The Accuracies of ANFIS and Genetic Algorithm with Tournament Selection on Classifying Hepatitis Data

W FA Siahaan¹, O S Sitompul² and Z Situmorang³

¹Graduate Program of Computer Science, Universitas Sumatera Utara, Medan, Indonesia
²Department of Information Technology, Universitas Sumatera Utara, Medan, Indonesia
³Department of Computer Science, Universitas Katolik Santo Thomas, Medan, Indonesia

Abstract - Classification is an important technique in data mining to create patterns and data modeling. Hepatitis is a disease that is dangerous to humans as this disease affects the human liver which is a vital organ. Early detection of hepatitis by diagnosing its occurrence on patients is needed to give them help, care, and treatment. Therefore, a classification technique is required to create patterns and models of data. To achieve accuracy in classifying hepatitis there would be requirements to separate relevant from irrelevant features. By only considering relevance features in classifying hepatitis, it is expected that the classification process will be faster, easier, and more accurate. In this research, the ANFIS algorithm and genetic algorithm are used in classifying the diagnosis of hepatitis based on a dataset from the UCI data mining repository. Results obtained showed that the genetic algorithm gave a slightly higher accuracy of 98.73% compared to 86.67% results obtained using ANFIS.

1. Introduction

Data mining techniques are always needed in various fields such as production, manufacturing, telecommunications, engineering, and medical. Classification is part of data mining to find patterns and modeling in data sets. Algorithms that are included in data mining contain fuzzy techniques, artificial neural network, ANFIS, etc. [6]. An important approach that could be incorporated in the classification is a selection of a feature that are relevant towards determining classification result and eliminating irrelevant features from data. Being considered as a dangerous disease for humans, hepatitis could cause liver damage in the human body. In humans, the cause of hepatitis is not only due to unhealthy lifestyles but also by viruses or bacteria. Early diagnosis is needed to provide help, care and prevention to prevent death of patients.

The selection features are meant to select the relevant features and eliminating irrelevant features from data. Some studies that have examined the ANFIS algorithm and genetic algorithm conducted their research using the Adaptive Neuro Fuzzy Inferences System (ANFIS) algorithm with data classification and reducing dimensions with Principal Component System (PCA) for diagnosing the risk of heart disease. Research data from UCI Machine Learning (UCI), the University of California with 303 samples. Training data = 200 samples and testing data = 103 samples. Research results obtained accuracy = 93.2%. [2]. The research uses Bayes, support vector system (SVM), and genetic algorithms with classification and extraction of features. Data from Yuhungdig Hospital, Yantai, Shandong Province, China for hepatitis with HBV virus for liver carcinoma patients after radiotherapy. Hepatitis B disease features with 30 features. Classification and extraction of features are divided into 10 features. The research results obtained accuracy with Bayes algorithm = 80% and...
Support vector system (SVM) = 80.56%. [7]. ANFIS is a combination of neural networks and fuzzy inference system with hybrid learning method where the gradient descent learning algorithm and least square estimation for error values. Gradient descent and least square estimation of learning algorithms on ANFIS. The weaknesses of ANFIS algorithms: the numbers of rule base inputs and the high computational costs in ANFIS. [3] Classification of features is needed to make it faster and easier and more accurate on ANFIS algorithms. Genetic algorithms are natural evolutionary methodologies. Genetic algorithms are used to classify relevant features so that they reduce and rational heuristic probabilities of genetic algorithms. Genetic algorithms are needed to classify relevant features and reduce irrelevant features as well as to increase accuracy in the ANFIS algorithm. Based on research that has examined the ANFIS algorithm and genetic algorithms are interesting things to study.

2. Algorithm

2.1. ANFIS (Adaptive Neuro Fuzzy Inferences System)

ANFIS is a combination of neural networks and fuzzy. The adaptive system in the ANFIS algorithm which includes hybrid learning to set the rules used for learning (gradient descent and least square) of the data set to make predictions easier [1]. Learning algorithm in ANFIS gradient descent and least square estimation for input features with outputs of membership function and error values in the ANFIS algorithm. The advantages of the ANFIS algorithm: adaptive and fast learning so that it is fast, easy, predictive, and modeling on input features. Classification of features is needed to produce the performance of the ANFIS algorithm. The architecture of the ANFIS algorithm between one node is connected to another. ANFIS architecture consists of 5 layers:

Each node is an input layer is an adaptive node function and output membership function. With the equation:

\[ O_i = \mu A \cdot x \]

\[ \mu (A) = \frac{1}{1 + \left( \frac{x-c}{a} \right)^2 b} \]  

where:  \( x= input \) fitur \( \{a, b, D\} \).

\( \mu (A) \) = membership function.

Each node is a fixed node or non-adaptive symbolized by \( \pi \). Multiplication of membership function in layer 2. With the equation:

\[ O_i^2 = \mu (A) (x) \cdot \mu (B) (y) \cdot i = 1,2, \ldots \]
- Each node in layer 3, normalization with the ratio of the firing strength. Output layer at layer 3. With the following equation:

\[ O_i^3 = \bar{w}_1 = \frac{w_i}{w_1 + w_2} \quad \text{for } i = 1,2,\ldots \ldots \quad (3) \]

- Each node i in the layer is adaptive. With the following equation:

\[ O_{k,t} = \bar{w}_1 \cdot f_1 = \bar{w}_1 (p_1 x + q_1 y + r_1) \]

where : \{p, q, r \} = features

- Computation of total outputs added together. The sum of all incoming signals. The symbol \( \sum \):

\[ O_i^5 = \text{overall output} = \sum_d \bar{w}_i f_i = \sum_{i} \frac{w_i f_i}{\sum w_i} \quad (5) \]

2.2. Genetic Algorithms

The genetic algorithm is a natural evolutionary methodology from Darwin’s theory. As for genetic algorithms, it is needed to get the best solution to problems.[3] The process in genetic algorithms is the search of the problems the best solution seen through stages in genetic algorithms.

Process in genetic algorithms: Encoding genes of hepatitis data string sets into chromosomes. Chromosomes are collected into populations. Population installation of several chromosome strings contained in classified individuals. The selection used is tournament selection. Then crossover and mutation are carried out to produce fitness value which is the best solution to the problem. The stages of genetic algorithms:

- Encoding: Classify relevant features and eliminate irrelevant features. Encoding relevant features into binary form 1 and 0 to become individuals.
- Population: Binary strings become individuals. Individuals are collected into populations.
- Selection: The selection in research papers using tournament selection. Tournament selection is to compare the value of fitness and the best fitness value is taken. Fitness function to evaluate the best individual and the best fitness value.
- Crossover: Crossing two parents so that it becomes offspring[12].
- Mutation: change individual gene bits 1 to 0 and vice versa.

3. Methodology

3.1. The Methodology of ANFIS Algorithm and genetic algorithm

The classification of relevant features and irrelevant features for diagnosing hepatitis. The methodology of the ANFIS algorithm and genetic algorithm is represented in Figure 2(a) and (b).

![Figure 2(a) and (b). Block Diagram of the ANFIS Algorithm and Genetic Algorithm.](image-url)
The information on the block diagram of the ANFIS algorithm and genetic algorithm:

- Data set from UCI machine learning. Total of data set = 155 samples.
- Data cleaning is performed. Cleaning data is to identify incomplete data and eliminate incomplete data. Incomplete data such as blank signs and question marks on hepatitis data. A good data set to measure the performance of ANFIS algorithms and genetic algorithms in diagnosing hepatitis. Hepatitis cleaning data set = 79 samples.
- The data set is divided into two namely: training data and testing data.
- After the training data and testing data, an error value is obtained for the ANFIS algorithm.
- Training data and testing data obtain the training data modeling and testing data.
- A genetic algorithm is needed to improve the performance of the ANFIS algorithm.
- After optimization with genetic algorithms, accuracy is obtained.

3.2. The classification of data sets

Total of data set = 155 samples with 19 features for hepatitis. Features in the hepatitis data set: Age, Sex, Stroid, Antivirals, Malaise, Anorexia, Liver big, Liver firm, Spleen Palpable, Spiders, Ascites, Varices, Bilirubin, Alk Phosphate, SGOT, Albumin, Protame, and Histology. The total of cleaning data is 79 samples, in Table1.

### Table 1. The features of hepatitis disease.

| No | Name of Features       | Domain Value        |
|----|------------------------|---------------------|
| 1. | Age                    | 10,20,30,40,50,60,70,80 |
| 2. | Sex (Gender)           | Male, Female        |
| 3. | Stroid                 | Yes, No             |
| 4. | Antivirals             | Yes, No             |
| 5. | Fatigue                | Yes, No             |
| 6. | Malaise                | Yes, No             |
| 7. | Anorexia               | Yes, No             |
| 8. | Liver Big              | Yes, No             |
| 9. | Liver Firm             | Yes, No             |
| 10.| Spleen Palpable        | Yes, No             |
| 11.| Spiders                | Yes, No             |
| 12.| Ascites                | Yes, No             |
| 13.| Varices                | Yes, No             |
| 14.| Bilirubin              | 0.39, 0.80, 1.20, 2.3, 4 |
| 15.| Alk Phosphate (ALP)    | 33, 80, 120, 160, 200, 250 |
| 16.| SGOT                   | 13, 100, 200, 300, 400, 500, 600 |
| 17.| Albumin                | 2, 1, 3, 8, 4, 5, 6 |
| 18.| Protame                | 10, 20, 30, 40, 50, 60, 70, 80, 90 |
| 19.| Histology              | Yes, No             |

The importance of selecting and classifying features that are needed and features that are not needed in order to obtain accuracy in diagnosis of hepatitis. The references to the researcher opinion for the classification of features that are relevant and irrelevant features. Relevant features: Bilirubin, ALP, SGOT, Albumin [9].
4. Results and discussion

4.1. The results and discussion of ANFIS Algorithm

The features in the ANFIS algorithm use R-Studio: Age, sex, Bilirubin, SGOT, Alkphosphate (ALP), and Albumin. The data set is divided into 2: train data = 50% and test data = 50%. Train data = 40 and test data = 39 for hepatitis. Training data for ANFIS Algorithm can be seen in Table 2:

| No | Age | Sex | Bilirubin | ALP | SGOT | Albumin |
|----|-----|-----|-----------|-----|------|---------|
| 1  | 34  | 1   | 0.9       | 95  | 28   | 4       |
| 2  | 39  | 1   | 1.3       | 78  | 30   | 4.4     |
| 3  | 32  | 1   | 1         | 59  | 249  | 3.7     |
| 4  | 41  | 1   | 0.9       | 81  | 60   | 3.9     |
| 5  | 30  | 1   | 2.2       | 57  | 144  | 4.9     |
| 6  | 38  | 1   | 2         | 72  | 89   | 2.9     |
| 7  | 40  | 1   | 0.6       | 62  | 166  | 4       |
| 8  | 38  | 1   | 0.7       | 53  | 42   | 4.1     |
| 9  | 38  | 1   | 0.7       | 70  | 28   | 4.2     |
| 10 | 22  | 2   | 0.9       | 48  | 20   | 4.2     |
| 11 | 27  | 1   | 1.2       | 133 | 98   | 4.1     |
| 12 | 31  | 1   | 1         | 85  | 20   | 4       |
| 13 | 42  | 1   | 0.9       | 60  | 63   | 4.7     |
| 14 | 25  | 2   | 0.4       | 45  | 18   | 4.3     |
| 15 | 27  | 1   | 0.8       | 95  | 46   | 3.8     |
| 16 | 58  | 2   | 1.4       | 175 | 55   | 2.7     |
| 17 | 61  | 1   | 1.3       | 78  | 25   | 3.8     |
| 18 | 51  | 1   | 1         | 78  | 58   | 4.6     |
| 19 | 39  | 1   | 2.3       | 280 | 98   | 3.8     |
| 20 | 41  | 2   | 0.7       | 81  | 53   | 5       |
| 21 | 26  | 2   | 0.5       | 135 | 29   | 3.8     |
| 22 | 35  | 1   | 0.9       | 58  | 92   | 4.3     |
| 23 | 23  | 1   | 1.3       | 194 | 150  | 4.1     |
| 24 | 42  | 1   | 1         | 85  | 14   | 4       |
| 25 | 65  | 1   | 0.3       | 180 | 53   | 2.9     |
| 26 | 52  | 1   | 0.7       | 75  | 55   | 4       |
| 27 | 33  | 1   | 1         | 46  | 90   | 4.4     |
| 28 | 56  | 1   | 0.7       | 71  | 18   | 4.4     |
| 29 | 28  | 2   | 1.8       | 191 | 420  | 3.3     |
| 30 | 36  | 1   | 0.8       | 85  | 44   | 4.2     |
| 31 | 38  | 1   | 0.7       | 125 | 65   | 4.2     |
| 32 | 44  | 1   | 0.6       | 110 | 145  | 4.4     |
| 33 | 40  | 1   | 1.2       | 85  | 31   | 4       |
| 34 | 30  | 1   | 0.7       | 50  | 78   | 4.2     |
Testing data for the ANFIS algorithm can be seen in Table 3.

| No | Age | Sex | Bilirubin | ALP | SGOT | Albumin |
|----|-----|-----|-----------|-----|------|---------|
| 1  | 49  | 1   | 0.8       | 103 | 43   | 3,5     |
| 2  | 39  | 1   | 1         | 34  | 15   | 4       |
| 3  | 32  | 1   | 1         | 55  | 45   | 4,1     |
| 4  | 34  | 2   | 0.6       | 30  | 24   | 4       |
| 5  | 34  | 1   | 1         | 72  | 46   | 4,4     |
| 6  | 36  | 1   | 0.7       | 62  | 224  | 4,2     |
| 7  | 30  | 1   | 0.7       | 100 | 31   | 4       |
| 8  | 62  | 2   | 1.3       | 141 | 156  | 3,9     |
| 9  | 28  | 1   | 1.6       | 44  | 123  | 4       |
| 10 | 38  | 1   | 0.6       | 76  | 18   | 4,4     |
| 11 | 50  | 2   | 0.9       | 230 | 117  | 3,4     |
| 12 | 59  | 1   | 1,5       | 107 | 157  | 3,6     |
| 13 | 40  | 1   | 0.6       | 40  | 69   | 4,2     |
| 14 | 30  | 1   | 0.8       | 147 | 128  | 3,9     |
| 15 | 47  | 1   | 2         | 84  | 23   | 4,2     |
| 16 | 48  | 1   | 4,8       | 123 | 157  | 2,7     |
| 17 | 27  | 1   | 2,4       | 168 | 227  | 3       |
| 18 | 51  | 1   | 4,6       | 215 | 269  | 3,9     |
| 19 | 47  | 1   | 1,7       | 86  | 20   | 2,1     |
| 20 | 54  | 1   | 1         | 155 | 225  | 3,6     |
| 21 | 33  | 1   | 0,7       | 63  | 80   | 3       |
| 22 | 42  | 1   | 0,5       | 62  | 68   | 3,8     |
| 23 | 50  | 1   | 1         | 139 | 81   | 3,9     |
| 24 | 20  | 1   | 1         | 160 | 118  | 2,9     |
| 25 | 50  | 1   | 1         | 85  | 75   | 4       |
| 26 | 34  | 2   | 0,7       | 70  | 24   | 4,1     |
| 27 | 50  | 1   | 2,8       | 155 | 75   | 2,4     |
| 28 | 54  | 1   | 1,2       | 85  | 92   | 3,1     |
| 29 | 57  | 1   | 4,6       | 82  | 55   | 3,3     |
| 30 | 72  | 1   | 1         | 115 | 52   | 3,4     |
| 31 | 38  | 1   | 0,4       | 243 | 49   | 3,8     |
| 32 | 25 | 1 | 1,3 | 181 | 181 | 4,5 |
|----|----|---|-----|-----|-----|-----|
| 33 | 38 | 1 | 1.6 | 130 | 140 | 3.5 |
| 34 | 47 | 1 | 1   | 166 | 30  | 2.6 |
| 35 | 45 | 1 | 1.3 | 85  | 44  | 4.2 |
| 36 | 49 | 1 | 1.4 | 85  | 70  | 3.5 |
| 37 | 31 | 1 | 1.2 | 75  | 173 | 4.2 |
| 38 | 53 | 2 | 1.5 | 81  | 19  | 4.1 |
| 39 | 43 | 1 | 1.2 | 100 | 19  | 3.1 |

Training data source code for the ANFIS algorithm can be seen as in the following.

```r
residuals(anfis3)[1:5,]
[1]  0.011354394 -0.014242212 -0.010291178
  0.013337535  0.005391769
> summary(anfis3)
ANFIS network
Training Set: 
dim(x)= 121x2
 dim(y)= 121x1
Arquitecture:  2 ( 5x5 ) - 25 - 75 ( 75x1 ) - 1
Last training error:  0.01916307
Call: trainHybridJangOffLine(object = anfis3, 
     epochs = 10)
Statistics for Off-line training
     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
          0.01916 0.02139 0.02358 0.02489 0.02735 0.03741
> print(anfis3)
ANFIS network
Training Set: 
dim(x)= 121x2
 dim(y)= 121x1
Arquitecture:  2 ( 5x5 ) - 25 - 75 ( 75x1 ) - 1
```

Testing data source code for the ANFIS algorithm can be seen as follows.

```r
coef(anfis3)$premises[[input=1]][[mf=1]]
MembershipFunction: Normalized Gaussian
Membership Function
Number of parameters: 2
 mu  sigma
 9.984987  2.038630
Expression: expression(exp(-1/2*((x - mu)/sigma)^2))
> coef(anfis3)$consequents[1:2,]
[1]  0.005011691  0.005011691
> fitted(anfis3)[1:5,]
[1]  0.0008394804  0.007514315  0.012824644 -0.003044622 -0.030125619
> residuals(anfis3)[1:5,]
[1]  0.011354394 -0.014242212 -0.010291178
  0.013337535  0.005391769
```

The learning ANFIS algorithm on train data features for diagnosing hepatitis. Training data features age, sex, Bilirubin, ALP, Albumin, SGOT. The software tool used by R-Studio 3.5.1 in the ANFIS algorithm. The learning gradient descent and least square estimation on the ANFIS algorithm to obtain an error value. The results obtained at ANFIS error: 0.05391769.
8

Figure 3. Learning ANFIS on training data.

Meanwhile the features of the test data: Age, sex, Bilirubin, ALP, SGOT, and albumin to diagnose hepatitis. The results obtained at ANFIS error: 0.01337553

Figure 4. Learning ANFIS on test data.

Gradient descent and least square estimation are the learning in the ANFIS algorithm. Gradient descent and least square estimation can be known for its optimization with the smallest error value.

Classification and modeling in the training data:

Figure 5. Classification and modeling of the ANFIS algorithm in the training data.

Relevant features are classified. The features are Age, sex, Bilirubin, ALP, SGOT, and albumin. Modeling of relevant features for diagnosing hepatitis. The features: Age, Sex, Bilirubin, ALP, SGOT, and Albumin in training data to predict the diagnosis of hepatitis.
The prediction of the ANFIS algorithm on the train data can be seen, representation in Figure 6.

**Figure 6.** Prediction simulation of ANFIS Algorithm for *train data* of hepatitis disease.

The features of the ANFIS algorithm for *train data*: Age, sex, Bilirubin, ALP, SGOT, and albumin with 40 samples. Classification and modeling of ANFIS algorithm with train data obtained modeling for albumin features ± 58% also simulation of albumin feature modeling = 86.59 %.

Classification and modeling of ANFIS algorithm for test data:

**Figure 7.** Classification and modeling of ANFIS algorithm with test data.

Testing data for hepatitis diagnose. The features of the test: Age, sex, Bilirubin, ALP, SGOT, and albumin with 39 samples. Classification and modeling of ANFIS algorithm with test data diagnosing hepatitis. Prediction and simulation of the ANFIS algorithm with test data can be seen representation in Figure 8.

**Figure 8.** Prediction simulation for *test data* in hepatitis disease.
The classification of features in test data: Age, sex, Bilirubin, ALP, SGOT, and albumin with 39 samples obtained simulations and predictions. Classification and modeling of test data obtained predictive modeling for albumin features ± 58% also predictive simulations = 84.72%.

4.2. *The results and discussion of genetic algorithms.*

After cleaning and classifying the relevant and irrelevant features, the algorithm shows the performance of the genetic algorithm. The genetic algorithm shows mathematical functions on the maximum and minimum functions that show the best solution to the problem as indicated by equation 6:

\[ f(x) = (x^2 + 1) \cdot \cos(x) \]  

(6)

![Figure 9. Classification of Genetic Algorithms.](image)

Features in genetic algorithms: age, sex, Bilirubin, Alk Phosphate (ALP), SGOT, Protime through the stages of genetic algorithms, namely: Encoding, selection, crossover, and mutation. The implementation of genetic algorithms by generating minimum and maximum functions with mathematical functions. The minimum and maximum functions in genetic algorithms show the best solution to the problem. Classification and the modeling of relevant features are representations in Figure10.

![Figure 10. Classification of relevant features in genetic algorithms.](image)
The feature classifications are Bilirubin, Alk Phosphate (ALP), SGOT, and Albumin to obtain the desired results. Modeling features is relevant in genetic algorithms. Classification uses features to obtain the diagnosis of hepatitis in Table 4.

| Algorithm          | Accuracy |
|--------------------|----------|
| ANFIS              | 86.66    |
| Genetic Algorithm  | 98.73    |

The classification of relevant features can improve the performance of the ANFIS algorithm so accuracy can obtain on ANFIS algorithms and genetic algorithms. Genetic algorithms are needed to improve the performance and accuracy of ANFIS algorithms.

5. Conclusion

Classification is needed for accuracy. Classification of hepatitis data by reducing irrelevant features without eliminating relevant features by using the ANFIS algorithm and genetic algorithm. A genetic algorithm is needed to improve the performance and accuracy of the ANFIS algorithm. Genetic algorithm accuracy is higher than the ANFIS algorithm. A genetic algorithm is the best solution to the problem.

References

[1]Jang JSR 1996. Input selection for anfis learning. IEEE Xplore Digital Library pp : 1493-1499
[2]Amma NGB 2013. An intelligent approach based on principal component analysis and adaptive neuro-fuzzy inference system for predicting the risk cardiovascular diseases. Fifth Internasional Conference on Advanced Computing. IEEE Xplore Digital Library pp 241-245
[3]Kalaivani P and Shunmuganathan 2014. An improved k-nearest neighbour algorithm using genetic algorithm. IEEE Xplore Digital Library pp 1657-1651
[4]Walia N,Singh H and Sherma A 2015. ANFIS: adaptive neuro-fuzzy inferences system-a survey. Internasional Journal of Computer Applications Vol.123 No.123 pp 32-38
[5]Arotaritei D and Rotariu C 2015 Artificial intelligence methods applied to parameter detection of atrial fillation. Journal of Physics: Conference Series 637. pp 1
[6]Aparna U R and Shaju P 2016. Feature selection and extraction in data mining. IEEE Xplore Digital Library
[7]Wu G,Liu Y,Wang S,Huang W,Liu T,and Yin Y 2016. The classification prognosis models of hepatitis b virus reaction based on bayes and support vector machine after feature extraction genetic algorithm .12 th Internasional Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery IEEE Xplore Digital Library pp 572-577
[8]Hernandez J J ,Gomez E G J,Counder C, and Ledesma 2017. Envolving aerodynamic air foils for wind turbine through a genetic algorithm. IOP Conf.Series. 792.
[9]Das R N 2018 Determinants of total bilirubin for liver patient Journal of liver research,disorders & therapy Vol 4 Issue 3 pp 127-131
[10]Jones A H,Paolo A,Danto A and Winarti 2018. Prediction of popular growth using sugeno adaptive neuroscience fuzzy inferences system:IOP Conf.Series:Material Science and Engineering.403.
[11]Solikh M,Nuryani N and Nugroho A S 2019.Identification of atrial fibrillation using descriptive statistic features and adaptive neuro-fuzzy inference system. IOP Conf.Series:Journal of Physics : Conference.Series 1153 pp.1
[12]Sabara M A, Somatri O, Nurcahaya H, Kurnia, Latifah U and Harsono 2019. Diagnosis classification of dengue neural network and genetic algorithm. IOP Conf.Series:Journal of Physics : Conf.Series.1175
[13] Sapuhtera Y E and Soegjino B 2019. *Optimation of AMC’S tensile properties using adaptive neuro fuzzy inferences system*. IOP Conf.Series : Materials Sciences and Engineering.515
[14] Solikah M, Nuryani N, Nugroho 2019. *Identification of atrial statistic features and adaptive neuro fuzzy inferences system*. IOP. Conf.Series : Journal of Physics : Conf.Series.1153.