TELKOMNIKA Indonesian Journal of Electrical Engineering
Vol.12, No.1, January 2014, pp. 314 ~ 322
DOI: http://dx.doi.org/10.11591/telkomnika.v12i1.3997

Precipitation's Level Prediction Based on Tree Augmented Naïve Bayes model

Xue Shengjun[a], Chen Jingyi[b], Xu Xiaolong[c], Li Mengying[d]
Nanjing University of Information Science & Technology, School of Computer and Software, Nanjing, China
*Corresponding author, e-mail: sjxue@163.com[a], cjy_1225@163.com[b], xlxu1988@gmail.com[c], lmy3612@163.com[d]

Abstract

At present, most of the precipitation's level predictions use the laws of nature to build the mathematical model which contains one or more series level to carry out the numerical simulation, as thus to analyze the causes and consequences of the evolution. Bayesian model is one kind of the foregoing said. In the Bayesian classification model, Naive Bayes model is known for its stability and easy to operate, but the established precedent assumption tends to be inadmissible. So here the article proposed a new precipitation's level prediction model based on the tree Augmented Naïve Bayes(we called TAN model for short hereafter), which improve the original Naïve Bayes model defects and increase the association between the leaf nodes on the basis of the original model. And we use the Dongtai station, Jiangsu province meteorology data to test the new precipitation model. The results show that the new precipitation prediction model's performance is superior to the traditional Naive Bayes model.

Keywords: *precipitation, prediction, naïve Bayes, TAN model*

1. Introduction

With the development of science and technology, the precipitation prediction affects people's plans in every minute. Predictions have already been an indispensable part of the life. Due to the precipitation plays a significant role in the biological, aviation, military, etc aspects, the precipitation prediction is especially important. The causes of the precipitation are very complex. It involves a variety of climatic factors, such as pressure, temperature, relative humidity, terrain, atmospheric circulation etc. These factors are obviously not independent of each other. However, the meteorological conditions are diverse and easy to change. It's not a stable and fixed process. It always has a certain degree of difficulty when we build a precipitation prediction model. Currently we have many precipitation prediction methods at home and abroad, such as grey theory [1], Markov chain, grey Markov chain etc [2]. The classified methods include the neural network [3], decision-making tree [4], nonparametric method [5], and bayesian classification. And the bayesian classification has the best performance [6]. The precipitation prediction mostly through the statistical analysis, and get the results based on the precipitation data. Bayesian algorithm is such a prediction. In the precipitation prediction, Bayesian algorithms prediction is applied very extensive and effective. Bayes theorem has a high position in the field of data mining. Bayesian classifier contains a variety of classification algorithms. [7] They are all based on Bayes theorem. Naive Bayes algorithm has a stable structure and easy to operate. It use the priori probability and the sample to calculate the probability of each class variable. Then it chooses the maximum probability of the predicted results.

However, Naive Bayes algorithm has a prerequisite assumption------the attributes are independent of each other. If the attribute variables are not independent to each other, the prediction accuracy of the results is always poor. The formation of precipitation is related to a variety of climatic factors, such as pressure, temperature, relative humidity, terrain, atmospheric circulation. Their relations are very close and influence mutually. If use the traditional Bayesian algorithms only, the results will not be so ideal. Some researches have proposed the improvement on NB model. [8] For example, using the Ordered Weighted Operator to be the weighted of the product of the probability. [9] Or the Weighted naïve bayes classifier proposed

Received June 24, 2013; Revised August 18, 2013; Accepted September 17, 2013

by Harry. [10] The numerous experiments show that the TAN model performed very well in most cases. [11~12] Therefore, this study will use tree augmented naïve bayes model algorithms (TAN model) to predict the precipitation's level. At last, the experimental results show that this method make an effective improvement on the traditional naïve bayes prediction method.

2. Meteorological data preparation

In meteorology, precipitation is divided into light rain, moderate rain, heavy rain, storm, large rain storm, and severe storm, the classification criteria of the following table (Table 1):

Table 1. Precipitation classification standard

Category	light rain	moderate rain	heavy rain	storm	large rain storm	severe storm
classification standard (0.1 mm)	≤100	100~250	250~500	500~750	750~2000	≥2000

Selecting Jiangsu Province Dongtai Station 1951 ~ 2005 Chinese terrestrial climate and high altitude climate information day's value meteorological data material as a sample of the experimental data, Dongtai is the coastal city of East China, the climate is northern subtropical and warm temperate monsoon climate, four seasons are very distinction. The place is full of sunshine. And the rainfall is abundant. It satisfies the condition as the experimental sample.

This experiment will select the pressure, temperature, extreme maximum temperature, relative humidity as a attribute variable. The predictors we selected should have a high correlation with the precipitation's level. The correlation is come from the covariance between two variables. Covariance is a reflection of the average condition of the abnormal relationship of the two meteorological elements. For two variables a and b, the calculated equation of the correlation coefficient r_{ab} is as follow:

$$r_{ab} = \frac{\sum_{i=1}^{n}(a_i - \bar{a})(b_i - \bar{b})}{\sqrt{\sum_{i=1}^{n}(a_i - \bar{a})^2}\sqrt{\sum_{i=1}^{n}(b_i - \bar{b})^2}} \quad (1)$$

According to this equation, the selected predict factors comply with the requirement of experiment after certified. According to the meteorological grading standards, classify the four attribute variables. The meteorological grading standards are as follows (Table 2):

Table 2. Meteorological grading standards

least	less	little	many	much	most
≤50%	-50%~-20%	-20%~0	0~20%	20%~50%	≥50%

According to this table, discretizing the attribute variables of the sample meteorological data from 1951 to 2005, the air pressure, air temperature, relative humidity, and extreme maximum temperature are broken into six categories. So we get a new data set and make a preparation for the TAN model prediction algorithm.

3. Bayesian Algorithm and TAN Model

In this chapter we will introduce two special models in Bayesian case, Naïve Bayes is popular of its stable, simple, and easy to construct. The TAN model is improved on the Naïve Bayes model. It compensate the Naïve Bayes model defect in some degree.

3.1. The Basic Idea of the Bayesian Algorithm

Bayesian classification is a kind of the statistical classification. It's an algorithm based on the probability [13]. Bayesian algorithm theory is rediscovered and perfected by Laplace, the basic idea is using of the known prior probability and conditional probability density parameter, based on Bayes theorem to calculate the corresponding posterior probability, and then obtained the posterior probability to infer and make decisions. the equation of the Bayesian is as follows (2): [14]

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)} \tag{2}$$

Wherein, A is a hypothetical, B is a set of evidence. Before considering the evidence B, the probability of occurrence of event A---- P (A) is known as the a priori probability. After considering the evidence, the probability of the event A occurs under the conditions B event is called posterior probability. Bayes theorem descript the relationship between a priori and a posteriori probability very clear. The equation can be proved out from the definition of conditional probability:

According to the definition of conditional probability, the equation of the occurrence probability of event A under the conditions of B event occurs is as follows (3):

$$P(A|B)=\frac{P(A\cap B)}{P(B)} \tag{3}$$

The equation of the occurrence probability of event B under the conditions of A event occurs is as follows (4):

$$P(B|A)=\frac{P(A\cap B)}{P(A)} \tag{4}$$

According to (3) and (4), we can find that they have a common factor P (A∩B), so we can integrate the two equations and get a new equation (5):

$$P(A|B)P(B)=P(B|A)P(A)=P(A\cap B) \tag{5}$$

On both sides we divide P (B), if P (B) is non-zero, we can get the Bayes' theorem just like the said before. In the case of the promotion, it is assumed that {A1, A2, A3 .... An} is a subset of event collection, the following equation also can be used for any An:

$$P(A_n|B)=\frac{P(B|A_n)P(A_n)}{\sum_j P(B|A_j)P(A_j)} \tag{6}$$

The models we constructed are all from the Bayesian network. Bayesian network consists of two parts: The first part is a directed acyclic graph, where each node represents a random variable, each directed edge on behalf of the dependency of the relative nodes. The case which has given parent node variables, each variable is conditionally independent on non-children nodes. The second part includes the joint probability distribution of all the variables; it expresses all possible conditional probability of the node and its parent node.

Bayesian Network Learning model is a process which gets the Bayesian model through the data analysis process. It consists of structure learning and parameter learning. The Structure Learning is when the model structure unknown, it combined with prior knowledge and find a training set which comply to the requirement as far as possible, it is not only to determine the structure to build, but also to determine the parameters of the constructed network [15]. The parameters learning is when the model structure is known, using the sample data to learn the joint probability distribution, in order to determine the structure of the network parameters [16].

3.2. Naive Bayes Model

Naive Bayes is a classification model based on the famous Bayes theorem. Naive Bayes is a tree Bayesian network which contains a root node, a plurality of leaf nodes. In which the leaf node is an attribute variable. It descripts the properties of the object to be classified. The root node is a class variable that describes the object's categories. The classification is to give a data point, computing the posterior distribution P (A | B1 = b1... Bn = bn), and then select the max probability value of A as the category of the data points. It is a very simple structure model [17].

Naive Bayes model is based on a very simple assumption: It assumes that under the given classification feature condition; the attribute variables are independent between each other. In the Naive Bayes classifier, the number of P (bi | Aj) items in the training set to be estimated is just the number of different attribute values multiplied by the number of different target value, according to the Big O complexity estimation, we can find the complexity of the estimation of the P (b 1, b 2 ..... A j) is much smaller than others.

Naïve Bayes model was first proposed by Warner; Naïve Bayes model is a tree Bayesian network which contains a root node, leaf node .

According to the requirements of this study, the Naive Bayesian model for precipitation prediction s is shown in Figure 1.

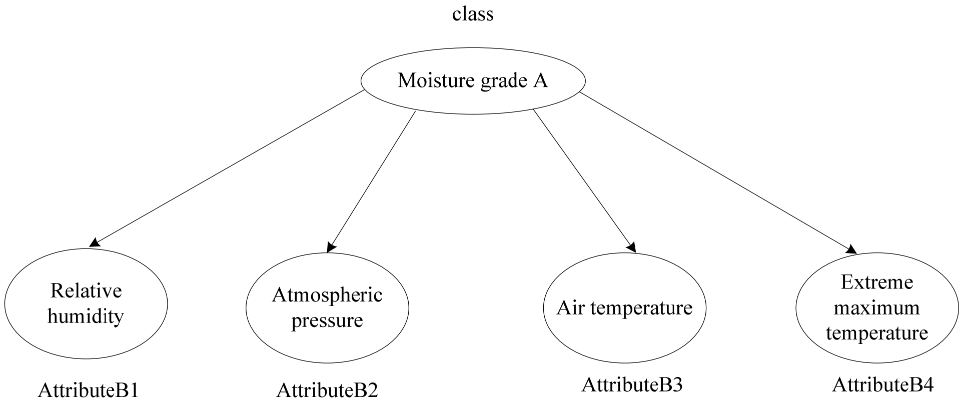

Figure 1. The precipitation's level prediction NB model

Figure 1 briefly shows the process of its operation, precipitation's level as the root node; it does not have a parent node. The leaf node, respectively, are relative humidity, pressure, temperature, extreme maximum temperature. It can be found that the leaf nodes as the predictor have no relation between each other. It contains an assumption that the attribute variables are independence mutually. According to the conditional probability distribution we find that the Naive Bayes model satisfies the equation as follows:

$$P(A, B_1 .....B_n) = P(A)\prod_{i=1}^{n} P(B_i \mid A) \tag{7}$$

Its structure is very simple. Through the Naïve Bayes model, the four predictors Naive Bayes algorithm calculation will be very small and get the max probability result of the occurrence.

According to the Naive Bayes we find that the data we need to compute is the sum times of precipitation in different level occurrence. So that we divide the times of all levels precipitation occurrence, we can get the prior probability in different levels. And to the conditional probability ,we need to know the times of the attributes and the class occurrence at the same time. We can use (4) to compute the probability.

3.3 TAN Model

Four predictors: pressure, temperature, relative humidity, extreme maximum temperature is not independent variables. The relative humidity is a percentage of the actual amount of water vapor contained and the maximum amount of water vapor contained in the air temperature.[18]According to the definition of relative humidity, we can find that it contact with the air temperature closeness. Under the normal circumstances, the pressure often decreases with the increasing height. From the geographical point of view, the pressure and temperature will be affected by the location, atmospheric circulation. For example, when the air is cooling and accumulation, it may form the cold anticyclone. The winter will affect China's cold air. So they are in highly interconnected. In order to simplify the relationship between the variables of the TAN model properties, we will use the dataset sample data to calculate the weights between the properties.

TAN model is a Tree Bayesian network which proposed by Friedman . In each leaf node, we add some edges to represent the relationship between the attributes. It is an extension of the Naive Bayes model. Owing to the complete Bayesian network is an NP problem, so it is equivalent to a compromise model, Friedman verified that the model performed excellent in most cases than the Naive Bayes model in his paper. The example in Figure 2, Shows the dependencies between properties in a simple TAN model. We order the collection U = {A, B1, B2 .... bn}, except the precipitation as the parent node of all the attribute variables, at most have one other attribute variables as the attribute's parent node. That is, each attribute have up to only two points to it directed edges.

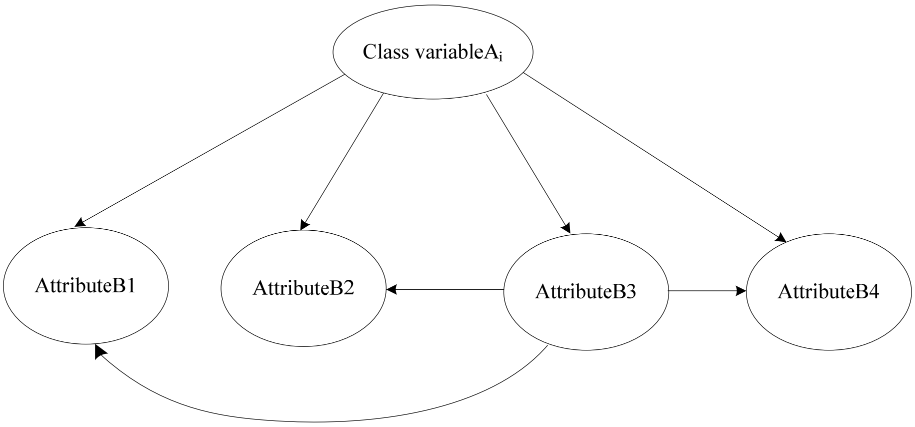

Figure 2. TAN model example

The TAN model joint probability distribution can be represented by the following equation:

$$P(A, B_1.....B_n) = P(A)\prod_{i=1}^{n} P(B_i \mid \pi(B_i)) \tag{8}$$

$\pi(B_i)$ not only include the precipitation, but also include other attributes of variables. that is, the four predictors are selected in this article.

3.4 The Construction of TAN Model

The key to use of the TAN model algorithm is how to determine the dependencies between attributes; we need to determine the parent node of a non-class attribute. Here we calculate the maximum weighted spanning tree of the precipitation's level prediction network which proposed by Friedman [19].

1. Calculating the conditional mutual information I(Bi;Bj|A), i≠j between attributes B1, B2.... Bn. The mutual information can be effectively expressing the correlation between two events, which is calculated as the equation (9):

$$I(B_i;B_j \mid A) = \sum_{b_i,b_j,a} P(b_i,b_j \mid a)\log\frac{P(b_i,b_j \mid a)}{P(b_i \mid a)P(b_j \mid a)} \tag{9}$$

A represents the precipitation.

2 Constructing the complete undirected graph. Each vertex are attributes B1, B2, B3, B4 showed on the figure. The edge which connect Bi and Bj sign the weight as I(Bi;Bj|A).

3 Constructing the maximum spanning tree of the precipitation classification network according to the weights to determine the direction of edges in the tree.

4. adding the class variable as the parent node of all the attributes variables, add the directed edges.

5. Constructing TAN model according to the weights of the results

4. Constructing the Precipitation Classification Prediction Model Based on the TAN Model

As shown on figure 3, firstly we need to construct the TAN model of the precipitation. According to the 3.4 section, we need to construct a specific TAN model. Before using the precipitation data over the years, we should handle the data to be the formation which can be used in the TAN model. So we according to the meteorological grading standards discrete the attribute variable data from 1951~2005. We just need to focus on the attributes that we used. We customize the method to realize the discretization. The specific operations of the method are as follows:

1. Creating a BufferedReader ( ) object to read the value in the file.

2. According to the meteorological file format, assigning each attribute variable value to the different container.

3. According to the meteorological grading standards, we classify the value of each container and output the txt file what we need to use.

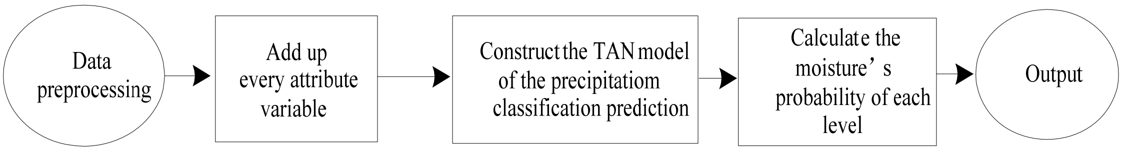
Data preprocessing
Add up every attribute variable
Construct the TAN model of the precipitatiom classification prediction
Calculate the moisture' s probability of each level
Output

Figure 3. The specific steps to construct the precipitation classification prediction model

After the end of the pre-processing of the data, we add up the size of the value of each attribute categories, according to the statistics to calculate the probability of occurrence of each value. Then we need to calculate the conditional mutual information of four attributes variable and sign the dependency relationship of the four properties. We can build the maximum weighted tree. After comparing the value of the conditional mutual information of four attribute variables, we get the TAN model.

Then we will make use of the joint probability distribution of the TAN model. That is, (8) and Bayes theorem to calculate the posterior probability after thinking about the training set. We regard the posterior probability as the probability of each precipitation classification level. Then the value returns to the maximum value. Thereby we get a classified prediction results.

Figure 4 descript the steps of the algorithm of precipitation classification prediction model, which atmospressure, temper, maxtemper, precipitation, rainfall. monthdate represent the pressure, the temperature, the extreme maximum temperature, the relative humidity, monthdate represent the maximum date monthly:

```
Algorithm: TAN_RainfallPrediction
Input: 1951 to 2005, Dongtai station, Jiangsu Province meteorological data and test data set.
Output: The prediction results from January to August, 2006
Begin
    Initialize atmospressure, temper, maxtemper, humidity, rainfall.monthdate
    I1←I (atmospressure, atmospressure | temper)to calculate the weight of atmospressure and temper;
    I2 ← I  (atmospressure, atmospressure | maxtemper) to calculate the weight of atmospressure and maxtemper;
    I3←I (atmospressure, atmospressure |humidity) to calculate the weight of the atmospressure and humidity;
    I4← I (temper, temper | maxtemper) to calculate the weight of the temper and maxtemper;
    I5 ← I (temper, temper | humidity) to calculate the weight of the temper and humidity;
    I6 ← I (maxtemper, maxtemper | humidity) to calculate the weight of the maxtemper and humidity ;
        for (i = 0, i <6, i + +)
            Imax1← getMax (I1, I2, I3)
            Imax2← getMax (I4, I5)
        end for
    Establish new TAN relations according to Imax1, Imax2, and I6;
        for (m = 0, m <month.date; m + +)
            Calculate P (rainfall | atmospressu-re, temper, maxtemper, precipitation, Imax1, Imax2, I6) accordi-ng to the equation;
            getMax (P(rainfall | ~);
        end for
    output the prediction results from January to August, 2006
end
```

5. Experimental Results

We use the precipitation data of the Dongtai station to predict the precipitation's level and compare the traditional NB classifier to the specific precipitation classification prediction model. We define the day as the unit and measure the level. Then we count the correct rate by monthly and analyze the results. The results are shown in figure 5. We can observe the model optimization intuitive. According to the comparison chart, we can find that TAN model's prediction rate is higher than the original Naïve Bayes model except February. Due to the changeable of the weather, it still not get a very high prediction accuracy generally. Due to it's very difficult to construct a fully Bayesian network. The TAN model already has a significant improvement. Although the prediction still not reach 90% or more, comparing to other precipitation predictive models, the prediction accuracy of TAN model still has many advantages.

At the same time, this study also defines the month as the unit and measure the level of every month. We found that the Naive Bayes forecast accuracy is 50.0% in short-term. However the TAN model is 62.5%. So it also improved in short-term. We all know to predict the amount of the precipitation is a sticky problem. The method gave in the article give a direction to the problem. The study gets some improvement on the meteorological prediction aspect.

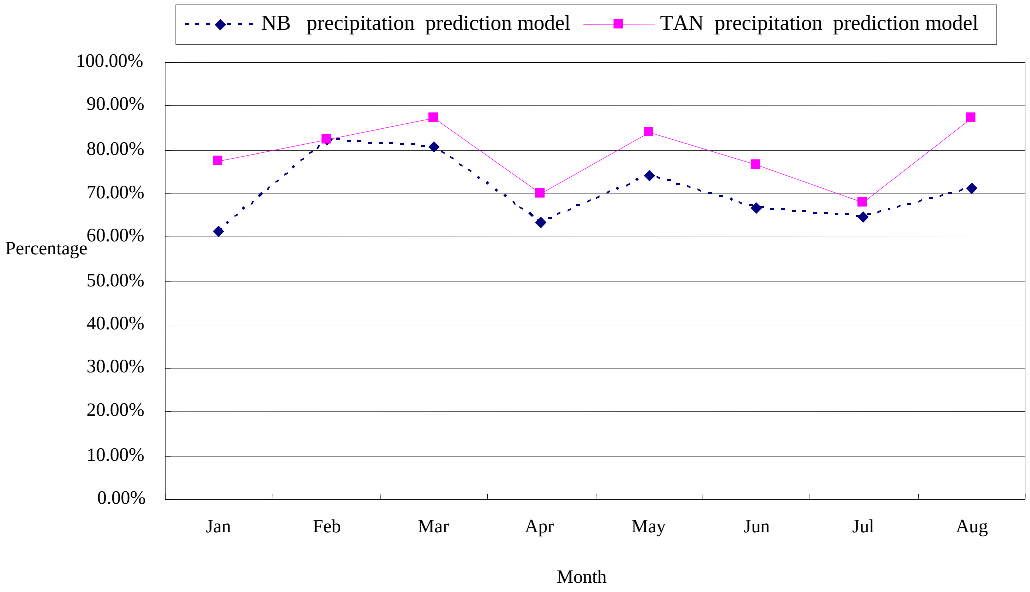

Figure 5. Predictability comparison chart of two precipitation prediction model

6. Conclusion

The paper finds the advantage of the TAN model in the precipitation prediction. The main work is as follows:

1. Construct the NB short-term precipitation prediction model and TAN short-term precipitation prediction model by using the meteorological data of Dongtai station from 1951~2005.

2. Make the 2006 meteorological data as the testing data and predict the precipitation level from January to August of two models. The prediction includes per-day and per-month. And make an experimental comparison between two models. And find that TAN model's prediction rate is higher than the original Naïve Bayes model except February.

In summary, because the model links the various predictors more closely, it is very suitable for meteorological factors. Atmospheric factors can not exist alone. TAN model grab the natural character in atmosphere, improve the original Naive Bayesian prediction model to optimize the experimental results.

But the TAN model still has some defects. The association between properties can be more complex. If we also consider the influence of the climate factors, such as rainy season---- the possibility of the light rain and the moderate rain happened surely higher than usual. Then we can get a more scientific prediction model.

Acknowledgments

This work is partly supported by National Natural Science Foundation of China (Grand Nos. 41275116) and Jiangsu Economic and Information Technology Commission project of China (Grand Nos.{2011}1178).

References

[1] Yeh YL, Chen TC. Grey degree and grey prediction of groundwater head. *Stochastic Environmental Research and Risk Assessmen (tSERRA)*. 2004; 18(5): 351-363.

[2] Rizkha Emillia N, Suyanto NFN, Maharani W. Isolated word recognition using ergodic hidden markov models and genetic algorithm. *TELKOMNIKA Telecommunication, Computing, Electronics and Control*. 2012; 10(1): 129-136.

[3] Shadaksharappa NM. Optimum Generation Scheduling for Thermal Power Plants using Artificial Neural Network. *International Journal of Electrical and Computer Engineering (IJECE)*. 2011; 1(2): 134-139.

[4] JR Quinlan. C4.5: programs for machine learning. San Mateo, Calif : Morgan Kaufman Publishers,1993

[5] RO Duda, PE Hart, DG Stork. Pattern Classification (2nd Edition). Wiley-Interscience. 2000.

[6] P Langley, W Iba and K Thompson. *An analysis of Bayesian classifiers*. In AAAI '90. 1992: 223-228.

[7] Written IH, Frank E. Data mining: Practical machine learning tools and techniques with Java implementation. Seattle: Morgan Kaufmann Publishers. 2006: 265-314.

[8] Zhihai Wang, GI Webb, F Zheng. *Adjusting Dependence Relations for Semi Lazy, TAN Classifiers*. Advances in Artificial Intelligence, Berlin Heidelberg: Spring-Verlag. 2003: 453-456

[9] Yager RR. An extension of the Naive Bayesian classifier. *Information Sciences*. 2006; 176: 577-588.

[10] Zhang H, Sheng S. *Learning weighted nave Bayes with accurate ranking*. The 4th IEEE International Conference on Data Mining. Chicago: IEEE Computer Society, 2004: 567-570.

[11] Cerquides J, De Màntaras R L. *Tractable Bayesian learning of tree augmented naïve Bayes models.* ICML. 2003: 75-82.

[12] Ramonim, Sebastianip. Robust Bayes classifiers. *Artificial Intelligence*. 2001; 125(1-2): 209-226.

[13] Luo Ke, Lin Mugang, Xi Dongmei. Review of classification algorithms in data mining. *Computer Engineering*. 2005; 31(1): 1-11.

[14] Y Zhang, CH Chu, Y Chen, H Zha, X Ji. Splice site prediction using support vector machines with a Bayes kernel. *Expert Systems with Application*. 2006; 30: 73-81.

[15] Chickering DM. Learning Bayesian networks is NP-complete. Learning from Data:AI and Statisitcs V. New York: Springer. 1996: 121-130.

[16] Chow CK, Liu CN. Approximating discrete probability distributions with dependence trees. *IEEE Transaction on Information Theory*. 1968; IT-14(3): 462-467.

[17] McCallum A, Nigam K. *A comparison of event models for naive bayes text classification*. AAAI-98 workshop on learning for text categorization. 1998; 752: 41-48.

[18] Cunningham WP, Saigo BW, Cunningham MA. Environmental science: A global concern. Boston, MA: McGraw-Hill, 2001.

[19] Friedman N, Koller D. Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks. *Machine learning*. 2003; 50(1-2): 95-125.