Dataset visualization based on a simulation of intermolecular forces

Jan Drchal[1], Pavel Kordík[1], Miroslav Šnorek[1]

[1]*Dept. of Computer Science and Engineering, Karlovo nám. 13, 121 35 Praha 2, Czech Republic*

drchaj1@fel.cvut.cz, kordip@fel.cvut.cz, snorek@fel.cvut.cz

Abstract. *The visualization is an important technique used in many stages of data mining process. This article deals mostly with visualization for preprocessing purposes. The aim of our approach is to visualize distances (Euclidean or others) between data samples. This can be helpful when taking picture of data clustering. In classification tasks it can be used to select outlayer for removal. In this paper we present a novel way of such visualization which is based on a physical system simulation. It is inspired by intermolecular forces and employs overall energy minimization. This minimization is done via known unconstrained optimization numerical methods such as Steepest Descent, Conjugated Gradients or Quasi-Newton. The proposed algorithm was originally designed and was found useful when interpretting diversity in evolutionary algorithms. Here, we show its properties on well-known datasets Iris and Ecoli.*

Keywords

Data mining, visualization, optimization.

1 Introduction

This article presents a visualization technique which can be useful in data mining, mostly in preprocessing. It is able to display a dataset, showing similarities (distances) between its data rows. In other words it performs a dimension reduction and a projection to 2-D or 3-D, approximating the original distances between data. The algorithm has been originally designed to visualize diversity of population in evolutionary algorithms [5].

We have organized the paper as follows. In Section 2 visualization algorithm is presented. In Section 3 we show experimental results. The last section concludes.

2 Visualization

2.1 Distance matrix

Our approach is based on a visualization of distances between all N data samples. We have mostly experimented with Euclidean metric but it is possible to use other – for example Manhattan metric. Distances can be represented in a matrix form. The distance matrix D is symmetric with zero diagonal:

$$d_{ii} = 0, \tag{1}$$

$$d_{ij} = d_{ji} \tag{2}$$

for $i, j = 1..N$. In a 2-D case for a given distance matrix D we are trying to find representative points (x_i, y_i) in the plane such that the 2–D distances $l_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$ *approximate as closely as possible* the original distances d_{ij}. An example is shown in the Figure 1. Our visualization does similar job to Sammon-projection [6]. Result comparison is now in progress.

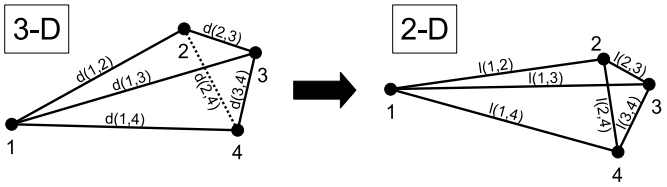

Fig. 1. Dataset visualization example. Four data rows can be projected to 3-D while preserving their mutual distances. Our algorithm performs dimension reduction: new distances l_{ij} approximate original distances d_{ij} in 2-D.

2.2 Physical system simulation approach

The solution we propose is based on modeling of a physical phenomena. It is inspired by intermolecular forces [7]. Atoms in molecules are exposed to attractive and repulsive forces. Attractive forces prevail for long distances while repulsive forces for short. We used the above scheme, associating data samples with atoms, and modified it to:

1. the shorter the distance d_{ij} is, the stronger the attractive force is,
2. the shorter the projected 2-D (3-D) distance l_{ij} is, the stronger the repulsive force is.

This system can be described by energy and force equations. The total energy of such system should be minimized:

$$\Phi_{TOTAL} = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \Phi_{ij}, \tag{3}$$

where Φ_{ij} is an energy contribution from data samples i and j. The situation is depicted in the Figure 2. The force contribution F_{ij} can be evaluated as a gradient of Φ_{ij} taken with minus sign. We have developed two different sets of energy and force equations (**A** and **B**). The first marked **A** is

$$\Phi_{ij}^{A} = \left(\frac{l_{ij}}{d_{ij} + d_0} - \frac{\arctan\left(\frac{l_{ij}}{a}\right)}{a} \right), \tag{4}$$

$$F_{ij}^{A} = -\frac{\partial \phi_{ij}^{A}}{\partial l_{ij}} = -\frac{1}{d_{ij} + d_0} + \frac{1}{l_{ij}^2 + a^2}. \tag{5}$$

The equations of **B** read

$$\Phi_{ij}^{B} = \frac{1}{l_{ij} + a} + \frac{l_{ij}}{(d_{ij} + a)^2} - \frac{2}{d_{ij} + a}, \tag{6}$$

$$F_{ij}^{B} = -\frac{\partial \phi_{ij}^{B}}{\partial l_{ij}} = \frac{1}{(l_{ij} + a)^2} - \frac{1}{(d_{ij} + a)^2}. \tag{7}$$

The difference between these sets is in the influence of d_{ij} for longer distances (in **A** the influence is stronger). Both **A** and **B** were found empirically, however, they are very similar to the real-world equations [7]. The elements d_{ij} of the distance matrix D are always normalized in order to keep parameter settings uniform.

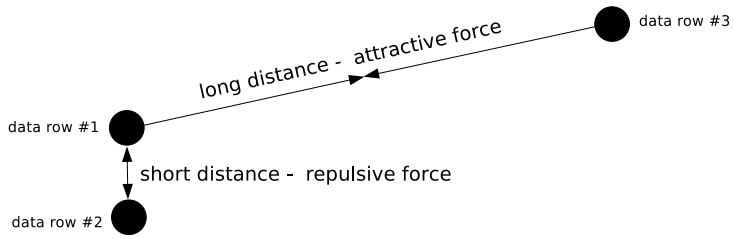

Fig. 2. Intermolecular forces inspiration. Atractive forces prevail for short atom distances, repulsive for long distances. We have associated data rows with atoms. Our algorithm searches for equilibrium of the system. The equilibrium is found via minimization of overall system energy.

2.3 Visualization algorithm

Our visualization algorithm is presented in the Figure 3. At first the dataset is standardized (mean values of all features become 0 and their standard deviations 1). Then the distance matrix D is computed using chosen metric (Euclidean, Manhattan, etc.). After this step we choose an equation set (**A** or **B**). Finally the energy of the system is minimized using an unconstrained optimization algorithm – each data row is found its projection coordinates (x_i, y_i) (or (x_i, y_i, z_i) in the 3-D case). We have tested Steepest Descent, Conjugated Gradient and Quasi-Newton [8, 2, 9, 4, 10]. Conjugated Gradient seems to be the best choice as it was fastest and most reliable for larger datasets [5].

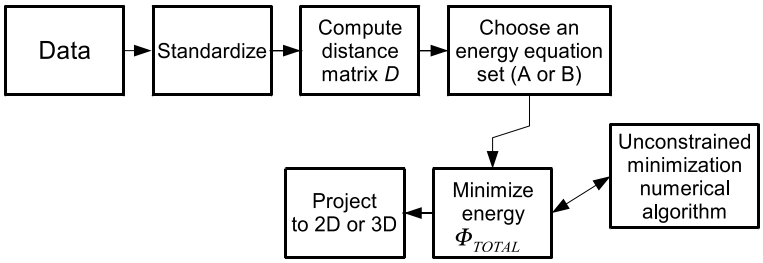

Fig. 3. Visualization algorithm flow. At first the dataset is standardized. Then, using selected metric (Euclidean or other) we compute a distance matrix D – distances between all data rows. We choose a set of equations (**A** or **B**) and continue with minimization of energy Φ_{TOTAL}. The minimization is done via an unconstrained minimization numerical algorithm. At last the data samples are visualized on a plane or in 3-D. Similar data rows are visualized as clusters. This information can be used to determine data clustering or to select outlayer for removal.

3 Experiments

In all following experiments we have used Euclidean metric. Conjugated Gradient as implemented in PAL (Phylogenetic Analysis Library) [1] was used to minimize energy.

At first, we have tested our algorithm on an artificially generated distance matrix D which reflected user supplied number and sizes of clusters. The distance matrix was generated so that intra-cluster distances were 0.1 and inter-cluster distances were 0.9. The result can be seen in the Figure 4. It shows that equation set **B** behaves better – the smallest cluster of size 2 is not broken in projection. On contrary we found the optimization of equation set **B** much slower than equation set **A**.

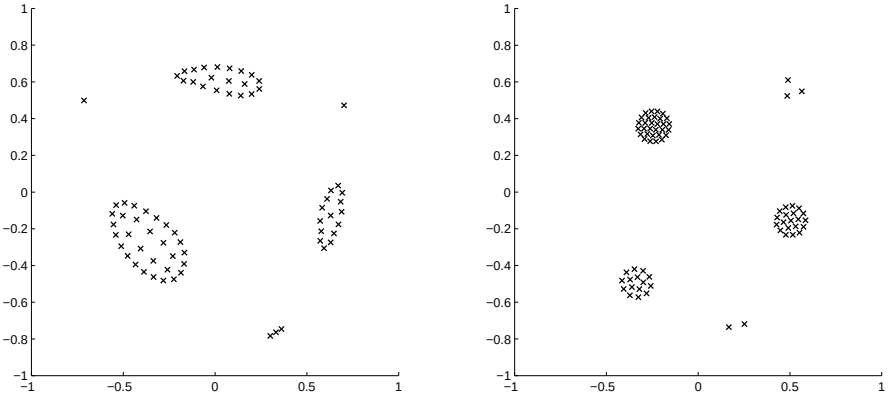

Fig. 4. Small coarse clusters. Equation set **A** (left), equation set **B** (right). Each mark represents a single data sample. Five clusters of different sizes $(2, 3, 15, 20$ and $30)$ were generated. The intra-cluster distance d_{ij} was set to 0.1 the inter-cluster distance was 0.9. Notice, that equation set **A** broke the smallest cluster of size 2.

The first experiment taken on real-world data was taken on a well-known Iris dataset [3]. The results are shown in the Figure 5. All three classes: Setosa (red), Versicolor (green) and Virginica (blue) are shown as distinct clusters. Versicolor and Virginica are known to be not linearly separable which is clearly reflected by the visualization (green and blue clusters overlap).

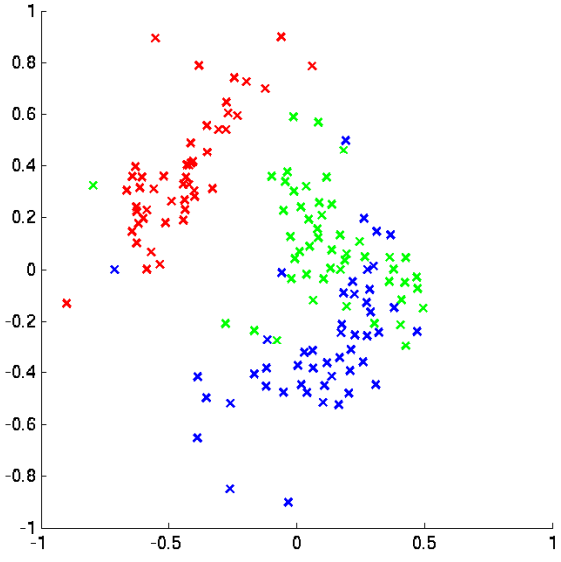

Fig. 5. Iris dataset, equation set **B**. Three classes: Setosa (red), Versicolor (green) and Virginica (blue) form clusters. Versicolor and Virginica are known to be not linearly separable which is clearly reflected by the visualization (green and blue clusters overlap).

In this place we have found that it might be useful to alter distance matrix so that small distances are made even smaller (closer to 0) and larger distances even larger (closer to 1). For this purpose we have used a slightly modified "squashing" function known from the area of neural networks:

$$d_{ij}^S = \frac{1}{1 + \varepsilon^{-\alpha(d_{ij}-o)}}, \tag{8}$$

where α is a small positive number (e.g. 2.0) called gain and o an offset (typically 0.5). The results are shown in the Figure 6. Now the clusters are even more evident.

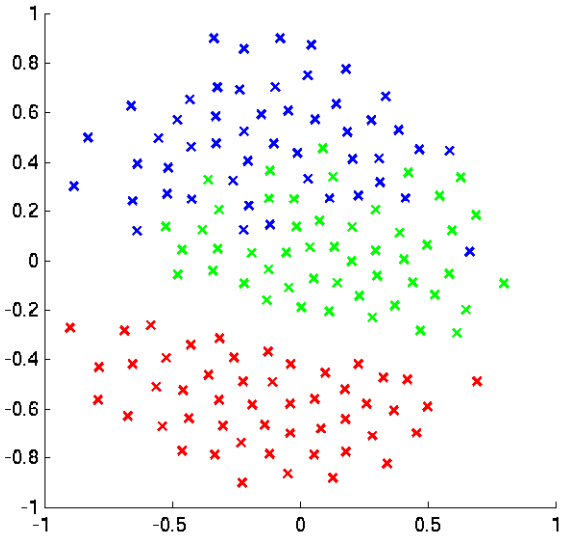

Fig. 6. Iris dataset with squashing function applied to the distance matrix D. The squashing function makes clustering more evident.

We have also performed experiments on Ecoli dataset [3]. The results are shown in Figures 7 and 8. Eight projected classes form distinct clusters again.

At last we have experimented with 3-D visualization. The results are shown in Figures 9, 10 and 11.

4 Conclusion

This article proposes a novel method of a dataset visualization in data mining based on modeling of intermolecular forces. We have created two sets of energy equations to simulate the system. Our experiments have shown visualization properties on both artificial and real-world data.

There are many areas to be explored. Our current focus is on comparison with Sammon-projection. Also, our algorithm is now able to visualize only datasets which contain up to several thousand rows – for 1000 data samples it has to optimize a search-space of dimension 2000 when projecting to 2-D and a search-space of dimension 3000 when projecting to 3-D. This drawback can be partialy solved by projecting only a selected subset of data. It would be also interesting to make experiments on metrics where feature significance is somehow weighted (adjusted).

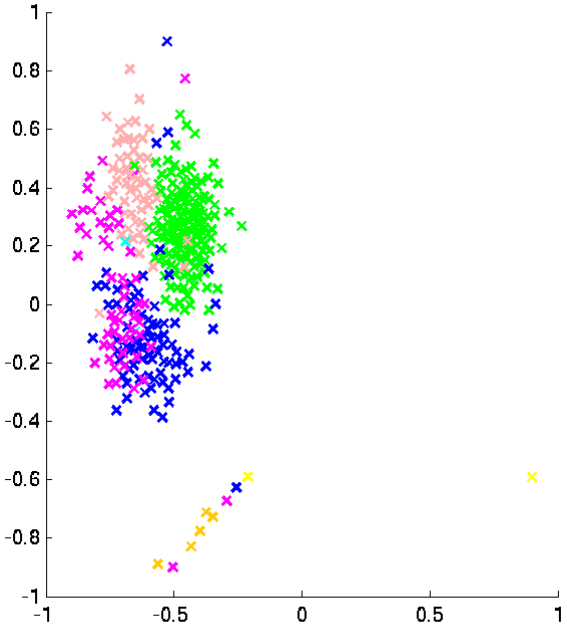

Fig. 7. Ecoli dataset, equation set **B**. Eight classes of different colors shown as clusters.

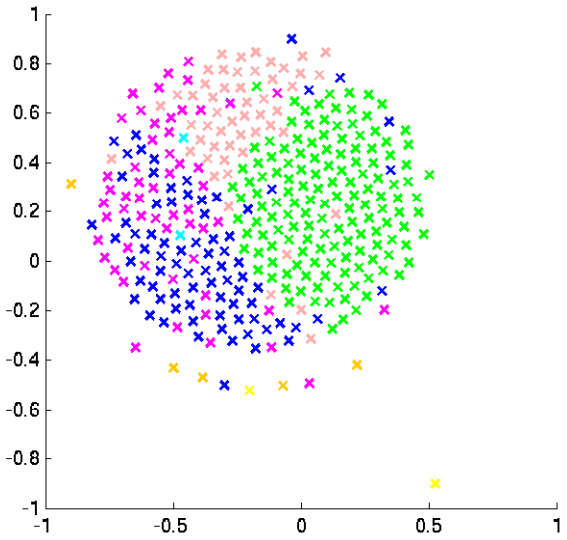

Fig. 8. Ecoli dataset with squashing function applied to the distance matrix D.

Acknowledgements

This research is partially supported by the grant Automated Knowledge Extraction (KJB201210701) of the Grant Agency of the Academy of Sciences of the Czech Republic, the research program "Trans-

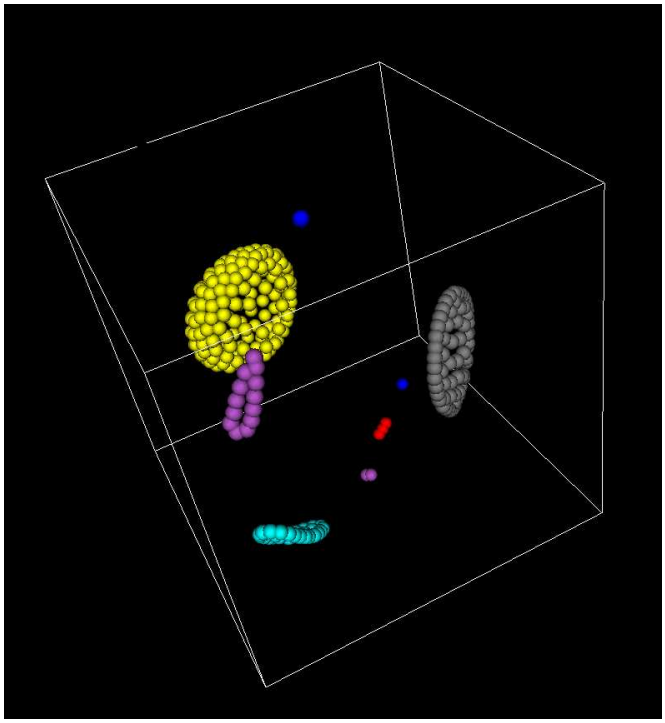

Fig. 9. Atificial coarse clusters in 3-D. Equation set **A**. Data samples are represented by spheres. Seven clusters of different sizes $(2, 3, 15, 20, 30\ 75\ 175)$ are distinguished by different colors. The intra-cluster distance d_{ij} was set to 0.1 the inter-cluster distance was 0.9.

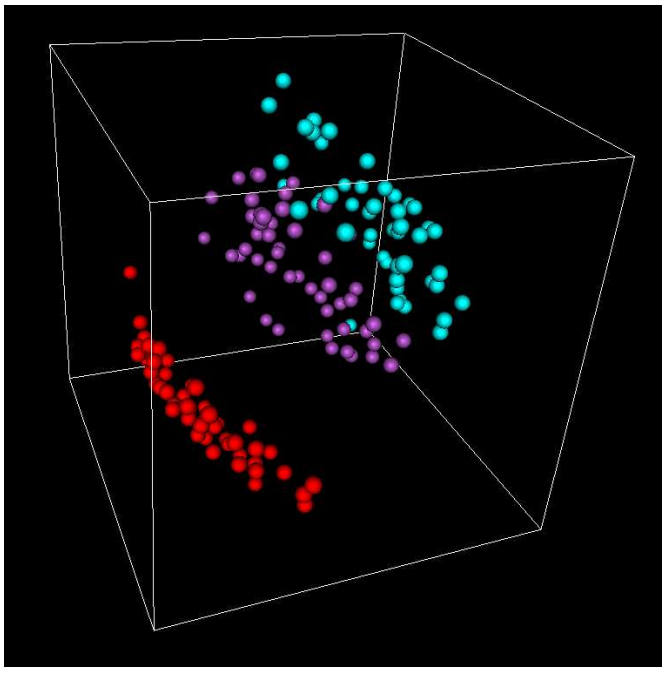

Fig. 10. Iris dataset in 3-D. Equation set **A**.

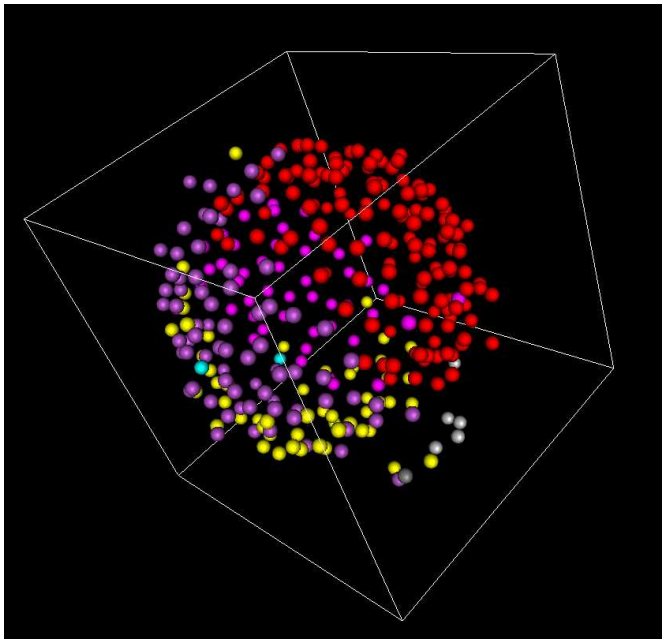

Fig. 11. Ecoli dataset in 3-D. Equation set **A**.

disciplinary Research in the Area of Biomedical Engineering II" (MSM6840770012) sponsored by the Ministry of Education, Youth and Sports of the Czech Republic and by the CTU IGS under grant CTU0707013.

References

[1] PAL: Phylogenetic analysis library. http://www.cebl.auckland.ac.nz/pal-project/index.html.

[2] J.F. Bonnans, J.C. Gilbert, C. Lemaréchal, and C.A. Sagastizábal. *Numerical Optimization: Theoretical and Practical Aspects*. Springer-Verlag, Berlin Heidelberg, Germany, 2003.

[3] C.L. Blake D.J. Newman, S. Hettich and C.J. Merz. UCI repository of machine learning databases, 1998. http://www.ics.uci.edu/∼mlearn/MLRepository.html.

[4] R. Fletcher. *Practical Methods of Optimization Vol.1: Unconstrained Optimization*. John Wiley & Sons, New York, USA, 1980.

[5] M. Šnorek J. Drchal. Diversity visualization in evolutionary algorithms. In J. Štefan, editor, *Proceedings of 41th Spring International Conference MOSIS 07, Modelling and Simulation of Systems*, pages 77–84. Ostrava: MARQ, 2007.

[6] J. W. Sammon Jr. A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, C-18(5):401–409, 1969.

[7] W.J. Moore. *Physical Chemistry*. Prentice Hall, New York, USA, 1972.

[8] J. Nocedal and S.J. Wright. *Numerical Optimization*. Springer-Verlag, New York, USA, 1999.

[9] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery. *Numerical Recipes in C: the art of scientific computing 2nd ed.* Cambridge University Press, Cambridge, 1992.

[10] R.B. Schnabel, J.E. Koontz, and B.E. Weiss. A modular system of algorithms for unconstrained minimization. *ACM Transactions on Mathematical Software*, 11(4):419–440, December 1985.