Revista Internacional de Investigación e Innovación Tecnológica

Página principal: www.riiit.com.mx

Análisis comparativo del Índice Entrópico en Bases de Datos Utilizando la Entropía de Tsallis y Renyi en Árboles de clasificación C4.5

Comparative Analysis of Entropic Index in Databases Using Tsallis and Renyi Entropy in C4.5 Classification Trees

De la Cruz-García, J.S., Ramírez-Arellano, A.

Instituto Politécnico Nacional, MÉXICO.
https://orcid.org/0000-0003-2127-7087; https://orcid.org/0000-0002-6782-9847.
susdcg@gmail.com; aramirezar@ipn.mx

Innovación tecnológica: mejora en la clasificación del árbol de la decisión C4.5.

Área de aplicación industrial: sistemas embebidos como sistemas de clasificación por reconocimiento de patrones de productos defectuosos.

Recibido: 25 marzo 2022
Aceptado: 22 septiembre 2022

Abstract

Machine learning gives to the systems the ability to learn from experience. This is achieved through the generation of machine learning models. One of the most widely used models is supervised learning, which employs classification models that allow a computer program to learn from the input data to obtain classifications. Input and output data are labelled for classification, providing a learning base for future data processing. The C4.5 algorithm is used to obtain classification models (from a database), called decision trees. This algorithm uses the concept of entropy defined by Shannon to calculate the gain ratio. In this study Tsallis and Renyi entropies (instead of Shannon) are used to construct a decision tree. In previous works, these entropies have shown better results than Shannon. These entropies have an additional parameter *q* that is used to affect the probability distributions. This research focuses on developing a method that obtains the value of *q* that will be applied to compute the information gain ratio in the C4.5 algorithm using Tsallis and Renyi entropies. The method obtains a network representation of the database; then, the box-covering algorithm is computed to obtain the minimum number of boxes to cover the network. The calculation of the parameter *q* will depend on the minimum network coverage. The results show

that the method to obtain *q* in some cases improves the classification and in some cases does not detract from it compared to the classical C4.5 algorithm.

Keywords: Algorithm C4.5, Tsallis, Renyi, parametric entropies, entropic index.

Resumen

El aprendizaje automático brinda a los sistemas la capacidad de aprender con base a la experiencia. Esto se logra mediante la generación de modelos de aprendizaje automático. Uno de los modelos más usados es el aprendizaje supervisado, el cual emplea modelos de clasificación que permiten que un programa informático aprenda de los datos de entrada para obtener clasificaciones. Los datos de entrada y salida son etiquetados para su clasificación, proporcionando una base de aprendizaje para el procesamiento futuro de datos. El algoritmo C4.5 se utiliza para obtener modelos de clasificación (de una base de datos), llamados árboles de decisión. Este algoritmo utiliza el concepto de entropía definido por Shannon para calcular la relación de ganancia. En este estudio se usan las entropías de Tsallis y Renyi (en lugar de Shannon) para construir un árbol de decisiones. En trabajos anteriores, estas entropías han mostrado mejores resultados que Shannon. Estas entropías tienen un parámetro adicional *q* que se usa para afectar las distribuciones de probabilidad. Esta investigación se centra en desarrollar un método que obtiene el valor de *q* que se aplicará para calcular la relación de ganancia de información en el algoritmo C4.5 mediante entropías de Tsallis y Renyi. El método obtiene una representación de red a partir de una vista minable; luego, se calcula el algoritmo de cobertura de cajas para obtener el número mínimo de cajas para cubrir la red. El cálculo del parámetro *q* dependerá de la cobertura mínima de la red. Los resultados muestran que el método para obtener *q* en algunos casos mejora la clasificación y en otros no la demerita, en comparación con el algoritmo C4.5 clásico.

Palabras clave: Algoritmo C4.5, Tsallis, Renyi, entropías paramétricas, índice entrópico.

I. INTRODUCCIÓN

El aprendizaje automático es un método que se basa en datos de entrenamiento los cuales se obtienen mediante la minería de datos. Uno de los métodos más usados en minería de datos son los árboles de decisión. Un árbol de decisión se puede representar como un diagrama de flujo en forma de una estructura arborescente, donde el nodo raíz corresponde al mejor predictor, las ramas son un resultado de las pruebas y el nodo hoja representa una decisión. Los árboles de decisión se han usado como apoyo en la toma de decisiones en diferentes áreas como astronomía, finanzas, medicina, manufactura y biología, entre muchas otras [1].

El algoritmo C4.5 fue desarrollado por Ross Quinlan [2] y es una evolución del algoritmo ID3 el cual fue desarrollado por el mismo Quinlan. Los árboles C4.5 generan árboles de decisión con base de un conjunto de datos de entrenamiento utilizando el concepto de entropía de información de Shannon [3] para calcular la ganancia. La ganancia se obtiene con base a las probabilidades de los valores de los atributos de las tuplas de entrenamiento. Cada nodo divide las clases en función de la obtención de información. El

atributo con la mayor ganancia de información se utiliza como criterio de división. El algoritmo C4.5 divide en sublistas más pequeñas de manera recursiva hasta formar el árbol. Una evolución del algoritmo C4.5 es el C5.0. Se trata de una versión con la capacidad de producir reglas de clasificación más precisas generando un rendimiento superior de memoria [4].

A. Ganancia de Información

La ganancia de información se calcula para todos los atributos. La finalidad es determinar cuál de todos los atributos se usará para dividir los registros a partir de un nodo dado. Es una medida para determinar cuánta información obtiene el modelo al dividir los registros en un nodo usando un determinado atributo. Se calcula utilizando la siguiente formula [5]:

$$G(A) = S(X) - S(X)_A, \qquad \textbf{(Ec. 1)}$$

Donde S es la entropía, A es un atributo dado en un conjunto de datos X.

B. Entropía de Shannon

La entropía de Shannon se define con la siguiente formula [3]:

$$S(X) = -\sum_i^n p_i \ln(p_i), \qquad \textbf{(Ec. 2)}$$

Donde pi es la probabilidad de ocurrencia de un evento, xi que es un elemento de X que puede tomar valores [x1... xn] [6].

C. Entropía de Tsallis, Renyi y el índice Gini

Constantino Tsallis [7] y Alfred Renyi [8] propusieron entropías generalizadas donde para ambas con q = 1 se reducen a Shannon. La entropía de Tsallis se define con la siguiente formula:

$$T_q = \frac{\sum_{i=1}^n p_i - p_i^q}{q-1}, \qquad \textbf{(Ec. 3)}$$

Donde n es el número total de posibilidades del sistema, pi es la probabilidad de ocurrencia de un evento y q es el índice entrópico [9]. El índice entrópico en la entropía de Tsallis define la propiedad de sub-extensividad (q>1), super-extensividad (q<1) y la extensividad q=1. Como se ha dicho antes cuando q=1, la entropía de Tsallis se reduce a la entropía de Shannon es extensiva al igual que la entropía de Renyi. Para la entropía de Tsallis cuando q→-∞ los eventos con mínima probabilidad tienen el máximo efecto, cuando q→0 todas las diferentes probabilidades tendrán el mismo efecto y cuando q→∞ los eventos con máxima probabilidad tendrán el máximo efecto [24].

Un caso particular de Tsallis cuando el índice entrópico es igual a 2 equivale al índice de Gini. El índice de Gini se trata de otra medida de impureza. La pureza aumenta mientras se cuenta con una mejor división de una clase. Si Gini vale cero el nodo es puro lo que indica que el nodo no se puede volver a dividir. El índice Gini se define como [10].

$$\text{Gini}(L) = 1 - \sum_{i=1}^N p_i^2, \qquad \textbf{(Ec. 4)}$$

Donde pi es la frecuencia relativa de la clase i en L.

Tsallis propuso el logaritmo q definido por:

$$\ln(x) = \frac{x^{1-q} - 1}{1-q} \qquad \textbf{(Ec. 5)}$$

Para introducir una entropía física dada por [11].

$$I_q^T = -\sum_{i=1}^{N} p_i \ln_q p_i = \frac{1}{q-1}\left(1 - \sum_{i=1}^{N} p_i^q\right) \quad \textbf{(Ec. 6)}$$

Si q=2, en la ecuación (6) tenemos que:

$$I_q^T = -\frac{1}{2-1}\left(1 - \sum_{i=1}^{N} p_i^2\right) = 1 - \sum_{i=1}^{N} p_i^2 \quad \textbf{(Ec. 7)}$$

Por lo tanto, podemos decir que cuando el índice entrópico de Tsallis es igual a 2 es equivalente al índice de Gini [12].

La entropía de Renyi se define con la siguiente formula [8]:

$$R_q = \frac{1}{1-q} ln \sum_{i=1}^{n} p_i^q \,, \quad \textbf{(Ec. 8)}$$

Donde *n* es el número total de posibilidades del sistema, p_i es la probabilidad de ocurrencia de un evento y *q* es el índice entrópico [9]. Para esta entropía cuando q→0 el resultado es el logaritmo del número de eventos, cuando q→1, la entropía de Shannon es recuperada y cuando q→∞ los eventos de mayor probabilidad determinan la entropía de Renyi [23].

Con el ajuste del índice entrópico en las entropías de Tsallis y Renyi se llegan a obtener mejores resultados que calculando la ganancia con Shannon [6]. En el trabajo [13] se realizan pruebas con árboles de decisión C4.5 utilizando un método híbrido donde la ganancia de información para atributos nominales se calcula con la entropía de Shannon y para atributos numéricos se calcula con diferentes entropías paramétricas. Los enfoques híbridos se definen como híbrido Renyi (HR), híbrido Tsallis (HT), híbrido Abe (HA) e híbrido Landsberg-Vedral (HLV). La evaluación de dicho método fue mediante el análisis del área bajo la curva ROC (AURC) y mediante el método de Exactitud (Acurracy). El método se comparó con diferentes algoritmos como SVM (Support Vector Machines), KNN (k-nearest neighbors), red neuronal, logit (Logistic Regression), TEIM (Tsallis Entropy Information Metric), ID3 y el C4.5 clásico concluyendo que HT y HR generan arboles de decisión menos complejo en comparación con el algoritmo TEIM con AURC similar. El HR presento mejores resultados para conjuntos de datos nominales, numéricos y mixtos ya que obtuvo un AURC igual o superior al C4.5 clásico.

En otro estudio [14] usaron las entropías parametrizadas de Tsallis y Renyi para calcular la ganancia de información en el algoritmo C4.5 en servicios de telecomunicaciones con la finalidad de analizar el problema de pérdida de clientes. Los árboles se comparan con los métodos de redes neuronales, máquinas vectoriales de soporte y regresión logística. Los resultados mostraron que los árboles basados en *q* arrojaron mejores resultados debido a que *q* aumenta las posibilidades de clasificación ayudando a generar modelos con una mayor capacidad.

De igual forma, en el estudio [15] se diseñan arboles de decisión C4.5 mediante el uso de las entropías de Shannon, Renyi y Tsallis para el sistema computacional tolerante a intrusiones. Las bases de datos usadas describen el tráfico de una red con diferentes tipos de intrusiones. Los resultados mostraron que utilizando las entropías paramétricas de Tsallis y Renyi se construyen arboles de decisión más eficientes y compactos siempre y cuando los valores de los parámetros sean adecuados.

En otra investigación [6] se generaron árboles de decisión basados en las entropías de Tsallis

y Renyi. Se observó una mejora usando Renyi de hasta el 88.5 ± 2.4% en comparación con el 81.4 ± 4.1% que se muestran usando Shannon. Así mismo, se observó una mejora usando Tsallis del 91.3±3.5% en comparación del 83.8±5.3 que se observó usando Shannon.

El algoritmo propuesto en [16] construye árboles de clasificación, basados en la entropía de Tsallis y Rényi que son aplicados al problema de rotación de clientes en la industria de las telecomunicaciones. Las medidas de calidad de los árboles obtenidos se comparan para diferentes valores del parámetro *q*. Se utilizó el algoritmo de árbol de decisión C4.5 para la clasificación y se analizó el rendimiento de ambas en el caso de que el conjunto de datos de aprendizaje estuviera equilibrado o desequilibrado. Los resultados demostraron que las entropías de Tsallis y Rényi, con parámetros *q* adecuados, pueden conducir a árboles de decisión compactos y eficientes, con medidas de alta precisión. Tsallis proporcionó una mejor generalización ya que los árboles resultantes no eran tan complejos como en el caso de Rényi. Se concluyó que el uso de Tsallis y Rényi hace que el análisis sea más flexible que el enfoque estándar con Shannon, ya que permite la exploración de la compensación entre la probabilidad de diferentes clases y la ganancia de información general.

Podemos concluir que al usar entropías paramétricas para calcular la ganancia se obtienen arboles de decisión con mayor precisión y eficiencia. Sin embargo, una limitante para su uso es fijar el valor del índice entrópico. En los trabajos antes mencionados el índice entrópico tiene un rango limitado de valores, es decir, es necesario buscar el valor de *q* de manera sistemática hasta encontrar el valor que genere mejores resultados. Si bien, se han implementado métodos para encontrar el valor de *q* [17] en diferentes áreas de conocimiento, aun no se ha trabajado en un método que sea aplicable a arboles de decisión.

Por lo anterior, se propone un método para obtener el valor de *q* en entropías paramétricas y aplicarlo al cálculo de la ganancia para la generación de árboles de decisión C4.5 en distintos conjuntos de datos. Las secciones restantes describen la metodología de la investigación seguido de los resultados y discusiones. Por último, presentaremos las conclusiones.

II. METODOLOGÍA

Dentro de la metodología recopilamos y evaluamos los datos numéricos con la finalidad de identificar patrones y correlaciones mediante técnicas estadísticas para realizar la comparación de los resultados. Por lo anterior el tipo de estudio que se realiza es cuantitativo.

A. Transformación de una base de datos a una red compleja

Como primer paso, desarrollamos un método que transforma una red a partir de una vista minable. Este método consiste en tomar una tupla y por cada nombre del atributo concatenar el valor que le corresponde. Considere los valores de la Tabla 1, para obtener los nodos de la primera tupla concatene los nombres de los atributos con el valor correspondiente, por ejemplo, NOMBRE.Jonh Smit (nodo 1), VECINDARIO.Clairemont (nodo 2) y TELEFONO.754-3010 (nodo 3). De esta forma se obtienen 3 nodos de la primera tupla. Para la segunda tupla se obtienen 2 nodos NOMBRE.Jane Williams (nodo 4), y TELEFONO.387-9827 (nodo 5). VECINDARIO.Clairemont (nodo 3) ya ha sido creado en la tupla 1. Repita el mismo método para las tuplas restantes.

El siguiente paso es generar las aristas de los nodos resultantes. Para esto, una los nodos

que se relacionan con los valores de cada una de las tuplas. En el ejemplo, NOMBRE.Jonh Smit tiene una arista a VECINDARIO.Clairemont y a su vez, VECINDARIO.Clairemont tiene una arista con TELEFONO.754-3010. Los nodos NOMBRE.Jonh Smit y TELEFONO.754-3010 pertenecen al mismo registro, por lo que deben estar conectados. Este paso se repite para los nodos generados en las tuplas restantes dando como resultado la red que se muestra en la Fig. 1.

Tabla 1. Ejemplo de una vista minable.

Nombre	Vecindario	Teléfono
Jonh Smit	Clairemont	754-3010
Jane Williams	Clairemont	387-9827
Harry Williams	Clairemont	387-9827

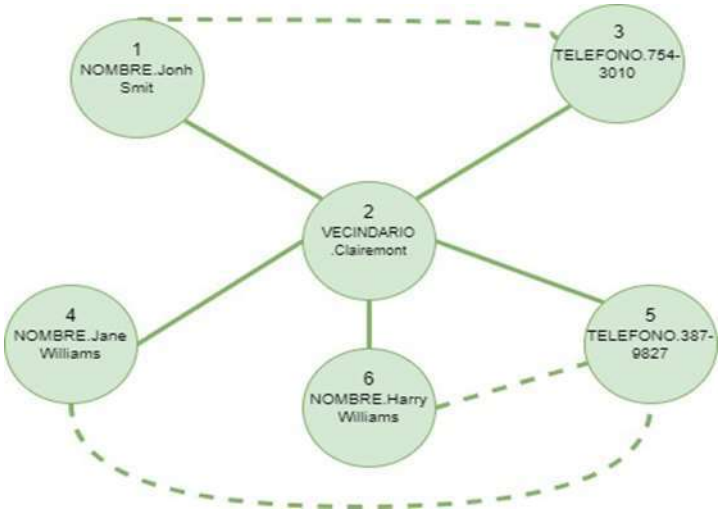

Figura 1. Representación de una red compleja a partir de atributos de una vista minada.

En la Fig. 2 se muestra un ejemplo de una red compleja transformada con los pasos descritos a partir de la base de datos "Vehicle" del repositorio UCI [18].

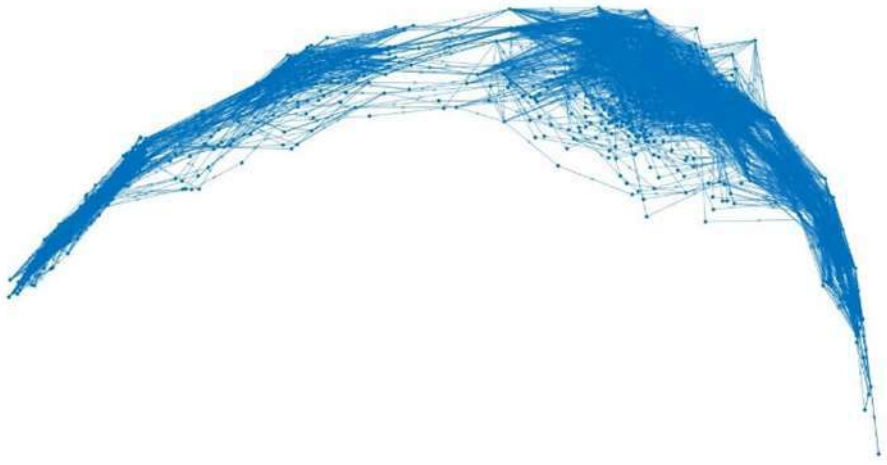

Figura2. Representación de una red compleja a partir del conjunto de datos Vehicle.

B. Cálculo del índice entrópico

Después de obtener la representación de un conjunto de datos en una red compleja, el siguiente paso es aplicar el algoritmo de cubrimiento de cajas para encontrar el valor de *q*. Considere la Fig. 3. Para calcular el número mínimo de cajas N_b de longitud $l=1$, observe que necesitamos el mismo número de nodos de la red, que cajas para cubrirla, $N_b\ (l = 1) = 4$. Ahora considere cubrir toda la red con una caja, para ello la longitud de la caja deberá ser el diámetro de la red (D) más uno, por lo tanto, $N_b\ (l = D + 1) = 1$ para la red de la Fig. 3. Estos dos casos extremos son triviales no, así los valores de $N_b(l)$ para $l = [2, D]$.

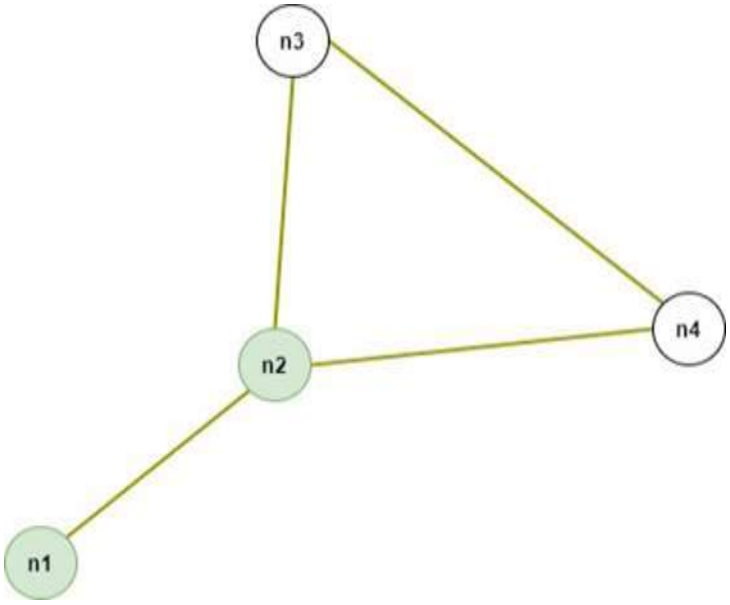

Figura 3. Ejemplo del cubrimiento de cajas con l=2, donde los nodos del mismo color pertenecen a la misma caja. Imagen basada en [19].

Para calcular el número mínimo de cajas de diámetro $l=[2, D]$, más allá de los dos casos triviales mencionados con anterioridad, se usa el algoritmo de cubrimiento de cajas [20]. Para este ejemplo tomemos l=2. A partir de la red original (G), ver Figura 3, calculamos una red dual (G'), ver Figura 4. Dada una distancia l; dos nodos i, j, en la red dual, están conectados si la distancia entre $l_{ij} >= l$. Iniciando a partir del nodo *n4* la distancia a *n1* es de 2 en G, por lo tanto, se conectan en *G'*. Note que el camino más corto desde *n4* hasta el resto de los nodos es uno, por lo tanto, no se conectan en la red dual. Ahora, elegimos el nodo *n3* como el nodo inicial. La distancia a *n1* es 2, por lo tanto, la conexión se coloca en *G'* como se muestra en la Fig. 4. Finalmente partiendo del nodo n2, no más conexiones se agregan a *G'*. En el siguiente paso se colorean los nodos con dos reglas: dos nodos conectados en *G'* no pueden tener el mismo color y usar el mínimo número de colores para el coloreado de nodos. El número de colores de *G'* representa el número mínimo de cajas $N_b(l)$ para un l dado. Los nodos de G, del mismo color, pertenecen a la misma caja. El procedimiento se repite hasta que $l = D + 1$ [19].

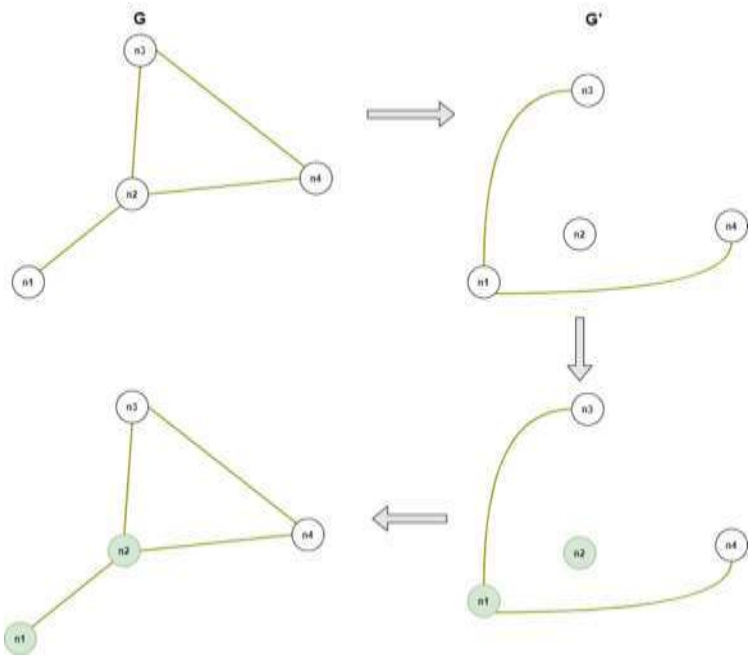

Figura 4. Cobertura de la red para un tamaño de caja l = 2. El número de cajas en esta red es Nb (2) = 2. Imagen basada en [19].

Por lo tanto, el cálculo del parámetro entrópico depende de la cobertura mínima de la red y se calcula de la siguiente forma [17]:

$$q_l = \frac{lN_b(l)}{D\sum_{l=2}^{D} N_b(l)}, \qquad \textbf{(Ec. 9)}$$

Donde, *Nb(l)*, *l* y *D*, ya se han descrito con anterioridad. Note que *q* es un vector de longitud *D* puesto que *l=[2, D]*.

A continuación, se propone el índice *α* de una red basado en el trabajo [21]:

$$\alpha_{l,i} = \frac{|G_i| innerdeg(G_i)}{n\sum_{i=1}^{N_b} innerdeg(G_i)}, \qquad \textbf{(Ec. 10)}$$

Donde *D* es el diámetro de la red, $N_b(l)$ es el número de cajas para cubrir la red para un *l* dado, *n* es el total de nodos de la red y $innerdeg(G_i)$ es el promedio de los enlaces entre los nodos que están en la caja G_i. Por ejemplo, en la Fig. 3, se observa que hay dos cajas (nodos verdes y blancos) $G_1=\{n1,n2\}$ y $G_2=\{n3,n4\}$. Si extraemos estos dos subgrafos G_1 y G_2 las conexiones de los nodos *n2*, *n3* y *n2*, *n4* se eliminan, por lo que $innerdeg(G_1)=1$ y $innerdeg(G_2)=2$. Los detalles del cálculo se describen en [21].

De igual forma *β* se define [21]:

$$\beta_{l,i} = \frac{outerdeg(G_i)l}{D\sum_{i=1}^{N_b} outerdeg(G_i)}, \qquad \textbf{(Ec. 11)}$$

Donde *l*, *D*, $N_b(l)$ y *l* se describieron con anterioridad. $outerdeg(G_i)$ es el número de enlaces entre las subredes G_i. Por ejemplo, considere la Fig. 5, se observa el proceso de renormalización [22] de la red de la Figura 3 para calcular $outerdeg(G_i)$ que consiste en convertir los nodos de una misma caja (G_i) en supernodos para luego calcular el grado de estos supernodos que serán el $outerdeg(G_i)$, en nuestro ejemplo $outerdeg(G_1)= outerdeg(G_2)=1$

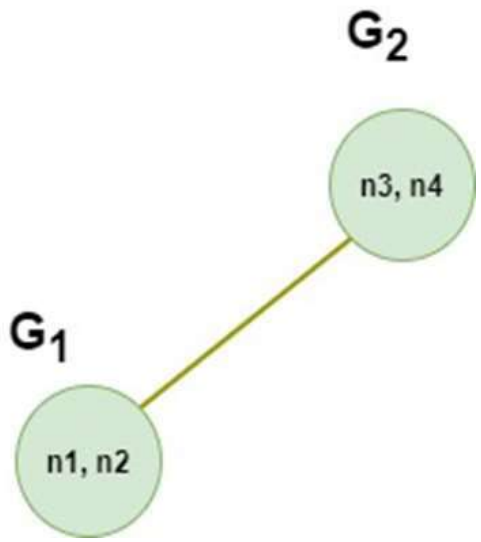

Figura 5. Proceso de renormalización para obtener el outerdeg(Gi) de la red de la Fig. 3.

C. Índice entrópico para una base de datos

Basados en q, α, y β, definidas para las redes complejas, la primera aproximación para el índice entrópico de base de datos que presentamos se define de la siguiente forma:

$$q_{\beta} = \frac{|q|}{|\beta|}, \qquad \textbf{(Ec. 12)}$$

Donde $|\beta|$ es el promedio de las entradas de la matriz $\beta_{l,i}$ de la ecuación 11.

De forma similar presentamos la segunda aproximación:

$$q_{\beta^{-1}} = \frac{|\beta|}{|q|}, \qquad \textbf{(Ec. 13)}$$

Donde $|\beta|$ es el promedio de las entradas de la matriz $\beta_{l,i}$ de la ecuación 11.

La tercera aproximación se presenta a continuación:

$$q_{\alpha} = \frac{|q|}{|\alpha|}, \qquad \textbf{(Ec. 14)}$$

Donde $|q|$ es el promedio de las entradas del vector q_l de la ecuación 9, $|\alpha|$ es el promedio de las entradas de la matriz $\alpha_{l,i}$ de la ecuación 10.

Por último, presentamos la cuarta aproximación de la siguiente forma:

$$q_{\alpha^{-1}} = \frac{|\alpha|}{|q|}, \qquad \textbf{(Ec. 15)}$$

Donde $|\alpha|$ es el promedio de las entradas de la matriz $\alpha_{l,i}$ de la ecuación 10, $|q|$ es el promedio de las entradas del vector q_l de la ecuación 9.

III. RESULTADOS Y DISCUSIÓN

Para llevar a cabo los experimentos se consideraron 13 conjuntos de datos del repositorio UCI [18]. Los experimentos se realizaron sin discretizar los datos. La descripción de los conjuntos de datos del repositorio UCI antes mencionados se muestran en la Tabla 2.

La columna “Tipo” especifica si los atributos del conjunto de datos son Nominales (N), Mixtos (M) o Numéricos (U). La columna "Balanceado" indica la distribución de cada etiqueta de la clase. Las representaciones de las redes para los conjuntos de datos antes mencionados se obtuvieron con el método anteriormente descrito. Así mismo se incluye el número de nodos y los arcos de las redes.

Tabla 2. Descripción de los conjuntos de datos del repositorio UCI.

Conjunto de datos	Tareas asociadas	Instancias	Atributos	Tipo	Clases	Balanceado	Nodos	Arcos
Breast-cancer	Clasificación	699	9	N	2	No	737	1276
Car	Clasificación	1728	6	N	4	Si	25	70
Chess	Clasificación	3196	36	N	36	Si	58	346
Cmc	Clasificación	1473	9	M	3	No	74	264
Glass	Clasificación	214	10	U	7	No	1159	1743
Haberman	Clasificación	306	3	U	2	No	94	395
Hayes	Clasificación	160	5	N	3	No	150	186
Image	Clasificación	2310	19	U	7	Si	12705	24411
Letter	Clasificación	20000	16	U	16	Si	282	2700
Scale	Clasificación	625	4	N	3	No	23	90
Vehicle	Clasificación	946	18	U	4	Si	1434	8064
Wine	Clasificación	178	13	U	3	No	1279	2239
Yeast	Clasificación	1484	9	M	10	No	1917	4907

La Tabla 3 describe los valores de los parámetros entrópicos $|q|$, q_α, q_α^{-1}, q_β y q_β^{-1} *que* se obtuvieron con las ecuaciones 9, 12, 13, 14 y 15 respectivamente.

Tabla 3. Resultados del cálculo de los parámetros entrópicos.

Conjunto de datos	$\|q\|$	α	β	q_β	q_β^{-1}	q_α	q_α^{-1}
Breast-cancer	0.013	0.007	0.007	1.926	0.519	1.752	0.571
Car	0.042	0.137	0.12	0.347	2.883	0.303	3.298
Chess	0.035	0.065	0.057	0.617	1.622	0.534	1.874
Cmc	0.020	0.066	0.060	0.328	3.051	0.297	3.367
Glass	0.01	0.006	0.005	1.910	0.523	1.798	0.556
Haberman	0.068	0.025	0.02	3.343	0.299	2.688	0.372
Hayes	0.044	0.020	0.017	2.641	0.379	2.208	0.453
Image	0.005	0.001	0.001	7.112	0.141	6.757	0.148
Letter	0.007	0.034	0.032	0.227	4.400	0.214	4.666
Scale	0.066	0.122	0.1	0.663	1.507	0.544	1.838
Vehicle	0.005	0.007	0.006	0.832	1.202	0.788	1.269
Wine	0.011	0.005	0.005	2.318	0.431	2.164	0.462
Yeast	0.017	0.003	0.002	7.495	0.133	6.747	0.148

El promedio del Área Bajo la Curva ROC (AROC) resultante para los 13 conjuntos de datos se muestra en la Tabla 4. En esta misma tabla se muestran los valores que fueron estadísticamente diferentes al comparar la entropía de Shannon (S) contra la de Renyi (R) para calcular la ganancia en el algoritmo C4.5.

Tabla 4. AROC comparada con el C4.5 con la entropía de Shannon vs entropía de Renyi.

Conjunto de datos	S	q_β	q_β^{-1}	q_α	q_α^{-1}
Breast-cancer	0.964	0.964	0.967	0.962	0.965
Car	0.983	0.983	0.500	0.983	0.500
Chess	0.993	0.992	0.990	0.993	0.990
Cmc	0.691	0.687	0.711*	0.686	0.705*
Glass	0.794	0.825*	0.781	0.818	0.770
Haberman	0.579	0.570	0.500	0.580	0.500
Hayes	0.892	0.900	0.881	0.898	0.881
Image	0.998	0.990	0.999	0.990	0.999
Letter	0.987	0.980	0.971	0.980	0.971
Scale	0.845	0.841	0.851	0.838	0.851
Vehicle	0.762	0.776	0.761	0.780*	0.753
Wine	1.000	1.000	0.986	1.000	0.993
Yeast	0.735	0.787*	0.707	0.789*	0.707
Promedio	0.863	0.869	0.816	0.869	0.814

*Estadísticamente diferentes con S con $p<0.05$.

El promedio de la Exactitud y los valores que fueron estadísticamente diferentes comparando Shannon con Renyi, se muestran en la Tabla 5.

Tabla 5. Exactitud comparada con el C4.5 con la entropía de Shannon vs entropía de Renyi.

Conjunto de datos	S	q_β	q_β^{-1}	q_α	q_α^{-1}
Breast-cancer	0.949	0.956	0.951	0.954	0.949
Car	0.923	0.925	0.700	0.926	0.700
Chess	0.982	0.982	0.979	0.982	0.979
Cmc	0.514	0.496	0.532*	0.494	0.530*
Glass	0.676	0.699	0.659	0.686	0.647
Haberman	0.710	0.743*	0.727	0.742*	0.726*
Hayes	0.794	0.790	0.794	0.782	0.794
Image	0.968	0.945	0.956	0.945	0.956
Letter	0.880	0.880	0.812	0.880	0.812
Scale	0.778	0.779	0.787	0.776	0.780
Vehicle	0.723	0.732	0.717	0.736*	0.710
Wine	0.932	0.918	0.940*	0.921	0.941*
Yeast	0.564	0.578*	0.526	0.582*	0.526
Promedio	0.800	0.802	0.775	0.800	0.773

*Estadísticamente diferentes con S con $p<0.05$.

De la misma forma, en la Tabla 6 se muestra el promedio del AROC usando la entropía de Tsallis (T).

Tabla 6. Aroc comparada con el C4.5 con la entropía de Shannon vs entropía de Tsallis.

Conjunto de datos	S	q_β	q_β^{-1}	q_α	q_α^{-1}
Breast-cancer	0.964	0.953	0.964	0.954	0.959
Car	0.983	0.981	0.983	0.981	0.983
Chess	0.993	0.992	0.990	0.992	0.990
Cmc	0.691	0.701	0.639	0.695	0.638
Glass	0.794	0.758	0.829*	0.756	0.820
Haberman	0.579	0.500	0.612*	0.500	0.608
Hayes	0.892	0.833	0.900	0.835	0.898
Image	0.998	0.500	0.995	0.500	0.995
Letter	0.987	0.979	0.870	0.979	0.864
Scale	0.845	0.845	0.849	0.844	0.851
Vehicle	0.762	0.748	0.730	0.743	0.710
Wine	1.000	0.983	0.993	0.958	0.997
Yeast	0.735	0.497	0.707	0.497	0.710
Promedio	0.863	0.790	0.851	0.787	0.848

*Estadísticamente diferentes con S con $p<0.05$.

El promedio de la Exactitud y los valores que fueron estadísticamente diferentes comparando Shannon con Tsallis, se muestran en la Tabla 7.

Tabla 7. Exactitud comparada con el C4.5 con la entropía de Shannon vs entropía de Tsallis.

Conjunto de datos	S	q_β	q_β^{-1}	q_α	q_α^{-1}
Breast-cancer	0.949	0.949	0.945	0.948	0.946
Car	0.923	0.918	0.922	0.919	0.922
Chess	0.982	0.982	0.978	0.982	0.978
Cmc	0.514	0.525	0.473	0.524	0.476
Glass	0.676	0.633	0.700	0.628	0.689
Haberman	0.710	0.712*	0.713	0.712*	0.709
Hayes	0.794	0.679	0.717	0.688	0.696
Image	0.968	0.143	0.961	0.143	0.962
Letter	0.880	0.878	0.361	0.878	0.350
Scale	0.778	0.778	0.780	0.778	0.776
Vehicle	0.723	0.714	0.658	0.710	0.646
Wine	0.932	0.909	0.940*	0.901	0.940*
Yeast	0.564	0.314	0.518	0.314	0.516
Promedio	0.800	0.703	0.744	0.702	0.739

*Estadísticamente diferentes con S con $p<0.05$.

Haciendo una comparativa de los valores obtenidos en las Tablas 4 y 6 podemos determinar cuál es la mejor aproximación del AROC comparando las entropías de Shannon (S), Renyi (R) y Tsallis (T). El resultado de esta comparación se muestra en la Tabla 8.

Tabla 8. AROC con mejor aproximación comparando las entropías de Shannon, Renyi y Tsallis.

Conjunto de datos	S	*Mejor Aproximación Renyi*	*Parámetro con mejor aproximación Renyi*	*Mejor Aproximación Tsallis*	*Parámetro con mejor aproximación Tsallis*
Breast-cancer	0.959	0.967	q_{β}^{-1}	0.964	q_{β}^{-1}
Car	0.981	0.983	q_{β}, q_{α}	0.983	$q_{\beta}^{-1}, q_{\alpha}^{-1}$
Chess	0.993	0.993	q_{α}	0.992	q_{β}, q_{α}
Cmc	0.691	0.711*	$q_{\beta}^{-1}, q_{\alpha}^{-1}$	0.701	q_{β}
Glass	0.794	0.825*	q_{β}	0.829*	q_{β}^{-1}
Haberman	0.574	0.580	q_{α}	0.612*	q_{β}^{-1}
Hayes	0.892	0.900	q_{β}	0.900	q_{β}^{-1}
Image	0.998	0.999	$q_{\beta}^{-1}, q_{\alpha}^{-1}$	0.995	$q_{\beta}^{-1}, q_{\alpha}^{-1}$
Letter	0.987	0.980	q_{β}, q_{α}	0.979	q_{β}, q_{α}
Scale	0.845	0.851	$q_{\beta}^{-1}, q_{\alpha}^{-1}$	0.851	q_{α}^{-1}
Vehicle	0.762	0.780*	q_{α}	0.748	q_{β}
Wine	1.000	1.000	q_{β}, q_{α}	0.997	q_{α}^{-1}
Yeast	0.735	0.789*	q_{α}	0.710	q_{α}^{-1}

*Estadísticamente diferentes con S con p<0.05.

De igual manera, hacemos una comparativa de los valores obtenidos en las Tablas 5 y 7 para determinar la mejor Exactitud comparando Shannon (S), Renyi (R) y Tsallis (T). El resultado de esta comparación se muestra en la Tabla 9.

Tabla 9. Exactitud con mejor aproximación comparando las entropías de Shannon, Renyi y Tsallis.

Conjunto de datos	S	*Mejor Aproximación Renyi*	*Parámetro con mejor aproximación Renyi*	*Mejor Aproximación Tsallis*	*Parámetro con mejor aproximación Tsallis*
Breast-cancer	0.949	0.956	$_{q}\beta$	0.949	q_{β}
Car	0.923	0.925	$_{q}\beta$	0.922	$q_{\alpha}^{-1}, q_{\beta}$
Chess	0.982	0.982	q_{α}, q_{β}	0.982	q_{α}, q_{β}
Cmc	0.514	0.532*,0.530*	$q_{\alpha}^{-1}, q_{\beta}^{-1}$	0.525	q_{β}
Glass	0.676	0.699	$_{q}\beta$	0.700	q_{β}^{-1}
Haberman	0.710	0.743*,0.742*	q_{α}, q_{β}	0.712*	q_{α}, q_{β}
Hayes	0.794	0.794	$q_{\alpha}^{-1}, q_{\beta}^{-1}$	0.717	q_{β}^{-1}
Image	0.968	0.956	$q_{\alpha}^{-1}, q_{\beta}^{-1}$	0.962	q_{α}^{-1}
Letter	0.880	0.880	q_{α}, q_{β}	0.878	q_{α}, q_{β}
Scale	0.778	0.787	q_{β}^{-1}	0.780	q_{β}^{-1}
Vehicle	0.723	0.736*	q_{α}	0.714	q_{β}
Wine	0.932	0.940*,0.941*	$q_{\alpha}^{-1}, q_{\beta}^{-1}$	0.940*	$q_{\alpha}^{-1}, q_{\beta}^{-1}$
Yeast	0.564	0.578*,0.582*	q_{α}, q_{β}	0.518	q_{β}^{-1}

*Estadísticamente diferentes con S con p<0.05.

En la Tabla 10 podemos ver los resultados obtenidos usando el índice de Gini comparados con Shannon mediante el AROC y la Exactitud.

Tabla 10. AROC y exactitud del árbol de decisión de Shannon (s) y árbol de decisión de Gini (g).

Conjunto de datos	S_{AROC}	G_{AROC}	$S_{ACURRACY}$	$G_{ACURRACY}$
Breast-cancer	0.959	0.963	0.949	0.949
Car	0.981	0.981	0.923	0.904
Chess	0.993	0.988	0.982	0.988
Cmc	0.691	0.58*	0.514	0.482
Glass	0.794	0.712*	0.676	0.635
Haberman	0.574	0.52*	0.710	0.721*
Hayes	0.892	0.871	0.794	0.830*
Image	0.998	0.988	0.968	0.930
Letter	0.987	0.962*	0.880	0.819
Scale	0.845	0.866*	0.778	0.794*
Vehicle	0.762	0.71*	0.723	0.669
Wine	1.000	0.932*	0.932	0.871
Yeast	0.735	0.728	0.564	0.508

*Estadísticamente diferentes con S con $p<0.05$.

IV. CONCLUSIONES

El método estándar para encontrar el valor del índice entrópico para arboles de decisión usando Tsallis o Renyi es mediante prueba y error. En minería de datos esta técnica no es viable ya que se cuenta con miles y millones de registros y encontrar el valor a prueba y error puede llevar mucho tiempo.

Es posible encontrar los valores apropiados a usar en el índice entrópico de Renyi o Tsallis mediante el método propuesto, el cual consiste en representar una base de datos como una red compleja y basándonos en esta red podemos obtener 4 posibles valores: q_β, q_β^{-1}, q_α y q_α^{-1}.

Los resultados obtenidos al aplicar el método a trece bases de datos muestran que con los valores q_β, q_β^{-1}, q_α y q_α^{-1} obtenemos mejores clasificaciones AUROC usando Renyi para 4 bases de datos, mientras que con Tsallis obtenemos una mejor clasificación en 2 bases de datos. Usando la Exactitud (Acurracy) obtenemos mejores clasificaciones para 5 bases de datos con Renyi, mientras que con Tsallis se obtiene mejora en dos bases de datos. En ambos casos para las bases de datos restantes el resultado no se vio demeritado.

Los árboles de decisión también se construyeron usando el índice de Gini. Los resultados no muestran una mejora significativa en la clasificación mediante el AUROC y la Exactitud.

De forma general podemos concluir que el método propuesto funciona para encontrar el valor del índice entrópico entre los parámetros q_β, q_β^{-1}, q_α y q_α^{-1} y de esta manera poder obtener una mejora en la clasificación de los árboles C4.5.

Ya que los árboles de decisión se han usado en distintas áreas como la medicina, fabricación y producción, análisis financiero, astronomía y biología molecular como herramienta en la toma de decisiones, el método propuesto en esta investigación puede ayudar a mejorar dicha toma de decisiones.

En esta investigación preparamos el camino para probar la efectividad del método en otras técnicas de minería de datos como K-means, MST genérico, MST de Kruskal y algoritmos para la reducción de dimensiones. También queda pendiente explorar la relación que pudiera tener los resultados con las características de los conjuntos de datos, así como, la representación de las redes complejas y la distribución de la probabilidad.

V. REFERENCIAS

[1] J. P. Jiawei Han, Micheline Kamber, *Data Mining: Concepts and Techniques*. 2011.

[2] S. L. Salzberg, "C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993", *Mach Learn*, vol. 16, no. 3, pp. 235–240, Sep. 1994, doi: 10.1007/BF00993309.

[3] Claude Elwood Shannon; Warren Weaver, *The mathematical theory of communication*. 1964.

[4] E. Ahmadi, G. R. Weckman, and D. T. Masel, "Decision making model to predict presence of coronary artery disease using neural network and C5.0 decision tree", *J Ambient Intell Humaniz Comput*, vol. 9, no. 4, pp. 999–1011, Aug. 2018, doi: 10.1007/s12652-017-0499-z.

[5] A. Alharan, Abbas & Alsagheer, Radhwan & Al-Haboobi, "Popular Decision Tree Algorithms of Data Mining Techniques: A Review", *International Journal of Computer Science and Mobile Computing*, vol. 6, pp. 133–142, 2017.

[6] T. Maszczyk and W. Duch, "Comparison of Shannon, Renyi and Tsallis Entropy Used in Decision Trees", in *Artificial Intelligence and Soft Computing – ICAISC 2008*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 643–651. doi: 10.1007/978-3-540-69731-2_62.

[7] C. Tsallis, RenioS. Mendes, and A. R. Plastino, "The role of constraints within generalized nonextensive statistics", *Physica A: Statistical Mechanics and its Applications*, vol. 261, no. 3–4, pp. 534–554, Dec. 1998, doi: 10.1016/S0378-4371(98)00437-3.

[8] A. Rényi, *On measures of entropy and information. In Proceedings of Berkeley Symposium on Mathematical Statistics and Probability*. California, 1961.

[9] Y. Wang and S.-T. Xia, "Unifying attribute splitting criteria of decision trees by Tsallis entropy", in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2017, pp. 2507–2511. doi: 10.1109/ICASSP.2017.7952608.

[10] S. R. and D. D. Turban Efraim, *Business Intelligence and Analytics: Systems for Decision Support*, 10th ed. 2014.

[11] C. Tsallis, "Possible generalization of Boltzmann-Gibbs statistics", *J Stat Phys*, vol. 52, no. 1–2, pp. 479–487, Jul. 1988, doi: 10.1007/BF01016429.

[12] Y. Wang, C. Song, and S.-T. Xia, "Improving decision trees by Tsallis Entropy Information Metric method", in *2016 International Joint Conference on Neural Networks (IJCNN)*, Jul. 2016, pp. 4729–4734. doi: 10.1109/IJCNN.2016.7727821.

[13] A. R. Arellano, J. Bory-Reyes, and L. M. Hernandez-Simon, "Statistical Entropy Measures in C4.5 Trees", *International*

Journal of Data Warehousing and Mining, vol. 14, no. 1, pp. 1–14, Jan. 2018, doi: 10.4018/IJDWM.2018010101.

[14] K. Gajowniczek, A. Orłowski, and T. Ząbkowski, "Entropy Based Trees to Support Decision Making for Customer Churn Management", *Acta Phys Pol A*, vol. 129, no. 5, pp. 971–979, May 2016, doi: 10.12693/APhysPolA.129.971.

[15] C. F. L. Lima, F. M. de Assis, and C. P. de Souza, "Decision Tree Based on Shannon, Rényi and Tsallis Entropies for Intrusion Tolerant Systems", in *2010 Fifth International Conference on Internet Monitoring and Protection*, May 2010, pp. 117–122. doi: 10.1109/ICIMP.2010.23.

[16] K. Gajowniczek, T. Ząbkowski, and A. Orłowski, "Comparison of Decision Trees with Rényi and Tsallis Entropy Applied for Imbalanced Churn Dataset", Oct. 2015, pp. 39–44. doi: 10.15439/2015F121.

[17] A. Ramirez-Arellano, L. M. Hernández-Simón, and J. Bory-Reyes, "A box-covering Tsallis information dimension and non-extensive property of complex networks", *Chaos Solitons Fractals*, vol. 132, p. 109590, Mar. 2020, doi: 10.1016/j.chaos.2019.109590.

[18] M. Lichman, "UCI Machine Learning Repository. University of California. Irvine: School of Information and Computer Sciences". 2013.

[19] A. Ramirez-Arellano, "Classification of Literary Works: Fractality and Complexity of the Narrative, Essay, and Research Article", *Entropy*, vol. 22, no. 8, p. 904, Aug. 2020, doi: 10.3390/e22080904.

[20] C. Song, L. K. Gallos, S. Havlin, and H. A. Makse, "How to calculate the fractal dimension of a complex network: the box covering algorithm", *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2007, no. 03, pp. P03006–P03006, Mar. 2007, doi: 10.1088/1742-5468/2007/03/P03006.

[21] A. Ramirez-Arellano, L. M. Hernández-Simón, and J. Bory-Reyes, "Two-parameter fractional Tsallis information dimensions of complex networks", *Chaos Solitons Fractals*, vol. 150, p. 111113, Sep. 2021, doi: 10.1016/j.chaos.2021.111113.

[22] C. Song, S. Havlin, and H. A. Makse, "Origins of fractality in the growth of complex networks", *Nature Physics*, vol. 2, no. 4, pp. 275–281, Apr. 2006, doi: 10.1038/nphys266.

[23] Duan, S., Wen, T., & Jiang, W. (2019). A new information dimension of complex network based on Rényi entropy. Physica A: Statistical Mechanics and its Applications, 516, 529-542. doi:https://doi.org/10.1016/j.physa.2018.10.045.

[24] Zhang, Q., Luo, C., Li, M., Deng, Y., & Mahadevan, S. (2015). Tsallis information dimension of complex networks. Physica A: Statistical Mechanics and its Applications, 419, 707-717. doi:https://doi.org/10.1016/j.physa.2014.10.071.