Clasificación de datos basado en compresión

Data classification based on compression

Avid Roman Gonzalez

TELECOM ParisTech, 46 rue Barrault, 75013 – Paris, Francia
German Aerospace Center – DLR, Remote Sensing Institute, Oberpfaffenhofen 82234 Wessling, Germany
Centre National d'Etudes Spatiales – CNES, Francia

RESUMEN

El incremento del volumen de datos en esta era digital es enorme, la tarea de analizarlos, procesarlos, identificarlos para luego poder clasificarlos y así tener un buen sistema de minería de datos donde poder indexar la información que contienen sin importar la cantidad y el tipo de dato, resulta una tarea nada fácil. Debido a esto, cada vez se hace más necesario el desarrollo de métodos más efectivos que faciliten estas tareas de manera automática. En este articulo se presenta un vista general de diferentes trabajos realizados a lo largo del mundo que utilizan técnicas de compresión de datos como base para el desarrollo de un método de clasificación, estas técnicas se basan en la Complejidad de Kolmogorov y la utilización de esta para implementar una medida de similaridad entre datos. El aporte principal de estos métodos es la no necesidad de un proceso de extracción de características para realizar la clasificación, lo cual hace que sea un método libre de parámetros, por lo que se puede aplicar a cualquier tipo de datos, ya sean texto, imágenes, audio, etc.
Descriptores: *clasificación, NCD, compresión de datos, similaridad métrica.*

ABSTRACT

The increased volume of data in this digital age is enormous, the task of analyzing, processing, identifying and classify them for to have a good data mining system where we can index the information contained regardless the amount and data type, it is no easy task. That is the reason for it is becoming more necessary to develop more effective methods to facilitate these tasks automatically. This paper presents an overview of different works performed throughout the world that use data compression techniques as a basis for developing a classification method, these techniques are based on Kolmogorov Complexity and use this complexity for implement a similarity metrics between data. The main contribution of these methods is, no need a feature extraction process for classification, which makes it a parameter-free method, so it can be applied to any type of data, whether text, images, audio, etc.
Keywords: *classification, NCD, data compression, metric similarity*

INTRODUCCIÓN

Los métodos tradicionales de clasificación de datos basan su funcionamiento en la extracción de uno o más características y/o parámetros. Esta técnica clásica se hace mucho más compleja cuando la diversidad y cantidad de datos aumenta considerablemente, ya que la característica y/o parámetro determinado no sirve para clasificar a toda la diversidad de datos que se tenga, por tanto el sistema carece de robustez y efectividad. Por lo anteriormente expuesto, surge la necesidad de encontrar un método de clasificación de datos libre de parámetros, el cual nos permita interactuar con gran diversidad de datos sin importar sus características.

La finalidad de este artículo es hacer una revisión y dar a conocer los diferentes trabajos relacionados con clasificación de datos usando técnicas de compresión, una de las bases más importantes para ello es presentada en [1]. Para entender como es que la compresión puede utilizarse como método de

clasificación es necesario realizar una revisión teórica de algunos conceptos.

Existen 2 enfoques para tratar y evaluar la información: el enfoque probabilístico de Shannon y el enfoque algorítmico de Kolmogorov, Solomonoff o Chaitin [2].

La Complejidad de Kolmogorov (complejidad algorítmica) *K(x)* de un objeto *x* es la cantidad de recursos computacionales necesarias para describir *x*. Conocido también en algunas áreas importantes como Teoría de la Información Algorítmica. La Complejidad de Kolmogorov (*KC*) es directamente asociado con otras entidades y áreas de estudio como: Entropía de Shannon & Teoría de la Información Clásica, Mínimum Description Length, Occam's Razor. La *KC* es no computable. La *KC* puede ser aproximado como en [1] y usado en varias aplicaciones como las presentadas en [2], [4], [5], [9], [10] y [11].

La estructura del presente artículo es como siguen: En la siguiente sección presentaremos las bases teóricas, luego presentaremos las aplicaciones e investigaciones que utilizan técnicas de compresión como método de clasificación, finalmente presentaremos las discusiones y conclusiones.

MARCO TEORICO

En esta parte se va a presentar las bases teóricas sobre las cuales están implementados los diferentes métodos de clasificación de datos basados en compresión.

A. La Entropía de Shannon

La entropía de la información de una variable *X* aleatoria, que puede tomar los valores 0 o 1. La entropía depende de la probabilidad *P (X=1)* de que *X* tome el valor 1. Cuando *P(X=1)=0.5*, todos los resultados posibles son igualmente probables, por lo que el resultado es poco predecible y la entropía es máxima, tal como se muestra en la figura 1.

La entropía de Shannon esta definida según la ecuación 1.

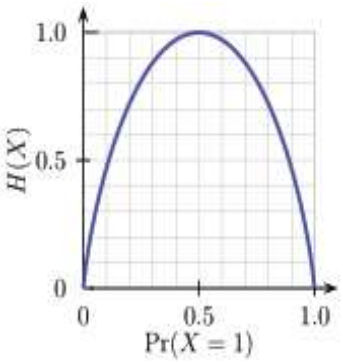


Figura 1: Entropía de una variable *X*.

(ec. 1) $$H(X) = -\sum_{x} p(x) \log p(x)$$

Donde:

H(X): Entropía de *X*

p(x): Probabilidad de *x*

La teoría de información de Shannon también puede utilizarse como un proceso de compresión. Por ejemplo si tenemos un mensaje *X = I N G E N I E R I A,* compuesto por un alfabeto *{A, E, G, I, N, R}*, como el alfabeto esta compuesto por 3 símbolos, se necesitaría de 3 bits para poder codificar cada uno, pero si estudiamos la probabilidad de uso de cada símbolo tendríamos el siguiente resultado:

X	I	E	N	A	G	R
P(x)	3/10	2/10	2/10	1/10	1/10	1/10

Si usamos códigos más cortos para los símbolos con mayor probabilidad, tenemos el siguiente resultado:

Símbolo	I	E	N	A	G	R
Código	00	01	100	101	110	111

Lo cual finalmente nos indica que en promedio se ha utilizado menos de 3 bits por símbolo, para este caso:

$$2\left(\frac{3}{10}+\frac{2}{10}\right)+3\left(\frac{2}{10}+\frac{1}{10}+\frac{1}{10}+\frac{1}{10}\right)=2.5bits$$

B. La Complejidad de Kolmogorov

Conocido también como la Complejidad de Kolmogorov-Chaitin o como la Complejidad Algorítmica. La Complejidad de Kolmogorov de *x, K(x),* es el programa mas corto *q* que puede dar como resultado la cadena *x* y corre en una maquina universal de Turing. *K*(*x*) es una medida de los recursos computacionales necesarios para especificar un objeto *x*. Pero existe un inconveniente, la característica principal de la Complejidad de Kolmogorov es que *K(x)* es una función no calculable.

(ec. 2) $$K(x)=\min_{q\in Qx}|q|$$

Donde:

Q_x *es el conjunto de códigos que generan instantáneamente* x.

Si queremos ver la complejidad de manera más práctica, pues analizamos las siguientes cadenas:

Cadena 1	000000000000000 15 x (Write 0)
Cadena 2	1010001010111011 Write 1010001010111011

La segunda cadena es más compleja que la primera.

C. La Distancia de Información Normalizada

La Distancia de Información Normalizada (*NID* por sus siglas en ingles) es la longitud normalizada del programa mas corto que puede calcular *x* conociendo *y*, así como calcular *y* conociendo *x*.

La *NID* es una medida de similaridad:

- *NID*(*x*,*y*) = 0 *Si* *x* = *y*
- *NID*(*x*,*y*) = 1 --> distancia máxima entre *x* & *y*

El *NID* minimiza todas las distancias a valores normalizados. La *NID* se calcula según la ecuación 3.

(ec. 3) $$\text{NID}(x,y)=\frac{K(x,y)-\min\{K(y),K(x)\}}{\max\{K(x),K(y)\}}$$

D. La Distancia de Compresión Normalizada

La Distancia de Compresión Normalizada (*NCD* por sus siglas en ingles) es una aproximación de la *NID* ya que la Complejidad de Kolmogorov *K*(*x*) es una función no calculable. *K*(*x*) representa un limite inferior de lo que se puede lograr con un compresor por lo que en [1] Vitányi sugiere la siguiente aproximación: *K(x)* -> *C(x)* donde C(x) es el factor de compresión (tamaño del archivo comprimido / tamaño del archivo original) de *x* con un compresor estandar (como el zip).

Para poder ver un poco la relación y aproximación de la complejidad y la compresión, se presenta en la figura 2 un par de imágenes con complejidad visual distinta lo cual se refleja en el factor de compresión, donde la imagen de los cachorritos es menos compleja ya que tiene un fondo constante y por ende se puede comprimir mas, mientras la otra figura de un nacimiento navideño contiene mucho detalle, es mas compleja y por ello no se puede comprimir mucho.

Figura 2: Compresión y complejidad.

Después de esta aproximación de *K(x)* con *C(x)* siendo este ultimo el factor de compresión de *x*, se obtiene la Distancia de compresión Normalizada NCD (Normalized Compression Distance).

(ec. 4) **Error! Objects cannot be created from editing field codes.**

Donde $C(x,y)$ $C(x,y)$ es un aproximación de la complejidad de Kolmogorov $K(x,y)$ y representa el tamaño del archivo al comprimir la concatenación de x e y.

E. Compresión de Datos

Con la compresión se pretende transportar la misma información, pero empleando la menor cantidad de espacio. Un ejemplo sobre compresión se muestra en la figura 3.

Figura 3: Ejemplo de compresión.

Cuantos más bits se empleen mayor será el tamaño del archivo.

Existen diferentes tipos de compresores: Compresores con pérdida y compresores sin pérdida, esta clasificación se muestra en la figura 4.

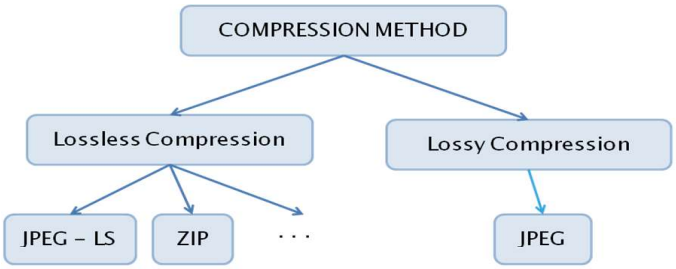


Figura 4: Métodos de compresión

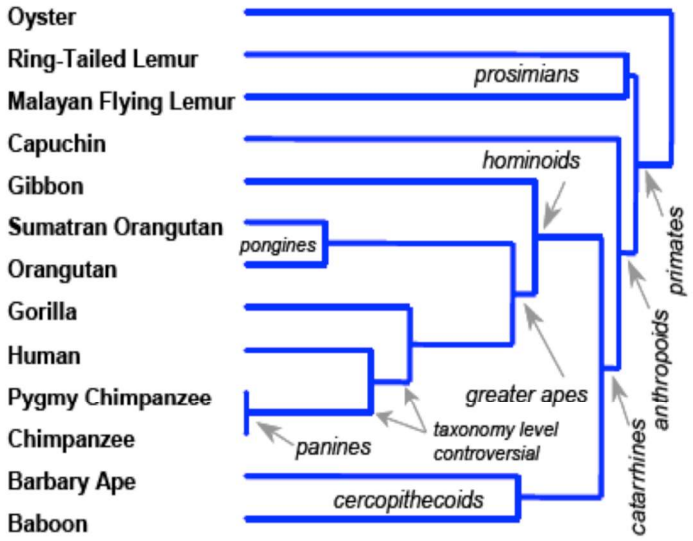
La mayoría de compresores están basados en el algoritmo LZW (Lempel-Ziv-Welch), se basan en un análisis inicial del texto para identificar cadenas repetidas para armar un diccionario de equivalencias, asignando códigos breves a estas cadenas. En una segunda etapa, se convierte el texto utilizando los códigos equivalentes para las cadenas repetidas. Esto requiere dos etapas, una de análisis y una segunda de conversión y también requiere que el diccionario se encuentre junto con el texto codificado, incrementando el tamaño del archivo de salida.

APLICACIONES EN CLASIFICACION DE DATOS

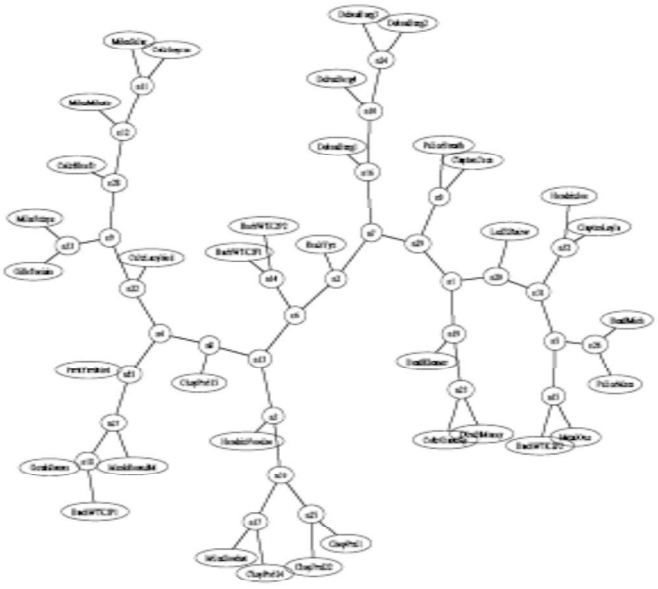
Utilizando técnicas de compresión y una clasificación jerárquica se han desarrollado diferentes aplicaciones como: Clasificación jerárquica de textos simples, diccionario de diferentes idiomas, clasificación de música, secuencias de ADN, secuencia de imágenes biológicas, cadena de caracteres, tablas, imágenes satelitales, etc.

Podemos mencionar algunos trabajos de investigación como:

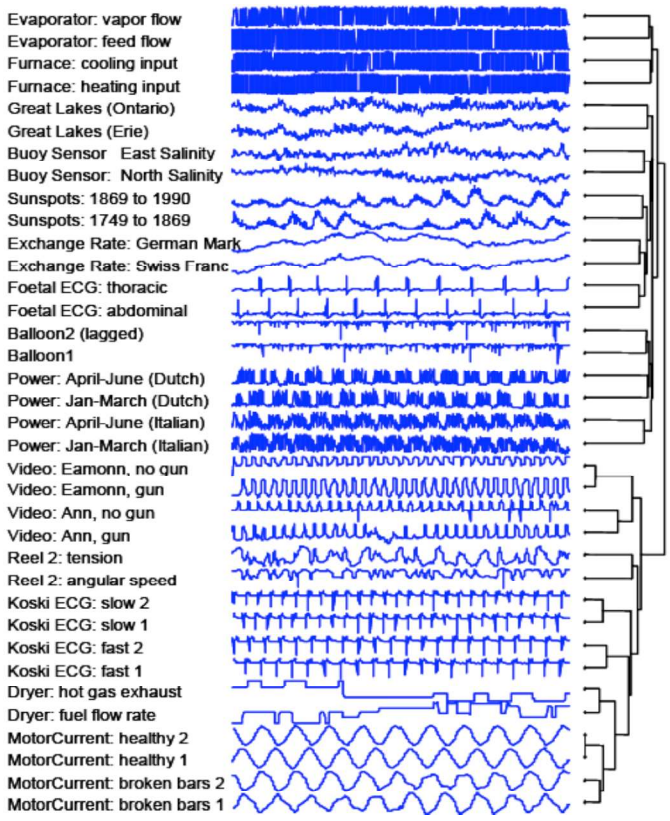
Clasificación de secuencias de ADN presentado en [4].



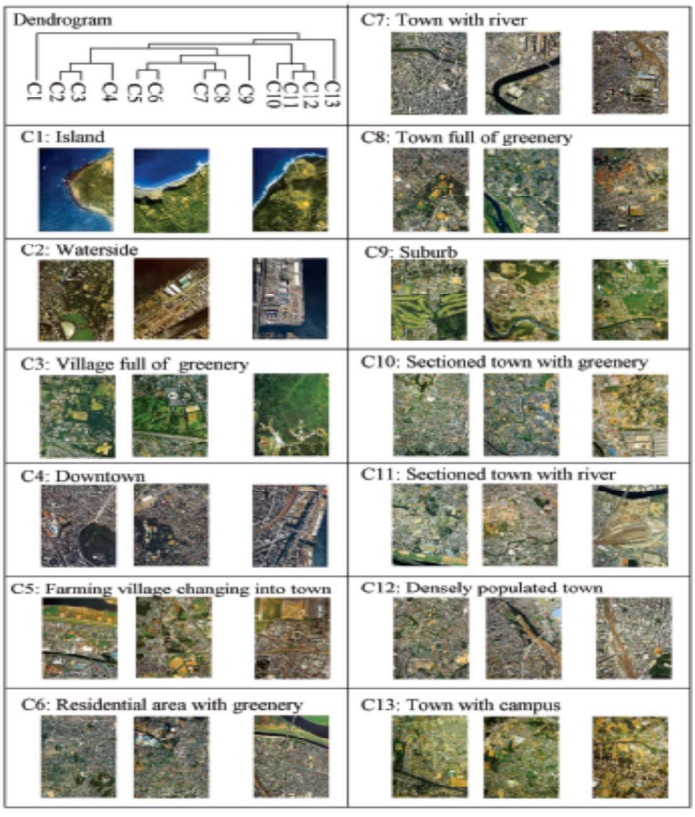
Clasificación de ritmos musicales presentado en [9].



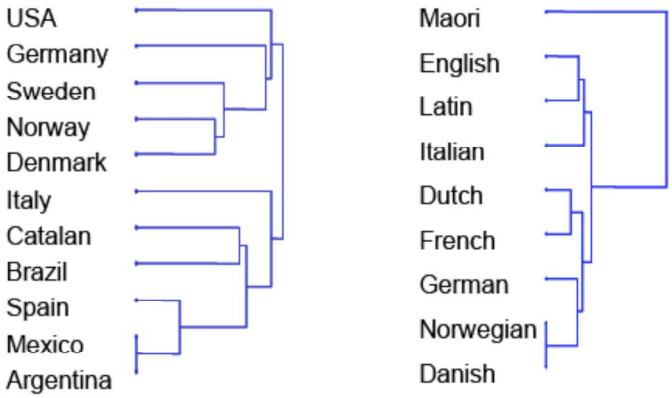
Clasificación de señales biomédicas presentado en [4].



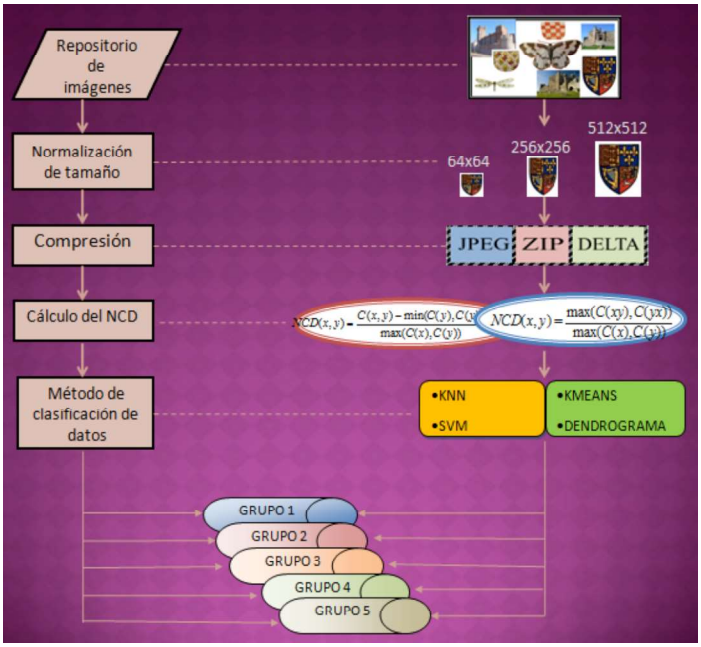
Clasificación de imágenes satelitales presentado en [11].



Clustering de idiomas presentado en [4].



Clasificación de imágenes presentado en [5] y [10].



CONCLUSIONES

Con el desarrollo de este método la clasificación de datos de realiza de una manera libre de parámetros de tal manera de no necesitar de un proceso previo de extracción de características, esto hace que el proceso sea más sencillo sin importar el tipo de dato sobre el cual se está trabajando. Así mismo aparte de las aplicaciones ya mencionadas que presentan muy buenos resultados, este método se podría aplicar para la detección de plagios, búsqueda de imágenes por contenido, detección de anomalías, etc.

REFERENCIAS

[1] M. Li and P. Vitányi, “The Similarity Metric”, IEEE Transaction on Information Theory, vol. 50, N° 12, 2004, pp. 3250-3264.

[2] M.C. Pinto, “Um Algoritmo para Comparação Sintática de Genomas Baseado na Complexidade Condicional de Kolmogorov”, Universidad Estadual de Campinas, Brasil 2002.

[3] Paul M. B. Vitanyi, Frank J. Balbach, Rudi L. Cilibrasi, and Ming Li, “Kolmogorov Complexity and its applications”, Springer, 1997.

[4] E. Keogh, S. Lonardi, Ch. Ratanamahatana, “Towards Parameter-Free Data Mining”, Department of Computer Science and Engineering, University of California, Riverside.

[5] B.J.L. Campana y E.J. Keogh, “A Compression Based Distance Measure for Texture”, University of California, Riverside, EEUU 2010.

[6] F. Tussell, “Complejidad Estocástica”, San Sebastián 1996.

[7] J. Rissanen, “Information and Complexity in Statistical Modeling”, Springer, 2007.

[8] D. McKay, “Information Theory, Inference, and Learning Algorithms”, Cambridge University Press, 2003.

[9] R. Cilibrasi, P. M. B. Vitanyi; “Clustering by Compression”, *IEEE Transaction on Information Theory*, vol. 51, N° 4, April 2005, pp 1523 – 1545.

[10] M. R. Quispe-Ayala, K. Asalde-Alvarez, A. Roman-Gonzalez, “Image Classification Using Data Compression Techniques”; *2010 IEEE 26th Convention of Electrical and Electronics Engineers in Israel – IEEEI 2010*; Eilat – Israel; November 2010, pp. 349-353.

[11] T. Watanabe, K. Sugawara, H. Sugihara, “A New Pattern Representation Scheme Using Data Compression” IEEE Transactions on pattern Analysis and Machine Intelligence, Vol. 24 N°5, May – 2002, pp. 579 – 590.

E-mail: avid.roman-gonzalez@ieee.org