

A Novel Chaotic Ant Swarm Based Clustering Algorithm for Clinical Prediction

Haohan Wang, Lixiang Li, Xi Yang and Chong Lian

Information Security Center in Beijing University of Posts and Telecommunications

P.O. Box 145, Beijing 100876, China

Abstract: In this study, we introduce an algorithm which is aimed at distinguishing patients of a certain disease out of healthy people when the thorough knowledge of this disease failed to be obtained. The algorithm deals with the physical parameters of each person in a group and when the information is sufficient, it separates the diseased people out. It is a clustering algorithm based on the Chaotic Ant Swarm (CAS). This algorithm is introduced in detail for researchers to use and has been experimented on four different real data sets. The results show that when sufficient data of each person is available, this new algorithm has a high precision.

Keywords: Algorithm, CAS, clinical prediction, clustering

INTRODUCTION

Requiring a deep research related to Clinical Prediction Rules (CPR), Clinical prediction is intended to facilitate clinical decision-making in the assessment and treatment of individual patient. CPR is a type of medical research study in which researchers try to identify the best combination of medical sign, symptoms and other findings in predicting the probability of a specific disease or outcome (McGinn *et al.*, 2000). CPRs are thought to be of great potential when they are developed and utilized for clinical conditions that involve complex clinical decision making (Haskins *et al.*, 2011). However, a process of validation for CPR is needed before the CPR is used (McGinn *et al.*, 2000, 2008; Reilly and Evans, 2006).

Insufficient knowledge of the characteristics of a particular disease can make it nearly impossible to perform clinical prediction with CPR. CPR cannot be used until a long time after a new disease breaks out. On this background, we introduce an algorithm to help with clinical decision making, especially when CPR is not ready. The algorithm is aimed to discriminate the healthy group and the diseased group only based on the physical features of each person, demanding no thorough knowledge of the disease.

To introduce this algorithm, we employ chaotic optimal solutions for finding the global optimal solutions, using chaotic variables to search the entire space. Inspired by the chaotic ant swarm (Li *et al.*, 2006), we proposed this chaotic ant swarm based clustering algorithm for clinical prediction.

We perform four experiments on real data set of clinical information to show that when provided with sufficient information (like thirty attributes of the physical

feature of a person), the algorithm has a relatively high precision for clinical use in emergency or for decision-making.

In this study, we introduce an algorithm named Chaotic optimal Solutions based Clustering (CAS-C) to cluster the people into two groups, healthy group and diseased group. In order to present the algorithm, we firstly introduce the background and overview of chaotic optimal solutions. Then we introduce the mathematical model and detailed workflow of CAS-C. We performed the experiment on four real data sets of disease. The results show when provided with sufficient data, the CAS-C algorithm shows a relatively high precision.

METHODOLOGY

Chaotic Ant Swarm based Clustering (CAS-C)

algorithm: In this section, we first introduce the overview of CAS algorithms briefly, then the formal mathematical model and algorithm process of Chaotic Ant Swarm based Clustering (CAS-C) will be given.

Overview of CAS algorithm: In 1990s, reference (Cole, 1991) discovered that a single ant shows low dimensional deterministic chaotic activity out of the periodic behavior of the ant colony. However, there is no further research on the relationship of individual chaotic behavior with the self-organization and foraging behaviors of the ant colony. In the view of dynamics, the interactions between the two behaviors must exist because these interactions are necessary for ants to survive. The solution of optimization problems can be adapted from these interactions. In this way, Chaotic Ant Swarm (CAS) (Li *et al.*, 2006) was developed for solving the problems

of optimization. This algorithm incorporates chaotic dynamics of ant, swarm organization and optimization principles.

In the algorithm of CAS, there are M ants in a D -dimensional search space S , trying to minimize a function J . Any point of S can be the solution and CAS is intended to rule the way the ants go as expected. The task can only be done after two phases of the colony: chaotic phase and organization phase.

First, they perform the chaotic behaviors and decreases with respect to the time in two phases respectively. In the whole process, each ant keeps exchanging information with the ants nearby. The changing process of position of ant i can be described as the following in math view (Li *et al.*, 2006):

$$\begin{aligned} y_i(t) &= y_i(t-1)^{1+r_i} \\ x_{id}(t) &= (x_{id}(t-1) + V_{id}) \times e^{1-e^{-ay_i/(t)3-\psi_d(x_{id}(t-1)+V_{id})}} \\ &+ (pbest_{id}(t-1) - x_{id}(t-1))e^{-2ay_i(t)+b} - V_{id} \end{aligned} \quad (1)$$

in which:

- t : Stands for the current iteration step; $(t-1)$ means the previous iteration step.
- $y_i(t)$: The organization variable for ant i in step t and $y_i(0) = 0.999$.
- $x_{id}(t)$: Current station of ant i in dimension d .
- $pbest_{id}(t-1)$: Stands for the best position found by ant i with the nearby ants in $(t-1)$ steps.
- V_{id} ($0 < V_{id} < 1$) : Determines the search region for ant i in dimension d .
- a : A positive constant and should be large enough. Generally 2000 is enough for a .
- b : A constant and $0 < b < 2/3$. It is recommended to be selected as $1/2$ or $2/3$.
- y_i : An organization variable that decreases through the time. The decrement of y_i makes the influence of it increase to control the chaotic behavior of an individual ant.

In addition, r_i and ψ_d are two parameters of importance in CAS. r_i is a positive constant less than 1, named as organization factor of ant i , in direct proportion with the convergence speed of CAS algorithm. r_i is usually determined by the concrete problem and runtime. The ants of the colony do not necessarily have the same r_i and we can assign r_i as $r_i = 0.1 + 0.2 \text{ rand}(1)$.

ψ_d determines the search region of CAS. An approximate formula $\omega \approx 7.5/\psi_d$ can be obtained if the interval of the search region is $[-\omega_d/2, \omega_d/2]$. Additionally,

$V_{id} = \omega_d/2$ ($0 < V_{id} < 1$) can shift the interval to $[0, \omega_d]$. Details of adjusting each parameter and the corresponding impact of the adjustment is deeply discussed in Li *et al.*, (2006).

CAS-C algorithm for clinical prediction: Our algorithm is introduced to deal with the physical features of a group of people when a new disease breaks out. It helps doctors in the way to identify the diseased people in the group and perform clinical prediction based on the diagnosed results. It functions as a colony of ants forage, searching for the food (the centroid of diseased group and healthy group). In the initial step, ants are randomly placed, in another statement, numbers of a person's physical information are randomly selected as the positions of ants. The ants search for the food in each step of the iteration and gradually converge to the centers.

Since our algorithm is introduced for an unfamiliar disease demanding a quick diagnose, we should not wait for all the ants get to the centers. The algorithm ends when a certain number of iterations, represented by $Istep$, are done.

Given the number of the groups as K , in order to find each center of the group $z_p(I, 2, \dots, k)$, the iteration function is as following:

$$\begin{aligned} y_i(t) &= y_i(t-1)^{1+r_i} \\ z_{pid}(t) &= (z_{pid}(t-1) + V_{id}) \times e^{(1-e^{-ay_i(t)})3-\psi_d(z_{pid}(t-1)+V_{id})} \\ &+ (zbest_{pid}(t-1) - z_{pid}(t-1)) \\ &\times e^{(-2ay_i(t)+b)-V_{id}} \end{aligned} \quad (2)$$

where,

- $z_{pid}(t)$: Stands for the current state of ant i in dimension d for the p^{th} desired center z_p
- $zbest_{pid}(t-1)$: Means the best position in dimension d for all ants within $(t-1)$ steps

Other parameters function the same as in (1)

After both of the centers are found, all the people should be diagnosed into two groups. The algorithm should guarantee that all the diagnosed healthy people are similar to each other; all the diagnosed diseased people are similar to each other; and the people from different groups should be dissimilar enough. In order to give a convincing diagnose, the algorithm calculates the similarity by calculating the distance. In a mathematics view, the cluster of the data is determined by:

$$\begin{aligned} C_i &= \{x_j \mid \|x_j - z_j\| \leq \|x_j - z_p\|, x_j \in S\}, \\ p &\neq i, p = 1, 2, \dots, k, \\ z_i &= \frac{1}{C_i} \sum_{x_j \in C_i} x_j, i = 1, 2, \dots, k \end{aligned} \quad (3)$$

where,

$\|\bullet\|$: Stands for the distance between two points
 x_j : Stands for the data clustered into C_i
 z_i : Means the center of C_i , which can be defined by the average of all the data in this cluster

The points nearest to z_i gather to form the C_i .

After all the people are diagnosed, if there is a new need to diagnose people, the first process can be eliminated. The diagnosing process can be finished by only calculating the distance between physical features of a person and the center.

A work flow is given to show how the CAS-C helps doctors with the diagnose of the physical features of people. With the cluster numbers K , the CAS-C will be performed as follows:

- **Initialization:** We should assign some parameters before the algorithm starts: the ants number M , the number of iteration steps $Istep$, organization factor r_i , organization variable y_i , search scope ψ_d according to the size of the data sample. Then, assign $t = 1$ and randomly place the ants in the data sample.
- **Iteration process:** At Step t , the best position found by ant i and its nearby neighbor within $t-1$ steps is worked out as $zbest_{pi}(t-1)$, then each ant moves as (3). After each step of iteration, calculate the $zbest_{pi}(t-1)$ for ant i and store it for the next iteration. If the iteration has performed $Istep$ steps, the cycle is terminated and move to Step (3). If not, repeat this Step.
- **Get the centers:** After the iteration, the ants will converge to two places of the data sample, where are the cluster centers for healthy and diseased group, respectively.
- **Partition the data and get the result:** After getting the center, calculate the distance of the personal information and allocate the person into the group. Mark the data with its label and end the process.

The detailed process of the algorithm is showed below:

Input: Data Set: $X = x_1, x_2, x_n$

Begin

- Initialize the search scope ψ_d , organization factor r_i and position z of M ants randomly, in which each single ant z_i contains 2 randomly generated centroid vectors: $z_i = \{z_{i1}, z_{i2}\}$
- for $t = 1: Istep_{max}$ do
- for $i = 1: M$ do
- calculate the objective function $J(i; t)$ with current $z_i(t)$
- $J_{last} = J(i; t-1)$
- $y_i(t) = y_i(t-1)^{1+r_i}$
- $z_i(t) = (z_i(t-1) + V_i) \exp((1-e^{ay_i(t)})(3-j(z_i(t-1)+V_i))) + (zbest_i(t-1)-z_i(t-1)) \exp(-2ay_i(t)+b)-V_i$
- Calculate $J(i; t)$ with current $z_i(t)$ according to (2).
- if $J(i; t) < J_{last}$ then
- $zbest_i(t) = z_i(t) /* zbest_i$ is the best position found so far for ant $i.*/$

- else
- $zbest_i(t) = zbest_i(t-1)$
- end if
- end for
- Update the global best position ($zbest_g$): Select the best $zbest_i$ from $zbest_1; zbest_2$ as $zbest_g$. /* $zbest_g$ represents the global best position in the neighborhood of each ant.*/
- $z_1, z_2 = zbest_g$
- end for
- for $j = 1: n$ do
- for $c = 1: 2$ do
- Calculate distance $dc = \|x_j - z_c\|$
- end for
- $d = d1, d2$
- Find the position p of $\min(d)$
- $C_p.add(x_j)$
- end for
- End
- Output: Cluster Result: C_1, C_2

After the process, the data will be labeled to indicate the group and the clinical diagnosis is performed. However, if there are other people in need of the clinical diagnosis, the first process of searching for the center can be omitted since the centers for each disease should be similar.

Experiments: In this section, the experiments and the results are showed to indicate that our algorithm works well on helping clinical diagnosis when sufficient information is provided. However, in order to show the result more clearly, we firstly introduce another algorithm, the classic clustering algorithm named K-means for comparison. If both CAS-C and K-means have a high precision in one experiment, it is probably because the data set is easy for clustering.

Algorithm for comparison: K-means (MacQueen, 1963), as an algorithm based on partition, is the most famous classic algorithms. It is widely used because of its simplicity and effectiveness.

The detailed flow of this algorithm is showed below:

Input: Data Set: $X = x_1, x_2, x_n$

Expected Groups: K

Begin

- Initialize K centers for C_1, C_2, C_k . Randomly select K data points from X as the initial centroid vectors
- repeat
- Assign each data point to its closest centroid and from K clusters.
- Recompute the centroid for each cluster
- until
- Centroid vectors do not change
- End
- Output: Clustered Results: $C1, C2, Ck$

Table 1: The denotation of clustering results

	In the group before clustering	Not in the group before clustering
In the group after clustering	a	b
Not in the group after clustering	c	d

EVALUATION METHODS OF THE RESULTS

In this section, we introduce two criteria to evaluate the results.

The first one is the classic criterion for measuring clustering algorithms: F-measure. It is a measurement generally used in the field of statistics verification and pattern recognition. It combines the most frequently used two criteria used in the field of searching and statistics-- *Precision* and *Recall*.

For each clustered group C_i , we can show the result in the Table 1.

In the table, a represents the number of the instances that are labeled in C_i before clustering and labeled in C_i after clustering. b represents the number of the instances that are not labeled in C_i before clustering but labeled in C_i after clustering. c represents the number of the instances that are labeled in C_i before clustering but not labeled in C_i after clustering. d represents the number of the instances that are not labeled in C_i before clustering and not labeled in C_i after clustering.

With the definition of a, b, c, d , we can get that:

$$precision = \frac{a}{a+b} \quad (4)$$

$$recall = \frac{a}{a+c} \quad (5)$$

The value of *precision* and *recall* are between 0 and 1. Often, we hope to get bigger values of *precision* and *recall*. However, this cannot be achieved for both *precision* and *recall*, when one of them is higher, the other is consequently lower. For this reason, F-measure is introduced to measure an algorithm:

$$F = \frac{(m^2 + 1).precision.recall}{m^2.precision + recall} \quad (6)$$

when $m = 1$, F can evenly reflect the status of *precision* and *recall*. F is valued from 0 to 1, inclusively, the higher the value of F is, the better the result will be. In order to reflect the *precision* and *recall* evenly, we select m as 1 for the testing.

However, since our algorithm is introduced for clinical prediction, not for clustering, F-measure cannot reflect the precision for clinical prediction. We also use the *precision* to evaluate the work of the algorithm. High precision means that when a person is diagnosed into a diseased group, it is more likely that he will need a

detailed medical observation. High precision can guarantee a good diagnosing results to be used for doctors or biologists.

Experiments on real data sets: In this section, the evaluation of the performance of CAS-C clustering groups of people of a certain disease is presented, through four real diseases data sets, in the comparisons with k-means algorithm. Though this algorithm is introduced to deal with unfamiliar diseases, the experiments should be based on the data of existing diseases so that we can ensure the data set is convincing in order to show the performance of this algorithm. We perform the experiments on four classic data sets. The performance should be portable for other data sets.

The first of these four data sets is the Breast Cancer Data Set from Frank and Asuncion (2010). The data set consists of 569 samples, each with three-cell nucleus that are featured by thirty attributes in total. The clustering result of these 569 samples should be showed as clustered into two groups: the malignant ones and the benign ones.

The second data set is the Parkinson Disease Data Set containing 197 instances of patients and healthy people. Each person has a record of twenty-three attributes. This data set is originally from Little *et al.* (2007), but some instances with missing attributes are removed in case of misleading the clustering process, now this data set is composed of 162 people, twenty-four without Parkinson's disease. The expected clustering result is that the diseased people are picked out from this group.

The third data set contains the results of benign disease study, which is originally studied in Hosmer and Lemeshow (1989). After removing the 28 samples with missing information, this data set is made up of 172 observations, each described with thirteen variables. Forty samples are suffering the benign disease, while others serve as the control group. The cases include the women with a biopsy-confirmed diagnosis of fibrocytes breast disease identified through two hospitals in New Haven, Connecticut. Controls are selected from among patients admitted to the general surgery.

The last data set is named Lupus Nephritis Data Set. It is a data set arising from eighty-seven persons with lupus nephritis. The original data set contains nearly fifty variables for each instance. However, the most popular data set only contains three attributes. We introduce this to show how the algorithm works when the information is not sufficient and not precise.

The results is shown in Table 2 and visually reflected in Fig. 1. For all the four real data sets, the new algorithm has shown an advantage in precision. For the first data set, the ants firstly search out the centers for each attribute, as shown in Fig. 2. However, since the domain of each attribute is variable. The figure cannot reflect the result clearly. The result is also showed in Table 3. Then, calculate the distance between the attributes of each

Table 2: The results of the experiments

	CAS-C precision	CAC-C F -measure	K-means precision	K-means F-measure
Breast cancer	0.9051	0.8546	0.6046	0.5966
Parkinson disease	0.8580	0.8955	0.7346	0.4557
Benign breast cancer	0.7401	0.8857	0.5085	0.6603
Lupus nephritis	0.7471	0.8500	0.7356	0.5773

Table 3: Clustering centers for clinical diagnose

C	H	D	C	H	D	C	H	D
1	13.1700	16.1600	11	0.2023	0.4322	21	14.9000	19.4700
2	18.2200	21.5400	12	0.6850	1.2650	22	23.8900	31.6800
3	84.2800	10.6200	13	1.2360	2.8440	23	95.1000	129.7000
4	537.3000	809.8000	14	16.8900	43.6800	24	68.7600	117.5000
5	0.7466	0.1008	15	0.0059	0.0049	25	0.1282	0.1395
6	0.0599	0.1284	16	0.0149	0.0195	26	0.1965	0.3055
7	0.0485	0.1043	17	0.0156	0.0222	27	0.1876	0.2992
8	0.0287	0.0561	18	0.0085	0.0092	28	0.1045	0.1312
9	0.1454	0.2160	19	0.0109	0.0153	29	0.2235	0.3480
10	0.0544	0.0589	20	0.0017	0.0024	30	0.0692	0.0762

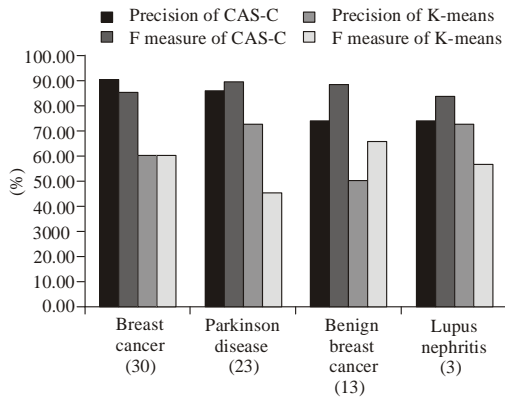


Fig. 1: Results of experiments

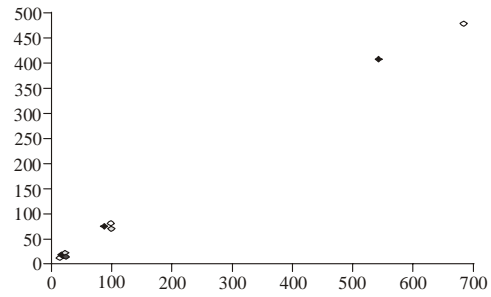
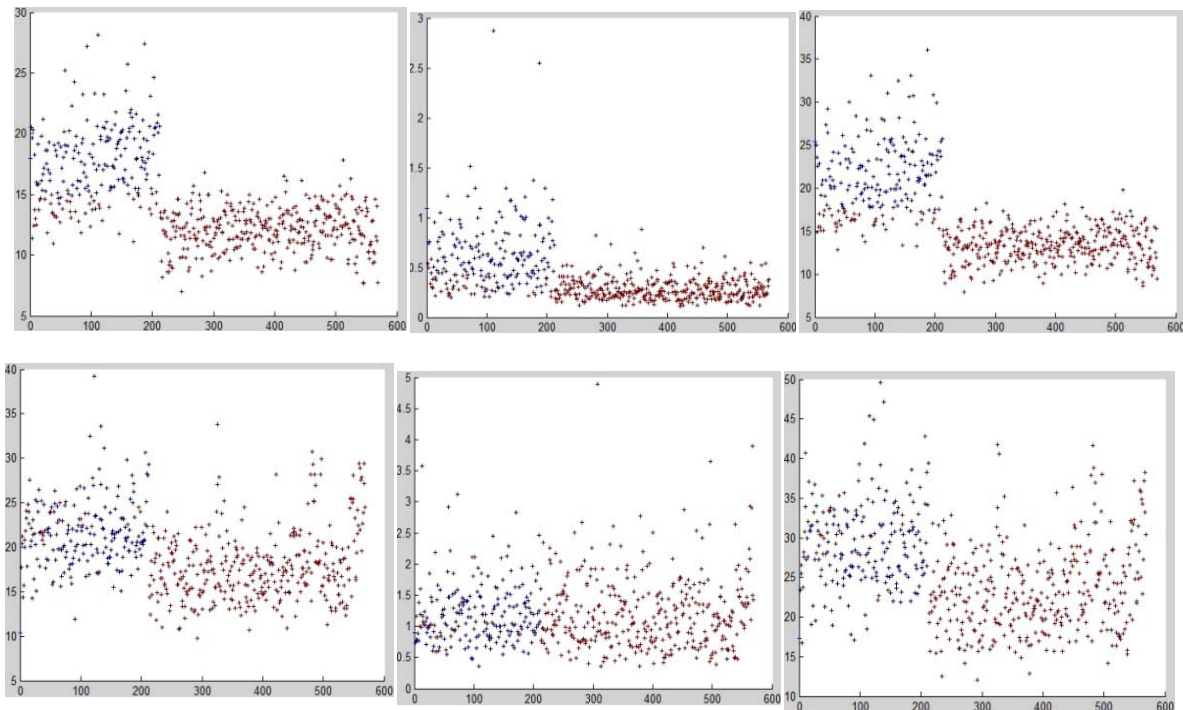
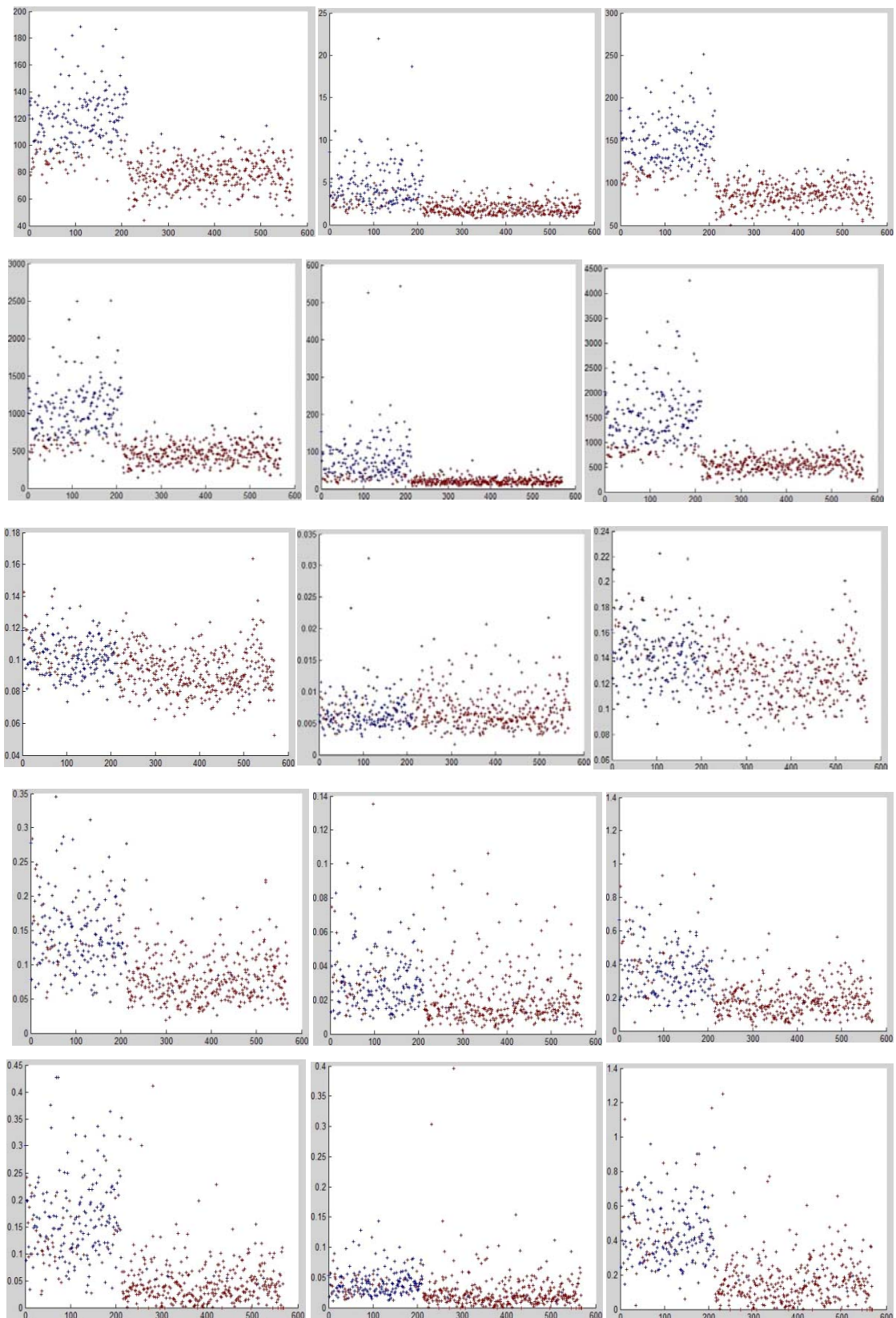


Fig. 2: The clustering centers, the centers for clinical diagnosis. in the figure, the x-coordinate reflects the healthy group while the y-coordinate reflects the diseased group

instance and the centers of each attribute. The clustering result is shown in Fig. 3. Each graph shows the result of the clustering result of an attribute. Most of the points are clearly divided into 2 groups, representing healthy people and diseased people, respectively. This result shows the outcome of our algorithm for clinical diagnosis. However, for clinical prediction, since the center of each attribute is calculated, the algorithm can just calculate the distance with Eq. (4) to cluster and to perform the clinical prediction.

Figure 1 shows that the precision of this data set is up to more than ninety percent, which indicates that when a person is clustered into the diseased group, it is very likely that she will have a breast cancer and therefore needs a specific observation. The Parkinson Disease Data Set shows a precision of about 85%, which is also satisfying.



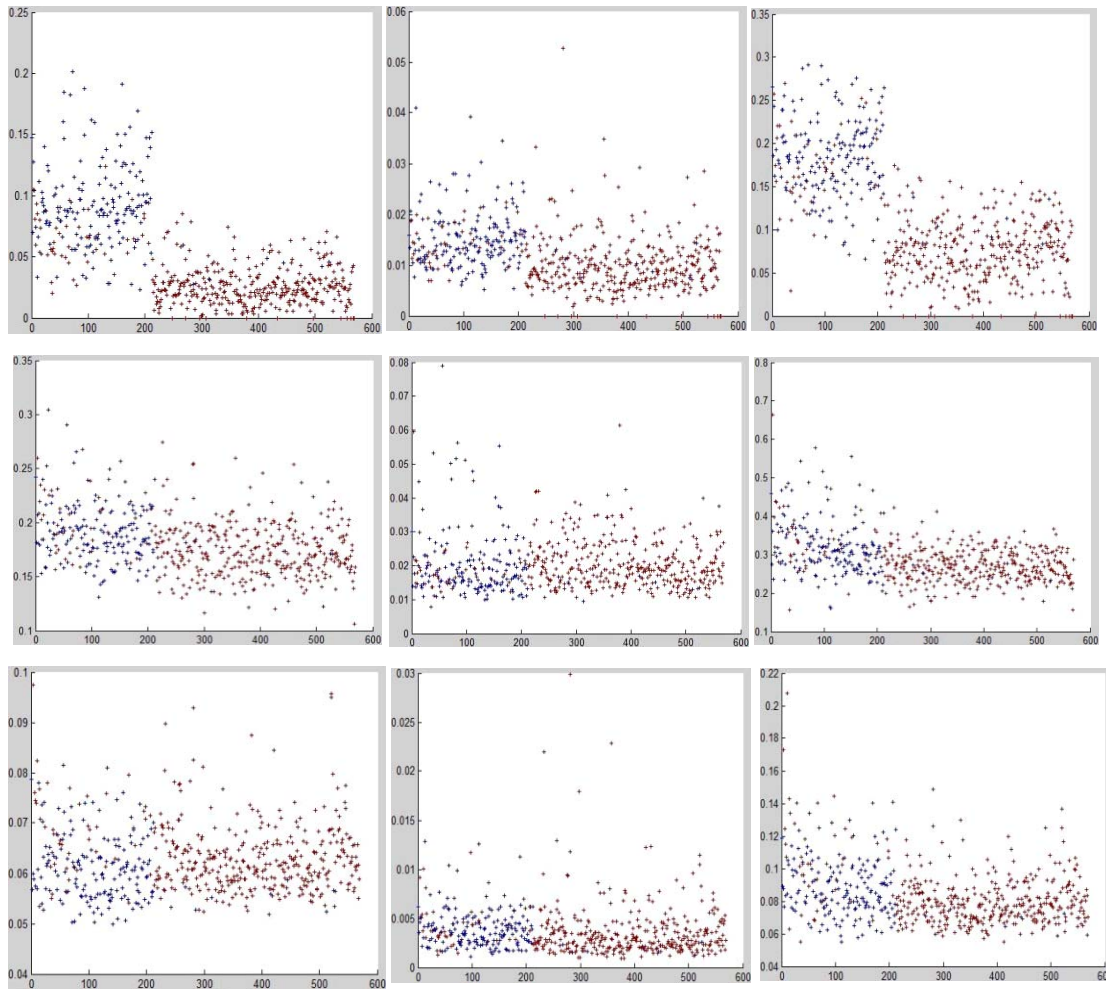


Fig. 3: The clustering result of features in breast cancer

The precision for the Benign Breast Cancer is relatively low. The precision about 75% cannot be used for clinical diagnosing, which means this algorithm cannot work well. However, all these 3 results of CAS-C are better than the result of K-means, indicating that these data sets are not easily for clustering by normal clustering algorithms. For the Lupus Nephritis Data Set, though the precision is up to 74%, the precision of K-means is nearly the same, which means that the precision of 74% is achieved mainly because this data set is easy for cluster. The precision goes down as the number of attributes declines. Ninety percent for 30 attributes, 85% for 23 attributes, 74% for attributes, but still higher than K-means. The CAS-C algorithm can make a good clinical prediction with 30 attributes, but it can never make the clinical decision with only 3 attributes. It is usually impossible for a doctor or biologist to judge whether a person is diseased or healthy by observing only 3 physical features.

The results of the experiment show that when sufficient data of one's physical feature is obtained, this

algorithm can be applied to clinical prediction with a precision about 90%. Though it still cannot predict the disease with a perfect precision, it can help the doctors in the process of clinical diagnosing.

CONCLUSION AND FUTURE WORK

In order to perform clinical diagnose and prediction for the disease with a lack of thorough knowledge, we introduce an algorithm named Chaotic optimal Solutions based Clustering (CAS-C) to cluster the people into 2 groups, healthy group and diseased group. In order to present the algorithm, we firstly introduce the background and overview of chaotic optimal solutions. Then we introduce the mathematical model and detailed workflow of CAS-C. We performed the experiment on four real data sets of disease. The results show when provided with sufficient data, the CAS-C algorithm shows a relatively high precision.

This high precision can guarantee that this algorithm can work well to help the doctors and biologists to make the clinical diagnose prediction, especially when a new disease breaks out. Besides this, when a plague strikes, there are hundreds of people in a hurry need to be diagnosed, only with a physical examination on each of them, the algorithm can diagnose them with a ninety precision of clustering process.

Though 90% is high enough to help the doctors make clinical decision, the precision can still be improved. In the future, we need to do more experiments to test its stability and consider improving its precision.

ACKNOWLEDGMENT

The authors would like to thank the anonymous referees for their valuable comments and suggestions to improve the presentation of this study. This study is supported by the the Foundation for the Author of National Excellent Doctoral Dissertation of PR China (FANEDD) (Grant No. 200951), the Program for New Century Excellent Talents in University of the Ministry of Education of China (Grant No. NCET-10-0239) and the Fok Ying-Tong Education Foundation, China (Grant No. 121062).

REFERENCES

- Cole, B.J., 1991. Is animal behavior chaotic? Evidence from the activity of ants. *Proc. R. Soc. Lond. B Biol. Sci.*, 144: 253-259.
- Frank, A. and A. Asuncion, 2010. {UCI} Machine Learning Repository. Retrieved from: <http://archive.ics.uci.edu/ml>.
- Haskins, R., D.A. Rivett and P.G. Osmotherly, 2011. Clinical prediction rules in the physiotherapy management of low back pain: A systematic review. *Man. Ther.*, 17(1): 9-21.
- Hosmer, D.W. and S. Lemeshow, 1989. *Applied Logistic Regression*. John Wiley and Sons, New York, Appendix 5.
- Li, L., Y. Yang, H. Peng and X. Wang, 2006a. An optimization method inspired by chaotic ant behavior. *Int. J. Bifurc. Chaos*, 28: 2351-2364.
- Li, L., Y. Yang, H. Peng and X. Wang, 2006b. Parameters identification of chaotic systems via chaotic ant swarm. *Chaos Solitons Fractals*, 28: 1204-1211.
- Little, M.A., P.E. McSharry, S.J. Roberts, D.A.E. Costello and I.M. Moroz, 2007. Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. *Biomed. Eng.*, 6(1): 23.
- McGinn, T.G., G.H. Guyatt, P.C. Wyer, C.D. Naylor, I.G. Stiell and W.S. Richardson, 2000. Users' guides to the medical literature XXII: How to use articles about clinical decision rules. *Evidence-Based Medicine Working Group, JAMA*, pp: 115-116.
- MacQueen, J., 1963. Some methods for classification and anylysis of multivariate observations. *Proceedings of the 5th Berkeley Symposium on Mathematical Statics and Probability*. Univ. of Calif. Press, 1: 281-297.
- McGinn, T., P. Wyer, J. Wisnivesky, P.J. Devereaux, I. Stiell, S. Richardson, *et al.*, 2008. Advanced topics in diagnosis: Clinical prediction rules. *A Manual for Evidence Based Clinical Practise*.
- Reilly, B.M. and A.T. Evans, 2006. Translating clinical research into clinical practice: Impact of using prediction rules to make decisions. *Ann. Int. Med.*, 144: 201-209.