ISBN: 978-1-948012-08-9

Contents List available at VOLKSON PRESS

International Symposium on Computer Science and Artificial Intelligence(ISCSAI)

The Research of Postoperative Life Expectancy of Lung Cancer Based on Semi-Naive Bayesian

Bei Hui[a], HongTing Zhou[a], YuNan Jiang[a], Lin Ji[b*], Jia Chen[a]

[a]*School of Information and Software Engineering, University of Electronic Science and Technology of China, North JianShe Road, ChengDu, China*
[b]*Department of radiology, West china hospital of SiChuan University, Guoxuexiang, ChengDu, China*
**Corresponding Author: Jilin2@sina.com*

ARTICLE DETAILS

Article History:

Received 02 october 2017
Accepted 06 october 2017
Available online 11 october 2017

Keywords:

Bayes, Info-NODE, lung cancer

ABSTRACT

Since lung cancer has a high fatality rate, it is really necessary to help surgeons predict the postoperative life expectancy of patients. In this paper, we are committed to discuss some semi-Naive Bayesian classifiers that will effectively predict the postoperative survival time of patients with lung cancer through combining the idea of information gain with assemble. The classifier with the best performance evaluation will be selected for helping surgeons to better judge their methods of treating patients.

1. Introduction

Lung cancer, also known as lung carcinoma, is a malignant lung tumor characterized by uncontrolled cell growth in lung tissue. Reported on, lung cancer happened to 1.8 million people and killed 1.6 million people around the world in 2012 [1,2]. Due to that data, lung cancer has been regarded as the most common cause of cancer-related death in men and the second most common in women after breast cancer.

Paper shows that of all people with lung cancer in the US, 16.8% survive for at least five years after diagnosis [3]. In England and Wales, the five-year overall survival rate of lung cancer is estimated at 9.5% in 2010. The English data suggests that around 70% of patients survive at least one year if diagnosed at the earliest stage, but it falls to only 14% for those who are diagnosed with the most advanced disease. Therefore, the prognosis of survival rate is significant for patients with lung cancer who received operation one year ago and the corresponding study is also very important.

In this paper, we want to establish a lung tumor computer-aided diagnosis model based on Bayesian classification to help surgeons predict the postoperative life expectancy of patients with lung cancer. We will elaborate on the method and compare its performance in the forecast with several other Bayesian Classifiers through the experiment.

In the second part, we will introduce the relevant principles of the Bayesian classifier and discuss the improved algorithm in detail. The third part describes the experimental environment and displays the results of the analysis. The fourth part is a summary and future work, and then the last part is acknowledgment.

2. BAYESIAN CLASSIFIER

2.1 NB Classifier

For given instance, the Naive Bayesian (NB) Classifier, regards that all attributes are independent of each other [4]. Under this assumption, we can approximate $P(X|y)$ as formula (1). Where x_i is the i-th attribute of the X attribute vector.

$$P(X|y) = \prod_{i=1}^{n} P(x_i|y) \qquad (1)$$

The working mechanism of the NB classifier is given by the formula (2). $\hat{P}(y)$ and $\hat{P}(x_i|y)$ are two evaluators, which calculate the probability of a class and the probability of each individual attribute in the category.

$$\underset{y}{\operatorname{argmax}}(\hat{P}(y)\prod_{i=1}^{n}\hat{P}(x_i|y)) \qquad (2)$$

3. Semi-Naive Bayesian Classifier

3.1 AODE Classifier

NB assumes that the attributes are independent of each other. Averaged One-Dependence Estimators (AODE), makes a weakening of the attributes independently [5]. It selects each attribute as a parent attribute one by one, that is, other attributes are dependent on the parent. The post probability of each class label is calculated under every parent. Then it averages post probability for each class label.

$$\underset{y}{\operatorname{argmax}}(\sum_{i:1\le i\le m)}\hat{P}(y,x)\prod_{j=1}^{n}P(x_i|y))) \qquad (3)$$

Formula (3) describes the AODE. The m is number of attributes. The parent function only needs to traverse all the other attributes without correlation analysis, so it improved a lot of performance.

2.2.2 Info-NODE

Based on AODE Classifier, the info-NODE method is proposed in this paper. Because computing resource is limited in some scenario, Info-Node releases restriction of the number of super-parents. It selects some attributes to become parent attributes involved in the calculation. Considering the selecting super-parent method, Info-NODE computes the information gain for all attributes and sorts them by their info-gains descending order. And then the attribute is divided into two sets—parent attributes (sp set) and general attributes (child set) by parameter N. We add the first N attributes of the sorted attribute array to sp set. The N means the number of parent attributes. Take 1 to 16 to calculate the best value for Info-NODE.

Table 1: Function: info-sort (Attrs)

1 :	**for** each $p_i \epsilon Attrs$ in turn **do**
2 :	$\mathrm{H}(Y) = -\sum_{i=1}^{n} p(y_i) \log_b p(y_i)$
3 :	$\mathrm{H}(\mathrm{Y}\vert\mathrm{X}) = \sum_x p(x) H(Y\vert X = x)$
4 :	$\mathrm{Gain}(\mathrm{pi}) = \mathrm{H}(\mathrm{Y}) - \mathrm{H}(\mathrm{Y}\vert\mathrm{X})$
5 :	**end for**
6 :	**sort(Attrs);**
7 :	**return Attrs;**

Table 1 shows how to sort the attributes. Traverse from the 1st to 5th line of the various attributes to calculate the info-gain. The entropy of the attribute is calculated at the 2nd line, and the entropy of the attribute is the measure of the uncertainty of the random variable in the information theory and the probability theory. The 3rd line calculates the conditional entropy of the attribute. Finally subtract the entropy from the entropy to obtain the information gain of the attribute. The 6th line calls the sorting algorithm to use the attribute's information gain size to sort the attributes.

In table 2, lines 1 to 3 compute NB by formula (2). The required posteriori probability is calculated by the previously known probabilities and conditional probabilities. And then call the info − sort function on the 4th line to sort the attributes of the dataset based on information gain. The input Attrs set is the original attribute sequence. The 6th to 11th rows pick the first N attributes after the sort as the parent set (sp) and take the above set of attributes in turn as a super-parent to calculate the probability of class label values to get the average value of the results. Meanwhile the TP(y) is the temporary probability. Finally, output the estimated probability $P[\ c_1 \dots c_k\ ]$.

Table 2: Algorithm: Info-NODE

1 :	**for** $y := c_1 \dots c_k$ **do**
2 :	**calculating the NB probability by formula(2)**
3 :	**end for**
4 :	**Call** $\boldsymbol{info - sort(Attrs)}$;
5 :	$count := 1$;
6 :	**for** $each\ sp \in \{a_1 \dots a_N\}$ **do**
7 :	$count++$;
8 :	**for** $y := c_1 \dots c_k$ **do**
9 :	$\boldsymbol{TP(y) = TP(y) + P(y\vert X, sp)}$
10 :	**end for**
11 :	**end for**
12 :	$P = TP(y)/count$
13 :	**return** P**;**

4. EXPERIMENT AND ANALYSIS

We use the Weka system to implement out algorithms [6]. Weka is a collection of machine learning algorithms for data mining tasks. And the data set will be divided into 10 subsets as 10-fold cross validations, which prove to be the most effective.

4.1 Experiment Data

The Thoracic Surgery dataset comes from the UCI Machine Learning Repository, which is a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms [7,8]. The table 3 shows that the data set has 470 instances. Each instance has 16 attributes and 1 class label.

Table 3: Thoracic Surgery Dataset

Ins.	Attr.	Class label	Missing
470	16	2	None

The data was collected retrospectively at Wroclaw Thoracic Surgery Centre for patients, who underwent major lung resections for primary lung cancer in the years from 2007 to 2011 [7,8].

4.2 Result and Analyze

There are two statistical values used to evaluate the performance of each classifier: the 0-1 loss classification error rate and Root Mean Squared Error(RMSE).

From the table 4, we can clearly see that most of the Bayesian classifier has a good performance. Except for NB, all other classifiers have a 0-1 loss ratio below 18%. The NB is 21.4894%. The AODE is 17.4468%. The lowest ODE is 17.0213 when parent is PRE8 or PRE30. The best results appear in the Info-NODE classifier with n = 9, which is 16.383%. On the other hand, in terms of RMSE, the best results still appear in the Info-NODE classifier with n = 10, which is 0.3581.

Table 4: Experiment result

Classifier		0-1 loss error	RMSE	Classifier		0-1 loss error	RMSE
NB		21.4894	0.3958	AODE		17.4468	0.3594
	n=1	17.8723	0.3711		DGN	17.8723	0.3711
	n=2	17.6596	0.3702		PRE4	17.8723	0.3636
	n=3	17.8723	0.3671		PRE5	17.8723	0.3636
	n=4	18.0851	0.3661		PRE6	17.4468	0.3624
	n=5	17.6596	0.363		PRE7	18.0851	0.3637
	n=6	16.8085	0.3621		PRE8	17.0213	0.3657
	n=7	16.8085	0.3619		PRE9	17.8723	0.3656
	n=8	16.383	0.36		PRE10	17.234	0.3649
I-NODE	n=9	16.383	0.3582	ODE	PRE11	17.234	0.3639
	n=10	16.8085	0.3581		PRE14	17.8723	0.3687
	n=11	16.8085	0.3583		PRE17	18.2979	0.3662
	n=12	17.4468	0.3586		PRE19	17.8723	0.3637
	n=13	17.4468	0.3588		PRE25	18.2979	0.3639
	n=14	17.4468	0.359		PRE30	17.0213	0.3629
	n=15	17.4468	0.3592		PRE32	18.0851	0.3635
	n=16	17.4468	0.3594		age	17.8723	0.3636

By combining the results of the two statistic comparisons, we obtained the best results at Info-NODE for n = 9. Although n = 10 RMSE has the best results, in statistical terms, the difference in the RMSE value of 0.01% has no effect. What's more, n = 10 consumes more resources in the calculation. So, in this case we are more inclined to choose n = 9.

5 CONCLUSION AND FUTURE WORKS

In summary, we constructed a good Bayesian classifier. First, it sorts the attributes by the information gain of each attribute. Then pick the first n attributes added to the parent attribute set. The n attributes in turn as the parent attribute calculated probability value and return the average as a result. Therefore, we recommend using the Info-NODE with parameter n = 9 to build a computer-aid treatment tool for this case.

In Info-NODE, we tried all the values of n. In later studies, we will explore whether the best value is constant in a particular interval and reduce the complexity as well as the consumption of the algorithm.

ACKNOWLEDGMENTS

The research work was supported by the Fundamental Research Funds for the Central Universities under Grant No. ZYGX2016J092, and the Fundamental Research Funds for the Central Universities under Grant No. ZYGX2015J068.

REFERENCES

[1] Kris, M.G., Johnson, B.E., Berry, L.D. 2014. Using multiplexed assays of oncogenic drivers in lung cancers to select targeted drugs. JAMA, 311 (19), 1998–2006.

[2] WHO. 2014. World Cancer Report 2014. World Health Organization Chapter 1.1. ISBN:92-832-0429-8.

[3] National Cancer Institute. 2016. Surveillance, Epidemiology and End

Results Program. https://seer.cancer.gov/statfacts/html/lungb.html.

[4] Karthick, G., Harikumar, R. 2017. Comparative performance analysis of Naive Bayes and SVM classifier for oral X-ray images. 4th International Conference on Electronics and Communication Systems (ICECS), 88–92.

[5] Webb, G. I., Boughton, J., Wang, Z. 2005. Not So Naive Bayes: Aggregating One-Dependence Estimators. Machine Learning, 58 (1), 5–24.

[6] Witten, I.H., Eibe, F., Hall, M.A. 2011. Data mining: practical machine learning tools and techniques—3rd ed. Morgan Kaufmann.

[7] Lubicz, M., Pawelczyk, K., Rzechonek, A., Kolodziej, J. 2013. Thoracic Surgery Data Data Set. http://archive.ics.uci.edu/ml/datasets/thoracic+surgery+data.

[8] Zięba, M., Tomczak, J.M., Lubicz, M., Świątek, J. 2014. Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients. Applied Soft Computing Journal, 14 (1), 99-108.

Cite the article: Bei Hui, HongTing Zhou, YuNan Jianga, Lin Ji, Jia Chen (2017). The Research of Postoperative Life Expectancy of Lung Cancer Based on Semi-Naive Bayesian ,*International Symposium on Computer Science and Artificial Intelligence*, 1(3) :17-19.