Hybrid Machine Learning Approach In Data Mining

Jyothi Bellary[#1], Bhargavi Peyakunta [*2], Sekhar Konetigari [#3]

[#]*Department of Computer Applications,*
Sri Venkateswara Universiy ,JNTU
17-154, Gandhi Road, Madanapalli.
[1]jyothibellary@rediffmail.com
sekar_sangam@yahoo.com

- *Department Of Computer Science,*
Madanapalle Institute of Technology and Science
Angallu.
pbhargavi18@yahoo.co.in

***Abstract*— In this paper we discuss various machine learning approaches used in mining of data. Further we distinguish between symbolic and sub-symbolic data mining methods. We also attempt to propose a hybrid method with the combination of Artificial Neural Network (ANN) and Cased Based Reasoning (CBR) in mining of data.**

***Keywords*— Data Mining, Machine Learning,Artificial Neural Networ (ANN), Case Based Reasoning(CBR), Hybrid Approach**

I. INTRODUCTION

Large collections of data are valuable resource from which potentially new and useful knowledge can be discovered through data mining. Data mining is an increasingly popular field including visualization, machine learning and other data manipulation and knowledge extraction techniques aimed at gain an insight into the relationships and patterns hidden in the data. For example hospitals and health care institutions are now well equipped with monitoring and other data collection devices, where data is collected and shared with other hospital information systems. Separated database or information system is now integrated as a large-scale information system. The increase in data volume causes difficulties in extortion useful information for decision support. The traditional manual data analysis has become insufficient and we need the technologies developed in the area of intelligent data analysis to analyze and acquire the required information from the large amount of data.

Intelligent Data Analysis

Intelligent data analysis (IDA) encompasses statistical, pattern recognition, machine learning, data abstraction and visualization tools to support the analysis of data and discovery of principles that are encoded within the data. IDA is largely related to Knowledge Discovery in Databases (KDD). The KDD is the non-trivial extraction of implicit and previously unknown but potentially useful information from data. The Knowledge Discovery Process consists of six stages.

- Data Selection: Selecting the right data for a KDD process.
- Cleaning: The removal of noise, errors and incorrect input from a database.
- Enrichment: New data is added to the existing data selection.
- Coding: Operations on a database to transform or simplify data in order to prepare it for a machine learning algorithm.
- Data Mining: Discovery phase of a knowledge discovery process.

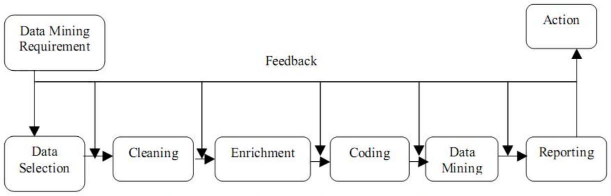

Figure 1. The Process of Knowledge Discovery

Data Mining Classifications

There are two different types of classification methods in data mining.

1) Symbolic methods

978-0-7695-3977-5/10 $26.00 © 2010 IEEE
DOI 10.1109/ICMLC.2010.57

2) Sub-symbolic methods

In problem solving it is important that a decision support system is able to explain and justify its decisions. Especially when faced with an unexpected solution of a new problem, the user requires substantial justification and explanation. Interpretation of known knowledge from past-solved cases is more important to understand and to accept the new solution for the user. In Symbolic data mining methods we can represent the solution by symbolic representation (for example, decision tree) form from data. In Sub-symbolic data mining method we can give the solution, but representation of how the solution learned is not possible.must be centered in the column. Large figures and tables may span across both columns. Any table or figure that takes up more than 1 column width must be positioned either at the top or at the bottom of the page.

Rule and tree induction, inductive logic, and case-based reasoning are symbolic data mining classification methods whereas artificial neural networks, induced based learning, Bayesian classifier are sub-symbolic data mining classification methods. In order to develop a hybrid machine learning approach, we limit our discussion to case-based reasoning and artificial neural networks.

Symbolic Method: Data Mining Through Case Based Reasoning

Case-based reasoning (CBR) uses the knowledge of past experience when dealing with new cases. A “case” refers to a problem situation. CBR relies on a database of past cases that has to be designed in the way to facilitate the retrieval of similar cases. CBR is a four-stage process

1. Given a new case to solve, find sets of similar cases are retrieved from the database.
2. The retrieved cases are reused in order to obtain a solution for a new case. This may be simply achieved by selecting the most frequent solution used with similar past cases, or, if an appropriate background knowledge or domain model exists, retrieved solutions may be adapted for a new case.
3. The solution for the new case is then checked by domain expert, and, if not correct, repaired using domain-specific knowledge or expert’s input. The specific revision may be saved and used when solving other new cases.
4. The new case, its solution, and any additional information used for this case that may be potentially useful when solving new cases are then integrated in the case database.

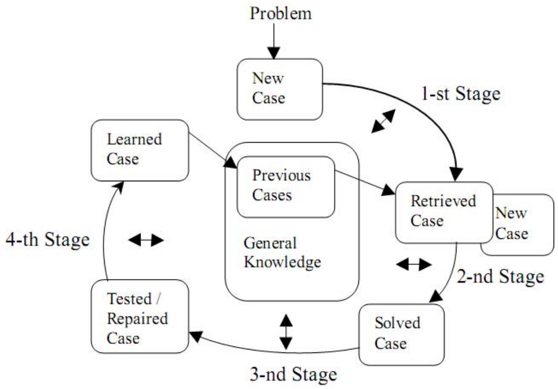

Figure 2. The CBR Cycle

Sub- Symbolic Method: Data Mining Through Neural Networks

Artificial neural networks imitate the structure of biological nervous system. It consists of a set of nodes. Input nodes to receive the input signals, output nodes to give the output signals and unlimited number of intermediate layers contain the intermediate nodes. There are various different architectures for neural networks and they each utilize different learning algorithms to perform tasks. Perceptron architecture consists of a simple three layered network with input unit, intermediate unit and output unit. It is very limited one but it can be used to perform simple classification tasks.

Multi layer Feedforward network (Backpropagation) consists of input and output unit with a set of input layers for hidden units. In its initial stage a backpropagation network has random weightings in its synapses. When we train the network, we expose it to a training set of input data. For each training instance, the actual output of the network is compared with the desired output that would give a correct answer, the weightings of the individual nodes and synapses of the network are adjusted. This process is repeated until the responses are more or less accurate. Once the structure of the network stabilizes, the learning stage is over, and the network is now trained and readies to categorize unknown input. It is a great improvement on the *perceptron* architecture. But it had various disadvantages.

1. They need an extremely large training set.
2. They do not provide with a theory about what they have learned. They are simply black boxes that give answers but provide no clear idea as to how they arrived at these answers.

To train the neural network system we have two different types of learning namely supervised and unsupervised learning. In supervised learning a set of input pattern and the target output are presented to the network repeatedly till the difference between the target output and the actual output of the network reaches a certain predetermined value. During the training process the difference between the actual output and the target output is compared. This difference is used to adjust the connection weights to the neurons in such away that the output of the network matches closely to the target output. *Perceptron* uses supervised training.

In unsupervised training output pattern for given input patterns is not required. The neural networks construct internal models that capture regularities in input pattern. The process of training the network consists of letting it discover salient features of the training set and using these features to group the inputs in to classes that it (the network) finds distinct.

Difference between supervised and unsupervised learning depends on whether the learning algorithm uses pattern-class information. Supervised learning assumes the availability of a teacher or supervisor who classifies the training examples into classes, where as unsupervised learning does not. Unsupervised learning must identify the pattern class information as a part of the learning process.

Proposed Hybrid Approach

Artificial neural networks can be used as black box classifiers lacking the transparency of generated knowledge and ability to explain the decisions. The back propagation network has non-transparent knowledge representation and thus in general can not easily explain their decisions. This is due to the large number of real valued weights, which influence the result. In some cases it is possible to extract symbolic rules from the trained neural network. However, the rules tend to be large and relatively complex. These rules are too complicated and hardly offer a useful explanation to a domain expert.

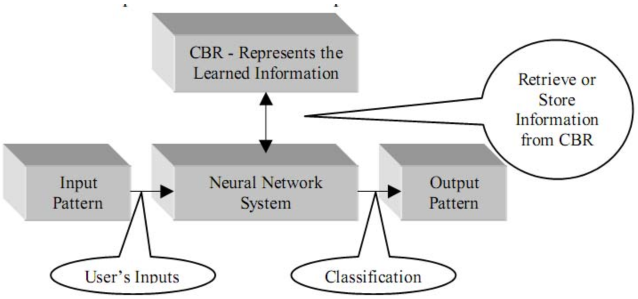
CBR - Represents the Learned Information
Retrieve or Store Information from CBR
Input Pattern
Neural Network System
Output Pattern
User's Inputs
Classification

Figure 3. Hybrid Model

One of the drawbacks of artificial neural network is that they do not provide the theory about how they learned. The end user is expecting the justification or explanation about the result from the system. So, our proposed hybrid approach, in figure 3, is to understand the representations formed by trained neural network to extract symbolic rules and represent that rule as a case using case-based reasoning method. Using cases user can understand the explanation and we can refer the past cases for the new problems occurred in the future.

REFERENCES

[1] Aamodt A. and Plaza E.: Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches: AI Communications, Vol. 7,1994.

[2] Craven M. W. and Shavlik J. W.: *Learning Symbolic Rules Using Artificial Neural Networks*, Machine Learning: Proceedings of the Tenth Intern. Conf., P.E. Utgoff (Ed.), Morgen Kaufmann, 1993.

[3] Lavrac N., Kerqvnou E. and Zupan B. (Eds.): *Intelligent Data Analysis in Medicine and Pharmacology*, 1997, Kluwer.