Zhao et al. BMC Genomics 2013, 14(Suppl 4):S6
http://www.biomedcentral.com/1471-2164/14/S4/S6

BMC Genomics

RESEARCH Open Access

BPLT$^+$: A Bayesian-based personalized recommendation model for health care

Jiashu Zhao[1,2], Jimmy Xiangji Huang[1,3*], Xiaohua Hu[4]

From IEEE International Conference on Bioinformatics and Biomedicine 2012
Philadelphia, PA, USA. 4-7 October 2012

Abstract

In this paper, we propose an Advanced Bayesian-based Personalized Laboratory Tests recommendation (BPLT$^+$) model. Given a patient, we estimate whether a new laboratory test should belong to a "taken" or "not-taken" class. We use the bayesian method to build a weighting function for a laboratory test and the given patient. A higher weight represents that the laboratory test has a higher probability of being "taken" by the patient and lower probability of being "not-taken" by the patient. For the sake of effectiveness and robustness, we further integrate several modified smoothing techniques into the model. In order to evaluate BPLT$^+$ model objectively, we propose a framework where the data set is randomly split into a training set, a validation input set and a validation label set. A training matrix is generated from the training data set. Then instead of accessing the training data set repeatedly, we utilize this training matrix to predict the laboratory test on the validation input set. Finally, the recommended ranking list is compared with the validation label set using our proposed metric $CorrectRate_M$. We conduct experiments on real medical data, and the experimental results show the effectiveness of the proposed BPLT$^+$ model.

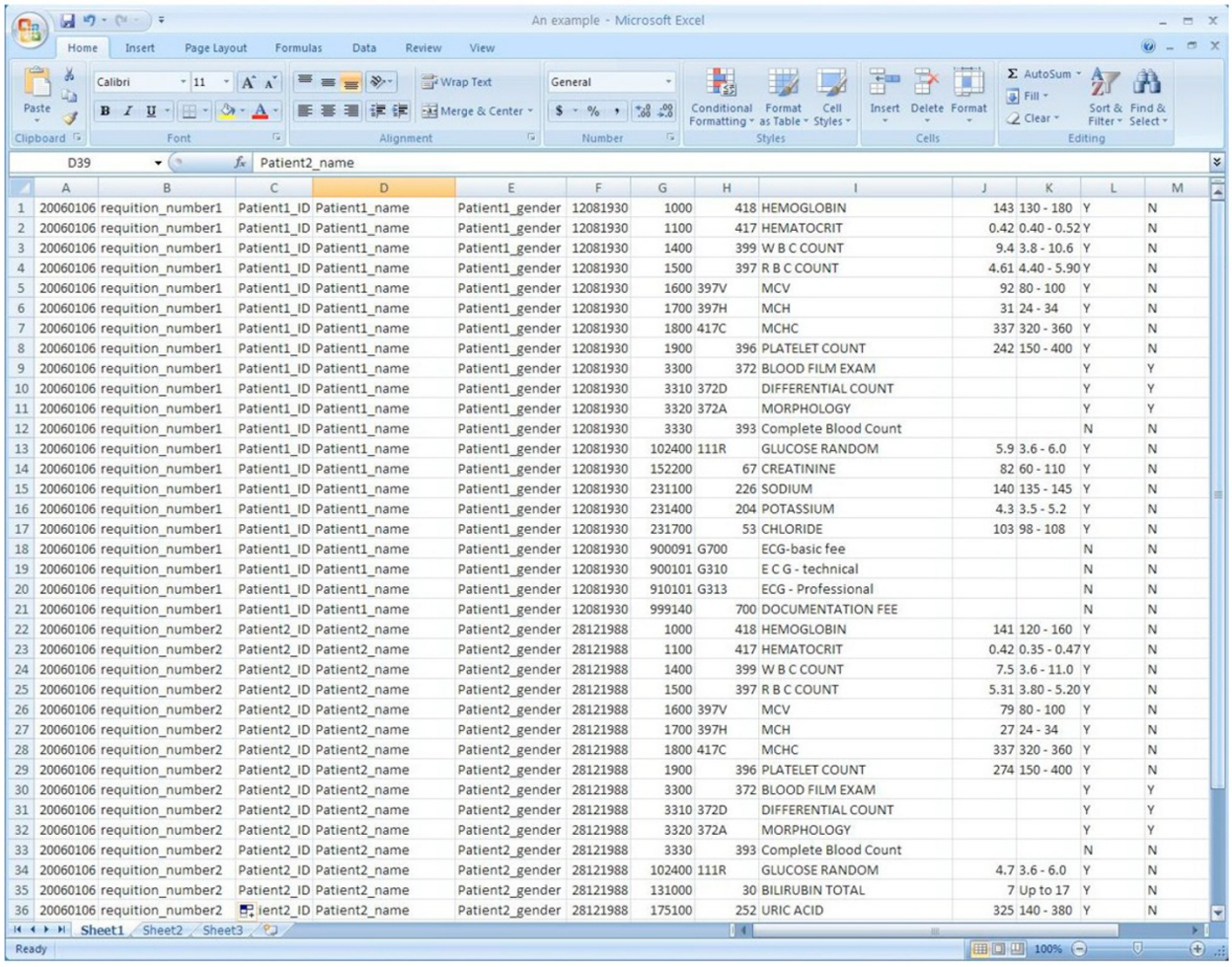
Background

Large amounts of clinic laboratory test data are collected and stored every day. Therefore, there is an increasing need for analyzing and utilizing the laboratory test data. The problem we are working on in this paper is to recommend laboratory tests for given patients. Health care recommendation problems have drawn researchers' attention for years. However, there are not a lot of studies conducted on the clinic laboratory test recommendation problem.

The medical data we are working on contains several years patients' laboratory test records. Figure 1 shows an example of the data format. Formally, the laboratory test prediction problem can be described as follows [1]: "Given a set of patients $P = \{p_1, p_2, ..., p_n\}$ and a set of laboratory tests $T = \{test_1, test_2, ... test_M\}$, each patient p_j has done tests $test_{j,1}, ..., test_{j,kj}$. If a doctor would like to assign a new test for patient p_j, which test in T should be chosen?"

The computer systems have been playing for an important role in health care for years [2-8]. Statistic algorithms [9-12] lead an important role in investigating health care data. [13,14] extracts chemical keywords from a query patent by analyzing word frequency and the word's effect over the data collection. Bayesian learning is a widely used algorithm that shows good performance [15-19]. A semantic-based association rule mining approach is proposed to model the medical query contexts in [20]. Using a novel classifier based on the Bayesian discriminant function, Raymer, M. L. [21] present a hybrid algorithm that employs feature selection and extraction to isolate salient features from large medical and other biological data sets. Martín and Pérez [22] analyze the robustness of the optimal action in a Bayesian decision making problem in the context of health care. [23,24] studies the association between two words by simulating the impact of words in documents in the context of information retrieval. A probabilistic survival model is derived from the survival analysis

* Correspondence: jhuang@yorku.ca
[1]Information Retrieval and Knowledge Management Research Lab, York University, Toronto, ON, M3J1P3, Canada
Full list of author information is available at the end of the article



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	20060106	requition_number1	Patient1_ID	Patient1_name	Patient1_gender	12081930	1000	418	HEMOGLOBIN	143	130 - 180	Y	N
2	20060106	requition_number1	Patient1_ID	Patient1_name	Patient1_gender	12081930	1100	417	HEMATOCRIT	0.42	0.40 - 0.52	Y	N
3	20060106	requition_number1	Patient1_ID	Patient1_name	Patient1_gender	12081930	1400	399	W B C COUNT	9.4	3.8 - 10.6	Y	N
4	20060106	requition_number1	Patient1_ID	Patient1_name	Patient1_gender	12081930	1500	397	R B C COUNT	4.61	4.40 - 5.90	Y	N
5	20060106	requition_number1	Patient1_ID	Patient1_name	Patient1_gender	12081930	1600	397V	MCV	92	80 - 100	Y	N
6	20060106	requition_number1	Patient1_ID	Patient1_name	Patient1_gender	12081930	1700	397H	MCH	31	24 - 34	Y	N
7	20060106	requition_number1	Patient1_ID	Patient1_name	Patient1_gender	12081930	1800	417C	MCHC	337	320 - 360	Y	N
8	20060106	requition_number1	Patient1_ID	Patient1_name	Patient1_gender	12081930	1900	396	PLATELET COUNT	242	150 - 400	Y	N
9	20060106	requition_number1	Patient1_ID	Patient1_name	Patient1_gender	12081930	3300	372	BLOOD FILM EXAM			Y	Y
10	20060106	requition_number1	Patient1_ID	Patient1_name	Patient1_gender	12081930	3310	372D	DIFFERENTIAL COUNT			Y	Y
11	20060106	requition_number1	Patient1_ID	Patient1_name	Patient1_gender	12081930	3320	372A	MORPHOLOGY			Y	Y
12	20060106	requition_number1	Patient1_ID	Patient1_name	Patient1_gender	12081930	3330	393	Complete Blood Count			N	N
13	20060106	requition_number1	Patient1_ID	Patient1_name	Patient1_gender	12081930	102400	111R	GLUCOSE RANDOM	5.9	3.6 - 6.0	Y	N
14	20060106	requition_number1	Patient1_ID	Patient1_name	Patient1_gender	12081930	152200	67	CREATININE	82	60 - 110	Y	N
15	20060106	requition_number1	Patient1_ID	Patient1_name	Patient1_gender	12081930	231100	226	SODIUM	140	135 - 145	Y	N
16	20060106	requition_number1	Patient1_ID	Patient1_name	Patient1_gender	12081930	231400	204	POTASSIUM	4.3	3.5 - 5.2	Y	N
17	20060106	requition_number1	Patient1_ID	Patient1_name	Patient1_gender	12081930	231700	53	CHLORIDE	103	98 - 108	Y	N
18	20060106	requition_number1	Patient1_ID	Patient1_name	Patient1_gender	12081930	900091	G700	ECG-basic fee			N	N
19	20060106	requition_number1	Patient1_ID	Patient1_name	Patient1_gender	12081930	900101	G310	E C G - technical			N	N
20	20060106	requition_number1	Patient1_ID	Patient1_name	Patient1_gender	12081930	910101	G313	ECG - Professional			N	N
21	20060106	requition_number1	Patient1_ID	Patient1_name	Patient1_gender	12081930	999140	700	DOCUMENTATION FEE			N	N
22	20060106	requition_number2	Patient2_ID	Patient2_name	Patient2_gender	28121988	1000	418	HEMOGLOBIN	141	120 - 160	Y	N
23	20060106	requition_number2	Patient2_ID	Patient2_name	Patient2_gender	28121988	1100	417	HEMATOCRIT	0.42	0.35 - 0.47	Y	N
24	20060106	requition_number2	Patient2_ID	Patient2_name	Patient2_gender	28121988	1400	399	W B C COUNT	7.5	3.6 - 11.0	Y	N
25	20060106	requition_number2	Patient2_ID	Patient2_name	Patient2_gender	28121988	1500	397	R B C COUNT	5.31	3.80 - 5.20	Y	N
26	20060106	requition_number2	Patient2_ID	Patient2_name	Patient2_gender	28121988	1600	397V	MCV	79	80 - 100	Y	N
27	20060106	requition_number2	Patient2_ID	Patient2_name	Patient2_gender	28121988	1700	397H	MCH	27	24 - 34	Y	N
28	20060106	requition_number2	Patient2_ID	Patient2_name	Patient2_gender	28121988	1800	417C	MCHC	337	320 - 360	Y	N
29	20060106	requition_number2	Patient2_ID	Patient2_name	Patient2_gender	28121988	1900	396	PLATELET COUNT	274	150 - 400	Y	N
30	20060106	requition_number2	Patient2_ID	Patient2_name	Patient2_gender	28121988	3300	372	BLOOD FILM EXAM			Y	Y
31	20060106	requition_number2	Patient2_ID	Patient2_name	Patient2_gender	28121988	3310	372D	DIFFERENTIAL COUNT			Y	Y
32	20060106	requition_number2	Patient2_ID	Patient2_name	Patient2_gender	28121988	3320	372A	MORPHOLOGY			Y	Y
33	20060106	requition_number2	Patient2_ID	Patient2_name	Patient2_gender	28121988	3330	393	Complete Blood Count			N	N
34	20060106	requition_number2	Patient2_ID	Patient2_name	Patient2_gender	28121988	102400	111R	GLUCOSE RANDOM	4.7	3.6 - 6.0	Y	N
35	20060106	requition_number2	Patient2_ID	Patient2_name	Patient2_gender	28121988	131000	30	BILIRUBIN TOTAL	7	Up to 17	Y	N
36	20060106	requition_number2	Patient2_ID	Patient2_name	Patient2_gender	28121988	175100	252	URIC ACID	325	140 - 380	Y	N

Figure 1 An example dataset. The format of the laboratory data sets is presented: the attributes from left to right are SDTE (SERVICE DATE), REQ# (REQUISITION NUMBER), PNUM (PATIENT HEALTH CARD#), PNAM (PATIENT NAME), PSEX (PATIENT SEX), BDTE (PATIENT DATE OF BIRTH), TSEQ (TEST SEQUENCE NUMBER), TEST (TEST CODE), DESC (TEST DESCRIPTION), RSLT (TEST RESULT), NORM (NORMAL RANGE), REXP (RESULT EXPECTED Y/N), EXRS (EXTENDED RESULT Y/N). The patient information in this table is fake due to privacy.

theory for measuring aspect novelty of genomics data [25]. A mixture markov model is proposed to investigate user navigation patterns so that a personalized recommendation system for each user can be built [26]. In our previous work [1], we propose a laboratory test prediction model, which would objectively determine whether a laboratory test is associated to a patient. This paper is a significant extension to [1].

Smoothing [27] is a technique to create an approximating function that attempts to capture important patterns in the data, while leaving out noise or other fine-scale structures/rapid phenomena. The smoothing techniques have been used in many realms to improve the accuracy [28]. Based on the basic Bayesian algorithm and smoothing techniques, we propose an Advanced Bayesian-based Personalized Laboratory Tests recommendation (BPLT$^+$) model, to investigate the correlation among laboratory tests for each patient. Evaluation is a crucial issue in the health care domain [29]. Some previous health care researchers do evaluation via patient interaction [30] or statistics [31]. We present a metric $CorrectRate_X$ by employing the idea of Mean Average Precision (MAP) [32] in Information Retrieval domain.

Four unique contributions are presented in this paper. Firstly, we learn the associations among laboratory tests and make personalized recommendations to patients without human interaction. Secondly, we integrate modified smoothing technologies to improve the personalized recommendation model and propose the BPLT+ model. Thirdly, we propose a framework to randomly generate a training data set, a validation input set and a validation label set. Fourth, we use a objective evaluation metric for personalized recommendation systems without patient interaction.

Methods

Bayesian-Based personalized laboratory tests recommendation (BPLT) model

Here we assume that the laboratory tests for a patient have associations among each other. For instance, if a patient is suspected to have diabetes, usually the doctor will assign both Hemoglobin test and Glucose Fasting test for this patient. We can see that there exists an association between Hemoglobin and Glucose Fasting with respect to some hidden information, diabetes in this case. On the other hand, if a patient is assigned Hemoglobin test, then it is very likely that this patient should also take Glucose Fasting test. In this section, we build a model for learning the associations of the laboratory tests, inferring the associations between patients and laboratory tests, and therefore recommending new laboratory tests to the patients. We regard the test recommendation problem as a special classification problem, where a test belongs to either a "taken" or "not-taken" class. We use Bayesian classifier as our basic classifier, and modify it to a personalized ranking model.

Basic concept: Bayesian classifier

A classification problem is the following [33]: given a set of training instances, each described with a set of n attributes and each belonging to exactly one of a certain number of possible classes, learn to classify new, unseen objects. In addition, each attribute has a fixed number of possible values. We use naive Bayesian classifier as our basic classifier in this paper, since it evaluates directly the probability of taking a test and the conditional probability among two tests. Moreover, naive Bayesian is easy to construct and has surprisingly good performance in classification, even though the conditional independence assumption is rarely true in real-world applications [34]. The probability model for a classifier is a conditional model

$$\Pr(C|F_1, ..., F_n) \tag{1}$$

where F_1, ..., F_n are attributes, and C is a class variable. By Bayesian criteria, it equals to

$$\frac{\Pr(C)\Pr(F_1, ..., F_n|C)}{\Pr(F_1, ..., F_n)} \tag{2}$$

The denominator is effectively constant, and the numerator is equivalent to the joint probability model

$$\begin{aligned}&\Pr(C,\ F_1, ..., F_n)\\ &= \Pr(C)\Pr(F_1|C)\Pr(F_2|C,\ F_1)\Pr(F_3|C,\ F_1,\ F_2)...\Pr(F_n|C, F_1, ..., F_{n-1})\end{aligned}$$

In naive Bayesian, it assumes the features are conditional independent

$$\Pr(F_i|C,\ F_{j,0}) = \Pr(F_i|C),\ for\ i \neq j$$

Therefore, the probability of a class C given feature F_1, ..., F_n is

$$\Pr(C|F_1, ..., F_n) = A\Pr(C)\prod_{i=1}^{n}(F_i|C) \tag{3}$$

where $A = \frac{1}{\Pr(F_1, ..., F_n)}$ is a constant.

The weighting function of BPLT model

In this Section, we describe the Bayesian-based Personalized Laboratory Tests recommendation (BPLT) model, which was proposed in our previous work [1]. More details are given in this paper. The purpose of BPLT model is to classify the laboratory tests for individual patients by their personal conditions. In the real world, it is often easier to obtain the patients' previous laboratory tests information. Therefore, the BPLT model recommends additional new laboratory tests to patients, given the previous laboratory tests that the patients have taken.

Suppose we have a set of M laboratory tests $T = \{test_1, test_2, ..., test_M\}$, and a patient p_j who has taken tests $T_j = \{test_{j,1}, ..., test_{j,kj}\}$ where $test_{j,i} \in T$ for all $1 \le i \le k_j$. We denote the events that tests in T_j are taken by p_j as $F_{j,1}$, $F_{j,2}$, ...$F_{j,M}$. For example, if we have 7 tests in T, and p_j has taken $test_3$, $test_5$ and $test_7$ could be represented as $(F_{j,1}, F_{j,2}, ..., F_{j,7}) = (0, 0, 1, 0, 1, 0, 1)$. Bayesian Classifier is employed to evaluate the association between p_j a new test $test_0$ where $test_0 \in T$ and $test_0 \notin T_j$. We use $F_{j,0}$ to represent the event of p_j should take t_0, and $F^c_{j,0}$ to represent the event of p_j should not take t_0. By Formula (3), the probability of $F_{j,0}$ given $F_{j,1}$, $F_{j,2}$, ...$F_{j,M}$ is

$$\Pr(F_{j,0}|F_{j,1}, F_{j,2}, \ldots F_{j,M}) \propto \Pr(F_{j,0})\prod_{i=1}^{M}\Pr(F_{j,i}|F_{j,0})$$

The probability of $F^c_{j,0}$ given $F_{j,1}$, $F_{j,2}$, ... $F_{j,M}$ is

$$\Pr(F^c_{j,0}|F_{j,1}, F_{j,2}, \ldots F_{j,M}) \propto \Pr(F^c_{j,0})\prod_{i=1}^{M}\Pr(F_{j,i}|F^c_{j,0})$$

In the BPLT model, we reward the tests with high probability of "taken" and low probability of "not-taken". The correlation between a new test $test_0$ and a given patient p_j is shown in Definition 1 [1].

Definition 1 *The correlation between a new test* $test_0$ *and a given patient* p_j *is defined as the log function of the probability of* p_j *should take test* $_0$ *divided by the probability of* p_j *should not take test* $_0$ *given* $F_{j,1}$, $F_{j,2}$, ... $F_{j,M}$.

$$corr(test_0, p_j) = \log\frac{\Pr(F_{j,0}|F_{j,1}, F_{j,2}, \ldots F_{j,M})}{\Pr(F^c_{j,0}|F_{j,1}, F_{j,2}, \ldots F_{j,M})} \tag{4}$$

We can see that higher value of $corr(test_0, p_j)$ indicates that $test_0$ has more association with p_j. The calculation of $corr(test_0, p_j)$ can be further simplified as follows

$$\begin{aligned} &corr(test_0, p_j) \\ &= \log \Pr(F_{j,0}|F_{j,1}, F_{j,2}, \ldots F_{j,M}) - \log \Pr(F^c_{j,0}|F_{j,1}, F_{j,2}, \ldots F_{j,M}) \\ &= \log \Pr(F_{j,0}) \prod_{i=1}^{M} \Pr(F_{j,i}|F_{j,0}) - \log \Pr(F^c_{j,0}) \prod_{i=1}^{M} \Pr(F_{j,i}|F^c_{j,0}) \qquad (5) \\ &= \log \frac{\Pr(F_{j,0})}{\Pr(F^c_{j,0})} + \sum_{i=1}^{M} \frac{\log \Pr(F_{j,i}|F_{j,0})}{\Pr(F_{j,i}|F^c_{j,0})} \end{aligned}$$

Moreover, a test either belongs to a "taken" class or a 'not taken" class. Thus, the following two formulas are held.

$$\Pr(F_{j,0}) + \Pr(F^c_{j,0}) = 1$$

$$\Pr(F_{j,i}|F_{j,0}) \Pr(F_{j,0}) + \Pr(F_{j,i}|F^c_{j,0}) \Pr(F^c_{j,0}) = \Pr(F_{j,i})$$

from which we can obtain $\Pr(F^c_{j,0})$ and $\Pr(F_{j,i}|F^c_{j,0})$

$$\Pr(F^c_{j,0}) = 1 - \Pr(F_{j,0})$$

$$\Pr(F_{j,i}|F^c_{j,0}) = \frac{\Pr(F_{j,i}) - \Pr(F_{j,i}|F_{j,0}) \Pr(F_{j,0})}{1 - \Pr(F_{j,0})}$$

Thus $\Pr(F^c_{j,0})$ and $\Pr(F_{j,i}|F^c_{j,0})$ in (5) can be eliminated in $corr\ (test_0, p_j)$, as shown below

$$\log \frac{\Pr(F_{j,0})}{1 - \Pr(F_{j,0})} + \sum_{i=1}^{M} \log \frac{\Pr(F_{j,i}|F_{j,0})(1 - \Pr(F_{j,0}))}{\Pr(F_{j,i}) - \Pr(F_{j,i}|F_{j,0}) - \Pr(F_{j,0})}$$

A joint probability for patient p_j take both of the tests $test_i$ and $test_0$ is

$$\Pr(F_{j,i}, F_{j,0}) = \Pr(F_{j,i}|F_{j,0}) \Pr(F_{j,0})$$

The definition of the correlation between $test_0$ and p_j is

$$\begin{aligned} &corr(test_0, p_j) \\ &= \log \frac{\Pr(F_{j,0})}{1 - \Pr(F_{j,0})} + \sum_{i=1}^{M} \log \frac{\Pr(F_{j,i}, F_{j,0})(1 - \Pr(F_{j,0}))}{\Pr(F_{j,0})(\Pr(F_{j,i}) - \Pr(F_{j,i}, F_{j,0}))} \\ &= (k-1) \cdot \log \frac{1 - \Pr(F_{j,0})}{\Pr(F_{j,0})} + \sum_{i=1}^{M} \log \frac{\Pr(F_{j,i}|F_{j,0})}{\Pr(F_{j,i}) - \Pr(F_{j,i}|F_{j,0})} \end{aligned}$$

which leads to the following Definition 2 [1].

Definition 2 *The weighting function for a laboratory test $test_0$ for a patient p_j is the simplified correlation between $test_0$ and p_j*

$$w(test_0, p_j) = (k-1) \cdot \log \frac{1-\alpha}{\alpha} + \sum_{i=1}^{M} \log \frac{\beta_{j,i}}{\gamma_{j,i} - \beta_{j,i}} \qquad (6)$$

where

$$\alpha = \Pr(F_{j,0}) = \frac{number\ of\ patients\ taken\ test_0}{number\ of\ patients}$$

$$\gamma_{j,i} = \Pr(F_{j,i}) = \frac{number\ of\ patients\ that\ F_{j,i}\ holds}{number\ of\ patients}$$

$$\beta_{j,i} = \Pr(F_{j,i}|F_{j,0}) = \frac{\Pr(F_{j,i}, F_{j,0})}{\Pr(F_{j,0})}$$

$$= \frac{1}{\alpha} \frac{number\ of\ patients\ that\ both\ F_{j,0}\ and\ F_{j,i}\ holds}{number\ of\ patients}$$

The new laboratory tests will be ranked in a list according to $w(test_0, p_j)$ for a given patient p_j. In the later section, we will present the evaluation environments for the laboratory test ranking list.

An advanced model: BPLT$^+$

To have a more robust and better performance model, we further propose an advanced model, BPLT$^+$, by improving the BPLT model using several smoothing techniques. There are two reasons for smoothing BPLT. One reason is that smoothing is a way to deal with noise within the data. Another reason is to avoid the mathematically meaningless. When $test^0$ laboratory test has not been observed in the previous visits, which means $\alpha = 0$, the first part of formula (6) will become an irrational number. Meanwhile, when the joint frequency of two laboratory tests is zero, which means $\beta_{j,i} = 0$, the second part of (6) will become an irrational number. Therefore, we introduce smoothing technologies to further improve BPLT model.

Smoothing techniques

In statistics, smoothing [27] is a technique to create an approximating function that attempts to capture important patterns in the data, while leaving out noise or other fine-scale structures/rapid phenomena. The main purpose of smoothing in this paper is to assign a non-zero probability to the unseen tests and improve the accuracy of test probability estimation in general.

The smoothing techniques are discussed based on the following definitions of a conditional probability [28].

$$\Pr(t|p) = \frac{c(t;p)}{\sum_{t \in T} c(t;p)} \qquad (7)$$

where c(t;p) is the count of a patient taking a test. Here are some commonly used smoothing methods. Since we have defined a ranking problem, which is similar to the problems in Information Retrieval (IR), we use some widely used smoothing methods in language model in IR. The general form of a smoothed model [35] is assumed to be the following:

$$\Pr(t|p) = \begin{cases} \Pr_t(t|p) \ \textit{if test t is observed} \\ \Pr(t|C) \ \textit{otherwise} \end{cases} \quad (8)$$

where $\Pr_t(t|p)$ is the smoothed probability of a test t given the patient with existing tests. $\Pr(t|C)$ is the probability of a test t given the whole data set.

A smoothing method may be as simple as adding an extra count to every test, which is called additive or Laplace smoothing, or more sophisticated as in Katz smoothing, where tests of different count are treated differently. Three representative methods that are popular and effective are:

- The Jelinek-Mercer method

$$\Pr_\lambda(t|p) = (1-\lambda)\Pr(t|p) + \lambda \Pr(t|C) \quad (9)$$

where λ is a balancing parameter ranges from 0 to 1.

- Bayesian Smoothing using Dirichlet Priors

$$\Pr_{\mu_0}(t|p) = \frac{c(t;p) + \mu_0 \Pr(t|C)}{\sum_{t\in T} c(t;p) + \mu_0} \quad (10)$$

where μ_0 is a balancing parameter, and $\mu_0 > 0$. The Laplace method is a special case of this technique.

- Absolute Discounting

$$\Pr_\delta(t|p) = \frac{max\,(c(t;p) - \delta, 0)}{\sum_{t\in T} c(t;p)} + \sigma p(t|C) \quad (11)$$

where $\delta \in [0, 1]$ is a discount constant and $\sigma = \delta|p|_u/|p|$, so that all probabilities sum to one. Here $|p|_u$ is the number of unique terms in document d, and $|p|$ is the total count of words in the documents.

BPLT⁺ with smoothing techniques

There are two parts in formula (6) that need smoothing. The first one is the conditional probability $\beta_{j,i} = \Pr(F_{j,i}|F_{j,0})$. Its smoothed format is as follows:

- BPLT⁺ with Jelinek-Mercer

$$\beta^{\lambda}_{j,i} = (1-\lambda)\beta_{j,i} + \lambda\gamma_{j,i} \quad (12)$$

- BPLT⁺ with dirichlet priors

$$\beta^{\mu}_{j,i} = \frac{\beta_{j,i} + \mu\gamma_{j,i}}{1+\mu} \quad (13)$$

- BPLT⁺ with absolute discounting

$$\beta^{\delta}_{j,i} = \frac{max(c(t;p) - \delta, 0)}{\sum_{t\in T} c(t;p)} + \delta\gamma_{j,i} \quad (14)$$

In Jelinek-Mercer BPLT⁺ and Absolute Discounting BPLT⁺, we use the existing smoothing method. The smoothing parameters λ, δ are within the range of [0, 1]. In Dirichlet Priors BPLT⁺, we modify the Dirichlet smoothing technique, by divide both the numerator and the denominator in (10) by $\sum_{t\in T} c(t;p)$, and normalize the parameter μ to the range of 0[1], where $\mu = \frac{\mu_0}{\sum_{t\in T} c(t;p)}$.

Another part in formula (6) needs smoothing is $\log\frac{\alpha}{1-\alpha}$, which is a simple division that could be smoothed

via Laplace smoothing as

$$\log(\frac{\alpha+\theta}{1-\alpha+\theta}) \quad (15)$$

where θ is a tuning parameter ranges from 0 to 1.

Evaluation environments

Datasets

The datasets in our experiment are obtained from Alpha Global IT [1,36]. Alpha Corporate Group provides laboratory, medical clinic, commercial electronic medical record and practice management software. The data set contains 78 monthly patient's laboratory test results. Our experiments use 6 month results, containing 1,048,575 patients' records, as a key study. Thousands of patients' records and more than 400 laboratory tests are included in our experiments. The data format is the same as the example shown in Figure 1. Our data set contains real patients' information, such as health card ID, age, gender, date of visit, laboratory test ID, laboratory test results. We only use the patient ID and laboratory ID attributes in this paper, and analyze the associations among these laboratory tests. In our future work, we will incorporate more attributes in the laboratory recommendation model.

Validation data and measure

To evaluate BPLT⁺ models objectively, we divide the data set into three components: a training set, a validation input set, and a validation label set. The data set is firstly randomly split into a training set and a validation set. In this step, we split based on the patients and do not split the records from a same patient. Then for the validation set, we randomly remove one test t^* from each patient p_j, and store the t^* in the validation label set. The ranked list returned by BPLT⁺ will be compared with t^* for each patient. To measure such comparison and finally evaluate the effectiveness of BPLT⁺, we use the following defined $CorrectRate_X$ [1]. Suppose the returned laboratory ranking list is $L = t'_{1,j}, \ldots t'_{l,p}$, $CorrectRate_X$ validates whether t^* appears in the top ranked tests. The measure is modified from Mean Average Precision (MAP) [32] evaluation metric.

Definition 3 *The $CorrectRate_X$ evaluates the accuracy of a laboratory tests prediction system. It is the number of patients with the desired (golden standard) test*

matching one of the top X tests generated by the system, divided by the total number of the patients.

$$CorrectRate_X = \frac{\sum_{j=1}^{n} TOP_{j,X}}{n} \quad (16)$$

where

$$TOP_{j,X} = \begin{cases} 1 \text{ if } t^* \text{ matches a test in } \{t'_{1,j}, \ldots t'_{X,j}\} \\ 0 \text{ otherwise} \end{cases}$$

n is the number of patients, X is a parameter indicating how many top tests are compared to the golden standard test t, which is set to be 1 or 3 in this paper.*

We present an example to show how the $CorrectRate_X$ evaluates the model in Table 1. Suppose the laboratory test sets includes 200 tests and there are 5 patients in the validation set. As we have introduced, the BPLT$^+$ model returns a ranked list for each patient. Here ">" represents that the weight of the left-side laboratory test is higher than the weight of the right-side laboratory test. In our example, 2 out of 5 patients have the desired test t^* ranked in the top 1 position of the list, then $CorrectRate_1$ equals 0.4. And 4 out of 5 patients have t^* appears within the top 3 positions of the returned ranking list, then $CorrectRate_3$ equals 0.8. We can see that the top 3 positions include the top 1 position, so the following statement is always true: $CorrectRate_1 \leq CorrectRate_3$.

BPLT$^+$ System Framework

The framework of BPLT$^+$ Model is shown in Figure 2. The data set in this framework is abstracted to contain only patient ID and laboratory test ID. The procedures in the proposed framework are described as follows.

- **Split**: First the data set is randomly split into a training set and a validation set.
- **Random Remove a test as label**: Since it is hard to objectively evaluate the performance of the BPLT$^+$ model, we further randomly remove a test for each visit of the patients from the validation set. These removed tests are regarded as labels of the validation set input. Our ultimate goal is to recommend the missing test for a patient's visit.
- **Build training matrix**: To avoid duplicate calculating the frequency of a test and the joint frequency between two tests, we build a training matrix out of the training data. This training matrix contains the frequency of co-occurrences of two laboratory tests. For example, if a patient in the training data did $test_1$ and $test_2$ together, then add 1 to F_{12} and F_{21}. We can see that the training matrix is a symmetric matrix.
- **BPLT$^+$ model**: The correlation of a given $test_0$ and a patient is calculated based on formula (6).
- **Evaluation via** $CorrectRate_X$: Finally, the evaluation criteria $CorrectRate_X$ evaluates if the model made the correct recommendations.

Results

We first show the overall performance under different training-validation proportion in Table 2[1]. We randomly take 40%, 50% and 60% of the data out of the raw data set as the training data and keep the rest as the validation data. In general, there is higher performance of BPLT$^+$ model on a larger training data set. This is because the larger training data set contains more information, and more knowledge can be learned. With the development of computer technology, larger amount of medical data will be available in practice. Therefore, we will use 60% of data as training data in the rest of this paper. As we have discussed before, $CorrectRate_3$ is always higher than $CorrectRate_1$. In general, the BPLT$^+$ model has promising performance with an accuracy of 0.7074 for $CorrectRate_1$ and an accuracy of 0.7840 for $CorrectRate_3$.

Then we investigate how the smoothing parameters affect the effectiveness in detail. We first consider smoothing $\beta_{j,i}$ only. There are three smoothing technologies utilized to smooth $\beta_{j,i}$. They are Jelinek-Mercer BPLT$^+$, Dirichlet Priors BPLT$^+$ and Absolute Discounting BPLT$^+$, with the corresponding parameters: λ, μ, $\delta \in [0, 1]$. We conduct experiments on these three methods individually. The change of $CorrectRate_1$ and $CorrectRate_3$ with respect to the parameters are shown in Figure 3, Figure 4, and Figure 5. We can see from the figures that the curve of $CorrectRate_1$ is always below the curve of $CorrectRate_3$, which is consistent as we have discussed Definition 3. With the increasing of parameters from 0.1 to 1, both $CorrectRate_1$ and $CorrectRate_3$ become higher at the beginning due to the

Table 1 An example of $CorrectRate_X$

	t^*	Recommendation list	$X = 1$	$X = 3$
p_1	$test_{104}$	$test_{104} > test_5 > test_{40} > \ldots$	$TOP_{1,1} = 1$	$TOP_{1,3} = 1$
p_2	$test_{30}$	$test_{30} > test_3 > test_{18} > \ldots$	$TOP_{2,1} = 1$	$TOP_{2,3} = 1$
p_3	$test_2$	$test_{95} > test_2 > test_{34} > \ldots$	$TOP_{3,1} = 0$	$TOP_{3,3} = 1$
p_4	$test_{95}$	$test_{78} > test_{19} > test_{58} > \ldots$	$TOP_{4,1} = 0$	$TOP_{4,3} = 0$
p_5	$test_{198}$	$test_{92} > test_{134} > test_{198} > \ldots$	$TOP_{5,1} = 0$	$TOP_{5,3} = 1$
All patients	–	–	$CorrectRate_1 = 0.4$	$CorrectRate_3 = 0.8$

This example contains a validation set of 5 patients, their desired laboratory test t^*, the recommendation list, and the corresponding evaluation results.

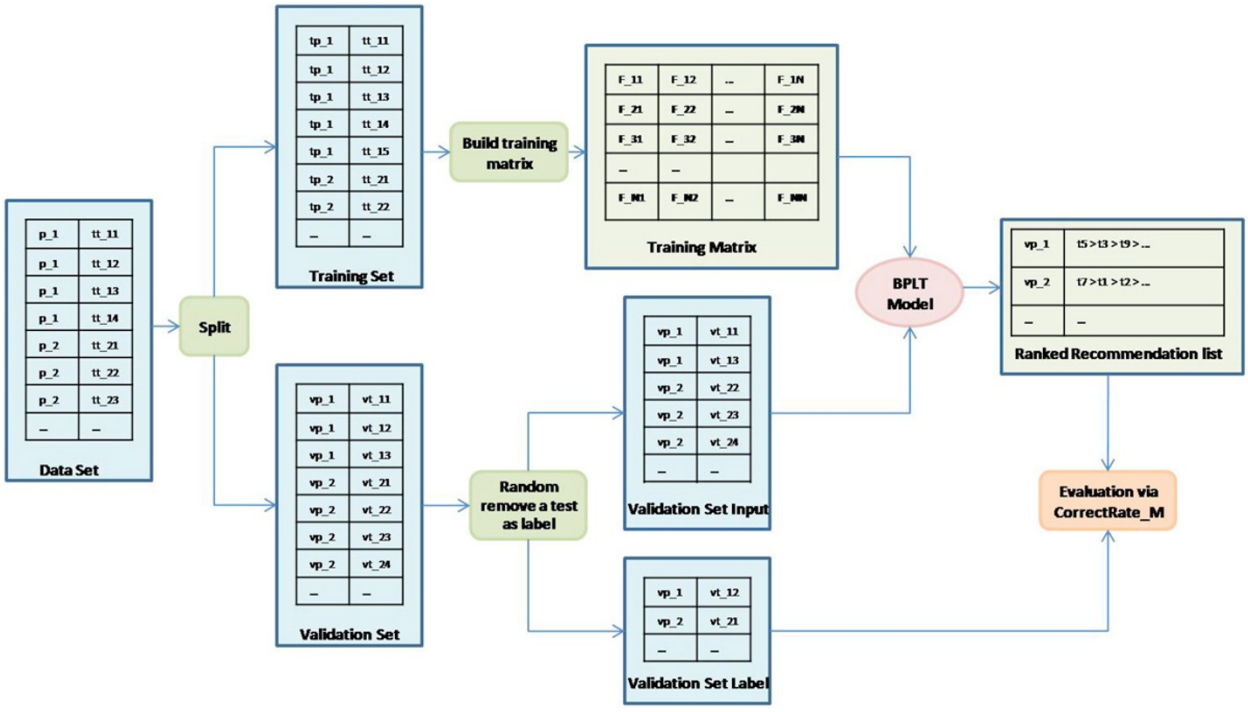


Figure 2 BPLT$^+$ System Framework. The procedures for processing the laboratory data and testing the BPLT$^+$ model are shown: (1) the rectangles represent the data sets; (2) the rounded rectangles present the implemented procedures; (3) the ovals show the personalized laboratory model; (4) the lines with arrows determine the directions through the framework.

incorporating of the smoothing portion. After reaching the maximum value, $CorrectRate_1$ and $CorrectRate_3$ become lower, since the weighing would tend to be more universal when too much smoothing is incorporated. All the smoothing parameters achieve their best performance at the value of 0.2. Comparing among these three methods, Jelinek-Mercer BPLT$^+$ obtains the best performance on both $CorrectRate_1$ and $CorrectRate_3$, which are 0.5569 and 0.6167. When it comes to the average value, Dirichlet Priors BPLT$^+$'s average performance on $CorrectRate_3$ is better than the other two, and Jelinek-Mercer BPLT$^+$'s average performance on $CorrectRate_1$ is the best.

We further discuss to smooth the second part of (6), where the Laplace smoothing parameter is θ. As we have discussed before, Jelinek-Mercer BPLT$^+$ has the best performance on both $CorrectRate_1$ and $CorrectRate_3$. We focus on investigating the sensitivity of θ by fixing

Table 2 Performance

Percentage of Training Data	$CorrectRate_1$	$CorrectRate_3$
60%	0.7074	0.7840
50%	0.6962	0.7837
40%	0.6823	0.7821

The overall performance of BPLT$^+$ with different training-validation proportions.

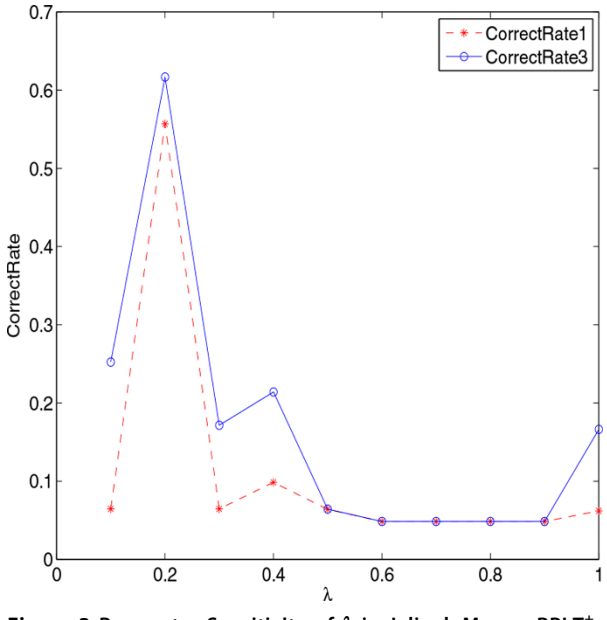


Figure 3 Parameter Sensitivity of λ in Jelinek-Mercer BPLT$^+$. The influence of parameter λ is investigated: (1) the stars represent the performance of Jelinek-Mercer BPLT$^+$ under the evaluation metric $CorrectRate_1$; (2) the circles represent the performance of Jelinek-Mercer BPLT$^+$ under the evaluation metric $CorrectRate_3$; (3) $CorrectRate_3$ is always higher than $CorrectRate_1$; (4) Jelinek-Mercer BPLT$^+$ achieves its best performance when $\lambda = 0.2$.

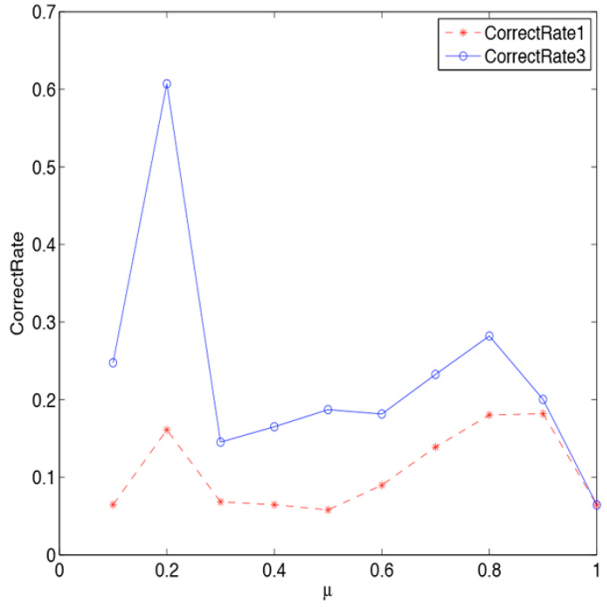


Figure 4 Parameter Sensitivity of μ in Dirichlet Priors BPLT$^+$. The influence of parameter μ is studied: (1) the stars represent the performance of Dirichlet Priors BPLT$^+$ under the evaluation metric $CorrectRate_1$; (2) the circles represent the performance of Dirichlet Priors BPLT$^+$ under the evaluation metric $CorrectRate_3$; (3) Dirichlet Priors BPLT$^+$ achieves its best performance when μ = 0.2.

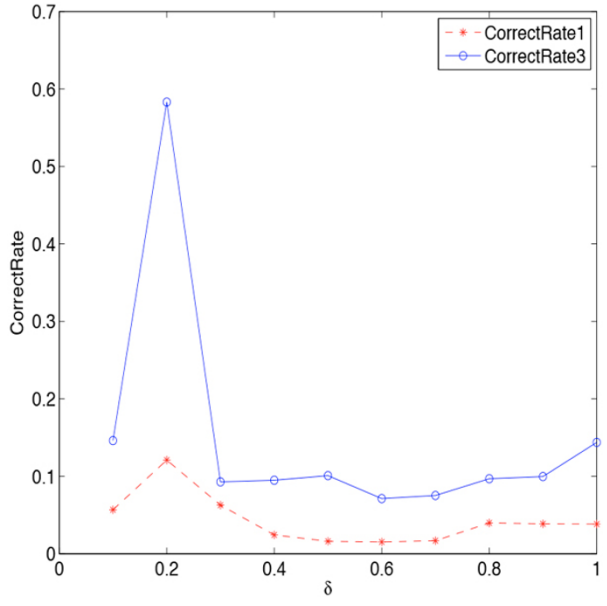


Figure 5 Parameter Sensitivity of δ in Absolute Discounting BPLT$^+$. The influence of parameter δ is investigated: (1) the stars represent the performance of Absolute Discounting BPLT$^+$ under the evaluation metric $CorrectRate_1$; (2) the circles represent the performance of Absolute Discounting BPLT$^+$ under the evaluation metric $CorrectRate_3$; (3) Absolute Discounting BPLT$^+$ achieves its best performance when δ = 0.2.

Jelinek-Mercer BPLT$^+$ with λ = 0.2. The results are shown in Figure 6. We can see that the $CorrectRate_1$ increases while θ is increasing, and the $CorrectRate_3$ decreases a little and then increases. Both of them reach the maximum and tend to be stable when θ is greater than 0.5.

Conclusions and future work

An Advanced Bayesian based Personalized Laboratory Tests recommendation (BPLT$^+$) model is proposed in this paper. Based on the assumption that hidden association could exist among laboratory tests, we employ a Bayesian approach to build a weighting function for scoring the correlation between a new laboratory test and a patient. To have a more robust and better performance model, we employ several enhanced smoothing technologies into the BPLT$^+$ model. The main purpose of smoothing in this paper is to assign a non-zero probability to the unseen laboratory tests and improve the accuracy of test probability estimation. We integrate existing smoothing techniques in the BPLT$^+$ model. In particular, we use three techniques, Jelinek-Mercer, Dirichlet Priors and Absolute Discounting approaches, to smooth the conditional probability of observing a patient taking an existing test when a new test $test_0$ is given (Formula 12-14). Also we use Laplace method to smooth the log function in the BPLT$^+$ model (Formula 15). We conducted experiments to discuss the performance of the BPLT$^+$ model and the sensitivity of smoothing parameters. We find that BPLT$^+$ is able to make accurate recommendations under proper smoothing parameters.

Further, we propose a novel framework for effectively implementing BPLT$^+$ model and objectively testing personalized recommendation systems without human interactions, shown in Figure 2. Based on the real patients' laboratory test data, we randomly generate a training data set, a validation input set and a validation label set. A training matrix containing the laboratory test statistics is calculated from the training data set and stored. For new patients (the validation input set), instead of processing the original training set, we utilize this training matrix to predict the laboratory test on the validation input set, and compare the ranking results with the validation label set.

There are a few future directions of this research work. As we can see from the data format in Figure 1, we have not make use of all the attributes. In the future, we would like to conduct a comprehensive investigation for the patients' profiles. For example, we can cluster the patients into groups and investigate the similarities of the patients in the same group. We can also analyze the associations among laboratory test results and therefore further enhance our proposed personalized

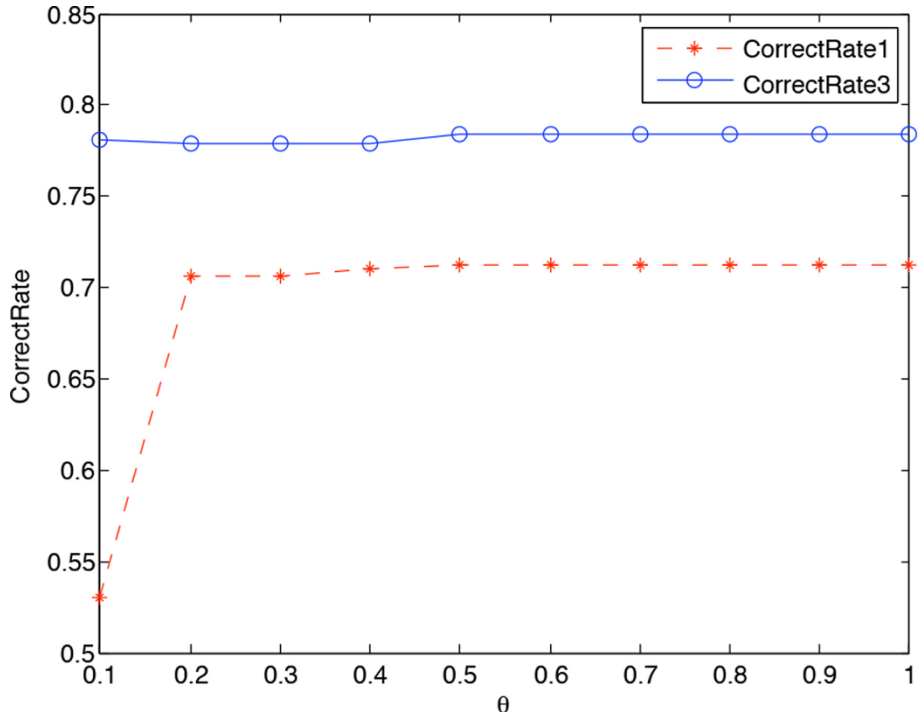


Figure 6 Parameter Sensitivity of θ. The influence of parameter θ is presented: (1) we use the best smoothing technique for the first part in Formula 6, which is Jelinek-Mercer BPLT$^+$; (2) the smoothing parameter λ is set to be optimal; (2) the stars represent the results of Jelinek-Mercer BPLT$^+$ under evaluation metric $CorrectRate_1$; (3) the circles represent the results of Jelinek-Mercer BPLT$^+$ under evaluation metric $CorrectRate_3$; (4) both metrics reach the maximum and tend to be stable when θ is greater than 0.5.

recommendation model. Moreover, we look forward to testing our proposed models in more real applications.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
JZ proposed BPLT$^+$ model, carried on the experiments and drafted the manuscript. JXH supervised the project and revised the manuscript. JXH also contributed in the study design and experiments. XH provides useful feedback. All authors read and approved the final manuscript.

Acknowledgements
This research is supported in part by the research grant from the Natural Sciences & Engineering Research Council (NSERC) of Canada and the Early Research Award/Premier's Research Excellence Award. The authors thank Dr. Joseph Kurian and Dr. William Melek from Alpha Global IT for their help and providing the data. In particular, we thank anonymous reviewers for their valuable and detailed comments on this paper.
Based on "A Bayesian-based prediction model for personalized medical health care", by Jiashu Zhao, Jimmy Xiangji Huang, Xiaohua Hu, C Joseph Kurian, and William Melek which appeared in Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference on. ©2012 IEEE, 579-582.

Declarations
The publication costs for this article were funded by the corresponding author.
This article has been published as part of *BMC Genomics* Volume 14 Supplement S4, 2013: Selected articles from the IEEE International Conference on Bioinformatics and Biomedicine 2012: Genomics. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcgenomics/supplements/14/S4.

Authors' details
[1]Information Retrieval and Knowledge Management Research Lab, York University, Toronto, ON, M3J1P3, Canada. [2]Department of Computer Science and Engineering, York University, Toronto, ON, M3J1P3, Canada. [3]School of Information Technology, York University, Toronto, ON, M3J1P3, Canada. [4]College of Information Science, Drexel University, Philadelphia, PA, 19104, USA.

Published: 1 October 2013

References
1. Zhao J, Huang JX, Hu X, Kurian J, Melek W: **A Bayesian-based prediction model for personalized medical health care.** *Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference on: 4-7 October 2012* 1-4.
2. Bates D, Cohen M, Leape L, Overhage J, Shabot M, Sheridan T: **Reducing the frequency of errors in medicine using information technology.** *Journal of the American Medical Informatics Association* 2001, **8(4)**:299-308.
3. Ogiela L, Tadeusiewicz R, Ogiela M: **Cognitive techniques in medical information systems.** *Computers in Biology and Medicine* 2008, **38(4)**:501-507.
4. Shortliffe E, Cimino J: *Biomedical informatics: computer applications in health care and biomedicine* Springer; 2006.
5. Melski J, Geer D, Bleich H: **Medical information storage and retrieval using preprocessed variables.** *Computers and Biomedical Research, An International Journal* 1978, **11(6)**:613.
6. Thoma G, Suthasinekul S, Walker F, Cookson J, Rashidian M: **A prototype system for the electronic storage and retrieval of document images.** *ACM Transactions on Information Systems* 1985, **3(3)**:279-291.
7. Frick S, Uehlinger D, Zenklusen R: **Medical futility: Predicting outcome of intensive care unit patients by nurses and doctors-A prospective comparative study*.** *Critical Care Medicine* 2003, **31(2)**:456-461.

8. Wu W, Bui A, Batalin M, Au L, Binney J, Kaiser W: **MEDIC: Medical embedded device for individualized care.** *Artificial Intelligence in Medicine* 2008, **42(2)**:137-152.
9. Kajíc V, Esmaeelpour M, Považay B, Marshall D, Rosin P, Drexler W: **Automated choroidal segmentation of 1060 nm OCT in healthy and pathologic eyes using a statistical model.** *Biomedical Optics Express* 2012, **3(1)**:86-103.
10. Kokol P, Pohorec S, Štiglic G, Podgorelec V: **Evolutionary design of decision trees for medical application.** *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2012, **2(3)**:237-254.
11. Pepe M: *The statistical evaluation of medical tests for classification and prediction* Oxford University Press, USA; 2004.
12. Rohian H, An A, Zhao J, Huang X: **Discovering temporal associations among significant changes in gene expression.** *Proceedings of IEEE International Conference on Bioinformatics and Biomedicine, IEEE 2009* 419-423.
13. Lupu M, Huang XJ, Zhu J: **TREC Chemical Information Retrieval - An Initial Evaluation Effort for Chemical IR Systems.** *World Patent Information Journal* 2011, **33(3)**:248-256.
14. Zhao J, Huang X, Ye Z, Zhu J.: **York University at TREC 2009: Chemical Track.** *Proceedings of the 18th Text REtrieval Conference* 2009.
15. Bernardo J, Smith A: **Bayesian theory.** *Measurement Science and Technology* 2001, **12**:221-222.
16. Chen J, Huang H, Tian F, Tian S: **A selective bayes classifier for classifying incomplete data based on gain ratio.** *Knowledge-Based Systems* 2008, **21(7)**:530-534.
17. Clèries R, Ribes J, Buxo M, Ameijide A, Marcos-Gragera R, Galceran J, Martínez J, Yasui Y: **Bayesian approach to predicting cancer incidence for an area without cancer registration by using cancer incidence data from nearby areas.** *Statistics in Medicine* 2012.
18. Huang X, Hu Q: **A Bayesian Learning Approach to Promoting Diversity in Ranking for Biomedical Information Retrieval.** *Proceedings of the 32nd Annual International Conference on Research and Development in Information Retrieval* 2009, 19-23.
19. Liechty J, Liechty M, Muller P: **Bayesian correlation estimation.** *Biometrika* 2004, **91**:1.
20. Babashzadeh A, Daoud M, Huang J: **Using semantic-based association rule mining for improving clinical text retrieval.** *Health Information Science* 2013, 186-197.
21. Raymer M, Doom T, Kuhn L, Punch W: **Knowledge discovery in medical and biological datasets using a hybrid bayes classifier/evolutionary algorithm.** *IEEE Transactions on Systems, Man, and Cybernetics, Part B* 2003, **33(5)**:802-813.
22. Martín J, Pérez C, Muller P: **Bayesian robustness for decision making problems: Applications in medical contexts.** *International Journal of Approximate Reasoning* 2009, **50(2)**:315-323.
23. Hu Q, Huang X: **Passage Extraction and Result Combination for Genomics Information Retrieval.** *Journal of Intelligent Information Systems* 2010, **34(3)**:249-274.
24. Zhao J, Huang JX, He B: **CRTER: using cross terms to enhance probabilistic information retrieval.** *Proceedings of the 34th international ACM SIGIR conference, ACM 2011* 155-164.
25. Yin X, Huang JX, Li Z, Zhou X: **A Survival Modeling Approach to Biomedical Search Result Diversification Using Wikipedia.** *IEEE Transactions on Knowledge and Data Engineering* 2013, **25(6)**:1201-1212.
26. Liu Y, Huang JX, An A: **Personalized recommendation with adaptive mixture of markov models.** *Journal of the American Society for Information Science and Technology* 2007, **58(12)**:1851-1870.
27. Titterington D: **Common structure of smoothing techniques in statistics.** *International Statistical Review/Revue Internationale de Statistique* 1985, 141-170.
28. Zhai C, Lafferty J: **A study of smoothing methods for language models applied to information retrieval.** *ACM Transactions on Information Systems* 2004, **22(2)**:179-214.
29. Kononenko I: **Machine learning for medical diagnosis: history, state of the art and perspective.** *Artificial Intelligence in Medicine* 2001, **23**:89-109.
30. Donabedian A: **Evaluating the quality of medical care.** *Milbank Quarterly* 2005, **83(4)**:691.
31. Cook N: **Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve.** *Clinical chemistry* 2008, **54**:17.
32. Sanderson M: **Information retrieval system evaluation: effort, sensitivity, and reliability.** *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, ACM 2005* 162-169.
33. Kononenko I: **Inductive and bayesian learning in medical diagnosis.** *Applied Artificial Intelligence* 1993, **7(4)**:317-337.
34. Zhang H, Su J: **Naive bayesian classifiers for ranking.** *Machine Learning: ECML* 2004, 501-512.
35. Chen S, Goodman J: **An empirical study of smoothing techniques for language modeling.** *Computer Speech and Language* 1999, **13(4)**:359-394.
36. Alpha Global IT:[http://www.alpha-it.com/].

doi:10.1186/1471-2164-14-S4-S6
Cite this article as: Zhao *et al.*: BPLT^{+}: A Bayesian-based personalized recommendation model for health care. *BMC Genomics* 2013 **14**(Suppl 4): S6.