Image Classification Using Data Compression Techniques

Martha Roxana Quispe Ayala, Krista Asalde Alvarez, Avid Roman Gonzalez

Image Classification Using Data Compression Techniques

Martha R. Quispe-Ayala * Krista Asalde-Alvarez * Avid Roman-Gonzalez * [+]
martharoxana_14@hotmail.com kristita@hotmail.com a.roman@ieee.org
* Universidad Nacional San Antonio Abad del Cusco – UNSAAC, Perú
[+] TELECOM ParisTech, 75013 Paris, France

***Abstract*—In this paper we propose a parameter free image classification method based on data compression techniques, so to calculate a measure of similarity between images based on the approximation of the compression to Kolmogorov Complexity. For this experiment we use two types of compressors, ZIP the general purpose compressor and JPEG compressor specialized for images. After developing the method and perform the relevant experiments to determine whether the proposed approach is useful or not for image classification**.

***Keywords*—Image Classification, Data Compression, NCD, Kolmogorov Complexity, JPEG, ZIP.**

I. Introduction

CURRENTLY image classification systems are based on the one or more features extraction to classify. This classic technique is much more complex when the diversity and number of images increases significantly as a determined property and / or parameter is not to classify the diversity of images, thus the system lacks robustness and effectiveness. For the foregoing reasons, the needs to find a parameter free image classification method, which allows us to interact with a variety of data regardless of its features.

The purpose of this work is to implement a tool for image classification using data compression techniques and thereby facilitate the search and classification of images. The project is to apply the Shannon information theory, Kolmogorov Complexity and its approximation to the Compression Factor; to use the concept of Normalized Compression Distance NCD for an effective image classification. Once determined the compression factor of each image (compressed image weight / weight original image) is necessary to measure the degree of similarity between images.

Paul Vitányi in 2004 approximated the Kolmogorov Complexity *K(x)* with the compression factor *C(x)* [1], this approach allows the use of NCD Normalized Compression Distance to measure the similarity degree between data. Before researches have used data compression techniques for classification and grouping of data such as:

- Syntactic Comparison of genomes based on the Conditional Kolmogorov Complexity [2].
- Analysis of the relationship between language and construction of a language tree that defines the languages those are more similar to other [3].
- Detection of plagiarism in programming code, based on a variant of the NID to measure the distance between two source code files.
- Classification of biomedical signals and detection of anomalies in the signals [4].
- Image Classification using MPEG compressor [5].

II. Theoretical Background

A. Shannon Information Theory

The Information Theory is a branch of mathematical theory of probability and statistics that studies the information and how to storage in digital form. It was developed by Claude E. Shannon in 1948 to find the fundamental limits in compression and reliable storage of data communication

According to probabilistic considerations is possible to establish a first principle of information measurement. It states that the more likely a message is less information provided. This can be expressed as follows:

$$I(x_i) > I(x_k) \Leftrightarrow P(x_i) < P(x_k) \quad (1)$$

Where:

$I(x_i)$: Amount of information provided by x_i

$P(x_i)$: Probability of x_i

According to this principle is the probability that a message be sent and not its content, which determines the amount of information. The content is only important insofar as it affects the probability. The amount of information that provides a message varies from one context to another, because the probability of sending a message varies from one context to another

A second principle states that if selected two messages *x* and *y* the amount of information from both messages is equal to the amount of information provided by *x* plus the quantity of information provided by *y*, given that *x* has already been selected

This can be expressed as

$$I(x_i, y_i) = fP(x_i) + fP(\frac{y_i}{x_i}) \quad (2)$$

Where:

$I(x_i, y_i)$: Amount of information provided by messages x_i and y_i.

f : Function

$P(x_i)$: Probability of x_i

$P(\frac{y_i}{x_i})$: Probability of y_j given that x_i has been selected

If there is a message x_1, with a occurrence probability $P(x_1)$, the information content can be expressed as:

$$I(x_i) = \log_2 {}^{1}\!/\!_{P(x_1)} \quad (3)$$

Where: $I(x_i)$ will has as unit the bit.

B. Information and Complexity

From what we saw in the previous section follows that $\log_2 p(x_i)$ is approximately the information contained in x. We imagine the following two strings of binary digits:

String 1	000000000000000
String 2	101000101011101

Table 1. Strings of Binary Digits

Both have the same number of binary digits, 15. If we imagine the set of all strings of 15 binary digits (there are 2^{15} different) and take them one randomly, either displayed is equally likely to appear: 2^{15}. However, we would assign to the first lower complexity than the latter. One way to rationalize this is that we pass first to a third party by a very short description: "fifteen zeros", while the latter requires a long description, which could hardly be shorter than the original string.

C. Kolmogorov Complexity

The Kolmogorov Complexity $K(x)$ of a string x is defined as the length of the shortest program capable of producing x on a universal machine, as Turing machine. Different programming languages give different values of $K(x)$ but can prove that the differences are only a fixed additive constant. Intuitively, $K(x)$ is the minimum amount of information needed to generate x by an algorithm.

$$K(x) = \min_{q \in Q_x} |q| \quad (4)$$

Q_x is the set of codes to generate instantly x. Since the programs can be written in different programming languages, $K(x)$ is measured to an additive constant not depending on the objects but in the Turing machine used. One interpretation is the amount of information needed to recover x from the beginning, the strings are recurring patterns that have low complexity, while the complexity of random strings is high and almost equal to its length. The main property of $K(x)$ is that it can not be computable.

The Kolmogorov Complexity $K(x|y)$ from x to y is defined as the length of the shortest program that computes x when y is given as an auxiliary input for the program. The function $K(x|y)$ is the length of the shortest program that produces the concatenation of x and y.

D. Data Compression

With compression is intended to convey the same information, but using the least amount of space. Data compression is based primarily on finding repetitions in sets of data to then store only the data with the number of times it is repeated. For example, if a file appears in a sequence like "AAAAAA", occupying 6 bytes, could be stored simply "6A" which occupies only 2 bytes.

We have two compression methods: lossless compression and lossy compression, as shown in figure 1.

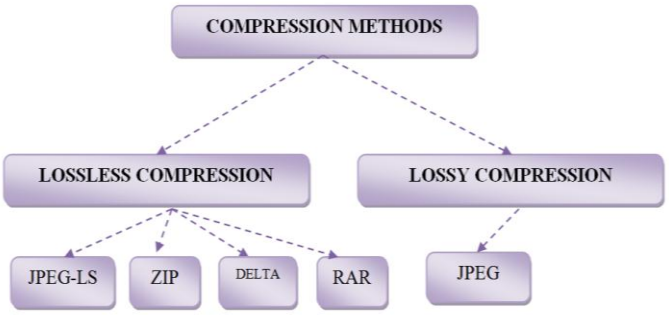

Fig. 1. Compression Methods

ZIP Compressor: This is a single step of encoding based on the combination of LZW code and Huffman code. The input file is divided into sequence of blocks where each block is compressed using the combination of codes, in figure 2 we can see the block diagram for a zip compressor.

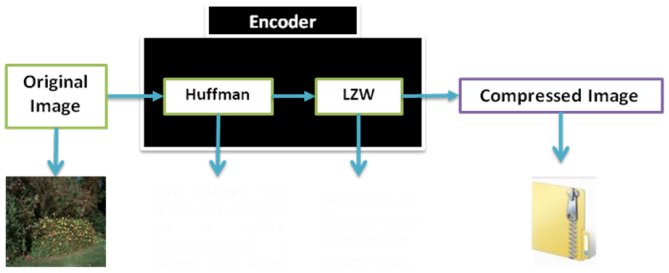

Fig. 2. ZIP Compressor

JPEG-LS Compressor: It is a form of lossless JPEG encoding. Although this is not widely used by the data processing community in general, is especially used for the transmission of medical images in order to avoid artifacts in the image (only dependent on the image and scanning) and confusing signs real disease. Thus, the compression is much less effective.

JPEG-LS compressor takes an image and for each pixel it try to predict the pixel value based on the neighborhood and then apply an entropy encoder, finally get

the compressed image. In figure 3 we can see the block diagram for a lossless JPEG compressor.

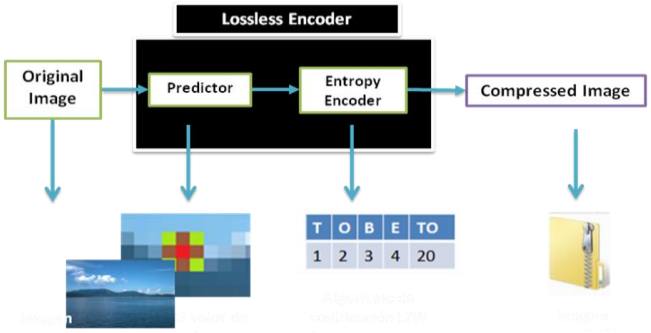

Fig. 3. JPEG-LS Compressor

JPEG Compressor: It is a lossy compression algorithm, this means that when decompressing the image does not get exactly the same picture we had before compression.

Lossy JPEG compressor takes a image and divides it into blocks of 8x8 pixels, for each block applies Discrete Cosine Transform (DCT) and then apply a quantifier, finally an entropy encoder for to get the compressed image; worth highlighting the loss of information is in the quantifier. In figure 4 we can see the block diagram for a lossy JPEG compressor.

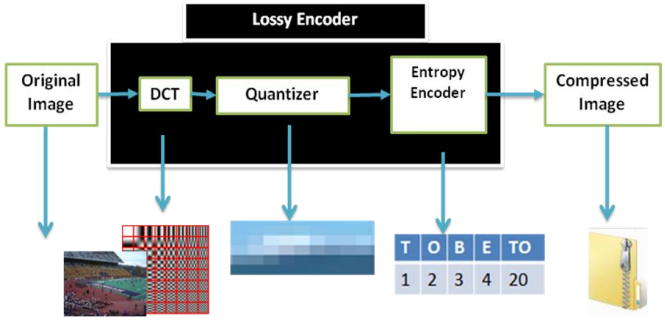

Fig. 4. JPEG Compressor

III. Parameter Free Method for Data Clasification

A. Normalized Information Distance:

One of the applications of Kolmogorov Complexity and the likely success of Algorithmic Information Theory is the latest estimate for sharing data between two objects: Normalized Information Distance NID, is a measure of similarity admissible minimizing any measurement proportional to the length of the shortest program that represents *x* given *y* as much as the shortest program that represents *y* given *x*. The calculated distance is normalized, meaning that its value is between 0 and 1, 0 when *x* and *y* are totally equal and 1 when the maximum difference between them.

$$NID(x,y)=\frac{\max\{K(x\mid y),K(y\mid x)\}}{\max\{K(x),K(y)\}}=\frac{K(x,y)-\min\{K(x),K(y)\}}{\max\{K(x),K(y)\}} \quad (5)$$

B. Normalized Compression Distance

As the complexity *K(x)* is not a computable function of *x*, a reasonable approximation is defined by Li and Vitanyi [1] which approximates *K(x)* with *C(x)* the latter being the compression factor of *x*. Based on this approach yields the Normalized Compression Distance NCD.

$$NCD(x,y)=\frac{C(x,y)-\min\{C(x),C(y)\}}{\max\{C(x),C(y)\}} \quad (6)$$

Where *C(x,y)* is an approximation of Kolmogorov Complexity *K(x,y)* and represents the file size by compressing the concatenation of *x* and *y*.

As the above NCD is asymmetrical, we propose a NCD_2 with symmetrical properties.

$$NCD_2(x,y)=\frac{\max\{C(xy),C(yx)\}}{\max\{C(x),C(y)\}} \quad (7)$$

C. Classification Methods:

For this work, we used the following classification methods:

KNN (K Nearest Neighbors): Supervised classification algorithm. Proceed to classify a data according to K (positive integer) nearest neighbors to the query data.

SVM (Support Vector Machine): Supervised classification algorithm. Take half of the data for training, dividing them into two groups, then create each tab that may be due to linear and nonlinear properties. A consultation data is classified according to which side of the separator is.

K-Means: Unsupervised classification algorithm. Classify data according to K (positive integer) or centroids groups as assigned. Calculate the minimum distance of these centroids to all other data and the group as the minimum distance

Dendrogram: Unsupervised classification algorithm. Group the data based on the minimum Euclidean distance, forming classification hierarchies.

IV. Testing and Results Analysis

A. Image Data Base

To determine test analysis, we create two image data base that from now on we will call as BD1 and BD2, each

containing five groups of 10 images each one, undefined size initially, a total of 50 images for each database.

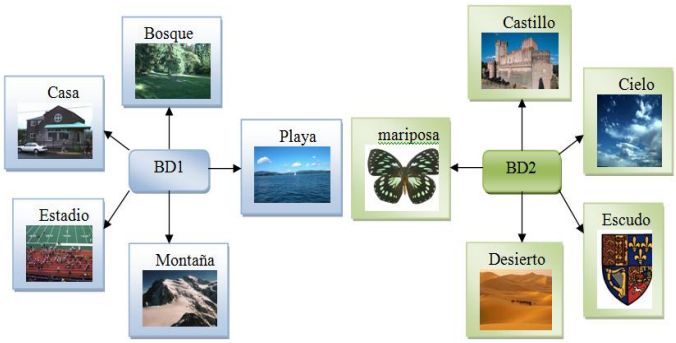

Fig. 5. Image Data Base Under Test

B. Experiment Process

Each element of BD1 and BD2 pass through the experimental process detailed in the diagram in figure 6. First resize images in our database that have the same size, we call this step as normalization of the image where we can leave or change color images to grayscale, the next step is the image compression with the two types of compressors (JPEG, ZIP) and then find the Symmetric and Asymmetric NCD and finally analyzed by each of the four data classification methods listed above.

C. Results

To show and analyze the results, it was calculated the confusion matrices for each test with each of the databases, with each standard size either grayscale or color, with different compressors calculating the symmetrical or asymmetric NCD, as well as for each classification method.

Below, we show some results.

BD2x64 COLOR – JPEG – NCD ASIMETRIC

KNN = 72%

3	2	0	0	0
0	4	1	0	0
0	1	4	0	0
0	0	0	3	2
0	0	0	1	4

SVM = 40%

1	0	0	0	4
0	1	0	0	4
0	1	1	0	3
0	0	0	3	2
0	0	0	1	4

K means = 56%

7	1	1	0	1
3	4	1	0	2
0	0	6	3	1
0	0	0	6	4
0	4	0	1	5

BD2x64 COLOR – ZIP-NCD ASIMETRIC

KNN = 64%

5	0	0	0	0
2	2	1	0	0
2	1	1	1	0
0	0	1	3	1
0	0	0	0	5

SVM = 20%

0	0	0	0	5
0	0	0	0	5
0	0	0	0	5
0	0	0	0	5
0	0	0	0	5

K MEANS = 46%

6	0	0	0	4
0	2	7	1	0
0	1	9	0	0
3	0	0	6	1
2	2	4	2	0

BD1x64 COLOR – ZIP-NCD SIMETRIC

KNN = 88%

5	0	0	0	0
0	5	0	0	0
0	1	4	0	0
0	0	0	5	0
0	2	0	0	3

SVM = 20%

0	0	0	0	5
0	0	0	0	5
0	0	0	0	5
0	0	0	0	5
0	0	0	0	5

Kmeans = 66%

8	0	1	1	0
0	10	0	0	0
0	10	0	0	0
0	0	0	10	0
0	0	5	0	5

BD1x512 GRAY SCALE – JPEG-NCD SIMETRIC

KNN = 76%

2	1	2	0	0
0	5	0	0	0
0	0	5	0	0
0	0	1	4	0
0	1	0	1	3

SVM = 72%

2	0	1	1	1
0	5	0	0	0
0	0	4	1	0
0	0	2	3	0
0	0	0	1	4

Kmeans = 52%

9	0	0	0	1
1	4	1	4	0
2	0	0	0	8
6	0	0	4	0
1	0	0	0	9

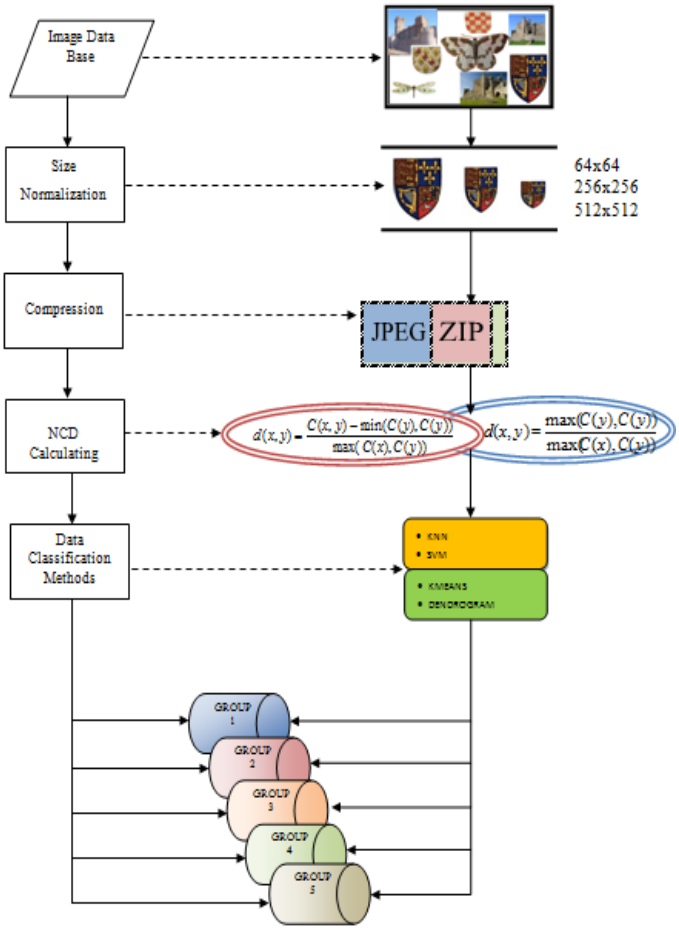

Fig. 6. Experiment Process

Below, we show two tables summarize all the experiments results:

	KNN			SVM			K MEANS		
	x64	x256	x512	x64	x256	x512	x64	x256	x512
JPEG-COLOR-NCD-ASIMETRIC	72%	60%	72%	40%	36%	40%	56%	46%	50%
JPEG-GRAY-NCD-ASIMETRIC	72%	68%	60%	40%	36%	40%	60%	46%	48%
ZIP-COLOR-NCD-ASIMETRIC	60%	60%	48%	20%	20%	20%	46%	52%	52%
ZIP-GRAY-NCD-ASIMETRIC	72%	72%	64%	20%	20%	20%	58%	58%	54%
JPEG-COLOR-NCD-SIMETRIC	72%	60%	72%	40%	36%	40%	54%	42%	54%
JPEG-GRAY-NCD-SIMETRIC	72%	68%	60%	40%	36%	40%	58%	46%	52%
ZIP-COLOR-NCD-SIMETRIC	64%	60%	48%	20%	20%	20%	54%	46%	50%
ZIP-GRAY-NCD-SIMETRIC	72%	72%	64%	20%	20%	20%	54%	56%	50%

Table 2. Summary of Success Rates of BD2

	KNN			SVM			K MEANS		
	x64	x256	x512	x64	x256	x512	x64	x256	x512
JPEG-COLOR-NCD-ASIMETRIC	64%	64%	80%	56%	56%	60%	54%	48%	56%
JPEG-GRAY-NCD-ASIMETRIC	72%	68%	76%	64%	74%	64%	40%	54%	50%
ZIP-COLOR-NCD-ASIMETRIC	80%	76%	72%	20%	20%	20%	42%	52%	48%
ZIP-GRAY-NCD-ASIMETRIC	**84%**	80%	80%	20%	20%	20%	56%	52%	50%
JPEG-COLOR-NCD-SIMETRIC	76%	64%	80%	56%	64%	64%	50%	50%	46%
JPEG-GRAY-NCD-SIMETRIC	80%	72%	76%	64%	68%	72%	52%	48%	52%
ZIP-COLOR-NCD-SIMETRIC	**88%**	80%	80%	20%	20%	20%	66%	74%	60%
ZIP-GRAY-NCD-SIMETRIC	80%	80%	88%	20%	20%	20%	36%	60%	60%

Table 3. Summary of Success Rates of BD1

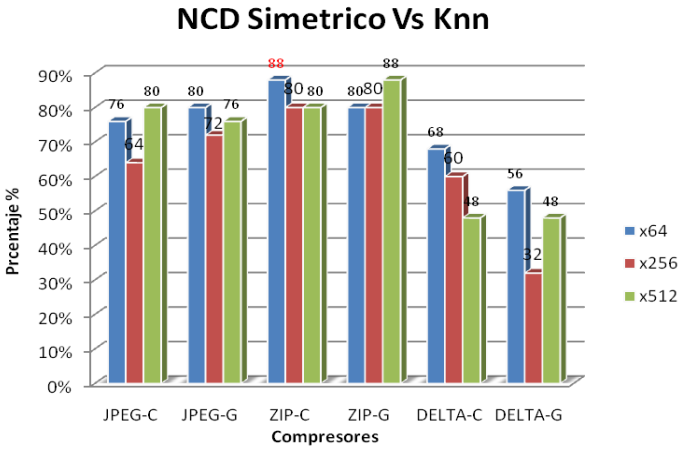

Fig. 7. Experiments Results

Finally by the bar chart in figure 7 we can see what parameters give the best results.

V. Conclusions

- We have developed a parameter free method for image classification, which uses ZIP compressor, the NCD Symmetric and KNN supervised classification method in standard images to 64x64 pixels to color, for optimum results of up to 88% success
- With regard to classification methods, we have better results with KNN is a supervised classification method compared to unsupervised methods such as K-MEANS
- The parameter free method based on the approximation of the Kolmogorov Complexity to data compression has a high degree of reliability when performing string comparison where the relationship between the data is horizontally, only in one dimension compared with data where the relationship is in two dimensions or more as in the case of images
- The results show a 88% success of data classification, this encouraging result, it is much higher than other results presented in classification systems and search for images based on content, so that it can be concluded that proposed method can be used to implement a system image search by content

VI. References

[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.