Austin Journal of Biomedical Engineering

Austin Publishing Group

Open Access Full Text Article

Review Article

Bayesian Models for Healthcare Data Analysis

Xiaoshan Xie[1]*, Gang Zhang[1], Ying Huang[1] and Shanxing Ou[2]

[1]School of Automation, Guangdong University of Technology, China

[2]Department of Radiology, Guangzhou General Hospital of Guangzhou Military Command, China

***Corresponding author:** Xiaoshan Xie, School of Automation, Guangdong University of Technology, Guangzhou, 510006, China

Received: May 15, 2014; **Accepted:** June 16, 2014; **Published:** June 18, 2014

Abstract

The rapid increasing amount of healthcare data poses great challenges to data mining and machine learning study and applications. Recently a large number of algorithms and models have been proposed to discover knowledge and information from large scale healthcare datasets. In medical applications, confidence measured by posterior probability is well accepted since it can quantify the certainty or severity of targets. In this article, we propose a sparse Bayesian model for healthcare data analysis. The proposed model utilizes a set of basic functions and it learns a sparse weight vector to combine them together. Our model is a fully Bayesian method which can incoporate a prior and derive a likelihood function from a given training data set. Working with the images of Pulmonary Embolism diagnosis dataset and Breast Cancer clinical dataset from KDDCup, our experiments demonstrate that the Bayesian approach lead to 83% and 80% test accuracy in modeling principles of healthcare data and it significantly improves the performance of its couterparts.

Introduction

With the increasing availability of biomedical and healthcare data with a wide range of sophisticated characteristics, healthcare data analysis has been an popular and challenging work in recent years. Therefore, a large number of algorithms in data mining have been proposed to model the uncertainties that come with the problem, including Decision Tree (DT), Neural Network (NN), Bayesian methods, association rule mining and so on. Currently, benefit from natural advantages of mining and learning in recognizing significant facts, relationships, trends and anomalies, mining and learning techniques have been widely applied in healthcare domain [1,2]. As early as 1997, to improve the quality of care as well as to help control spiraling costs in healthcare industry, Rogers et al. [3] applied the SAS technology to solve critical bussiness solutions with the healthcare industry. Moreover, Sellappan et al. [4] developed a web-based Intelligent Heart Disease Prediction System (IHDPS) by using Decision trees, Naive Bayes and Neural Network, which was considered as one of a prominent model [5]. And it also can be implemented to better understand key indicators involving quality outcomes and encounters of care. Liu Peng et al. [6] proposed to utilize decision tree, Naive Bayesian classifiers and feature selection methods to predict inpatient length of stay. A PSO-SVM based on association rules in automatic detection of erthemato-squamous diseases obtain higher accuracy [7,8]. And detection of fraudulent insurance claims, making better health policy, forecasting treatment costs are also applications of data mining in healthcare domain [9,10]. Nevertheless, according to the survey of [11], few data mining methods are treated as practically valuable tools for clinical purposes. To better solve these issues, Bayesian method has been attached more importance in theorical study and some new algorithms based on it have been proposed to solve practical problems.

Bayesian method is the powerful one that emerges as a method for discovering patterns in biomedical data and has better speed and accuracy for huge datasets [8,12,13]. Naive Bayesian Classifier (NBC) uses probability to represent each class and trends to find the most possible class for each sample, which always performs well in practice [6]. And the Naive Bayesian Imputation (NBI) proposed in [14] is used for missing data handling. Zhao et al. [15] proposed a Bayesian-based Personalized Laboratory Test prediction (BPLT) to predict laboratory tests for a given group of patients. By considering the aquisition of data from different sources, Martijin described a new formalism named multilevel Bayesian networks for the analysis of hierarchical health care data [16].

In this article, we attempt to construct a sparse model based on Bayesian learning methods. The proposed method tries to model the generative principles of the target data set. Mathematically, we often express a generative model as following:

$$y = w^t \Phi(x) + \varepsilon \qquad (1)$$

where $y \varepsilon R$ is a target variable and $x \in R^d$ is a d– dimension feature vector. Φ is a set of basic functions, where $\Phi_i(x): R^d \to R$. ε is Gaussian noise with zero mean and unknown variance. In

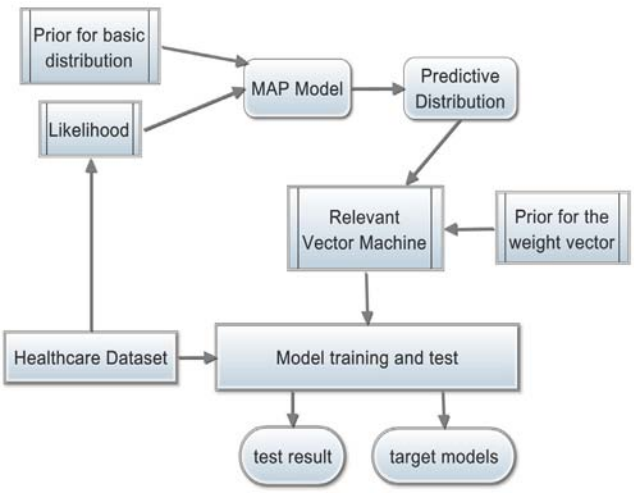

Figure 1: The main idea of sparse Bayesian learning.

Xie et al. © All rights are reserved

Citation: Xie X, Zhang G, Huang Y and Ou S. Bayesian Models for Healthcare Data Analysis. Austin J Biomed Eng. 2014;1(3): 1013.

this work, we limit Φ to be a set of random initialized Gaussian distributions. The goal is to derive the posterior distribution $p(w|D)$ given training dataset D and the predictive distribution $p(y|x,D)$ given training dataset D and a test example x. Moreover, to cut down the computational cost of both training and test, a sparse combination is preferred, meaning that in the weight vector w, there are a lot of elements are zero. We will show that the problem of finding a sparse weight vector can be solved by a Relevant Vector Machine (RVM), which is a Sparse Bayesian Learning (SBL) model [17,18]. Figure 1 sketches the main idea of this article.

The remainder of this article is organized as following. In Section 2 we formally give the sparse Bayesian model. In Section 3 we present the evaluation results of the propsed model compared with some recent methods. And finally we conclude the article in Section 4.

The Sparse Bayesian Model

We aim at building a model to capture the predictive distribution of the prediction target. Let $\Phi = \{\varphi_1, \varphi_2, ..., \varphi_m\}$ be a set of known basic distributions. Given a training dataset $D = \{(x_i, y_i), i = 1, ..., N\}$, the likelihood can be expressed as $p(y|x,w)$. And if we introduce a proper prior, we can get the Maximum a Prior (MAP) distribution. Thus the optimal w^* can be expressed as:

$$w^* = \arg\max_w p(y/w,x)p(w) = \Phi^T(\lambda I + \Phi\Phi^T)^{-1}y \quad (2)$$

where λ is a square ratio between the variance of ε and w. $p(y|w,x)$ is the likelihood function and $p(w)$ is the prior for the weights w. It is well known that in SBL an ARD prior is often imposed on w by introducing a set of parameter α [19]. Thus we have:

$$p(w/\alpha) = \prod_{i=0}^{N} N(w_i/0, \alpha_i^{-1}) \quad (3)$$

where α is the precison controlling each element of w. If α_i is infinity, the corresponding w_i would be driven to zero.

The likelihood $p(y|x,w)$ can be written as following:

$$p(y/x,w) = \prod_{i=0}^{N} p(y_i/x_i,w) \quad (4)$$

where $p(y_i/x_i,w)$ is the probability of each sample. Since the distribution is controlled by two parameters, we can use point estimation method to evaluate them.

According to Bayesian formula, we can derive the posterior distribution with likelihood, prior and the evidence as following:

$$p(w|D,\alpha,\varepsilon)=N(w|m,\Sigma) \quad (5)$$

where $m = \varepsilon\Sigma\Phi^T y$ and $\Sigma = (diag(\alpha_i) + \varepsilon\Phi^T\Phi)^{-1}$. Note that Φ can be a kernel matrix if we make use of a Gaussian process prior. In our work, we use a radius basic function kernel over D to generate such prior, such that $(\Phi)_{ij} = k_{RBF}(x_i, x_j)$ where $k_{RBF} = \exp(-d(x_i, x_j)^2)$. The optmial α and ε can be determined by solving a type-2 maximum likelihood problem, where we have:

$$p(y/x,\alpha,\varepsilon) = \int p(y/x,w,\varepsilon)p(w/\alpha)dw \quad (6)$$

According the formula of the convolution of two normal distribution, the above marginal likelihood can be solved analytically in its logarithm form as following:

[Sorry. Ignored \begin{aligned} ... \end{aligned}] (7)

Where $C = \varepsilon^{-1}I + \Phi diag(\alpha_i)^{-1}\Phi^T$. The optimal α and ε can be obtained through a iteration procedure, where we have:

$$\alpha_i^* = \frac{\gamma_i}{m_i^2} \quad (8)$$

$$(\varepsilon^*)^{-1} = \frac{\| y - \Phi m \|^2}{N - \Sigma_i \gamma_i} \quad (9)$$

$$\gamma_i = 1 - \alpha_i \Sigma_{ii} \quad (10)$$

An important thing should be noticed is that during the iteration procedure, a large number of α_i would be driven to infinity, leading to only a small amount of non-zero w_i. Hence we obtain a sparse model. The model sparsity is originated from the fitness between basic distributions and the groundtruth distribution implied in the training dataset D. If a basic distribution does not go along with some direction of the groundtruth, it will be gradually driven out by increasing the corresponding α.

After we find the optimal α and ε, we can derive the predictive distribution given a test sample x_η. The predictive distribution is the marginal of w. It is the integration on w with the convolution of posterior and prior. We have:

[Sorry. Ignored \begin{aligned} ... \end{aligned}] (11)

Where $\delta^2(x_\eta) = (\varepsilon^*)^{-1} + \varphi(x_\eta)^T\Sigma(x_\eta)$. Hence we can get a predictive distribution of a given test example x_η. However, there is still one thing to be addressed. The proposed model is naturally based on a regression setting, meaning that the target variable $y \varepsilon R$. In healthcare data analysis, the target variables are discreate in many cases. To make the proposed method suitable for classification, we introduce a sigmoid function into our model. A sigmoid function can typically be expressed as $f(x) = \frac{1}{1+e^{-x}}$, whose range is (0,1]. Each input can be compressed to a standard range. We use sigmoid function to convert the output of our model to (0,1] and then use a discriminate function to get a class label.

Evaluation and Results

Dataset description

We evaluate the proposed method on two healthcare datasets from KDDCup, a famous data mining competition. The first is a Pulmonary Embolism (PE) diagnosis dataset, and the seccond is a Breast Cancer clinical data set. We briefly denote these two datasets as $M1$ and $M2$. For $M1$, the target is to classify whether an individual has PE given an image. A total of 4429 candidates were identified in the candidate generation procedure: 3038 candidates appear in the training set, and 1391 appear in the test set. Each candidate is a cluster of voxels (the 3-D analog of pixels) with gray values for each voxel in the cluster. Each candidate was then l abeled as a PE or not based on proximity to a 3-D landmark provided by an expert. The image is expressed as a 69-ary feature vector. The detail of $M1$ is listed in Table 1. For $M2$, the analysis target is to identify whether a patient has breast cancer. A breast cancer screen typically consists of 4 X-ray images; 2 images of each breast from different directions (MLO and CC). Each image is represented by several candidates . For each candidate, there are image ID and the patient ID, (x,y) location, several features, and a class label indicating whether or not it is malignant. For convenience, the dataset has been preprocessed and each sample has a vector-form.

Table 1: Description of *M*1.

No.	Name	Type	Description
1	Patient ID	Number	4 bits
2	Label	Boolean	whether or not there is PE
3	Size feature	Real	image size
4	spatial shape feature	Real	3-ary vector
5	Location feature	Real	2-ary vector
6	neighborhood intensity feature	Real	2-ary vector
7	simple intensity statistic	Real	4-ary vector
8	neighborhood feature	Real	4-ary vector
9	neighborhood intensity feature	Real	18-ary vector
10	shape feature	Real	26-ary vector
11	anatonical feature	Real	4-ary vector
12	neighborhood feature threshold	Real	2-ary vector
13	intensity contrast feature	Real	5-ary vector
14	Shape neighbor feature	Real	34-ary vector

Table 2: Description of *M*2.

No.	Name	Type	Description
1	Ground truth label	Boolean	+1/-1
2	Image-Finding-ID	Number	a unique non-negative identifier
3	Study-Finding-ID	Number	identify a lesion
4	LeftBreast	Boolean	if the candidate was generated from the left breast
5	MLO	Boolean	from an MLO image or not
6	X-location	Number	X-pixel location
7	Y-location	Number	Y-pixel location
8	X-nipple-location	Number	X-pixel nipple location
9	Y-nipple-location	Number	Y-pixel nipple location

Table 3: Examples of raw clinical data in *M*1.

Feature examples	F1	F2	F3	F4	...	F115
No.1	3000	0	336	284	...	0.024697524
No.2	3000	1	328	287	...	0.006415986
No.3	3002	-1	250	275	...	-0.31899535

Table 2 lists the detail of *M*2. And three examples of raw clinical data have been presented in Table 3. The first one and the second one set are from train data, and the third one is from the test data. Note that the first two columns supply the patient identifier and the PE identifier. The PE indentifier is also our target lable variable. If it is a PE, the lable is a positive number, if it is not a PE, the label is set to 0. In the test data, all labels are set to -1, which means unknown.

Evaluation

We perform two kinds of evaluation to illustrate the effectiveness of the proposed method. The first is to evaluate the performance of the proposed model compared with the traditional Bayesian Model (BM) and Support Vector Machine (SVM) classifier. The second is to illustrate how sparse our model is, and the relationship between the model sparsity and the performance. For both cases, we use the same setting as following. The whole dataset is randomly divided into training set and test set with ratio 3:7. We only consider two classes

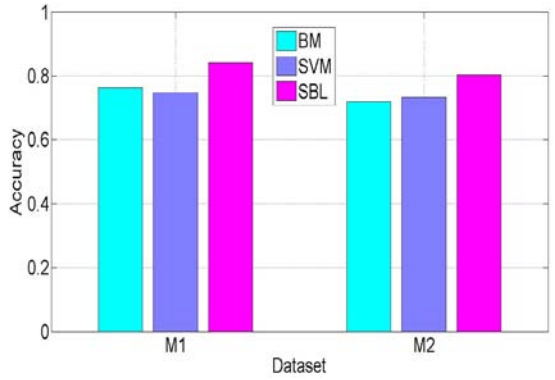

Figure 2: Evaluation results on two datasets.

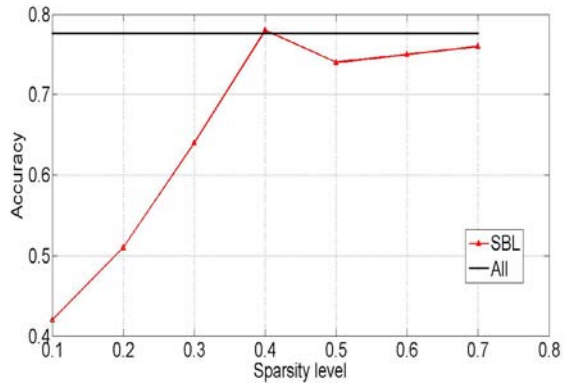

Figure 3: Relationship between model sparsity and accuracy.

classification problem. Since the two datasets are multiple classes, we convert them into *m* sub problems, each problem is a one-versus-rest classification problem. For the compared Bayesian model, a flat prior is used in the model. For SVM, the default parameter is set and radius basic function is used as kernel function. In both cases the distance function between samples is Euclidean distance function. Zero-one loss is used as the loss function for accuracy evaluation. Figure 2 shows the compared results for the first case. From Figure 2, we can see that compared with the traditional Bayesian model and SVM classifier, our SBL model has the best performance in both datasets. It leads to 83% and 80% accuracy respectively.

Note that the model sparsity can be controlled by a threshold, i.e. prune off the elements of *w* based on their absolute values. To show the relationship between the model sparsity and accuracy, we vary the threshold to obtain different sparsity levels. There are 500 basic distributions in the candidate set. Figure 3 shows the accuracy at different sparsity levels of the model. In Figure 3, SBL stands for the proposed method and all stands for the combination of all candidate distribution with equal weights. We see that a sparise combination performs a little better than the whole, which obeys the idea of selective ensemble learning. And when the sparsity level equals to 0.4, it can achieve the best accuracy [20].

Conclusion

In this article, we propose to use sparse Bayesian learning methods to analyse healthcare data. The goal is to model the underlying

principles of a given healthcare dataset. And use the learned model to classify unseen samples. The proposed method is a pure Bayesian solution. It is benefit from a Gaussian process prior and ARD prior. The optimization problem can be coverted into a standard relevant vector machine problem, which guarantees the sparsity of the target model. The proposed model can be applied in large scale healthcare data analysis tasks and real-time analysis.

Acknowledgement

publ This work is supported by the National Natural Science Foundation of China (No. 61273249, 81373883), the College Student Career and Innovation Training Plan Project of Guangdong Prov. (yj201311845015, yj201311845023 and yj201311845031).

References

1. Wang J, Hu X, Zhu D. Applications of Data Mining in the Healthcare Industry. Encyclopedia of Health-care Information Systems. 2008.
2. Srinivas BS, Govardhan A, Kumar CS. Data Mining Issues and Challenges in Healthcare Domain. In International Journal of Engineering Research and Tech-nology, Volume 3, ESRSA Publications. 2014.
3. Rogers G, Joyner E. Mining your data for health care quality improvement. In Proceedings of the Twenty-Second Annual SAS Users Group International Conference. San Diego, CA, Springer Verlag. 1997; 641-647.
4. Palaniappan S, Awang R. Intelligent heart disease prediction system using data mining techniques. In Computer Systems and Applications, 2008. AICCSA 2008. IEEE/ACS International Conference on, IEEE 2008: 108-115.
5. Desikan P, Hsu S, Srivastava J. Data mining for health care management. In 2011 SIAM International Con-ference on Data mining 2011.
6. Liu P, Lei L, Yin J, Zhang W, Naijun W, El-Darzi E. Healthcare data mining: predicting inpatient length of stay. 2006.
7. Abdi MJ, Giveki D. Automatic detection of erythemato-squamous diseases using PSO {SVM based on association rules. Engineering Applications of Artificial Intelligence. 2013; 26: 603-608.
8. Tomar D, Agarwal S. A survey on Data Mining approaches for Healthcare. International Journal of Bio-Science & Bio-Technology. 2013; 5: 241-266.
9. Canlas Jr RD. Data Mining in Healthcare: Current Applications and Issues. [MS in Information Technology thesis]. 2009.
10. Shukla D, Patel SB, Sen AK. A Literature Review in Health Informatics Using Data Mining Techniques 2014.
11. NIAK_SU O, KURASOVA O. Data Mining Applications in Healthcare: Research vs Practice.
12. Lucas P. Bayesian analysis, pattern analysis, and data mining in health care. Curr Opin Crit Care. 2004; 10: 399-403.
13. Bandyopadhyay S, Wolfson J, Vock DM, Vazquez-Benitez G, Adomavicius G, Elidrisi M, et al. Data mining for censored time-to-event data: A Bayesian network model for predicting cardiovascular risk from electronic health record data. arXiv preprint arXiv:1404.2189. 2014.
14. Liu P, El-Darzi E, Lei L, Vasilakis C, Chountas P, Huang W. An analysis of missing data treatment methods and their application to health care dataset. In Advanced Data Mining and Applications, Springer. 2005; 3584: 583-590.
15. Zhao J, Huang JX, Hu X, Kurian J, Melek W. A Bayesian-based prediction model for personalized medical health care. In Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference on, IEEE. 2012: 1-4.
16. Lappenschaar M, Hommersom A, Lucas PJ, Lagro J, Visscher S, et al. Multilevel Bayesian networks for the analysis of hierarchical health care data. Artif Intell Med. 2013; 57: 171-183.
17. Zhang W, Liu J, Niu YQ, Wang L, Hu X. A Bayesian regression approach to the prediction of MHC-II binding affinity. Comput Methods Programs Biomed. 2008; 92: 1-7.
18. Zhou Y, Kantarcioglu M, Thuraisingham B. Sparse Bayesian Adversarial Learning Using Relevance Vector Machine Ensembles. In Proceedings of the 2012 IEEE 12th International Conference on Data Mining, ICDM '12, Washington, DC, USA: IEEE Computer Society. 2012: 1206-1211.
19. Bishop CM. Pattern Recognition and Machine Learning (Information Science and Statistics). Secaucus, NJ, USA: Springer-Verlag New York, Inc. 2006.
20. Zhou ZH, Tang W, hua Zhou Z, Tang W. Selective Ensemble of Decision Trees. In Lecture Notes in Article Intelligence, Springer. 2003; 2639: 476-483.

Austin J Biomed Eng - Volume 1 Issue 3 - 2014
ISSN : 2381-9081 | www.austinpublishinggroup.com
Xie et al. © All rights are reserved

Citation: Xie X, Zhang G, Huang Y and Ou S. Bayesian Models for Healthcare Data Analysis. Austin J Biomed Eng. 2014;1(3): 1013.