Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 10, October 2014, pg.732 – 740

RESEARCH ARTICLE

Coal Mining Data for Data Mining Techniques

V.Renugadevi, M.Phil. (Computer Science) **K.Rajeswari,** M.Sc., M.Phil.

Research Scholar, Assistant Professor in Computer Science

Vivekanandha College for Women, Unjanai, Tiruchengode, India

Abstract: Classification is an important problem in data mining. Given a database of records, each with a class label, a classifier generates a concise and meaningful description for each class that can be used to classify future records whose classes are unknown. A number of popular classifier exists like naïve bays classifier, Neural Network, SVM classifier, CARD etc. In this paper we applied classification algorithm for coal mines dataset. We also discussed decision tree, nearest neighbor classifier, NN classifier for prediction of class label.

Keywords: Data Mining algorithm, C4.5 algorithm, Naïve Bayes Algorithm, Intelligent Data mining for coal data

1. Introduction

Classification is an important problem in data mining. Under the guise of supervised learning, classification has been studied extensively by the machine learning community as a possible solution to the "knowledge acquisition" or "knowledge extraction" problem. The input to the classifier construction is a training set of records, each of which is tagged with a class label. A set of attribute values defines each record. Attributes with discrete domains are referred to as categorical, while those with ordered domains are referred to as numeric. The goal is to induce a concise model or description for each class in terms of the attributes. The model is then used by the classifier to classify (i.e., assign class labels to) future records whose classes are unknown. It thus, reduces outlier problem. The classification model of dataset

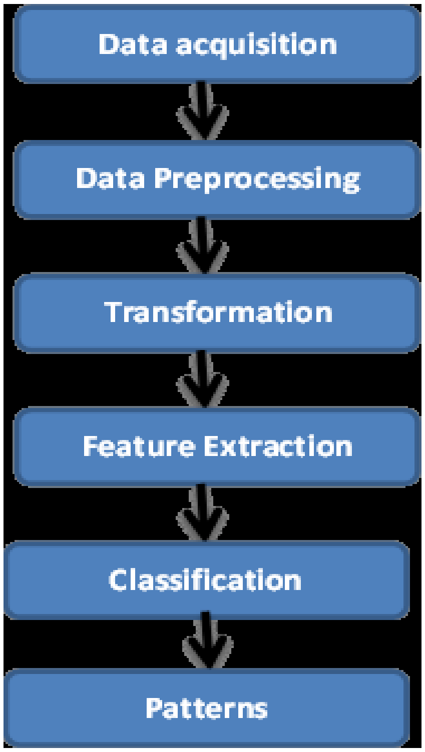


Fig. Data mining process

1.1. Data Acquisition process

The data is acquired from coal mines where the blasting processes for mines are performed. The spacing (M) denotes the space between holes. The class label Boulder denotes after the blasting the size of boulder produced in blasting.

1.2. Data Pre-processing

Data pre-processing is an often neglected but important step in the data mining process. The phrase “Garbage in, Garbage Out” is applicable in data mining and machine learning. Data gathering methods are loosely controlled, resulting in out of range values, impossible data combination, missing values. Analyzing data that has not been carefully screen for such problems can produce misleading results. The representation and quality of data is first and foremost before running an analysis. In our dataset we filled the missing value by mean of whole data set of specified attributes. Some data values the range is given, we computed the mean value and it is replaced by mean of the data e.g. for stemming (M) the range is given 3.5 to3.75 that data is replace by 3.625.

1.3. Data Transformation

In data transformation, the data are transformed or consolidated into forms appropriate for mining. Data transformation processes are follows.

Normalization- where the attribute data are scaled in the range such as -1.0 to +1.0.

1.4. Feature Selection

The main idea of feature selection is to choose a subset of input variables by eliminating features with little or no predictive information. Feature selection can significantly improve the comprehensibility of the resulting classifier models and often build a model that generalizes better to unseen points. In machine learning and statistics, feature selection also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features for use in model construction. The central assumption when using a feature selection technique is that the data contains many redundant or irrelevant features. Redundant features are those which provide no more information than the currently selected features, and irrelevant features provide no useful information in any context. Feature selection techniques are a subset of the more general field of feature extraction. Feature extraction creates new features from functions of the original features, whereas feature selection returns a subset of the features. Feature selection techniques are often used in domains where there are many features and comparatively few samples (or data points).

2. Classification

There are various classifiers to classify the data like NN, SVM, Decision tree and others.

2.1. Nearest Neighbour

Nearest neighbour (NN) also known as Closest Point Search is a mechanism that is used to identify the unknown data point based on the nearest neighbor whose value is already known. It has got a wide variety of applications in various fields such as Pattern recognition, mage databases, Internet marketing, Cluster analysis etc. A fast algorithm that finds the nearest neighbor (NN) of an unknown sample from a design set of labeled samples is proposed. This algorithm requires a quite moderate preprocessing effort and a rather excessive storage, but it accomplishes substantial computational savings during classification. The performance of the algorithm is described and compared to the performance of the conventional one.

The NN algorithm can also be adapted for use in estimating continuous variables. One such implementation uses an inverse distance weighted average of the k-nearest multivariate neighbors. This algorithm functions as follows:

1. Compute Euclidean or Mahalanobis distance from target plot to those that were sampled.

2. Order samples taking for account calculated distances.

3. Choose heuristically optimal k nearest neighbor based on RMSE done by cross validation technique.

2.2. Decision Tree

Decision Tree Classifier a decision tree is a class discriminator that recursively partitions the training set until each partition consist entirely or dominantly of examples from one class. Each leaf node of the tree contains one or more attributes and determines how the data is partitioned.

A decision tree classifier is built in two phases growing phase followed by pruning phase. In growth phase the tree is built by recursively partitioning the data until each partition is either “pure” or sufficiently small. To prevent over fitting: The MDL principle is applied to prune the tree build in the growing phase and make it more general.

Using the fewest number of bits. To find the sub tree of the tree that can be encoded with the least number of bits.

The algorithm for building tree.

```
Procedure buildTree (S)
Initialize root node using dataset S
Initialize queue Q to contain root node
While Q is not empty do {
    dequeue the first node N in Q
    if N is not pure {
        for each attribute A
            Evaluate splits on attributes A
        Use best split to split node N into N1 and N2
        Append N1 and N2 to Q
    }
}
```

2.3. Bayesian Network

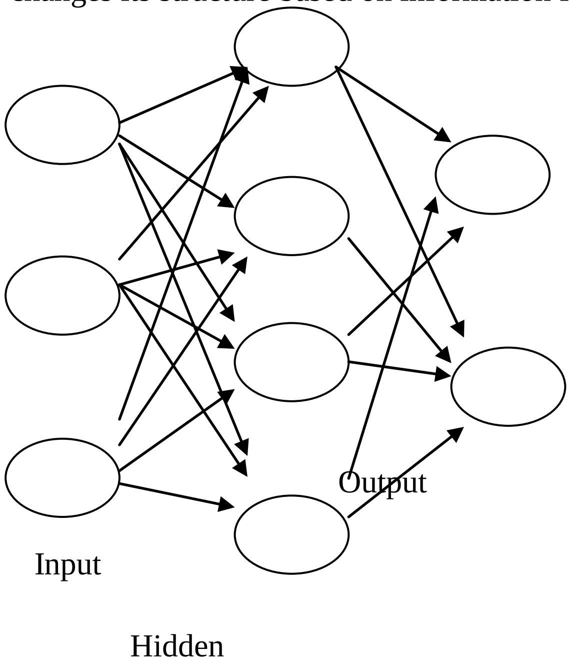
Bayesian network (BN) is also called belief networks. A BN is a graphical representation of probability distribution. It belongs to the family of probabilistic graphical models .This BN consist of two components. First component is mainly a directed acyclic graph (DAG) in which the nodes in the graph are called the random variables and the edges between the nodes or random variables represents the probabilistic dependencies among the corresponding random variables.

Second component is a set of parameters that describe the conditional probability of each variable given its parents. The conditional dependencies in the graph are estimated by statistical and computational methods. This prior expertise about the structure of Bayesian network algorithm work as follows.

1. Declare that a node is root node.
2. Declare that a node is leaf node.
3. Declaring that a node has direct effect of another node.
4. Declaring that a node is not directly connected to another node.
5. Declaring that two nodes are independent, giving a condition set.
6. Providing partial ordering among the nodes.

2.4. Artificial Neural Network

Artificial Neural Network (ANN) is a computational model based on biological neural network. ANN also called Neural Network [ANN]. It contains interconnected group of artificial neurons and processes the information by a connectionist approach. ANN is an adaptive system because it changes its structure based on information flow during the learning phase.



Architecture of ANN

Actual algorithm for a 3-layer network (only one hidden layer):

```
Initialize the weights in the network (often randomly)
Do
For each example e in the training set
O = neural-net-output(network, e) ; forward pass
t = teacher output for e
1) Calculate output , y=x1w1+x2w2,
2) Calculate error (T - O) at the output units E= $(t-y)^2$
3) Compute delta_wh for all weights from hidden layer to output layer ; backward pass
4) Compute delta_wi for all weights from input layer to hidden layer; backward pass continued.
5) Update the weights in the network
```

The different classification techniques and their comparative charts.

DIFFERENT CLASSIFICATION TECHNIQUES

Method	**Generative /Discriminative**	**Loss Functions**
K-Nearest Neighbour	Discriminative	-log P(X,Y) or Zero one
Decision tree	Discriminative	Zero –One loss
Bayesian Network	Generative	-log O(X,Y)
Neural Network	Discriminative	Sum-Squared Error

3. Proposed Work

3.1. Support Vector Machine Classifiers:

SVMs, also support vector networks are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

A good classification method:

Avoid underfitting: small training error

Avoid overfitting: small testing error

1. Conduct simple scaling on the data

2. Consider RBF kernel K(x, y) = e

−γkx−yk

3. Use cross-validation to find the best parameter C and γ

4. Use the best C and γ to train the whole training set

5. Test

4. Experimental Result

The experiments were perform in weka machine learning software in pentium IV machine with 1 GB RAM. No other application are running while performance computation. The dataset contain

47 instances, 6 attributes i.e. Number_of_holes, Burden, Depth, Primer, SME and Class. In test mode:10-fold cross-validation are perform. All the numeric data is converted to categorical data all <50 values denotes „Effective‟ Boulders and remaining are converted as „Not-Effective‟ Boulders.

5. Conclusion

In this paper we applied data mining process in coal mining data. The mining process start with data preparation is an important issue for data mining, as real world data tends to be incomplete, noisy and inconsistent. Data preparation includes data cleaning, data integration, data transformation and data reduction. In order to mine the effective size of Boulders (Nos.) the attribute effective are analysed. Decision tree model gives 87% of accuracy to correctly predict the class label where as Neural Network model predict 84% correct class label. we try to implement support vector machine classifier in order to improve the accuracy of classifier.

References

[1] YongSeog Kim, W. Nick Street, and Filippo Menczer. Feature Selection in Data Mining.

[2] http://en.wikipedia.org/wiki/Feature_selection

[3]. Ms. Aparna Raj, Mrs. Bincy G,, Mrs. T.Mathu. Surveyon Common Data Mining Classification Techniques. International Journal of Wisdom Based Computing, Vol. 2(1), April 2012

[4] J. Ross Quinlan. C4.5: Programs for Machine Learning. Morgan Kaufman, 1993.