Optimization of Decision Rules in Fuzzy Classification

Renuka Arora
M.tech student
B.R.C.M. College of Engineering & Technology,
Bahal (Bhiwani)

Sudesh Kumar
Phd, Associate Professor
B.R.C.M. College of Engineering & Technology,
Bahal (Bhiwani)

ABSTRACT

There are various advances in data collection that can intelligently and automatically analyze and mine knowledge from large amounts of data. World Wide Web as a global information system has flooded us with a tremendous amount of data and information Discovery of knowledge and decision-making directly from such huge volumes of data contents is a real challenge. The Knowledge Discovery in Databases (KDD) is the process of extracting the knowledge from huge data collection. Data mining is a step of KDD in which patterns or models are extracted from data by using some automated techniques. Discovering knowledge in the form of classification rules is one of the most important tasks of data mining. Discovery of comprehensible, concise and effective rules helps us to make right decisions. Therefore, several Machine Learning techniques are applied for discovery of classification rules. Recently there have been several applications of genetic algorithms for effective rules with high predictive accuracy.

Keywords:

Classification, Genetic Programming, Evolutionary Algorithms

1. INTRODUCTION

In the last several decades, human capabilities of both generating and collecting data have increased rapidly. Contributing factors include the computerization of many business, scientific and government transactions, and advances in data collection tools ranging from scanned text and image platforms to satellite remote sensing systems. In addition, popular use of the World Wide Web as a global information system has flooded us with a tremendous amount of data and information Discovery of knowledge and decision-making directly from such huge volumes of data contents is a real challenge. To cope with such complexities people have generated an urgent need for new techniques and various tools that can intelligently assist them in transforming the vast amounts of data into useful information and knowledge. These intelligent systems provide an infrastructure that identifies hidden patterns in the gathered data which could discover useful knowledge from data and thereby empower managers to make more effective decisions.

2. KNOWLEDGE DISCOVERY IN DATABASE (KDD)

KDD is the process of finding useful information and patterns in data. It starts with the understanding of the problem and concludes with the analysis and assessment of the results. The input to this process is the data, and the output is the useful information required by the user.

KDD Process:

The KDD process includes two steps:

1. Preprocessing [or Data Preparation] step

The goal of data preparation methods is to transform the data to facilitate the application of given data mining algorithms.

2. Post processing [or Knowledge Refinement] step

The goal of knowledge refinement methods is to validate and refine discovered knowledge.

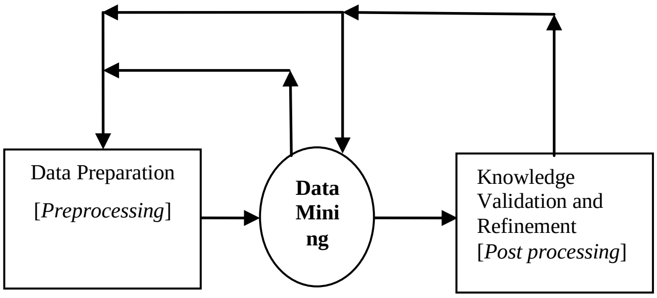

Fig. 1: The iterative nature of the knowledge discovery process

The KDD process is both interactive and iterative, involving numerous steps with many decisions being made by the user. KDD is iterative because the output of each step is often feedback to previous steps as shown in Figure and typically many iterations of this process are necessary to extract high-quality knowledge from data.

3 DATA MINING TASKS

Data mining tasks are also referred to as data mining outcomes or types and can be classified into two categories: descriptive and predictive. Descriptive mining tasks characterize general properties of the data in the database. Examples include association rule discovery and clustering. On the other hand, predictive mining tasks perform inference on the current data in order to make predictions. Examples of predictive mining tasks include classification and regression. In general, the main data mining methods includes: classification, regression, link analysis, segmentation and deviation detection.

3.1Classification

Classification involves mapping data into one of several predefined or newly discovered classes. In the illustration shown in Fig. 2, there are three groups or classes of data, (A), (B), and (C). The classification rule may specify minimum

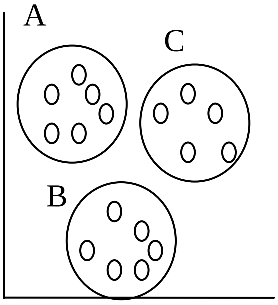

Fig. 2: Classification

proximity to the center of a particular group, as defined by numerical range or statistical spread.

3.2 Regression

Data mining based on regression method involves assigning data a continuous numerical variable based on statistical methods. The main intent behind using regression methods is to extrapolate trends from data samples. In the Fig. 3, the extrapolation formula is a simple linear function of the form:

y = mx + b

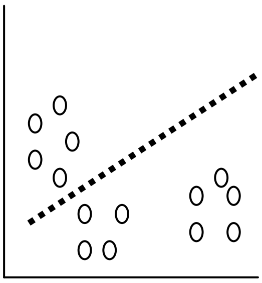

Fig. 3: Regression

where, x and y are coordinates on the plot, m is the slope of the line, and b is a constant. In practice, more complex extrapolation formulas are used to describe data trends.

3.3 Link Analysis

Link analysis evaluates apparent connection or links between data in the database or data warehouse. Link analysis highlights correlation in data that can suggest linkage, but not causality. In the illustration depicted in Fig. 4, the two pairs of data points are apparently linked, in that the value of one data element in the pair can be predicted by the value of the other data point in the pair

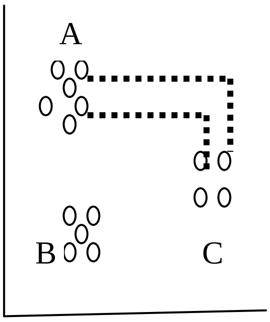

Fig. 4: Link Analysis

3.4 Deviation Detection

Deviation detection represented in Fig. 5, identifies data values that lies outside the norm, as defined by the existing

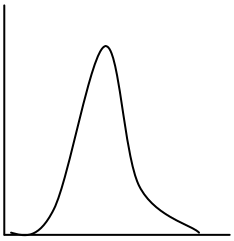

Fig. 5: Deviation Detection

models. The outlier in the illustration is an example of data value outside the expected spread of data in a sample.

3.5 Segmentation

Segmentation based data mining identifies classes or groups of data that behaves similarly, according to some metric. Segmentation is akin to link analysis applied to groups of data

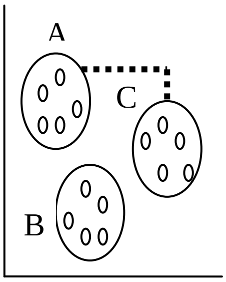

Fig. 6: Segmentation

instead of individual data points. In the Fig. 6 groups (A) and (C) behave similarly.

4. DATA MINING METHEDOLOGY

These are the steps which one would follow to do mining. Once the outcomes are determined then data mining can be done. There are two approaches for data mining: top down and bottom up. One could combine the two and have a hybrid approach. The top down approach starts with some idea or a pattern or hypothesis. In the bottom up approach of data mining there is no hypothesis to be tested. This is much harder as the tool has to examine the data and then come up with pattern. The bottom up approach could be directed or undirected. The hybrid approach is a combination of both top down and bottom up mining. In this the tool can switch between top down and bottom up mining and again between directed and undirected mining.

5. METHODS FOR CLASSIFICATION RULES

Data classification represents an important theme and is perhaps the most commonly applied data mining technique .The classification problem becomes very hard when the number of possible different combinations of parameters are so high that techniques based on exhaustive search of the parameter space rapidly become computationally infeasible. Thus, it is natural to devote attention to a heuristic approach to the classification problem.

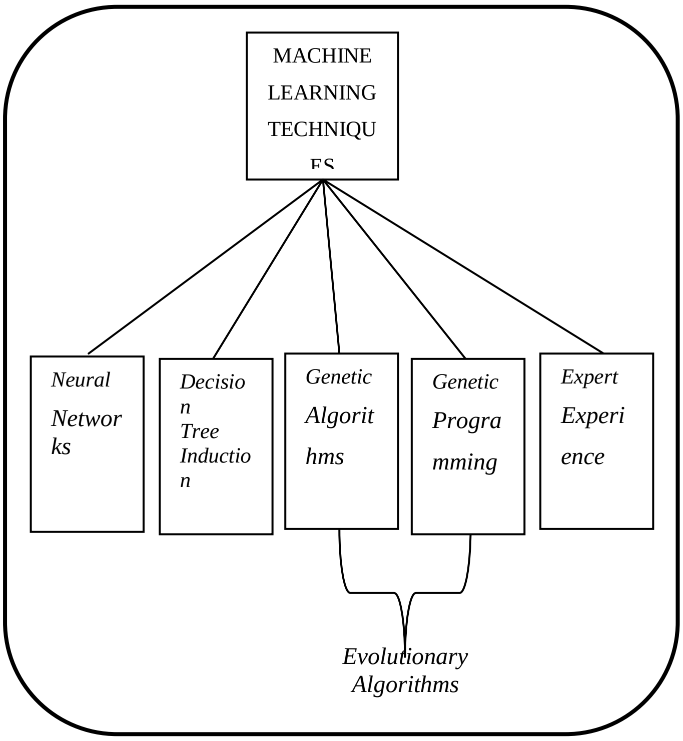
MACHINE LEARNING TECHNIQUES
Neural Networks
Decision Tree Induction
Genetic Algorithms
Genetic Programming
Expert Experience
Evolutionary Algorithms

Fig. 7: Taxonomy of Machine Learning Techniques

5.1 Neural Networks:

Neural Networks (Lippmann, 1987) can also be used to generate fuzzy decision rules as both neural networks and fuzzy systems are functionally equivalent despite of different structures. This functional equivalence has been explained and proved by Buckley et al. (1993) in the sense that any continuous, layered feed forward neural net can be approximated to any degree of accuracy by a fuzzy logic system and any continuous, discrete fuzzy logic system can be approximated to any degree of accuracy by a three-layered, feed forward neural net. Hayashi and Imura (1990) suggested a two-step procedure to extract fuzzy rules. In the first step, a neural net is trained from sample data and in the subsequent step an algorithm is used to automatically extract fuzzy rules from the trained neural net. Many other methods of generating fuzzy rules through neural networks have been suggested. Kosko (1992) proposed a system called Fuzzy Cognitive Maps, which integrates neural network and fuzzy logic.

5.2 Decision Tree Induction:

Another widely used machine learning method is the induction of decision trees (Quinlan, 1986; Safavian and Landgrebe, 1991). The search method decision tree induction employs fuzzy entropy to find the most efficient decision nodes (Weber, 1992; Yuan and Shaw, 1995). Although in most of the cases this method works well but still is rendered inefficient for generating fuzzy decision tree, as it may not be able to generate best tree due to one-step ahead node splitting without backtracking. Moreover, the best tree may not be able to yield the best set of rules.

5.3 Evolutionary Algorithms:

The abovementioned search techniques, neural networks and decision tree induction have problems of being trapped into local optimal. Thus, application of EAs for discovering comprehensible IF-THEN classification rules in the fuzzy environment is preferred over other searching techniques. There has been very extensive research on EAs for discovering FCRs. The motivation for this approach is twofold. First, it can be regarded as a form of incorporating some background knowledge into the rule discovery system. The user can specify membership functions that are sensible, according to his knowledge of the application domain and the meaning of data being mined. As a result, the membership functions will be consistent with the user's previous knowledge which adds to the improvement in the comprehensibility of discovered knowledge (Pazzani, 2000).Second, this approach avoids the computationally-expensive process of trying to optimize the shape and number of membership functions.

5.4 Genetic Programming:

Besides using GA, Genetic Programming (GP) which has emerged as an extension of GA proposed by Cramer (1985) and Salustowicz and Schmidhuber (1997) is also applicable for discovering FCRs when a direct search is impossible. The main difference between GA and GP lies in the representation of the structure they manipulate and the meaning of the representation. Furthermore, GA usually operates on a population of fixed-length binary strings whereas; GP typically operates on population of parse trees or syntactic trees to represent the problem (Michalewicz, 1996; Cordon, 2004). The application of standard GP to the classification task is relatively straightforward, as long as all the attributes are numeric. In this case we can include in the function set several kinds of mathematical function appropriate to the application domain and include in the terminal set the predicting attributes and possibly a random-constant generator. Once we apply functions in the internal nodes of a GP individual to the values of the attributes in the leaf nodes of that individual, the system computes a numerical value that is output at the root node of the tree.

5.5 Expert Experience:

Another approach to construct classification system using Genetic-based machine learning approach is to apply expert experience. In the method proposed by Li et al. (2007), experts experience was translated into fuzzy sets by similarity measures and was integrated into fuzzy genetic based learning mechanism. This classification algorithm was able to achieve better accuracy and interpretability. The main advantage behind applying expert experience to fuzzy classification system is that knowledge of man and machine is similar, so it is convenient to design effective classification system. Moreover, applying expert experience will reduce the computing complexity. Pittsburgh (Smith, 1980) and Michigan (Holland, 1986) approaches are the two genetic-based machine learning approaches for rule discovery. In the former each rule set is handled as an individual thus it represents a complete solution whereas in latter each rule is handled as an individual thus it represents a partial solution. The choice between two approaches depends on rule to be discovered. But in order to translate expert experience into fuzzy sets, the former is preferred over latter as it can directly optimize fuzzy rule-based system. For this, modifiers are used which act on atomic words and modify membership of fuzzy sets.

6. EXPERIMENTAL SET UP

The proposed GA approach is implemented using GALIB247 on a Pentium core 2 duo processor with Ubuntu release 9.10 as operating system. The performance of the suggested approach is validated on two real-valued datasets publically available at UCI [University of California at Irvine] machine learning repository and its corresponding site is ftp://ftp.ics.uci.edu/pub/machine-learning-databases/. The experimental datasets are described in Table 1.

Table 1: Description of datasets used for experimentation

Sr. No.	Name of the Dataset	No. of Examples	No. of Attributes	No. of Classes
1.	Balloon	16	4	2
2.	Poker	25010	11	10

Description of the Datasets:

1. Balloon

There are four data sets representing different conditions of an experiment.

All have the same attributes.

a. adult-stretch.data Inflated is true if age=adult or act=stretch

b. adult+stretch.data Inflated is true if age=adult and act=stretch

c. small-yellow.data Inflated is true if (color=yellow and size = small) or

d. small-yellow+adult-stretch.data Inflated is true if

(color=yellow and size = small) or (age=adult and act=stretch)

2.Poker

Each record is an example of a hand consisting of five playing cards drawn from a standard deck of 52. Each card is describe using two attributes (suit and rank), for a total of 10 predictive attributes. There is one Class attribute that describes the “Poker Hand”. The order of cards is important, which is why there are 480 possible Royal Flush hands as compared to 4.

7. PERFORMANCE EVALUATIONS

The GA statistics consist of two types of performance evaluations namely offline and online performance. Offline performance of genetic algorithm is taken as the average of best scores and average of worst scores whereas online performance gives the average of all scores over the number of generations.

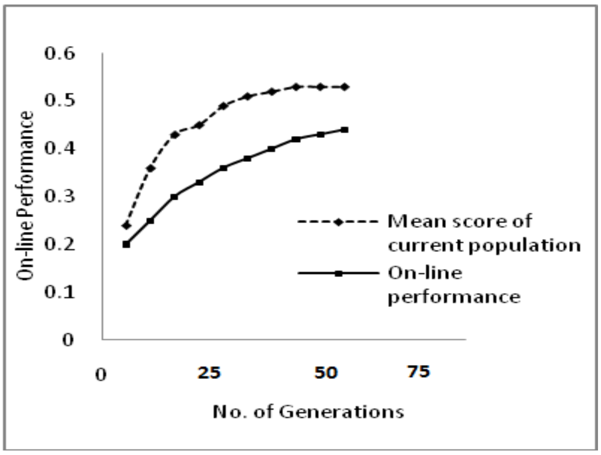

Fig. 8: Balloon Dataset

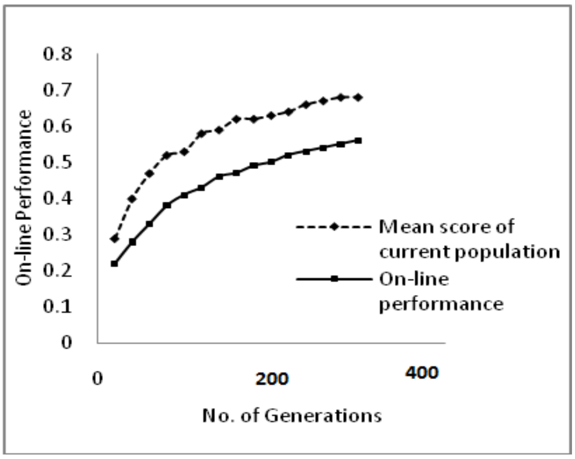

Fig. 9: Poker Dataset

As we are interested in an optimal rule set and not in a single best rule, we have considered plot of online performance and the mean score of current population with respect to number of generations for all the two datasets.

8. CONCLUSION

The knowledge representation is capable of handling uncertainty to decision making support systems. A genetic approach is proposed for the optimization of decision rules in Fuzzy Classification that can efficiently cope with large data which do not have crisp boundaries between them. The proposed scheme has flexible chromosome encoding and appropriate crossover and mutation operators are suggested. Keeping in view the basic constraints on classification, appropriate fitness function is formulated so as to make task of rule mining easier. This work has integrated fuzzy logic with genetic algorithm approach for the optimization of Decision Rules. The performance of proposed algorithm is tested across two real world datasets and the results are quite encouraging and have established the effectiveness of the proposed algorithms. The scheme provides a mechanism to discover concise and appropriate classification rules.

9. REFERENECES

[1] Freitas, “Data Mining and Knowledge Discovery with Evolutionary Algorithm”, Natural Computing Series, Springer-Verlag, New York, USA, 2002.

[2] Freitas, “A survey of Evolutionary Algorithms for Data Mining and Knowledge Discovery”, Advances in Evolutionary Computation Theory and Applications, Springer-Verlag, New York, USA, pp. 819-845, 2003.

[3] Kosko, “Neural Networks and Fuzzy Systems”, Prentice-Hall, Englewood Cliffs, NJ, 1992.

[4] Liu, W. Hsu and S. Chen, “Using General Impressions to Analyze Discovered Classification Rules”, In Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining (KDD-97), Newport Beach, CA, USA, AAAI Press, Portland Oregon, USA, pp. 31-36, 1997.

[5] Mota, H. Ferreira and A. Rosa, “Independent and simultaneous evolution of fuzzy sleep classifiers by genetic algorithms”, Proc. Genetic and Evolutionary Computation Conf. (GECCO-99), Morgan Kaufmann, pp. 1622-1629, 1999.

[6] T. Lin and C. S. G. Lee, “Neural-Network-based fuzzy logic control and decision system”, IEEE Trans. Comput., pp. 1320-1336, 1991.

[7] E. Goldberg, “Genetic Algorithms in Search, Optimization and Machine Learning”, Addison-Wesley Publishing Company, Inc. MA, New York, 1989.

[8] Walter and C. K. Mohan, “ClaDia: A fuzzy classifier system for disease diagnosis”, In Proc. Congress on Evolutionary Computation (CEC-2000), La Jolla, CA, USA, vol. 2, 2000.

[9] Noda, Alex A. Freitas and H.S. Lopes, “Discovering Interesting Prediction Rules with a Genetic Algorithm”, In Proc. Congress on Evolutionary Computation (CEC-99), Washington D.C., USA, pp. 1322-1329, July 1999.

[10] Eghbal G. Mansoori, Mansoor J. Zolghadri and Seraj D. Katebi, “SGERD: A Steady-State Genetic Algorithm For Extracting Fuzzy Classification Rules From Data,” IEEE Trans. Fuzzy Syst., vol. 16, no. 4, pp. 1061-1071, Aug. 2008.

[11] Rothlauf, Representations for Genetic and Evolutionary Algorithms, Physica-Verlag, Heidelberg, Germany, 2002.

[12] J. Klir and B. Yuan, “Fuzzy Sets and Fuzzy Logic: Theory and Applications”, Prentice-Hall, 1995.

[13] Ishibuchi, T. Nakashima and T. Murata, “Performance evaluation of fuzzy classifier systems for multi-dimensional pattern classification problems”, IEEE Trans. Syst.,Man, Cybern., Part B, vol. 29, pp. 601-618, 1999.

[14] Ishibuchi, T. Nakashima and T. Kuroda, “A Hybrid Fuzzy GBML Algorithm for Designing Compact Fuzzy Rule-Based Classification Systems”, In Proceedings of the 9th IEEE Int. Conf. Fuzzy Systems (FUZZ IEEE-2000), San Antonio, TX, USA, pp. 706-711, 2000.

[15] H. M. Chen and S. Y. Ho, “Designing an Optimal Evolutionary Fuzzy Decision Tree for Data Mining”, In Proc. of the Genetic and Evolutionary Computation Conference (GECCO-2001), San Francisco, California, USA, Morgan Kaufmann, San Francisco, California, USA, pp. 943-950, 2001b.

[16] D. Falco, A. D. Cioppa, A. Iazzetta and E. Tarantion, “An Evolutionary Approach for Automatically Extracting Intelligible Classification Rules”, Knowledge and Information Systems, Springer-Verlag, New York, USA, vol. 7, no. 2, pp. 179-201, 2005.