A Cooperative Biomimetic Approach for High Dimensional Data Mining

Lydia Boudjeloud
LITA EA 3097
Laboratoire d'Informatique Théorique et Appliquée
Université Paul Verlaine - Metz
Ile du Saulcy, F-57045 Metz Cedex 01, France
boudjeloud@univ-metz.fr

Hanane Azzag
LIPN UMR CNRS 7030
Laboratoire d'Informatique de Paris-Nord
Institut Galilée, 99 Av. J.B. Clément
F-93430 Villtaneuse, France
hanane.azzag@lipn.univ-paris13.fr

ABSTRACT

We propose in this paper an original alternative to solve the problem of search space visualization to discover the complex structure of data, while respecting topology. Our cooperative approach provided a multi-dimensional visualization from the data. The first method is the subspace selection from whole data space. This selection is obtained by a genetic algorithm reducing the data dimension space by simply determining the most relevant dimensions evaluated by a distribution measure. Once a subspace selected we construct a neighborhood graph using artificial ants algorithm.

Categories and Subject Descriptors

I.2 [**ARTIFICIAL INTELLIGENCE**]
; I.5.3 [**Clustering**]

General Terms

Algorithms

Keywords

Visual data mining, Evolutionary approach, Ant algorithm, High dimensional data.

1. INTRODUCTION

The growth of databases has far outpaced the human ability to interpret this data creating a need for automated analysis of databases. Knowledge Discovery in Databases (KDD) is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [2]. The KDD process is interactive and iterative, involving numerous steps. Data mining is one of the steps of KDD which has attracted a lot of research. This paper focus on a combined approach which uses first a genetic algorithm-based search technique that can be used in data mining and more particularly in attribute selection for high dimensional application (with up one million dimensions) [1]. Secondly we use artificial ants model to build and to visualize a large proximity graphs of the used data [4].

GECCO'11, July 12–16, 2011, Dublin, Ireland.
ACM 978-1-4503-0690-4/11/07.

2. A COOPERATIVE APPROACH

2.1 Evolutionary algorithm

To select a pertinent dimension subset, we use in this work an evolutionary search [1]. Our genetic algorithm starts with a population of 60 individuals (chromosomes), every individual is made of n genes (n is user-defined), $n < 10$. The individuals are made of a subset of dimensions that describe the dataset. We evaluate each chromosome of the population by SE to maximize.

$$SE = \frac{(\beta^2 + 1).CH.\frac{R_S}{R_T}}{(\beta^2.CH) + \frac{R_S}{R_T}} \tag{1}$$

$$R_S = \sum_{i=1}^{k} \frac{N}{N_{i_S}} \tag{2}$$

$$R_T = \sum_{i=1}^{k} \frac{N}{N_{i_T}} \tag{3}$$

$CH = (SSB/(k-1))/(SSW/(N-k))$, k represents the cluster number, N the whole dataset cardinality, SSW refers then to the within group sum of squares and SSB refers to the between group sum of squares.

R_S and R_T measures represent the inverse of the harmonic mean of the data point distribution in different clusters (N_{i_S}) in the subspace S and (N_{i_T}) in the whole dataset T. R_T has a fixed value and we search the subspace S that obtains the best R_S value according to the user task (clustering or outlier detection).

If $RS/RT = 1$, we obtain the same data distribution in the clusters in T and S, when RS/RT is around 1, only some elements (they are near the frontiers) swap between clusters. In this case, S is the optimal attribute subset that represents (T) the whole dataset in term of clustering. This subspace represents more clearly the data distribution. But when we search the maximal value of RS/RT we obtain clusters that can contain outliers.

b is a weighting parameter controlling the relative importance of the two aims in the evaluation. If $b = 1$ for instance, SE gives a same weight for $\frac{R_S}{R_T}$ and CH.

Once the population is evaluated and sorted, we operate a crossover on two parents chosen randomly. Next, one of the children is mutated with a probability of 1/10, and an individual is substitute randomly in the second part of population, under the median. The genetic algorithm finishes after a maximum number of iterations, or after a maximum number of crossovers or of mutations, without improvement of the solution.

2.2 Graph proximity and visualization algorithm

To build the graph proximity, we use in this work an artificial ants model, called AntGraph [4], which is a generalization of the previous building principle for construction of a proximity graph with the same desired performances. Initially AntGraph offers a visual partitioning of a dataset under the form of a proximity graph. AntGraph builds a proximity graph with artificial ants from a dataset. First we consider a set of data as a stream where each data is an ant. Initially an ant denoted as a_1 is selected as fixed support (point of entrance) in the graph. It is the starting point from which the graph will be built. Then while it exists unconnected ants in the stream (data not connected in the graph), we consider an ant denoted ai which is inserted in the graph at a_1. This ant moves from ant to ant until it connects to the ant which is the most locally similar (and then the next ant is simulated). Thus an ant follows the path of maximum similarity. The ants optimize their moving and can quickly cross some large amounts of data. It allows the algorithm to cut through the complexity of large datasets: when adding one ant/data, one considers only a single path rather than all data.

Once a proximity graph is built with AntGraph from a dataset, we visualize it with a force layout algorithm [3], obtaining an esthetic display to exploit knowledge information. This approach defines a set of forces which are exerted between nodes of the graph: some repulsion forces between all nodes whereas attraction forces are applied between adjacent/similar nodes/data. We finally profit an aesthetical disposition of graph nodes. We also visualize the graph structure with a distinction between the global shape and some local shapes. In the first case, we perceive the general form of the proximity graph. In the other case, we can discover with a content-based navigation the real proximity between data. The user may easily visualize the graph and interact above. Several interactions are available for the user (i.e. zoom, selection, navigation) to act on nodes/data.

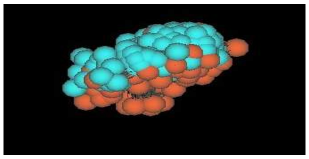
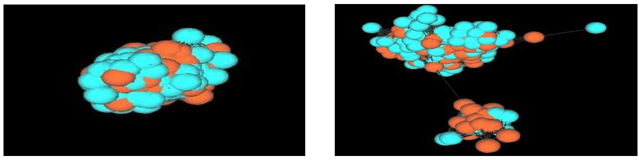


Figure 1: Visualization of Ovarian data set neighborhood graph in whole dimension set.

3. TESTS AND RESULT VISUALIZATION

In follows, we try to see the effective cooperation of two biomimetic approaches in terms of knowledge discovery and effeciency. In this work, we are interested to the form of graphs and not the clustering result. Once the dimensions selection procedure performed, we present to the *AntGraph* algorithm a data subspace to produce neighborhood (proximity) graph visualization. Therefore we compare the visualizations obtained on the whole dimensions set and those obtained in different dimensions subsets. We note that in some cases (depending on the selected sub-space and the b parameters of the Subset Evaluation) we approach the global solution, and in other cases we obtain subspaces that reveal characteristics of data can't be seen in the total data space (more homogeneous groups, outliers . . .).



(a) Ovarian 9 dimensions. (b) Ovarian 5 dimensions.

Figure 2: Visualization of Ovarian data set neighborhood graph in different dimensions subsets.

4. CONCLUSION

We presented in this paper a new way to solve the problem of finding a space visualization to explore complex structure of data. We have presented a genetic algorithm for dimension selection in high dimensional data combined with artificial ants. This combined approach offers a visual partitioning of a dataset using a proximity graph. Tests conducted on several datasets with high dimensions show the effectiveness of combined biomimetic approaches. We obtain subspaces that reveal some characteristics of data that we can't clearly seen in the total space of data (more homogeneous groups, outliers). These preliminary experiments are encouraging and several prospects may arise from this work. Interactive clustering can be obtained, user will evaluate interactively according to the information that he wishes to obtain on such dataset. We could also extend interactivity to identify outliers. We also want to provide a cooperative method for merging the obtained graph to find the graph that best represent the data.

5. REFERENCES

[1] L. Boudjeloud and F. Poulet. Attributes selection for high dimensional data clustering. In *proc. of International Symposium on Applied Stochastic Models and Data Analysis*, pages 387–395, 2005.

[2] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. In *AI Magazine*, volume 17-3, pages 37–54, 1996.

[3] T. Fruchterman and E. Reingold. Graph drawing by force-directed placement. *Software–Practice and Experience*, 21(11):1129–1164, 1991.

[4] J. Lavergne, H. Azzag, C. Guinot, and G. Venturini. Incremental construction of neighborhood graphs using the ants self-assembly behavior. *Tools with Artificial Intelligence, IEEE International Conference on*, 1:399–406, 2007.