International Journal of Security and Its Applications
Vol.9, No.10 (2015), pp.425-432
http://dx.doi.org/10.14257/ijsia.2015.9.10.39

Application of Data Mining based on Classifier in Class Label Prediction of Coal Mining Data

Haixu Xi, Dan Guo and Hongjin Zhu

Jiangsu University of Technology, No.1801, Zhongwu Avenue, Changzhou City, Jiangsu Province, China
jsut_xhx@126.com, zhuhongjin0427@hotmail.com

Abstract

For issue that coal mining data tends to be incomplete, noisy and inconsistent, some popular classifiers are applied to predict class label of coal mining dataset. Noise and bad points are rejected from coal mining data which will be exchanged to input format suitable for mining. Then different classifiers are used to classify class label after extracting features. In the end, classification results are analyzed and knowledge assimilation is done. Experiment results show that decision tree model gives 88% of accuracy to correctly predict class label whereas neural network model predicts 85% correct class label. This research provides a powerful class label prediction tool as well as increasing knowledge of data classification models.

Keywords: *Coal mining data; Data mining; Class label prediction; Naïve bays classifier; Artificial neural network; Decision tree model*

1. Introduction

Assortment [1, 2] is also called supervised learning, and has important significance in the Data Mining [3, 4], which is widely used in machine learning [5, 6], because it may be a solution of *knowledge gain* or *knowledge extract* problem. Assortment's objective is basing on attribute value to set up a simple model or describe for every class, with this model classifier can classify for the unknown class of future record, to reduce to generate the problem of outlier [7].

This article is applied to Coal Mining Data with classifier of the Data Mining Process [8]. In the Coal Data Mining Process, because of the problems of Coal mining data's incomplete, noisy, inconformity and so on, which make data preparation become an very important problem, and the data preparation includes data cleansing, data integration, data conversion and data reduction. This study not only has deepened the understanding of Data Class Label Prediction Classifier, but also provided strong Class Label Forecasting Tool for Coal Mining.

2. Coal Data Mining

2.1 Coal Data Mining Process

All data acquired from performing the blasting process. Interval (M) means the spacing between holes, Class label boulders means stone size after blasting. The whole process of the Stone size is in the Figure 1.

Aim at Coal Mining Data of Coal Data Mining, it mainly follows these steps:

1) Data Acquisition

Aim at Coal Mining Data, we will extract data, which has closely connection with mining object from Spatial Data Warehouse, and build Spatial Data Warehouse to store.

ISSN: 1738-9976 IJSIA

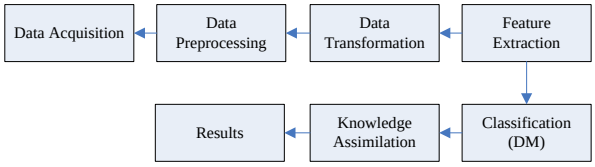


Figure 1. Coal Data Mining Process

2) Data Preprocessing

The main duty of the Data Preprocessing is screening out useful data, filtering out the useless or affected the mining data as far as possible, after preprocessing operating; the noise and dead pixel will be rejected in the data.

3) Data Transformation

The operations of the Data Transformation make the data organize Data Mining algorithm of input format.

4) Feature Extraction

The main duty of the Feature Extraction [9] is choosing an input variable of subset by eliminating few features, or which has no predictive information in the input variable.

5) Classification (Data Mining)

The key to Classification's stage is classifying the data, which after feature extracting with all sorts of classifiers.

6) Results

This stage is explaining, assessing and revealing the mining results.

7) Knowledge Assimilation

The Knowledge Assimilation means making the knowledge from Data Mining integrate and insert the Coal Mining Data, to exert the Data Mining knowledge of the commercial value.

2.2 Data Acquisition

Coal Mine Data Set is the Table 1; it is mainly made up these parts: Interval (m), Pressure (m), Depth (m), and Stemming (m). The second part is explosive term; it has the next three parts: Booster (kg), Primer (kg), and Stone Size (Nos.).

Table 1. Coal Mine Data Set

Inte-rval (m)	Pre-ssure (m)	Dept-h(m)	Stemming (m)	Boos-Ter (kg)	Pri-mer (kg)	Ston-e Size(Nos.)
5	3.25	6.25	3	10.5	0	1340
5.75	3.5	7	3.5	15.3	294	420
5.75	3.5	7	3.5	10.5	0	1200
6.25	4	8.25	3.5	35.2	0	625
5	3.25	6.25	4.0	13.2	0	1075

3. Classifier

Now there are a lot of classifiers to classify the data, such as the nearest neighbor classifier, the SVM and the decision tree, etc.

3.1 Nearest Neighbor Classifier

Nearest neighbor (NN) classifier [10] is also known as recently some query, it is a kind of mechanism which can identify the location data points the based on the known value's nearest neighbor. Nearest neighbor classifier have a wide range of applications in many areas, such as pattern recognition, mage database, Internet marketing, cluster analysis, etc.

NN algorithm can also be used to estimate the continuous variable. A similar implementation uses the K-nearest multiple neighbor inverse distance weighted average. This algorithm features as follows:

1. Calculate the Euclidean or markov distance from the target domain to sampling domain.
2. Arrange samples to calculate the distance.
3. Choose the optimal K neighbor through cross-validation technique selection on the basis of the RMSE.

3.2 Decision Tree

In Decision Tree (DT) [11] classifier, a Decision Tree is a kind of discriminator. It divides into a training set recursively until each part is a whole or a sample of the dominant class. Each leaf node of the tree contains one or more attributes and determine the data how to divide.

A decision tree consists two phases: growing phase, phase of pruning. In the growth stage, it partitions data recursively until each part is the "pure" or small enough, eventually it establish a spanning tree.

Tree-building algorithm is as follows:

```
Procedure builds Tree (S)
1)  Initialize root node using dataset S
2)  Initialize queue Q to contain root node
3)  While Q is not empty do {
4)    Dequeue the first node N in Q
5)    If N is not pure {
6)      For each attribute A
7)      Evaluate splits on attributes A
8)      Use best split to split node N into N1 and N2
9)      Append N1 and N2 to Q
10)   }
11) }
```

3.3 Bayesian Network

Bayesian network (BN) [12] is also called belief networks, BN is a graphical representation of a probability distribution, and it is a kind of probability graph model. BN has two parts. The first part is mainly a directed acyclic graph. The nodes in the graph have become random variables. Random variables probability of the node or random variables represents probabilistic dependency between the random variables. The second part is a set of parameters, it is used to describe each of the variables of a given parent conditional probability. The conditional dependencies in the diagram can estimate through the statistics and calculation method [13, 14].

The above description algorithm structure of Bayesian network is as follows:

1. Declare a root node.

2. Declare a leaf node.
3. Declare a direct influence on the node to another node.
4. Declare no direct influence on the node to another node.
5. Declare two nodes were independent of each other, given a set of conditions.
6. Give partial order between the nodes.

3.4 Artificial Neural Network

Artificial Neural Network (ANN) is also called neural networks. It contains artificial neurons and manages related groups of information through contact method. ANN changes its structure based on learning phase flow, so it is a kind of adaptive system.

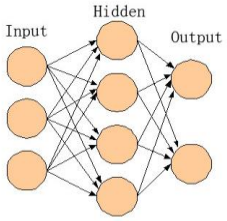


Figure 2. ANN Structure

Three layer network (only one hidden layer) the actual algorithm is as follows:
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,
2) Calculate error (T - O) at the output units
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
Comparison between different classifications technologies are shown in Table 2.

Table 2. The Comparison of the Different Classifications

Classification technology	Discriminative/Generation	Loss function	Parameter estimation algorithm
K-nearest neighbor	Discriminative	-log P(X, Y) or Zero one loss function	All date in forecast
Decision Tree	Discriminative	Zero one loss function	C4.5
Bayesian Network	Generation	-log O(X,Y)	Variable selection
neural network	Discriminative	WGSS	Forward-prop-agating

4. Experimental Result

The experiment runs on weka machine learning software in the 4 GB RAM Pentium IV machines. During the experimental process, no other program is running.

The experimental data set contains Forty-seven instance, five attributes values (hole number, pressure, depth, primer, class). Test mode is as follows: performing 10 times cross validation, all digital data into classification data, that is less than 50 values for efficient stone size, the rest is invalid stone size. Data is from January 2012 to December, the experimental results are such as Table 3, 4, 5, and 6.

Table 3. Decision Tree Models Results

Correctly classified instances	41	88.234%
Error classified instances	6	12.766%
Mean Absolute Error	0.2365	
Root Mean Squared Error	0.337	

Table 4. Decision Tree's Confusion-matrix

	Valid	Invalid
Valid	0	6
Invalid	0	41

Table 5. The Results of the Neural Network Model

Correctly classified instances	40	85.1064%
Error classified instances	7	14.8936%
Mean Absolute Error	0.1568	
Root Mean Squared Error	0.345	

Table 6. Confusion-matrix

	Valid	Invalid
Valid	0	4
Invalid	0	38

We can see that the correct prediction rate was 85.1064% through the neural network model to predict the class label and the correct prediction rate was 88.234% by using the decision tree model to predict the class label from table 3 to 6.

5. Conclusions

In this paper, the process data mining based on classifier applies to coal mining data and analyzes the properties of the utility in order to obtain the effectiveness of the stone size. We can see that the correct prediction rate was 85% through the neural network model to predict the class label and the correct prediction rate was 88% through using the decision tree model to predict the class label. The study not only deepen the understanding of the data class label prediction classifier, but also at the same time it provides a powerful class label forecasting tool for the coal mining.

In future we will try to realize classifier based on support vector machine (SVM) to further improve the accuracy of the classifier and provide more powerful prediction tool.

Acknowledgements

This work was supported by Natural Science Fund of Jiangsu Province (BK20130235), Natural Science Fund of Changzhou (CJ20140049), China National Natural Science Fund of China (61302124) and Scientific Research Foundation of Jiangsu University of Technology (KYY14018, KYY14020) and Applied Basic Research Program of Jiangsu University of Technology (KYY14005).

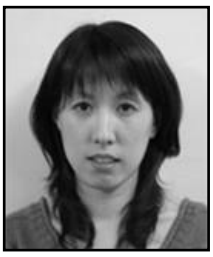
References

[1] ZHANG Xin-meng, JIANG Sheng-yi, "An algorithm for clustering uncertain categorical data based on similarity probability", Journal of Shandong University: Engineering Science, vol. 41, no. 3, **(2011)**, and pp.12-16.
[2] ZHU Jia-jun, ZHENG Jian-guo, LI Jin-bing, "Rough classification algorithm for uncertain extension group decision making", Control and Decision, vol. 27, no. 6, (2012), pp.32-36.
[3] Li Xiongfei, Li Jun, Qu Chengwei, etc, "Balancing Method for Skewed Training Set in Data Mining", Journal of Computer Research and Development, vol.49, no.2, **(2012)**, pp. 346-353.
[4] QING Xiao-Xia, XIAO Dan, WANG Bo, "A real-time monitoring method of energy consumption based on data mining", Journal of Chongqing University: science and Technology, vol.35, no. 7, **(2012)**, pp. 133-137.
[5] Liu Dayou, Chen Huiling, Qi Hong, etc, "Advances in Spatiotemporal Data Mining", Journal of Computer Research and Development, vol.50, no.2, **(2013)**, pp. 225-239.
[6] ZHU Xiao-Dong, XIAO Fang-Xiong, HUANG Zhi-Qiu, etc, "Description Logic Based Extended Predictive Model Markup Language EPMML", Chinese Journal of Computers, vol. 35, no. 8, **(2012)**, pp.1644-1654.
[7] HE Ping, XU Xiao-Hua, CHEN Ling, "Supervised Spectral Space Classifier", Journal of Software, vol.23, no.4, **(2012)**, pp.748-764.
[8] BAI Wen-Bin, JIAO Xiao-Yan, WANG Li-Ge, etc, "Effects of coal mining subsidence on the community structure of soil macro-fauna in central China", Chinese Journal of Eco-Agriculture, vol.20, no.4, **(2012)**, pp.123-135.
[9] XIE Xudong, LI Ning, PENG Liangrui, etc, "Baseline-independent feature extraction for Arabic writing", Journal of Tsinghua University: Sience and Technology, vol. 52, no.12, **(2012)**, pp.1682-1686.
[10] Hayat M, Khan A, "Discriminating outer membrane proteins with fuzzy K-nearest neighbor algorithms based on the general form of Chous PseAAC", Protein and peptide letters, vol.19, no.4, **(2012)**, pp. 411-421.
[11] YANG Zhe, LI Ling-zhi, JI Qi-jin, "Network traffic classification using decision tree based on minimum partition distance", Journal on Communications, vol.33, no.3, **(2012)**, pp. 90-102.
[12] WANG Zhong-Feng, WANG Zhi-Hai, "An Optimization Algorithm of Bayesian Network Classifiers by Derivatives of Conditional Log Likelihood", Chinese Journal of Computers, vol.35, no.2, **(2012)**, pp.364-374.
[13] Duan Yufeng, Hei Zhenzhen, Ju Fei, etc, "Semantic Annotation of Species Description Text in Chinese Literature by Naïve Bayes Classifier", Journer China Society for Scientific and Technical information, vol.31, no.8, **(2012)**, pp.805-812.
[14] CHEN Dongning, YAO Chengyu, "Reliability Analysis of Multi-state System Based on Fuzzy Bayesian Networks and Application in Hydraulic System", Journal of Mechanical Engineering, vol.48, no.16, **(2012)**, pp.175-183.
[15] LIN Guangdong,WANG Xufa, "A dynamic control model for modulating using artificial neural networks using the artificial endocrine system", Journal of University of Science and Technology of China, vol.42, no.2, **(2012)**, pp.148-153.

Authors

Haixu Xi, he received the master's degree in Educational Technology from Nanjing normal University in 2006. Currently, he is an lecturer at the School of Computer School of Computer Engineering in Jiangsu University of Technology. His interests are in digital teaching resource development and multimedia technology.

Dan Guo, she received the master's degree in Educational Technology from Nanjing normal University in 2006.Currently, she is an associate professor at the School of Computer School of Computer Engineering in Jiangsu University of Technology. Her interests are in Teaching Information and multimedia technology.



Hongjin Zhu, she received the M.Sc. and Ph. D. from the Yamgata University of Japan in 2007 and 2010, respectively. She was employed as a special researcher in the Department of Engineering, Yamagata University of Japan in 2010. She is currently an associate professor in Jiangsu University of Technology, Changzhou, China. Her research interests include image processing, computer vision, pattern recognition and evolutionary computation.