Research Journal of Applied Sciences, Engineering and Technology 5(3): 986-989, 2013
ISSN: 2040-7459; E-ISSN: 2040-7467
© Maxwell Scientific Organization, 2013
Submitted: June 23, 2012 Accepted: July 28, 2012 Published: January 21, 2013

3-Layered Bayesian Model Using in Text Classification

Chang Jiayu and Hao Yujie
Department of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China

Abstract: Naive Bayesian is one of quite effective classification methods in all of the text disaggregated models. Usually, the computed result will be large deviation from normal, with the reason of attribute relevance and so on. This study embarked from the degree of correlation, defined the node's degree as well as the relations between nodes, proposed a 3-layered Bayesian Model. According to the conditional probability recurrence formula, the theory support of the 3-layered Bayesian Model is obtained. According to the theory analysis and the empirical datum contrast to the Naive Bayesian, the model has better attribute collection and classify. It can be also promoted to the Multi-layer Bayesian Model using in text classification.

Keywords: 3-layered Bayesian model, coefficient of correlation, degree, matrix of correlation, naive Bayesian

INTRODUCTION

Classification is an important research topic in Machine Learning, which finds specified class tag and finally realizes the purpose of classification by constructing a Classifier. Such examples include decision tree, neural network, Bayesian method, etc. Bayesian method has become a popular research topic in Machine Learning because of factors including strong theoretical foundation in mathematics, prior probability information and rich sample information. The Bayesian network technology developed over the last decades is suitable to express and analyze things with uncertainty. With certain similarity to event tree or fault tree method in terms of reasoning mechanism and state description and with the ability to describe time polymorphism and non-deterministic logic, the Bayesian Model is very suitable for analyzing and determining whether they are two things. For this reason, it can be considered to apply in spam filtering. Bobbio *et al*. (1999, 2001), Xie *et al*. (2004) and Guang-Yan *et al*. (2004) discussed conversion methods from fault tree to Bayesian Model and proposed a method to convert AND gate, OR gate and Voting gate to Bayesian Model as well as a method to give probabilities of events on each node in Bayesian Model obtained from conversion. Xie *et al*. (2004) have a research of the improvement of faulty tree analysis by Bayesian networks. Guang-Yan *et al*. (2004) study the fault tree analysis based on Bayesian networks. Zhong-Bao *et al*. (2006) discussed conversion method from other logic gate to Bayesian Model in fault tree analysis. The solutions for minimal path sets and minimal partition sets importance degree based on Bayesian Model are provided. Hong-Bo *et al*. (2004) proposed a text classification algorithm based on restrictive 2-layered Bayesian Model, in which the first layer has only two nodes. However, nobody so far has performed research on the application of multi-layered Bayesian Model or Bayesian Model with multiple nodes in the first layer in text filtering. In this study, we'll take the advantage of multi-layered Bayesian Model to determine the spam to address the shortcomings of traditional methods.

This study embarked from the degree of correlation, defined the node's degree as well as the relations between nodes, proposed a 3-layered Bayesian Model. According to the conditional probability recurrence formula, the theory support of the 3-layered Bayesian Model is obtained. According to the theory analysis and the empirical datum contrast to the Naive Bayesian, the model has better attribute collection and classify. It can be also promoted to the Multi-layer Bayesian Model using in text classification.

TRANSFORMATION OF BAYESIAN FORMULA

Let $X_1, X_2,\ldots, X_n$ be components of vector X, i.e., $X = \{X_1, X_2,\ldots, X_n\}$ and $\{x_1, x_2,\ldots, x_n\}$ is one of values. From the Chain Guideline in probability theory $P(x_1, x_2,\ldots, x_n \mid A) = P(x_1 \mid x_2,\ldots, x_n, A) \cdot P(x_2,\ldots, x_n \mid A)$ the form can be expressed as formula (1):

$$p(x_1,...,x_n \mid A) = \prod_{i=1}^{n-1} p(x_i \mid x_{i+1},\cdots,x_n, A)\cdot p(x_n \mid A) \quad (1)$$

Definition: Let the attribute set $X = \{X_1, X_2,\ldots, X_n\}$ and the class variables C, $G_1 = \{X_{k_1}, X_{k_2}, \ldots,$

Corresponding Author: Chang Jiayu, Department of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China

$X_{k_m}\}$, $G_2 = \{X_{l_1}, X_{l_2}, \ldots, X_{l_m}\}$, $G_3 = \{X_{i_1}, X_{i_2}, \ldots, X_{i_{n-m-l}}\}$. Attribute set G_1, G_2, G_3 are partition of the attribute set X, where $X = G_1 \cup G_2 \cup G_3$ and $G_i \cap G_j = \emptyset$ $1 \leq i < j \leq 3$ both hold. Class C, connecting every node in Bayesian Model, represents that the occurrence of all nodes must be in the condition that Class C occurs. And the single arrows connecting between nodes are to indicate the order of attribute occurs.

Obviously, according to formula (1), the following can be obtained as formula (2):

$$p(G_1, G_2, G_3 \mid A) = p(G_3 \mid G_2, G_1, A) \cdot p(G_2 \mid G_1, A) \cdot p(G_1 \mid A) \quad (2)$$

According to the variant formula of Bayesian Theorem[6], it can be expressed as:

$$p(A \mid G_1, G_2, G_3) = \frac{p(G_3 \mid G_2, G_1, A) \cdot p(G_2 \mid G_1, A) \cdot p(G_1 \mid A) \cdot p(A)}{p(G_1, G_2, G_3)} \quad (3)$$

Considering that the value of $P(G_1, G_2, G_3)$ is fixed when given a set of fixed values of $x_1, x_2, \ldots, x_n$, let $\alpha = \frac{P(A)}{P(G_1, G_2, G_3)}$, we have:

$$p(A \mid G_1, G_2, G_3)$$

$$= \alpha \cdot p(G_3 \mid G_2, G_1, A) \cdot p(G_2 \mid G_1, A) \cdot p(G_1 \mid A) \quad (4)$$

where, α is a regularization factor. Thus, we have:

$$p(A \mid x_1, x_2, \cdots, x_n) = \alpha \prod_{i=1}^{n} p(x_i \mid F(x_i), A) \quad (5)$$

where, $F(x_i)$ is the set of father nodes of x_i. The mode with maximum relevant is selected criteria for direct computing, with the case of multiple parents nodes be considered. And no more than 3 father nodes have been chosen under considering the difficulty of calculation. Thus, the difficulty is decreased while making the implementation more convenient. For given instance $\{x_1, x_2, \ldots, x_n\}$, we should establish 3 attribute sets, G_1, G_2 and G_3, such that $\prod_{i-1}^{n-1} P(x_i \mid x_{i+1}, \ldots, x_n, A)$ is at its maximum value. The Bayesian Network under this condition is called Bayesian Optimal Network and the solution we get is the 3-layered Bayesian network optimal solution.

LAYERED BAYESIAN MODEL

Bayesian Model is the result of combination of probability theory and directed graph in graph theory, whose essence is a model of directed graph used to express and deduce uncertainty things. In fact, Bayesian Model is a weighted directed graph from the standpoint

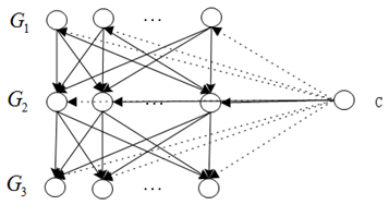

Fig. 1: 3-layered Bayesian model structure

of graph theory, in which the 'direct' between nodes means causal relationship between them. In the graph, both causes and consequences are represented by nodes that have their own probability distribution.

Definition: According to the definition of G_1, G_2, G_3 in Section 1, as well that any pair of nodes in G_3 are independent with each other. Any two nodes in G_2 have weak relevance, while any two nodes in G_1 have stronger relevance (Note that there are also edges in node sets G_1 and G_2, although we didn't draw them. Figure 1 shows the structure of 3-layered Bayesian Model.

Notice, edges in Fig. 1 are not necessarily to be existing. The figure is just a form of expression), node C is ancestor of every other node in the graph. The correlation between nodes represents occurrence condition, while the weight means strength of the correlation.

CONSTRUCTION OF 3-LAYERED BAYESIAN NETWORK MODEL

For construct 2-layered, 3-layered or multi-layered Bayesian Model, we first need to establish a relatively good classification model to split node set to three or multiple fairly ideal node sets. As for text attribute set $X_1, X_2, \ldots, X_n$, there are some correlations between attribute, or even some attribute occur simultaneously in text, i.e., linear dependence. In the light of statistics (Liang *et al*., 2002), we can use correlation coefficients of the sample to represent correlation between 2-dimensional whole population samples, that is, use a statistic R_{ij} to represent it, where R_{ij} is the sample correlation coefficient of 2-dimensional whole population (X_1, X_2). Thus we have the attribute correlation matrix *R*. Let:

$$R = \begin{bmatrix} R_{11}, R_{12}, \cdots, R_{1n} \\ R_{21}, R_{22}, \cdots, R_{2n} \\ R_{n1}, R_{n2}, \cdots, R_{nn} \end{bmatrix}$$

Definition: Each attribute has a correlation coefficient with other attribute. Let the sum of absolute values of all the correlation coefficients of one attribute with

other attribute be called as degree, denoted as D, then we have the degree of a node is $D_i = \sum_{k-1}^{n} | R_{ij} |$.

According the definition of degree, it's obvious we have following property:

Lemma 1: The lower of the degree of a node, the more independent this node from other nodes.

Lemma 2: The higher of the degree of a node, the more affected the node is by other nodes.

According to above theory, to attribute set node $X = \{X_1, X_2,\ldots, X_n\}$, the constructing steps of 3-layered Bayesian Model were written as following:

(1) Let $G_i = \varnothing$ (i = 1, 2, 3) and set threshold values ε_1, ε_2, where ε_1 is a miniature positive number and $\varepsilon_2 = n \times (1 - \varepsilon_1)$ and threshold value ε.

(2) To sample training set, abstract its Support Vector Machine (SVM). Let the vertex corresponding to each sample be $(x_{i1}, x_{i2},\ldots, x_{in})$ (i = 1, 2,…, n), where x_{ij}(j = 1, 2,…, n) is the value of the time of attribute i occurs in the sample times corresponding component.

(3) According to the refined sample vector space, statistically calculate correlation coefficient R_{ij}(i, j = 1, 2,…, n) between any pair of attribute in attribute set, to get relevant matrix R.

(4) Calculate each degree D_i (i = 1, 2,…, n).

(5) Conduct scans comparison to all nodes. Set the initial values and partition node set as following: If $D_i < \varepsilon_1$ then $X_i \in D_3$ else if $D_i < \varepsilon_2$ then $X_i \in D_1$ else $X_i \in D_2$.

(6) To any node $X_{k_i} \in D_3$ and any node $X_{m_i} \in D_2$, if $R_{k_i m_i} > \varepsilon$, then let X_{m_i} be the father node of X_{k_i}, with the weight of the corresponding edge $R_{k_i m_i}$; Likewise, do the same about the relationship between node pair in D_2, D_1, with 1.5 times correlation coefficient as the edge weight. There is no edge with both nodes are in D_3. As for any nodes X_i, X_j in D_2, if $R_{ij} > \varepsilon$, then let the node with larger degree be the father node of the other one, with the weight of $2 \times R_{i,j}$. Likewise, do the same operation to nodes in D_1, but the weight should be multiplied by 3 this time. This way we construct a 3-layered Bayesian Model.

(7) Calculate the performance of current classification and save the result *count*. Set terminal conditions *end1* = true, *end2* = true

(8) If *end1* then choose the node with maximum degree in D_2 and add it into node set D_1. Repeat Step 6 and 7, let the result be *count1*. If *count>count1* then *count* = *count1*, record this classification; else *end1* = false

(9) If end2 then choose the node with minimum degree in D_2 and add it into node set D_3. Repeat Step 6 and 7, let the result be *count2*. If *count>count2* then *count* = *count2*, record this classification; else *end2* = false

(10) Repeat Step 8 and 9, until getting an optimal classification solution to eventually form the structure of 3-layered Bayesian Model.

Table 1: Comparison between two kinds of Bayesian models

Model	Sample number	Max. attribute number	Chosen attribute number	Accuracy
Naive Bayesian	20000	64	40	64.2%
3-layered Bayesian	20000	120	90	75.3%

PERFORMANCE AND EXPERIMENTAL ANALYSIS

The two complexities is affected by the sample scale and used by the entire text filter algorithm base on SUM, so it is not considered here. The time complexity of Step 3 is O (n^2) and those of Step 4 and 5 are O(n), while Step 6 need O(n^2). In the best situation, we only need to run it for once to find the optimal classification, thus we have a O(n^2) time complexity in this situation. In the worst situation, however, it is possible that the value of ε_1 is too miniature such that everything should be re-selected, which lead us to at most $\frac{n}{2}$ - 1 iterations, thus the running time is O(n^3). This time complexity we calculated here is very close to that of other text filtering algorithm, so considering how it deals with data, especially miniature data, this time complexity is acceptable.

All of our experiments, running on Linux, are conducted upon attribute set and sample training set provided under Chinese rules. Experimental data is listed in following Table 1, which give the comparison between two kinds of Bayesian models:

When choosing attribute set, we consider mainly about wrong results caused by phenomena such as overflow due to the calculation process of a computer, thus we don't think we can provide a big data set to support Naive Bayesian Model, which is solved by Multi-layered Bayesian Model. In situations with relatively large number of attribute, we can handle potential calculating problems by increasing the number of attribute classifications to solve problems. In SVM theory, the more attribute set, the more ideal character express of text. From this point of view, Multi-layered Bayesian Model has more advantage than Naive Bayesian Model, which is also shown by experiment result.

CONCLUSION

When Naive Bayesian is applied in text classification, because of issues like correlations

between attribute and the attribute set produced by values of the covariance matrix in calculation are too small, we considered to classify attribute set into several categories before calculating. Starting from correlation, this study defines the concept of node degree and the relationship between nodes and classifies attribute sets in term of node degree, reduce the nodes that have low attribute concentration ratio and increase the nodes that have high attribute concentration ratio. By doing this, we strengthened the nodes that have strong constraints between attribute and weakened nodes that have weak constraints between attribute and thus we proposed 3-layered Bayesian Network Model structure. We also analyzed the usability of this model. With analysis based on experimental data, we conclude that this model has better attribute set to support and better classification result, with easy extensibility to multi-layered Bayesian Network Model.

REFERENCES

Bobbio, A., L. Portinale, M. Minichino and E. Ciancamerla, 1999. Comparing fault trees and Bayesian network for dependability analysis [A]. Proceeding of the 18th International Conference on Computer Safety, Reliability and Security [C]. Toulouse, France, pp: 310-322.

Bobbio, A., L. Portinatle, M. Minichino and E. Ciancamerla, 2001. Improving the analysis of dependable systems by mapping fault trees into Bayesian networks [J]. Reliab. Eng. Syst. Saf., 71(3): 249-260.

Guang-Yan, W., M.A. Zhi-Jun, H.U. Qi-Wei, 2004. The fault tree analysis based on bayesian networks [J]. Syst. Eng. Theory Pract., 6: 76-83.

Hong-Bo, S., W. Zhi-Hai, H. Hou-Kuan and L. Xiao-Jian, 2004. A restricted double-level bayesian classification model [J]. J. Software, 15(02): 193-199.

Xie, B., Z. Ming-Zhu, Y. Yu-Xian, 2004. Improvement of faulty tree analysis by bayesian networks [J]. J. Mianyang Normal University, 23(2): 29-33.

Zhong-Bao, Z., D. Dou-Dou and Z. Jing-Lun, 2006. Application of bayesian networks in reliability analysis [J]. Syst. Eng. Theory Pract., 26(6): 95-100.