Iganfis Data Mining Approach for Forecasting Cancer Threats

Benard Nyangena Kiage[1]
[1]Student
[1]Department of Computing and Information Technology
[1]Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya

Abstract— Healthcare facilities have at their disposal vast amounts of cancer patients' data. Medical practitioners require more efficient techniques to extract relevant knowledge from this data for accurate decision-making. However the challenge is how to extract and act upon it in a timely manner. If well engineered, the huge data can aid in developing expert systems for decision support that can assist physicians in diagnosing and predicting some debilitating life threatening diseases such as cancer. Expert systems for decision support can reduce the cost, the waiting time, and liberate medical practitioners for more research, as well as reduce errors and mistakes that can be made by humans due to fatigue and tiredness. The process of utilizing health data effectively however, involves many challenges such as the problem of missing feature values, the curse of dimensionality due to a large number of attributes, and the course of actions to determine the features that can lead to more accurate diagnosis. Effective data mining tools can assist in early detection of diseases such as cancer. In This paper, we propose a new approach called IGANFIS. This approach optimally minimizes the number of features using the information gain (IG) algorithm which is usually used in text categorization to select the quality of text. The IG will be used for selecting the quality of cancer features by virtue of reducing them in number. The reduced number quality features dataset will then be applied to the Adaptive Neuro Fuzzy Inference System (ANFIS) to train and test the proposed approach. ANFIS method of training is ideally the hybrid learning algorithm which uses the gradient descent method and Least Square Estimate (LSE) for computing the error measure for each training pair. Each cycle of the ANFIS hybrid learning consists of a forward pass to present the input vector calculating the node outputs layer by layer repeating the process for all data and a backward pass using the steepest descent algorithm to update parameters, a process called back propagation.

Key words: Data Mining, Clustering, selection, classification accuracy, neural networks, Fuzzy Inference system, Information gain

I. INTRODUCTION

Medical Databases today can range in size into hundreds of millions of terabytes. Within these masses of data lies hidden information of strategic importance. Drawing meaningful conclusions about this vast data has always been a challenge to healthcare practitioners. Data mining (DM) solves this problem. DM is a non trivial extraction of implicit, previously unknown, and imaginable useful information from data. DM finds important information hidden in large volumes of data. DM is the reasoning of data. It is the use of software techniques for finding patterns and consistency in sets of data [12]. Although computational, the utility of data mining algorithms can be used as qualitative tools to analyze quantitative data, particularly the large, complex databases being created by the health informatics community. Many countries have embraced the global healthcare system. This is done by standardizing healthcare in communication and building electronic healthcare records (EHR).

Health records may include a range of data such as general medical records, patient examinations, patient treatments, medical history, allergies, immunization status, laboratory results, Radiology images and other useful medical information for examination. This rich information may help researchers in examining and diagnosing diseases using computer techniques. EHR are capable of being shared across healthcare providers in various countries [1].

Data stored in hospital warehouses range from quantitative to analog to qualitative data; however well structured, these data conceal implicit patterns of information which cannot readily be detected by conventional analysis techniques.

Cancer diagnosing based on machine intelligence and previous history can be a step towards the reduction of the suffering of cancer patients in the entire world over. What is required however is a reliable, accurate and efficient approach for identifying diagnostic features that best describe data for the purpose of differentiating malignant and benign form of cancer, determining how missing feature values can improve prediction in determining the performance achieved by the data mining technique used and Investigating how classification accuracy and missing values can improve results by fusing the existing data mining algorithms for cancer diagnosis.

In this paper we propose a new technique (IG-ANFIS) which combines Adaptive neuro-Network based Fuzzy Inference System (ANFIS) and the Information Gain method (IG). ANFIS will be used to build an input-output mapping using both human knowledge and machine learning ability while IG method is to reduce the number of input features to ANFIS. In this study, sets of computations will be performed to evaluate benchmark attributes selection methods on well-known publicly available dataset from UCI machine learning repository and Wisconsin Breast Cancer (WBC) dataset. The structure of this paper is as follows; section 2 describes cancer diagnosis based on ANFIS, section 3 describes an overview of information gain methodology, section 4 described the IG –ANFIS experimental approach, section 5 provides the IG-ANFIS experimental results, section 6 describes related work , section 7 describes conclusions and future works and section 8 provides the references used.

II. ANFIS STRUCTURE

ANFIS is a combination of two learning approaches: Neural Network (NN) and Fuzzy Inference System (FIS) [89]. The purpose for ANFIS is to build an input-output mapping using both human knowledge and machine learning ability. ANFIS exploit the advantages of NN and FIS by combining the human expert knowledge (FIS rules) and the ability to

adapt and learn. FIS has a rule base made up of a selection of fuzzy rules; a database defining membership functions and a reasoning mechanism for outputting inferences.

Our approach applies Sugeno fuzzy rules. A common rule set for two fuzzy if-then - Sugeno rules can be:

Rule 1: if x is A_1 and y *is* B_1, then $f_1 = p_{1x} + q_{1x+}r_1$ (1)

Rule 2: if x is A_1 and y *is* B_2, then $f_2 = p_{2x} + q_{2x+}r_2$ (2)

Figure 13 (a) shows the fuzzy reasoning and (b) shows the corresponding structure of ANFIS.

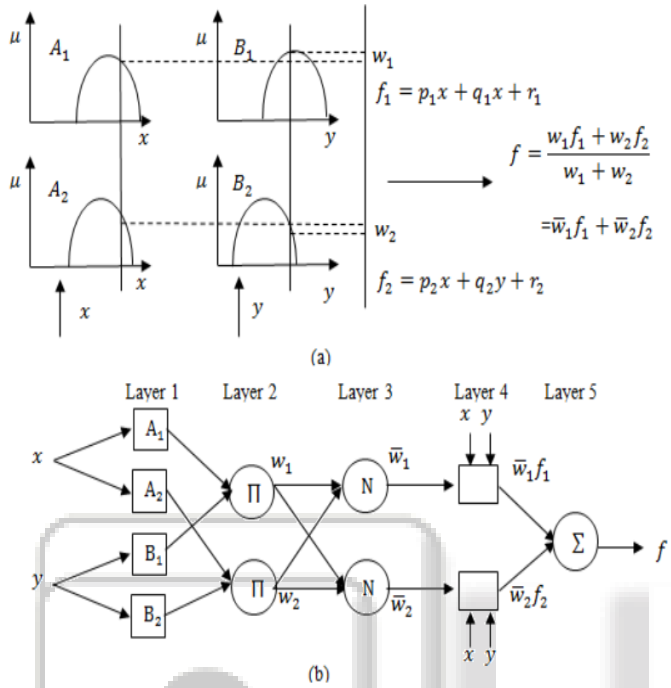

Fig. 1: (a) A two Sugeno fuzzy inference system with two rules, (b) An equivalent ANFSI Architecture.

An ANFIS network of five layers is demonstrated with the equivalent Sugeno fuzzy inference system in Figure 1 above.

ANFIS learns first the structure and then learns the parameters. Structure-learning includes space classifying of fuzzy input and rule-extracting. Accordingly clustering is done by extracting a set of rules that models the data behavior to classify the training sample space. If the space is clustered into n_i classes, then there will be corresponding n_i fuzzy rules. Hence, initial input parameters of membership functions for each class are determined by the clustered center coordinates and its radius length. In Figure 1 (b), the node function in each layer can be described as follows:

Layer 1: Each node i (represented as a square) in this layer accepts input and computes the membership $\mu_{A_i}(x)$

$$o_i^1 = \mu_{A_i}(x) \quad (3)$$

Where x is the input to node i, and A_i is the label (small, large, etc.) associated with this node. In other words, o_i^1 is the membership function of A_i and it specifies the degree to which the given x satisfies the quantifier. $\mu_{A_i}(x)$ is chosen to be bell-shaped with values between 0 and 1, such as the generalized bell function:

$$\mu_{A_i}(x) = exp\left[-\left(\frac{x-c_i}{a_i}\right)^2\right] \quad (4)$$

Where a_i and c_i are two parameters called premises

Layer 2: Every node in this layer (represented by a circle) takes the corresponding outputs from Layer 1 and multiplies them to generate a weight:

$$w = \mu_{A_i}(x) \times \mu B,\ i=1, 2 \quad (5)$$

Layer 3: Every node in this layer is a circle node labeled *N*. This layer normalize the weight of a certain node in comparison to the sum of other nodes weights (The ratio of weight) then compute the implication of each output member function.

$$\overline{w_i} = \frac{w_i}{\sum_j w_j}\ ,\quad i=1, 2.\ j=2 \quad (6)$$

Layer 4: Every node in this layer is illustrated with a square. Based on Sugeno inference system, the output of a rule can be written on the following linear format:

$$0_i^4 = \overline{w_i} f_i = \overline{w_i}(p_i x + q_i y + r_i) \quad (7)$$

Layer 5: This layer called the aggregation layer, which computes the summation of rules, the proposed algorithm produce a single output (centroid):

$$0_i^5 = finaloutput = \sum_i \overline{w_i} f_i = \frac{\sum_i w_i f_i}{\sum_i w_i} \quad (8)$$

A. *Training ANFIS Model:*

The method to train ANFIS is the hybrid learning algorithm. This algorithm uses the gradient descent method and Least Square Estimate (LSE). Each cycle of the hybrid learning consists of a forward pass and a backward pass. In the forward pass the signal travels forward until Layer 4 and the consequent parameters are identified using the LSE method. In the backward pass the errors are propagated backward and the premise parameters are updated by gradient descent. The process is repeated until it achieves the lowest error or a predefined threshold. In other words; the total parameter set is split into three: S = set of total parameters, S1= set of premise (nonlinear) parameters, S2= set of consequent (linear) parameters. So, ANFIS uses a two pass learning algorithm: where S1 is unmodified and S2 is computed using a LSE algorithm. In Backward Pass, S2 is unmodified and S1 is computed using a gradient descent algorithm such as back propagation. So, the hybrid learning algorithm uses a combination of steepest descent and least squares to adapt the parameters in the adaptive network.

III. INFORMATION GAIN

The information gain method was generally proposed to approximate the quality of each attribute using the entropy by estimating the difference between the prior entropy and the post entropy. This is one of the simplest attribute ranking methods and is often used in text categorization. If x is an attribute and c is the class, the following equation gives the entropy of the class before observing the attribute:

$$H(x) = -\sum_x p(x) log_2 P(x) \quad (9)$$

The conditional entropy of c given x (post entropy) is given by:

$$H\ c|x = -\sum_x p(x) \sum_c p(c/x) log_2 p(c/x) \quad (10)$$

Where p (c) is the probability function of variable c

The information gain (the difference between prior entropy and postal entropy) is given by the following equations:

$$H(c,) = H(c) - H(c|x) \quad (11)$$

$$H\ c,\ \text{x}=$$
$$-\sum_c p(\text{c})log_2\text{p(c)}\sum_x(-\ p(\text{x})\sum_c p(^c/_x)log_2\text{p}(^c/_x))(12)$$

IV. THE IG –ANFIS APPROACH

This approach combines the information gain (IG) method and ANFIS method. The IG selects the quality of attributes producing as output a set of features with high ranking values, which eventually becomes input for ANFIS. The selected features are applied to ANFIS for training and testing the proposed approach. The structure of the proposed approach is shown in Figure 2, where X= $\{x_1, x_2,\ldots,x_n\}$ are the original features in dataset, Y=$\{y_1, y_2,\ldots,y_k\}$ are the features after applying the information gain, and Z is the final output after applying Y on ANFIS (the diagnose).

X= $\{x_1, x_2,\ldots,x_n\}$ → $H(c,) = H(c) - H(c|x)$ →

Y=$\{y_1, y_2,\ldots,y_k\}$ → ANFI → Z

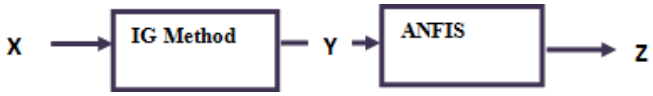

Fig. 2: The structure of the proposed approach

This process involves a number of stages: The first stage selects the most important features that may lead to more accurate results. The second stage is to construct the fuzzy inference system (FIS). In our case we used the most known MATLAB Type Sugeno-FIS which is computationally efficient and works well with optimization and adaptive techniques. Sugeno Fuzzy Inference system has been used to map feature to feature membership functions, feature membership function to rules, rules to a set of output, output to output membership functions, and the output membership function to a single output. In our proposed approach, the rules have been defined from the real data. The rules express the weight of each feature by giving higher priority for features that have the highest rank. The proposed approach contains 81 rules (Number of rules = $x^y$where x is the Number of member functions and y is the number of features i.e.3^4=81 rules). The following figure represents examples of rules used in the proposed approach

1. If (UCSIZE < =2,BN<=3) Then Diagnosis = Benign
2. If (UCSIZE <= 2,BN>3,CT<=3) Then Diagnosis = Benign
3. If (UCSIZE <= 2.5,BN>3,CT>3,BC<=2,MA<=3) Then Diagnosis = Malignant
4. If (UCSIZE <= 2.5,BN>3,CT>3,BC<=2,MA>3) Then Diagnosis = Benign
5. If (UCSIZE <= 2.5,BN>3,CT>3,BC>2) Then Diagnosis = Malignant
6. If (UCSIZE >2,UCSHAPE<=3,CT<=5) Then Diagnosis = Benign
7. If (UCSIZE >2,UCSHAPE<=3,CT>5) Then Dia = Malignant
8.If(UCSIZE>2,UCSHAPE>2,UCSIZE<=4,BN<=2,MA<=) Then Diagnosis = Benign
9.If(UCSIZE >2,UCSHAPE>2,UCSIZE<=4,BN<=2,MA>3) Then Diagnosis = Malignant
10. If (UCSIZE >2UCSHAPE>2,UCSIZE<=4,BN>2) Then Diagnosis = Malignant
11. If (UCSIZE >2,UCSHAPE>2,UCSIZE>4) Then Diagnosis = Malignant

Fig. 3: Examples of Rules generated in the proposed approach

V. THE IG-ANFIS EXPERIMENTAL RESULTS

In this work the database has been divided into training and testing datasets. 341 records used for training and 342 records for testing. 16 records have been ignored since they posses missing values. Normalization has been done to class attributes [0=Benign, 1=Malignant]. Information gain method has been used to select the quality of attributes. Table1shows the ranking of attributes after applying attribute evaluator and the searching method Ranker-T-1 using WEKA on WBC dataset

Attribute Name	Rank
Uniformity of Cell Size (UCSize)	0.636
Uniformity of Cell Shape (UCSshape)	0.633
Normal Nucleoli (NN)	0.555
NNs Bare Nuclei(BN)	0.538
Single Epithelial Cell Size (SECS)	0.421
Clump Thickness (CT)	0.411
Marginal Adhesion (MA)	0.394
Bland Chromatin (BC)	0.316
Mitoses(MI)	0.278

Table 1**:** Information Gain Ranking Using WEKA on WBC

This approach selects a certain number of features based on features rank. The most significant change in ranking is as shown by the slope point in graph of figure 4. The biggest drop is just after the feature number 4 (BN). Respectfully, features; UCSize, UCShape, NN, BN are the 4 features selected to train and test the model.

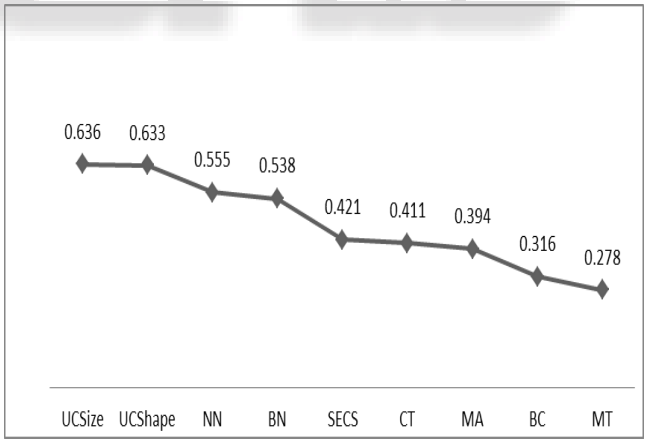

Fig. 4: Information Gain Ranking on WB

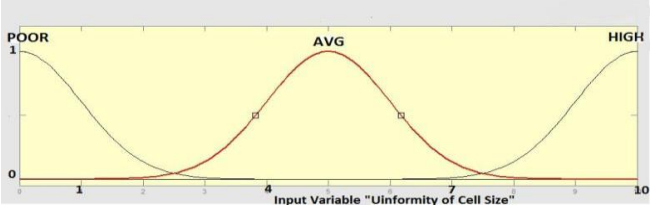

Fig. 5**:** Input Membership Function for the feature "Uniformity of Cell Size"

In the third and final stage, the constructed Fuzzy Inference System and the new features set were loaded to ANFIS which will train and test the proposed approach.

Table 11 is a representation of Comparison of classification accuracy between IG-ANFIS and some previous work which can also be represented in the bar graph as shown in the figure 6 below-:

The approach	Accuracy
AdaBoost	57.60%
ANFIS	59.90%
SANFIS	96.07%
FUZZY	96.71%
FUZZY- GENETIC	97.07%
ILFN	97.23%
NNs	97.95%
ILFN and FUZZY	98.13%
IG-ANFIS	98.24%
SIANN	100.00%

Table 2: Comparison of classification accuracy between IG-ANFIS and some previous work

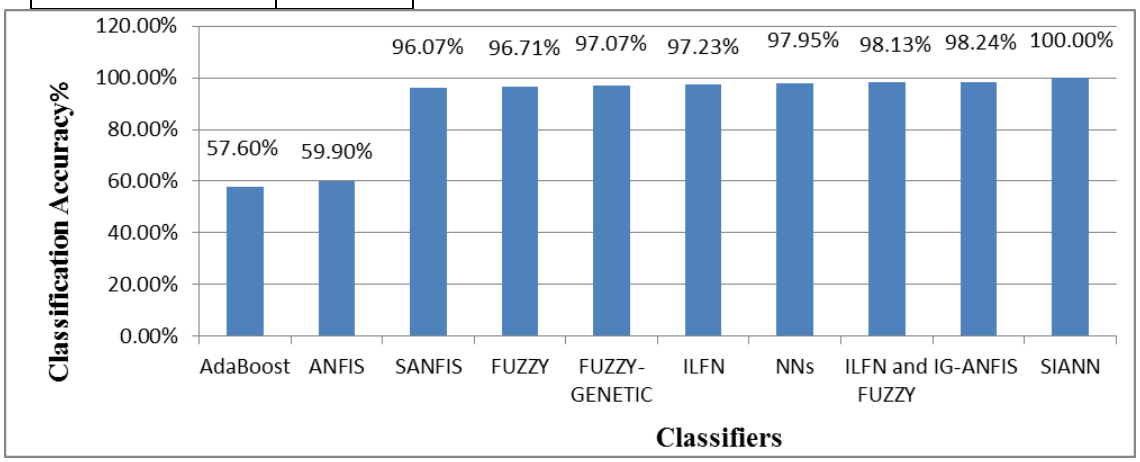

Fig. 6: Comparison of classification accuracy between IG-ANFIS and some previous work

VI. Related Work

Basing our discussion on breast cancer because of the readily available data for research purposes, we now bring forth some of the related prior work on data mining methods and machine learning for cancer diagnosis.

Meesad and Yen (2003) proposed a hybrid Intelligent System (HIS) which integrates the Incremental Learning Fuzzy Network (ILFN) with the linguistic knowledge representations. The linguistic rules were determined based on knowledge embedded in the trained ILFN or been extracted from real experts. In addition, the method also utilized Genetic Algorithm (GA) to reduce the number of the linguistic rules that sustain high accuracy and consistency. After being completely constructed, the system could incrementally learn new information in both numerical and linguistic forms. The proposed method was evaluated using Wisconsin Breast Cancer Dataset (WBC). The results showed that the proposed HIS performed better than some well-known methods.

Setiono (2006) proposed a method to extract *classification rules* from *trained neural networks* and discussed its application to breast cancer diagnosis. He also explained how the pre-processing of datasets can improve the accuracy of the neural network and the accuracy of the rules since some rules could be extracted from human experience, and may be erroneous. The data pre-processing involves the selection of significant attributes and the elimination of records with missing attribute values from Wisconsin Breast Cancer Diagnosis dataset. The rules generated by Setiono's method were more brief and accurate than those generated by other methods mentioned in the literature.

On their new approach that was based on artificial intelligence technology, Song et al. (2010) presented an automatic breast cancer diagnosis. This is a hybrid system for diagnosing new breast cancer cases in collaboration between Genetic Algorithm (GA) and Fuzzy Neural Network. They also showed that many problems that have high complexity and strong non-linearity with huge data to be analyzed, can use inputs reduction i.e. Features selections methods.

Arulampalam and Bouzerdoum (2011) proposed a method for diagnosing breast cancer named Shunting Inhibitory Artificial Neural Networks (SIANNS). SIANN is a neural network stimulated by human biological networks in which the neurons interact among each other's via a nonlinear mechanism called shunting inhibition. The feed forward sianns have been applied to several medical diagnosis problems and the results were more favourable than those obtained using Multilayer Perceptions (MLPS). They also investigated a reduction in the number of inputs.

VII. Conclusions and Future Work

To go by The results shown by our approach, then further attempts to engage in the application of information technology in cancer patients diagnosis, can lead to a breakthrough in the provision of efficient, timely and decent healthcare services in many states. A Comparison of classification accuracy between IG-ANFIS and some other methods showed an improvement of upto 98.24% for our approach.

Large databases that are used in the medical sector still have a concern of Missing features values. The description of future work can be as follows; the information gain ranking before considering them as inputs for ANFIS was the major investigation carried out in this paper. However classification accuracy can be considered in measuring how the proposed approaches perform. In the next paper we will proposes a new machine learning approach for constructing missing feature values and feature

distance metrics in determining classification accuracy, engage in the speed of classifiers to produce the desired results. In my opinion the faster the classifier can produce the results required, the better and effective the technique. Future work will focus on Classifier fusion and the cost of computation.

VIII. ACKNOWLEDGMENT

This proposal is as a result of inspirational assistance and guidance of my Dad, mentors, lecturers, professionals, and the administrative staff at the Jomo Kenyatta University of Agriculture and Technology as well as at the Machakos university college, my work place. First and foremost, I am grateful to my supervisors; Dr. George Okeyo and Dr. Wilson Cheruiyot, for their invaluable and continuous guidance during the conception of my proposal. My other regards goes Dr. Kimwele, who has always kept me on toes and encourage on this research. They have all given me an undisputed considerable help in every way possible.
I owe you all!

REFERENCES

[1] Lloyd-Williams, Empirical studies of the knowledge discovery approach to health information analysis. Informatica, 2013. 31: p. 249-253.
[2] Gunter, D.T. and P.N. Terry, the Emergence of National Electronic Health Record Architectures in the United States and Australia: Models, Costs, and Questions. J Med Internet Res, 2005. 7(1).
[3] Duda et al., An introduction to variable and feature selection. J. Mach. Learn. Res., 2003. 3: p. 1157-1182.
[4] Harold Doty and William (2013) "Theory Building: Toward Improved Understanding and Modeling" Machine Learning, 1986. 1(1): p. 81-106.
[5] World Health Organization Assesses the World's Health Systems. World Health Organization, 2010 website: http://www.who.int/whr/2000/media_centre/press_release/en/index.html.
[6] Priddy and Keller (2005): Artificial neural networks, an introduction. Washington: SPIE.
[7] Frawley W. j (2012) "knowledge discovery in databases: an overview". AI Magazine, Fall 2012, 213-228.
[8] Han, J. and K. M, "Data Mining Concepts and Techniques". Vol. 3. 2012, San Franscisco: Morgan Kaufmann.
[9] Kotsiantis, S., Supervised Machine Learning: a Review of Classification Techniques. Informatica, 2007. 31: p. 249-268.
[10] Daniel T. Larose (2013). "Discovering Knowledge in Data" Uniqueness of medical data mining. Artif. Intell. Med., 2013. 26(1-2): p. 1-24.
[11] Setiono, R., Generating Concise and Accurate Classification Rules for Breast Cancer Diagnosis, Artificial Intelligence in Medicine, 2006. 18(3): p. 205-219.
[12] Saeys, Y., I. Inza, and P. Larrañaga, A review of feature selection techniques in bioinformatics. Bioinformatics, 2007. 23(19): p. 2507-2517.
[13] Hall, M.A. and G. Holmes, Benchmarking Attribute Selection Techniques for Discrete Class Data Mining. IEEE Transactions on Knowledge and Data Engineering, 2003. 15(3)
[14] Information on UCI Machine Learning Repository -thewebsite: http://archive.ics.uci.edu/ml/about.html.
[15] Leach, M., (2012) "Parallelizing Feature Selection Algorithms". University of Manchester: Manchester.
[16] Larose, D., Discovering Knowledge in Data: An Introduction to Data Mining. 2005, New Jersey: John Wiley & Sons, Inc.

