

A Fuzzy-ACO Method for Detect Breast Cancer

Amin Einipour (Corresponding author)

Department of Computer, Andimeshk Branch, Islamic Azad University, Andimeshk, Iran

Daneshgah Street, Andimeshk, Khozestan, Iran

Tel: 98-642-424-0821 E-mail: a.einipour@gmail.com

Received: June 16, 2011 Accepted: July 7, 2011 doi:10.5539/gjhs.v3n2p195

Abstract

Data mining usually means the methodologies and tools for the efficient new knowledge discovery from databases. It is also a form of knowledge discovery essential for solving problems in a specific domain. For instance, the data mining approaches are applied in the filed of medical diagnosis recently. A major class of problems in medical science involves the diagnosis of disease, based upon various tests performed upon the patient. When several tests are involved, the ultimate diagnosis may be difficult to obtain, even for a medical expert. This has given rise, over the past few decades, to computerized diagnostic tools, intended to aid the physician in making sense out of the welter of data. Specifically, where breast cancer is concerned, the treating physician is interested in ascertaining whether the patient under examination exhibits the symptoms of a benign case, or whether her case is a malignant one. In this paper, we have focused on breast cancer diagnosis by combination of fuzzy systems and evolutionary algorithms. Fuzzy rules are desirable because of their interpretability by human experts. Ant colony algorithm is employed as evolutionary algorithm to optimize the obtained set of fuzzy rules. Results on breast cancer diagnosis data set from UCI machine learning repository show that the proposed approach would be capable of classifying cancer instances with high accuracy rate in addition to adequate interpretability of extracted rules.

Keywords: Data mining, Breast cancer detection, Fuzzy rule extraction, ACO algorithm

1. Introduction

Data mining usually means the methodologies and tools for the efficient new knowledge discovery from databases. It is also a form of knowledge discovery essential for solving problems in a specific domain. For instance, the data mining approaches are applied in the filed of medical care recently. A major class of problems in medical science involves the diagnosis of disease, based upon various tests performed upon the patient. When several tests are involved, the ultimate diagnosis may be difficult to obtain, even for a medical expert. This has given rise, over the past few decades, to computerized diagnostic tools, intended to aid the physician in making sense out of the welter of data (Abbass, 2002, 25, P.265-281 & Chou, 2004, P.133-142 & Ohmann, 1996, P.23-36).

A prime target for such computerized tools is in the domain of cancer diagnosis. Specifically, where breast cancer is concerned, the treating physician is interested in ascertaining whether the patient under examination exhibits the symptoms of a benign case, or whether her case is a malignant one (Pendharkar, 1999, P.223-232 & Bellazzi, 1998, P.5-28 & Fogel, 1998, P.317 & Carlos, 1999, P.131-155).

Earlier studies, the statistical related techniques were most commonly used data mining approaches to construct classification models. However, as the breast cancer classification problem is highly nonlinear in nature, it is hard to develop a comprehensive model taking into account all the independent variables using conventional statistical modeling techniques. Furthermore, traditional ad hoc mixtures of statistical techniques and data management tools are no longer adequate for analyzing the vast collection of data. For the needs of improving the prediction accuracy in breast cancer diagnosis, more and more researchers have tried to apply artificial intelligence related approaches for breast cancer prediction (Subhash, 2003, P.25-34 & Ta-Cheng, 2005, P.1-8).

A good computerized diagnostic tool should possess two characteristics. First, the tool must attain the highest possible performance, i.e. diagnose the presented cases correctly as being either *benign* or *malignant*. Moreover, it would be highly desirable to be in possession of a so-called degree of confidence: the system not only provides a binary diagnosis (benign or malignant), but also outputs a numeric value that represents the degree to which the

system is confident about its response. Second, it would be highly beneficial for such a diagnostic system to be human-friendly, exhibiting so-called interpretability. This means that the physician is not faced with a black box that simply spouts answers (Albeit correct) with no explanation; rather, we would like for the system to provide some insight as to how it derives its outputs.

Some experimental studies reported that success of artificial neural networks in breast cancer prediction, but there is a major drawback in building and using a model in which the user cannot readily comprehend the final rules that neural networks models acquire. In other words, the results of training a neural network are internal weights distributed throughout the network. These weights provide no more insight into why the solution is valid than asking many human experts why a particular decision is the right decision. For example, the weights are not readily understandable although, increasingly, sophisticated techniques for probing into neural networks help provide some explanation. It is also some recently studies used genetic and K-NN approaches to breast cancer diagnosis which interpretability of these approaches higher than of neural networks (Carlos, 1999, P.131-155 & Subhash, 2003, P.25-34 & Ta-Cheng, 2005, P.1-8).

In this paper we combine two methodologies—fuzzy systems and ACO algorithm—so as to automatically produce systems for breast cancer diagnosis. The major advantage of fuzzy systems is that they favor interpretability; however, finding good fuzzy systems can be quite an arduous task. This is where Ant-Colony algorithms step in, enabling the automatic production of fuzzy systems, based on a database of training cases. There are several recent examples of the application of fuzzy systems and evolutionary algorithms in the medical domain which combine both methodologies in a hybrid way, but our fuzzy-ACO approach produces systems exhibiting two prime characteristics: first, they attain high classification performance, with the possibility of attributing a confidence measure to the output diagnosis; second, the resulting systems involve a few simple rules, and are therefore (human-) interpretable.

To the best of our knowledge the use of Ant Colony algorithms as a method for discovering classification rules, in the context of data mining, is a research area still unexplored by other researchers. Actually, the only Ant Colony algorithm developed for data mining that we are aware of is an algorithm for clustering (Monmarche, 1999, P. 23-26 & Rafael, 2002, P.190-208) which is, of course, a data mining task very different from the classification task addressed in this paper. Also, Cordón et al. have proposed another kind of Ant Colony Optimization application that learns fuzzy control rules, but it is outside the scope of data mining (Cordon, 2000, P.13-21).

We believe the development of Ant Colony algorithms for data mining is a promising research area, due to the following rationale. An Ant Colony system involves simple agents (ants) that cooperate with one another to achieve an emergent, unified behavior for the system as a whole, producing a robust system capable of finding high-quality solutions for problems with a large search space. In the context of rule discovery, an Ant Colony system has the ability to perform a flexible, robust search for a good combination of logical conditions involving values of the predictor attributes.

This paper is organized as follows. In Section two we describe the Wisconsin breast cancer diagnosis (WBCD) problem, which is the focus of our interest in this paper. The third section describes the fuzzy systems based classification. Section 4 describes our particular fuzzy-ACO approach to the WBCD problem. The fifth section reports on computational results evaluating the performance of the proposed system. Finally, the sixth section concludes the chapter.

2. The Wisconsin breast cancer diagnosis problem

A major class of problems in medical science involves the diagnosis of disease, based upon various tests performed upon the patient. When several tests are involved, the ultimate diagnosis may be difficult to obtain, even for a medical expert. This has given rise, over the past few decades, to computerized diagnostic tools, intended to aid the physician in making sense out of the welter of data. In this section we describe the Wisconsin breast cancer diagnosis problem, which is the focus of our interest in this paper.

Breast cancer is the most common cancer among women, excluding skin cancer. The presence of a breast mass is an alert sign, but it does not always indicate a malignant cancer. Fine needle aspiration (FNA) of breast masses is a cost-effective, non-traumatic, and mostly non-invasive diagnostic test that obtains information needed to evaluate malignancy.

The Wisconsin breast cancer diagnosis (WBCD) database is the result of the efforts made at the University of Wisconsin Hospital for accurately diagnosing breast masses based solely on an FNA test. Nine visually assessed

characteristics of an FNA sample considered relevant for diagnosis were identified, and assigned an integer value between 1 and 10. The measured variables are as follows:

1. Clump thickness (V1);
2. Uniformity of cell size (V2);
3. Uniformity of cell shape (V3);
4. Marginal adhesion (V4);
5. Single epithelial cell size (V 5);
6. Bare nuclei (V 6);
7. Bland chromatin (V 7);
8. Normal nucleoli (V 8);
9. Mitosis (V 9).

The diagnostics in the WBCD database were furnished by specialists in the field. The database itself consists of 699 cases, with each entry representing the classification for a certain ensemble of measured values (see table 1). Note that the diagnostics do not provide any information about the degree of benignity or malignancy.

3. Fuzzy system based classification

Fuzzy If-Then rules for pattern classification problem with c pattern and n feature are written in the following form:

$$\text{Rule } R_j : \text{If } x_1 \text{ is } A_{j1} \text{ and } \dots x_n \text{ is } A_{jn} \text{ Then Class } C_j, j=1..N \quad (1)$$

Where R_j is the label of the j -th fuzzy rule, $x = (x_1, \dots, x_n)$ is an n -dimensional pattern vector, A_{ji} is a fuzzy antecedent term, such as *small*, *medium*, *large* which given for the i -th attribute, C_j is a consequent class, and N is number of fuzzy If-Then rules. When K fuzzy terms are given for each of the n attributes, we have K^n linguistic rules of the form (1-1).

Fuzzy If-Then rules with certainty factor also can be used for pattern classification problem:

$$\text{Rule } R_j : \text{If } x_1 \text{ is } A_{j1} \text{ and } \dots x_n \text{ is } A_{jn} \text{ Then Class } C_j \text{ with } CF_j, j=1..N \quad (2)$$

Where CF_j is a rule weight (i.e., certainty factor) of the j -th fuzzy rule R_j . The rule weight CF_j , which is a real number in the unit interval $[0,1]$, denotes the strength of the fuzzy rule R_j . Fuzzy rules with the maximum rule weight 1.0 have the largest effect on the classification of new patterns. On the other hand, fuzzy rules with the minimum rule weight 0.0 have no effect on the classification of new patterns.

4. Proposed fuzzy-acoapproach

An artificial Ant Colony System (ACS) is an agent-based system which simulates the natural behavior of ants and develops mechanisms of cooperation and learning. ACS was proposed by Dorigo et al. (Dorigo, 1996, P.1-13) as a new heuristic to solve combinatorial-optimization problems. This new heuristic, called Ant Colony Optimization (ACO), has been shown to be both robust and versatile – in the sense that it can be applied to a range of different combinatorial optimization problems. In addition, ACO is a population-based heuristic. This is advantageous because it allows the system to use a mechanism of positive feedback between agents as a search mechanism. Recently there has been a growing interest in developing rule discovery algorithms based on other kinds of population-based heuristics – mainly evolutionary algorithms (Rafael, 2002, P.190-208 & Dorigo, 1999, P.137-172).

In the next subsection we provide a brief overview of Artificial Ant Colony, and then describe the proposed algorithm which named FUZZY-ACO.

4.1 Artificial ant colony

An Artificial Ant Colony for pattern classification problem as followed:

Number of ants in Ant Colony: number of fuzzy rules for classification.

Feature of Ants: features and attributes in breast cancer domain.

The amount of pheromone associated with each Ant: The amount of evaluation function associated with each rule.

Pheromone updating: optimize the current classifier fuzzy rule-base.

Current Ant Colony: current rule-base.

Modify Current Ant Colony: modify current rule-base.

New Ant Colony: new current rule-base.

Calculate the fitness of new Ant Colony: Calculate the evaluate function for new rule base.

Admission of new Ant Colony with specific probability and improvement of fitness: Admission of new rule-base with specific probability and improvement of evaluation function.

4.2 Proposed FUZZY-ACO algorithm

Outline of the proposed algorithm based ACO is as follows:

Step1: Preprocessing

Normalization

Fuzzification

Step2: Generate an initial set of fuzzy if-then rules. (Initialization)

Step3: Evaluate cost of current rule-base using evaluation function.

Step4: Modify current rule-base using modify one of rules randomly and generate new rule-base.

Step5: Evaluate cost of new rule-base using evaluation function.

Step6: Admission of new rule-base with specific probability and improvement of evaluation function, then replace a current rule-base with new rule-base, and save best evaluation function and rule-base.

Step7: For specific iteration repeat Step4, Step5, Step6

Step8: Return best rule-base

In next subsections describes details of these steps.

5. Experimental results

This approach is implemented by using C++ programming language. All value of input feature normalized to [0, 1]. We use 10-CV technique for evaluate proposed approach. In this technique, WBCD data set divided to 10 parties, nine parties for train set and one party for test set. This process runs for ten times. Results of these run is indexed in table 2.

Average of train set accuracy and test set accuracy is 98.21 and 96.99, in sequence. Average of rules length is 1.235 which denotes that this system has a high interpretability.

Proposed approach is compared with some algorithms, such as C4.5, *k*-NN, Naïve Bayes, SVM, MLP. Result of comparison is indexed in table 3.

According to Table 4, train set accuracy of FUZZY-ACO algorithm (proposed algorithm) only less than SVM, and also test set accuracy of FUZZY-ACO algorithm better than other algorithms except 5-NN which accuracy is equal. However, proposed algorithm has a main advantage, which is high interpretability.

6. Conclusion

In this paper, we focused on breast cancer diagnosis by combination of fuzzy systems and ACO algorithm. The proposed method performs the classification task and extracts required knowledge using fuzzy rule based systems which consists of fuzzy if-then rules. Ant colony algorithm is employed to optimize the obtained set of fuzzy rules. The proposed system has two main features of data mining techniques which are high reliability and adequate interpretability, and is comparable with several well-known algorithms. Results on *breast cancer diagnosis* data set from UCI machine learning repository show that the proposed FUZZY-ACO (Ant Colony based Fuzzy System) would be capable of classifying cancer instances with high accuracy rate in addition to adequate interpretability of extracted rules.

References

- Abbass, H. A. (2002). *Artif Intell Med. An evolutionary artificial neural networks approach for breast cancer diagnosis*, 25, 265-281.
- Bellazzi, R., Ironi, L., Guglielmann, R., & Stefanelli, M. (1998). *Artif Intell Med. Qualitative models and fuzzy systems: an integrated approach for learning from data*, 14 (1-2), 5-28.

Carlos Andres Pena-Reyes & Moshe Sipper. (1999). *Artif Intell Med. A fuzzy-genetic approach to breast cancer diagnosis*, 17, 131-155.

Chou, S-M., Lee, T-S., Shao, Y. E., & Chen, I-Fei. (2004). *Expert system with applications. Mining the breast cancer pattern using artificial neural networks and multivariate adaptive regression splines*, 27, 133-142.

Cordon, O., Cassillas, J., & Herrer, F. (2000). *Second International Workshop on Ant Algorithms. Learnin Fuzzy Rules using Ant Colony Optimization*, 13-21.

Dorigo, M., Colorni, A., & Maniezzo, V. (1996). *IEEE Transactions on Systems, Man, and Cybernetics-Part B. The Ant System: Optimization by a colony of cooperating agents*, 26 (1), 1-13.

Dorigo, M., Di Caro, G., & Gambardella, L. M. (1999). *Artificial Life. Ant algorithms for discrete optimization*, 5 (2), 137-172.

Fogel, D. B., Wasson III, E. C., Boughton, E. M. & Porto, V. W. (1998). *Artif Intell Med. Evolving artificial neural networks for screening features from mammograms*", 14 (3), 317.

Monmarche, N. & Freitas, A. A. (1999). *AAAI Workshop. On data clustering with artificial ants*, 23-26.

Ohmann, C., Moustakis, V., Yang, Q., & Lang, K. (1996). *Artif Intell Med. Evaluation of automatic knowledge acquisition techniques in the diagnosis of abdominal pain*, 8, 23-36.

Pendharkar, P. C., Rodger, J. A., Yaverbaum, X. X., Herman, N., & Benner, M. (1999). *Expert Systems with Application. Association, statistical mathematical and neural approaches for mining breast cancer patterns*, 17, 223-232.

Rafael, S. Parpinelli, Heitor, S. Lopes, & Alex, A. Freitas. (2002). *An Ant Colony Algorithm for Classification Rule Discovery. Idea Group Publishing*, 190-208.

Subhash, C., Sikha Bagui, Kuhu, Pal., Nikhil, R. Pal. (2003). *Pattern Recog. Breast cancer detection using rank nearest neighbor classification rules*, 36, 25-34.

Ta-Cheng, Chen & Tung-Chou, Hsu. (2005). *Expert System with Applications. A GAs based approach for mining breast cancer pattern*, 1-8.

Table 1. WBCD Database in UCI repository

case	V ₁	V ₂	V ₃	V ₄	V ₅	V ₆	V ₇	V ₈	V ₉	Diagnosis(Class)
1	5	1	1	4	2	1	3	2	1	Benign
2	5	4	4	5	7	10	3	2	1	Benign
699	4	8	8	5	4	5	10	4	1	Malignant

Table 2. Results of proposed approach

Run	1	2	3	4	5	6	7	8	9	10
Train	614 /629	614 /629	619 /629	613 /629	622 /629	624 /629	615 /629	619 /629	620 /629	616/630
Test	68 /70	69 /70	67 /70	69 /70	67 /70	69 /70	68 /70	68 /70	67 /70	66/69
Rules	20	20	20	20	20	20	20	20	20	20
Length	1.3	1.4	1.3	1.1	1.55	1.05	1.2	1.4	1.3	0.75

Table 3. Results Comparison of some algorithms

Algorithm	MLP	SVM	Naïve Bayes	5-NN	C4.5	FUZZY-ACO
test accuracy	95.42	96.56	96.13	96.99	94.56	96.99
Train accuracy	97.45	98.41	97.57	98.09	97.29	98.21