Impact of Data Mining in Drought Monitoring

Anil Rajput [1] , Ritu Soni[2], Ramesh Prasad Aharwal [3], Rajesh Sharma[4]

[1]Principal, Bhabha Engineering Research Institute-MCA,

Bhopal-26, India

[2]Asstt Prof., Department of Computer Science, N.R.I. Group of Institute

Bhopal (M.P.) India
[3]Asstt. Prof., Department of Mathematics and Computer Science, Govt. P.G. College

Bareli (M.P.), India

[4]Asstt Prof. Shree Institute of Science and Technology, Bhopal (M.P.),India

Abstract

This paper focuses on association rule mining and decision tree classification in the rainfall and temperature data. We have used Apriori association rule mining algorithm with data mining tool WEKA. We have to try to generate Association rules and decision tree model.

Keywords: Data Mining, Classification, Drought, Association Rule.

1. Introduction

Data mining is the process of discovering actionable and meaningful patterns, profiles and trends by sifting through data using pattern recognition technologies such as neural networks and machine learning and genetic algorithms. Author [5] also mentions that data mining tools perform data analysis and may uncover important data patterns, contributing greatly to business strategies, knowledge bases, and scientific, social and medical research. Data mining has evolved into an important and active area of research because of theoretical challenges and practical applications associated with the problem of discovering previously unknown knowledge from very large real-world databases. It comes from the idea that large databases can be viewed as data mines containing valuable information that can be discovered by efficient knowledge discovery techniques. Han and Kamber [5] also said that KDD process consists of an iterative sequence of steps. The additional steps in the KDD process, such as data preparation, data selection, data cleaning, incorporation of appropriate prior knowledge, and proper interpretation of the result of mining are essential to ensure that useful knowledge is derived from the data. KDD is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [3].

2. Data Mining Techniques

The main goals of data mining are prediction and description identifying patterns and relationships in data, and generates new knowledge about the data set. The basic data mining tasks are divided as follows:

Predictive: Classification, Regression, Time series analysis, Prediction

Descriptive: Clustering, Summarization, Association rules, Sequence discovery (sequential analysis), Estimating the model

2.1 Classification

The model can be represented as classification rules in the form of "if...Then", decision trees, neural networks. A training set and a test set will be used Han and Kamber[5]. Some examples would be classifying countries based on climate, classifying credit risks. Commonly used techniques for this data mining function are:

- Discriminant analysis.
- Rule induction methods.
- Neural networks.
- Decision tree learning.
- K Nearest Neighbour.
- Case-based reasoning.
- Genetic algorithms.

2.2 Association Rule Mining

Association rule mining is one of the best studied models for data mining. In recent years, the discovery of association rules from databases is an important and highly active research topic in the data mining field. Association rule mining searches for interesting association or correlation relationships among items in a given dataset. Association rule mining algorithm was first proposed by Agrwal et al [1] in 1993 which is used to discover patterns in transactional databases. The idea to discover association rules is also named "market basket analysis" because it looks for associations among items that a customer purchases in a retail shop.

The Apriori algorithm is proposed by Agrawal et al [1], which is used for finding frequent itemsets. It generates the candidate itemsets in one pass through only the itemsets with large support in the previous pass, without considering the transactions in the database. An itemset with support larger than or equal to the minimum support is called a frequent itemset. The idea of the Apriori algorithm lies in the "downward-closed" property of support, which means if an itemset is a frequent itemset, then each of its subsets is also a frequent itemset.

2.3 Decision tree

Decision trees are often used in classification and prediction. The models produced by decision trees are represented in the form of tree structure. A leaf node indicates the class of the examples. The instances are classified by sorting them down the tree from the root node to some leaf node [3]. The algorithm commonly used to construct decision tree is known as recursive partitioning. This paper will focus on using J48 which is the advanced of C4.5 in building the decision tree. Decision trees represent a supervised approach to classification.

3. Use of Data Mining to identify drought

As large collections of historical data are essential in identifying relationships among different climatic and meteorological parameters, proper handling of data and quick processing becomes necessary. This highlights that efficient data management and information extraction is important in drought research. The difficulty of handling huge amount of data, understanding the complex results derived from atmospheric phenomenon and extracting the spatial and

temporal pattern of drought can be resolved with data mining techniques

4. WEKA

The WEKA software was developed in the University of New Zealand. A number of data mining methods are implemented in the WEKA software. Some of them are based on decision trees like the J48 decision tree, some are rule-based like ZeroR and decision tables, and some of them are based on probability and regression. We have used Weka [6] for our experiments.

5. Experimental Datasets

In this paper we have used Rainfall and Tempreture data of Sagar District M.P. from 1997 to 2010.

5.1 Data Preparation

The process of data cleaning and preparation is highly dependent on the specific data mining algorithm and software chosen for the data mining task. The researcher attempted to prepare the data according to the requirements of the selected data mining software, Weka and selected data mining algorithm, apriori. Weka is multi-functional data mining software. The major data mining functions incorporated in the software are data preprocessing, classification, association, clustering and visualizing input and output. Apriori is the only association rule algorithm implemented in Weka.

5.2 Screen Shots during Experiment

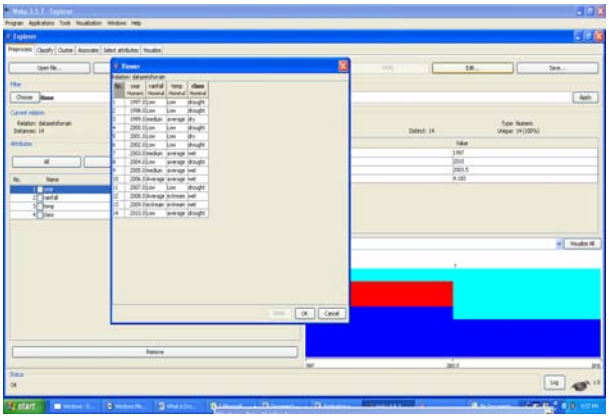

Figure 1 WEKA Explorer

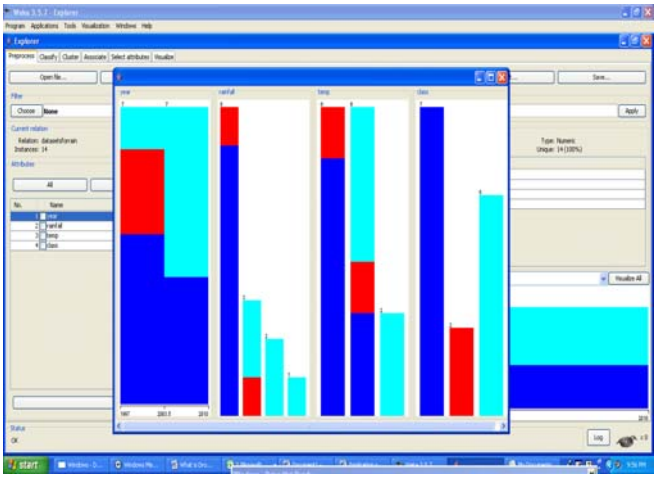

Figure 2 Visualization of attributes

6. First Experiment for generating precious association rules from dataset

=== Run information ===

Scheme: weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

Relation: datasetsforrain-weka.filters.unsupervised.attribute.Remove-R1

Instances: 14

Attributes: 3

rainfall

temp

class

=== Associator model (full training set) ===

Apriori

=======

Minimum support: 0.1 (1 instances)

Minimum metric <confidence>: 0.9

Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 10

Size of set of large itemsets L(2): 18

Size of set of large itemsets L(3): 8

6.1 Best Rules found

1. class=drought 7 ==> rainfall=Low 7 conf:(1)

2. temp=Low 6 ==> rainfall=Low 6 conf:(1)

3. temp=Low class=drought 5 ==> rainfall=Low 5 conf:(1)

4. rainfall=medium 3 ==> temp=average 3 conf:(1)

5. rainfall=Average 2 ==> class=wet 2 conf:(1)

6. temp=extream 2 ==> class=wet 2 conf:(1)

7. temp=average class=drought 2 ==> rainfall=Low 2 conf:(1)

8. rainfall=Low temp=average 2 ==> class=drought 2 conf:(1)

9. rainfall=medium class=wet 2 ==> temp=average 2 conf:(1)

10. rainfall=extream 1 ==> temp=extream 1 conf:(1)

7. Second Experiment for Decision Tree

=== Run information ===

Scheme: weka.classifiers.trees.J48 -U -M 2

Relation: datasetsforrain-weka.filters.unsupervised.attribute.Remove-R1

=== Classifier model (full training set) ===

J48 unpruned tree

----

rainfall = Low: drought (8.0/1.0)

rainfall = medium: wet (3.0/1.0)

rainfall = Average: wet (2.0)

rainfall = extream: wet (1.0)

Number of Leaves : 4

Size of the tree : 5

== Evaluation on training set ===

=== Summary ===

Correctly Classified Instances 12 85.7143 %

Incorrectly Classified Instances 2 14.2857 %

=== Detailed Accuracy By Class ===

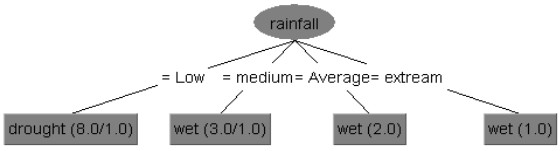

Figure 3: Generated decision tree from WEKA

8. Conclusion

Initially the experiments have been conducted on the whole dataset. Graphical Visualizations have been performed in order to make it easier to understand the data

itself. This is shown in figure 1 and 2. The reason for selecting the J48 decision tree algorithm is because the algorithm has the ability to handle data with missing attribute values better than ID3 decision tree algorithm result of first and second experiment as shown above in the section 6.1 and section 7.

9. References

[1] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases," Proc. Of ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'93) 1993, pp 45-51, Washington, DC.

[2] U. Fayyad., G.Piatetsky-Shapiro and P.Smyth ,Advances in Knowledge discovery and data mining, AAAI/MIT press 1996 .

[3] Fahad Shahbaz Khan, Rao Muhammad Anwer, Olof Torgersson, and Göran , Data Mining in Oral Medicine Using Decision Trees, Falkman Proceedings of world Academy of science, Engineering and technology volume 27 february 2008.

[4] **Gregory Piatetsky-Shapiro** (1991). Knowledge Discovery in Real Databases: A Report on the IJCAI-89 Workshop. AI Magazine, 1991, 68-70.

[5] J. Han and M. Kamber, Data Mining : Concept and Techniques , first edition, Harcoart India private Limited (2001).

[6] Weka Website: Data Mining Software in Java, http://www.cs.waikato.ac.nz/ml/weka/