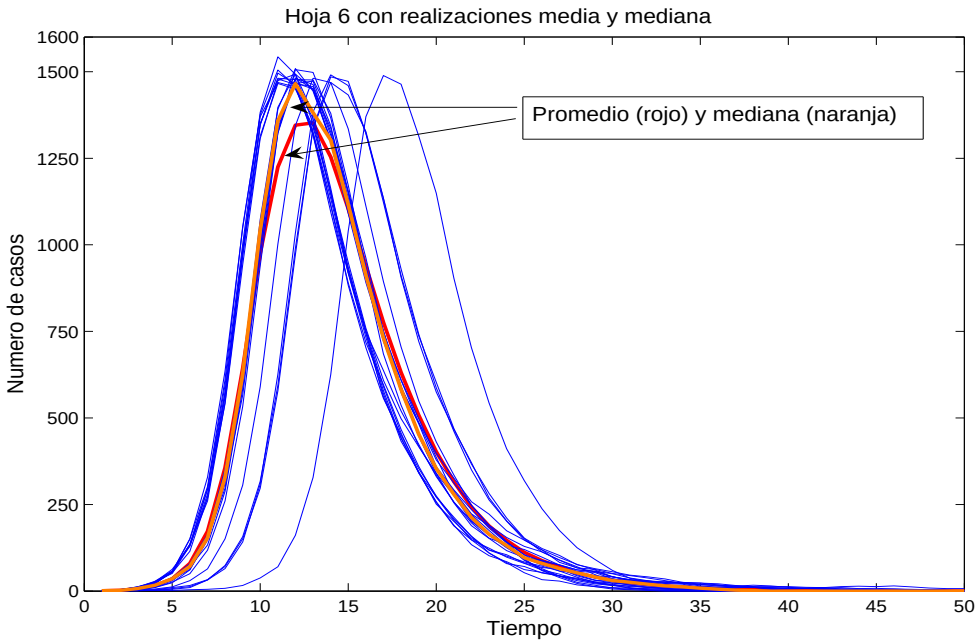
El Nodo 2 (44 realizaciones) se parte en los nodos 4 (hoja) y 5 mediante el atributo 1. La devianza del nodo 4 es 4.837, tiene 14 realizaciones y el valor promedio del pico epidémico es y=1.483,7.

El nodo 5 (27 realizaciones) tiene una devianza de 55.984, y se parte según el atributo 1 en los nodos 10 y 11, ambos hojas. La hoja 10 tiene una devianza de 679, 17 realizaciones y un valor y=1.807. La hoja 11 tiene una devianza de 80, 10 realizaciones e y=1.901.

El nodo 3 (31 realizaciones) se parte en los nodos 6 (hoja) y 7 mediante el atributo 1. La devianza del nodo 6 es 6.300, tiene 9 realizaciones e y=705.

El nodo 7 (22 realizaciones) tiene devianza 7.266,5 y se parte mediante el atributo 2 en los nodos 14 y 15, ambos hojas. La devianza del nodo 14 es

Figura 19:



11, tiene 13 realizaciones e y=1,61, mientras que la devianza del nodo 15 es 4.948,2, tiene 9 realizaciones e y=22,44.

Notemos las fuertes caídas de la devianza en cada partición, siempre de al menos un orden de magnitud. El proceso de partir cada nodo se detiene por el requerimiento de población mínima de 8 casos por nodo antes que por la pureza del mismo.

Con este árbol testeamos los 10 realizaciones de la muestra T. Recordemos que en regresión a cada realización se le asigna como etiqueta el valor del máximo de su curva casos vs. tiempo. La tabla 6.2.2 muestra los valores de los picos de las realizaciones (yindiv), el valor medio del pico de la hoja a la que fue asignado (yhoja) y el número de dicha hoja.

Figura 20:

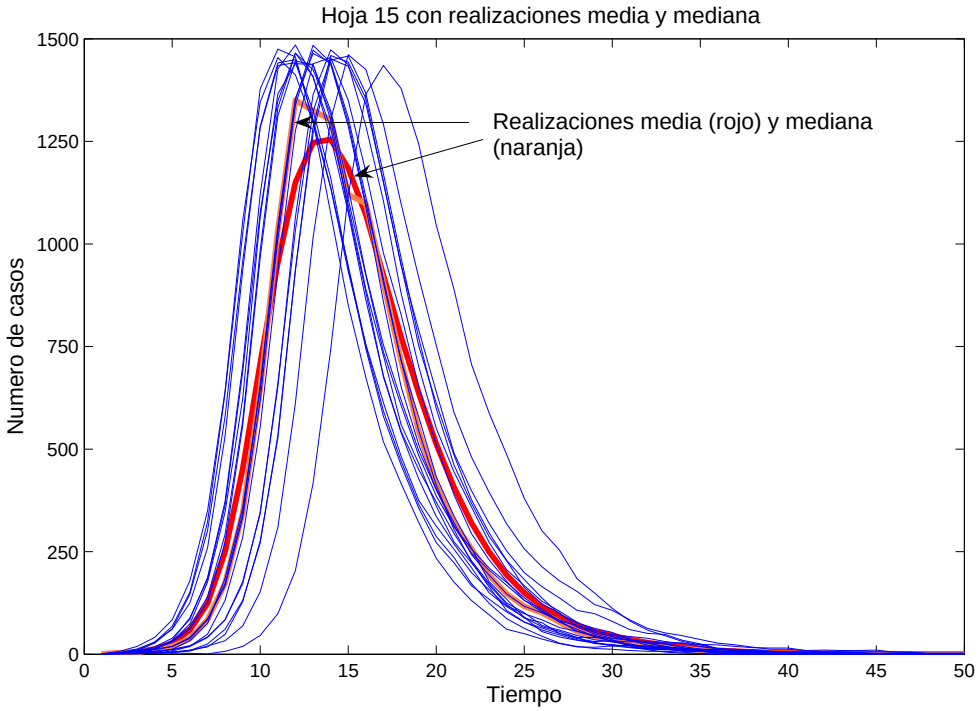
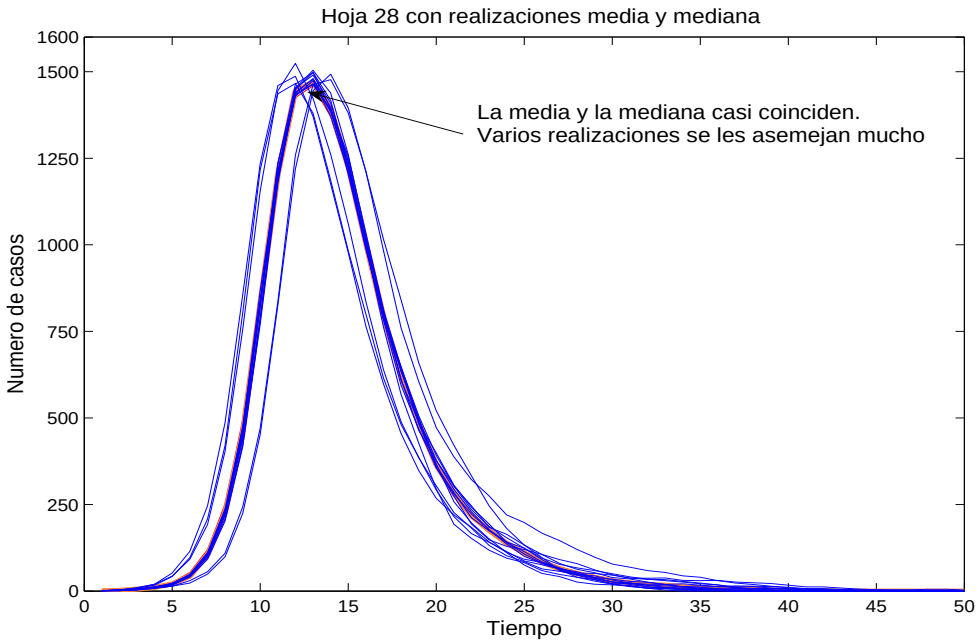


Tabla 6.1.3										
hoja	14	15	15	6	4	4	10	10	11	11
yindiv	1	34	8	678	1493	1498	1822	1818	1905	1915
yhoja	2	22	22	705	1484	1484	1808	1808	1901	1901

Figura 21:



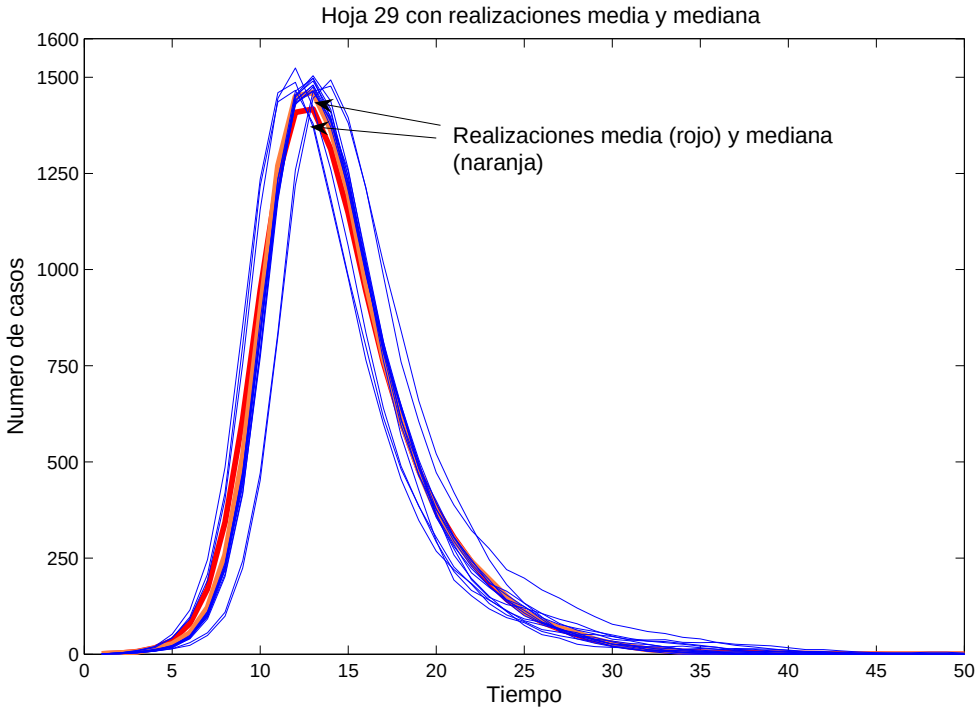
6.2.3. Vacunación

Los mejores blancos para vacunar deberían ser aquellos grupos de nodos con mayor participación en cuanto a aristas en el total de aristas del grafo. Si además fuera importante el tamaño de la población a vacunar (por disponerse de una cantidad limitada de vacunas, por ejemplo), parece razonable preferir - entre grupos con igual peso en cuanto a aristas - al de menor peso en cuanto a nodos.

Una medida para comparar grupos teniendo en cuenta ambos aspectos puede ser el cociente entre el aporte de aristas del grupo y el aporte de nodos del mismo. Un grupo porcentualmente menor en cuanto a realizaciones pero con gran participación de aristas tendrá un cociente mayor que otro grupo con más nodos pero más ralo en cuanto a aristas.

Sin embargo no importa solamente la cantidad de aristas, también la "calidad"de las mismas. Esto es, si un grupo de nodos tiene muchas aristas pero

Figura 22:



estas son mayoritariamente "internas"o "locales"(vinculan nodos del mismo grupo de modo que dicho grupo está muy cerca de conformar un grafo completo), entonces el vacunar a dichos nodos no va a minar la estructura de aristas del grafo en la medida en que lo podría hacer otro grupo de nodos cuyas aristas fueran mayoritariamente hacia nodos de otros grupos (ver sección 2.3.2).

Es como tener un grupo cerrado de individuos que solamente se relacionan entre sí. Aunque la cantidad de contactos (aristas) entre ellos pueda ser alta, son aristas internas o endogámicas, por lo que la remoción del citado grupo no afectará de manera perceptible al resto del grafo.

Es de esperar que los nodos de grado más alto participen más de la epidemia que los nodos de grado más bajo, pero en las epidemias más virulentas, si

comparamos el grado de los nodos infectados con los de la población entera, la distribución de frecuencias no parece ser muy distinta.

La epidemia no parece preferir a los nodos con un gran número de aristas (grado), por cuanto el porcentaje de los mismos respecto de la población infectada es prácticamente el mismo que el de todos los nodos con igual grado respecto de la población total antes de la epidemia (Figuras 23 y 24).

Surge entonces la duda sobre si no será la fuerza de la epidemia la que al contagiar a casi toda la población produce ese efecto igualitario, según el cual el grado de los nodos no parece ser relevante (aunque sí podría serlo la topología del grafo).

Podría pensarse que en lugar de vacunar a la población con mayores contactos (los nodos de grado más alto), una alternativa podría ser enfocarse en la población que representa la mayoría de los nodos infectados, por cuanto si se elige un individuo al azar, es más probable que pertenezca a dicho grupo.

La epidemia se propaga por las aristas, pero la vacunación tiene lugar en los nodos. Desde el punto de vista de minar la estructura de aristas del grafo ¿dará lo mismo eliminar un cierto porcentaje de las mismas inmunizando los nodos de mayor grado que si lo hacemos con nodos de grado más bajo? Cabe pensar que si el costo de vacunar no es relevante, es mejor juntar dicho porcentaje de aristas entre los nodos de grado más bajo (que son los más numerosos) que entre los de grado alto (que son menos).

Investigaremos el punto mediante simulaciones en los grafos 2000B y 2000d. Aplicaremos dos estrategias de vacunación distintas. Primero vacunaremos todos los nodos de grado más alto, hasta detectar en que momento dejan de ocurrir epidemias.

A continuación calculamos el porcentaje de aristas que dichos nodos representan en el total de aristas del grafo, y vacunamos una parte del grupo mayoritario (en frecuencia) hasta igualar el aporte de aristas.

Luego corremos 100 simulaciones de cada caso y comparamos la epidemia media de cada caso con la media cuando no hay vacunación alguna. Haremos esto para epidemias desde las más virulentas hasta las más suaves. Hasta lo mejor de mi conocimiento este tipo de comparación de estrategias de vacunación y su relación con la topología del grafo, introducen un enfoque alternativo en el estudio de este tema.

Grafo 2000B

En el grafo 2000B, la distribución de nodos y otras características según su grado se recoge en la tabla 6.2.3:

Figura 23:

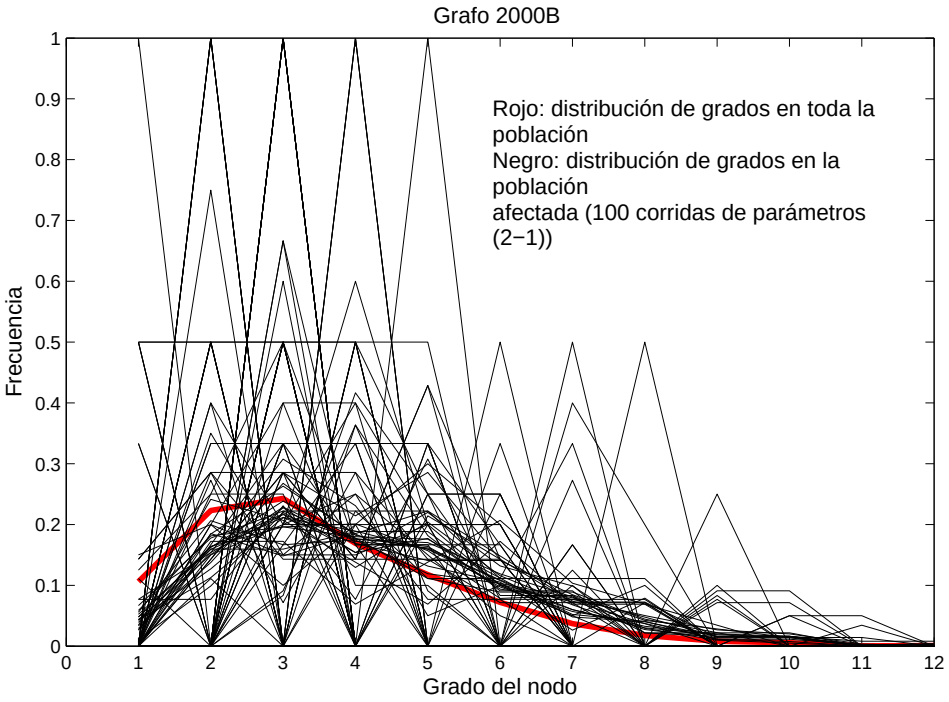
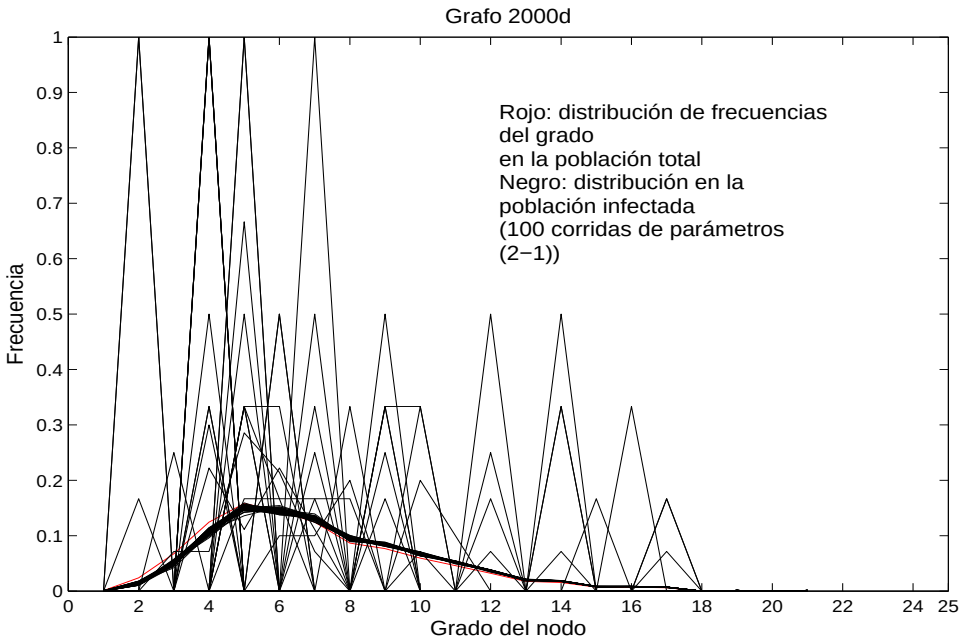

Grafo 2000B
Rojo: distribución de grados en toda la población
Negro: distribución de grados en la población
afectada (100 corridas de parámetros (2−1))
Frecuencia
Grado del nodo
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
1
2
3
4
5
6
7
8
9
10
11
12

Figura 24:



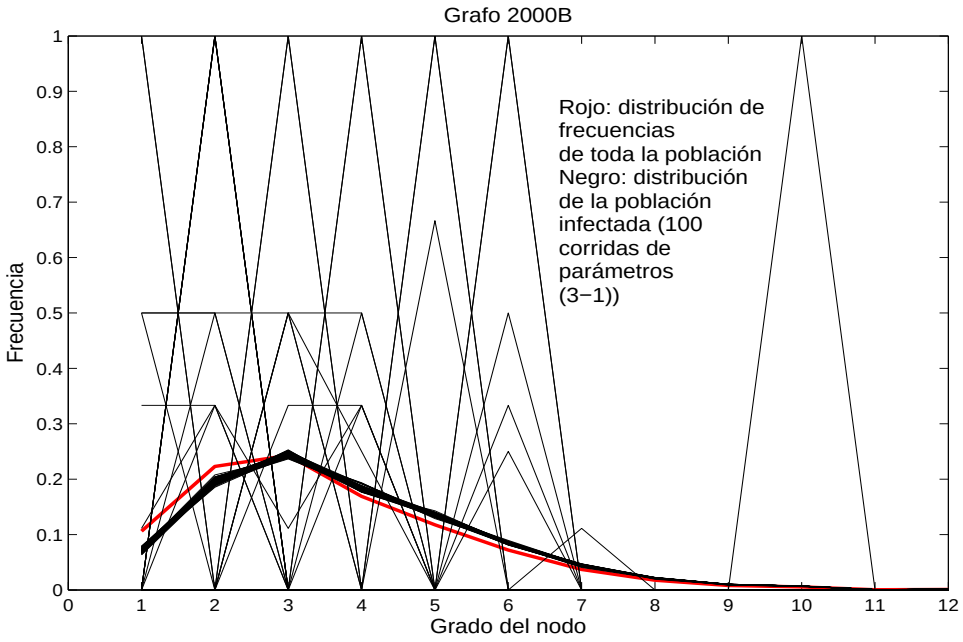
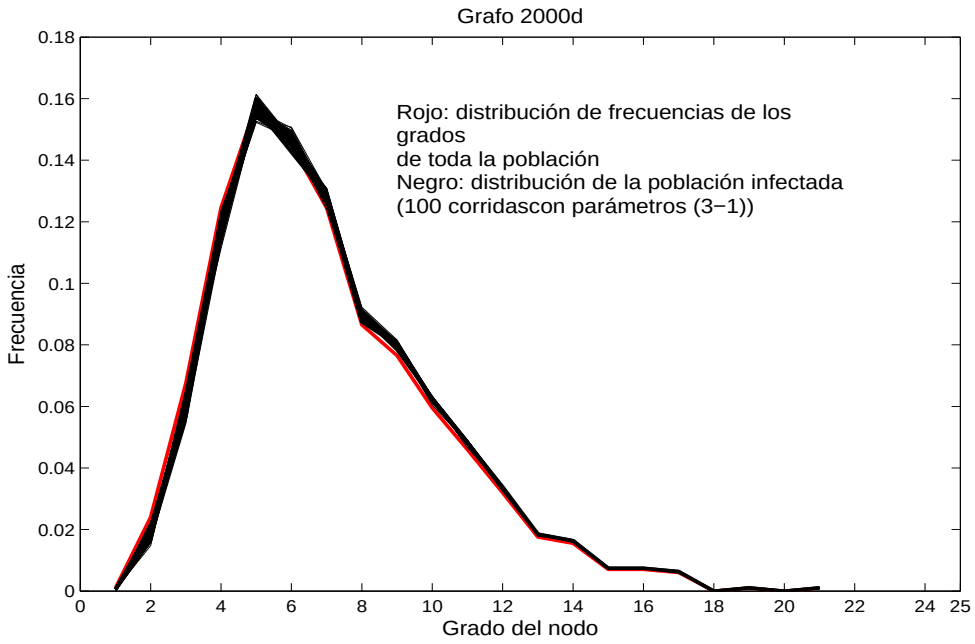


Figura 25:

Figura 26:



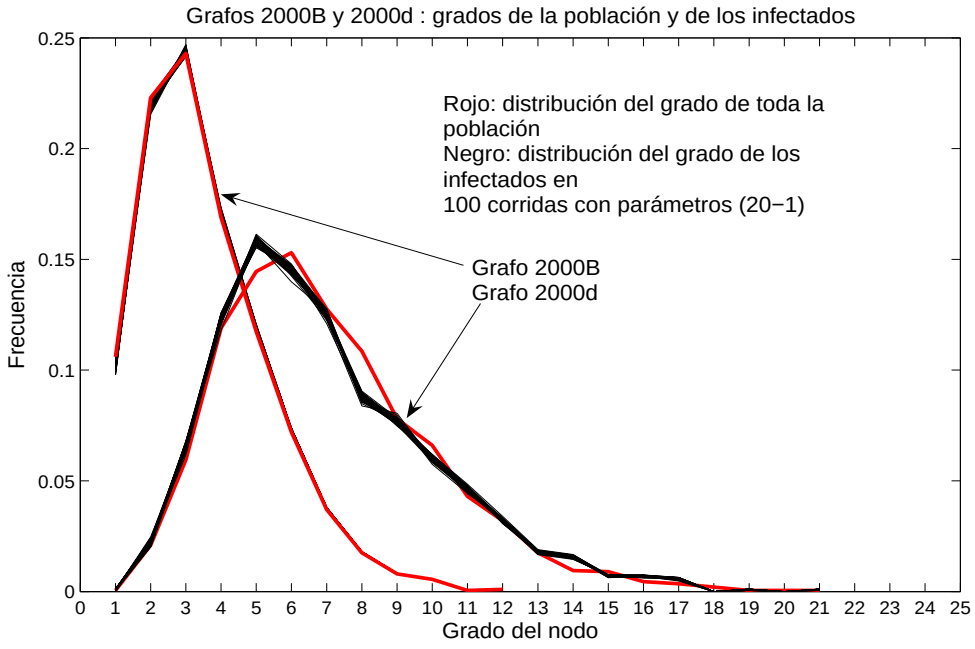


Figura 27:

Tabla 6.1.4					
Grado	Nro.nodos	frec.nodos	Nro.aristas	frec.aristas	frec.aristas/frec.nodos
1	212	0.1060	212	0.0301	0.2840
2	446	0.2230	892	0.1267	0.5682
3	486	0.2430	1458	0.2071	0.8523
4	338	0.1690	1352	0.1920	1.1361
5	235	0.1175	1175	0.1669	1.4204
6	144	0.0720	864	0.1227	1.7042
7	74	0.0370	518	0.0735	1.9865
8	35	0.0175	280	0.0397	2.2686
9	16	0.0080	144	0.0204	2.5500
10	11	0.0055	110	0.0156	2.8364
11	1	0.00055	11	0.00156	2.8364
12	2	0.0010	24	0.0034	3.4000
Efglobal	0.1707	Eflocal	0.0026		

Los pesos relativos de cada grupo de nodos de igual grado respecto del total de nodos del grafo, así como de las aristas de dicho grupo en el total de aristas del grafo se ven en la Figura 28.

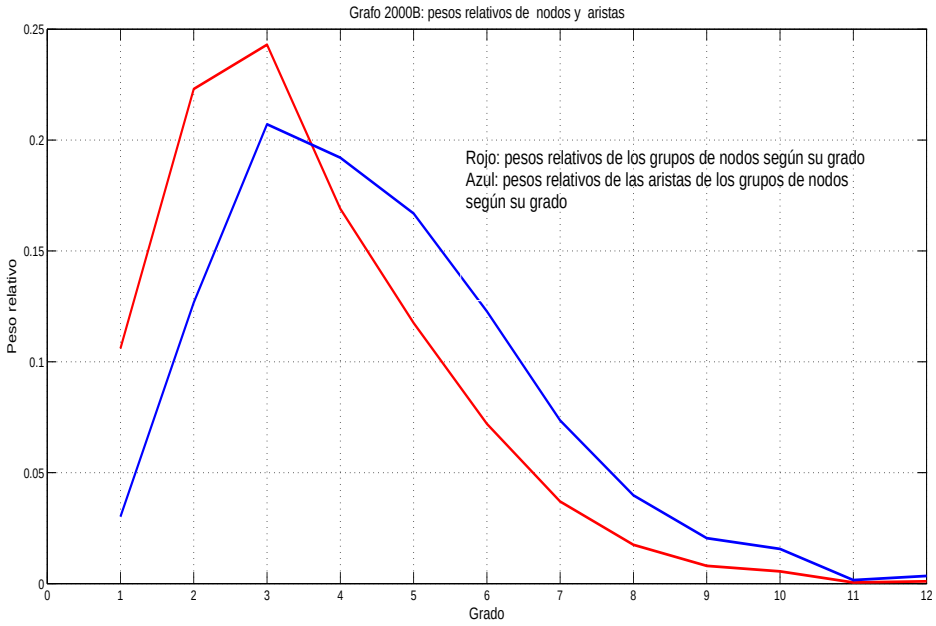
Los nodos de grado alto pesan más en cuanto a aristas que en cuanto a número. El grupo mayoritario en número de nodos es también el que tiene mayor peso individual en aristas.

Si inmunizamos a todos los nodos de grado mayor o igual a 5, entonces no ocurren epidemias. Lo mismo ocurre cuando inmunizamos a todos los de grado mayor o igual a 6 (representan el 14.15 % de la población de nodos y el 27.68 % de la de aristas). Recién comienzan a ocurrir epidemias cuando inmunizamos a todos los nodos de grado mayor o igual a 7.

Esto ya nos dice el importante rol jugado por el conjunto de nodos de grado más alto: Los nodos de grado mayor o igual a 6 son 283 (el 14.15 % de la población), con cuya inmunización no ocurren epidemias.

Si por el contrario inmunizamos 283 nodos del grupo mayoritario (el 30.36 % de los nodos de grado 2 y 3 y el 14.15 % del total) y corremos 100 réplicas de la epidemia, observamos en primer lugar que ocurren epidemias. éstas son de menor cuantía, comparables e incluso menores al caso de parámetros $(10, 1)$ como se aprecia en la Figura 29 (el pico epidémico de la epidemia media es de 1805 casos). Esto sugiere que a igual peso en cuanto a cantidad de realizaciones, es más eficiente vacunar a aquellos de grado más

Figura 28:



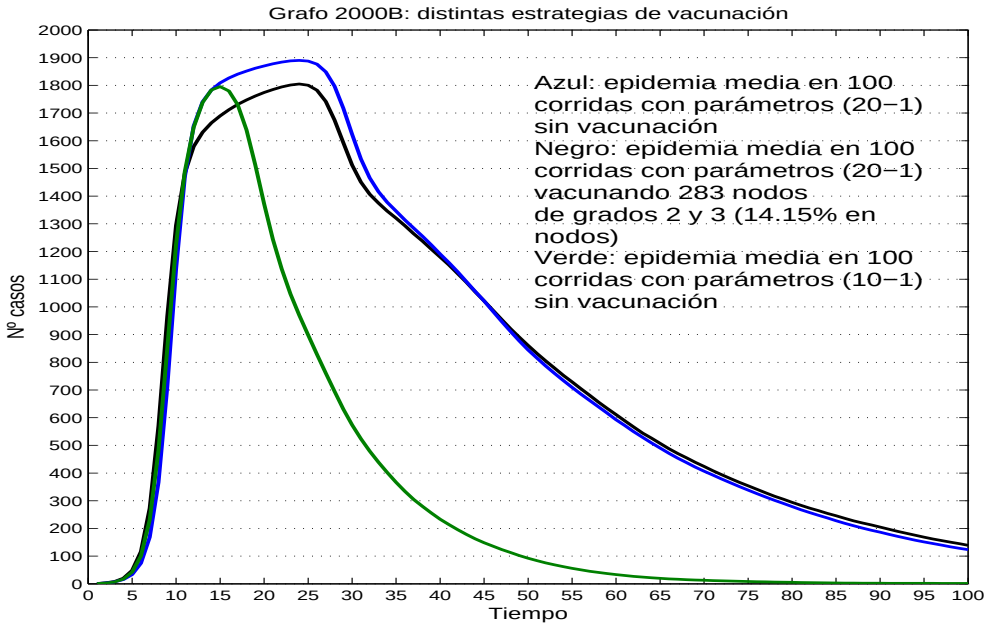


Figura 29:

alto (no ocurrieron epidemias en las 100 corridas).

Ahora vacunaremos grupos iguales en cuanto a peso en aristas. Para forzar la epidemia, inmunizaremos a los nodos de grado mayor o igual a 7. Los mismos son 139 nodos con 1087 aristas que representan el 6.95 % del total de nodos y el 15.44 % del total de aristas. Para igualar el peso en aristas en los dos grupos mayoritarios en cuanto a nodos (los de grado 3 y 2 respectivamente), inmunizamos el 46.25 % de los mismos (431 nodos, el 21.55 % de la población) y corremos 100 simulaciones de cada caso con parámetros $(20, 1)$. Las epidemias medias se muestran en la Figura 30.

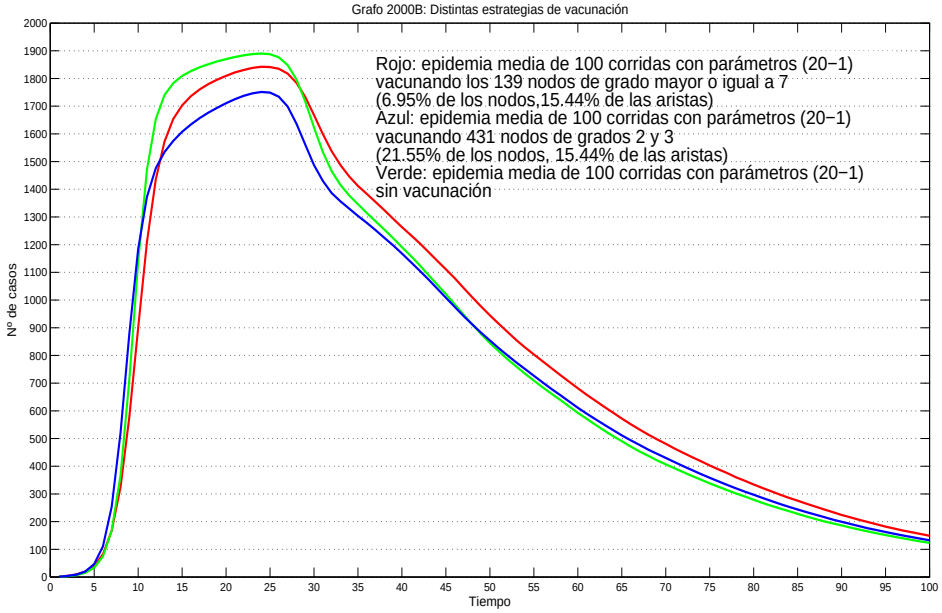
La epidemia media para la vacunación del grupo de grados 2 y 3 es menor aunque se inmunizó al 21.55 % del total de nodos frente al 6.95 % que representan los nodos de grado mayor o igual a 7. Tenemos una diferencia de 97 casos (un 5.40 % menos) contra el triple de realizaciones vacunados. Ahora veamos el comportamiento en el caso de parámetros $(10, 1)$ (Figura 31).

En epidemias menos virulentas, donde el efecto igualador de la fuerza de la epidemia es menor, la diferencia a favor de vacunar los nodos de mayor grado debería ser mayor.

Veremos para el caso de parámetros $(5, 1)$ (Figura 32).

Notamos que la diferencias entre los resultados de ambas políticas de vacunación es ahora menor que en el caso $(20, 1)$. Cuando vacunamos dentro del grupo mayoritario de realizaciones (los nodos de grados 3 y 2) el pico epidémico se da antes que cuando vacunamos los nodos de grado más alto. Esto sugiere que la conectividad del grafo se ve menos afectada en el primer caso que en el segundo.

Figura 30:



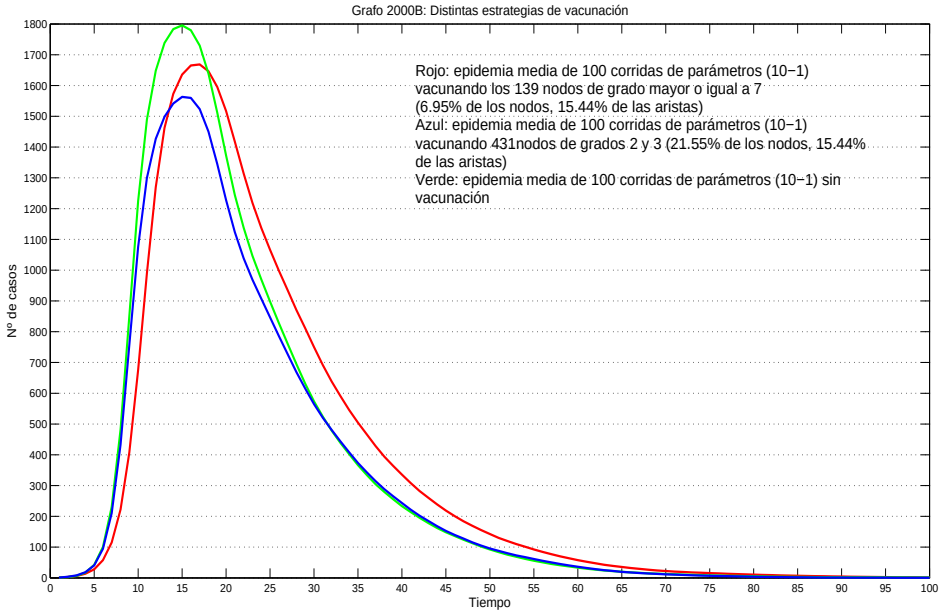
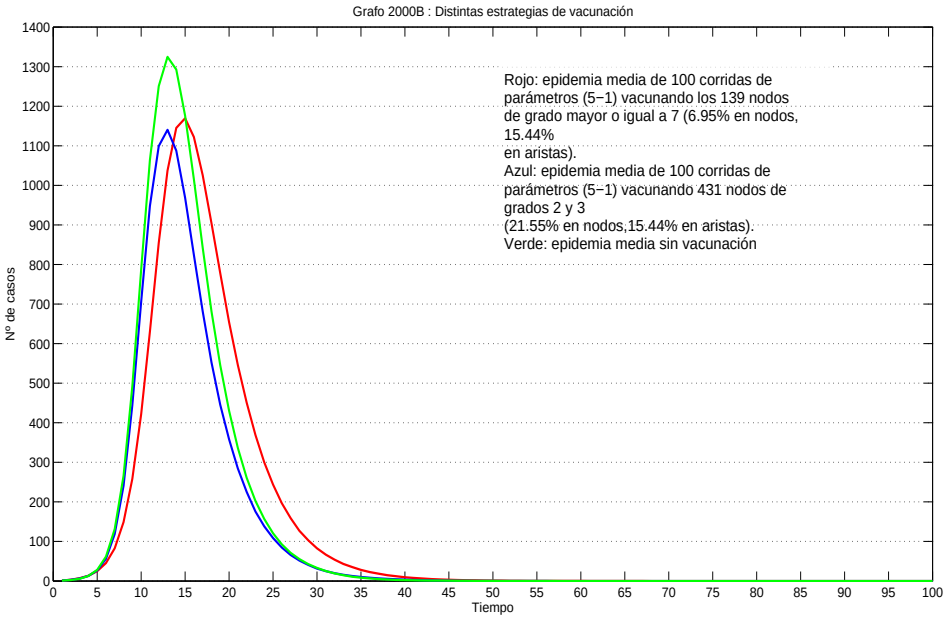


Figura 31:

Figura 32:



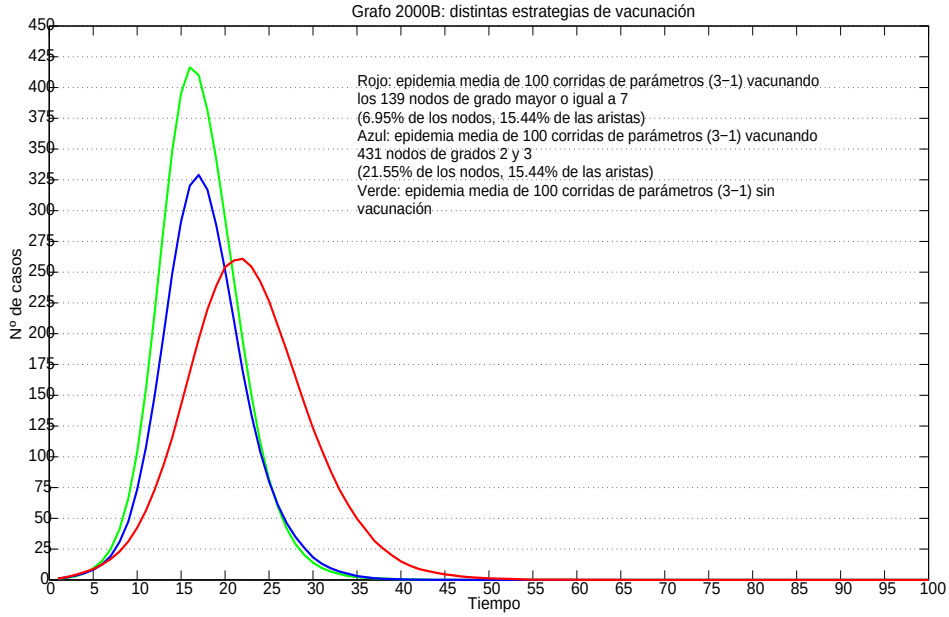


Figura 33:

Ahora el caso de parámetros $(3, 1)$ (Figura 33).

Notamos que a medida que la epidemia es menos virulenta, la estrategia de vacunar los nodos de grado más alto es cada vez más efectiva frente a la de vacunar nodos de los grupos mayoritarios. Asimismo la curva media cuando vacunamos al grupo mayoritario suele tener su máximo en tiempos anteriores al de la media, mientras que al vacunar el grupo con mayor grado los tiempos se retrasan. La epidemia avanza y concluye antes para la primera estrategia que para la segunda. Esto sugiere que el grafo queda más desconectado al remover un porcentaje de aristas a través de nodos de grado alto que el mismo porcentaje a través de nodos de grado bajo.

Grafo 2000d En el grafo 2000d, la distribución de nodos y otras características según su grado se recoge en la tabla 6.1.5

Tabla 6.1.5					
Grado	Nro. nodos	frec.nodos	Nro.aristas	frec.aristas	frec.aristas/frec.nodos
1	2	0.0010	2	0.0001	0.1000
2	48	0.0240	96	0.0069	0.2875
3	135	0.0675	405	0.0291	0.4311
4	249	0.1245	996	0.0717	0.5759
5	316	0.1580	1580	0.1137	0.7196
6	290	0.1450	1740	0.1252	0.8634
7	249	0.1245	1743	0.1254	1.0072
8	173	0.0865	1384	0.0996	1.1514
9	153	0.0765	1377	0.0991	1.2954
10	119	0.0595	1190	0.0856	1.4387
11	92	0.0460	1012	0.0728	1.5826
12	64	0.0320	768	0.0553	1.7281
13	35	0.0175	455	0.0327	1.8686
14	31	0.0155	434	0.0312	2.0129
15	14	0.0070	210	0.0151	2.1571
16	14	0.0070	224	0.0161	2.300
17	12	0.0060	204	0.0147	2.4500
18	0	0	0	0	0
19	2	0.0010	38	0.0027	2.700
20	0	0	0	0	0
21	2	0.0010	42	0.0030	3
Efglobal	0.2565	Eflocal	0.0079		

Los pesos relativos de cada grupo de nodos de igual grado respecto del total de nodos del grafo, así como de las aristas de dicho grupo en el total de aristas del grafo se ven en la Figura 34.

De nuevo en el caso más virulento ($\mu = 20, \sigma = 1$) no ocurren epidemias si removemos los nodos de grado mayor o igual a 6 (que ahora representan el 62.5 % del total de nodos y el 77.85 % del total de aristas). Si removemos los nodos de grado mayor o igual a 7 (920 nodos con 9081 aristas, que representan el 48 % del total de nodos y el 65.33 % del total de aristas),entonces ocurren epidemias. Esto implica un fuerte grado de conexión en el grafo, mayor que el del grafo 2000B.

Para igualar el porcentaje de nodos acudiendo a los nodos de los grupos mayoritarios en frecuencia, miramos los nodos de grados 4 al 7, (1104 nodos con 6059 aristas que representan el 55.2 % del total de nodos y el 43.59 % del total de aristas).

Marcamos como removidos 920 nodos de dicho grupo (83.33 % del grupo, 48 % de la población total) y corremos 100 réplicas de la epidemia. Los resultados se muestran en la Figura 35, junto con la epidemia media cuando no hay vacunación.

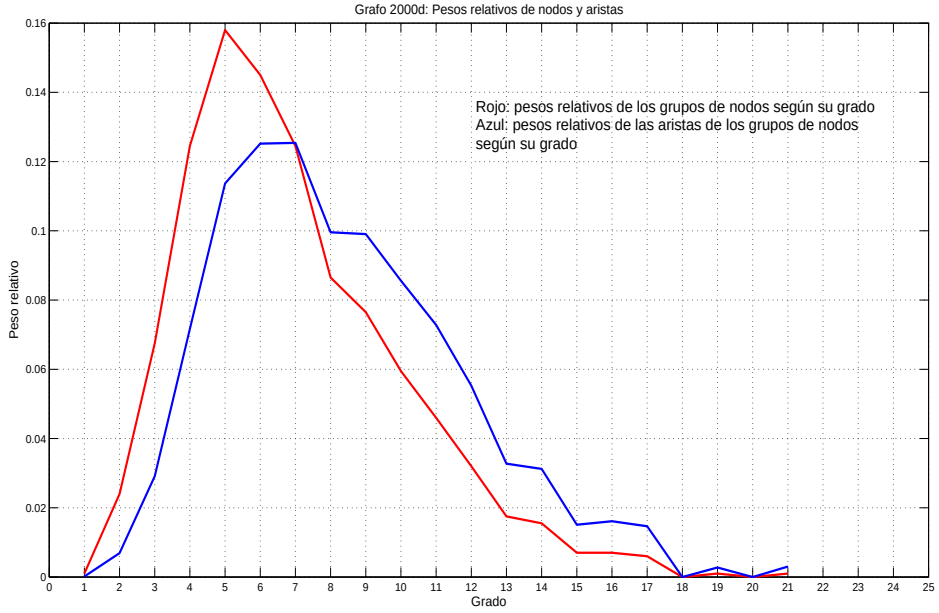
Al igual que en el grafo 2000B, la estrategia de vacunar los grupos mayoritarios en cuanto a nodos ofrece mejores resultados que vacunar los nodos con mayor cantidad de aristas, siempre en el contexto de forzar la epidemia. Es claro que si vacunamos todos los nodos de grado mayor o igual a 6, no ocurren epidemias.

Una posible explicación puede ser que aunque los nodos de grado mayor o igual a 7 representan más aristas que los nodos de grados 4 al 7, muchas de estas arístas son de carácter "interno ", por lo que su eliminación mediante la vacunación de los nodos correspondientes no mina tanto la conectividad del grafo como cuando removemos los nodos de grados 4 al 7.

Claramente la disminución es más espectacular cuando removemos 920 nodos elegidos entre aquellos pertenecientes a los grupos de mayor frecuencia (los de grado 4 al 7) que si removemos la misma cantidad exclusivamente entre los de grado 7 en adelante. Esto sugiere que los nodos de grado más alto tienden a conectarse entre ellos, por lo que gran parte de las aristas eliminadas al inmunizar a dichos nodos serían aristas "internas".

Para epidemias de parámetros $(10, 1)$, de nuevo al remover todos los nodos de grado mayor o igual a 6 no ocurren epidemias. Las mismas comienzan a darse cuando removemos todos los nodos de grado mayor o igual a 7. Repetimos las estrategias de vacunación: removemos todos los nodos de grado

Figura 34:



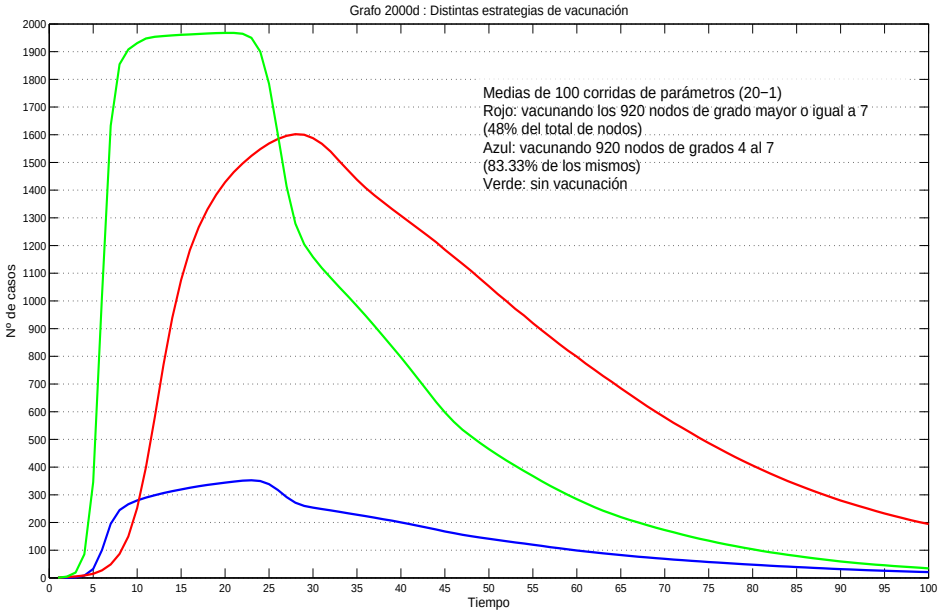


Figura 35:

mayor o igual a 7, y removemos el 83.33 % de los nodos de grado 4 al 7, comparando ambas estrategias. Los resultados se muestran en la Figura 36.

Para epidemias de parámetros $(5, 1)$ obtenemos la Figura 37.

El panorama es similar al caso anterior, lo que era de esperarse si en efecto la topología es la causante. Recién para epidemias menos virulentas (donde el grafo formado por los nodos infectados y las aristas por las cuales se dan los contagios difieren más del grafo preexistente sobre el cual se da la epidemia), cabe esperar que la influencia de la topología sea menor.

Para los casos de parámetros $(3, 1)$ y $(2, 1)$, las Figuras 38 y 39 nos muestran los resultados.

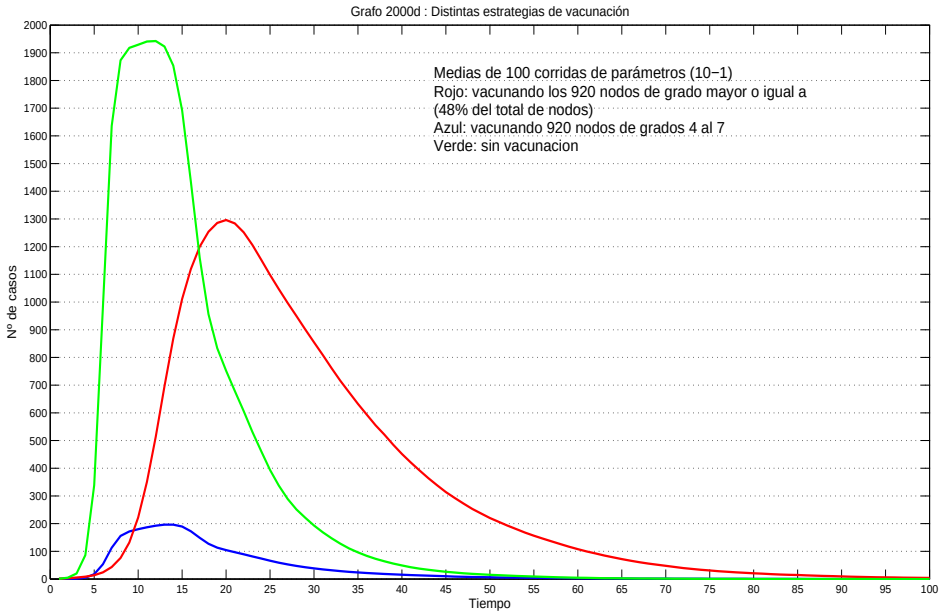
En la medida que la epidemia es menos virulenta y la influencia de la topología del grafo es menor, la estrategia de vacunar los nodos de mayor grado pasa a ser más eficiente que la de vacunar en los grupos mayoritarios.

Grafo 2000 la distribución de nodos y otras características según su grado se recoge en la tabla 6.1.6.

Tabla 6.1.6					
Grado	nro. nodos	frec.nodos	nro.aristas	frec.aristas	frec.aristas/frec.nodos
1	38	0.0190	38	0.0038	0.1998
2	194	0.0970	388	0.0388	0.3996
3	324	0.1620	972	0.0971	0.5994
4	345	0.1725	1380	0.1379	0.7992
5	374	0.1870	1870	0.1868	0.9990
6	272	0.1360	1632	0.1630	1.1988
7	196	0.0980	1372	0.1371	1.3986
8	123	0.0615	984	0.0983	1.5984
9	59	0.0295	531	0.0530	1.7982
10	28	0.0140	280	0.0280	1.9980
11	25	0.0125	275	0.0275	2.1978
12	11	0.0055	132	0.0132	2.3976
13	4	0.0020	52	0.0052	2.5974
14	4	0.0020	56	0.0056	2.7972
15	0	0	0	0	0
16	3	0.0015	48	0.0048	3.1968
Efglobal	0.2158	Eflocal	0.0043		

Los pesos relativos de cada grupo de nodos de igual grado respecto del

Figura 36:



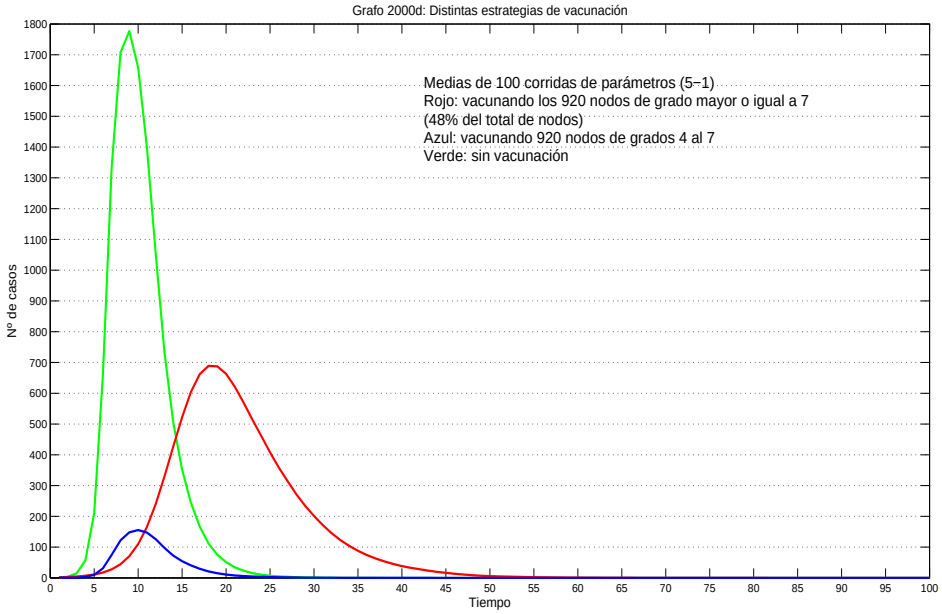
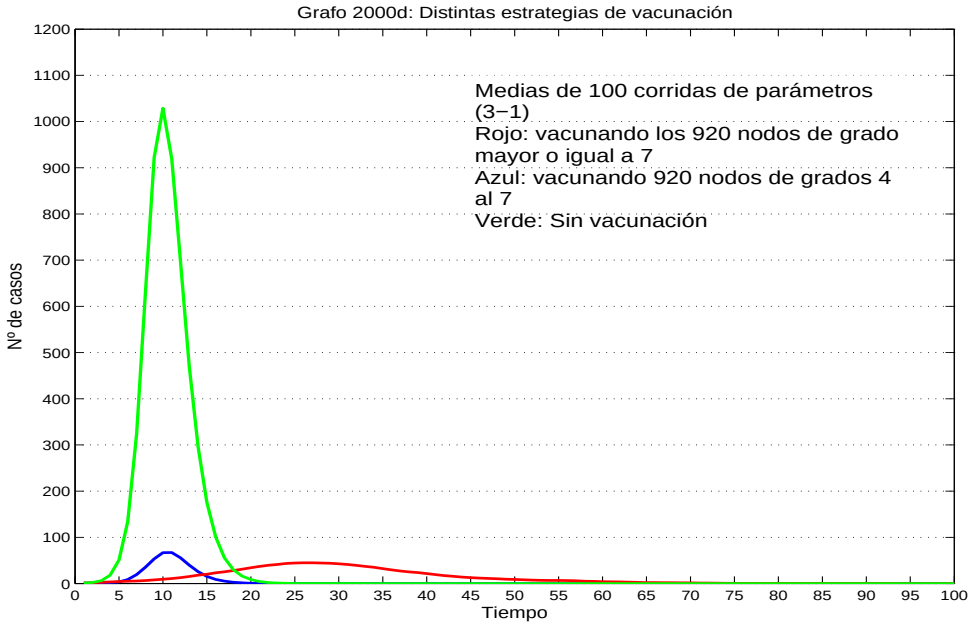


Figura 37:

Figura 38:



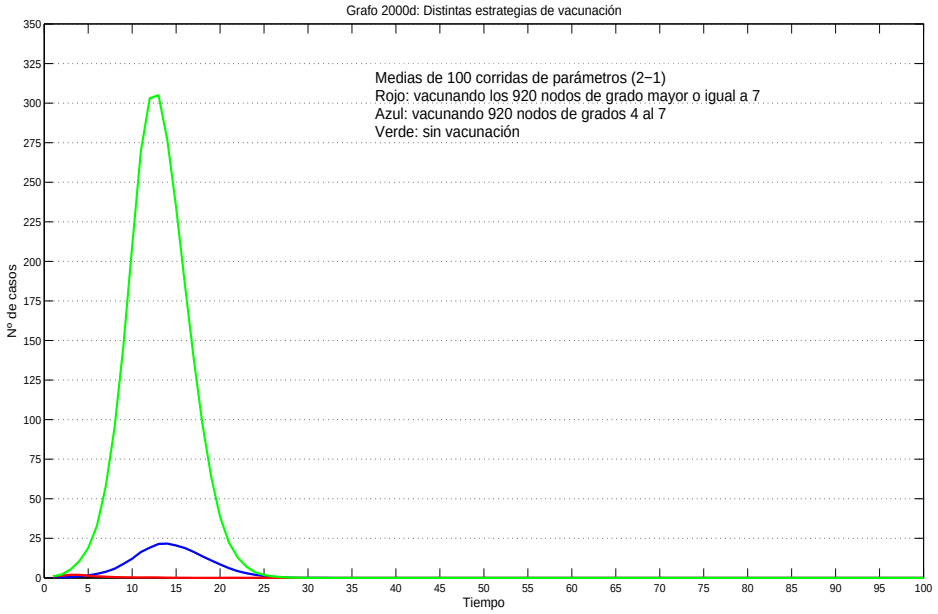
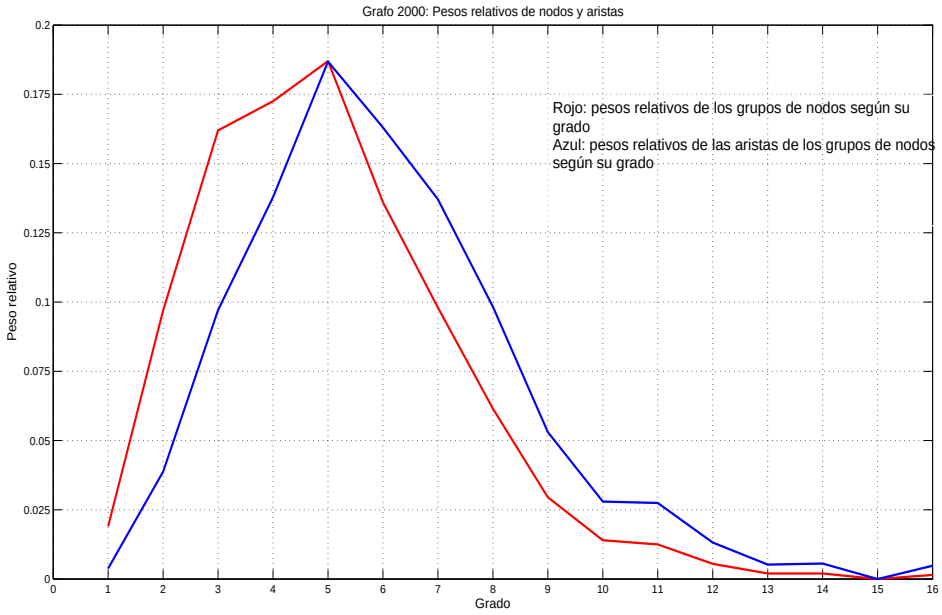


Figura 39:

Figura 40:

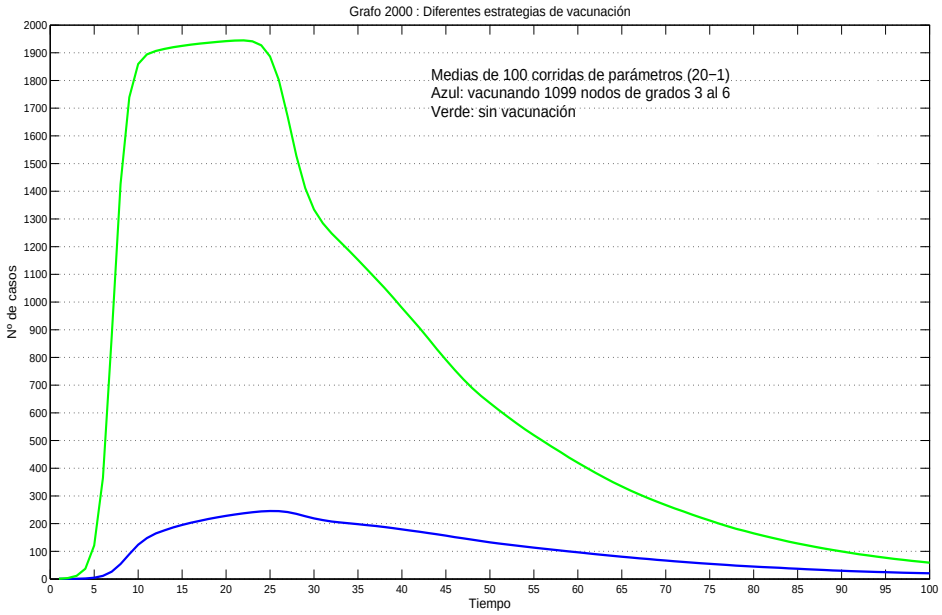


total de nodos del grafo, así como de las aristas de dicho grupo en el total de aristas del grafo se ven en la Figura 40.

En el caso más virulento (parámetros $(20, 1)$), cuando removemos todos los nodos de grado mayor o igual a 5 no ocurren epidemias. Estos nodos (1099, un 54.95 % del total de nodos) tienen 7232 aristas, un 72.25 % del total de aristas. Si removemos 1099 nodos de los grupos mayoritarios (los nodos de grados 3 al 6, que son 1315) y corremos 100 simulaciones para el caso de parámetros $(20, 1)$, obtenemos el resultado medio que se aprecia en la Figura 41.

Cuando removemos todos los nodos de grado mayor o igual a 6, ocurren epidemias. En este panorama forzado estos nodos (725 nodos, un 36.25 % del total), representan 5362 aristas, un 53.57 % del total de aristas. También removeremos 725 nodos de los grupos mayoritarios, los nodos de grados 3 al 6 (el 55.13 % de ellos, con aproximadamente 3227 aristas). Los resultados se

Figura 41:



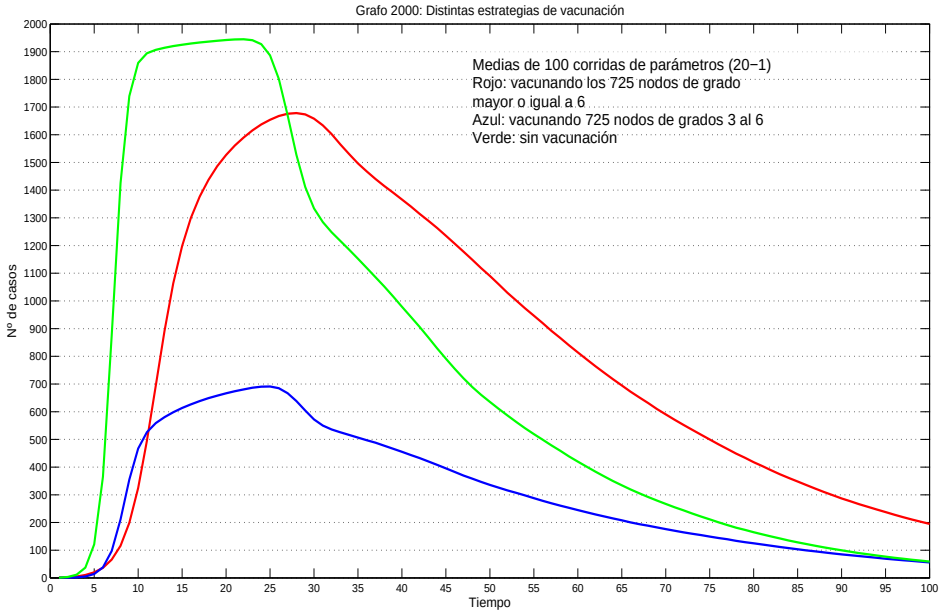


Figura 42:

aprecian en la siguiente Figura 42.

Repetimos para parámetros $(10, 1)$ (Figura 43).

Ahora para parámetros $(5, 1)$ (Figura 44).

Al igual que en los casos anteriores, en la medida que la epidemia es menos virulenta la diferencia entre las dos estrategias se va atenuando. De nuevo puede interpretarse como síntoma de una gran proporción de aristas internas entre los nodos de mayor grado.

Para el caso de parámetros $(3, 1)$ la Figura 45 muestra los resultados.

Y para el caso $(2, 1)$, la Figura 46 muestra los resultados.

6.2.4. Conclusiones

Si bien la estrategia de remover aquellos nodos con mayor número de aristas es la más efectiva, no lo es tanto como cabría esperar. El éxito de la estrategia alternativa, donde se remueven los nodos mayoritarios en número y no tanto en contribución de aristas al grafo, puede explicarse en base a la baja eficiencia local de los tres grafos aleatorios considerados.

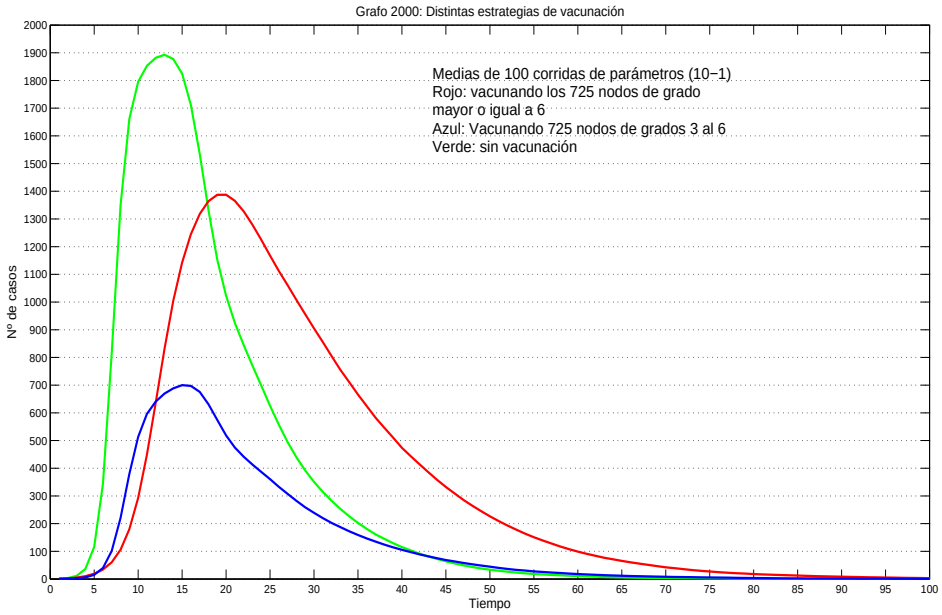
Esta característica, más propia de grafos puramente aleatorios que de redes *small-world* (ver sección 2.3.2) nos dice que dichos grafos son altamente sensibles a la remoción de vértices, sin importar demasiado el grado del mismo. Cuando la epidemia es menos virulenta, el peso de la topolgía subyacente a la red disminuye, y entonces la estrategia de remover los nodos con mayor grado vuelve a ser efectiva.

6.3. Simulaciones en una red de grafos aleatorios

6.3.1. Descripción

Ahora formamos una red de 5 grafos aleatorios (2000B,2000d,2000,2000B,2000) a los que llamaremos Sitio 1, Sitio 2, Sitio 3, Sitio 4 y Sitio 5 respectivamente. En cada iteración existe una probabilidad de que un nodo de un grafo se conecte aleatoriamente con un nodo de otro grafo (si pensamos en ciudades e realizaciones, es la probabilidad que tiene un individuo de viajar a otra ciudad).

Figura 43:



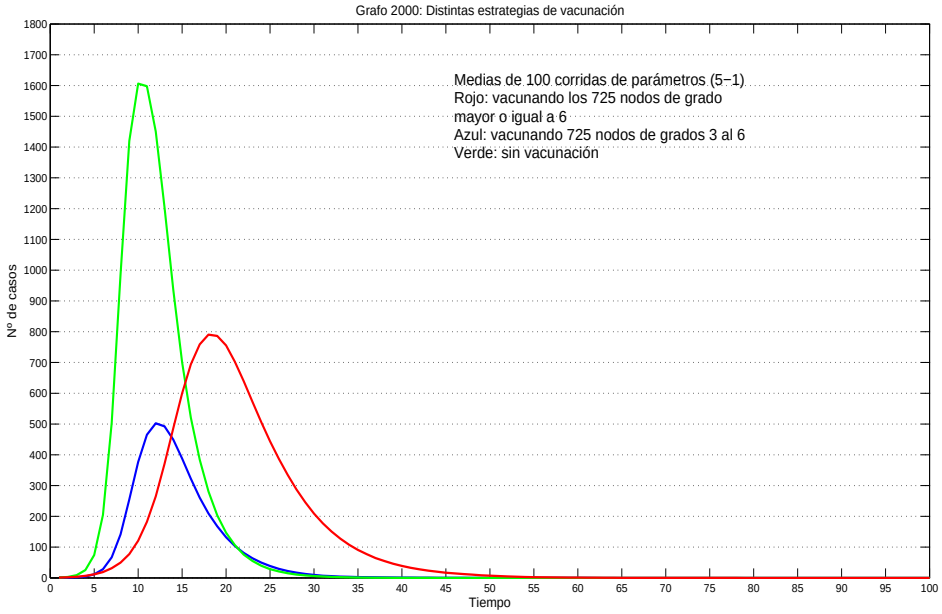
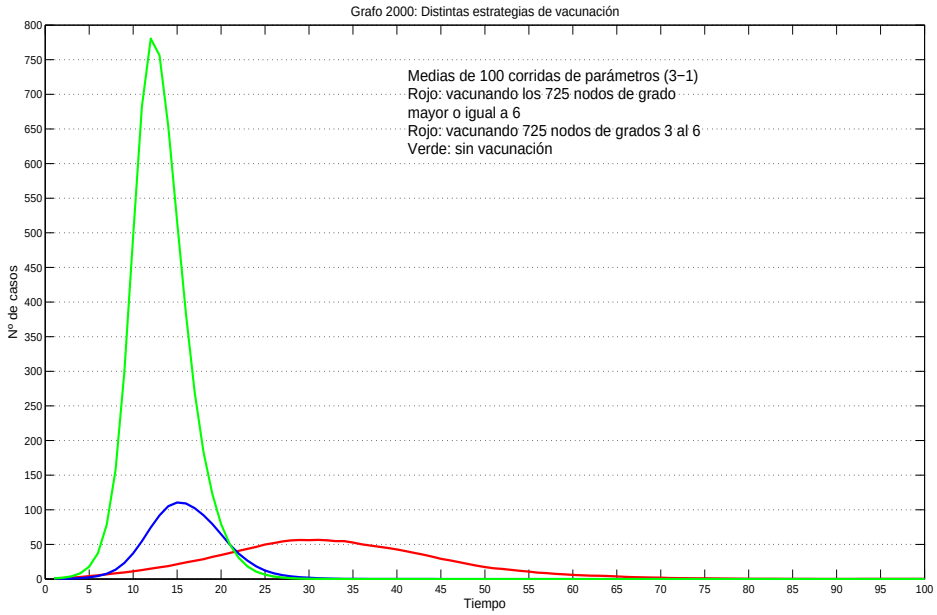


Figura 44:

Figura 45:



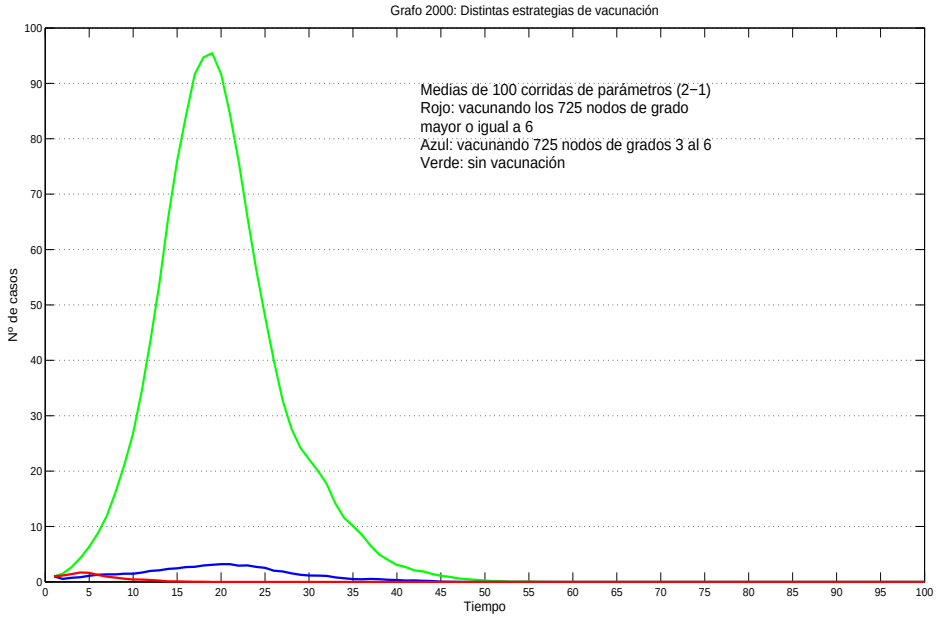


Figura 46:

La matriz de probabilidades es

$$P = \begin{pmatrix} 0{,}9970 & 0{,}0007 & 0{,}0010 & 0{,}0007 & 0{,}0006 \\ 0{,}0007 & 0{,}9910 & 0{,}0017 & 0{,}0050 & 0{,}0017 \\ 0{,}0010 & 0{,}0017 & 0{,}9948 & 0{,}0013 & 0{,}0013 \\ 0{,}0007 & 0{,}0050 & 0{,}0013 & 0{,}9922 & 0{,}0008 \\ 0{,}0006 & 0{,}0017 & 0{,}0013 & 0{,}0008 & 0{,}9957 \end{pmatrix}$$

Corremos 100 simulaciones para cada uno de los cinco casos. Esto es, se introduce un infectado (caso cero) elegido al azar entre los 10000 nodos (o sea elegido al azar en una de las 5 redes o ciudades), con un tiempo de vida como infeccioso sorteado de la distribución normal de parámetros $(\mu, 1)$ donde μ toma los valores $2, 3, 5, 10, 20$. Se toma el máximo entre dicho valor y cero para evitar tiempos de vida negativos.

Ahora cada realización es el resultado de simular una epidemia sobre una red formada por 5 grafos interconectados, al cual le asociaremos como descriptores las gráficas número de infectados vs. tiempo para cada grafo. De este modo cada realización tiene 5 atributos que son la evolución temporal de la epidemia en cada grafo.

Como las trayectorias para cada grafo en general se extinguen ya para $t = 60$ (Figura 48), truncamos los 5 atributos y los dejamos de longitud 60. Construímos árboles de clasificación, donde ahora la similitud entre realizaciones estará dada por la suma de las similitudes para cada atributo. La figura 47 muestra una realización para cada uno de los casos $(2, 1), (3, 1), (5, 1), (10, 1)$ y $(20, 1)$, mientras que la figura 48 muestra todas las realizaciones.

Esto es, dos realizaciones serán similares si la suma de las DTW (ver sección 3.3.2) entre sus correspondientes atributos es pequeña. De este modo para que dos corridas (realizaciones) sean similares, la evolución temporal de la epidemia en el grafo 1 tiene que haber sido similar en ambas corridas, lo mismo en el grafo 2, etc.

Así, no alcanza con que en dos corridas se haya observado una conducta similar en alguno o algunos de los 5 grafos, sino que pediremos un parecido global, parecido en cada grafo. Si pensamos que cada corrida es una fotografía de 5 ciudades, querremos que dos fotos sean similares si son similares en cada ciudad. En la Figura 49 se muestran dos realizaciones del tipo $(3, 1)$.

La población consiste de 500 realizaciones, cada uno con 5 atributos de longitud 60. Tomamos una muestra al azar de 100 realizaciones con la cual

Figura 47:

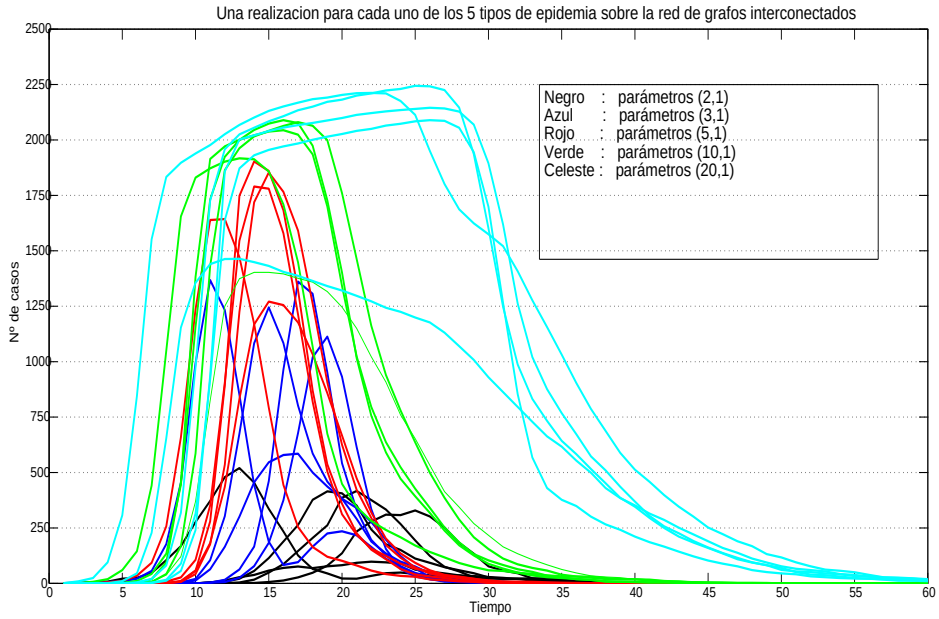


Figura 48: 100 realizaciones de todos los tipos

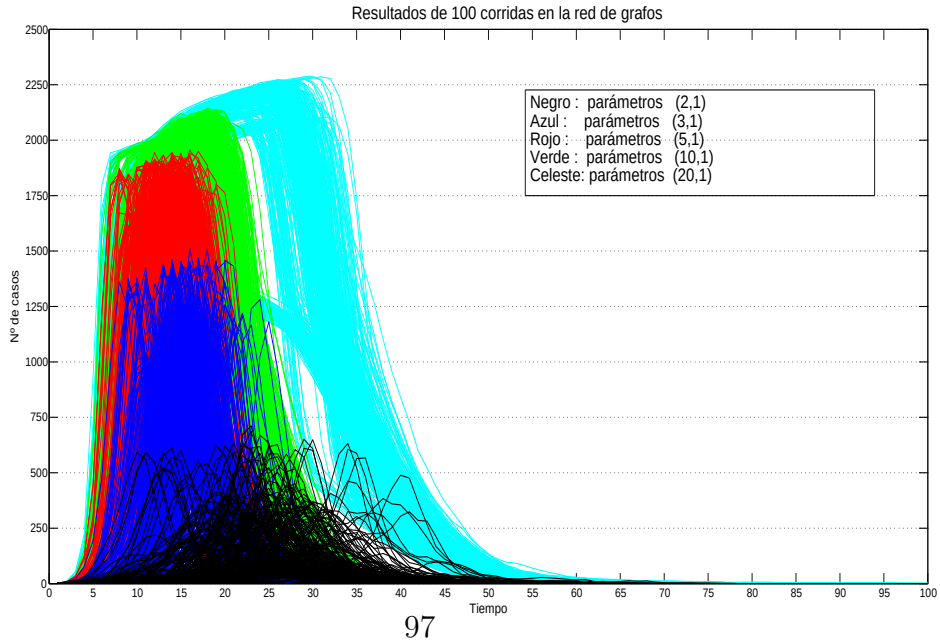
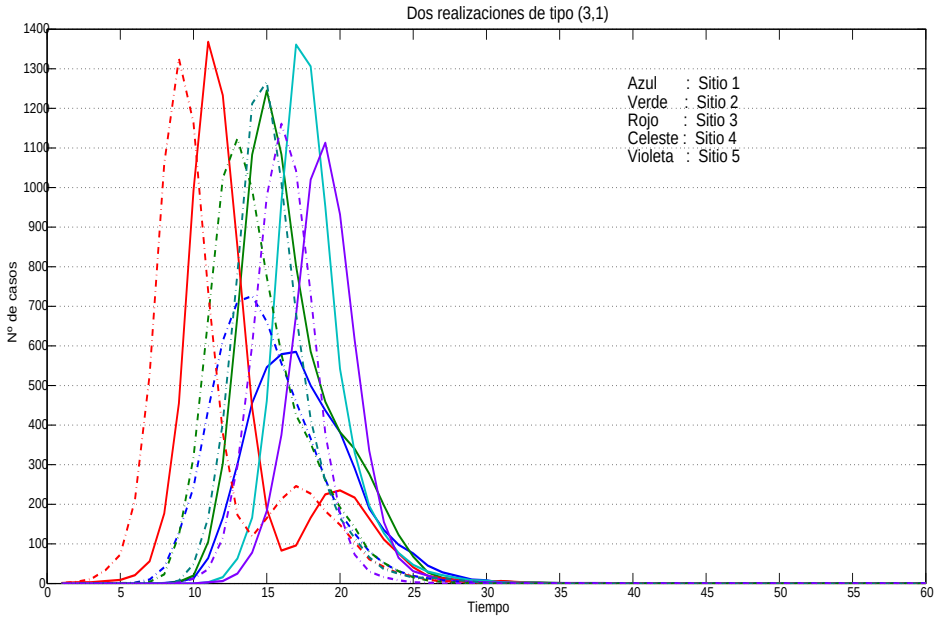


Figura 49: Dos realizaciones (3,1)



hacemos *Bagging* (sección 6.2.1). Recordemos que partimos a la población en una muestra L de entrenamiento y otra muestra T de testeo.

Hacemos n remuestras de L y con cada una de ellas fabricamos el predictor (árbol de clasificación). Luego usamos dicho árbol para clasificar a los realizaciones de la muestra T (la que permanece fija) y contamos los errores de clasificación (usamos el error duro, sección 6.2.1). Luego promediamos los errores.

Por ejemplo, para L y T con 65 y 35 realizaciones respectivamente, si repetimos 10 veces el procedimiento anterior ($n = 10$) , con el parámetro hojamin = 10 (esto es, si un nodo es tal que al partirlo resulta uno de sus hijos con menos de 10 hojas, no se le parte). El error duro promedio es de 0.028 (una realización del tipo $(3, 1)$ que es erróneamente clasificada como $(2, 1)$). Algunos de los 10 árboles que votaron se muestra en las Figuras 50 y 51.

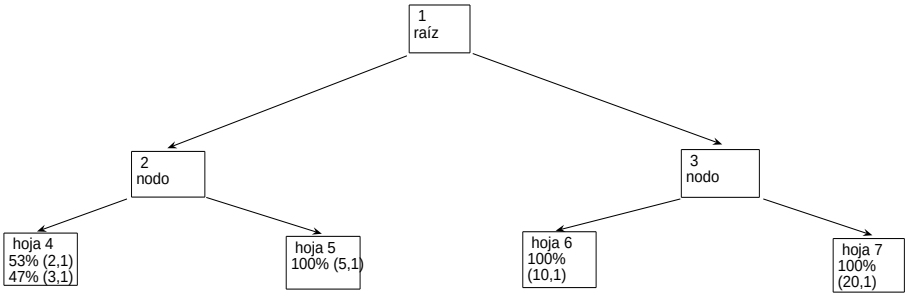
Si ahora tomamos L y T con 70 y 30 realizaciones respectivamente, $n = 15$ y hojamin=10, el error duro promedio es de 0.066 (dos realizaciones de tipo $(2, 1)$ que son clasificadas como $(3, 1)$). Las Figuras 52, 54 y 57 muestran algunos de los 10 árboles que votaron.

6.3.2. Observaciones

A pesar de lo exigüo de las muestras y de que en este caso pedimos que para considerar similares a dos realizaciones los 5 atributos de uno de ellos se parezcan uno a uno a los correspondientes atributos del otro, encontramos que el clasificador agrupa correctamente los distintos tipos de epidemias. Los errores de clasificación (estamos usando el error duro) son entre realizaciones de tipos similares, la agregación de predictores o "comité de expertos"(sección6.2.1) formado por los n clasificadores no confunde las epidemias suaves con las fuertes.

En el capítulo 7 brindaremos las conclusiones finales de este trabajo en base a los resultados experimentales obtenidos. Introduciremos además algunos puntos a tener en consideración como trabajo futuro; puntos que consideramos relevantes para enriquecer la línea de investigación aquí establecida.

Figura 50:



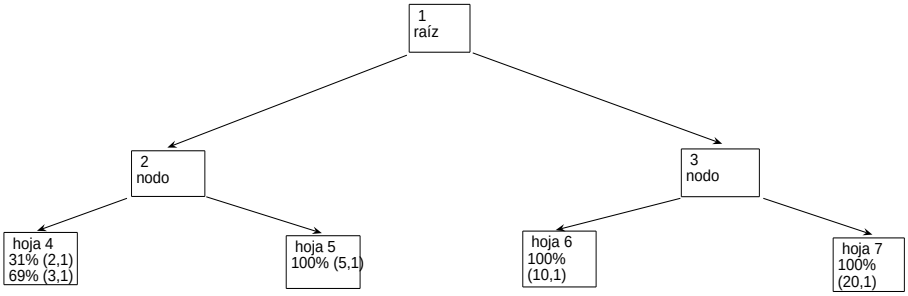


Figura 51:

Figura 52:

Figura 53: Árbol con L = 70, hojamin = 10

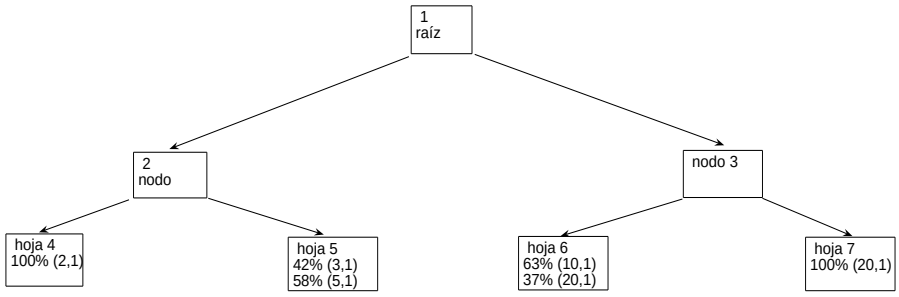


Figura 54:

Figura 55: Árbol con L = 70, hojamin = 10

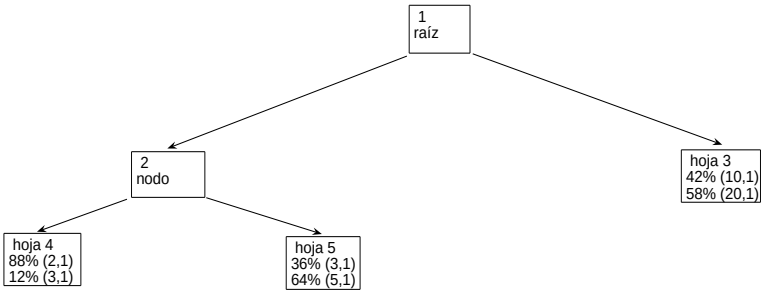


Figura 56: Árbol con L = 70, hojamin = 10

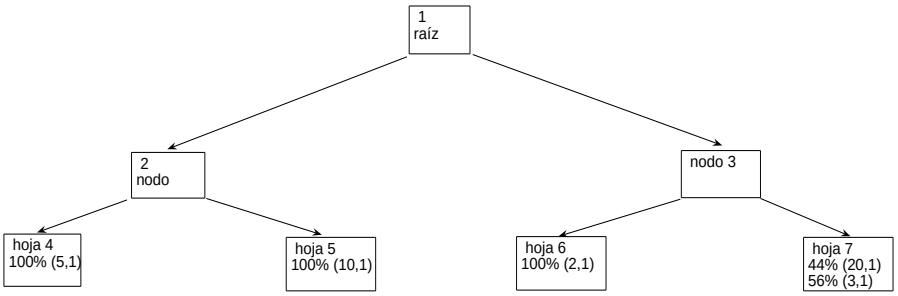


Figura 57:

7. Conclusiones

Los resultados (tanto de clasificación como de regresión) validan la idea de asociar a un proceso estocástico un vector de atributos o descriptores funcionales a los cuales se puede aplicar luego herramientas de machine learning. Esto permite hacer grupos (*clusters*) de epidemias, así como predicciones mediante regresión.

Los resultados muestran que los descriptores utilizados (distribución de los grados de los nodos infectados y numero de casos/tiempo) son adecuados para describir y caracterizar una epidemia sobre un grafo aislado, al menos si la topología de la red permanece fija durante el proceso.

La topología de la red juega un rol fundamental. Una explicación del éxito relativo de vacunar tomando como población objetivo los nodos más abundantes en vez de aquellos con mayor grado es la combinación de eficiencias globales altas y eficiencias locales bajas (sección 2.3.2).

En particular, como la mera distribución de frecuencias de los grados de los nodos no alcanza para caracterizar la topología, es necesario contar con algoritmos que permitan generar grafos aleatorios cuya estructura local refleje adecuadamente diversos tipos de redes sociales, con alta eficiencia global y local.

Asimismo las herramientas utilizadas muestran una buena performance al utilizarlas para comparar epidemias cuando en lugar de un grafo aislado tenemos varios grafos con una interconexión aleatoria.

Como trabajos a futuro quedan la aplicación y profundización de estas herramientas, así como el desarrollar algoritmos que permitan generar grafos aleatorios que no solamente tengan una distribución de grados prefijada, sino también características tales como la eficiencia global y local, permitiendo capturar mejor las diversas características de redes reales.

En la medida que las redes sociales reales puedan ser adecuadamente modeladas por grafos aleatorios, las simulaciones permitirán estudiar las versiones virtuales de epidemias reales. Esto permitirá ensayar medios de prevención y control, insumo indispensable para diseñar estrategias adecuadas.

Referencias

[1] Amaral L.A.N., Scala A., Bartélémy M., Stanley H.E. "Classes of small-world networks" *Proceedings of the National Academy of Sciences of the United States of America* published on line sep. 26 (2000)

[2] "Epidemics in a population with social structure" Math. Biosci. 140, 79-84

[3] Barrat A., Weigt M. Eur.Phys. J.B. **13**,547 (2000)

[4] Bayati M., Han J., Saberi A. "A Sequential Algorithm for Generating Random Graphs" Lecture Notes In Computer Science; Vol. 4627 *Proceedings of the 10th International Workshop on Approximation and the 11th International Workshop on Randomization, and Combinatorial Optimization. Algorithms and Techniques*

[5] Bollobás B. *Random Graphs* (Academic, London, 1985)

[6] Bradbent S.R., Hammersley, J.M. Percolation processes: I. Crystals and mazes, *Proc. Cambridge Philos. Soc.* **53**, 629-641 (1957)

[7] Chartrand G. , Lesniak L. "Graphs & Digraphs" Second Edition. Wadsworth & Brooks/Cole. Mathematics Series ISBN 0-534-06324-1

[8] Devroye L., Györfi L., Lugosi G. "A Probabilistic Theory of Pattern Recognition" Applications of Mathematics, Springer ISBN 0-387-94618-7

[9] Erdös P., Rényi, A. "On the evolution of random graphs" *Publications of the Mathematical Institute of the Hungarian Academy of Sciences* **5**,17-61 (1960)

[10] Fromont E., Artois M., Pontier D. "Cat population structure and circulation of feline viruses.ᴬcta Oecol. 17,609-620 (1996)

[11] Grimaldi Ralph P. "Matemáticas Discreta y Combinatoria "3^a Ed. Addison-Wessley Iberoamericana ISBN 0-201-65376-1

[12] Hammersley, J.M. Percolation processes: II The connective constant, *Proc. Cambridge Philos. Soc.* **53**,642-645 (1957).

[13] Hethcote H.W. "The mathematics of infectious diseases" *SIAM Review* **42**,599-653 (2000)

[14] Keogh E.J., Pazzani M.J. "Scaling up Dynamic Time Warping for Datamining Applications" *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*(pp. 285-28). New York : ACM (2000)

[15] Latora V., Marchiori M. *Physical Review Letters* Vol. 87, N^o 19 (2001)

[16] Lopez de Mántaras R. "A Distance-Based Attribute Selection Measure for Decision Tree Induction" *Kluver Academic Publishers* Boston (1991).

[17] Nerini D., Ghattas B. "Classifying densities using functional regression trees: Applications in oceanology" Computational Statistics & Data Analysis (2006), doi: 10.1016/j.csda.2006.09.028

[18] Newman M.E.J., Strogatz S.H. and Watts D.J. "Random graphs with arbitrary degree disributions and their applications" *Phys. Rev. E* **64**, 026118 (2001).

[19] Newman M.E.J. "The spread of epidemic disease on networks" *Phys. Rev. E* **66**, 016128 (2002).

[20] Newman M.E.J. "The structure and function of complex networks" *SIAM Review* **45**, 167–256 (2003).

[21] Pastor-Satorras R. and Vespignani A. "Epidemic spreading in scale-free networks" *Phys. Rev. Lett.* **86**,3200-3203 (2001)

[22] Shirley M.D.F., Rushton S.P. "The impacts of network topology on disease spread" Ecological Complexity 2 (2005) 287-299

[23] Watts D.J., Strogatz S.H. Nature (London) **393**, 440 (1998)

[24] Yamada Y., Suzuki E., Yokoi H., Takabayashi K. "Decision-tree Induction from Time-series Data Based on a Standard-example Split Test". *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003)* Washington DC, 2003 (2003)