Proceedings of the 5th WSEAS International Conference on Applications of Electrical Engineering, Prague, Czech Republic, March 12-14, 2006 (pp171-175)

Design Classification Algorithm in Data Mining Prototype System and Application in Unit's Bidding Ability of Electricity Market [1]

Hongwen Yan , Rui Ma
Changsha University of Science and Technology
Chiling Road 45, Changsha , 410076
China

Abstract: - Design and implementation of classification algorithm in data mining prototype system is described in this paper. This function analyzes a set of training data, constructs a model for each class based on the features in the data, and adjusts the model based on the test data. The architecture of data mining prototype system is defined and the algorithms including ID3,C4.5,SLIQ and Bayesian is discussed. A method based on Naive Bayesian classification is applied to the generation unit's bidding decision system of electricity market. The knowledge that the ability of unit bidding is gained, Taking the market's demand, bidding price and the capacity of bidding unit into consideration,.This knowledge is very useful in supporting the generating bidding unit to make decisions and the electric agency, PX and ISO to design an optimal trade project

Key-Words: - Classification Algorithm, Data Mining , Prototype System, Naive Bayesian, Unit's Bidding Ability ,Electricity Market

1 Introduction

Classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends. We are given a set of example records called a training set where each record consists of several fields or attributes, one of the attributes called the classifying attribute indicates the class to which each example belongs to.[1-2] Applications in Data mining for biomedical analysis and financial data analysis and the industry are developed in past years Although data mining is a young field with many issues that still need to be researched in depth, there are already a great many data mining system products and specific data mining application softwares available. But how to design a data mining prototypes system and how to choose a data mining systems are appropriate for our task. Design and implementation of classification algorithm in data mining prototype system is described in this paper. This function analyzes a set of training data, constructs a model for each class based on the features in the data, and adjusts the model based on the test data. The algorithms including ID3,C4.5,SLIQ and Bayesian is discussed. The advantage of classification method of big sample is investigated. In the end the research and implementation for Naive Bayesian classification is given.

The rest of the paper is organized as follows. In section 2 , the system architecture of data mining prototype system is describes. In section 3, The classification algorithm.and Naive Bayesian classification are discussed . Application based on Naive Bayesian classification in Unit's Bidding Ability of Electricity Market is described in section IV. Finally, Section V contains our conclusions

2 Architecture of Data Mining Prototype

The architecture of data mining prototype system is shown in Fig. 1,which takes data from a relational database, integrates and transforms them into a multidimensional database, and then performs multidimensional on-line analytical processing and on-line analytical mining based on the user's processing requests.

The core module of the architecture is an OLAM engine, The OLAM engine in the Miner system performs multiple data mining tasks, including concept description, association, classification, prediction, clustering, and time-series analysis. Classification method is discussed in this paper.

[1] This work is partially supported by the Hunan Province Science Foundation Grant #05JJ40088 and the Foundational Research Funds ofChangsha University of Science and Technology

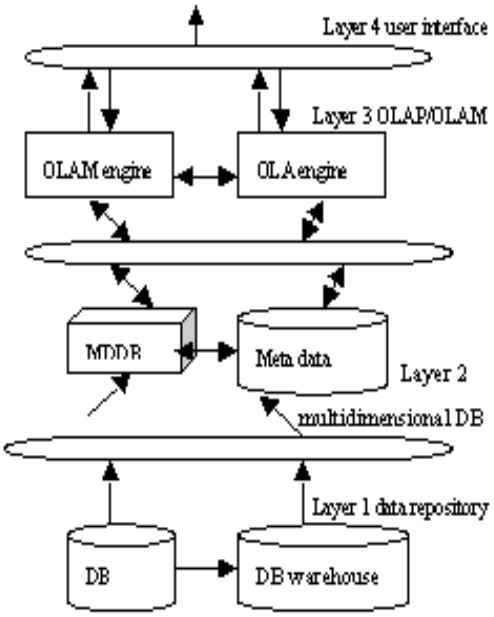

Fig. 1 Miner architecture of data mining prototype system

3 Classification Algorithm

3.1 Classification by Decision Tree

3.1.1 Algorithm ID3 and C4.5

The algorithm ID3 and C4.5 [3]are greedy algorithm that constructs decision trees in a top-down recursive divide-and-conquer manner. The information gain measure is used to select the test attribution at each node in tree.Many enhancements to the ID3 has been proposes in algorithm C4.5. The knowledge can be extracted and represented in the form of classification IF-THEN rules. The design and application of a decision tree algorithm C4.5 is presented. The algorithm C4.5'speed is fast and has robustness, but this algorithm has some disadvantages, such as, information gain measure is biased in that it tends to prefer attributes with many values, but attributes with many values is always not the best, this two algorithms have been established for relatively small data sets. Efficiency and scalability become issues of concern when these algorithms are applied to the mining of very large real-world databases.

Algorithms for the induction of decision trees from large training sets include SLIQ and SPRINT that address the scalability issue have been propose.

3.1.2 Algorithm SLIQ

SLIQ is described in 1996 by Mehta, Agrawal and Rissanen, it is a decision tree Classifier ,designed to classify large training data. It uses a pre-sorting technique in the tree growth-phase[4]. This sorting procedure is integrated with a breadth-first tree growing strategy to enable classification of disk-resident datasets. SLIQ can obtain higher accuracies by classifying larger training data sets which cannot be handled by other classifiers. SLIQ can split for numeric attributes and for categorical attributes.

A splitting index evaluates the goodness of the alternative splits for an attribute. Several splitting indices have been proposed in the past, SLIQ use the gini index instead of information, originally proposed in [5][6][7]

.Data set S contains examples from *N* classes gini(S) is defined as：

$$Gini(S) = 1 - \sum_{j=1}^{n} p_j^{\,2} \qquad (1)$$

where p_j is the relative frequency of class *j* in *S*. If data set *S* contains examples from S_1 and S_2, then $Gini$ is defined as:

$$Gini\_split(s) = n_1 / n * Gini(s_1) + n_2 / n * Gini(s_2) \quad (2)$$

where n is the recorder numbers,n_1 is the recorder numbers of s_1, n_2 is the recorder numbers of s_2,the $Gini\_split(s)$ is the lowest ,the information is the highest. Typically, the midpoint of $Gini\_split(s)$ is chosen as the split point. SLIQ is binary decision trees, first, we evaluate of splits for each attribute and the selection of the best split. Second, create of partitions using the best split. If we splits for numeric attributes, The first step is to sort the training examples based on the values of the attribute being considered for splitting, Let $v_1,v_2,\ldots v_n$ is to be the sorted values of a numeric attribute A. Since any value between vi and vi_{+1} will divide the set into the same two subsets, we need to examine only *n-1* possible splits. Typically, the midpoint of each interval $vi - vi_{+1}$ is chosen as the split point, The highest information gain is chosen as the split point. If *S(A)* is the set of possible values of a categorical attribute *Ai*, then the split test is of the form $A \in S'$,where $S' \in S$ .Since the number of possible subsets for an attribute with *n* possible values is 2^n ,the search for the best subset can be expensive, Therefore, We should use a fast algorithm for subset selection for a categorical attribute .

3.2 Bayesian Classification and Naïve Bayesian Classification

3.2.1 Bayesian Classification

The Bayes decision rule (Duda and Hart, 1973;Langley et al., 1992) is a classification method based on the Bayes theorem. It performs an approximate calculation of the probability that an example belongs to a class given the values of predictor variables. The simple naivve Bayes classifier is one of the most successful algorithms on many classification domains. In spite of its simplicity, it is shown to be competitive with respect to other more complex approaches in several specific domains.[8-12]

Bayesian classification is based on Bayes theorem, Let X be a data sample whose class label is unknown. Let H be some hypothesis, we want to determine $P(H|X)$,the probability that the hypothesis H hold given the observed data sample X. Bayes theorem is

$$P(H|X)= \frac{P(X|H)P(H)}{P(X)} \quad (3)$$

3.2.2 Naïve Bayesian Classification

The naïve Bayesian classifier works as follows:

Step 1):This classifier learns from training data the conditional probability of each variable Xn given the class label C. Classification is then done by applying Bayes rule to compute the probability of C given the particular instance of $X1; \ldots ; Xn$,

$$P(C_i|X)>P(C_j|X),1\leq j\leq m,J\neq i \quad (4)$$

Step 2)Naiv Bayes is founded on the assumption that variables are conditionally independent given the class. Therefore, posterior probability of the class variable is formulated as follows,

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) \quad (5)$$

If A_k is categorical, then $P(X_k|C_i) = s_{ik}/s_i$,where s_{ik} is the number of training samples of class C_i having the value x_k for A_k ,and s_i is the number of training samples belonging to C_i .

If A_k is continuous valued,then the attribute is assumed to have a Gaussian distribution so that

$$P(x_k|C_i) = g(x_k, \mu_{C_i}, \sigma_{C_i}) = \frac{1}{\sqrt{2\pi}\sigma_{C_i}} e^{-\frac{(x_k-\mu_{C_i})^2}{2\sigma_{C_i}^2}} \quad (6)$$

Step 3)In order to classify an unknown sample X, $P(X|C_i)P(C_i)$ is evaluated for each class C_i, Sample X is then assigned to the class C_i only if

$$P(X|C_i)P(C_i) > P(X|C_j)P(C_j), \quad \text{for } 1\leq j\leq m, j\neq I \quad (7)$$

It is assigned to the class C_i for which $P(X|C_i)P(C_i)$ is the maximum.

4 Application Based on Naive Bayesian classification in Unit's Bidding Ability of Electricity Market

4.1 Background

Recent years, the electricity industry has undergone drastic changes due to a world wide deregulation or privatization process that has significantly affected energy markets. With past and current difficulties in building new transmission lines and the significant increase in power transaction associated with competitive electricity markets, maintaining system security is more than before.

We have proposed a new data-mining framework for competitive bidding assessment in deregulated power market. This bidding mechanism is consistent with features of electric energy production and consummation, which is more convenient for operating power markets[13], A novel multi-objective block and group bidding model for hydrothermal power market is also proposed by the author[14],Taking the market's demand, bidding price and the capacity of bidding unit into consideration, the author illustrates that indicates the procedure of calculation and some knowledge about load's rate and the attributes mentioned above. In this way, we can get the data of bidding ability.

We use multi-objective optimal block bidding model with Market Operator(MO) purchase low-price and low-emission power to supply load demand in [15].

Data preprocessing is an important issue for data mining, as real-word data tend to be incomplete, noisy, and inconsistent. We can clean integrate and transform those data. Data cleaning can used to fill in missing values, smooth noisy data, identify outliers, and correct data inconsistencies. Elements including taking the market's demand, bidding price and the capacity of bidding unit into consideration affect the

data of bidding ability.

In this paper, we can use discretization techniques to reduce the number of values for a given continuous attribute. We divide into the capacity of bidding unit into high medium and low. Bidding price is numeric attributes. Bayesian classification can splits for numeric attributes and for categorical attributes.

Taking the market's demand, bidding price and the capacity of bidding unit into consideration, the author illustrates an example that indicates the procedure of calculation and some knowledge about load's rate and the attributes mentioned above. In this way, we can get the data of bidding ability based Bayesian classification

4.2 Example Analysis

Samples of generation bidding training data tuples are shown in table 1

Table 1The Samples of generation bidding

No	market's demand	Bidding price	generator capacity	Class load's rate
1	low	low	big	high
2	medium	medium	small	medium
3	medium	medium	big	medium
4	low	medium	small	medium
5	medium	low	big	high
6	low	medium	big	medium
7	high	high	small	high
8	low	high	big	medium
9	high	high	big	High
10	low	high	small	low
11	high	low	big	high
12	high	low	small	high
13	medium	high	small	low
14	medium	low	small	high

Predicting a class label using naive Bayesian classification. The data samples are described by the attributes market's demand, bidding price, generator capacity. The class label attribute, load's rate ,has three distinct values. Let C_1 ,correspond to the class load's rate="high" and C_2 ,correspond to the class load's rate="medium" and C_3 ,correspond to the class load's rate="low". The unknown sample we wish to classify is

X=(market's demand="low", bidding price="low", generator capacity="big")

We need to maximize $P(X|C_i)P(C_i)$,

for i=1,2,3 $P(C_i)$,The prior probability of each class ,can be computed based on the training samples:

P（load's rate="high"） = 6/14 = 0.429

P（load's rate="medium"） =5/14 = 0.357

P（load's rate="low"） =3/14 = 0.214

.To compute $P(X|Ci)$， i=1,2,3, we compute the following conditional probabilities:

P（market's demand ="low" | load's rate="high"） =1/5=0.2

P（market's demand ="low"| load's rate="medium"） =3/5=0.6

P（market's demand ="low"| load's rate="low"） =1/5=0.2

5 Conclusion

Data mining prototype system is an on-line analytical mining system, developed for interactive mining of multiple-level knowledge in large relational databases and data warehouses. Design and implementation of classification algorithm in data mining prototype system is described in this paper. This function analyzes a set of training data, constructs a model for each class based on the features in the data, and adjusts the model based on the test data. The algorithms including ID3,C4.5,SLIQ and Bayesian is discussed. The advantage of classification method of big sample is investigated. In the end, the example analysis based naive Bayesian classification is given., Taking the market's demand, bidding price and the capacity of bidding unit into consideration, the author illustrates an example that indicates the procedure of calculation and some knowledge about load's rate and the attributes mentioned above. This knowledge is very useful in supporting the generating bidding unit to make decisions and the electric agency, PX and ISO to design an optimal trade project

References:

[1] J R. Kohavi, G.H. John, Wrappers for feature subset selection, Artif. Intell. 97 Press, 1997 pp. 273–324.

[2] F. Pernkopf, P. O`Leary, Visual inspection of machined metallic high precision surfaces, Eurasip J. Appl. Signal Process. Special Issue on Applied Visual Inspection 2002, pp 667–678

[3] J.C Schlimmer and D. Fisher. A case study of incremental concept induction. In Proc.

5th Natl. Conf. Artificial Intelligence(AAAI'86), San Mateo:Morgan Kaufmann 1986, pp496-501

[4] Manish Mehta, Rakesh Agrawal and Jorma Rissanen. SLIQ: A Fast and Scalable Classifier for Data Mining. IBM Almaden Research Center, 1996 ,pp.18-32

[5] Riese, Martin. "The GINI-index as a measure of the goodness of prediction", *Bulletin of Economic Research*, Vol. 49, pp.127-135, Jan.1997

[6] L. Breiman et .al .*Classification and Regression Trees*,Wadsworth, Belmont 1984,p80

[7] M. Mehta, J.Rissanen and R,Agrawal. MDL based decision tree pruning, InInt'l Conf.on Knowledge Discovery in Databases and Data Mining (KDD-95).Montreal. Canada, Aug.1995,pp.404-416

[8] Langley, P., Iba, W., Thompson, K., An analysis of Bayesian classifiers. In: Proceedings of 10th Nationa lConference on Artificial Intelligence, 1992. , pp. 223− 228

[9] P.Domingos and M.Pazzani. Beyyond independence: Conditions for the optimality of the simple Bayesian classifier. In Proc. 13th Intl.Conf. Machine Learing,1996, pp.105-112

[10] Langely, P., Sage, S.,. Induction of selective Bayesian classifiers. In: Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence, 1994, pp. 399− 406.

[11] Friedman, N., Goldszmidt, M., Building classifiers using Bayesian networks. In: Proceedings of the National Conference on Artificial Intelligence, 1996., pp. 1277− 1284.

[12] Friedman, N., Geiger, D., Goldszmidt, M., Bayesian network classifiers. Machine Learning, vol.29, pp131− 163., Apr.1997

[13] Jian Geng, Xifan Wang, Xiaoying Ding, etal. "Models of block bidding in power market and comparison with hourly bidding". Proceedings of the CSEE, vol.24, pp.18-23, Mar. 2004

[14] Rui Ma, Renmu He, Hongwen Yan. Novel Multi-objective Optimal Block and Group Bidding Model for Hydrothermal Power Market. Proceeding of the CSEE, vol.24, pp.53-57,Nov.2004

[15]Ma Rui, A Novel Bi-objective Fuzzy Optimal Model of Short-Term Tradeplanning Considering Envionmental Protection and Economic Profit in deregulated Power System, Proceeding of the CSEE, vol.22, pp.104-108,Apr.2002