Fast Feature Selection using Fractal Dimension - Ten Years Later

Caetano Traina Jr.[1], Agma Traina[1], Christos Faloutsos[2]

[1] Department of Computer Science – University of São Paulo at São Carlos – Brazil
{caetano,agma}@icmc.usp.br
[2] Department of Computer Science – Carnegie Mellon University – USA
christos@cs.cmu.edu

Abstract. Here we comment about the works that the original paper published in the 2000 Brazilian Symposium on Databases fostered in the Database and Images Group – GBdI, what by their turn motivated other researches abroad. It is shown that the Fractal Theory is indeed helpful to a large spectrum of activities required to manage large amounts of data. Research derived from the original paper includes speeding up similarity queries, designing of cost models and selectivity estimation for similarity queries, sampling on databases, performing attribute selection, identifying clusters of correlated attributes, as well as correlation clustering on large, high dimensional datasets.

Categories and Subject Descriptors: Information Systems [**Database Management**]: Database applications

Keywords: Fractal Theory, Intrinsic dimensionality, Multi-scale space mapping algorithm

1. INTRODUCTION AND MOTIVATION

The original paper [Traina Jr. et al. 2000], published in the 2000 Brazilian Symposium on Databases, has two main directions that are, from our point of view, the reasons why it attracted the researchers' attention: the way it uses the Fractal Theory concepts to analyze large datasets; and the multi-scale space division algorithm proposed there. Both directions leaded to interesting works and present, until now, open areas for research. Following, we discuss some of these works, using the concepts and symbols already defined in the original paper.

2. USING THE FRACTAL THEORY TO ANALYSE LARGE DATASETS

One intuitive way to think about the data stored in relational databases is imagining that each tuple in a relation is a point in the space represented in the relation, where the number of dimensions E in the space is the number of attributes composing the meaningful key of the relation (by 'meaningful' we mean attributes that have a meaning for people, not the artificially generated tuple codes or 'object identifiers'). Saying that a relation has E key attributes means that its degree of freedom is E. For data traditionally stored in databases, the number of meaningful key attributes and thus the dimensionality of the data stored is usually low, most of them less them five or so. However, data from scientific experiments, characteristics extracted from complex data such as images, videos, genetic sequences, etc., often require that most or even all the attributes must be part of the key. Therefore, the number of key attributes can be very high, sometimes in the order of thousands.

The research reported in this paper has been funded by the Brazilian agencies FAPESP, CNPq and CAPES, and the American NSF agency. Any opinions, findings, and conclusions or recommendations expressed in this material are solely from the authors and not from the supporting entities.

Those who already worked with multidimensional data are aware of the "dimensionality curse", a problem arising when the data dimensionality scales up, which makes data processing too expensive in terms of the required computational power. However, the "real", "intrinsic" space providing by the data can have a much smaller dimensionality than the "embedded" dimension of the space where the data is represented. An example can help here. Let us suppose that we are measuring the movements of a performer artist arm, sensing the movements of his/her joints. It is common to assume that the human arm has eight degrees of freedom, so the arm positions can be stored in a relation with eight attributes, each measuring a joint angle. However, the position of a sensor in the tip of a finger can always be reported by just three numbers: either two angles and a radius or three coordinates in the 3-dimensional space where the arm lives in. This example shows that whereas the embedding space is $E = 8$, the intrinsic space is $D = 3$.

In fact, we cannot reach every point in the sphere centered at the elbow (our joints impose some restrictions on the allowable movements), so the dimensionality of the intrinsic space is lower than 3, whereas greater than 2. This intrinsic dimensionality is what the original paper [Traina Jr. et al. 2000] proposed to approximate using the fractal dimensionality, and developed a method to perform attribute selection employing it on relations without any special characteristics.

Works that followed explored both the spatial characteristics of the data and their organization in a Database Management System. For example, the use of the fractal dimensionality was explored to analyze data streams [Sousa et al. 2007], such as climate and remote sensing data streams [Romani et al. 2009] and results of association rule mining [Sousa et al. 2006]. We also employed the fractal concepts to define selectivity and cost models for similarity queries over both traditional data (numbers and short character strings) and multimedia data [Baioco et al. 2007].

Special characteristics of the data were also explored, for example considering that not every attribute is correlated to every other but, instead, there are groups of correlated attributes. Thus, the attribute selection task can be better achieved if the attribute groups can be spotted, for example, using the technique described in [Sousa et al. 2007], which also studied how individual restrictions of the attributes affect the global dimensionality. Going a step further, it was recognized that the data itself has clusters, and that the most important attributes for a given cluster can be distinct from the attributes that are important for other clusters. This situation is currently under study in our group, and preliminary results can be seen in [Cordeiro et al. 2010].

3. THE MULTI-SCALE SPACE DIVISION ALGORITHM

Two well-known techniques have been employed to evaluate the cumulative density function of the number of points reached from each point in the dataset for a varying radius. The two techniques are: calculate the pairwise distance between every pair of points from the dataset; and to approximate the number of points at each vicinity using a multi-resolution grid over the space covered by the dataset. Although both techniques are quadratic in the number of points N in the dataset, the second one is much faster, because does not require computation of distances.

A big improvement brought by the multi-scale space division algorithm proposed in the original paper was that it was the first (and up to now the only) algorithm in the literature able to evaluate the cumulative density function with a liner cost on N (the number of points in the dataset). The linear complexity of the algorithm motivated the use of the fractal dimension in several other applications, where speed requirements could preclude its use. An example can be seen in the original paper itself, as the fractal dimension of original space projections are evaluated iteratively. The linear complexity was fundamental for other applications as well, such as finding groups of correlated attributes [Sousa et al. 2007], where successive attribute rotations in increasingly higher-dimensionality projections of the original dataset demand the computation of their intrinsic dimension.

Another example where the linear complexity of the algorithm was fundamental appeared in [Traina

Jr. et al. 2006], where the fractal dimensionality of the subspace around the center of a similarity query is evaluated to define a limiting range radius for k-NN queries. Having a limiting radius significantly speeds up a k-NN query, but this is effective only if the fractal dimensionality can be evaluated really fast, otherwise it disturbs the search efficiency.

The concepts in the multi-scale space division algorithm was employed once more to perform density aware dataset sampling [Appel et al. 2007]. In that work, the idea of superimposing a grid over the space covered by the dataset is used not to evaluate its fractal dimensionality, but instead, to impose a minimum and maximum density of points at each space region, guaranteeing that even small clusters are well-represented.

4. FUTURE RESEARCH ISSUES

The last ten years has shown that the ideas of the original paper have broader application than its original intent on a variety of domains. However, there are several issues remaining. We discuss those unsolved issues that have struck worse, as follows.

The Intrinsic Dimensionality Curse. – The dimensionality curse is greatly reduced when it is possible to take advantage of the intrinsic dimension instead of the embedded one. However, the techniques to handle the fractal dimensionality are effective only if the intrinsic dimension itself is low, i.e., not higher than six or so. Therefore, novel techniques to tame the "intrinsic dimensionality curse" are yet to be developed.

Clusters of Fractals. – The fractal dimension is a single number describing the data distribution of the whole dataset. Although it is able to provide an indication that the data distribution is not uniform, and even what kind of distribution it is, the fractal dimension is not sufficient to spot if distinct dataset regions have distinct distributions. Thus, associating the fractal dimension concept to clustering techniques is an intriguing research area.

Inability to handle Small Datasets. – The techniques based on the fractal theory requires recognizing the fractal represented by the data. Real datasets usually include noise, usually in a small amount, but nonetheless enough to introduce instability in the algorithms, which can be overpassed only if large volumes of data are employed. This problem gets stronger as the fractal dimension increases. Thus, datasets with high intrinsic dimension can be analyzed if they are large in the number of points too. As the research focus is now pointing to clustered data, and a large amount of clusters means a smaller number of points in each cluster, the problem is expected to be even more relevant. It is, however, interesting to note that this problem runs in the opposed direction from the majority of the other analysis techniques: whereas large datasets are a problem for the competing techniques, the fractal-based techniques suffers only when the datasets are too small.

REFERENCES

Appel, A. P., Paterlini, A. A., Sousa, E. P. M. d., Traina Jr., C., and Traina, A. J. M. A density-biased sampling technique to improve cluster representativeness. In *Proceedings of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases.* Varsovia, Polonia, pp. 366–373, 2007.

Baioco, G. B., Traina, A. J. M., and Traina Jr., C. Mamcost: Global and local estimates leading to robust cost estimation of similarity queries. In *Proceedings of the 19th International Conference on Scientific and Statistical Database Management.* ACM Press, Banff, Canada, pp. 6–16, 2007.

Cordeiro, R. L. F., Traina, A. J. M., Faloutsos, C., and Traina Jr., C. Finding clusters in subspaces of very large, multi-dimensional datasets. In *Proceedings of the 26th IEEE International Conference on Data Engineering.* IEEE Computer Society, Long Beach, California, pp. 625–636, 2010.

Romani, L. A., Sousa, E. P. M. d., Ribeiro, M. X., Zullo Jr., J., Traina Jr., C., and Traina, A. J. M. Employing fractal dimension to analyze climate and remote sensing data streams. In *Proceedings of the First SIAM SDM Workshop on Multimedia Data Mining.* Vol. 1. SIAM, Sparks, Nevada, pp. 5–16, 2009.

Sousa, E. P. M. d., Ribeiro, M. X., Traina, A. J. M., and Traina Jr., C. Tracking the intrinsic dimension of evolving data streams to update association rules. In *Proceedings of the 3rd International Workshop on Knowledge Discovery from Data Streams.* Pittsburgh, PA, pp. 643 – 648, 2006.

Sousa, E. P. M. d., Traina, A. J. M., Traina Jr., C., and Faloutsos, C. Measuring evolving data streams'behavior through their intrinsic dimension. *New Generation Computing* 25 (1): 33–60, 2007.

Sousa, E. P. M. d., Traina Jr., C., Traina, A. J. M., Wu, L., and Faloutsos, C. A fast and effective method to find correlations among attributes in databases. *Data Mining and Knowledge Discovery* 14 (3): 367–407, 2007.

Traina Jr., C., Traina, A. J. M., Vieira, M. R., Arantes, A. S., and Faloutsos, C. Efficient processing of complex similarity queries in rdbms through query rewriting. In *Proceedings of the ACM 15th International Conference on Information and Knowledge Management*. ACM Press, Arlington - VA, USA, pp. 4–13, 2006.

Traina Jr., C., Traina, A. J. M., Wu, L., and Faloutsos, C. Fast feature selection using fractal dimension. In *Proceedings of the Brazilian Symposium on Databases*. João Pessoa, pp. 158–171, 2000.