The Research of Meteorological Data Mining Using Discrete Bayesian Networks Classifier Based on Hadoop

Zhang Yongjun[a], Sun Jing[b]

BUPT, No.10 Xitucheng Road, Haidian District, Beijing

[a]yjzhang@bupt.edu.cn, [b]sunjing_tyut@163.com

Keywords: Bayesian Networks; Predictive Ability; Classified Prediction; Data Mining; Hadoop

Abstract. The method of Native Bayesian classification data mining in weather forecast has some defects, such as there is not independent of each other between predictors, but a certain relevance which results in the decrease of prediction accuracy. This paper explores an improved algorithm which is based on the theory of discrete Bayesian Networks, and combines with Hadoop distributed file system and parallel processing programming models to predict rainfall. The experiments show that the improved algorithm not only makes the classification prediction more reliable but also improves the efficiency greatly. In addition, it provides a solution of huge amounts of data mining in the other fields.

Introduction

Bayesian Networks is one of the efficient models in the field of uncertain knowledge expression and inference, widely used in the field of data mining. It has the following characteristics: the expression form of graph model, partial and distributed study mechanism and directly perceived inference. Bayesian Networks is able to use incomplete, inaccurate or uncertain knowledge and information to make effective reasoning.

The method of Native Bayesian classification data mining[1] in rainfalls prediction mainly do the pretreatment, model training, accuracy assessment on pre-selected predictors and the target factor. It aims to infer the maximum possible value of the target factor in the case of the known predictors. Conditional independence assumption is the precondition of Native Bayesian classifier, however this assumption is usually not set up in real life, so the prediction accuracy may be reduced.

This paper explores an improved algorithm in rainfalls prediction. It analyzes the correlation of pre-selected predictors and the target factor using correlation coefficient to select predictors and rainfall factors as the target factor, do training model of Bayesian Networks classifiers and computing conditional probability tables(CPT) from huge amounts of data. In addition, this algorithm uses Hadoop distributed file system and parallel processing programming models[2], so it is suitable for handling massive data with high efficiency, stability and reliability.

Theoretical background

Bayesian Networks classifier.

(1)Bayesian theory: The prior probability P(h) represents the initial probability of h before being trained. P(D) represents the prior probability of training data to be observed. P(D|h) represents the probability of D assuming h establishes. The posterior probability P(h|D) represents the probability of h when given the training data D. There:

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)} \tag{1}$$

Maximum a posterior(MAP) is finds is the most likely hypothesis h(h∈H) finding from the candidate set H assuming given the data D. When the following equation establishes, h_{MAP} is called MAP assumption:

$$h_{MAP} = \arg\max_{h \in H} P(h \mid D) = \arg\max_{h \in H} \frac{P(D \mid h)P(h)}{P(D)} \tag{2}$$

(2)Bayesian Networks theory[3]: Discrete Bayesian Networks is based on the Bayesian probability. It is composed of binary group (Bs, Bp), Bs is a Bayesian Networks structure, including a node set A of the network structure and a directed edge set E of direct relationship between nodes of A. Bp is CPT of the nodes when the states of their parent nodes has been given, it show the probability of the dependencies between the node and its parent nodes. When the node X has no parent nodes, there is prior probability p(X) in CPT of X; When the node X has k parent nodes $\{Y_1, Y_2, \ldots, Y_k\}$, the CPT of node X is the conditional probability $P(X \mid Y_1, Y_2, \ldots, Y_k)$.

Bayesian Networks classifier[4]: There is an implicit conditions independent relationship in Bayesian Networks structure. In addition to the child nodes and the parent nodes of the node, the node and the rest of the nodes are conditional independence. So joint probability can be described as following, assume that $\{X_1, X_2, \ldots, X_n\}$ is a finite set:

$$p(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p(x_i \mid \pi_{x_i})\} \tag{3}$$

$p(x_1, x_2, \ldots, x_n)$ represents the probability of a specific combination of X, π_{x_i} represents the parent node of node x_i. Assume that C is a finite set and cj is the jth decision attribute of C, n attributes variables of Collection Ω are $x_1, x_2, \ldots, x_n$, $\{x_1, x_2, \ldots, x_n, c_j\}$ is a collection of samples with known class. The formula can be obtained by Bayesian theory and Bayesian Networks theory:

$$c = \arg\max_{c_j(x_1, x_2, \ldots, x_n)} \{p(c) \prod_{i=1}^{n} p(x_i \mid \pi_{x_i}, c)\} \tag{4}$$

Predictive ability theory[5]

Definition 1: $F(X_{m_1}, X_{m_2}, \ldots, X_{m_t} \to X_i)$ is the predictive ability of（$X_{m_1}, X_{m_2}, \ldots, X_{m_t}$）on X_i,

$$F(X_{m_1}, X_{m_2}, \ldots, X_{m_t} \to X_i) = \sum_{X_{m_1}} \ldots \sum_{X_{m_t}} p(X_{m_1}, \ldots, X_{m_t}) \max_{X_i(X_{m_1}, \ldots, X_{m_t})} \{p(X_i \mid X_{m_1}, \ldots, X_{m_t})\}$$

$$m_j \neq i, j = 1,2,\ldots t \tag{5}$$

Definition 2: $\hat{F}(X_{m_1}, X_{m_2}, \ldots, X_{m_t} \to X_i)$ is the estimate value of $F(X_{m_1}, X_{m_2}, \ldots, X_{m_t} \to X_i)$, it can be described as following:

$$\hat{F}(X_{m_1}, X_{m_2}, \ldots, X_{m_t} \to X_i) = \sum_{X_{m_1}} \ldots \sum_{X_{m_t}} p(X_i) \max_{X_i(X_{m_1}, \ldots, X_{m_t})} \{p(X_{m_1}, \ldots, X_{m_t} \mid X_i)\} \quad m_j \neq i, j = 1,2,\ldots t \tag{6}$$

(1) Build the initial Bayesian Networks structure

Assuming that $\rho_{in} = \rho_{out} = 1.1$

$\frac{\hat{F}(X_j \to X_i)}{\hat{F}(X_i \to X_i)} > \frac{\hat{F}(X_i \to X_j)}{\hat{F}(X_j \to X_j)}$, and $\max\{\frac{\hat{F}(X_j \to X_i)}{\hat{F}(X_i \to X_i)}, \frac{\hat{F}(X_i \to X_j)}{\hat{F}(X_j \to X_j)}\} > \rho_{in}$, add the arc $X_j \to X_i$;

$\frac{\hat{F}(X_i \to X_j)}{\hat{F}(X_j \to X_j)} > \frac{\hat{F}(X_j \to X_i)}{\hat{F}(X_i \to X_i)}$, and $\max\{\frac{\hat{F}(X_i \to X_j)}{\hat{F}(X_j \to X_j)}, \frac{\hat{F}(X_j \to X_i)}{\hat{F}(X_i \to X_i)}\} > \rho_{in}$, add the arc $X_i \to X_j$;

$\frac{\hat{F}(X_j \to X_i)}{\hat{F}(X_i \to X_i)} < \rho_{out} \cap \frac{\hat{F}(X_i \to X_j)}{\hat{F}(X_j \to X_j)} < \rho_{out}$, add the arc $X_j \to X_i$ or $X_i \to X_j$;

(2) Regulate the initial Bayesian Networks structure

Assume that $X_{m_1^1}, X_{m_2^1}, \ldots, X_{m_t^1}$ is a minimum cut set between X_i and X_j

$$\frac{\hat{F}(X_{m_1},...,X_{m_t},X_j \to X_i)}{\hat{F}(X_{m_1},...,X_{m_t} \to X_i)} > \frac{\hat{F}(X_{m_1},...,X_{m_t},X_i \to X_j)}{\hat{F}(X_{m_1},...,X_{m_t} \to X_j)},$$

and $\max\{\frac{\hat{F}(X_{m_1},...,X_{m_t},X_j \to X_i)}{\hat{F}(X_{m_1},...,X_{m_t} \to X_i)}, \frac{\hat{F}(X_{m_1},...,X_{m_t},X_i \to X_j)}{\hat{F}(X_{m_1},...,X_{m_t} \to X_j)}\} > \rho_{in}$, add the arc $X_j \to X_i$;

$$\frac{\hat{F}(X_{m_1},...,X_{m_t},X_i \to X_j)}{\hat{F}(X_{m_1},...,X_{m_t} \to X_j)} > \frac{\hat{F}(X_{m_1},...,X_{m_t},X_j \to X_i)}{\hat{F}(X_{m_1},...,X_{m_t} \to X_i)},$$

and $\max\{\frac{\hat{F}(X_{m_1},...,X_{m_t},X_i \to X_j)}{\hat{F}(X_{m_1},...,X_{m_t} \to X_j)}, \frac{\hat{F}(X_{m_1},...,X_{m_t},X_j \to X_i)}{\hat{F}(X_{m_1},...,X_{m_t} \to X_i)}\} > \rho_{in}$, add the arc $X_i \to X_j$;

$\frac{\hat{F}(X_{m_1},...,X_{m_t},X_j \to X_i)}{\hat{F}(X_{m_1},...,X_{m_t} \to X_i)} < \rho_{out} \cap \frac{\hat{F}(X_{m_1},...,X_{m_t},X_i \to X_j)}{\hat{F}(X_{m_1},...,X_{m_t} \to X_j)} < \rho_{out}$, delete the arc $X_i \to X_j$.

(3) Loop checking

If there are no parent and child nodes of the node in the network, delete the node and its connected arc, then do this operation on the sub-network. If that, each node still has a parent and child nodes the sub-network. Then there is the loop, otherwise there is no loop.

MapReduce.

MapReduce[6] is a model which can process massive data in parallel on a large computer cluster. As a simplified model of parallel computation, it abstracts the details of concurrent processing, fault tolerance and data distribution to a library, and divides the data processing into Map stage and Reduce stage:

Map stage: A group of <key, value> as the data input of Map function are mapped and aggregated according to the key, then producing a set of intermediate results <key1, value1>.

Reduce stage: The output of Map function is passed to Reduce function in which the intermediate results with the same Key are merged to produce the final result <key2, value2>.

MapReduce provides programmers with a powerful yet simple API, programmers do not care about the background complex task scheduling and load balancing problem. In addition, application can be deployed on the cluster formed by the ordinary PC with high performance.

The key algorithm

First, it analyzes the correlation of pre-selected predictors and the target factor using correlation coefficient to select predictors and rainfall factors as the target factor. Then it does model training to get Bayesian Networks classifiers and compute conditional probability tables (CPT) from huge amounts of data. Finally, it does accuracy assessment.

Pretreatment.

The source data contains the inconsistent format and invalid data, so it is necessary to unified data format and remove the invalid data. Selection of predictors are the key factors influencing the efficiency and reliability of prediction method, it should choose the predictors using the correlation coefficient which have high correlation with rainfall. In addition, the output data needs to be organized in order to reconstruct the record set property.

Therefore, the pretreatment has two parallel tasks: one is the selection of discrete interval for each predictor; the other is to organize output data and identify the target factor. It can be completed by two MapReduce jobs. Job1 finds the maximum and minimum of each predictor adds the identifiers to predictors, getting the discrete interval width w=(Max-Min)/k ($k=\sqrt[6]{N}$, Where N is the total number of samples), so the cut-off points collection of each predictor is: $\{A_{min}+w,$

$A_{min}+2w,\ldots,A_{min}+(k-1)w\}$. Job2 organizes the data, and make each row of output data is the data of predictors three days in the past and rainfall one day in the future, the reduced data is prepared for the next model training.

Model training.

In order to evaluate the classifier's accuracy, the data set is divided into two parts: the training set and the test set. The former is used to obtain the Bayesian Networks classifier and the latter to test the classifier's accuracy. Model training process is divided into two tasks, one is to gather Statistical frequency of each predictor and the target factor, the other is calculating predictive ability according to the statistical frequency, the latter process depends on the former results. So it needs two MapReduce[7] processes to complete the task. Finally, according to the theory of Bayesian Networks theory and prediction ability theory, it gets the Bayesian Networks structure and the CPT.

Accuracy assessment.

In this process Map function divides the test set into many blocks, each of which is predicted by the classification model. Prediction results will be compared with the actual data to verify whether the prediction is correct or not, and will be output to Reduce function. In Reduce function, the results of Map function are used to calculate the accuracy rate.

Experimental results and analysis

Experimental environment and data.

This paper does the research based on Hadoop cloud computing platform which is composed by eleven computers with the same configuration and the configuration is as follows: 3.4GHz Dual-core CPU, 4GBMemory, 150GBHard disk, CentOS6.0 system, Hadoop 1.0.2. The data downloaded from China meteorological data sharing service system from 1951 to 2014 is used as the environment data which include Average pressure, precipitation, average temperature, average vapor pressure, average wind speed, sunshine time and other factors.

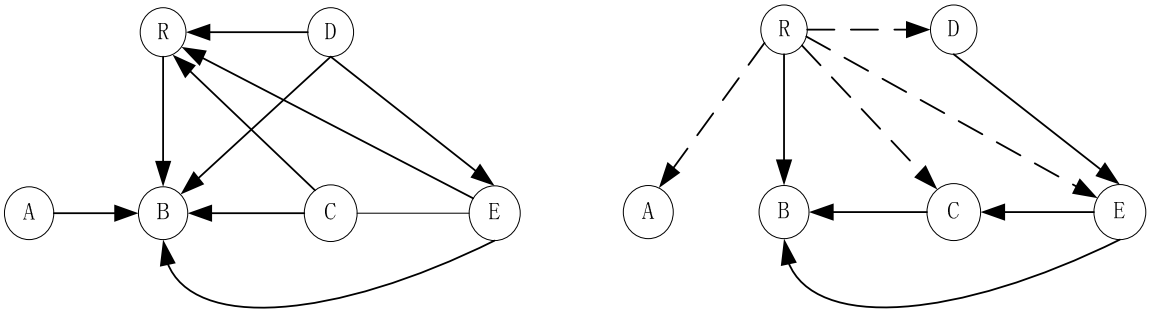
Experimental results.

(1)Correlation analysis: In this paper, average pressure, average temperature, average vapor pressure, sunshine time, small evaporation are chosen as the predictors and rainfall as the target factor. The predictors and the target factor added the identifiers as follows:

Table 1 factors and identifiers

predictors and target factor	20-20 rainfall	average pressure	average temperature	average vapor pressure	sunshine time	small evaporation
identifier	R	A	B	C	D	E

(2)Model training: Model training results as shown in the Fig.1. It shows that compared to the initial Bayesian Networks structure, the regulated Bayesian Networks structure adds the arc $R \rightarrow A$, deletes the arc $A \rightarrow B$ and $D \rightarrow B$, regulates the direction of the arc $R \rightarrow C$, $R \rightarrow D$ and $R \rightarrow E$. This paper puts the regulated Bayesian Networks structure as the classification model.



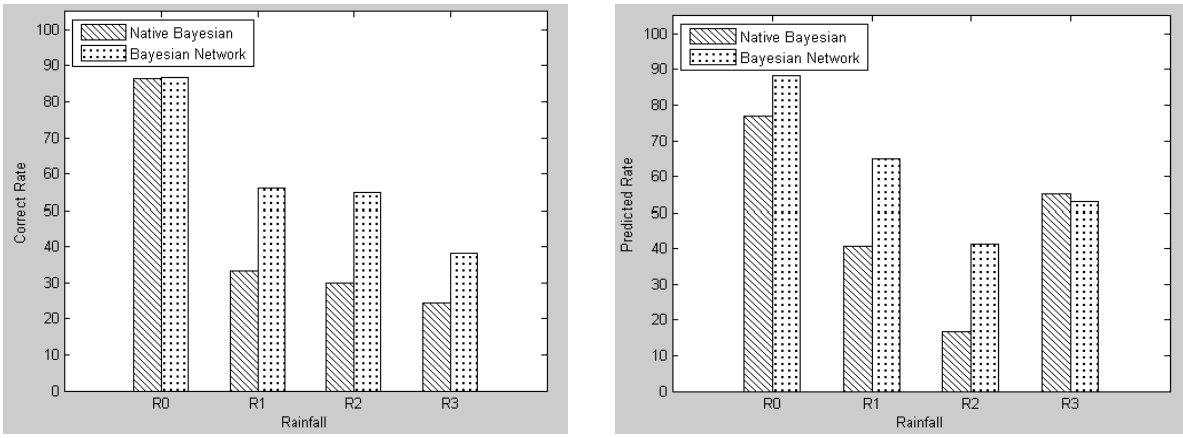
(a) (b)

Fig.1 Initial Bayesian Networks (a) and Regulated Bayesian Networks (b)

(3)Accuracy assessment: This paper does the accuracy assessment on the test set based on Bayesian Networks classification model. The correct rate and the predicted rate are shown in Table 2, R0 represents no rain, R1 represents light rain, R2 represents moderate rain, R3 represents heavy rain.

Table 2 the correct rate and the predicted rate

	R0	R1	R2	R3
the correct rate(%)	88.68	56.08	54.76	38.1
the predicted rate(%)	88.30	64.65	41.21	53.1



(a) (b)

Fig.2 The correct rate (a) and predicted rate (b) comparison

Fig.2 shows that the correct rate and the predicted rate of the Bayesian Networks classification are considerable, and prediction accuracy is significantly higher than the short-term climate prediction of using Naive Bayesian classification.

(4)Efficient testing: Three data sets are randomly selected from the original data set, the size of the three data sets is 10G, 20G, 50G, respectively. When the number of nodes in the cluster is 3, 5, 10, algorithm running time is shown in Fig.3 It can be seen that with the nodes' number increasing, the running time reduce more, it mean that that making using of Hadoop will have better efficiency.

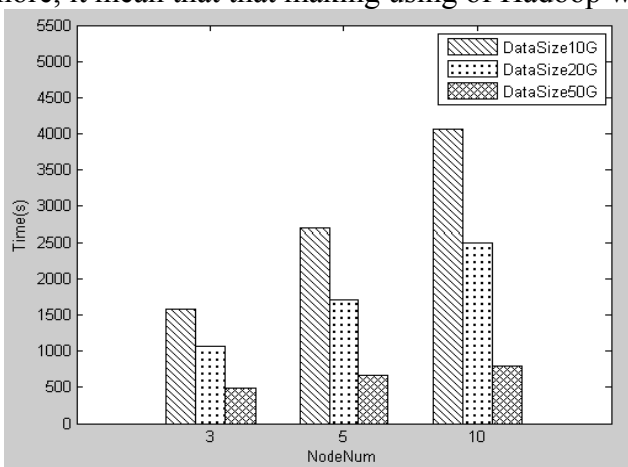


Fig.3 the running time of the algorithm

Summary

This paper explores an improved algorithm in rainfalls prediction. It uses predictive ability training Bayesian Networks structure on Hadoop platform. The rainfall experiment proves that the improved algorithm has better accuracy in the massive rainfall data set classification, and it can make full use of cluster resources, improve the efficiency of the large data mining. It not only can save processing time, optimize the algorithm's efficiency, improve the reliability of the

classification results, but also provides a solution of huge amounts of data mining in the other fields, such as radio and television networks.

References

[1] Hu Banghui, Yuan Ye, Wang Xuezhong, etal. Thunderstorm prediction based on Bayesian classification method[J]. Journal of PLA University of Science and Technology(Natural Science Edition), 2010, 11(5).

[2] Lam C. *Hadoop in action*[M]. Maning publication, 2010.

[3] Friedman N, Goldszmidt M. Building classifiers using Bayesian networks[C]. Proceedings of the 13th National Conference on Artificial Intelligence (AAAI), 1996,2:1277-1284.

[4] Heckerman D. Bayesian Networks for Data Mining[J]. Machine Learning, 1997(3):213-244.

[5] Wang Hui, Zhang Jianfei, Wang Shuangcheng. Learning Bayesian networks structure based on prediction ability[J].Computer Engineering and Applications, 2005, 37(1).

[6] White T. *Hadoop:the definitive guide*[M].Yahoo Press,2010.

[7] Apache Hadoop Main 2.7.0 API on http://hadoop.apache.org/docs/current/api/.