A coarse to fine granular tree based on density peaks

Xukun Li[1, a], Jie Yang[1, 2, b]

[1]Chongqing Key Laboratory of Computational Intelligence, Chongqing University of Posts and Telecommunications, Chongqing, 400065, China

[2]School of Physics and Mechanical and Electrical Engineering, Zunyi Normal College, Zunyi, 563002, China

[a]email:Wingkzy@qq.com, [b]email:yj530966074@foxmail.com

Keywords: Multi-granularity; Density peaks; Hierarchical clustering; coarse to fine

Abstract. The granular computing is a methodology aiming at simulating the process and structure of the human cognition in the world. A new promising clustering algorithm, the density peak clustering (DPC) was proposed recently, which whereas, just takes procedures at a single granularity and could be conditionally ineffective by the inaccurate judgment by the decision graph. In this paper, we expand the DPC to the multi-granularity space and construct a coarse to fine granular tree. The structure deeply simulates the cognizing framework of human from a view of global to local of conceptions, making an innovative enlightenment to hierarchical clustering and cognitive computing in fields like robotics. Experiments show that the method includes every possible conclusion of the DPC by variable peaks picked from the decision graph, thus avoiding the limitations by uncertain artificial selections, at the same time, providing a competitive granular tree framework that could be analogically transplanted to other hierarchical algorithms.

Introduction

Clustering is a general methodology and a rich conceptual and algorithmic framework for data analysis and interpretation that gathers objects into groups [1]. From the perspective of recognition, clustering has a natural connection with granular computing (GrC). The information unit could be seen as the origin that can't be decomposed to more detailed atoms. Data tuple represented by informative dimensions are usually seen as the prime source from artificial dataset or the real world. Human process the original units and forms the coarser knowledge based on their transcendent experience. And the process is always represented by clustering methods. However, some data in the real world is naturally hierarchically structured. A drawback in most clustering is that the result is flat, leading to a one-side understanding of the knowledge behind the data. The limitation can be overcome by multi-granularity clustering, which provides a hierarchy clustering result, which is more informative than flat clustering [2].

GrC has emerged as one of the fastest growing information-processing paradigms in the domain of computational intelligence and human-centric systems [3]. GrC is often regarded as an umbrella term to cover theories, methodologies, techniques, and tools that make use of granules in complex problem solving [4]. Recently, GrC has been listed as one of the fundamental techniques and technologies to analyze big data [5]. Generally, there are two aspects to study the process of GrC. One is that we simulate human cognizing process that firstly the coarser impression is constructed of the problem figure, then after deeper insight, we expand the figure to a more detailed sub-structures iteratively, forming the coarse to fine architecture until the essence of the problem reached. Oppositely, the other is that we run the fine to coarse process which means knowledge induction in mankind's cognition.

Many works of granular clustering have been studied recently. Researches mainly focus on two aspects: one displays a brand-new method and the other combines the hierarchical methodology with plat clustering algorithms. However, both sides exhibit similar limitations: some are not robust to the shape of the clusters, while others are not efficient enough at forming the resulting hierarchy [6].

Recently, Rodriguez and Laio proposed DPC in the Science Journal, which can efficiently and accurately cluster datasets of any shape [7]. The DPC, based on density clustering, is a promising and competitive clustering method because it has concise running process which conforms to human recognizing logics and is of little methodological complexity to be understood. However, DPC is a flat clustering method that returns one partition on the dataset per run. In this paper, we improve the DPC to form coarse to fine granular levels, which finally construct an effective hierarchical tree following the decision graph. Due to the separation of data points from different clusters, it offers a more efficient and precise count of the density of the points, leading to optimizing distribution of the dataset on the graph. Besides, it overcomes the failure of the selection of density peaks due to artificial mistakes by including every possible clustering conclusion of the DPC per run on peak selection drift. At the end of the paper, we'll display the experiment to illustrate the hierarchical method as well as the granular conceptions, which, you can see, conforms to human intuitive cognition and covers a clear and complete granular levels.

Density peak clustering

It is necessary to provide a core introduction to the DPC which our work is primarily based on. There are two main parameters that construct the decision graph, from which, the density peaks are selected. Suppose the dataset to be clustered is $S = \{x_i\}_{i=1}^{n}$ and $d_{ij} = \mathrm{dist}(x_i, x_j)$ is the Euclidean distance between the sample points x_i and x_j. One main parameter ρ_i is defined as following:

$$\rho_i = \sum_{j \in I_s \setminus \{i\}} \chi(d_{ij} - d_c) \tag{1}$$

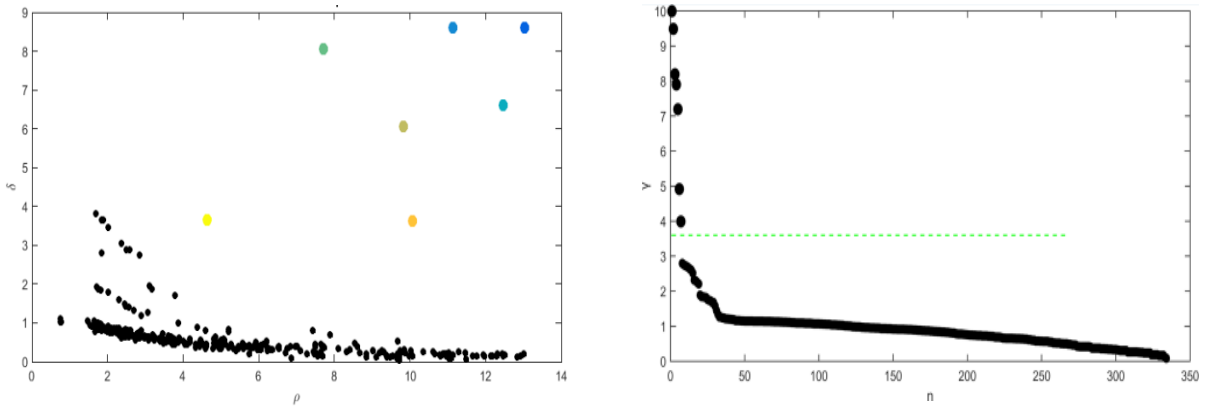
$$\chi(x) = \begin{cases} 1, x < 0 \\ 0, x \geq 0 \end{cases} \tag{2}$$

Where d_c is the cutoff distance, ρ_i the density of each point x_i, which is affected by its connection with other points in the set S. Another primary parameter δ_i is defined as:

$$\delta_i = \min_{j:\rho_j > \rho_i}(d_{ij}) \tag{3}$$

It is measured as the relative distance of x_i to the nearest point x_j whose density is larger than i.

According to (ρ_i, δ_i) of each point in the set S, we draw the decision graph in Figure 1 (a).



(a) The decision graph (b) γ distribution

Fig.1. An example of dataset S

The colored dots in in Figure 1 (a) have both relatively larger ρ_i and δ_i. To have a clear and quantitative observation of the decision graph, we draw an extension Figure 1 (b) of the dataset where γ is defined as:

$$\gamma_i = \rho_i * \delta_i \tag{4}$$

The points are realigned by their γ value decreasingly. The seven most foregoing points in the figure represents the corresponding colored dots in Figure 1 (a). By observations of naked eyes or

some numerical method, we could find a cut of slope after the first seven dots in the γ distribution map. Therefore, they would be selected as the cluster centers and then run the clustering process afterwards.

The granular tree framework

There are two main limits of the DPC depicted above. The first is that it is a flat clustering which can't reflect the hierarchical clustering process of humanity. The other is, when we discuss the ρ_i in Formula (1), it is affected by the points even far away from the point i in the calculation. To avoid the intensive influence by the far-away points, usually we need a carefully constructed function to compute ρ_i, like using Gaussian kernel [8]. However, this kind of functions have common characters of a quasi-exponential decay, finally making it just a density locally.

The granular tree has a root node on behalf of the whole dataset. It is generally a concept standing for a field at the coarsest level. In other words, it is a whole cluster undifferentiated and misty in our cognition.

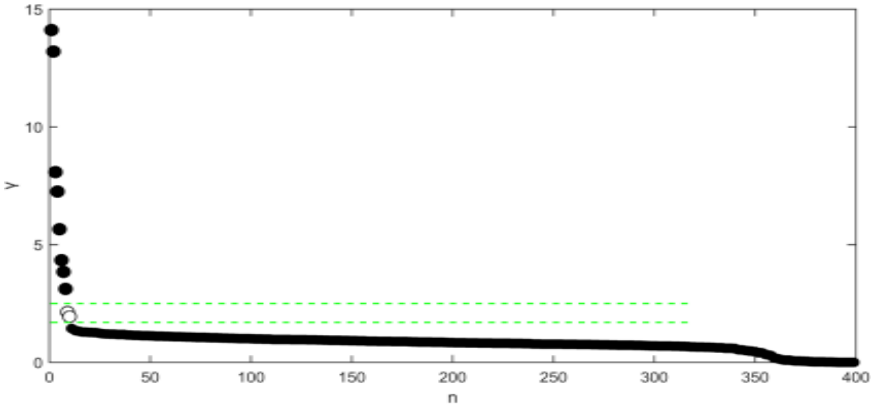
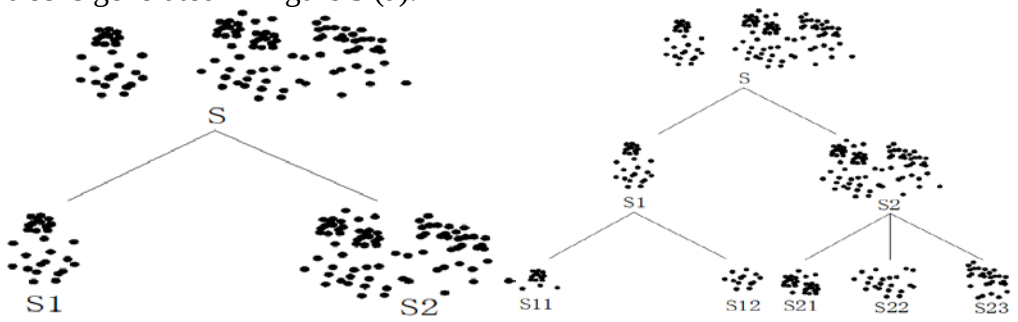


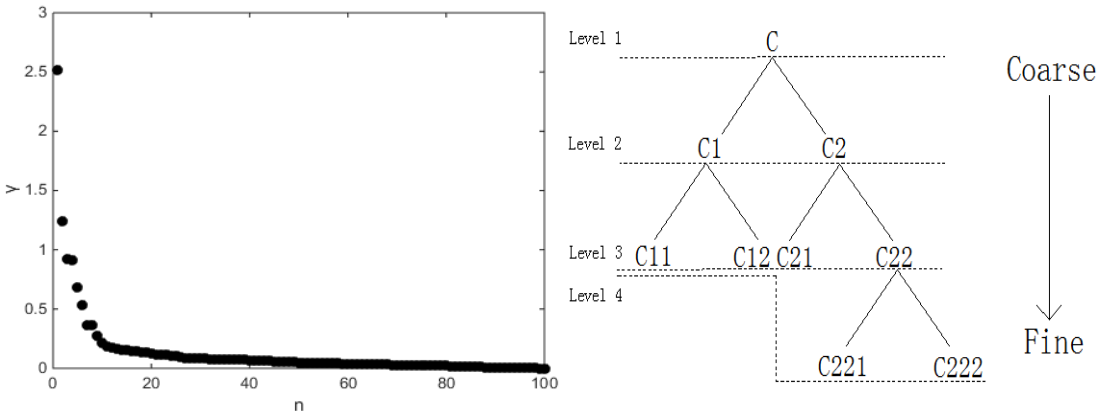
Fig.2. The split process based decision graph

To get finer conceptions, the root node must be split into child nodes based on the decision graph. This should be a carful process. Take Figure 2 as an example. It could be seen as a genetic γ distribution of a typical dataset S. By observation, we find 4 steep change of the slope in the Figure which obviously happens after the second, the fourth, the eighth and the tenth point. On the perspective of DPC, it is difficult for us decide how many density peaks should be picked as the cluster centers are exhibited in 4 cut and could not be confirmed even by sophisticated mathematic methods. Nonetheless, if we split the γ distribution at the first cut, namely, after the second point in this example, the set S would be split into two sub sets labeled S_1 and S_2 and the first two points should be the children of the root. The prudent division ensures every possible concept layer at a least cost splitting, reflecting our deduction step by step. Besides, the two parts together include all points in the original set S. If we take S_1 and S_2 as the new sets, and calculate their γ distribution, the other possible peaks after the first cut and before the last one would be presented again in the new maps and their natural bigger ρ_i and δ_i would make them the new cut points repetitively. The current tree is generated in Figure 3 (a).



(a) The first spit of root S (b) The extension of the tree

Fig.3. The split process based decision graph



(a) The typical γ feature of leaf nodes (b) The granularities of the tree

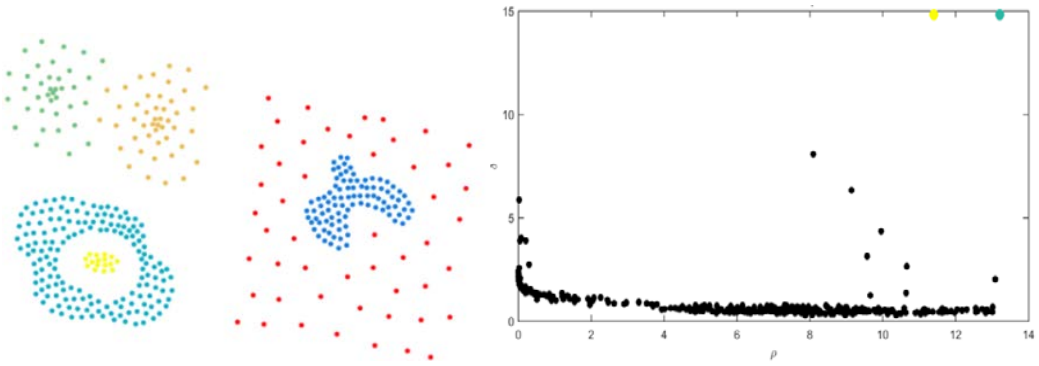
Fig.4. The granular tree

As S_1 and S_2 are independent sets, thus when the density of data is recounted, the procedure excludes the points of the other, allowing us to introduce more precise functions to fill in the defects of exponential analogical ways. Similarly, we extend the granular tree referring to the identical procedures iteratively using S_1 and S_2 which would be the fresh parent nodes as shown in Figure 3 (b), until the γ distribution of the split new child node has no obvious cut, or a continuous distribution like that shown in Figure 4 (a).

Despite the value next to the first point in the figure, the γ distribution has a smooth and continuous feature. No distinct cut of slope could be found, which means the end of the extension of the granular tree.

The granular levels of the tree are determined by the depth of the tree. If the tree is not balanced, the nodes on the deeper branch should construct the granularity with leaf nodes on other branches. Such as that in Figure 4 (b), C_{221} and C_{222} at the finest level don't have peers on the branches of C_1. Thus leaf nodes C_{11} and C_{12} on level third are demoted to the fourth level, taking charge of the pairing with them in order to include all data points at the granularity. In principle, as we deal with the division of the parent nodes conservatively, the hierarchical tree composes all possible conclusions of the DPC, overcoming the limitations of the decision graph depicted above and density calculation mentioned above.

Experiment results

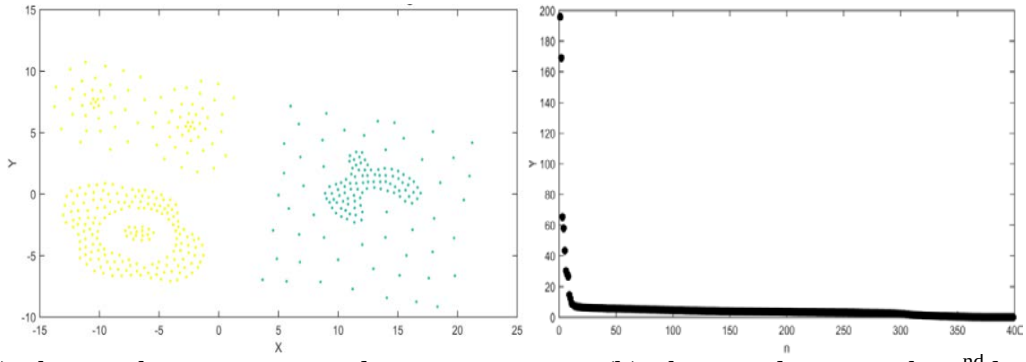


(a) The mapping of data points (b) The decision graph

Fig.5. The Compound dataset

This section gives the illustrative experimental results for the granular tree. The environmental configuration is Intel i5-2430M, 8G memory, Windows7 64bit desktop for the operating system and MATLAB2014 for the development. The Zahn's Compound dataset [9] is utilized representatively to expand the tree. The data points in the two dimensional European space and their corresponding decision graph are displayed in Figure 5 (a) and (b).

The six colors represent six original clusters in the dataset. Draw its γ distribution based on the decision graph, which is shown in Figure 6 (a).

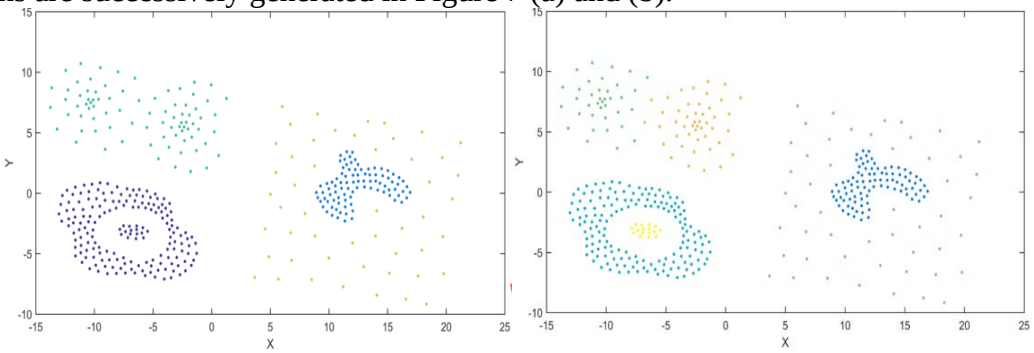


(a) The γ at the coarsest granularity (b) The two clusters at the 2nd level

Fig.6. The first extension of the granular tree

From the figure, the first cut happens just after the second point. Around the two child nodes two clusters are constructed at the left and right respectively in Figure 6 (b). The result is a macroscopic perspective at the second coarsest level. It generally stands for the rather fuzzy concept.

Regard the two nodes as parent nodes. We continue expanding the tree iteratively, the conclusions are successively generated in Figure 7 (a) and (b).



(a) The third level (b) The fourth level

Fig.7. The extension of the tree of Compound

At the third granularity level, the two nodes standing for the respective clusters on the right reaches the splitting end, because their decision graph lacks any cut. However, the left nodes continue the expansion and develop the tree to the fourth level, where the tree stops growing based on our rules. All the leaf nodes construct the perfect and fine clusters which are identical to the original classification the author suggested.

As shown above, the granular tree with a depth of 4 levels contributes to observe the set at different perspectives or concepts. Hence it could be also called a kind of genetic conceptual tree [10].

Conclusion

Based on the DPC, the brand-new and promising clustering, the main contributions of this paper are: (1) we proposed a coarse to fine granular tree which expands the original flat clustering method to multi-granularity levels; (2) we put forward a granularity decomposition mechanism that simulates the process of human cognition in problem solving space from the decision graph. The hierarchical concepts reflect the natural exploration from macro fuzziness to micro details. The idea this paper presents could be expanded to other hierarchical clustering methods or granular computing strategies, thus heuristic at the fields of artificial intelligence [11], cloud models [12] [13] etc. The experiment as an illustration of the structure shows that the tree could commendably discover the granular framework inside the dataset and the result conforms to human intuition of classification. The further work is to construct the inverse tree, i.e. the fine to coarse granular tree in order to form a complete two-way cognitive model based on the DPC.

Acknowledgement

In this paper, the research was sponsored by the ChongQing Postgraduate Scientific Research and Innovation projects under grant CYB16106.

References

[1] W. Pedrycz, Knowledge-Based Clustering: From Data to information Granules, John Wiley & Sons, 2005.

[2] C.D. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval, Cambridge University Press, 2009.

[3] Wang G, Yang J, Xu J. Granular computing: from granularity optimization to multi-granularity joint problem solving[J]. Granular Computing, 2016:1-16.

[4] Y.Y. Yao, Perspectives of granular computing, in: Proceedings of IEEE International Conference Granular Computing Beijing, China, 1, 2005, pp. 85–99.

[5] C.P. Chen, C.Y. Zhang, Data-intensive applications, challenges, techniques and technologies: a survey on big data, Inf. Sci. 275 (2014) 314–347.

[6] Ji Xu, Guoyin Wang, DenPEHC: Density peak based efficient hierarchical clustering, Information Science 373(2016) 200-218.

[7] A. Rodriguez, A. Laio, Clustering by fast search and find of density peaks, Science 344 (6191) (2014) 1492–1496.

[8] Cheng Y. Mean Shift, Mode Seeking, and Clustering[J]. Pattern Analysis & Machine Intelligence IEEE Transactions on, 1995, 17(8):790-799.

[9] Zahn, C.T , Graph-theoretical methods for detecting and describing gestalt clusters. IEEE Transactions on Computers, 1971. 100(1): p. 68-86.

[10] Jiang R, Li D., Fan J. Automatic generation of pan-concept-tree on numerical data[J]. Chinese Journal of Computers, 2000.

[11] Russell S J, Norvig P. Artificial intelligence: a modern approach[J]. Applied Mechanics & Materials, 2009, 263(5):2829-2833.

[12] Mell P M, Grance T. SP 800-145. The NIST Definition of Cloud Computing[M]. National Institute of Standards & Technology, 2011.

[13] Yang J, Wang G, Li X. Multi-granularity Similarity Measure of Cloud Concept[M]// Rough Sets. Springer International Publishing, 2016.