Asian Journal of Information Technology 5 (3) : 280-284, 2006

The Application of Data Mining Technologies for Routing in Communication Networks

[1]S. Behin Sam, [2]Amna Omar, [3]A. Kannan and [2]P. Vivekanandan
[1]Department of Mathematics, Anna University, Chennai, India
[2]Department of Information Technology, Higher College of Technology, Muscat
[3]Department of Computer Science and Engineering, Anna University, Chennai, India

Abstract: Generally data communication is done by using the routing algorithms that uses metrics such as path length, communication cost etc., to determine the optimal path to a destination. This paper approaches the problem of routing in communication network from a new angle by using data mining technologies to determine the optimal path in a network. The new routing model combines several scalable data mining methods, such as decision tree, association rules and neural networks. The model predicts the quality of service in the communication paths, which can be used by routing algorithms to determine the optimal path to a destination.

Key words: Data mining, back propagation neural network, decision trees, association rules

INTRODUCTION

A computer communication network is a set of several interconnected computers. The main objective of a computer communication network is to provide an efficient communication among various geographically distant computers in order to increase their utility and making their service available to more users. Computer communication systems made their first noticeable appearance in the form of packet switching with the ARPANET.

Routing involves two basic activities which are determining optimal routing paths and transporting information groups (typically called packets) through a communication network.

Routing protocols use metrics to evaluate what path will be the best for a packet to travel. A metric is a standard of measurement, such as path bandwidth, that is used by routing algorithms to determine the optimal path to a destination. To aid the process of path determination, routing algorithms initialize and maintain routing tables, which contain route information. Route information varies depending on the routing algorithm used.

Data mining has been defined as The non-trivial extraction of implicit, previously unknown and potentially useful information from data[1]. It uses machine learning, statistical and visualization techniques to discover and present knowledge in an easily comprehensible form.

This study focuses the problem of metric calculation with the help of data mining techniques. The quality of service is the metrics which is predicted by using the data mining techniques, viz., decision tree, association rules and neural networks. Routing algorithm can use this metric for optimal path selection.

DATA MINING TECHNIQUES

Data mining is a knowledge discovery process and has become one of the hottest research areas in recent years. Data Mining is not so much as a single technique, the idea that there is more knowledge hidden in the data then shows itself on the surface. Any technique that helps extract more out of your data is useful, so data mining techniques form quite a heterogeneous group. Various different techniques available are query tools, statistical techniques, visualization online analytical processing (OLAP), case based learning (K-nearest neighbor), decision trees, association rules, neural networks, genetic algorithms etc. In the following subsections, we present a brief introduction to the three techniques used in our prediction based routing method, viz., network, decision trees and association rules.

Neural network: BP Neural Network is a back propagation multi-layer perceptron[9]. Generally speaking, it consists of an input layer, an output layer and several hidden layers. Each layer is made up of certain amount of artificial neurons. Two adjacent layers are connected with links, each with an associated weight. Fig. 1 shows the general structure of the artificial neuron.

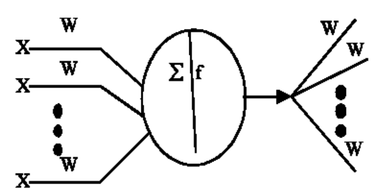

Fig. 1: Structure of an artificial neuron

Corresponding Author: S. Behin Sam, Department of Mathematics, Anna University, Chennai, India

There are a number of variations on the output function f. Two commonly used functions are the threshold function,

$$y = f(x) = \text{sgn}(\sum_{i=1}^{n} w_i x_i - \theta),$$

where θ is the threshold and the sigmoid function is given by

$$y = f(x) = \frac{2}{1+e^{-2x}} - 1, y \in (-1,1).$$

For a multi-layer artificial neural network, the outputs of the neurons in the first layer become the inputs of the next layer and so on. The weights in the network are initially assigned as random values; as the network is trained with the input samples, errors are propagated backwards and the weights are corrected iteratively, hence the term for this type of networks, Back Propagation.

BP Neural Network is widely used for prediction with a time series data set. It can also be trained as a classifier to conduct classification analysis.

Decision trees: Decision trees are the most popular model in classification analysis, for the mined results can be conveniently visualized with a tree structure, providing the user with the knowledge that is easy to understand[3]. Some well-known decision tree algorithms include C4.5, CART and CHAID[4-6].

Formally, let S be the data set. There is an attribute (or feature) A_k in S that can take c values from v_1 to v_c. Then S can be split into c non-overlapping subsets according to the values on A_k. That is $S = S_1 \cup S_2 \cup ... S_c$. Here A_k is called Classification Attribute and v_i(I = 1………c) is called a classification.

Decision trees can be used in classification prediction. Put a data record with unknown classification to the root of the decision tree. Let it move down along the path of measuring attributes, level by level. When it reaches a leaf node, the maximum classification of the leaf would be the classification prediction of the unknown data. Moreover, every path that originates from the root to a leaf gives rise to a classification rule. A set of such rules makes a decision table, which is also a frequently used method in prediction.

Association rules: Association rules analysis[7, 8] aims at discovering the hidden relationship within a data set. Frequently occurring patterns and unusual patterns are identified by iteratively scanning the data set, with the results cast into a number of rules that take a form such as X⇒Y or if X Then Y.

More formally let $l = l_1, l_2, ..., l_n$ be the set of several nominal attributes. Each attribute is called an Item. A data record in the data set can be represented by an n-dimensional vector. This vector

$$\tilde{T} = \{v_{t_1}, ...v_{t_i}, ...v_{t_m}\}, \{1 \le m \le n, \forall I_{t_i} \in I, v_{t_i} \in \text{Values}(I_{t_i})\},$$

is made up of values of several items in I. It is called an Itemset. Here m is the size of this itemset. If a certain itemset *T* comes in existence on a data record, we say the record satisfies itemset T. Let N be the number of records in the data set and Count (T) denote the number of records that satisfy itemset T.

Then if X and Y are two itemsets and there are no common items in them, we could write down an association rule with the confidence and support constraints in this form $X \Rightarrow Y \backslash_{c,s}$.
Here confidence is defined as,

$$c = \frac{\text{Count}(X \cup Y)}{\text{Count}(X)}$$

which measures the accuracy of a rule and support is defined as

$$s = \frac{\text{Count}(X \cup Y)}{N},$$

which describes the frequency of the rule[9].

DATA ATTRIBUTES AND COLLECTION

Data mining is based on a sufficiently large amount of data. Therefore, a data set constructed on a set of attributes, also called indices, is needed. Thus, analysis is on the traffic flow status in the different paths available in the communication network. Data was collected through the sniffer software. Two types of indices i.e. static and dynamic indices are listed as follows.

Static Indices: These indices have values that are fixed or relatively stable in a period of time.

a. Number of Communicat-ion links
b. Capacity of the link

Dynamic Indices: These indices have numeric values that change with time

a. No of units transmitted.
b. Delay across the links

The information about the static indices can be acquired by the network administrator or the network designer. The dynamic indices are measured in a time interval, ranging from 1 to 5 min.

BUILDING THE PREDICTION BASED ROUTING MODEL

Prediction attributes and prediction model: The purpose of building the prediction based routing model is to be able to predict the quality of service in the next period of time by processing and analyzing historical and current data of the link. Most of the time, it is of little meaning to get an accurate numeric value of the traffic flow by prediction. It is always the traffic flow status and the levels of traffic change in the next period of time extent get our attention. We can calculate the current reliability r from the data set and then categorize it into quality of service, shown in Table 1.

Quality of service is discrete thus more macroscopic in a sense, dependent upon traffic conditions. It will be taken as the main prediction index value in the model. A mapping between the ten classes of quality-of-service to numbers 9 to 0 can be built very easily, thereby preserving the sequence of information of the traffic flow status and making it more convenient for the model to process.

After the prediction index is determined, the prediction model of a single path in the network can be constructed. In essence, this model is a BP neural network with decision trees constraints and association rules constraints as shown in Fig 2. It only predicts the quality of service in the next period through dynamic indices. It also takes into account the relationships between dynamic and static indices and sequential patterns. So, compared with the traditional neural network prediction model, this model is more reasonable and accurate.

Prediction class network: The predicting class neural network focuses on the real time prediction of the Quality of Service, which is very important to the whole model. Its input is the Quality of Service in the time period t, obtained from the real time software at the path. And it outputs the prediction of the Quality of Service in period t+I, which is $S_{p(t+I)}$.

The structure and the corresponding parameters of the prediction class neural network need to be determined in advance. Some historical data are needed to train the network. As the network converges and become stable, its structure can be saved and used to predict the Quality of Service. For the correctness of the prediction, the training data set must contain a large amount of data.

Table 1: Quality of service criteria for communication links

Quality of Service	Reliability
A	$0.9 \le r < 1$
B	$0.8 \le r < 0.9$
C	$0.7 \le r < 0.8$
D	$0.6 \le r < 0.7$
E	$0.5 \le r < 0.6$
F	$0.4 \le r < 0.5$
G	$0.3 \le r < 0.4$
H	$0.2 \le r < 0.3$
I	$0.1 \le r < 0.2$
J	$0 \le r < 0.1$

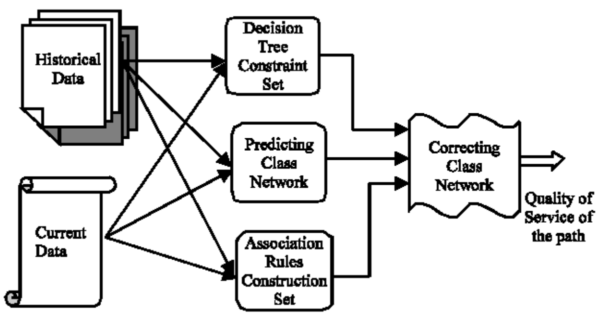

Fig 2. Structure o f the prediction model

Decision trees constraints set: The prediction class network only predicts quality of service from the quality of service itself and takes no consideration about the influence from those static indices. To analyze this type of influences, decision trees are used.

First select several sets of static indices. Multiple training data sets are thus constructed by combining the data of a set of static indices in period *t* (for the static attributes values can still change in some situations) and the quality of service in t+I. Each training set generates a decision tree T_i, which takes the quality of service as the classification attribute to examine the influences from different combination of static indices. Then these decision trees are saved into a set DTset={….., T_i,….}, which is called a Decision trees constraints set.

Association rules constraints set: The idea and function of an association rules constraints set is similar to a decision tree constraints set. They differ in that the association rules constraints set takes into account the influences from dynamic indices at the path, rather than the static indices, as the decision trees do. So a data set can be constructed by combining the dynamic indices data of the current path and its neighbors in period t, with the quality of service of the current path in period t+I, thereby generating association rules from historical data.

If the association rules are expected to constrain the quality of service, then it needs to make certain constraints in rules generation. In other words, it is to find out every association rule $R_{i\,with}$ the form $X \Rightarrow S_{P(t+1)} \backslash_{c,s}$, which satisfies

$$c = \frac{Count(X \cup S_{P(t+1)})}{Count(X)} \geq c_{min}$$

and.

$$s = \frac{Count(X \cup S_{P(t+1)})}{S} \geq S_{min}$$

This avoids the generation of those useless rules and saves computing resources by setting conditions while generating candidate item sets in the association rule algorithm[7, 8].

Those generated rules form the association rules constraints set Arset = {..., R_i,...}. It is a dynamically updated rules table in essence. And its function in constraining is similar to the decision trees constraints set.

Correcting class network: The use of the correcting class neural network is to compare the corrected quality of service $S_{p(t+1)}$ with the real quality of service $S_{r(t+1)}$ collected. By propagating errors it adjusts the weight W_N of the raw prediction result from the predicting class network, the weight W_t of the correction from the decision trees constraints set and the weight W_r of the correction from the association rules constraints set. These weights represent the degree of the influence on the final prediction result.

There is a structural uncertainty in the prediction of neural networks, that is, many trials of their structures and parameters need to be done before the best network appears. Factors like the size and distribution of the training set and the way networks are trained will influence the correctness of the predicting class network. If the predicting class network is well trained, then $S_{P(T+1)}$ will be greatly dependent on $S_{P(t+1)}$ and W_t and W_r will turn to be very small in this situation. But if there is quite some difference $S_{p(t+1)}$ and $S_{r(t+1)}$, then W_t and W_r will be increased. The target of placing the correcting class network is to compensate the failing accuracy in case of a terribly trained predicting class network.

Extending and updating the model: The prediction ability of this model at a collecting point depends on the way the training historical data is constructed. If the prediction of the quality of service after n periods is needed, a training data set that consists of the static and dynamic indices value, the quality of service in period t and the quality of service in period t+ n should be joined together. Thus the trained model could reveal the variation of the quality of service with an n-periods interval. If predictions of variant intervals are needed, there is a need to train and save multiple models.

The collected data of all collecting points in the network should be saved to a central data warehouse. All components of the prediction model are in general updated through historical data with a fixed time interval. Besides the periodic update, the model should be updated, even when the topology is reconstructed, once there are addition and deletion of nodes in the network.

IMPLEMENTATION OF THE MODEL

In the implementation, the association rules constraints set was used to analyze the dynamic indices of one path and its neighboring ones. Then the decision tree constraints set was used to take into account the influences of the static indices. The above two steps could be regarded as a kind of data pre-processing. When the processing was finished, the data was then imported to the neural networks for training.

The structure of the neural networks needed to be determined before training. The BP neural network is used in this implementation because it is technically mature. It provides quite satisfactory predicting accuracy with simple structure. The following graph shows the accuracy of the neural network training. Fig 3 shows the total squared error for the prediction process.

From Fig 3, it can be seen that the sum of squared error suddenly drops down after fifth epoch of training. It means that the neural network predicts the quality of service very quickly.

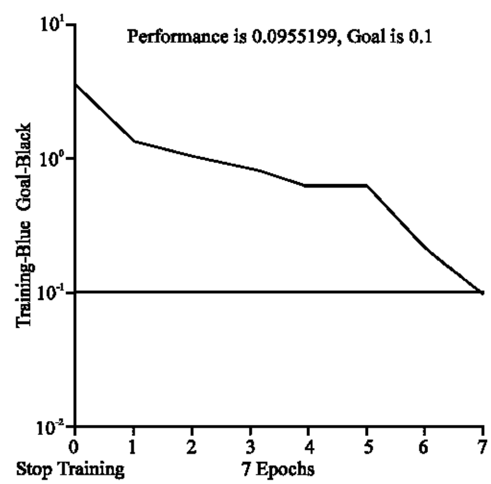

Fig 3: Squared error

CONCLUSION

With the help of several data mining techniques, this study proposes an analytical and prediction model for predicting the data traffic in a path, which is based on dynamic and static indices data. It is implemented through integrating the predicting class network, the decision trees constraints set and the association rules constraints set into the model. Also a correcting class network is set up to correct the prediction errors. Thus this new scheme will be the next step towards the use of data mining technologies for reliable routing in computer communication network.

REFERENCES

1. Frawley, W., G. Piatetsky-Shapiro and C. Matheus, 1992. Knowledge Discovery in Databases: An Overview. AI Magazine, pp: 213-228.
2. Zhaoqi, B. and Z. Xuegong, 2000. Pattern Recognition (2nd Edn). Tshnghua University Press, Beijing.
3. Jiawei, H. and K. Micheline, 2000. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers.
4. Kass, G.V., 1980. An exploratory technique for investigating large quantities of categorical data. Applied Statistics, 29: 119-27.
5. Brieman, L., J.H. Frieman R.A. Olshen and C.J. Stone, 1984. Classification and Regression Trees. Wadsworth, Inc.
6. Ross, Q.J., 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers.
7. Rakesh, A. and S. Ramakrishnan, 1994. Fast algorithm for mining association rules. In proceedings of the 20th Intl. VLDB Conference, Santiagom Chile, www.cs.buffalo.edu.
8. Rakesh, A., I. Tomasz and S. Arun, 1993. Mining Association rules between sets of items in large databases. In proceedings of the International conference on management of data (ACM SIGMOD), www.portal.acm.org.