

Clustering based on genetics algorithm

Dariusz Mazur

Silesian University of Technology,
Faculty of Organisation and Management,
ul. Roosevelta 26, 41-800 Zabrze, Poland
e-mail: dmazur@polsl.gliwice.pl

Keywords: clustering, algorithm, decision list, information retrieval

1. Introduction

In this paper is proposed a new approach for clustering categorical data, different from previous techniques in several fundamental respect:

1. No a priori quantization.
2. Computation without need of similar function.
3. Ability do discover clusters with different shapes.
4. New method for genetics clustering.

There is adopted fundamentally different, approach to the problem of clustering sets; it is motivated by genetics methods, a powerful methods for *searching* problems.

2. Genetically guided clustering

Genetic Algorithms (GAs) have been fairly successful at solving problems of this type that are too ill-behaved (such as multi modal and/or non-differentiable) for more conventional hill-climbing and derivative based techniques. They are not guaranteed to find the global optimum solution to a problem, but they are generally good at finding acceptably good solutions to problems acceptably quickly. There are controlled by several inputs, such as size of population, ways to encode a potential solution as a chromosome, choice of modification operators. Such of these choices are better suited to a particular problem than others, and no single choice is the best for all problems. GAs have had a great measure to success in search and optimization problems. The reason for a great part of this success is their ability to exploit the information accumulated about an initially unknown search space in order to bias subsequent searches into useful subspaces, i.e., their *adaptation*.

This paper introduces an evolutionary algorithm of clustering based on decision list. There has been several proposals of genetic operators designed particularly for rule discovery. Although these genetic operators have been used mainly in the classification task, in general they can be also used in other tasks that involve rule discovery, such as dependence modeling.

Mutation is a common reproduction operator used for finding new points in then search space to evaluate. When a chromosome is chosen for mutation, a random choice is made of some of the genes of the chromosome, and these genes are modified. It is proposed to introduce certain variant of mutation, which is based on random choosing two elements

from the list and swapping them (there is sort of permutation). The observed feature of algorithm was used in order to increase efficiency of such mutation. Part of then rules list is inactive because during transcribing process only leading rules are taken into consideration and there is possibility to define which rule is the last and divides then list into two parts: active and inactive. The order of the rules in the second, inactive part does not matter for the transcribing process since these rule are not participate in the process. In order to use this characteristic the mutation function guarantees that one of the chosen element always comes form active part of the list.

The key difference between this operator and classic mutation operator is the information which each attempts to preserve during recombination. For the clustering problem the important information would seem to be the adjacency information. This operator explicitly preserves adjacency and relative order information. Information about absolute positions appears to be relatively unimportant.

3. Decision list

Decision lists play a substantial role in machine learning. They were first suggested in [16]. Decision rules are in the form of decision rules sets in which order has been established, i.e. the sequence in which the rules are to be used to classify the examples. The hypothesis represented by such a set of rules assigns a category connected with the first rule in the sequence to the classified pattern, which entails the given pattern. Decision lists can also be treated as a degenerated decision tree for which out of every node two branches are produced, one coming to a leaf and the other to another node.

4. Clustering representation

In hierarchical algorithms the solution is represented by a graph in form of an upturned tree (dendogram). In such a tree the leaves represent particular objects and the branches reflect particular groups. The number of groups depends on the height of the tree. The groups determine all branches at a particular level. The leaves belonging to different branches represent objects belonging to different groups.

In the proposed method a significant alteration is suggested: the graph will be created on the basis of a list of discriminant descriptions, analogously as in CN2 algorithm [6]. The same quality measure based on entropy function from information theory is also applied. As mentioned above, the obtained solution has the form of an ordered decision rules

list. The notion of rules is based on the theory of *symbolic classification*. The form of the rules is:

$$D_j ::= K_i, \quad (1)$$

where D_j is the descriptive rule, K_i represents the cluster, and $::=$ the assignation operator. The rules can be interpreted as follows: *If a given object fulfills the D_j condition it then belongs to K_i cluster.* In order to adapt the notion of rules in document set clustering the following assumption was made. The rule will have the form of the occurrence of a particular term in the text, that is, if the term used in rule D_j occurs in the document X_n it is then assigned to K_i cluster. The rules are represented in the form of an ordered list. The process of assigning objects to cluster has the following procedure.

5. Conclusions

The clustering problem is a very important problem and has attracted much attention of many researches. Genetics algorithms are providing themselves in solving real problems in data mining, especially in cases where data are noisy, requires the solution of multi-objective optimization problem or data are too ill-behaved (such as multimodal and/or non-differentiable) for more conventional hill-climbing and derivative based techniques. In this paper is shown new approach to genetically based clustering using decision lists.

A genetic clustering based on genetics algorithms is proposed. The traditional neighborhood clustering algorithm usually needs the user to provide distance d for the clustering. But a unique d for a set of objects often cause problems because there may be some natural cluster in which the objects are not close to one another within the distance d . Proposed algorithm avoids this kind of problem by processing the data in a global view.

Clustering through the use of attributes brings in a highly innovative quality in terms of interpretation and further analysis of results of clustering. In place of a long table of object assignations (of equal length to the number of objects) a relatively short list of rules is obtained. The fact of decreasing the script of the result has a positive influence on the possibility of further adaption of the result by a man. In addition, the list is arranged so that it allows assigning certain relevance features to particular rules. This quality can be very effectively used for comparing the results of clustering in sets of minor differentiation. A good example for such a notion is comparing the same phenomenon but at different time intervals. The sets of objects representing the given phenomenon are altered and modified with time. Data mining aims at understanding and explanation of these changes. One of the methods is the analysis of the differences between the starting and final states of a given phase. If this notion is interpreted in terms of the clustering presented above, the analysis of the obtained lists of rules, given the fact of the relative position of the rules, will make it possible to interpret and explain by using notions instead of examples.

References

- [1] R. Agrawal, T. Imielinski, and A. N. Swami, *Mining association rules between sets of items in large databases*, Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data (Washington) (P. Buneman and S. Jajodia, eds.), 26–28 1993, pp. 207–216.
- [2] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A.I. Verkamo, *Fast discovery of association rules*, Advances in Knowledge Discovery and Data Mining (1996), 307–328.
- [3] S. Brin, R. Motwani, J. D. Ullman, and S. Tsur, *Dynamic itemset counting and implication rules for market basket data*, ACM SIGMOD International Conference on Management of Data, 1997, pp. 255–264.
- [4] Ming-Syan Chen, Jiawei Han, and Philip S. Yu, *Data mining: an overview from a database perspective*, IEEE Trans. On Knowledge And Data Engineering **8** (1996), 866–883.
- [5] P. Cichosz, *Systemy uczone sie*, WNT, Warszawa, 2000.
- [6] P. Clark and T. Niblett, *The CN2 induction algorithm*, Machine Learning **3** (1989), 261–283.
- [7] G. Das, H. Mannila, and P. Ronkainen, *Similarity of attributes by external probes*, Knowledge Discovery and Data Mining, 1998, pp. 23–29.
- [8] R. O. Duda and P. E. Hart, *Pattern classification and scene analysis*, Wiley, New York, 1973.
- [9] D. Fisher, *Iterative optimization and simplification of hierarchical clusterings*, Journal of Artificial Intelligence Research **4** (1996), 147–180.
- [10] A. Freitas, *A survey of evolutionary algorithms for data mining and knowledge discovery*, 2001.
- [11] L. O. Hall, I. B. Özyurt, and J. C. Bezdek, *Clustering with a genetically optimized approach*, IEEE Trans. on Evolutionary Computation **3** (1999), no. 2, 103–112.
- [12] E. Han, G. Karypis, V. Kumar, and B. Mobasher, *Clustering based on association rule hypergraphs*, Research Issues on Data Mining and Knowledge Discovery, 1997.
- [13] A. K. Jain and R.C. Dubes, *Algorithms for clustering data*, Prentice Hall, New Jersey, 1988.
- [14] R. Krishnapuram and J. Keller, *A possibilistic approach to clustering*, IEEE Transactions on Fuzzy Systems **1** (1993), no. 2, 98–110.
- [15] R. T. Ng and J. Han, *Efficient and effective clustering methods for spatial data mining*, 20th International Conference on Very Large Data Bases (Los Altos, USA) (J. Bocca, M. Jarke, and C. Zaniolo, eds.), Morgan Kaufmann Publishers, 1994, pp. 144–155.
- [16] R. L. Rivest, *Learning decision lists*, Machine Learning **2** (1987), no. 3, 229–246.
- [17] E. H. Ruspini, *A new approach to clustering*, Information and Control **15** (1969), no. 1, 22–32.