International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:2, No:8, 2008

Classifying Bio-Chip Data using an Ant Colony System Algorithm

Minsoo Lee, Yearn Jeong Kim, Yun-mi Kim, Sujeung Cheong, and Sookyung Song

***Abstract*—**Bio-chips are used for experiments on genes and contain various information such as genes, samples and so on. The two-dimensional bio-chips, in which one axis represent genes and the other represent samples, are widely being used these days. Instead of experimenting with real genes which cost lots of money and much time to get the results, bio-chips are being used for biological experiments. And extracting data from the bio-chips with high accuracy and finding out the patterns or useful information from such data is very important. Bio-chip analysis systems extract data from various kinds of bio-chips and mine the data in order to get useful information. One of the commonly used methods to mine the data is classification. The algorithm that is used to classify the data can be various depending on the data types or number characteristics and so on. Considering that bio-chip data is extremely large, an algorithm that imitates the ecosystem such as the ant algorithm is suitable to use as an algorithm for classification. This paper focuses on finding the classification rules from the bio-chip data using the Ant Colony algorithm which imitates the ecosystem. The developed system takes in consideration the accuracy of the discovered rules when it applies it to the bio-chip data in order to predict the classes.

***Keywords*—**Ant Colony System, DNA chip data, Classification.

I. Introduction

THE widely used gene analysis methods are Southern and Northern blot which use the characteristics of the bonding between the DNA and DNA or between the RNA and DNA. Most of the Southern blots are used to find out the similarity of the gene information and the Northern blots are used to calculate the expression for specific genes. These methods have the difficulty in searching extremely large gene data sets at once.

Bio-chip is the newly developed method to overcome the problems of the existing gene analysis methods. A bio-chip is combines technologies from the existing molecular biology field and computer engineering field. It enables at least thousands to millions of the DNA to be placed on the small spaced area. Biologists can experiment with genes by easily using bio-chips. After the experiments, the analysis process is needed. The analysis process uses the data extracted from the bio-chip and mines the data to find patterns for similarity among genes, etc.

In this paper, we mainly focus on the mining part of the analysis process. The analysis system takes the preprocessed bio-chip data and classifies it to predefined classes. An algorithm that imitates the ecosystem is used as the classification algorithm. As bio-chip data deals with at least thousands to millions of records, in order to achieve high performance the algorithm for classification should be suitable for processing extremely large sets of data and most of the algorithms that imitate the ecosystem are suitable for this kind of computation. In our system, we used the Ant Colony algorithm for classification. The analysis system creates the rules by using the Ant Colony algorithm and the best rule will be chosen as the rule to be applied. A prediction is made on the test data set on which class the data value belongs to and the accuracy of the predictions are calculated.

The organization of this paper is as follows. Section 2 provides a survey of related work regarding classification and the algorithms that imitate the ecosystem. Section 3 gives an overview of the bio-chip analysis system. Section 4 explains the Ant Colony based classification system. Section 5 describes the implementation and experimental results of the Ant Colony based classification system. Section 6 gives the summary and conclusion.

II. Related Research

Classification is perhaps the most familiar and most popular data mining technique. Examples of classification applications include image and pattern recognition, medical diagnosis, loan approval, detecting faults in industry applications and classifying financial market trends.

Classification is a two-step process. In the first step, a model is built describing a predetermined set of data classes or concepts. The model is constructed by analyzing database tuples described by attributes. The data tuples analyzed to be built the model collectively form the training data set. The individual tuples making up the training set are referred to as training samples and are randomly selected from the sample population. Since the class label of each training sample is provided this step is also known as supervised learning. The learned model is represented in the form of classification rules,

Manuscript received July 25, 2006. This work was partially supported by the Brain Korea 21 program, and partially supported by the Ministry of Commerce, Industry and Energy.

Minsoo Lee is with the Dept of Computer Science and Engineering, Ewha Womans University, 11-1 Daehyun-Dong, Seodaemoon-Ku, Seoul, Korea 120-750 (corresponding author phone: +82-2-3277-3401; fax: +82-2-3277-2306; e-mail: mlee@ewha.ac.kr).

Yearn Jeong Kim, Yun-mi Kim, Sujeong Cheong, and Sookyung Song are with the Dept of Computer Science and Engineering, Ewha Womans University, 11-1 Daehyun-Dong, Seodaemoon-Ku, Seoul, Korea 120-750 (e-mail: {inverno, cherish11, bloom01, happymint}@ewhain.net).

decision trees, or mathematical formulae.

In the second step, the model is used for classification. First, the predictive accuracy of the model is estimated. The accuracy of a model on a given test set is the percentage of test set samples that are correctly classified by the model. For each test sample, the known class label is compared with the learned model's class prediction for that sample. The accuracy of the model is estimated based on the training data set. If the accuracy of the model is considered acceptable, the model can be used to classify future data tuples or objects for which the class label is not known [3].

Prediction can be viewed as the construction and use of a model to assess the class of an unlabeled sample, or to assess the value or value ranges of an attribute that a given sample is likely to have.

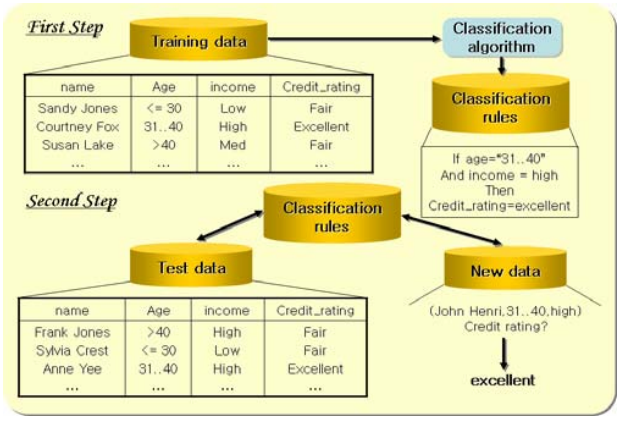


Fig. 1 Classification Process

The most popular algorithm that imitates the ecosystem is a genetic algorithm [2, 9]. A genetic algorithm is a search technique used in computer science to find approximate solutions to optimization and search problems. Specifically it falls into the category of local search techniques and is therefore generally an incomplete search. Genetic algorithms are a particular class of evolutionary algorithms that use techniques inspired by evolutionary biology such as inheritance, mutation, selection, and crossover.

Genetic algorithms are typically implemented as a computer simulation in which a population of abstract representations of candidate solutions to an optimization problem evolves toward better solutions. Traditionally, solutions are represented in binary as strings of 0s and 1s, but different encodings are also possible. The evolution starts from a population of completely random individuals and happens in generations. In each generation, the fitness of the whole population is evaluated and multiple individuals are stochastically selected from the current population based on their fitness, and modified by mutation or crossover to form a new population. The new population is then used in the next iteration of the algorithm.

Neural Network [1,9,10] is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. The key element of this paradigm is the novel structure of the information processing system. It is composed of a large number of highly interconnected processing elements called neurons working in unison to solve specific problems. Neural networks, like people, learn by example. Neural networks are configured for a specific application, such as pattern recognition or data classification, through a learning process. Learning in biological systems involves adjustments to the synaptic connections that exist between the neurons. This is also same in the neural network as well.

III. Bio-Chip Analysis System

Bio-chips in other words microarrays are the tools for gene expression analysis. The bio-chip is consisted of the probe that is a single strand DNA which is printed on the solid substrate simply called a chip. The types of the chip or the name of the chip depends on the method of the chip fabrication.

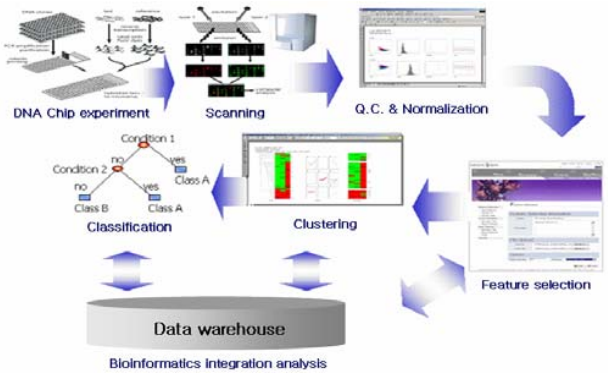


Fig. 2 Bio-chip analysis system

The idea of the bio-chip is the DNA in the solution that contains sequences complementary to the sequences of the DNA deposited on the surface of the array will hybridized to those complementary sequences. Usually, this interrogation is done by washing the array with a solution containing ssDNA, called a target. The key to the interpretation of the microarray experiment is in the DNA material that is used to hybridize on the array. Since the target is labeled with a fluorescent dye, a radioactive element, or another method, the hybridization spot can be detected and quantified easily. After the hybridization process, the scanner scans the spots and converts quantity to numeric values namely expression values. This process is called image processing. Quality control and Normalization process is performed in order to get rid of unnecessary data and the data that can influenced to the whole expression values. It also adjusts for any bias which arises from variation in the microarray technology rather than from biological differences between the RNA samples of the printed probes [8]. By the processing of quality control and normalization, the data are filtered and as a result the data that are meaningful and significant remain. These adjusted and arranged data are used as an input to create the classification rules. Classification is done by the algorithm that imitated the ecosystem in this case the Ant Colony algorithm is used. The rules are applied to the

test set of data and the data are predicted to which classes it belongs according to the rules and the accuracy of predicting a class is calculated.

The analysis system we explain in the following sections will be focused on the last step in the process, classifying and predicting the class including calculating the accuracy.

IV. Ant Colony based Classification

The main goal of the analysis system is classifying the sample data into predefined classes. The bio-chip data used in the analysis system is a two-dimensional microarray where one axis represents the sample or experiments data and the other represents the gene or probe data.

A. System Overview

The analysis system classifies the bio-chip data using an Ant Colony algorithm [5, 6]. It gains input data from the database which contains bio-chip data that pass through the normalization and quality control process. The system accesses the database and runs the classification algorithm based on the Ant Colony algorithm to create classification rules. In order to create the classification rules it needs to be trained and the training data sets are gained from an input file format with an extension of *.arff. The algorithm used for training is the Ant Colony algorithm. After the training process, many rules are produced and the best rules would be chosen as a final set of rules. Then the rule is applied to the test set and from the predicted class results we can compare them with the correct answers provided by the training data set and get the accuracy of the prediction. Fig. 3 shows an overview of the analysis system.

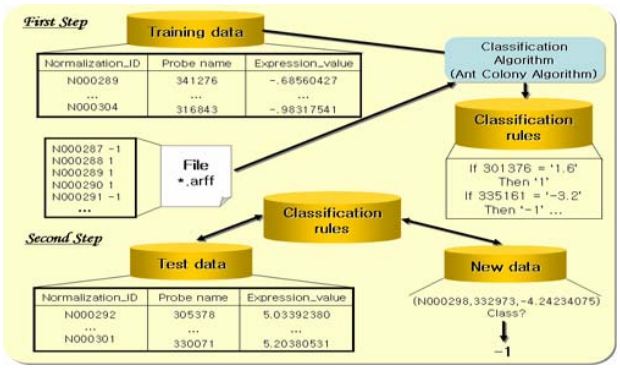


Fig. 3 Overview of the Analysis System

The analysis system can be divided into two-phases. The first phase is creating classification rules and the second phase is performing classification using the discovered rules. Each phase is explained in the following subsections.

B. Discovering Classification Rules

This phase is composed of two steps. The first step is the input pre-processing step where the analysis system gets data from the database and converts it into an appropriate data format so that it can be used in the analysis system. Additionally, the binning process is done in this step. The second step is the training step using the Ant Colony algorithm. The Ant Colony algorithm is the imitation of the ants' behavior of finding foods through the shortest path. The Ant Colony algorithm used in the training process is applied by instead of foraging for food the ants forage for classification rules.

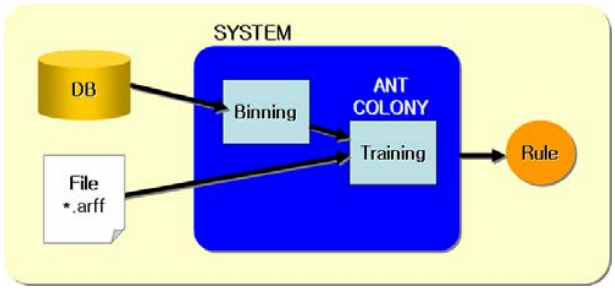


Fig. 4 Process of creating classification rules

1. Input Data Pre-Processing

The analysis system gets the input data from two sources. One is the database and the other is from a file. The data in the database contains the whole information about the samples/experiments and genes/probes expression values while the file contains only the information about the predefined classes which samples/experiments belong to. As the data in the database has a large domain for values, it needs to be converted into certain interval data to be dealt with by the rules. Therefore, the data from the database are rounded to the second places of decimals.

In order to increase the performance and make the data more useful, the binning process is needed. The binning process is the process of creating the bins. As the data are distributed so widely without any overlap it needs to be cut into intervals so that the distribution of the data would could be simplified and thus let the performance of the algorithm increase. This interval is called a bin. Fig. 5 briefly shows the binning process. From this step the analysis system gets the data that are ready to be used for training.

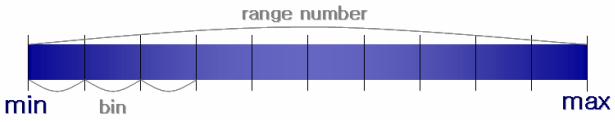


Fig. 5 Binning process

2. Training Process using Ant Colony Algorithm

In nature ants are seen creating "highways" to and from their food, often using the shortest route. Each ant lays down an amount of pheromone and the other ants are attracted to the strongest scent. As a result, ants tend to converge to the shortest path. This is because a shorter path is faster to transverse, so if an equal amount of ants follow the long path and the sort path, the ants that follow the short path will make more trips to the food and back to the colony. If the ants make more trips when following the shorter path, then they will deposit more pheromone over a given distance when compared to the longer

path. This is a type of positive feedback and the ants following the longer path will be more likely to change to follow the shorter path, where scent from the pheromone is stronger.

Ant Colony algorithm used in the analysis system takes the idea from the Ant Colony paradigm. Instead of foraging for food the ants in the Ant Colony algorithm forage for classification rules, and the path they take correspond to a conjunction of attribute-value pairs. A high-level pseudo-code of the Ant Colony algorithm is shown in Algorithm 1.

```
TrainSet = {all training cases};
DiscoveredRuleList = [];
REPEAT
   Initialize all trails with the same amount of
pheromone;
   REPEAT
      An ant incrementally constructs a
         classification rule;
      Prune the just-constructed rule;
      Update the pheromone of the all trails;
   UNTIL (stopping criteria)
   Choose best rule out of all rules constructed by
      all ants;
   Add the best rule to DiscoveredRuleList;
   TrainSet = TrainSet – {cases correctly covered
               by best rule};
UNTIL (stopping criteria)
```

Algorithm 1 Ant Colony Algorithm for Classification

Ant Colony Algorithm starts by initializing the training set to the set of all training cases, for example, gene and the class where gene belongs, and initializing the discovered rule list to an empty list. Each iteration of the loop discovers one classification rule. The first step of the loop is to initialize all trails with the same amount of pheromone, which means that all terms have the same probability of being chosen by the current ant to incrementally construct the current classification rule. The second step consists of pruning the just-constructed rule terms that do not improve the predictive accuracy of the rule. The third step is updating the pheromone of all trails by increasing the pheromone in the trail followed by the ant, proportionally to the rule's quality. In other words, the higher the quality of the rules, the higher the increase in the pheromone of the terms occurring in the rule antecedent. Once inner loop is over, the algorithm chooses the highest-quality rule out of all the rules constructed by all the ants in the inner loop, and it adds the chosen rule to the discovered rule list. Next, the algorithm removes from the training set all cases correctly covered by the rule. Hence, the next iteration of the outer loop starts with a smaller training set, consisting only of cases which have not been correctly covered by any rule discovered in previous iterations. The outer loop is performed until some stopping criterion is satisfied [7].

3. Classifying by using the Discovered Rules

The rules discovered from the previous step are used to predict the classes for the test data set. The process involves two steps. First, the analysis system gets the test data set from the database and applies the rules to decide which class the samples in the test set belong to. Second, compare the predicted test set with the answer set that contains the actual class of the test set. The accuracy of the rules is then calculated.

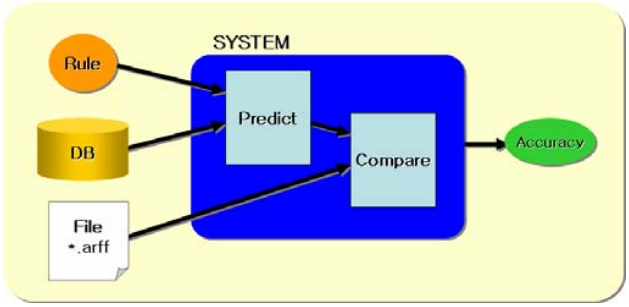


Fig. 6 Process of classifying by using the rules

V. Implementation and Experimental Results

The Ant Colony based classification analysis system is implemented with Java. It uses as input the two-dimensional bio-chip data. It also requires a file used for training with an extension of *.arff . The Ant Colony based classification analysis system works with the following setup.

System environment	
Java environments	J2SDK 1.4.02
Database	Oracle 9i

The file used for training and contains the answer class for accuracy computation needs to also contain the information about the database addresses, range value, id and password. The range value is used for the binning process. The bin size is decided by dividing between of the maximum and minimum expression value in the database according to the range value. The file also contains the information about the samples/experiments with its class. For example, N00287 -1 means that sample N00287 belongs to class -1. The sample data should be sorted in increasing order. The system only classifies the sample/experiment data.

We used the AB 1700 mouse chip (1- dye) which was provided from Macrogen Inc. as input data for the Ant Colony based classification analysis system for training. The chip data, 1,000 data of different probe_name from each 24 normalization_ID, was imported to the database and only the class of each sample is written in the file with the access information of database. Table I refers to the data in the database, and Fig. 7 refers to the class information in file.

TABLE I
AB 1700 Mouse Chip Data in Database

Normalization_ID	Probe_name	Expression_value
N000289	341276	-.68560427
N000289	341287	-1.4155682
…	…	…
N000304	316843	-.98317541

```
@relation AB 1700 mouse chip
@address 203.255.177.137:1521:rome
@id bio
@pw ant
@range 10

@class
N000287 -1
N000288 1
N000289 1
   …
N000310 1
```

Fig. 7 Database and class information in file

According to the results shown in Table II, the accuracy of 4th cross validation produced the most reliable rules. Therefore the rule in cross validation 4th is chosen as the rule for classification. Fig. 8 shows the chosen rule. The accuracy of the rule is 54.2%. Fig. 9 shows the interface of the Ant Colony based classification analysis system.

VI. Future Work and Conclusion

Our developed Ant Colony based classification analysis system classifies the sample/experiment DNA chip data to the predefined classes. It shows that the ecosystem imitating algorithm, ant colony algorithm, can be effectively applied to the analysis of bio-chip data. Because the bio-chip data is extremely large and has continuous values, we use a binning process to filter the data to get high performance in discovering the classification rules.

We plan to further work on enhancing the accuracy of the rules and apply more optimizations to enhance the speed of the algorithm when used for classification on Bio-chip data.

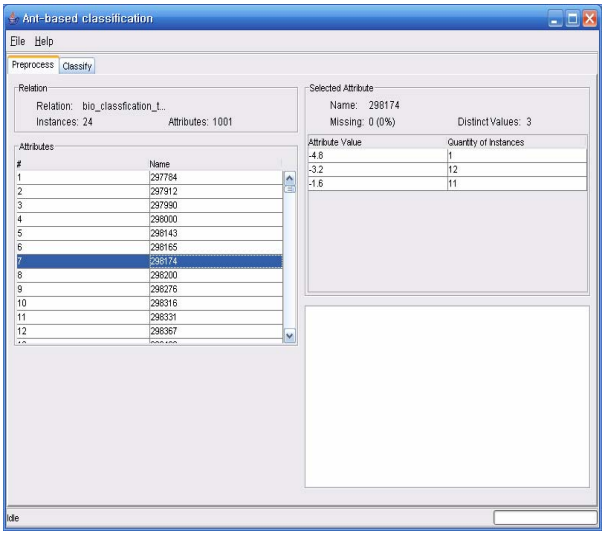
TABLE II
Result of the Ant Colony based Classification Analysis System

Cross Validation #	Training set	Test set	Rule #	Accuracy (training set)	Accuracy (test set)	Time taken
1	21	3	3	95.2%	67%	1244s
2	21	3	3	95.2%	33%	3554s
3	22	2	3	90%	50%	2551s
4	22	2	3	95.4%	50%	1907s
5	21	3	3	95%	0%	3103s
6	22	2	3	90%	50%	1728s
7	22	2	3	95%	0%	3482s
8	22	2	3	90%	100%	2774s
9	22	2	3	95%	0%	2838s
10	21	3	3	95%	33%	1712s

```
IF 298428 = '-1.6' AND 298941 = '-1.6' AND
302279 = '3.2' AND 304455 = '-1.6' THEN '1'
IF 306410 = '-3.2' AND 307847 = '0.0' THEN
'-1'

default 1
```

Fig. 8 Classification rule



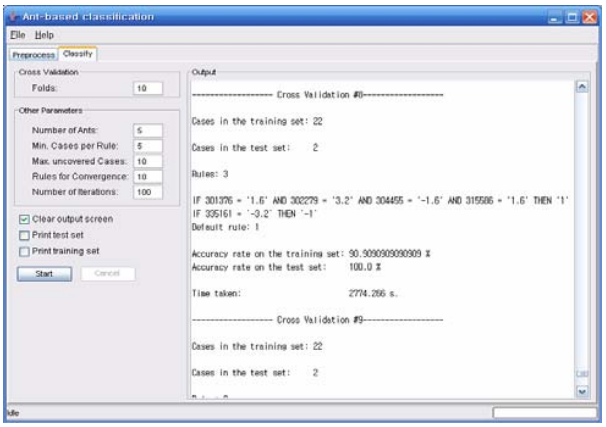


Fig. 9 Interface of Ant Colony Based Classification System

References

[1] Barbara Comes, Arpad Kelemen. Probabilistic neural network classification for microarraydata. IEEE, 2003.
[2] J. Bala, J. Huang , H. Vafaiem K. DeJong and H. Wechsler. Hybrid Learning Using Genetic Algorithms and Decision Trees for Pattern Classification. IJCAI conference, 1995.
[3] Jiawei Han, Micheline Kamber. Data Mining Concepts and Techniques. Morgan Kaufmann, 2001.
[4] Lizhuang Zhao, Mohammed J. Zaki, TriCluster: An Effecitive Algorithm for Mining Coferent Clusters in 3D Microarray Data. SIGMOD , Baltimore, Maryland, USA, June(2005).
[5] Marco Dorigo, Vittorio Maniezzo, and Alberto Colorni. The Ant System: Optimization by a colony of cooperating agents. IEEE Transactions on Systems, Vol.26, No.1, 1996.
[6] Marco Dorigo, and Luca Maria Gambardella. Ant colonies for the traveling salesman problem. BioSystems, 1997.
[7] Nicholas Holden and Alex A. Freitas. Web Page Classification with an Ant Colony Algorithm. Parallel Problem Solving from Nature - PPSN VIII, LNCS 3242, pages 1092-1102. Springer-Verlag, September 2004.
[8] Sorin Draghici. Data Analysis Tools for DNA Microarrays. Chapman & Hall, 2003.
[9] Wikipedia, http://www.wikipedia.org/
[10] Yi-Shiou Chen and Tah-Hsiung Chu, A Neural Network Classification Tree, IEEE, 1995.