Engineering, 2013, 5, 549-552
http://dx.doi.org/10.4236/eng.2013.510B113 Published Online October 2013 (http://www.scirp.org/journal/eng)

On Clustering Algorithms for Biological Data*

Xiaowan Li, Fei Zhu#
School of Computer Science and Technology, Soochow University, Suzhou, China
Email: 1027401004@suda.edu.cn, #zhufei@suda.edu.cn

Received 2013

ABSTRACT

Age of knowledge explosion requires us not only to have the ability to get useful information which represented by data but also to find knowledge in information. Human Genome Project achieved large amount of such biological data, and people found clustering is a promising approach to analyze those biological data for knowledge hidden. The researches on biological data go to in-depth gradually and so are the clustering algorithms. This article mainly introduces current broad-used clustering algorithms, including the main idea, improvements, key technology, advantage and disadvantage, and the applications in biological field as well as the problems they solve. What's more, this article roughly introduces some database used in biological field.

Keywords: Clustering; Algorithms; Biologiocal Data; Applications; Database

1. Introduction

We humans are now in an era of information which are stored or represented as data. People have found that it is an effective way to classify or group things into a set of categories or clusters in order to analyze them. Because when encountering new things, we always try to seek features that can describe them by comparing them with objects we already knew and things in the same cluster because things in one clustering always share similar features and have similar functions. And data reflect those features and functions. So we can cluster data to analyze further knowledge. What's more, things are not single or isolated but have verities of links. So when the number of objects gets bigger the relationship will become more complex. It will be easier to analyze high connected things other than a single one.

So clustering is a promising way to analyze biological data. From one hand, biological data contains large amount of knowledge which may be unknown but useful. However the quantity and complexity of biological data make it hard to analyze those data in a certain study for that knowledge. So we need to cluster them first to simplify the process. From other hand, biological function is not determined by a single gene or protein. There are complex relationships to consider, we should analyze data in high connected subgraphs.

Different clustering algorithms are designed on the basis of different applications to solve different problems. Some of them proved to be good in biological field by achieving good results or accelerating the process. In the meantime, with the research on biological field goes in-depth, clustering algorithms are improved to favor the new need [1].

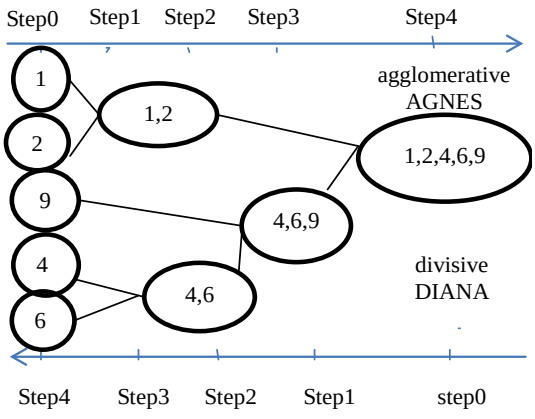
2. Conventional Clustering Algorithms

2.1. Hierarchical Clustering Algorithm

There are two of Hierarchical Clustering Algorithms, AGNES and DIANA [2].

AGNES: Make every object a single cluster and then merge the nearest two, refresh the distance of the original and new one, then repeat it until all of objects are one cluster or reach the result wanted.

DIANA: Make all objects one then divide it, until every objects is a single cluster or reach the result wanted.



*This work is supported by fund NO.5731512811. This work is supervised by Fei Zhu.
#Corresponding author.

The hierarchical clustering algorithm is used frequently because it is simple and it can handle large data. But it has high compute and complex calculates and the data's class cannot be adjusted once settled.

Liming Wang and Xiaodong Wang propose a non-parametric Bayesian clustering algorithm based on the hierarchical Dirichlet processes (HDP) which capture the hierarchical features prevalent in biological data. They conduct experiments on the yeast galactose datasets and yeast cell cycle datasets by comparing their clustering results to the standard results. The proposed clustering algorithm is shown to outperform several popular clustering algorithms by revealing the underlying hierarchical structure of the data [3].

2.2. Fuzzy Clustering Algorithms

Classification problem in practice is often fuzzy and many papers prove that fuzzy clustering algorithm is an effective tool for the analysis of biological data [4,5].

FCM: 1974, Dunn proposed this algorithm, and Bezdek promoted it. The standard FCM as follows, in which the Euclidean or L_2 norm distance function is used [6].

1) Select appropriate values for m, c, and a small positive

Number c. Initialize the prototype matrix M randomly. Set step variable $t = 0$.

2) Calculate (at $t = 0$) or update (at $t > 0$) the membership Matrix U by

$$u_{ij}^{(t+1)} = 1 \Big/ \sum_{t=1}^{c} \left(\frac{D_{lj}}{D_{ij}} \right)^{1/(1-m)}$$

for $i = 1, \cdots c$ and $j = 1, \cdots N$.

3) Update the prototype matrix M by

$$m_i^{(t+1)} = \left(\sum_{j=1}^{N} \left(u_{ij}^{(t+1)} \right)^m x_j \right) \Big/ \left(\sum_{j=1}^{N} \left(u_{ij}^{(t+1)} \right)^m \right)$$

for $i = 1,\ldots c$.

4) Repeat steps 2)-3) until $\left\| M^{(t+1)} - M^{(t)} \right\| < \varepsilon$.

FCM algorithm and other visions of it are robust to the scaling transformation of dataset, while others are sensitive to such transformation [7]. So they are used widely. Compared with K-means, it is more efficient for the needless of iteration. But it has a problem of being easily influenced by isolated points, and it has difficulties of determining clustering number.

2.3. Graph Clustering Algorithm

In the post-genome era, it is a current challenge to mine the hidden knowledge stored in the biological networks [8]. As an important means for knowledge discovery, graph clustering does better in the analysis of complex biological networks [9].

The advantage of graph clustering is that it is relatively straight forward to see the highly connected subgraphs because of the network community structure.

However there are some problems remain to be solved. For example, some of these complex networks have so many nodes that it is difficult to compute effectively. What's more, many typical networks have high in homogeneity, which will make graph clustering algorithm lose its advantages.

Jain created the main idea of graph clustering algorithm that made a minimal spanning tree (MST) about data, and then delete the longest branch of the minimum tree to cluster. The main algorithms include Random, Walk, CHAMELEON and AUTOCLUST.

3. Biological Applications

- Two clustering algorithms are used by Jing Yang to analyze the data searched from the Nucleotide database of National Center for Biotechnology Information (NCBI) and they have similar results, which could give big support on study of PD (Parkinson's disease) [10].
- Based on the aiNer model an important model of Artificial Immune System Jun Wang and Xinyu Liu presented aiNHA. This algorithm can be used in clustering the arbitrary shapes of data sets with fast discovering speed high efficiency [11], and insensitive about noise.
- Gene Ontology (GO), a novel and effective method named significant clustering analysis based on GO (ScaGO) was presented to improve individual GO term analysis algorithm for detecting differential gene expression. Compared to individual GO term analysis, ScaGO was turned out to be more sensitive when applied to the acute lymphoblastic leukemia expression dataset and yeast Rapl DNA-binding mutant dataset, and some novel differential expression changes which were mostly reported were mined successfully [12].
- Juan Men came up with a high-performed graph clustering algorithm: CD (contraction-dilation), which can be applied to analyze large networks. This algorithm focusing on complex biological networks is proved to be efficient to discover more stable community structures with higher modularity scores and accuracies at lower expenses of both CPU time and memory. CD is superior to spectral clustering algorithm and MCL algorithm because it can detect protein remote homology successfully at the meantime. The results show that sequence similarity carry significant information on remote homology, which can be mined by using CD algorithm.
- Yuan Wei and Zhu Shanfeng clarify problems of current clustering method by analyzing their reliability

and parameters, and put forward a solution: ensemble clustering. And they use Mesh ontology as knowledge to improve the clustering, which contains a wealth of knowledge of biology. This algorithm based on the distance between the MeSH is proved to be better in clustering results compared with other methods [13].

- Yuan Yinli proposed a revised fuzzy algorithm to solve the problem that it is difficult to select hidden node centers in the study of RBF network. And applied it to strengthen the robustness of the outliers by the network with the effectiveness proved [14].
- Cuifang GAO proposed a new algorithms: CKFCM (collaborative kernel fuzzy c-means clustering), in which the function of collaborative relationship was incorporated into kemel fuzzy c-means clustering (KFCM). By enlarging the difference among the samples and implementing on several subsets can be processed together with an objective function, CKFCM achieves better classification and is effective clustering with better performance [15].
- An improved algorithm of weighted fuzzy kemel clustering (WFKCA) is proposed to overcome its shortcoming of liability to stick to local optimum. To reduce the possibility of local optimum the idea of iterative self-organizing data analysis techniques algorithm (ISOODATA) is introduced into the WFKCA, and initial center vectors are adjusted by the intermediate results from splitting and/or merging of clustering centers. It achieves more stable performance of clustering for using match-able measurement from feature space, and increases the adjustment range of clustering centers [16].
- Damodar Reddy Edla and Prasanta K. Janawe propose a new clustering algorithm which is based on Voronoi diagram. The algorithm uses a real valued function defined by the radii of Voronoi circles. This function enables to deal with the inner points of the clusters followed by the boundary points. The proposed scheme is applied on various artificial and biological data. The experimental results of the proposed method are also compared with K-means and a few existing clustering techniques [17].
- Take noise into account, there are several means to deal with it. For example, Roman Sloutsky, Nicolas Jimenez, S. Joshua Swamidass and Kristen M. Naegle explore several methods of accounting for noise when analyzing biological data sets through clustering [18].

4. Introductions to Bioinformatics Databases

- GenBank: A complete database of DNA sequences contains almost all of the Protein sequences and DNA sequences that have been found as well as the relative paper. Each data record has a simple description, such as scientific name, references, table of the feature and the sequence itself.
- GDB: preserve and deal with the gene data for Human Genome Project, contains the human genome region, the human genome map and the genetic variation. It provides read or write access directly.
- PIR and PSD: An overall, annotated, no redundant database for protein sequence, including some protein sequences come from dozens of integrated genes. Thus, almost 99% of the data have been classified in a certain protein family. And cross-reference can be achieved in the annotation.
- COG: Attempt on a phylogenetic classification of the proteins encoded in 21 complete genomes of bacteria, archaea and eukaryotes, constructed by applying the criterion of consistency of genome-specific best hits to the results of an exhaustive comparison of all protein sequences from these genomes. The database comprises 2091 COGs that include 56% - 83% of the gene products from each of the complete bacterial and archaeal genomes and ~35% of those from the yeast *Saccharomyces cerevisiae* genome [19].

5. Conclusion

From the introduction we can know that every clustering algorithm has advantages, disadvantages and scope of application. So it is necessary to analyze each clustering algorithm to use them better. Thus the truth is that it is useful to apply clustering algorithms on biological data. What's more, the deeper the research goes, the higher the demand becomes. So we should also analyze each case to know the exact requirement for the algorithms and then improve the current algorithm to get a better result.

6. Acknowledgements

Xiaowan LI, ID Number: 1027401004, currently is an undergraduate student of Computer Science and Technology School of Soochow University, majoring in computer science and technology.

REFERENCES

[1] J.-G. Sun, J. Liu and L.-Y. ZHao, "Clustering Algorithms Research."

[2] M. X. Duan, 2009-5-1.

[3] L. M. Wang and X. D. Wang, "A Non-Parametric Bayesian Clustering for Gene Expression Data," *IEEE Workshop on Statistical Signal Processing* (*SSP*), Ann Arbor, 5-8 August 2012, pp. 556-559.

[4] M. Zhang and J. Yu, "Fuzzy Partitional Clustering Algorithms."

[5] L. Wang, H. Peng, J.-S Hu and H.-F. Liang, "Fuzzy Clustering Applied in Genetic Differentiation Analysis,"

Control & *Automation*, Vol. 22, No. 3, 2006, pp. 172-174.

[6] R. Xu, Survey of Clustering Algor*ithms," IEEE Transactions on Neural Networks*, Vol. 16, No. 3, 2005, pp. 645-678.

[7] M. X. Duan and L. M. Yang "The Improvement of Hierarchical Clustering Algorithm."

[8] Y. Tian, D. Y. Liu and B. Yang, "Application of Complex Networks Clustering Algorithm in Biological Networks," *Journal of Frontiers of Computer Science and Technology*, Vol. 4, No. 4, 2010, pp. 330-337.

[9] E. Hartuv and R. Shamir, "A Clustering Algorithm Based on Graph Connectivity," *Information Processing Letters*, Vol. 76, No. 4-6, 2000, pp. 175-181.

[10] J. Yang, "A Study on the Clustering Analysis for Parkinson-Relates Genes," Master's Thesis, Tianjin Medical University, Tianjin, 2007.

[11] J. Wang and X. Y. Liu, "Hierarchical Clustering Algorithm Based on the aiNet Model of Artificial Immune System," *Computer Engineering and Applications*, Vol. 42, No. 24, 2006, pp. 167-169.

[12] Q. J. Tang, T. Xu, D. Wang, L. J. Li and L. F. Du, "Clustering GO Term Applied to Differential Gene Expression Detection," *Chinese Journal of Applied and Environmental Biology*, No. 3, 2011, pp. 422-426.

[13] W. Yuan and S. F. Zhu, "Study on Biological Text Clustering Algorithm Based on Metric Learning," *Computer Applications and Software*.

[14] Y. L. Yuan, "Improved Fuzzy C-means Clustering Algorithm."

[15] C. F. Gao, "Novel Fuzzy Clustering Algorithms and Applications," Ph.D. Thesis, Jiangnan University, Wuxi, 2011.

[16] X. Wang, X. B. Yang and L.-L. Zhou, "An Algorithm of Hierarchical Clustering Based on Correcting Class Center," *Microelectronics & Computer*, Vol. 28, No. 10, 2011.

[17] D. R. Edla and P. K. Jana, "A Novel Clustering Algorithm using Voronoi Diagram," *Seventh International Conference on Digital Information Management* (*ICDIM*), Macau, 22-24 August 2012, pp. 35-40.

[18] R. Sloutsky, N. Jimenez, S. Joshua Swamidass and K. M. Naegle, "Accounting for Noise When Clustering Biological Data," *Briefings in Bioinformatics*, Vol. 14, No. 4, 2013, pp. 423-436.

[19] R. L. Tatusov, M. Y. Galperin, D. A. Natale, L. V. Grakavtsev, T. A. Tatusova, U. T. Shankavaram, B. S. Rao, B. Kiryutin, M. Y. Galperin, N. D. Fedorova and E. V. Koonin. "The COG Database: New Developments in Phylogenetic Classification of Proteins from Complete Genomes," *Nucleic Acids Research*, Vol. 29, No. 1, 2001, pp. 22-28.

