Global Journals LaTeX JournalKaleidoscope™
Artificial Intelligence formulated this projection for compatibility purposes from the original article published at Global Journals. However, this technology is currently in beta. *Therefore, kindly ignore odd layouts, missed formulae, text, tables, or figures.*

Classification Rules and Genetic Algorithm in Data Mining

Mr. Puneet Chadha[1]

[1] Department of Computer Science, D.A.V. College, Sector-10, Affiliated to Panjab University, Chandigarh-160010

Received: 12 December 2011 Accepted: 31 December 2011 Published: 15 January 2012

Abstract

Databases today are ranging in size into the Tera Bytes. It is an information extraction activity whose goal is to discover hidden facts contained in databases. Typical applications include market segmentation, customer profiling, fraud detection, evaluation of retail promotions, and credit risk analysis. Major Data Mining Tasks and processes include Classification, Clustering, Associations, Visualization, Summarization, Deviation Detection, Estimation, and Link Analysis etc. There are different approaches and techniques used for also known as data mining models and algorithms. Data mining algorithms task is discovering knowledge from massive data sets. In this paper, we are focusing on Classification process in Data Mining.

Index terms—

1 Introduction

atabases today are ranging in size into the Tera Bytes. It is an information extraction activity whose goal is to discover hidden facts contained in databases. Typical applications include market segmentation, customer profiling, fraud detection, evaluation of retail promotions, and credit risk analysis. Major Data Mining Tasks and processes include Classification, Clustering, Associations, Visualization, Summarization, Deviation Detection, Estimation, and Link Analysis etc. There are different approaches and techniques used for also known as data mining models and algorithms. Data mining algorithms task is discovering knowledge from massive data sets. In this paper, we are focusing on Classification process in Data Mining.

The management and analysis of information and using existing data for correct prediction of state of nature for use in similar problems in the future has been an important and challenging research area for many years. Information can be analyzed in various ways. Classification of information is an important part of business decision making tasks. Many decision making tasks are instances of classification problem or can be formulated into a classification problem, viz., prediction and forecasting problems, diagnosis or pattern recognition. Classification of information can be done either by statistical method or data mining method.

2 II.

3 Classification

Classification is a form of Data Analysis that can be used to construct a Model, which can be further used in future to predict the Class Label of new Datasets. Learned Decision Trees can be used for Classification. Given a tuple X for which the associated Class Label is unknown, the attribute values of the tuple are tested against the decision tree. A path is traced from the root to a leaf node, which holds the class prediction for that tuple. Decision trees can be easily converted to Classification Rules.

ii. Bayesian Classification Classifiers made using Bayesian Classification can predict the probability that a given tuple belongs to a particular Class.

Baye's Theorem: Using Baye's theorem we can predict Posterior Probability, P(H,X) from P(H),P(X|H) and P(X).Here X is a data tuple. Baye's Theorem is D P(H|X)=P(X|H) P(H)

4 P(X)

Where H ->Hypothesis such as that the data tuple X belongs to a specified Class C P(H|X) ->Probability that hypothesis H exists for some given values of X's attribute. P(X|H) ->Probability of X conditioned on H P(H) ->Probability of H P(X) ->Probability of X Let D be the Training data set. Let X be a tuple. If a Rule is satisfied by X, the rule is said to be triggered. Rule fires by returning the Class Prediction.

5 Rule Extraction from a Decision Tree

To Extract Rules from a Decision Tree, One Rule is created for each path from the root to a leaf node. Each Splitting Criterion along a given path is Logically ANDed to form the Rule Antecedent(IF part).The Leaf node holds the Class Prediction, forming the Rule Consequent(Then part).

Rule Induction using a Sequential Covering Algorithm

Here the Rules are Learned Sequentially, One at a time (for one Class at a time) directly from the Training Data (i.e without having to generate a Decision Tree first) using a Sequential Covering Algorithm.

6 iv. Classification by Backpropagation

Backpropagation is the most popular Neural Network Learning Algorithm. Neural Network is a set of connected input/output units, in which each connection has a weight associated with it. During the learning phase, the Network learns by adjusting the weights so as to be able to predict the Correct Class Label of the Input Tuples. Backpropagation performs on Multilayer Feed-Forward Neural Network. Several techniques have been developed for the Extraction of Rules from Trained Neural Networks. These factors contribute toward the usefulness of Neural Networks for Classification and Prediction in Data Mining.

7 v. Support Vector Machines

Support Vector Machines is a promising new method for the Classification of both Linear and Non Linear Data. Support Vector Machine is an algorithm that uses a Non Linear Mapping to transform the original training data into a higher dimension. Within this new dimension, it searches for Linear Optimal Separating Hyperplane (that is, a decision boundary separating the tuples of one class from another).The SVM finds this Hyperplane using Support Vectors (essential training tuples) and Margins (defined by the support vectors).

8 vi. Associative Classification

In Associative Classification, Association Rules are generated and analyzed for use in Classification. The general idea is that we can search for Strong Associations between Frequent Patterns (conjunctions of attribute-value pairs) and Class Labels. Because Association Rules explore highly confident Associations among Multiple Attributes, this approach may overcome some constraints introduced by Decision-Tree Induction which considers only one attribute at a time. In

9 III.

10 Genetic algorithm

A genetic algorithm (GA) is a search technique used in computing to find exact or approximate solution to optimization and search problems. Genetic algorithms are categories as global search heuristics.

Genetic algorithms are a probabilistic search and evolutionary optimization approach. Genetic algorithms are inspired by Darwin's theory about evolution. Solution to a problem solved by genetic algorithms is evolved.

Algorithm is started with a set of solutions (represented by chromosomes) called population. Solutions from one population are taken and used to form a new population. This is motivated by a hope, that the new population will be better than the old one. Solutions which are selected to form new solutions (offspring) are selected according to their fitness -the more suitable they are the more chances they have to reproduce. This is repeated until some condition (for example number of populations or improvement of the best solution) is satisfied. The construction of a classifier requires some parameters for each pair of attribute value where one attribute is the class attribute and another attribute is selected by the analyst. These parameters may be used as intermediate result for constructing the classifier. Yet, the class attribute and rest all attributes that analyst considers as relevant attributes must be the attributes of the tables that might be used for analysis in future. Hence, attribute values of class attribute are always frequent. When pre-computing the frequencies of pairs of frequent attribute values, the set of computed frequencies should also include the frequencies that a potential application needs as values of the class attribute and relevant attribute are typically frequent.

A framework for Genetic Algorithm to be implemented for Classification is [1]


OPEN
ASSOCIATION
OF RESEARCH
SOCIETY, USA
RESEARCH | DIVERSITY | ETHICS

Figure 1:

Naive Bayesian Classification
Let d be a training set of tuples and
X=(x 1 ,x 2 ,x 3 ??.x n) are the n attributes.
Let there be m classes
C 1 ,C 2 ?????..C m .Naive Bayesian Classifier predicts
that tuple X belongs to the class Ci if and only if
P(Ci|X) > P(Cj|X) for 1<=j<=m, j = i
i.e X belongs to the Class having the Highest
Posterior Probability.
Bayesian Belief Networks
The Naive Bayesian Classifier assumes Class
Conditional Independence, but in practic
dependencies can exist between Attributes (variables)
or the tuple ie a particular categorization can depend on
the values of two attributes.
Bayesian Belief Networks specify Joint
Conditional Probability Distributions.
A Belief Network is defined by two components-
A Directed Acyclic Graph and a Set of Conditional
Probability Tables.
Each Node in the Graph correspond to actual
attributes given in the data or to hidden variables. Each
Arc represents a Class Label.
The Classification Process can return a
Probability Distribution that gives the Probability of each
Class.
iii. Rule Based Classification
Rule Based Classifiers uses a set of IF-Then
Rules for Classification.
IF Condition Then Conclusion.

Figure 2:

[Weber and Mateas ()] *A Data Mining Approach to Strategy Prediction*, Ben G Weber , Michael Mateas . 2009. 2009. IEEE. p. .

[Luo ()] *Advancing Knowledge Discovery and Data Mining" Knowledge Discovery and Data Mining*, Qi Luo . 2008. 2008. 2008.

[Kantarc?oglu et al. ()] 'Classifier evaluation and attribute selection against active adversaries'. Murat Kantarc?oglu , Xi , Bowei , Chris Clifton . *Data Min Knowl Disc* 2011. 2011. 22 p. .

[Piatetsky ()] 'Data mining and knowledge discovery 1996 to 2005: overcoming the hype and moving from "university" to "business" and "analytics'. Gregory Piatetsky . *Data Min Knowl Disc* 2007. 2007. 15 p. .

[Das et al. ()] 'Data Quality Mining using Genetic Algorithm'. Sufal Das , Saha , Banani . *International Journal of Computer Science and Security* 2009. (3) . (IJCSS))

[Kriegel et al. ()] 'Future trends in data mining'. Hans-Peter Kriegel , Karsten M Borgwardt , Kröger , Peer . *Data Min Knowl Disc* 2007. 2007. 15 p. .

[Weiss et al. ()] 'Guest editorial: special issue on utility-based data mining'. Gary M Weiss , Bianca Zadrozny , Maytal Saar-Tsechansky . *Data Min Knowl Disc* 2008. 2008. 17 p. .

[Kamble (2010)] 'Incremental Clustering in Data Mining using Genetic Algorithm'. Atul Kamble . *International Journal of Computer Theory and Engineering* 2010. June, 2010. 2 (3) .