A clustering algorithm using DNA computing based on three-dimensional DNA structure and grid tree

Jie Xue,Xiyu Liu
School of Management Science and Engineering
Shandong Normal University
Shandong,Jinan
China
xiaozhuzhu1113@163.com,sdxyliu@163.com

Abstract: - Clustering is an important technique for data analysis, conventional methods include hierarchical clustering, Density-based clustering, Subspace clustering, etc. In this paper, we utilize DNA computing using three-dimensional DNA structure(also called k-armed DNA structures) and grid tree to execute the clustering algorithm. In our study, we will design grid tree ,the process of clustering will become a parallel bio-chemical reaction and three-dimensional DNA structure are also adopted . Two examples are showed to offer a detailed insight into the performance of our method. The new method provides a new idea for traditional clustering.

Key-Words: - clustering, DNA computing, three dimensional; k-armed DNA structure; Grid tree;

1 Introduction

1994 ,Adleman[1] computed the seven vertices of Hamiltonian path problem with DNA molecules in test tube, which shows a great power in combinational problems by DNA computing. DNA computing has three advantages:(1)huge parallelism(2) high speed (3)largestorage,1 bit information can be stored in 1 nm^3[10].

However ,traditional DNA structure and their models have some restricts, such as single or double strands can only link data they represent into chains ,they can not denote graphs which are beyond two dimension, besides ,sometimes they may add time for coding and reactions.

Three dimensional DNA structure can come over the drawbacks that conventional DNA structure has in some degrees. Three dimensional DNA structure is exist in nature, just as the holliday intermediates, which is being studied widely. This structure can be divided into 3 forms, 2-armed DNA structure ,3-armed DNA structure and 4-armed DNA structure. These structures described different three-dimensional DNA graph[20] and have already proved the 3-armed DNA structure and the 4-armed DNA structure are stable. The 3' end of k-armed DNA structure is single a sequence with 30~45 base.

Clustering is an assignment of a set of data into subsets so that data in the same cluster are similar and data between clusters are different, which is used in many fields, including machine learning, data mining, pattern recognition, image analysis, etc. There are many types of clustering, such as hierarchical clustering, Density-based clustering, Subspace clustering, etc. However, when the number of clusters is unknown and the data set become huge ,these algorithms exhibit polynomial or exponential complexity, which make the problem being more challenging.[12]

DNA computing has been used in many fields, but there has not many researches in clustering. Bakar and Watada presented some ideas to use DNA computing to solve clustering problems [5][6][7][8][9]. They proposed a new DNA approach to solve clustering problem based on k-means and Fuzzy C-means algorithm [15].Kim and Watada (2009) gave the similar method for heterogeneous coordinate data. Zhang Hongyan ,Liu xiyu[24] presented another research on clustering based on the idea of using DNA computing to find Hamilton circuit. There have not been any researches on clustering by three dimensional DNA structure.

In this paper, we provide a new method in clustering using three dimensional DNA structures (k-armed DNA structures).We propose the basic idea of using DNA computing to achieve the clustering algorithm, meanwhile we present three dimensional DNA as well as biochemical operations design which is a creative point. Different from the traditional techniques of data processing ,we provide grid tree to do the pre-work, which can store many characteristics of data for us. We also use two examples to illustrate our idea and compare time complexities and correctness of the new algorithm with other clustering solutions .

2 Three Dimensional DNA Structure and Computing

We know that traditional DNA sequences are classified into two forms. They are single sequence and double ones, single sequence can become double ones and the reverse manipulate is also feasible through the biology reaction: denaturation and annealing. We show the double DNA sequence in figure 1.

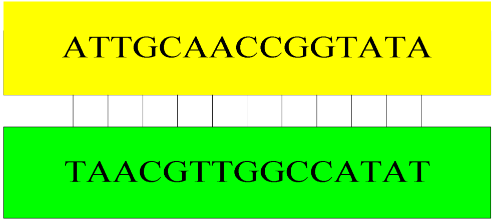


Fig.1 double DNA sequence

So far, Adleman-Lipton model is the most popular research in DNA field. It solved so many problems in computer science area, just as the Hamilton path problems. In this conventional model, we always encode the vertex of a graph as a single DNA sequence, which length is determined by the complex of the graph, each vertex can be seen as two sections ,the first section and the second section. Every edge is comprised by three parts, first one is the complementary sequence of the front vertex's second part DNA sequence, the second part is a single sequence which stands for the weight of edge, the third part is the complementary sequence of the latter vertex's first part DNA sequence. This can be seen in figure 2.

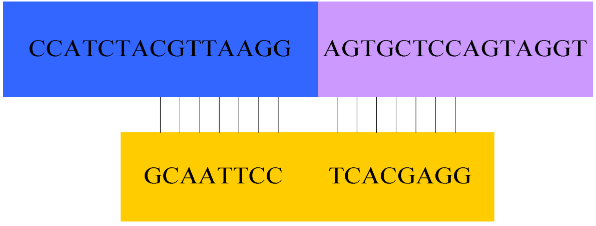


Fig.2 coding method of two vertexes and their edge

There are some other models of DNA computing ,like stick model, which is based on a coding scheme, using double and single sequence to stand for 0 and 1,then do the computation; splicing model, which is proposed by Tom Head
[25] proposed and based on formal language theory.

In our paper, we use three dimensional DNA structure, for distinguishing their classes, we also use the name k-armed DNA structure below, so our coding method is different from traditional ones, k-armed DNA structure obeys the Waston-Crick rules as well. It has three forms, they are two-armed DNA structure ,three-armed DNA structure and four-armed DNA structure as showed in figure 3

The first discrete three dimensional structures were reported in 1999 by Seeman. A topological cube was generated by ligation of two closed, interlocked DNA rings into a belt-like molecule, followed by a series of ligation, purification, and reconstitution steps [26]. These pioneering experiments demonstrated the feasibility of using DNA as a useful building block for three dimensional structures[27].As traditional DNA sequence, three dimensional structures also has those manipulate: gel electrophoresis, denaturation and annealing, but for polymerase chain reaction , three dimensional DNA structure has not grasp over.

Three dimensional DNA structure was used to solve combinational problems was first demonstrated by Natasa Jonoska in 1999[17], which can simplify the code process and reduce the time and steps .There are lots of problems being solved by this structure ,such as SAT problems, 3-vertex-colorability[20].etc. figure 3 shows the structure of our DNA sequences and figure 5 is the three dimension graph figure 4 creating by k-armed DNA..

Each arm is figure 6 double DNA sequence with sticky end that they can stick with their complementary DNA sequence . We can see from the figure that sticky end of left arm v_1 is the complementary sequence of that in v_2.In our coding strategy ,only the arms of fluorescent mark cores linking to a same units can be complementary .The up and below arms are as same as the arms mentioned above. This method ensures that only the neighboring marked units can be linked.

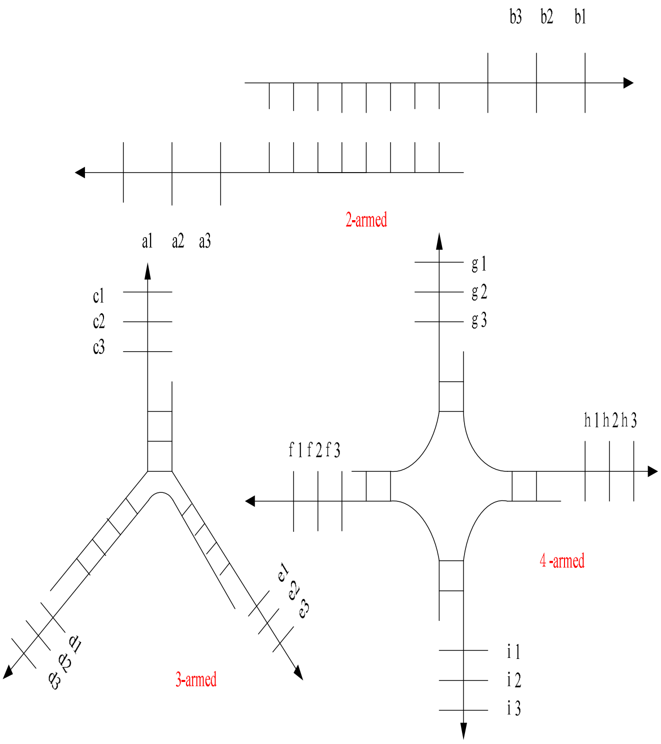


Fig.3 k-armed DNA structure

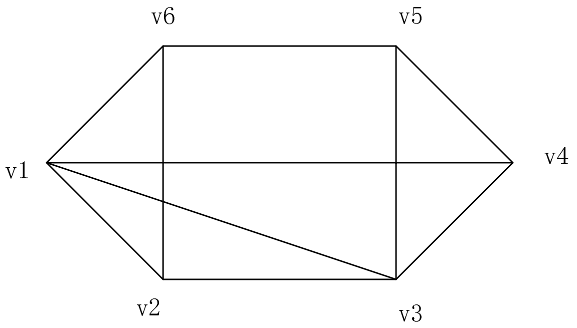


Fig.4 a graph of six vertexes

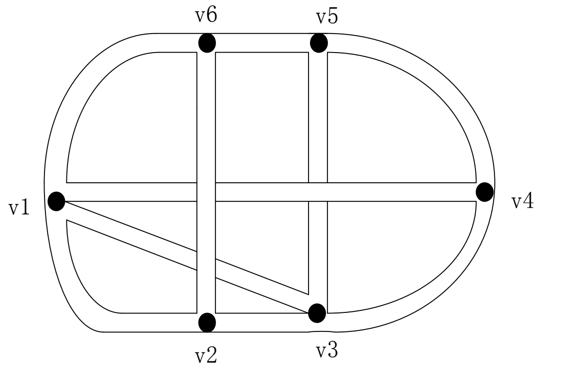


Fig.5 three dimensional graph of Fig.4

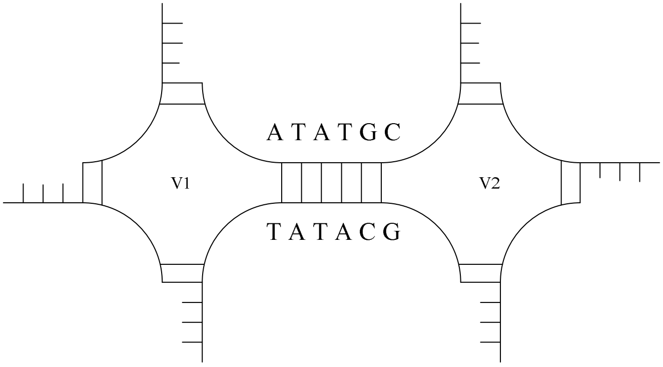


Fig.6 connection of two k-armed DNA structure

3 Grid-based algorithm

Grid-based algorithm uses a multi-resolution grid structure(also called cell) containing the data objects which is limited space to quantify the number of units and acts as operands of clustering performance. The advantage of this method is its fast processing speed. Time of this algorithm is independent of data object's number, only depends on the quantitative dimension of space in each unit number. Traditional approaches include STING, which uses

the information stored in grids; WaveCluster , which clusters data by wave change; CLIQUE, an algorithm who combines density clustering with Grid-based algorithm.

Common Grid-based algorithm also has disadvantages, in the process in the process of grid partition, those characteristics about data can be achieved, blank grid exist in every division step of algorithm, which add time complexity to ,moreover, much data is missing inevitably.

Next, we propose grid tree to do grids partition, which base on the common grids division, add data information in every node of tree, reduce time complexity by deleting blank grid nodes, keep every grid node who has data.

3.1 A grid tree definition

Suppose the data set is X={x_1,x_2,x_3,….,x_n} $\subseteq R^n$ It is bounded by a rectangle D_0 in R^n. A grid is a tree T, where each node of T called a cell and is represented by a triad n=(D,c,s,),where D is a rectangle with four smaller units, as showed in figure7., c is the center of node, s is the number of data in node n, here s$\neq$0.

If a cell n' is a unit of cell n, then, n' is defined as child cell of n. For two adjacent cells n_i=(D_i,c_i,s_i), i=1,2.There exist three situations, shows in figure 8:
1.n_1,n_2 come from a same parent cell n .
2.n_1,n_2 come from different parent cells but their parent cells have a same edge of grid, they are beside the edge both their parent cells have.
3.n_1,n_2 come from different parent cells but their parent cells have a same point of grid.

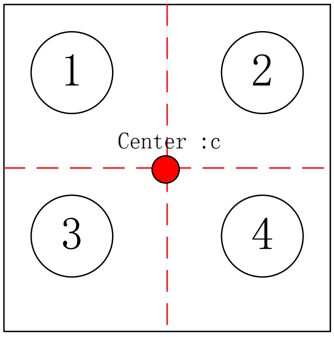


Fig7.a cell of grid tree

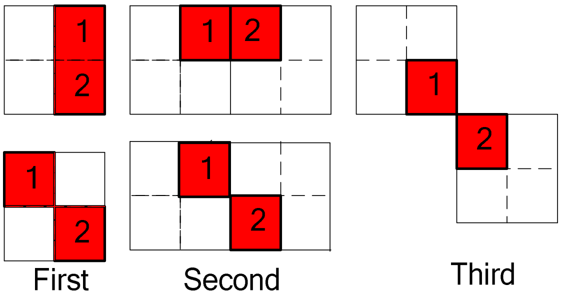
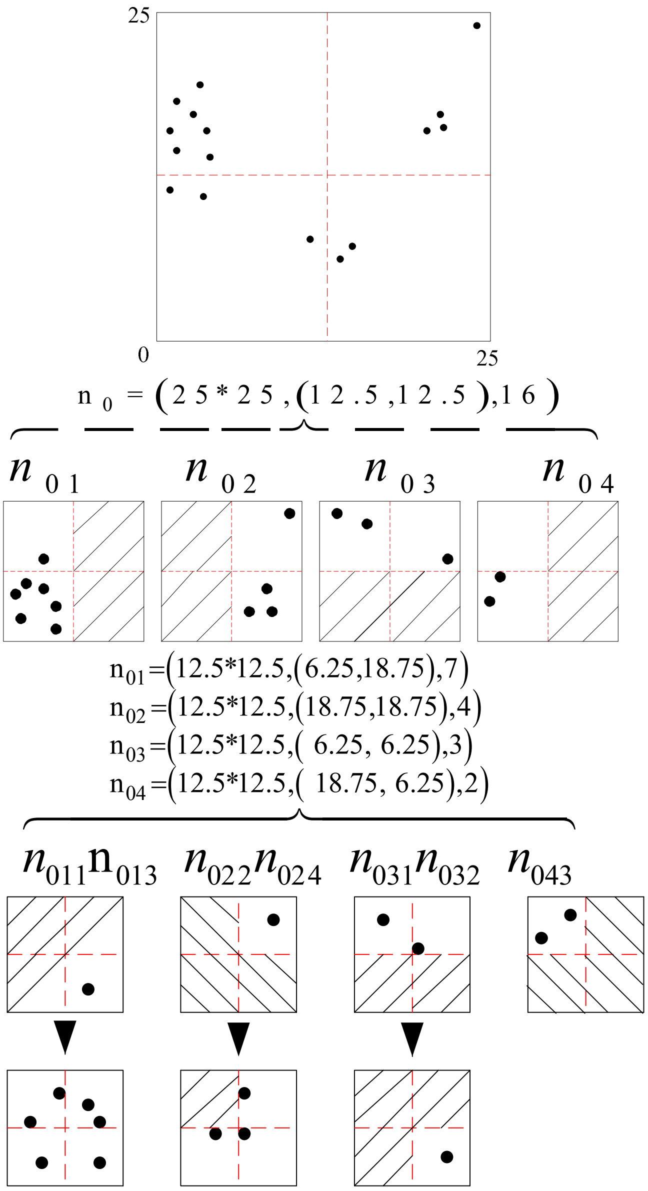


Fig8. Three situations of adjacent cells

To construct the grid tree, we need the initial length and height of the rectangle D_0. Then the tree is constructed iteratively. We start with the first node

$n_1 = (D_0, c_1, s_1)$ where D_0 is the original rectangle, c_0 is the center of D_0,s_1 is data initial data number of D_0 ,$s_1 \neq 0$,then D_0 is divided by four parts denoted as D_{01}, D_{02} ,D_{03}, D_{04} ,each of them is a child cell of D_0 denoted as n_{01} ,n_{02} ,n_{03}, n_{04} if their s=0,then they will be deleted. Next, each cell kept in this level will be divided as steps to D_0. Steps continue until data in cells is satisfied with conditions. The resulting tree is called a grid tree.

We present an example to show the data set and grid tree generated by the above algorithm as figure 9.


25
0
25
n 0 = (2 5 * 2 5 ,(1 2 .5 ,1 2 .5),1 6)
n 0 1
n 0 2
n 0 3
n 0 4
n01=(12.5*12.5,(6.25,18.75),7)
n02=(12.5*12.5,(18.75,18.75),4)
n03=(12.5*12.5,(6.25, 6.25),3)
n04=(12.5*12.5,(18.75, 6.25),2)
n011n013
n022n024
n031n032
n043

n_{011}=(6.25*6.25,(3.125,21.875),1)
n_{013}=(6.25*6.25,(3.125,15.625),6)
n_{022}=(6.25*6.25,(21.875,21.875),1)
n_{024}=(6.25*6.25,(21.875,15.625),3)
n_{031}=(6.25*6.25,(3.125,9.375),2)
n_{032}=(6.25*6.25,(9.375,9.375),1)
n_{043}=(6.25*6.25,(15.625, 3.125),2)

Fig.9 the process of generating grid tree

3.2 Transform for clustering

When the tree is constructed, the clustering problem is converted into grouping leaf cells of the tree into clusters.For the purpose of this paper, here, we will give a different transform .Firstly, we set up a fluorescent mark core for every leaf cell in the location of their c_i, we knows that leaf cells are divided into four units, define arms as a diagonal from fluorescent mark core to another horn of units where have data in .By these steps, the tree is changed into a special graph as the example showed in figur10.

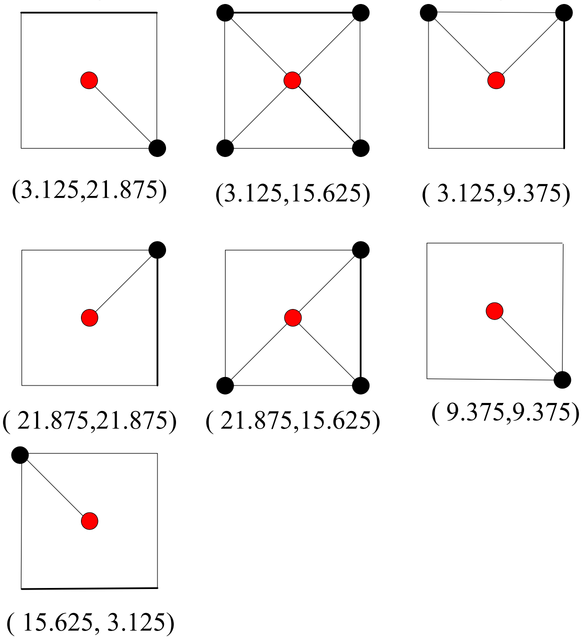


Fig.10 transform of grid tree

Now structures present in figure10 are stands for the initial data . The clustering problem is transform Ed into finding those structures who are adjacent.

In this example ,we can see that n_{011} is adjacent with n_{013},n_{013}, is adjacent with n_{031} ,and n_{032} is the neighbor of n_{043},.n_{022} and n_{024} are alone

4 Clustering by three dimensional DNA structure

4.1Strategy

Here we use grid tree to deal with our data set, those data who will be clustered are divided by grids tree firstly, the process of generating grid tree and deal with data are showed in figure9.

In this strategy, all the cells are considered as structure showed in figure10 and the adjacent location as figure8.

The clustering strategy is to discover the neighboring leaf cells and link them to become cliques.

We use DNA computing to aggregate DNA structures indicating the neighboring cells. We will g

-et all the possible combinations of the neighboring leaf cells and the clustering result is appeared by new DNA cliques.

Each leaf cell can be coded into k-armed DNA structure with a fluorescent mark core and their arms , leaf cells linked by their arms which linked to the same point in location. All of k-armed DNA structures are put into the test tube for ligation and hybridization. We achieve some cliques DNA structures after the process. These cliques are contained with the k-armed DNA structures whose density is large enough. Those units in the same clique can be seen as in a cluster .

In our graph ,x and y are the center positions of leaf cells, we use C(x,y) to stand for the leaf cells and Ci(x,y),i-1,2,3,4 to represent the ith arms from left to right, up to below .

Evidently, the problem is changed into the combinational problem and we all know DNA computing can solve the class of these problem. Therefore, our clustering method is feasible. It is just a graph connection problem [1].

4.2DNA coding

Encoding the leaf node of grid tree (cells) is one of the most important step in our algorithm. Here we have a restriction to elaborate: Only the leaf node of grid tree can be encode as k-armed DNA structure and linked. This restriction ensure that al the vertexes which stands for the original data can be encoded as k-armed DNA structure and take part in the experiments.

Each leaf cell of grids which has points with them is encoded as a k-armed DNA structure ,whose arms have sticky ends .The length of their arms is based on the scale of the data set. Just like data in figure10, there are seven leaf cells . Arms of leaf cells who are adjacent and their arms linked to the same point can be encoded as the complementary sequence

For example, the DNA codes of figure10 show in table1. According to the data set ,we use 16 DNA base to encode the arms of leaf cells, basing on Hamming distances and similar temperature.

In table1,C4(21.875,21.875)is respectively the complementary sequence of C2(21.875,15.675), C3(21.875,15.675)and C4(21.875,15.675) are the complementary sequence of C1(3.125,9.375) and C2(3.125,9.375), C4(9.375,9.375) is respectively the complementary sequence of C1(15.625,3.125) . so they can do the hybridization reactions.

Leaf cells	Arms	Arms coding	
C(3.125,21.875)	C4	ATATCGCG TATAGCGC	CAACGTGC
C(3.125,15.625)	C1	GCAGTTGA CGTCAACT	AGAGCTGG
	C2	TATAGTAC ATATCATG	GTTGCACG
	C3	AACCTGGT TTGGACCA	ACCTAGCT
	C4	TGGTTTGG ACCAAACC	CCAGGTCT
C(21.875,21.875)	C2	TTTAGCGC AAATCGCG	CTCGTCGA
C(3.125,9.375)	C1	GTCGTAGA CAGCATCT	TGGATCGA
	C2	CTGCTCTG GAGAAGAC	GGTCCAGA
C(21.875,15.625)	C2	ATGCTGGG TACGACCC	CGAGATCA
	C3	GTTACGTG CAATGCAC	GATTCCAG
	C4	GTACACAG CATGTGTC	CGATCGTA
C(9.375,9.375)	C4	AATTGGCC TTAACCGG	TGGTATCT
C(15.625, 3.125)	C1	AAAACTGA TTTTGACT	ACCATAGA

Tab.1 DNA coding

4.3DNA program and experiments

After all the leaf cells being encoded by k-armed DNA structure, the biology reaction is beginning:

Step 1: Combine multiple copies of the leaf cells (all fluorescent mark cores with all their arms) in a single tubeT$_0$

Step 2: Allow the complementary ends to hybridize and be ligated in T$_1$

Step 3:Remove those structures which have not exactly match and have open-ends by exonuclease enzyme ,which are denoted by S ,put the remaining structures denoted by P into test tube T$_2$

Step 4:Select the DNA cliques comprised by different vertexes by gel electrophoresis and keep them in test tube T$_3$

Step 5:Find the DNA cliques which contain fluorescent mark cores DNA for each grid by their fluorescent mark (here, sometimes we delete some DNA sequence that the patterns they stand for being seen as yawp),then, keep them into test tube T$_4$

In the test tube T$_3$is the solution of clustering. If the number of the cliques DNA is k, there are K groups of clustering T$_3$.

We can obtain several structures of the DNA sequences, just as the instance figure9,there are three structures, showed in figure10-13 and the DNA programs are as table2.

Input(T$_0$).

All k-armed DNA sequences are placed in empty test tube T0.

merge(T$_0$,T$_1$,T$_1$)

Make all sequences in T0 mixed together and execute ligation process.

After hybridization process, all possible combinations of DNA sequences happen in T0,and they were put into test tubeT1.

T_2 ← +(T_1; P)

Select only those DNA strands which have matched exactly and have not have any open ends from T1 and keep them into empty test tube T2.

separate(T_2,T_3).

Using gel electrophoresis to find the DNA cliques in test tubeT2.

Put them in an empty tube T3.

This is the original solution of clustering the problem.

for i = 1 to N do {begin T_3 ← +(T_3; c_i) end;}T_4 ← T_3.

end for;

Select all k-armed DNA structures that contain all the n cores c1,…,cn in test tube T3. Put them in empty test tube T4.

return(T_4)

END

Tab.2 DNA program

The first structure is four-armed DNA linking with 4-armed DNA, because four units of leaf cells are all filled with data .

The second structure is a three-armed DNA,because there are three units of leaf cells are filled with data

The third structure is two-armed DNA, it is double DNA sequence.

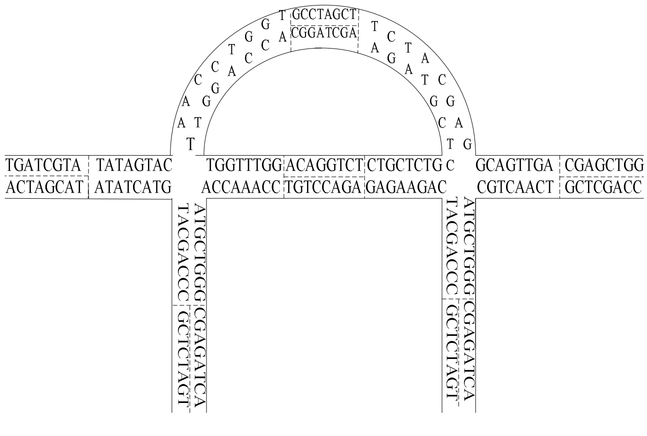


Fig.11 the result of four-armed DNA

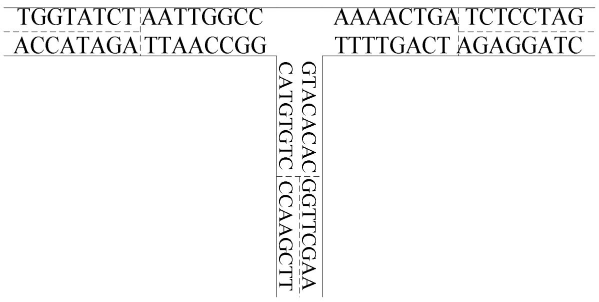


Fig.12 the result of three-armed DNA

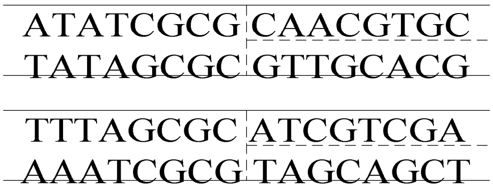


Fig.11 the result of two-armed DNA

To cluster data in figure9,we put enough DNA structures into test tube T_0,then we added oligonucleotide ,enzyme and gave other conditions which are needed in experiments . After reaction adequately, we used cut enzymes to remove DNA sequences which are not content with our demands. Test tube T_2 are the DNA sequences we want, the arm C4(21.875,21.875) linked with C2(21.875,15.675) , C3(21.875,15.675) andC4(21.875,15.675) are linked with C1(3.125,9.375) and C2(3.125,9.375) ,C4(9.375,9.375) combined with C1(15.625, 3.125), so the data being represented by leaf cells C(3.125, 21.875) , C(3.125,15.625) and C(3.125,9.375) became a cluster. Data being represented by C(9.375,9.375) and C (15.625, 3.125) became a cluster.

C(21.875,21.875) and C(21.875,15.625)became two single clusters because of not linking with others .Therefore, at last, we identified the DNA cliques by fluorescent mark cores. Data set in figure 9 has four clusters, the result of reaction shows in figure 14 and results in figure 15.

Next part ,we use the most popular data set ,the Iris data set, which has 150 patterns with four dimensions. We knew that this data set should be divided into three clusters. However, it is famous for the difficult of clustering because the data in this set are alternative and have not have any obvious bounds. So it is easily to put the data into wrong

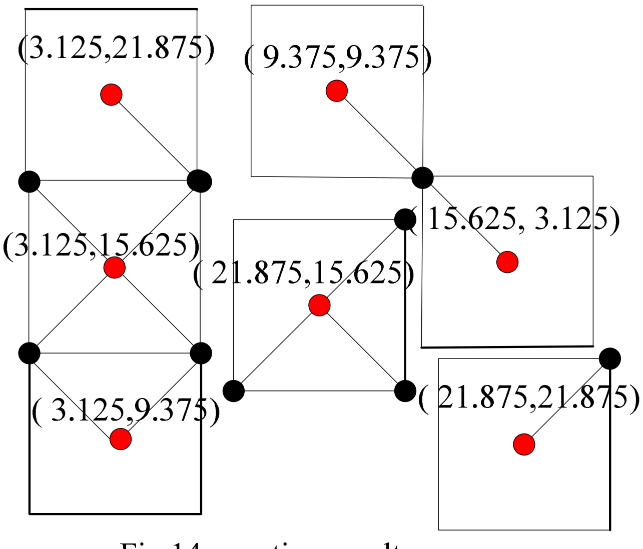


Fig.14 reaction result

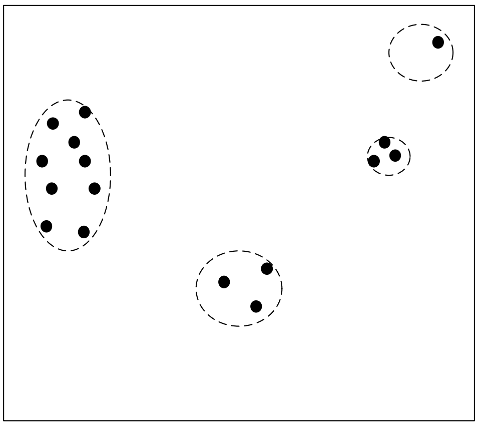


Fig.15 the result of clustering

clusters ,and this is the reason why there are many experts study in the clusters of the Iris data set .In order to suit for our algorithm ,we use two dimensions of this set firstly ,here, we choose two choice to complement our idea: the first-second dimension and second-third one ,data are showed in figure16 and figure18.

To the first-second dimension data set, we choose fit grid node in every step pg the tree for our data, according to the distribution of the two dimension data, at last, we have 100 leaf grid cells to divide data , threshold was defined as the unit having data. Arms are encoded as 16 base, problems were solved by step 1~step 5 mentioned above, standard chain reaction showed in table2. .Unfortunately,11 patterns were discard in the end .The result is in figure17.Data set was clustered into 3 cliques. There are 31 patterns are divided into wrong clusters.

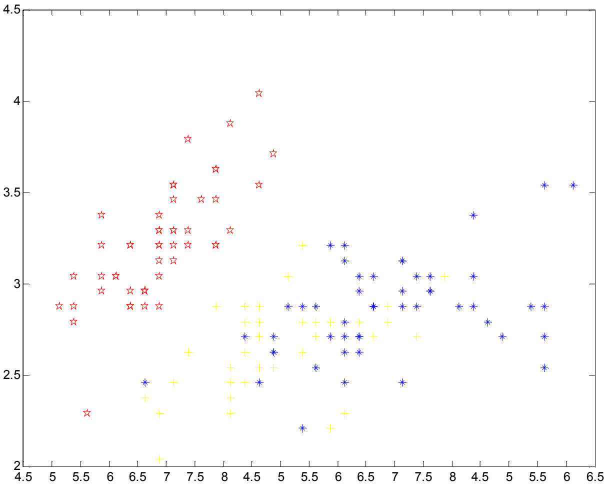


Fig.16 one-two dimension of Iris data set

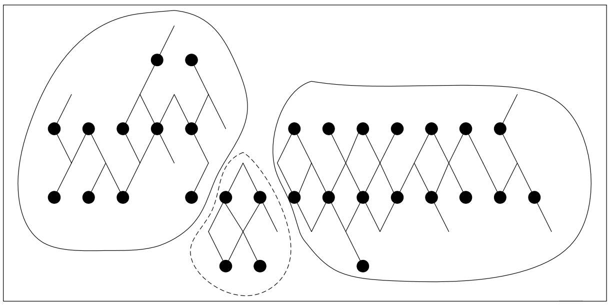


Fig.17 the clustering result of one-two dimension of Iris data set

To the second-third dimension data set, we get 180 leaf grid cells,data are also clustered into three groups, this time, there are 15 patterns in a wrong group and 23 patterns being discarded in figure19.

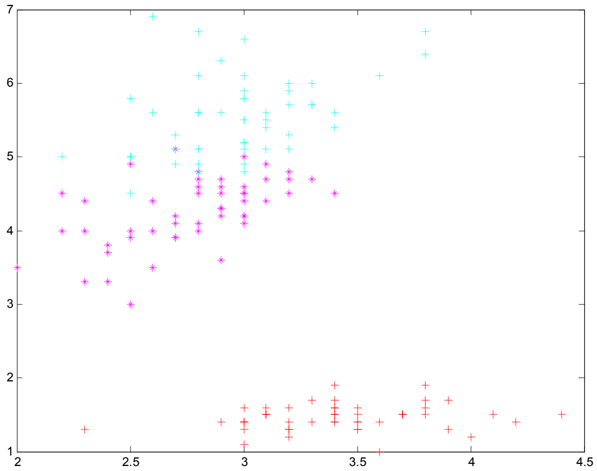


Fig.18 two-three dimension of Iris data set

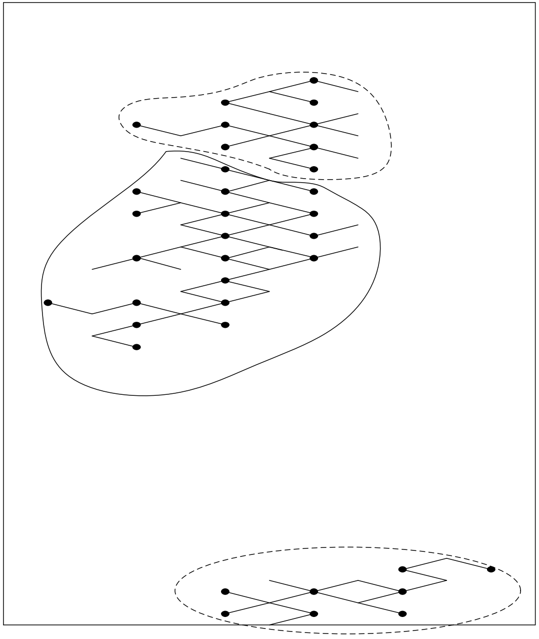


Fig.19 the clustering result of two-three dimension of Iris data set

5 Discussion

5.1Comparison with other clustering algorithm

The time complexity can be calculated by the steps showed in section 4(biology reaction).We assume that time consumed by coding is k, every annealing reaction time of two single DNA sequence is n. Everybody knows the biology reaction conducts in parallel. Therefore, the time consumed by annealing reaction of all the DNA arms is n. The remaining operations are just some simple work for finding which are helped by gel electrophoresis, DNA sequence measuring etc. All the time consumed by them can be seen as m(m approximately equal ton).So the whole time complexity of our algorithm is k+n+m, which means O(n).

There are many types of clustering, such as hierarchical clustering, Density-based clustering, Subspace clustering ,etc. The time complexities of the more frequent and useful algorithm are demonstrated in table3[13]It is obvious that the new clustering algorithm consumes less time than that show in table3.

clustering algorithm	time complexity
K-means	O(nkl)
K-median	$O(n^{i+\varepsilon)}$
Hierarchical agglomerative algorithms	$O(n^2 logn)$
MST	$O(n^2 logn)$
DBSCAN	O(nlogn)

Tab.3 The time complexities of the more frequent and useful algorithm

All the processes of our method are going in test tubes, which is a combination between biology and computer science .So it is a challenge of traditional silicon computer and provide a creative idea and novel thought to data mining and other field, the important significance is obvious.

In the other hand, we use the Iris data to testify the feasibility and validity of our algorithm, we know that the famous k-means algorithm and Particle Swarm Optimization algorithm are very effective on the clustering in data mining. However ,it can be seen in figure20 and figure21that k-means can not divide the data into three groups and PSO also can not do this, it is to say that these two algorithm give wrong clusters and their mistakes are more that 50 patterns .Hence, our algorithm is more useful. But we have to say that in our process we lost many patterns ,this is the place which we will focus on to improve.

5.2Comparison with clustering algorithm using DNA computing

So far, the clustering algorithms using DNA computing are the algorithm based on k-means and Fuzzy C-means algorithm [15] the CLIQUE algorithm based on closed-circle DNA sequences[24], the algorithm based on minimum spanning tree[23] ,their time complexities are all much less except for considering the time wastes for DNA coding. Actually, coding consumes much time because of they all use double DNA sequences and single ones. Compared with them ,our k-armed DNA structure is easy to encode and it can demonstrate those vertexes and edges in graphs more clearly.

On the other hand ,we know that data sets needed to be clustered are always masses of groups, the algorithms above use DNA chains to indicate the result of clustering ,which seems to lose the structures of cluster and sometimes may appear deviations. The clusters of our algorithm are groups which are similar to the original data sets because of the k-armed DNA structure, so it is much practical.

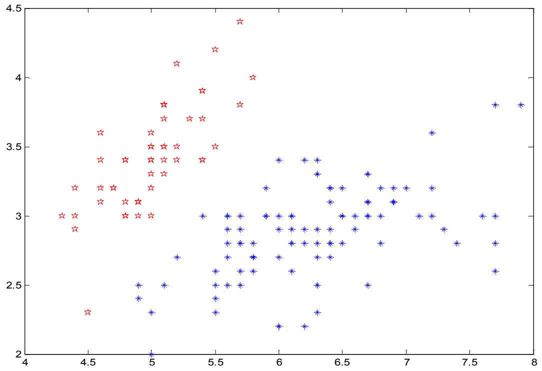


Fig.20 the clustering result of 1-2 dimension Iris data set by k-means and PSO algorithm

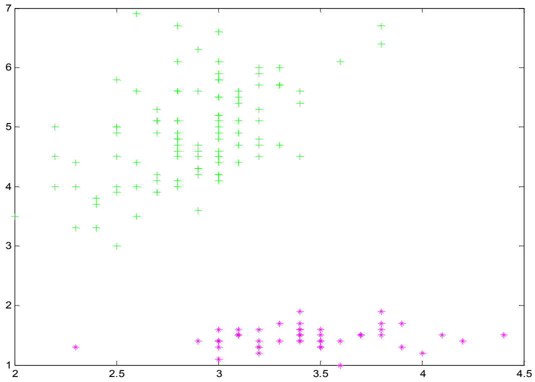


Fig.21 the clustering result of 2-3 dimension Iris data set by k-means and PSO algorithm

6 Conclusion

The greatest advantages of DNA computing are high parallelism and huge storage. Since Adleman's experiment, DNA computing techniques are considered to be suitable to solve NP-complete problems especially the combinatorial problems [4].

In this paper, we propose a new strategy for the clustering algorithm using DNA computing. We use gird tree to divided data and consider each leaf node cell of grid tree to be four smaller units, each unit is linked with the cell core which is a k-armed DNA structure. Therefore, the combination problem of the leaf cell becomes a problems to find several k-armed DNA structures who have same vertexes in a graph. We give a coding strategy to make the k-armed DNA structures to satisfy the requirements in Section 4.1.We also present the DNA coding methods in Section 4.2 and the biology reaction process and an realization example to illustrate it in Section 4.3.Finally, we discuss the time of the complexities of our methods with other clustering algorithm from two aspects. It is found that the simulated approach is faster than processing using silicon-computer and the DNA sequences is much suitable than the DNA chain in solving clustering problem because of its parallel characters and three dimensional DNA structure.

Although we give the process of our algorithm and add an instance to prove its feasibility, our work is still theoretical and there is much work for bio-chemical techniques to do. Three dimensional structure can represent three dimension structures and the CLIQUE algorithm also can solve spatial data, but at the current stage we just use them to cluster the two-dimensional data. In the future, we will continue to research how to using DNA computing techniques to cluster the three-dimensional or spatial data and at the same time we look forwards to proceeding with the bio-chemical experimentation.

References:

[1] Adleman L M., Molecular Computing of Solutions to Combinatorial Problems, *Science*, 1994, 266(5187): 1021-1023.

[2] Adleman, L.M., 1998. Computing with DNA, *Scientific American*,1998,54-61.

[3] Amos, M. et al ,Topics in the theory of DNA computing ,*J. Theor. Comput*.Sci. ,2002,287, 3-38.

[4] Bach, E., Condon, A.et al, DNA models and algorithm for NP-complete problems, *Proceedings of the 11th Annual IEEE Conferences on Computingal Complexity (CCC'96)*, 1996,290.

[5] Bakar, R.B.A., Watada, J.,A DNA computing approach to cluster-based logistic design ,*Proceedings of the 2nd International Conference on Innovative Computing*,2007,383-1383.

[6] Bakar, R.B.A., Watada, J.,Biological clustering method for logistic place decision making, *CKnowledge-Based Intelligent Information and Engineering Systems*, 2008,5179,136-143.

[7] Bakar, R.B.A., Watada, J.,A biologically inspired computing approach to solve cluster-based determination of logistic problem,*Biomedical Soft Computing and Human Sciences*, 2008,13, 59-66.

[8] Bakar, R.B.A., et al.,A DNA computing approach to data clustering based on mutual distance order, *Proceedings9th Czech-Japan Semin ar*,2006,13,139-145.

[9] Bakar, R.B.A.,Watada, J., Pedrycz,W., DNA approach to solve clustering problem based on a mutual order, *Biosystems*,2008,91,1-12.

[10] Ezziane,Z., DNA computing:applications and chanllenges, *Nanotechnology*,2005,17,27-39.

[11] Faulhammer, et al.,Molecular computing: RNA solutions to chess problems *Proceedings of the Natl. Acad. Sci*., 2000, 1328-1330.

[12] Han J. and M. Kamber, Data Mining, Concepts and techniques, *Higher Education Press*, Morgan Kaufmann Publishers, Beijing, 2000.

[13] Jain, A.K., Murty,M.N., Flynn, P.J., Data Clustering: *A Review,ACM Computer Surveys*, 1999, 264-323.

[14] Jain, A.K., Law, M., Data clustering: a user's dilemma, Proceedings of *International Conference on Pattern recognition and machine intelligence* (PReMI),

[15] Kim, S.Y., Lee, J.W., Bae, J.S., E_ect of data normalization on fuzzy clustering of DNA microarr ay data , BMC *Bioinformatics*, 7, 134.

[16] Lipton, R.J., DNA solution of hard computing problems, *Science,* 1995, 268(28), 542-545.

[17] Jonoska N., Karl S.A., Saito M., Three dimensi onal DNA structures in computing, *Biosystems*, 1999(52)143-153.

[18] Ouyang, Q., et al., NA solution of the maximal clique problem, *Science*,1997, 278, 446-449.

[19] P˘aun G, Rozenberg G., and Salomaa A., DNA Computing, New Computing Paradigms*, Springer-Verlag*, Berlin Heidelberg, 2010.

[20] Seeman, N.C., The perils of polynucleotides: the experimental gap between the design and assem - bly of unusual DNA structures, Landweber, L.Ba um. (Eds.),1999,215-233.

[21] W.L. Chang, M. Guo, Solving the set-cover problem and the problem of exact cover by 3-sets in the Adleman-Lipton's model, *BioSystems*, 2003(72) :263-275.

[22] W.L. Chang, Fast parallel DNA-based algorith -ms for molecular computing: the set-partit ion pro blem, IEEE Transactions on *Nanobioscience*(2007) 346-353.

[23] Xue Jie ,Liu xiyu ,Applying DNA Computing to Clustering in Graph,*2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce*,2011,986-989.

[24] Zhang Hongyan, Liu xiyu, A CLIQUE algorithm using DNA computing techniques based on closed-circle DNA sequences, *BioSystem s*,105(2 011)1-12.

[25] Head Tom, Formal language theory and DNA: an analysis of the generative capacity of specific recombinant behaviors, *Bulletin of Mathematical Biology*, 1987(49):737-759,.

[26]Chen JH, Seeman NC: Synthesis from DNA of a molecule with the connectivity of a cube. *Nature* 1991(350):631-633.

[27]Zhang Y, Seeman NC: The construction of a DNA truncated octahedron. *J Am Chem Soc* 1994,(116):1661-1669