HHS Public Access

Author manuscript

Proc Data Compress Conf. Author manuscript; available in PMC 2017 October 16.

Published in final edited form as:

Proc Data Compress Conf. 2017 April ; 2017: 455–. doi:10.1109/DCC.2017.82.

Compressing Tabular Data via Pairwise Dependencies

Dmitri S. Pavlichin[1], **Amir Ingber**[2], and **Tsachy Weissman**[1]

[1]Stanford University

[2]Yahoo! Research

Tabular datasets, such as server logs, business transactions, social media interactions and more, are very commonly generated and maintained across industries and organizations. Generic lossless compression algorithms, such as Lempel-Ziv and variants, are fast and robust, but do not exploit the unique structure of the data that can be learned from the source file. Previous work in this setting includes [1, 2] and references therein.

We propose a method and algorithm for lossless compression of tabular data, based on a method known as a Chow-Liu tree [3] with a minimum description length-like criterion for graph selection. The "vanilla" Chow-Liu tree approach captures pairwise dependencies between different columns (or more generally fits a Bayesian network to the dataset) by entropy coding with respect to a maximum spanning tree Bayesian network model on the features, with edge weights given by the empirical mutual informations $\hat{I}(X_i; X_j)$ between the features X_i, X_j. We improve on the Chow-Liu choice of tree by modifying the edge weights to account for the space to store the model description "metadata" – the pairwise joint empirical distributions – since in practice this cost can be large. Our choice of Bayesian network graph is T^*, optimized over all forest graphs $T = (V, E)$ on the features:

$$T^* = arg \max_{T=(V,E)} \sum_{(i \to j) \in E} \left(n\hat{I}(X_i;X_j) - \sum_{(i \to j) \in E} |c_n(\hat{p}_{i,j})| \right), \quad (1)$$

where n is the number of rows in the file and $|c_n(\hat{p}_{i,j})|$ denotes the length of an encoding c_n of an empirical pairwise joint histogram $\hat{p}_{i,j}$.

Our algorithm benefits from several features combined in a novel way: 1) efficient encoding of the (often sparse) empirical distributions by a combination of Golomb and arithmetic coding; 2) memory-efficient empirical mutual information approximation using hashing; 3) special handling of the (often many) values that occur only once in a file.

We test the algorithm on several datasets, and demonstrate an improvement in the compression rates of between 2X and 5X compared to gzip, columnar gzip, and bzip2. The larger improvements are observed for very large datasets, such as the Criteo click prediction dataset which was published as part of a recent Kaggle competition: 736MB vs 3.76GB for `gzip`.

References

1. Vo BD, Manku GS. RadixZip: Linear time compression of token streams. VLDB 2007. 2007:1162–1172.
2. Gao Y, Parameswaran A. SQUISH: Near-optimal compression for archival of relational datasets. Conference on Knowledge Discovery in Databases (KDD). 2016
3. Chow CK, Liu CN. Approximating discrete probability distributions with dependence trees. IEEE Transations on Information Theory. 1968; IT-14:462–467.