ISBN: 978-1-948012-08-9

Contents List available at VOLKSON PRESS

Computer Science and Artificial Intelligence(CSAI)

DOI : http://doi.org/10.26480/iscsai.01.2017.20.22

The Research of predicting the Cervical Cancer Basing on Bayesian Algorithm

Bei Hui[a], YuNan Jiang[a], HongTing Zhou[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, 610054, 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:

Cervical cancer, Bayesian Classification, Decision Tree, SVM

ABSTRACT

Cervical cancer remains a significant cause of mortality in low-income country. It is necessary to find a most effective machine learning algorithm to predict the cervical cancer. Our research attempts to help doctor to judge whether a patient develops cervical cancer. Thus, we use Bayesian Classification, decision tree and Support Vector Machine to analysis the dataset. As a result, we compared 0-1 loss classification error rate of these algorithm, whose error rate is the least so it is the best. And the conclusion is that AODE is the most suitable algorithm of these.

1. Introduction

Cervical cancer is a cancer rising from the cervix basing on [1]. It is due to the abnormal growth of cells that have the ability to invade or spread to other part of the body. Worldwide, cervical cancer is both the fourth-most common cause of cancer and the fourth-most common cause of death from cancer in women. In 2015, there are 429200 cases of cervical cancer in China, with 281400 deaths referencing [2]. This is about 8% of the total cases and total deaths from cancer. About 70% of cervical cancers occur in developing countries. In low-income countries, it is the most common cause of cancer death. In developed countries, the widespread use of cervical screening programs has dramatically reduced rates of cervical cancer.

Because of the seriousness of cervical cancer, we did this research to find the best algorithm of the prediction of indicators/diagnosis of cervical cancer. Some Machine Learning methods are applied to this research, including Bayesian Classification, Support Vector Machine(SVM), Decision Tree, etc.

In the second section, three algorithms are introduced which are Bayesian Classification including Naive Bayes and Averaged One-Dependence Estimators (AODE), Decision Trees, SVM and and write their pseudo code. What's more, in the third part, we will make a brief description of the experiment, including environment, dataset and result analysis. In the last part, we give a conclusion and think about more research work in future.

2. ALGORITHMS

2.1 Bayesian Classification

In machine learning, Bayesian classification is a collection of probabilistic classifiers based on applying Bayes' theorem between the features. There are many kinds of Bayesian classification, and we will use two of them in this research, which are Naive Bayes and AODE.

2.2 Naive Bayes

Because the calculating scale of Bayes rule will grow rapidly with the increasing number of attributes. It often uses Naive Bayes, instead of Bayes rule [3]. The NB release the dependence of attributes. It regards that every attribute is independence each other.

Formal (1) gives the classifier model basing on NB. Among y is the event that patients develop cervical cancer and X is patients vector, for example, $X =< x_1, \cdots, x_m >$. What's more, x_1 is the value of the i[th] attribute.

$$argmax_y\left(\hat{P}(y)\prod_{i=1}^{n}\hat{P}(x_i|y)\right) \quad (1)$$

$\hat{P}(y)$ and $\hat{P}(x_i|y)$ are estimates of the respective probabilities derived from the frequency of their respective arguments in the experiment dataset. Table 1 gives the detail of NB algorithm.

Table 1: NB Algorithm

```
1   int attIndex = 0;
2   while (attributes are existing) {
3     if (current instance has not a missing attribute value) {
4       for (possible class value){
5         calculate the probability
6         multiply all the probability of each attributes—put this in array
    prob[j]
7       }
8     }
9     attIndex++;
10  }
11  return NB probabilities P̂(y)
```

2.2.1 AODE

However, except in the case of applying NB to classify small numbers of examples for dataset, this is achieved at considerable computation cost. AODE, is a technique that weaken NB's attribute independence assumption [4]. The classifier selects the class by formula (3). And algorithm detail is showed in Table 2.

$$\text{arg}max_y\left(\sum_{i:1\le i\le n\wedge F(x_i)\ge m}\hat{P}(y,x_i)\prod_{j=1}^{n}\hat{P}(x_j|y,x_i)\right) \quad (3)$$

Table 2: AODE Algorithm

```
1   for (each possible class value) {
2     for (each parent attributes) {
3       determine correct index for the parent in m_CondiCounts matrix
4       if (the attribute value doesn't have a frequency of m_Limit or
    greater) {
```

```
5          continue;
6        }
7        calculate the prior probability using formula (3)
8        for (each attributes) {
9          calculate the probability using formula (3)
10       }
11       add this probability to the overall probability
12       unblock the parent
13     }
14     if (at least one is not parent) {
15       do plain naive bayes conditional probability
16     } else {
17       divide by numbers of parent attributes to get the mean
18     }
19   }
20   return AODE probabilities P(y)
```

2.3 Decision Trees

A decision trees is a decision support tool that uses a tree-like graph of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

There are three kinds of decision trees algorithms, which are ID3, C4.5 and CART. In our research, C4.5 algorithm, showing in table 3, is used to selects leaf node by calculate information gain ratio.

Table 3: C4.5 Algorithm

```
1   while (current node is not pure) {
2     Calculate entropy of information.
3     Calculate entropy of information of each attributes.
4     Calculate information gain.
5     Calculate split information.
6     Calculate information gain ratio.
7   }
8   end while
9   Select current node as leaf node.
10  return sorts of leaf node
```

2.4 SVM

Support Vector Machine is a category of supervised learningmodels with associated learning algorithms that analyze data used for classification and regression analysis ,which is proposed in [5].

Given a set of training instances, each marked as belonging to one class label y_i , an SVM training algorithm builds a model that assigns new instance to one class label, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting). An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New instance is then mapped into that same space and predicted to belong to a class label based on which side of the gap they fall.

3 EXPERIMENT AND RESULTS ANALYSIS

3.1 Dataset

We use a dataset which focuses on the prediction of indicators of cervical cancer in UCI machine learning repository [6]. The dataset was collected at 'Hospital Universitario de Caracas' in Caracas, Venezuela, as described in Table 4. The dataset comprises demographic information, habits, and historic medical records of 858 patients. In this dataset, the class label is biopsy, which has two values –0 and 1. While biopsy value is 0, it means that the patient does not develop cervical cancer in the current instance. On the contrary, the patient has developed cervical cancer if biopsy value is 1. 793 of 858 patients do not develop cervical cancer and 65 of 858 patients have cervical cancer.

Table 4: Information of this data set

Number of instances	Number of Attributes	Number of cervical cancer patient	Percentage of cervical cancer patient
858	36	65	7.6%

Several patients decided not to answer some of the questions because of privacy concerns (missing values).

3.2 Experiment Environment

During this experiment, we use waikato Environment for Knowledge Analysis(weka) to exploring data. Weka, is open source software issued under the GNU General Public License [7]. It is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. Ten cross-validation is used in the research due to it proved to be the most effective.

4 RESULTS AND ANALYSIS

There are various approaches to determine the performance of classifiers. We used six of them in this paper, which are 0-1loss classification error rate, kappa statistics, mean absolute error, Root mean squared error, Relative absolute error and Root relative squared error. 0-1 loss classification also called classification accuracy. It counts the proportion of correctly predicted examples in a test dataset. The more correct examples, the smaller 0-1 loss classification error rate is. The kappa statistic measures the agreement of prediction with the true class -- 1.0 signifies complete agreement. The error values that are shown, e.g., the root of the mean squared error, indicate the accuracy of the probability estimates that are generated by the classification model. Table 5 shows different results of these three algorithms.

Algorithm	0-1 loss classification error rate	kappa statistics	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error
NB	0.113054	0.4052	0.1123	0.3102	0.928106	1.266546
AODE	**0.036131**	**0.6799**	**0.0435**	**0.1709**	**0.359264**	**0.697497**
Decision trees	0.048951	0.6118	0.0593	0.1934	0.490121	0.789499
SVM	0.064103	0	0.0641	0.2532	0.529885	1.033603

Table 5: Results of this reach

By comparing with each parameter, we can draw a conclusion that AODE is the best algorithm because it has the least 0-1 loss classification error rate between theses four algorithm which is 3.6131%. Also, its mean absolute error is 0.0435, root mean squared error is 0.1709, relative absolute error is 0.359264 and root relative squared error is 0.697497 and these parameters are the least one. These parameters are all the smaller the better.

Furthermore, its kappa statistics is the biggest one, which is 0.6799 and the bigger the better.

5. FUTURE WORKS

In this work, we do a comparison with some algorithms to find the best way of predicting cervical cancer. Future we expect to use more algorithms to estimate the dataset in order to find a better way than AODE.

ACKNOWLEDGEMENT

The research work was supported by the Fundamental Research Funds for the Central Universities under Grant No. ZYGX2016J092, the Sichuan Science and Technology Project under Grant No. 2017GZ0318, and the Fundamental Research Funds for the Central Universities under Grant No. ZYGX2015J068.

REFERENCES

[1] Cervical Cancer Treatment. 2014. (PDQ®)". NCI. 2014-03-14. Archived from the original on 5 July 2014. Retrieved 24 June.

[2] Chen, W. 2015. Cancer statistics in China, 2015". Doi: 10.3322/caac.21338

[3] Hand, D.J., Yu, K. 2001. Idiot's Bayes — not so stupid after all?. International Statistical Review, 69 (3), 385–399. ISSN 0306-7734. doi:10.2307/1403452.

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

24. doi: 10.1007/s10994-005-4258-6

[5] Cortes, C., Vapnik, V. 1995. Support-vector networks. Machine Learning. 20 (3), 273–297. doi:10.1007/BF00994018.

[6] http://archive.ics.uci.edu/ml/datasets/Lymphography

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