Research Journal of Applied Sciences, Engineering and Technology 6(13): 2470-2473, 2013
DOI:10.19026/rjaset.6.3724
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: December 23, 2012 Accepted: January 25, 2013 Published: August 05, 2013

Research Article

Hidden Markvo Models Based Research on Lung Cancer Progress Modeling

Hui-Min Li, Li-Ying Fang, Pu Wang and Jian-Zhuo Yan
College of Electronic Information and Control Engineering, Beijing University of Technology, Beijing

Abstract: Considering of the requirements of medical clinical longitudinal data modeling, research is conducted on lung cancer post-surgical operation progress based on Hidden Markvo Models (HMM). This algorithm can do better analysis both in quality and quantity and experiments based on lung cancer followed up longitudinal data were performed, results demonstrate that it is an effective integrated analysis methods and is suitable for longitudinal data modeling and prognosis.

Keywords: Followed up materials, hidden markvo models, longitudinal data, lung cancer progress, mathematical modeling

INTRODUCTION

General clinical behavior always includes several observation based on time-signal, such as followed up data. Longitudinal data analysis based on observation is the key point of medical analysis (Liu and Meng, 2003). Nowadays, lung cancer is the most serious malignant tumor in the world; it is paid more and more close attention to quantity of patients by medical experts. Lung cancer has long way of cure and inflammation appears repeatedly, the feature is sophisticated during the process of lung cancer, some factors of disease always change with time elapse, so classical statistical methods are insufficient (Geert and Eert, 2006) on medical longitudinal data analysising.

Classical longitudinal analysis prefers describing total trend to individual average trend of cases; it can't analysis's and gives a suitable explanation to diversities of individuals; especially for missing data and observation with different interval (Yang *et al.*, 2011a, b). Considering of above factors, a new lung cancer progress modeling method based on Hidden Markvo Models is proposed.

LITERATURE REVIEW

HMM is a stochastic process, having firm foundation of statistics and widely applied to time series data processing. HMM is proposed by Leonard Baum in 1960s' and is used in speech recognition for the first time (Rabiner, 1986, 1989), the application domain of HMM also includes DNA analysis, machine translation, textual extraction and user interest drifting, etc. HMM is made up of five parts and which can be expressed as $\lambda = (S, V, A, B, \pi)$, S is a set of status and $S = \{l, 2,3, ..., N\}$; V is a set of observation and $V = \{v_l, v_2,v_3,...,v_m\}$; A is the matrix of status changing and A can be expressed as $A = [a_{ij}]$, where $a_{ij} = p(q_{t+1} = j|q_t = i)$, s.t. $(1< = i, j< = N)$; B is observation probability distribution and $B = \{b_j\ (k)\}$, where $b_j\ (k)$ indicates the observation probability v_k during status j, $b_j(k) = p(v_k|j)$, s.t. $(1< = k< = M,1< = j< = N)$; π is initial status probability distribution and $\pi = \{\pi_i\}$, which denotes the selecting probability of some status in time 1, $\pi = p(q_1 = i)$. We can see, observation series O can be generated by HMM and $o = (o_1, o_2, ..., o_T)$, where o_i denotes observation value in time i. Therefore, the HMM can be defined as follows:

Step 1: Initial status is selected via initial status probability distribution, $q_1 = i$
Step 2: $t = 1$
Step 3: Observation value o_t is selected according to observation probability distribution $b_i(k)$ in status i
Step 4: The followed status q_{t+1} is selected via status changing probability a_{ij} and $q_{t+1}= j$
Step 5: If $t < T$, $t = t + 1$ are given and jump to step 3, otherwise algorithm will finish.

HMM include 3 algorithms, the first is forward-backward algorithm, which is used for calculating observation series; Viterbi, the second algorithm is used to generate hidden status series, on the condition that HMM and observation series is existence; the last algorithm of HMM is Baum Welch, which applies to generate the optimum model from observation series (Steven, 2005).

Three parameters are necessary for HMM, they are status changing probability matrix A, status outputting

Corresponding Author: Hui-Min Li, College of Electronic Information and Control Engineering, Beijing University of Technology, Beijing

probability B and initial status probability distribution π, i.e., the problem is how to select $\lambda = (A, B, \pi)$ and makes the observation series probability $P(O|\lambda)$ maximum, so observation series is used as training samples and A, B, π is used as unknown parameters, that is to say, parameters evaluation is the key for HMM modeling. Baum Welch algorithm is a kind of repeat processing of estimation, which can generates the optimum parameters of model through iteration. Evaluation of 3 parameters is as follows:

$$\bar{\pi} = \gamma_1(i) \tag{1}$$

It denotes the expecting time of i under the condition of t=1;

$$\bar{a}_{ij} = [\sum_{t=1}^{T-1} \xi_t(i,j)] / \sum_{t=1}^{T-1} \gamma_t(i) \tag{2}$$

The formula denotes the ratio of expecting changing time from status i to j and all the changing time from status i:

$$\bar{b}_j(k) = [\sum_{t=1}^{T} \gamma_t(j) \times \delta(o_t, v_k)] / \sum_{t=1}^{T} \gamma_t(j) \tag{3}$$

$\Delta(o_t, vk) = 1$, when $o_t = vk$ and $\delta(o_t, v_k) = 0$, when $o_t \neq v_k$, numerator denotes expecting time from status j to v_k and denominator denotes expecting time in status j. the formulas subject to:

$$\sum_{t=1}^{N} \pi_i = 1 \tag{4}$$

$$\sum_{j=1}^{N} a_{ij} = 1 \text{, s.t. } 1 \le i \le N \tag{5}$$

$$\sum_{k=1}^{M} b_j(k) = 1 \text{, s.t. } 1 \le j \le N \tag{6}$$

Considering of formula (1)(2)(3), a new model parameters can be generated, then, these parameters would be looked as the new start for next iteration for evaluation, this process would not stop until the parameters convergence to some stable value.

THE PROPOSED ALGORITHM

The data source comes from 508 lung cancer patients' followed up data by Beijing Hospital of Traditional Chinese Medicine from 2010 to 2012, the series include 4~18 time investigation via 2 month interval, the leading index are astriction, fervescence, diarrhea, cough, blood in coughing, panting, expectorate, hypodynamia, anhelation, Hoarse, anorexia, Chest pain, the preprocessing for 12 index include missing data filling, data integration, data transformation and each index are expressed by 11,22,33,44, which denote four status like none, light, middle and serious.

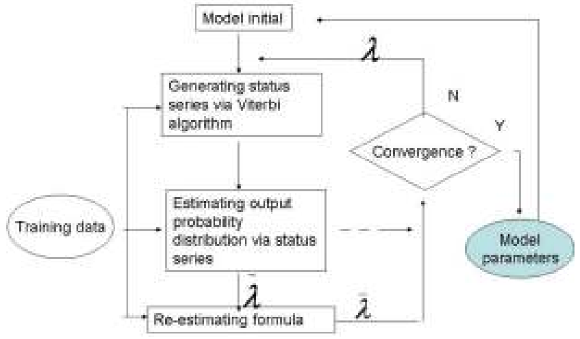

Fig. 1: HMM parameters estimation method

General speaking, the base value of parameters π and A is not important and can not impact the result badly, so they can be selected in random, s.t. $0 \le a_{ij} \le 1$, $\sum_{j=1}^{N} a_{ij} = 1$, $0 \le \pi_i \le 1$, $\sum_i \pi_i = 1$, but for parameter B, the initial value of B can influence the model and need to select carefully. Figure 1 is the estimation of parameters in HMM.

The parameters estimation formula is only suit to one series, considering several training series and data overflow problem, evaluation formula should be revised as followed:

$$\bar{\pi}_i = \sum_{l=1}^{L} \alpha_1^l(i) \beta_1^l(i) \text{, s.t. } 1 \le i \le N \tag{7}$$

$$\bar{a}_{ij} = \frac{[\sum_{l=1}^{L} \sum_{t=1}^{T_l-1} \alpha_t^l(i) a_{ij} b_j(o_{t+1}^l) \beta_{t+1}^l(j) / \phi_{t+1}]}{[\sum_{l=1}^{L} \sum_{t=1}^{T_l-1} \alpha_t^l(i) \beta_t^l(i)]} \text{, s.t. } 1 \le t, j \le 1 \tag{8}$$

$$\bar{b}_{jk} = \frac{[\sum_{l=1}^{L} \sum_{t=1, o_t=v_k}^{T_l} \alpha_t^l(j) \beta_t^l(j)]}{\sum_{l=1}^{L} \sum_{t=1}^{T_l} \alpha_t^l() \beta_t^l(j)} \text{, s.t. } 1 \le j \le N, 1 \le k \le M \tag{9}$$

where, $\bar{a}_{ij}$ and $\bar{b}_{jk}$ denotes forward variable and backward variable, respectively.

Via aboving analysis, the HMM modeling can be describe as follows, to begin with, a novel segmental K means cluster is used to divide observation vectors, the initial model parameters is confirmed by learning, then Baum Welch algorithm is used repeatedly for HMM parameters optimization and make the distance between two time is minimum, until the parameters weren't changing anymore.

SIMULATION RESULTS

Based on theoretical analysis, a number of experiments were performed to verify the lung cancer

Table 1: Lung cancer progress model based on HMM

State 0-light	π	0.30010351	A_{ij}	0.043 0.957 0	Multi-variate Gaussian distribution-Mean	[12.845 14.506 10.121 14.309 17.286 16.653 14.223 10.762 13.471 11.918 12.644 11.109]
State l-middle	π	0.29999894	A_{ij}	0.051 0.211 0.738	Multi-variate Gaussian distribution-Mean	:[14.215 14.542 11.663 13.201 22.754 11.271 24.681 16.464 18.339 12.847 27.654 13.2261]
State 2-serious:	π	0.39979512	A_{ij}	0.915 0 0.085	Multi-variate Gaussian distribution-Mean	[14.684 16.107 13.154 17.596 32.237 25.706 21.902 23.574 26.631 13.857 20.527 17.289]
The HMM model for Chinese medicine aid in post-surgical group: HMM with 3 state (s) in MATLAB 7.1.						
State 0-light	π	0.19995999	A_{ij}	0.063 0.848 0.089	Multi-variate Gaussian distribution-Mean	[13.425 18.702 16.071 15.631 14.236 17.124 23.327 13.466 15.l76 l8.017 14.913 l7.923]
State 1-middle	π	0.20164235	A_{ij}	0.815 0.049 0.136	Multi-variate Gaussian distribution-Mean	[16.823 18.276 14.626 15.098 31.382 29.384 27.603 17.839 26.156 18.709 14.028 18.119]
State 2-serious	π	0.59735746	A_{ij}	0.084 0.194 0.723	Multi-variate Gaussian distribution-Mean	[14.922 21.304 16.25 13.535 26.616 22.282 36.827 27.713 31.017 15.337 24.702 15.831]

progress model by HMM, 12 vectors are selected in experiments, patients are divided into two group, Western Medicine in Post-surgical (WMP) and Chinese Medicine Assisted in Post-surgical (CMAP) and the output status include 0-light, 1-middle, 2-serious for disease (Fieuws and Verbeke, 2004). Programming is conducted in MATLAB, the results is as follows:

The HMM model for western medicine in post-surgical (WMP) group:

HMM with 3 state (s) in MATLAB 7.1

From Table 1, we can see, π is initial status probability distribution; A_{ij} denotes status transformation probability distribution; Multi-variate Gaussian distribution-Mean denotes observation probability distribution function. The transformation probability matrix of two groups is list as follows.

WMP group:

$$\begin{bmatrix} 0.043 & 0.957 & 0 \\ 0.051 & 0.211 & 0.738 \\ 0.915 & 0 & 0.085 \end{bmatrix}$$

CMAP group:

$$\begin{bmatrix} 0.063 & 0.848 & 0.089 \\ 0.815 & 0.049 & 0.136 \\ 0.084 & 0.194 & 0.723 \end{bmatrix}$$

Fig. 2: The matrix of status transformation probability

From the status transformation matrix of WMP goup, down-triangle show the better-trend transformation probabilities, Fig. 2 shows better-trend 0.051 from middle to light, while 0.815 is list in CMAP group corresponding, the explanation is that it has better-trend for lung cancer patients after surgical operation, which can enhance physical fitness under the aid of Chinese medicine. In the same way, up-triangle denotes the worsen-trend transformation probabilities, Fig. 2 shows the worsen-trend is 0.738 in WMP group, while only 0.136 is shown in CMAP group corresponding, the explanation is that it has fewer worsen-trend probability, when Chinese medicine aid is conducted after surgical operation, which can prolong the existing time of patients and release ache feeling (Ming-Rui *et al*., 2011; Agrawal *et al*., 2011).

In addition, using HMM model can predict the disease progress in the future, as discussing above, a observation series can be forecasted via probability distribution function, after model parameters is calculated by iteration learning, so the most likely status series can be calculated, the same to the probability of observation series.

CONCLUSION

Hidden Markvo Model can be generated by training time-dependent data set and the transformation probability matrix is computed at the following time, better-trend and worsen-trend form HMM are used to forecast disease situation in future, the optimum and difference is discussed for two curing methods via analyzing behavior of patients. Observation probability distribution function is used to show the difference of two cure methods and model can forecast the status in the future. Experiments show lung cancer progress modeling based on HMM is effective and efficient and is a novel methods for medicine longitudinal data analysis.

HMM model for longitudinal data is suitable on the condition of some hypothesis, such as each status only impact by previous status. In trials, only 12 vectors are used for modeling, so using part symptoms to analysis and modeling is not sufficient in some way, however it is also impossible to select all the vectors in modeling, so how to improve the quality of selected vectors is the key problem in the future work.

ACKNOWLEDGMENT

This study is supported by National Science Fund of China (No.61174109); the New Century National Hundred, Thousand and Ten Thousand Talent Project and got the cooperation with Beijing Hospital of Traditional Chinese Medicine affiliated to CPUMS. Special thanks have been given there.

REFERENCES

Agrawal, A., S. Misra, R. Narayanan, L. Polepeddi and A. Choudhary, 2011. A lung cancer mortality risk calculator based on SEER data. Proceeding of IEEE 1st International Conference on Computational Advances in Bio and Medical Sciences (ICCABS), pp: 233-237.

Fieuws, S. and G. Verbeke, 2004. Joint modeling of multivariate longitudinal profiles: Pitfalls of the random-effects approach. Statist. Med., 23: 3093-3104.

Geert, M. and V. Eert, 2006. Models for discrete longitudinal data. J. Am. Stat. Assoc., 101: 1307-1317.

Liu, H. and Q. Meng, 2003. Longitudinal analysis. Psychol. Sci. Prog., 11(5): 586-592.

Ming-Rui, Z., L. Ying-Xu, J. Yi-Chen, Z. Sun and P. Yang, 2011. Model based user interface design for predicting lung cancer treatment outcomes. Proceeding of 33rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS), Computer Science Department, Winona State University, Winona, MN 55987, USA, pp: 75-78.

Rabiner, J., 1986. An introduction to hidden markov models. IEEE ASSP Mag., 4-16.

Rabiner, L.R., 1989. A tutorial on hidden markov models and selected applications in speech recognition. Proc. IEEE, 77: 257-286.

Steven, G.C., 2005. Hidden markov models for longitudinal comparisons. J. Am. Statist. Assoc., 100: 359-369.

Yang, Y., L. Xue and X. Wang, 2011a. Variable selection in high-dimensional partly linear models. J. Beijing Univ. Technol., 37(2): 291-295.

Yang, Y., L. Xue, X. Wang, X. Luo, Y. Li, J. Ma, *et al*., 2011b. Semi-parameter partly linear regression models on missing data. Chinese J. Math. Phys., 30(1): 71-85.