Evaluating Data Minability Through Compression – An Experimental Study

Dan Simovici
Univ. of Massachusetts Boston,
Boston, USA,
dsim at cs.umb.edu

Dan Pletea
Univ. of Massachusetts Boston,
Boston, USA,
dpletea at cs.umb.edu

Saaid Baraty
Univ. of Massachusetts Boston,
Boston, USA,
sbaraty at cs.umb.edu

***Abstract*—The effectiveness of compression algorithms is increasing as the data subjected to compression contains repetitive patterns. This basic idea is used to detect the existence of regularities in various types of data ranging from market basket data to undirected graphs. The results are quite independent of the particular algorithms used for compression and offer an indication of the potential of discovering patterns in data before the actual mining process takes place.**

***Keywords*-data mining; lossless compression; LZW; market basket data; patterns; Kronecker product.**

I. INTRODUCTION

Our goal is to show that compression can be used as a tool to evaluate the potential of a data set of producing interesting results in a data mining process. The basic idea that data that displays repetitive patterns or patterns that occur with a certain regularity will be compressed more efficiently compared to data that has no such characteristics. Thus, a pre-processing phase of the mining process should allow to decide whether a data set is worth mining, or compare the interestingness of applying mining algorithms to several data sets.

Since compression is generally inexpensive and compression methods are well-studied and understood, pre-mining using compression will help data mining analysts to focus their efforts on mining resources that can provide a highest payout without an exorbitant cost.

Compression has received lots of attention in the data mining literature. As observed by Mannila [7], data compression can be regarded as one of the fundamental approaches to data mining [7], since the goal of the data mining is to "compress data by finding some structure in it".

The role of compression developing parameter-free data mining algorithms in anomaly detection, classification and clustering was examined in [4]. The size $C(x)$ of a compressed file x is as an approximation of Kolmogorov complexity [2] and allows the definition of a pseudo-distance between two files x and y as

$$d(x,y) = \frac{C(xy)}{C(x)+C(y)}.$$

Further advances in this direction were developed in [8][5][6]. A Kolmogorov complexity-based dissimilarity was successfully used to texture matching problems in [1] which have a broad spectrum of applications in areas like bioinformatics, natural languages. and music.

We illustrate the use of lossless compression in pre-mining data by focusing on several distinct data mining processes: files with frequent patterns, frequent itemsets in market basket data, and exploring similarity of graphs.

The LZW (Lempel-Ziv-Welch) algorithm was introduced in 1984 by T. Welch in [9] and is among the most popular compression techniques. The algorithm does not need to check all the data before starting the compression and the performance is based on the number of the repetitions and the lengths of the strings and the ratio of 0s/1s or true/false at the bit level. There are several versions of the LZW algorithm. Popular programs (such as Winzip) use variations of the LZW compression. The Winzip/Zip type of algorithms also work at the bit level and not at a charater/byte level.

We explore three experimental settings that provide strong empirical evidence of the correlation between compression ratio and the existence of hidden patterns in data. In Section II, we compress binary strings that contain patterns; in Section III, we study the compressibility of adjacency matrix for graphs relative to the entropy of distribution of subgraphs. Finally, in Section IV, we examine the compressibility of files that contain market basket data sets.

II. PATTERNS IN STRINGS AND COMPRESSION

Let A^* be the set of strings on the alphabet A. The length of a string w is denoted by $|w|$. The null string on A is denoted by λ and we define A^+ as $A^+ = A^* - \{\lambda\}$.

If $w \in A^*$ can be written as $w = utv$, where $u, v \in A^*$ and $t \in A^+$, we say that the pair (t, m) is an occurrence of t in w, where m is the length of u.

The occurrences (x, m) and (y, p) are overlapping if $p < m + |x|$. If this is the case, there is a proper suffix of x that equals a proper prefix of y. If t is a word such that the sets of its proper prefixes and its proper suffixes are disjoint, there are no overlapping occurrences of x in any word. The number of occurrences of a string t in a string w is denoted by $n_t(w)$. Clearly, we have $\sum\{n_a(w) \mid a \in A\} = |w|$. The *prevalence* of t in w is the number $f_t(w) = \frac{n_t(w)\cdot|t|}{|w|}$ which gives the ratio of the characters contained in the occurrences of t relative to the total number of characters in the string.

 ISBN: 978-1-61208-242-4

The result of applying a compression algorithm C to a string $w \in A^*$ is denoted by $C(w)$ and the *compression ratio* is the number

$$\mathsf{CR}_C(w) = \frac{|C(w)|}{|w|}.$$

In this section, we shall use the binary alphabet $B = \{0, 1\}$ and the LZW algorithm or the compression algorithm of the package `java.util.zip`.

We generated random strings of bits (0s and 1s) and computed the compression ratio strings with a variety of symbol distributions. A string w that contains only 0s (or only 1s) achieves a very good compression ratio of $\mathsf{CR}_{jZIP}(w) = 0.012$ for 100K bits and $\mathsf{CR}_{jZIP} = 0.003$ for 500K bits, where $jZIP$ denotes the compression algorithm from the package `java.util.zip`. Figure 1 shows, as expected, that the worst compression ratio is achieved when 0s and 1s occur with equal frequencies.

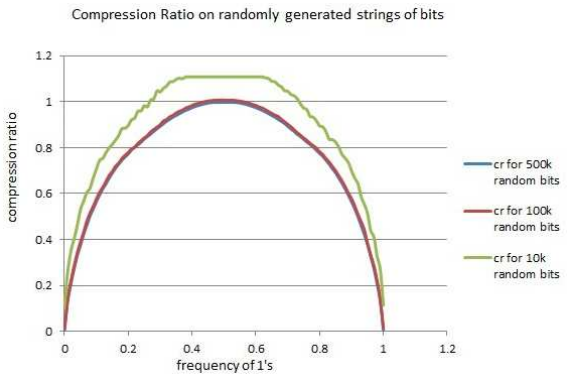


Figure 1. Baseline CR_{jZIP} Behavior

For strings of small length (less than 10^4 bits) the compression ratio may exceed 1 because of the overhead introduced by the algorithm. However, when the size of the random string exceeds 10^6 bits this phenomenon disappears and the compression ratio depends only on the prevalence of the bits and is relatively independent on the size of the file. Thus, in Figure 1, the curves that correspond to files of size 10^6 and $5 \cdot 10^6$ overlap. We refer to the compression ratio of a random string w with an $(n_0(w), n_1(w))$ distribution as the *baseline compression ratio*.

We created a series of binary strings $\varphi_{t,m}$ which have a minimum guaranteed number m of occurrences of patterns $t \in \{0,1\}^k$, where $0 \leq m \leq 100$. Specifically, we created 101 files $\varphi_{001,m}$ for the pattern 001, each containing 100K bits and we generated similar series for $t \in \{01, 0010, 00010\}$. The compression ratio is shown in Figure 2. The compression ratio starts at a value of 0.94 and after the prevalence of the pattern becomes more frequent than 20% the compression ratio drops dramatically. Results of the experiment are shown in Table I and in Figure 3.

Table I
PATTERN '001' PREVALENCE VERSUS THE CR_{jZIP}

Prevalence of '001' pattern	CR_{jZIP}	Baseline
0%	0.93	0.93
10%	0.97	0.93
20%	0.96	0.93
30%	0.92	0.93
40%	0.86	0.93
50%	0.80	0.93
60%	0.72	0.93
70%	0.62	0.93
80%	0.48	0.93
90%	0.31	0.93
95%	0.19	0.93
100%	0.01	0.93

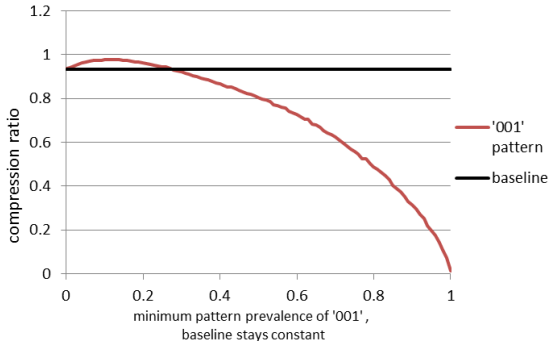


Figure 2. Variation of compression rate depends on the prevalence of the pattern '001'

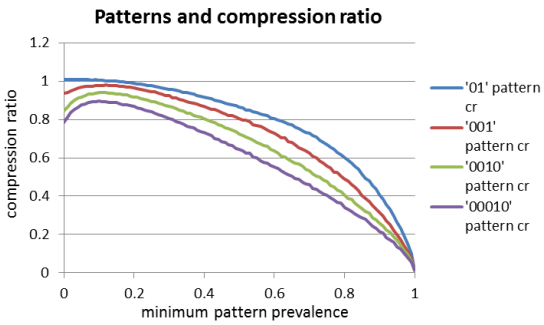


Figure 3. Dependency of Compression Ratio on Pattern Prevalence

We conclude that the presence of repeated patterns in strings leads to a high degree of compression (that is, to low compression ratios). Thus, a low compression ratio for a file indicates that the mining process may produce interesting results.

III. RANDOM INSERTION AND COMPRESSION

For a matrix $M \in \{0,1\}^{u \times v}$ denote by $n_i(M)$ the number of entries of M that equal i, where $i \in \{0, 1\}$. Clearly, we have $n_0(B) + n_1(B) = uv$. For a random variable V which

ranges over the set of matrices $\{0,1\}^{u\times v}$ let $\nu_i(V)$ be the random variable whose values equal the number of entries of V that equal i, where $i \in \{0,1\}$.

Let $A \in \{0,1\}^{p\times q}$ be a $0/1$ matrix and let

$$\mathcal{B} : \begin{pmatrix} B_1 & B_2 & \cdots & B_k \\ p_1 & p_2 & \cdots & p_k \end{pmatrix},$$

be a matrix-valued random variable where $B_j \in \mathbb{R}^{r\times s}$, $p_j \geq 0$ for $1 \leq j \leq k$, and $\sum_{j=1}^{k} p_j = 1$.

Definition 3.1: The random variable $A \leftarrow \mathcal{B}$ obtained by the *insertion* of $\mathcal{B}$ into A is given by

$$A \otimes \mathcal{B} = \begin{pmatrix} a_{11}\mathcal{B} & \ldots & a_{1n}\mathcal{B} \\ \vdots & \ddots & \vdots \\ a_{m1}\mathcal{B} & \ldots & a_{mn}\mathcal{B} \end{pmatrix} \in \mathbb{R}^{mr\times ns}$$

In other words, the entries of $A \leftarrow \mathcal{B}$ are obtained by substituting the block $a_{ij}B_\ell$ with the probability p_ℓ for a_{ij} in A. □

Note that this operation is a probabilistic generalization of Kronecker's product for if

$$\mathcal{B} : \begin{pmatrix} B_1 \\ 1 \end{pmatrix},$$

then $A \leftarrow B$ has as its unique value the Kronecker product $A \otimes B$.

The expected number of 1s in the insertion $A \leftarrow \mathcal{B}$ is

$$E[\nu_1(A \leftarrow \mathcal{B})] = n_1(A) \sum_{j=1}^{k} n_1(B_j)p_j$$

When $n_1(B_1) = \cdots = n_1(B_k) = n$, we have $E[\nu_1(A \leftarrow \mathcal{B})] = n_1(A)n$.

In the experiment that involves insertion, we used a matrix-valued random variable such that $n_1(B_1) = \cdots = n_1(B_k) = n$. Thus, the variability of the values of $A \leftarrow \mathcal{B}$ is caused by the variability of the contents of the matrices $B_1, \ldots, B_k$ which can be evaluated using the entropy of the distribution of $\mathcal{B}$,

$$\mathcal{H}(\mathcal{B}) = -\sum_{j=1}^{k} p_j \log_2 p_j.$$

We expect to obtain a strong positive correlation between the entropy of $\mathcal{B}$ and the degree of compression achieved on the file that represents the matrix $A \leftarrow \mathcal{B}$, and the experiments support this expectation.

In a first series of compressions, we worked with a matrix $A \in \{0,1\}^{106\times 106}$ and with a matrix-valued random variable

$$\mathcal{B} : \begin{pmatrix} B_1 & B_2 & B_3 \\ p_1 & p_2 & p_3 \end{pmatrix},$$

where $B_j \in \{0,1\}^{3\times 3}$, and $n_1(B_1) = n_1(B_2) = n_1(B_3) = 4$. Several probability distributions were considered, as shown in Table II. Values of $A \leftarrow \mathcal{B}$ had $106^2 * 3^2 = 101124$ entries.

In Table II, we had 39% 1s and the baseline compression rate for a binary file with this ratio of 1s is 0.9775. We also computed the correlation between the CR_{jZIP} and the Shannon entropy of the probability distribution and obtained the value 0.9825 for 3 matrices. In Table III, we did the same experiment but with 4 different matrices of 4×4. A strong correlation (0.992.) was observed between CR_{jZIP} and the Shannon entropy of the probability distribution.

Table II
MATRIX INSERTIONS, ENTROPY AND COMPRESSION RATIOS

Probability distribution	CR_{jZIP}	Shannon Entropy
$(0,1,0)$	0.33	0
$(1,0,0)$	0.33	0
$(0,0,1)$	0.33	0
$(0.2,0.2,0.6)$	0.77	1.37
$(0.6,0.2,0.2)$	0.74	1.37
$(0.33,0.33,0.34)$	0.79	1.58
$(0,0.3,0.7)$	0.7	0.88
$(0.9,0.1,0)$	0.51	0.46
$(0.8,0,0.2)$	0.61	0.72
$(0.49,0.25,0.26)$	0.77	1.5
$(0.15,0.35,0.5)$	0.78	1.44

Table III
KRONECKER PRODUCT AND PROBABILITY DISTRIBUTION FOR 4 MATRICES

Probability distribution	CR_{jZIP}	Shannon Entropy
$(0,1,0,0)$	0.23	0
$(0,1,0,0)$	0.23	0
$(0.2,0.2,0.2,0.4)$	0.69	01.92
$(0.25,0.25,0.25,0.25)$	0.69	2
$(0.4,0,0.2,0.4)$	0.53	1.52
$(0.3,0.1,0.2,0.4)$	0.65	1.84
$(0.45,0.12,0.22,0.21)$	0.61	1.83

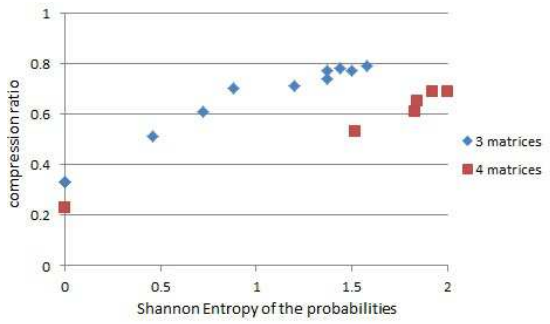


Figure 4. Evolution CR_{jZIP} and Shannon Entropy of Probability Distribution.

In Figure 4, we have the evolution of CR_{jZIP} on the y axis and on the x axis the Shannon Entropy of the probability distribution for both experiments. We can see clearly the linear correlation between the two.

 ISBN: 978-1-61208-242-4

This experiment proves us again that in case of repetitions/patterns the CR_{jZIP} is better than in the case of randomly generated files.

Next, we examine the compressibility of binary square matrices and its relationship with the distribution of principal submatrices. A binary square matrix is compressed by first vectorizing the matrix and then compressing the binary sequence. The issue is relevant in graph theory, where the principal submatrices of the adjacency matrix of a graph correspond to the adjacency matrices of the subgraphs of that graph. The patterns in a graph are captured in the form of frequent isomorphic subgraphs.

There is a strong correlation between the compression ratio of the adjacency matrix of a graph and the frequencies of the occurrences of isomorphic subgraphs of it. Specifically, the lower the compression ratio is, the higher are the frequencies of isomorphic subgraphs and hence the worthier is the graph for being mined.

Let $\mathcal{G}_n$ be an undirected graph having $\{v_1, \ldots, v_n\}$ as its set of nodes. The adjacency matrix of $\mathcal{G}_n$, $\mathbf{A}_{\mathcal{G}_n} \in \{0,1\}^{n \times n}$ is defined as

$$(\mathbf{A}_{\mathcal{G}_n})_{ij} = \begin{cases} 1 & \text{if there is an edge between } v_i \text{ and } v_j \text{ in } \mathcal{G}_n \\ 0 & \text{otherwise.} \end{cases}$$

We denote with $\mathsf{CR}_C(\mathbf{A}_{\mathcal{G}_n})$ the compression ratio of the adjacency matrix of graph $\mathcal{G}_n$ obtained by applying the compression algorithm C. Define the *principal subcomponent* of matrix $\mathbf{A}_{\mathcal{G}_n}$ with respect to the set of indices $S = \{s_1, \ldots, s_k\} \subseteq \{1, 2, \ldots, n\}$ to be the $k \times k$ matrix $\mathbf{A}_{\mathcal{G}_n}(S)$ such that

$$\mathbf{A}_{\mathcal{G}_n}(S)_{ij} = \begin{cases} 1 & \text{if there is an edge between } v_{s_i} \text{ and } v_{s_j} \\ & \text{in } \mathcal{G}_n \\ 0 & \text{otherwise.} \end{cases}$$

The matrix $\mathbf{A}_{\mathcal{G}_n}(S)$ is the adjacency matrix of the subgraph of $\mathcal{G}_n$ which consists of the nodes with indices in S along with those edges that connect these nodes. We denote by $\mathcal{P}_n(k)$ the collection of all subsets of $\{1, 2, \ldots, n\}$ of size k where $2 \leq k \leq n$. We have $|\mathcal{P}_n(k)| = \binom{n}{k}$.

Let $(\mathbf{M}_1^k, \ldots, \mathbf{M}_{\ell_k}^k)$ be an enumeration of possible adjacency matrices of graphs with k nodes where $\ell_k = 2^{\frac{k(k-1)}{2}}$. We define the finite probability distribution

$$P(\mathcal{G}_n, k) = \left(\frac{\mathsf{n}_1^k(\mathcal{G}_n)}{|\mathcal{P}_n(k)|}, \ldots, \frac{\mathsf{n}_{\ell_k}^k(\mathcal{G}_n)}{|\mathcal{P}_n(k)|} \right),$$

where $\mathsf{n}_i^k(\mathcal{G}_n)$ for $1 \leq i \leq \ell_k$ is the number of subgraphs of $\mathcal{G}_n$ with adjacency matrix $\mathbf{M}_i^k$. The Shannon entropy of this probability distribution is:

$$\mathcal{H}_P(\mathcal{G}_n, k) = -\sum_{i=1}^{\ell_k} \frac{\mathsf{n}_i^k(\mathcal{G}_n)}{|\mathcal{P}_n(k)|} \log_2 \frac{\mathsf{n}_i^k(\mathcal{G}_n)}{|\mathcal{P}_n(k)|}.$$

If $\mathcal{H}_P(\mathcal{G}_n, k)$ is low, there are to be fewer and larger sets of isomorphic subgraphs of $\mathcal{G}_n$ of size k. In other words, small values of $\mathcal{H}_P(\mathcal{G}_n, k)$ for various values of k suggest that the graph $\mathcal{G}_n$ contains repeated patterns and is susceptible to produce interesting results. Note that although two isomorphic subgraphs do not necessarily have the same adjacency matrix, the number $\mathcal{H}_P(\mathcal{G}_n, k)$ is a good indicator of the frequency of isomorphic subgraphs and hence subgraph patterns.

We evaluated the correlation between $\mathsf{CR}_{jZIP}(\mathbf{A}_{\mathcal{G}_n})$ and $\mathcal{H}_P(\mathcal{G}_n, k)$ for different values of k.

As expected, the compression ratio of the adjacency matrix and the distribution entropy of graphs are roughly the same for isomorphic graphs, so both numbers are characteristic for an isomorphism type. If ϕ is a permutation of the vertices of $\mathcal{G}_n$, the adjacency matrix of the graph $\mathcal{G}_n^\phi$ obtained by applying the permutation is defined by $\mathbf{A}_{\mathcal{G}_n^\phi}$ is given by

$$\mathbf{A}_{\mathcal{G}_n^\phi} = P_\phi \mathbf{A}_{\mathcal{G}_n} P_\phi^{-1}.$$

We compute this adjacency matrix of $\mathbf{A}_{\mathcal{G}_n^\phi}$, the entropy $\mathcal{H}_P(\mathcal{G}_n^\phi, k)$ the compression ratio $\mathsf{CR}_{jZIP}(\mathbf{A}_{\mathcal{G}_n^\phi})$ for several values of k and permutations.

We randomly generated graphs with $n = 60$ nodes and various number of edges ranging from 5 to 1765. For each generated graph, we randomly produced twenty permutations of its set of nodes and computed $\mathcal{H}_P(\mathcal{G}_n^\phi, k)$ and $\mathsf{CR}_{jZIP}(\mathbf{A}_{\mathcal{G}_n^\phi})$.

Finally, for each graph we calculated the ratio of standard deviation over average for the computed compression ratios, followed by the same computation for distribution entropies.

The results of this experiment are shown in Figures 5 and 6 against the number of edges. As it can be seen, the deviation over mean of the compression ratios for $n = 60$ does not exceed the number 0.05. Also, the deviation over average of the distribution entropies for various values of k do not exceed 0.006. In particular, the deviation of the distribution entropy for the graphs of 100 to 1500 edges falls below 0.001, which allows us to conclude that the deviations of both compression ratio and distribution entropy with respect to isomorphisms are negligible.

For each $k \in \{3, 4, 5\}$, we generated randomly 560 graphs having 60 vertices and sets of edges whose size were varying from 10 to 1760. Then, the numbers $\mathcal{H}_P(\mathcal{G}_n, k)$ and $\mathsf{CR}_{jZIP}(\mathbf{A}_{\mathcal{G}_n})$ were computed. Figure 7 captures the results of the experiment. Each plot contains two curves. The first curve represents the changes in average $\mathsf{CR}_{jZIP}(\mathbf{A}_{\mathcal{G}_n})$ for forty randomly generated graphs of equal number of edges. The second curve represents the variation of the average $\mathcal{H}_P(\mathcal{G}_n, k)$ for the same forty graphs. The trends of these two curves are very similar for different values of k.

Table IV contains the correlation between $\mathsf{CR}_{jZIP}(\mathbf{A}_{\mathcal{G}_n})$ and $\mathcal{H}_P(\mathcal{G}_n, k)$ calculated for the 560 randomly generated graphs for each value of k.

 ISBN: 978-1-61208-242-4

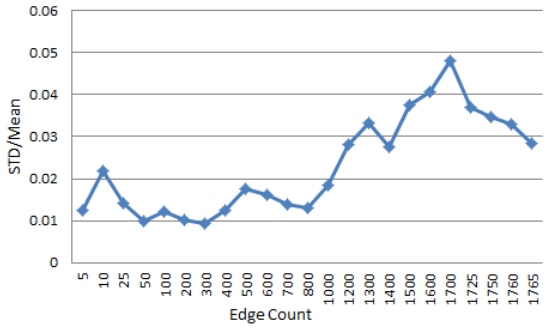


Figure 5. Standard deviation vs. average of the $\mathsf{CR}_{jZIP}(\mathbf{A}_{\mathcal{G}_n})$ for a number of different permutations of nodes for the same graph. The horizontal axis is labelled with the number of edges of the graph.

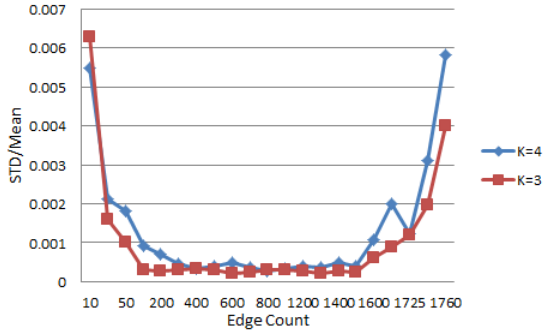


Figure 6. Standard deviation vs. average of the $\mathcal{H}_P(\mathcal{G}_n, k)$ of a number of different permutations of nodes for the same graph. The horizontal axis is labelled with the number of edges of the graph. Each curve corresponds to one value of k.

IV. Frequent Items Sets and Compression Ratio

A market basket data set consists of a multiset T of *transactions*. Each transaction t is a subset of a set of items $I = \{i_1, \ldots, i_N\}$. A transaction is described by its characteristic N-tuple $t = (t_1, \ldots, t_N)$, where

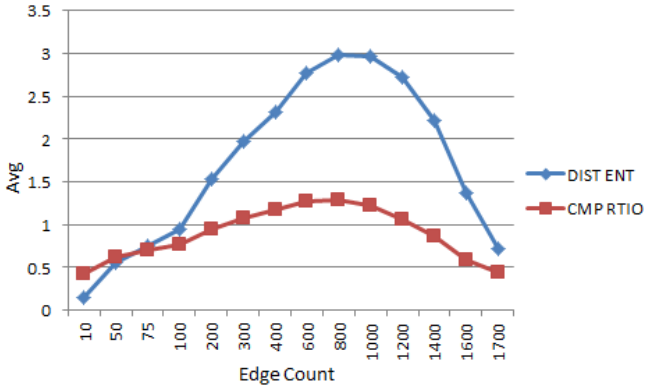
$$t_k = \begin{cases} 1 & \text{if } i_k \in t. \\ 0 & \text{otherwise,} \end{cases}$$

for $1 \leq k \leq N$. The length of a transaction t is $|t| = \sum_{k=0}^{N} t_k$, while the average size of transactions is $\frac{\sum\{|t| \mid t \text{ in } T\}}{|T|}$.

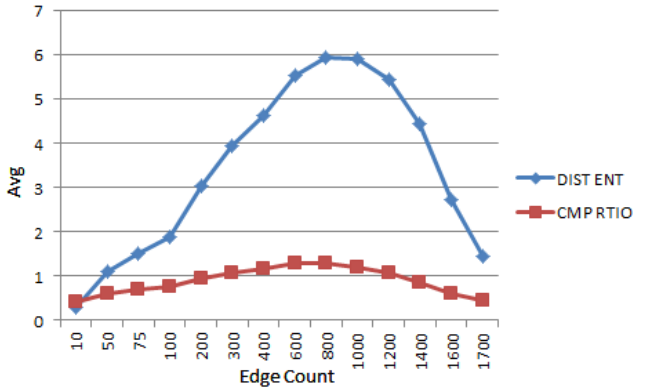
The support of a set of items K of the data set T is the number $\mathsf{supp}(K) = \frac{|\{t \in T \mid K \subseteq t|}{|T|}$. The set of items K is s-frequent if $\mathsf{supp}(K) > s$.

The study of market basket data sets is concerned with the identification of association rules. A pair of item sets (X, Y) is an association rule. Its support, $\mathsf{supp}(X \to Y)$ equals $\mathsf{supp}(X)$ and its confidence $\mathsf{conf}(X \to Y)$ is defined as

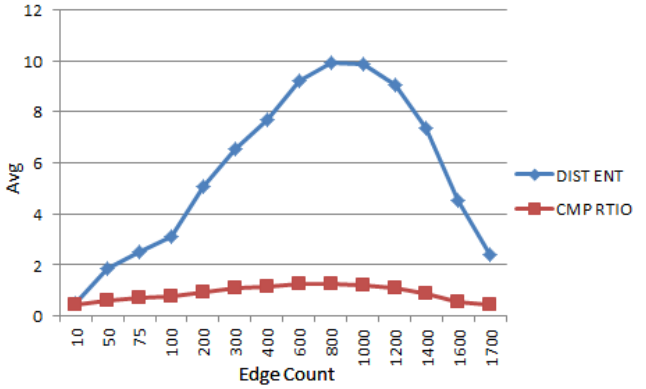
$$\mathsf{conf}(X \to Y) = \frac{\mathsf{supp}(XY)}{\mathsf{supp}(X)}.$$



$n = 60$ and $k = 3$



$n = 60$ and $k = 4$



$n = 60$ and $k = 5$

Figure 7. Plots of average $\mathsf{CR}_{jZIP}(\mathbf{A}_{\mathcal{G}_n})$ (CMP RTIO) and average $\mathcal{H}_P(\mathcal{G}_n, k)$ (DIST ENT) for randomly generated graphs $\mathcal{G}_n$ of equal number of edges with respect to the number of edges.

 ISBN: 978-1-61208-242-4

Table IV
CORRELATIONS BETWEEN $\mathsf{CR}_{jZIP}(\mathbf{A}_{\mathcal{G}_n})$ AND $\mathcal{H}_P(\mathcal{G}_n, k)$

k	Correlation
3	0.92073175
4	0.920952812
5	0.919256573

Using the artificial transaction ARMiner generator described in [3], we created a basket data set. Transactions are represented by sequences of bits $(t_1, \cdots, t_N)$. The multiset of M transactions was represented as a binary string of length MN obtained by concatenating the strings that represent transactions.

We generated files with 1000 transactions, with 100 items available in the basket, adding up to 100K bits.

For data sets having the same number of items and transactions, the efficiency of the compression increases when the number of patterns is lower (causing more repetitions). In an experiment with an average size of a frequent item set equal to 10, the average size of a transaction equal to 15, and the number of frequent item sets varying in the set $\{5, 10, 20, 30, 50, 75, 100, 200, 500, 1000\}$, the compression ratio had a significant variation ranging between 0.20 and 0.75, as shown in Table V. The correlation between the number of patterns and CR was 0.544. Although the frequency of 1s and baseline compression ratio were roughly constant (at 0.75), the number of patterns and compression ratio were correlated.

Table V
NUMBER OF ASSOCIATION RULES AT 0.05 SUPPORT LEVEL AND 0.9 CONFIDENCE

Number of Patterns	Frequency of 1s	Baseline compression	Compression ratio	Number of assoc. rules
5	16%	0.75	0.20	9,128,841
10	17%	0.73	0.34	4,539,650
20	17%	0.73	0.52	2,233,049
30	17%	0.76	0.58	106,378
50	19%	0.75	0.65	2,910,071
75	18%	0.75	0.67	289,987
100	18%	0.75	0.67	378,455
200	18%	0.75	0.70	163
500	18%	0.75	0.735	51
1000	18%	0.75	0.75	3

Further, there was a strong negative correlation (-0.92) between the compression ratio and the number of association rules indicating that market basket data sets that satisfy many association rules are very compressible

V. CONCLUDING REMARKS

Compression ratio of a file can be computed fast and easy, and in many cases offers a cheap way of predicting the existence of embedded patterns in data. Thus, it becomes possible to obtain an approximative estimation of the usefulness of an in-depth exploration of a data set using more sophisticated and expensive algorithms. The use of compression as a measure of minability is illustrated on a variety of paradigms: graph data, market basket data, etc. Recent investigations show that identifying compressible areas of human DNA is a useful tool for detecting areas where the gene replication mechanisms are disturbed (a phenomenon that occurs in certain genetically based diseases.

REFERENCES

[1] B. J. L. Campana and E. J. Keogh. A compression based distance measure for texture. In *SDM*, pages 850–861, 2010.

[2] R. Cilibrasi and P. M. B. Vitnyi. Clustering by compression. *IEEE Transactions on Information Theory*, 51:1523–1545, 2005.

[3] L. Cristofor. ARMiner project, 2000.

[4] E. Keogh, S. Lonardi, and C. A. Ratanamahatana. Towards parameter-free data mining. In *Proc. 10th ACM SIGKDD Intnl Conf. Knowledge Discovery and Data Mining*, pages 206–215. ACM Press, 2004.

[5] E. Keogh, S. Lonardi, C. A. Ratanamahatana, L. Wei, S. Lee, and J. Handley. Compression-based data mining of sequential data. *Data Mining and Knowledge Discovery*, 14:99–129, 2007.

[6] E. J. Keogh, L. Keogh, and J. Handley. Compression-based data mining. In *Encyclopedia of Data Warehousing and Mining*, pages 278–285. 2009.

[7] H. Mannila. Theoretical frameworks for data mining. *SIGKDD Exploration*, 1:30–32, 2000.

[8] L. Wei, J. Handley, N. Martin, T. Sun, and E. J. Keogh. Clustering workflow requirements using compression dissimilarity measure. In *ICDM Workshops*, pages 50–54, 2006.

[9] T. Welch. A technique for high performance data compression. *IEEE Computer*, 17:8–19, 1984.

 ISBN: 978-1-61208-242-4