A Survey of Knowledge Extraction Classification using Different Techniques

Mona Shrivastava
RKDFIST Bhopal

Piyush Singh
RKDFIST Bhopal

Gaurav Shrivastava
RKDFIST Bhopal

ABSTRACT
Knowledge extraction is the method of extracting some useful information from a set of databases such that the extracted information can be used in a wide variety of applications. Here a brief survey of different techniques of classification for the knowledge extraction is given. Although there are many technique used for the classification but here the knowledge extraction for useful information techniques is presented**.**

Keywords
Decision Tree, Fuzzy Logic, Genetic Algorithm, Knowledge Extraction.

1. INTRODUCTION
In the current days, intelligent and automatic systems are essential to assist specialists in the knowledge extraction process to help in the decision making. In the real life, the data or operations with anomalies can have many common characteristics with normal data so that simple methods of pattern recognition usually are not enough to detect these anomalies in an efficient way. This problem grows up when the anomalies perception is only over human specialists without any computational tool.

Some tasks like selection, fraud detection, pattern recognition, medical diagnostics, prognostics and many others has been supported by computational techniques. But a problem appears when working with manual rules. Normally, the systems are based in rules and parameters that, when violated, can indicate a possible anomaly. These rules and parameters have been defined by specialists with a great experience in the investigated area. But when the number of rules and parameters are large, the manual process of rule creation and update is critical, affecting the classification quality and efficiency. The second problem occurs with some tasks where the system has to be able to show not only good numerical results, but also it's clear that the system has to show results with good interpretability, in other words, the specialist has to know the reason of the given decision[4].

So, it's desirable to have an automatic system to generate interpretable rules since a human specialist cannot be able to express the knowledge about the problem in a rule form or just there are no available specialists with a sufficient knowledge in the problem domain.The objective of this work is to show a system able to automatically generate Fuzzy Rule (FR) sets with less human participation in order to get better classification of instances or transactions in an efficient way in direction of both numerical and interpretable results as, in some cases, it's better to have results with less efficiency in a numerical terms but with more efficiency in terms of interpretability. In this context, an interpretable result means a result where the specialist can know the reasons about the given classification by the system[3].

2. FUZZY LOGIC
Fuzzy systems have shown their utiliy in wide range of problems in different applications domain.Fuzzy logic is a form of many valued logic or probabilistic logic,it deals with reasoning that is approximate rather than fixed and exact. Fuzzy logic has been extended to handle the concept of partial truth, where the truth value may range between completely true and completely false. Furthermore, when linguistic variables are used, these degrees may be managed by specific functions.

In Classical Boolean Logic a variable can only be 0 (False) or 1 (True). Fuzzy Logic is an extension of the Classical Boolean Logic proposed by Lotfi Zadeh in 1965 where a variable can assume any value between 0 (False) and 1 (True) . A set of logical operations such as "fuzzy AND" and "fuzzy OR" can then be defined. Using these definitions one can deduce the concepts of "fuzzy sets" and "membership functions". For a classical set the output of a membership function can only be 0 or 1, while for a fuzzy set it can be any real number between 0 and 1.

Figure1 illustrates the difference between classical and fuzzy sets. If the temperature of an object is 10 (in a certain temperature scale), in the classical view this temperature receives the label "low temperature" since the degree of pertinence of "temperature = 10" is 1 to the classical set of "low temperatures". In the fuzzy view, 'temperature = 10" is simultaneously "low" and "medium" since the degrees of pertinence to the fuzzy sets "low temperature" and "medium temperature" are 0.2 and 0.5 respectively[1].

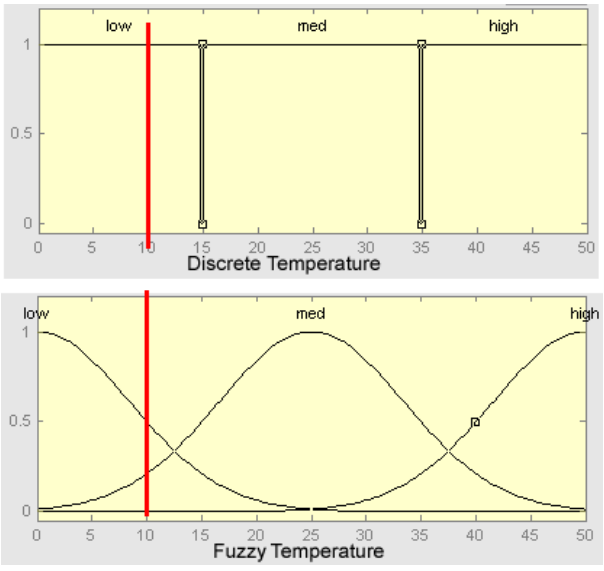

Fig.1 Membership functions for classical (top) and fuzzy (bottom) sets

Fuzzy Logic can be applied to a Decision Tree to generate a Fuzzy Rule-Based System (FRBS)[2][10]. This is particularly useful when at least some of the attributes that are tested in the decision nodes are numerical. In this case, the tests can be formulated using labels (e.g., "if Temperature is low") and the children nodes will be partially activated. As a consequence, two or more (possibly conflicting) terminal nodes will be partially activated and an aggregation method should be used to generate the final output/conclusion of the FRBS [4][6].

3. GENETIC ALGORITHM

Genetic Algorithm (GA) is a search heuristic that mimics the process of natural evolution. This heuristic is routinely used to generate useful solution to optimization and search problems. Genetic algorithms belong to the larger class of Evolutionary Algorithm (EA), which generate solutions to optimization problems using techniques inspired by natural evolution, such as inheritance, selection,mutation and crossover.

Genetic Algorithms (GA) are based on the "survival of the fittest" principle (also known as "natural selection") to solve hard nonlinear optimization problems.It is believed that the evolution of a group of individuals of any biological species in nature is based fundamentally in three phases: mutation, crossover and selection of the fittest. Mutation is a small random variation in the genetic material of some individuals in the current generation. Crossover is the combination of the genetic material of any two individuals in the current generation to generate the individuals of the next generation. Figure 2 shows a possible crossover implementation [1].

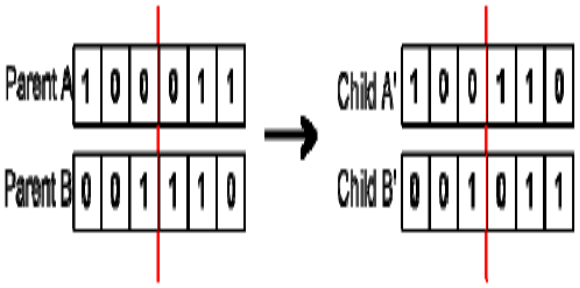

Fig.3 Example of a Decision Tree

In the "selection of the fittest" phase, the environments where the individuals live impose a set of conditions. The individualthat are not able to satisfy these conditions are eliminated from the population [3].

The work presents design and development of a system to automatically evolve rules through genetic-fuzzy approach. The work highlights the advantages of genetic and fuzzy hybridization and proposes a framework to evolve rules automatically. The objective of the framework is to reduce the development effort and guide the development procedure in user friendly fashion. The architecture presents a novel approach which integrates fuzzy logic with system design as well as back end of the framework is used to evolve the fuzzy rules using genetic algorithm approach. The architecture of evolving rule based model using genetic-fuzzy approach can also be applied to various domains like advisory systems, decision support systems, data mining systems, and control and monitoring systems, etc. The same approach can be used to provide training for teachers, planning for resources and many more. We have analyzed the different measures proposed in the different quadrants. Since the interpretability of linguistic FRBSs is still an open problem, this will help researchers in this field determine the most appropriate measure depending on the part of the KB in which they want to maintain/improve interpretability.

4. DECISION TREE

Decision tree support tool that uses tree-like graph or models of decisions and their consequences [7][8], including event outcomes, resource costs, and utility, commonly used in operations research, in decision analysis help to identify a strategy most likely to reach a goal. In data mining and machine learning, decision tree is a predictive model that is mapping from observations about an item to conclusions about its target value. The machine learning technique for inducing a decision tree from data is called decision tree learning. . Figure 1 shows an example of a decision tree that can be used to decide if one should go outside to play golf considering the weather conditions [1].

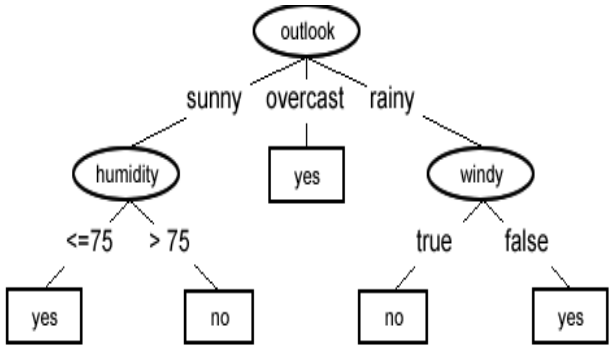

Fig.3 Example of a Decision Tree

In a decision tree the conclusion can be reached using the following procedure:

1) The root node is assumed as the current decision node.

2) Perform the test at the current decision node.

3) Using the test result, activate the corresponding child and take this node as the current node,

4) If the current node is a decision node, go to step 2.

5) If the current node is a terminal node, then stop.

5. FEATURE OF THE ALGORITHM

ID3 Algorithm:

The ID3 algorithm (Inducing Decision Trees) was originally introduced by Quinlan in [5] and is described below in Algorithm. Here they briefly recall the steps involved in the algorithm. For a thorough discussion of the algorithm we refer the interested reader to [6].

Require: R, a set of attributes.

Require: C, the class attribute.

Require: S, data set of tuples.

1: if R is empty then
2: Return the leaf having the most frequent value in data set S.
3: else if all tuples in S have the same class value then
4: Return a leaf with that specific class value.
5: else
6: Determine attribute A with the highest information gain in S.
7: Partition S in m parts S(a1), ..., S(am) such that a1, ..., am are the different values of A.
8: Return a tree with root A and m branches labeled a1...am, such that branch i contains ID3(R − {A}, C, S (ai)).
9: end if

ID3 Decision Tree:

Iterative Dichotomiser is an algorithm to generate a decision tree invented by Ross Quinlan, based on Occam's razor. It prefers smaller decision trees (simpler theories) over larger ones. However, it does not always produce smallest tree, and therefore heuristic. The decision tree is used by the concept of Information Entropy [9]. The ID3 Algorithm is:

1) Take all unused attributes and count their entropy concerning test samples

2) Choose attribute for which entropy is maximum

3) Make node containing that attribute

ID3 (Examples, Target _ Attribute, Attributes)

- Create a root node for the tree
- If all examples are positive, Return the single-node tree Root, with label = +.
- If all examples are negative, Return the single-node tree Root, with label = -.
- If number of predicting attributes is empty, then
- Return the single node tree Root, with label = most common value of the target attribute in the examples.
- Otherwise Begin
- A = The Attribute that best classifies examples.
- Decision Tree attribute for Root = A.
- For each possible value, vi, of A,
    - Add a new tree branch below Root, corresponding to the test A = vi.
    - Let Examples (vi), be the subset of examples that have the value vi for A
        - If Examples (vi) is empty common target value in the examples
    - Else below this new branch add the sub tree ID3 (Examples(vi), Target_ Attribute, Attributes – {A}
- End
- Return Root

6. RELATED WORK

In 2012 Rogério Ishibashi, Cairo Lúcio Nascimento Júnior proposed a fuzzy rule-based system which performs a classification task by using a genetic algorithm and a fitness function it gives the accuracy of the model and its interpretability. It creates a decision tree using any tree induction algorithm such as CART, ID3 or C4.5.The parameters of the membership functions are adjusted by the genetic algorithm.[1**]**

In 2011 M.J. Gacto, R. Alcalá, F. Herrera proposed an overview of the proposed interpretability measures and techniques for obtaining more interpretable linguistic fuzzy rule-based system. It proposes a taxonomy based on a double axis Complexity versus semantic interpretability considering the two main kinds of measures; and ''rule base versus fuzzy partitions'' which considers the different components of the knowledge base to which both kinds of measures can be applied. It provides a well established framework. [2]

In 2005 Francisco Herrera briefly reviews the classical models and the most recent trends for Genetic Fuzzy Systems. It gives specification about some critical considerations of recent developments and to the suggestion of potential research future directions.[3]

In 2005 Hanli Wang, Sam Kwong, Yaochu Jin, Wei Wei, and Kim-Fung Man [4] proposed an agent-based evolutionary approach to extract interpretable rule-based knowledge. In the multiagent system, each fuzzy set agent autonomously determines its own fuzzy sets information. In this scheme Pittsburgh-style approach is applied to extract fuzzy rules that take both the accuracy and interpretability of fuzzy systems into consideration and fuzzy set agents can cooperate with each other to exchange their fuzzy sets information and generate offspring agents.

In 2011 Kunjal Mankad1, Priti Srinivas Sajja2 and Rajendra Akerkar proposed a approach which uses encoding strategy of rules, suitable genetic operators and fitness function which evaluates fuzzy rules for the selected system. The rules, evolved generations and output results are also described.[5]

In 2012 Shakti Kumar1, Parul Narula & Tazeem Ahme and Sohna [6] proposed biogeography based optimization (BBO) for the rule base generation of Mamdani type fuzzy logic based systems. Biogeography is the geographical distribution of biological organisms. It is a burgeoning nature inspired technique to find the optimal solution of the problem.

In BBO, habitats represent the problem solutions, and species migration represents the sharing of features between solutions according to the fitness of the habitats. BBO is a very promising optimizing algorithm for evolving fuzzy logic based systems.

In 2012 N. Alavi proposed an approach of Mamdani fuzzy rule-based which evaluates and classify date fruits was presented. This approach gives the classification accuracy of the MFIS model was 86%. This method is more exact than experts, and provides a better representation of date grading. [7]

In 2004 JUN LIU, JIAN-BO YANG, JIN WANGb, HOW-SING SIIb, and YING-MING WANG [8] proposed a protocol for safety of an engineering system by using fuzzy rule-based evidential reasoning (FURBER) approach. In the approach the parameters used to define the safety level, including failure rate, failure consequence severity and failure consequence probability, are described using fuzzy linguistic variables. The offshore platform is used to illustrate the application of the proposed approach.

In 2009 Jos´e M. Alonso, Manuel Oca˜na, Miguel A. Sotelo, Luis M. Bergasa, and Luis Magdalena proposed a protocol which uses robot localization inside buildings using WiFi signal strength measure. The WiFi signal strength of all visible Access Points (APs) are collected and stored in a database or Wifi map. The protocol uses of Fuzzy Rule-based Classification to obtain the robot position during the estimation Stage. This protocol is easily adaptable to new environments where triangulation algorithms cannot be applied since the AP physical location is unknown.[9]

In 2008 Jos´e M. Alonso, Luis Magdalena, Serge Guillaume presented a protocol for fuzzy system modeling which maximizes the interpretability while keeping high accuracy. it combines of both expert knowledge and knowledge extracted from data. Both types of knowledge are represented using the fuzzy logic formalism. The integration process is made carefully at both levels variables and rules, avoiding contradictions and/or redundancies. It can generate highly interpretable knowledge bases with a good accuracy as compared to that achieved by other methods.[10]

In 2012 Teresa Garcia-Valverde, Alberto Garcia-Sola, Antonio Gomez-Skarmeta and Juan A. Botia, Hani Hagras, James Dooley and Victor Callaghan [11] invented a protocol in which system receives WiFi signals from a big number of existing WiFi Access Points with no prior knowledge of the access points locations and the environment. This scheme is able to adapt online incrementally in a lifelong learning mode to deal with the uncertainties and changing conditions and in simulated and real environments this system has given high accuracy to detect the user in the given AIE.

In 2011 Jingjing Cao and Sam Kwong proposed scheme consist a multi-objective evolutionary hierarchical algorithm to obtain a non-dominated fuzzy rule classifier set and a reduce-error based ensemble pruning method to decrease the size and enhance the accuracy. In this each chromosome represents a fuzzy rule classifier and compose of three different types of genes: control, parameter and rule genes. Similar classifiers are removed to preserve the diversity of the fuzzy system. This approach can maintain a good trade-off among accuracy, interpretability and diversity of fuzzy classifiers.[12]

7. PROPOSED WORK

7.1 Tree Generation of Decision Tree using Horizontal Partition based Id3 Decision

***Input Layer*:**

Define P_1, P_2… P_n Parties. (Horizontally partitioned).

Each Party contains R set of attributes A_1, A_2, …., A_R.

C the class attributes contains c class values C_1, C_2, …., C_c.

For party P_i where i = 1 to n do

If R is Empty Then

Return a leaf node with class value

Else If all transaction in $T(P_i)$ have the same class Then

Return a leaf node with the class value

Else

Calculate Expected Information classify the given sample for each party P_i individually.

Calculate Entropy for each attribute (A_1, A_2, …., A_R) of each party P_i.

Calculate Information Gain for each attribute (A_1, A_2,…., A_R) of each party P_i

- Calculate Total Information Gain for each attribute of all parties (TotalInformationGain( )).
- $A_{BestAttribute} \leftarrow$ MaxInformationGain( )
- Let V_1, V_2, …., V_m be the value of attributes. $A_{BestAttribute}$ partitioned P_1, P_2,…., P_n parties into m parties
- $P_1(V_1)$, $P_1(V_2)$, …., $P_1(V_m)$
- $P_2(V_1)$, $P_2(V_2)$, …., $P_2(V_m)$
- . .
- . .
- $P_n(V_1)$, $P_n(V_2)$, …., $P_n(V_m)$
- Return the Tree whose Root is labelled $A_{BestAttribute}$ and has m edges labelled V_1, V_2, …., V_m. Such that for every i the edge Vi goes to the Tree
- NPPID3(R – $A_{BestAttribute}$, C, ($P_1(V_i)$, $P_2(V_i)$, …., $P_n(V_i)$))
- End.

7.2 Generation of Fuzzy Rule Based System

The process is simplified because the variables in the Decision tree generation phase were previously categorized. The fuzzy rules can be extracted from the decision tree rules by travelling each path of the tree from the leaf to root node. In this phase the algorithm is concerned only in the fuzzy rule extraction [6][7].

8. CONCLUSION

Here in this paper a description of each technique related to the decision tree and classification and fuzzy rules is given. A survey of all the techniques is given here in this paper on the basis of analysis of all the technique a new proposed work has been implemented which is more efficient as compared to the all the existing techniques that we have analyzed so far.

Here an efficient technique is used for the classification of the dataset used using horizontal partition based id3 decision tree and then applying fuzzy rules on the decision tree created

to minimize the error rate of the classified and unclassified instances. The efficiency of the proposed technique implemented here is more as compared to the algorithms implemented for the classification of the knowledge extraction such as C4.5 or CART.

9. REFERENCES

[1] Rogério Ishibashi, Cairo Lúcio Nascimento Júnior "Knowledge Extraction using a Genetic Fuzzy Rule-based System with Increased Interpretability" 2012 SAMI 10th IEEE Jubilee International Symposium on Applied Machine Intelligence and Informatics,January 2012.

[2] M.J. Gacto, R. Alcalá, F. Herrera "Interpretability of linguistic fuzzy rule-based systems: An overview of interpretability measures" 2011 Information Sciences 181 , 2011.

[3] Francisco Herrera "Genetic Fuzzy Systems: Status, Critical Considerations and Future Directions" 2005 International Journal of Computational Intelligence Research, ISSN 0973-1873 Vol.1, No.1, pp. 59-67, 2005

[4] Hanli Wang, Sam Kwong, Yaochu Jin, *Se*, Wei Wei, and Kim-Fung Man "Agent-Based Evolutionary Approach for Interpretable Rule-Based Knowledge Extraction" 2005 IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART C: APPLICATIONS AND REVIEWS, VOL. 35, NO. 2, MAY 2005.

[5] Kunjal Mankad1, Priti Srinivas Sajja2 and Rajendra Akerkar "EVOLVING RULES USING GENETIC FUZZY APPROACH - AN EDUCATIONAL CASE STUDY" 2011 International Journal on Soft Computing(IJSC), Vol.2, No.1, February 2011.

[6] Shakti Kumar, Parul Narula, Tazeem Ahmed, "Knowledge extraction from numerical data for the Mamdani type fuzzy systems: a BBO approach," in proceedings of the international conference on Innovative Practices in Management and Information Technology for Excellence, Jagadhri, India, May 2010.

[7] N. Alavi "Date grading using rule-based fuzzy inference system" 2012 Journal of Agricultural Technology, Vol. 8, pp:1243-1254, ISSN 1686-9141.

[8] JUN LIU, JIAN-BO YANG, JIN WANGb, HOW-SING SIIb, and YING-MING WANG "FUZZY RULE-BASED EVIDENTIAL REASONING APPROACH FOR SAFETY ANALYSIS" 2004 International Journal of General Systems, Vol. 33, pp. 183–204, June 2004.

[9] Jos´e M. Alonso, Manuel Oca˜na, Miguel A. Sotelo, Luis M. Bergasa, and Luis Magdalena "WiFi Localization System Using Fuzzy Rule-Based Classification" 2009 Book Computer Aided Systems Theory - EUROCAST, pp. 383-390, 2009.

[10] Jos´e M. Alonso, Luis Magdalena, Serge Guillaume "Designing Highly Interpretable Fuzzy Rule-Based Systems with Integration of Expert and Induced Knowledge" 2008 In proceeding of 12th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU), Volume: 682-689, 2008.

[11] Teresa Garcia-Valverde, Alberto Garcia-Sola, Antonio Gomez-Skarmeta and Juan A. Botia, Hani Hagras, James Dooley and Victor Callaghan " An Adaptive Learning Fuzzy Logic System for Indoor Localisation using Wi-Fi in Ambient Intelligent Environments" 2012 IEEE World Congress on Computational Intelligence (WCCI 2012) June, 2012.

[12] Jingjing Cao and Sam Kwong proposed "Combining Interpretable Fuzzy Rule-Based Classifiers via Multi-Objective Hierarchical Evolutionary Algorithm" 2011 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 1771 – 1776, 2011.