Journal of the Korean Data & Information Science Society
2021, **32**(5), 1135–1141

http://dx.doi.org/10.7465/jkdi.2021.32.5.1135
한국데이터정보과학회지

Classification analysis using hidden Markov model[†]

Yeji Cheon[1] · Hayoung Choi[2] · Yongku Kim[3]

[1]Korea Real Estate Board
[23]Department of Statistics, Kyungpook National University
Received 30 July 2021, revised 15 August 2021, accepted 17 August 2021

Abstract

A hidden Markov model (HMM) provides useful representations of dependent heterogeneous phenomena. So it becomes a popular method for modelling stochastic processes and time-dependent sequences, and is primarily applied in many different fields such as language, handwriting recognition, and molecular biology. Especially, in the sequence classification case, classification among known hidden Markov models is known to be accomplished with a classifier that minimizes the probability of error. In this paper, we first generate variables for the hidden state using the hidden markov model and then analyze the state using various classification methods. It differs from the existing analysis method by using the state variable and the mixture distribution based on the state rather than using the observed value directly in the analysis. In addition, it can be used to identify the relevance in the underlying process. As as illustration, we used the annual production of Matsutake mushroom data observed in five regions from 1997 to 2016.

Keywords: Classification analysis, hidden Markov moddel, Matsutake mushroom, time-dependent process.

1. Introduction

Hidden Markov models (HMMs) were first introduced by Baum and his co-workers in a series of papers (Baum and Petrie, 1966). A hidden Markov model provides useful representations of dependent heterogeneous phenomena. So, a hidden markov model becomes the method of choice for modeling stochastic processes and time-dependent sequences, and is primarily applied in many different fields such as language, handwriting recognition, and molecular biology. In many applications, the marginal distribution of observations is clearly a multi-model, indicating that observations have occurred in a mixture of different distributions related to different regimes (Kim and Oh, 2019; Kang, 2019). Exactly such behavior

[†] This research is based on the part of Yeji Cheon's Master thesis and was supported by the Research Grants of Korea Forest Service (Korea Forestry Promotion Institute) project (No.2019149A00-2123-0301).
[1] Assistant manager, Korea Real Estate Board, Daegu 41068, Korea.
[2] Graduate student, Department of Statistics, Kyungpook National University, Daegu 41566, Korea
[3] Corresponding author: Associate professor, Department of Statistics, Kyungpook National University, Daegu 41566, Korea. E-mail: kim.1252@knu.ac.kr

is a key property of HMM. So, we apply HMM to detect the regimes with different distributions and find a hidden state. Then, various classification analysis is performed using the found status, not the observation. Also, use the found status as an independent variable to explain the volatility of the dependent variable (Keroglou and Hadjicostis, 2014; Bicego *et al.*, 2004). Any inference in the HMM relies heavily on maximum likelihood or Bayesian approaches, but the dependence structure of the HMMs causes more difficulty of computation compared to the regular mixtures (Titterington *et al.*, 1985; Biebolt and Robert, 1994). Robert *et al.* (1993) provided an efficient Bayesian estimation of the HMMs through Gibbs sampling. Chib (1996) especially introduces a widely used state simulation procedure.

When analyzing the data, there are cases where it is difficult to find factors that affect real life. This is because factors that act in combination or influence factors are not directly observable. Analysis of the hidden states found by HMMs without using observed values as a dependent variable can be effective in finding significant variables if no significant variables can be found directly. It differs from the existing analysis method by using the state variable and the mixture distribution according to the state without directly using observations in the analysis. In addition, this method is meaningful in that it can be used to identify the relationship even when it is difficult to show the relevance in the existing but underlying model. Therefore, in this paper, we propose to find the factors influencing the hidden state indirectly using the HMM when it is difficult to find the influence factor.

In Korea, Matsutake mushroom has not been developed aquaculture technique until now, So they mainly depend on the extraction of natural acids that grow in Pinus densiflora forest. The impact of climate factors on Matsutake mushroom yield has not been studied in detail. Although much efforts have been made to interpret the occurrence of clusters and weather relations in Korea, it has not achieved great results in that it can not find the influencing meteorological factors. Therefore, it is applied to the production of Matsutake mushroom to find the factors indirectly affecting it.

2. Hidden Markov models

A hidden Markov Model $\{\mathbf{Y}_t : t \in \mathbb{N}\}$ is a particular kind of dependent mixture. In its most general form, the hidden Markov model is defined by the following assumptions:

$$\Pr\left(C_t \mid \mathbf{C}^{(t-1)}\right) = Pr\left(C_t \mid C_{t-1}\right), \quad t = 2, 3, ..., \mathrm{T} \tag{2.1}$$

$$\Pr\left(Y_t \mid \mathbf{Y}^{(t-1)}, \mathbf{C}^{(t)}\right) = Pr\left(Y_t \mid C_t\right), \quad t = 1, 2, ..., \mathrm{T}, \tag{2.2}$$

where Y_t is a observation at the time t and C_t is an unobserved state at the time t. $\mathbf{C}^{(t)}$ and $\mathbf{Y}^{(t)}$ represent the unobserved state from time 1 to time t and the observation from time 1 to time t, respectively. T is the number of total observation. Note that hidden Markov dependence for C_t can induce "regime-like" behavior in (2.1) even though Y_t's are conditionally independent.

Unlike typical mixture models, the HMM assumes that observed data are generated through a finite valued unobserved process. This unobserved process is assumed to be in one of a finite number of discrete states at each discrete point in time and to transition stochastically in a Markov fashion given the previous state or states. Observed data at each

point in time depend only on the value of the corresponding hidden state and are independent of others. The heterogeneity of data is represented by hidden Markov states. That is, a pair (c_t, y_t) with the state $c_t \in \{1, ..., k\}$ and its random value $y_t \mid c_t \sim f_{c_t}(y_t)$. Here y_t is assumed to be conditionally independent given c_t. The HMM derives its name from the following two defining properties. First, c_t is distributed as a (finite-state) Markov chain. secondly, c_t is not actually observed.

For a classification analysis, we need to generate (hidden) state variables which is longitudinal binary outcomes. Now we introduce algorithm to find the single best state sequence $\mathbf{C} = \{C_1, C_2, \ldots, C_{\mathrm{T}}\}$ for given observation sequence $\mathbf{Y} = \{Y_1, Y_2, \ldots, Y_{\mathrm{T}}\}$. Generally, global decoding is mainly used.

- Global decoding: to find the state sequence $\mathbf{C} = \{C_1, C_2, \ldots, C_{\mathrm{T}}\}$ that maximizes the joint probability $\Pr\left(C_1, C_2, \ldots, C_{\mathrm{T}} | \mathbf{Y}^{(\mathrm{T})}\right)$.
- Local decoding: to find the state sequence $\mathbf{C} = \{C_1, C_2, \ldots, C_{\mathrm{T}}\}$ that maximizes the marginal probability $\Pr\left(C_i | \mathbf{Y}^{(\mathrm{T})}\right)$ for each $1 \leq i \leq \mathrm{T}$.

Note that hidden state classification is similar whether using global or local decoding in general (Jackson *et al.*, 2015).

3. Statistical analysis

Matsutake mushroom is a product that contributes to the increase of rural income as forest products as they can raise net income by 100 percent, mainly because it extracts wild natural which is naturally grown in Pinus densiflora forest. In Korea, however, Matsutake mushroom is a natural food that has not been raised because of the lack of farming technology. The impact of climate factors on Matsutake mushroom yield has not been studied in detail. Although much efforts have been made to interpret the occurrence of clusters and weather relations in Korea, it has not achieved great results in that it can not find the influencing meteorological factors. In this chapter, we find hidden states using HMMs for specific regions of Gyeongsangbuk - do, which account for about 65% of the national Matsutake mushroom yield, and use them to find the meteorological factors indirectly affecting them.

3.1. Data description

The data used in this study are the annual production of Matsutake mushroom (kg) data observed in five regions of Bonghwa-gun, Uljin-gun, Yeongju-si, Andong-si and Yeongyang-gun from 1997 to 2016. Matsutake mushroom is collected from late August to late October in Korea, but the most active period is about 10 days from late September to early October. In addition, the production of Matsutake mushroom greatly increases or decreases depending on the weather. So also consider the meteorological factors data on May, June, July and August from 1997 to 2016. Meteorological factors were combined and analyzed in terms of time and space, such as the production of Matsutake mushroom. Detailed variables are summarized in Table 3.1.

Table 3.1 Variables and their descriptions

variables	Description
m_temp[5/6/7/8]	mean temperature during [May/June/July/August](°C)
h_temp[5/6/7/8]	maximum temperature during [May/June/July/August](°C)
l_temp[5/6/7/8]	minimum temperature during [May/June/July/August](°C)
mdew_temp[5/6/7/8]	mean dew point temperature during [May/June/July/August](°C)
ms_pre[5/6/7/8]	mean spot atmospheric pressure during [May/June/July/August](hPa)
msea_pre[5/6/7/8]	mean sea-level pressure during [May/June/July/August](hPa)
hsea_pre[5/6/7/8]	maximum sea-level pressure during [May/June/July/August](hPa)
lsea_pre[5/6/7/8]	minimum sea-level pressure during [May/June/July/August](hPa)
mwat_pre[5/6/7/8]	mean water vapor pressure during [May/June/July/August](hPa)
hwat_pre[5/6/7/8]	maximum water vapor pressure during [May/June/July/August](hPa)
lwat_pre[5/6/7/8]	minimum water vapor pressure during [May/June/July/August](hPa)
sun[5/6/7/8]	amount of sunlight during [May/June/July/August] (hour)
daylight[5/6/7/8]	daylight ratio on [May/June/July/August](%)
m_humid[5/6/7/8]	mean relative humidity on [May/June/July/August](%)
l_humid[5/6/7/8]	minimum relative humidity on [May/June/July/August](%)
precip[5/6/7/8]	total precipitation during [May/June/July/August] (mm)
day_precip[5/6/7/8]	the largest daily precipitation on [May/June/July/August] (mm)
hhour_precip[5/6/7/8]	the largest amount of precipitation in an hour on [May/June/July/August](mm)
hmin_precip[5/6/7/8]	the largest amount of precipitation in ten minutes on [May/June/July/August](mm)
m_wind[5/6/7/8]	mean wind speed during [May/June/July/August](m/s)
h_wind[5/6/7/8]	maximum wind speed during [May/June/July/August] (m/s)
h_instwind[5/6/7/8]	the maximum instantaneous wind speed during [May/June/July/August](m/s)
lgra_temp[5/6/7/8]	minimum grass temperature during [May/June/July/August](°C)
mground_temp[5/6/7/8]	mean ground temperature during [May/June/July/August](°C)

3.2. Statistical model

Hidden markov model assumes that the production of Matsutake mushroom is in one of two states such as a lean year and a bumper year. Each hidden state has an independent distribution of total Matsutake mushroom production, assumed to be Gaussian distributions. The persistence in each year state varies because it is governed by the state transition probabilities. Let C_t denote the (unobserved) annual year state at time t (i.e., C_t=1 for a bumper year, C_t=0 for a lean year). Let Y_t be a (observed) total Matsutake mushroom production amount at time t for $1 \leq t \leq$ T. The total Matsutake mushroom production amount Y_t is conditionally independent given the current year state C_t.

The first assumption (2.1) states that given the history of the annual year state up to $t-1$, the year state at time t depends on the previous year state. This is simply first-order Markov assumption applied to the year state process, which is homogeneous in time. The second one (2.2) states that The total Matsutake mushroom production amount Y_t is conditionally independent given the current year state C_t. In other words, all the temporal persistence in total productions is captured by the persistence in the year state.

3.3. Data analysis

We estimated the state sequence corresponding to the Matsutake mushroom production using global decoding, which is commonly used for each of the five regions (see Figure 3.1). The gray solid line shows the annual Matsutake mushroom production amount per year for each region, and the solid red line shows the year state (hidden) corresponding to the Matsutake mushroom production for each year. When it is located below, it means a lean year, and when it is located above it means a bumper year.

We obtain the hidden state transition probability matrix which shows the probability of

moving from each state to each state, and estimate the mean and standard deviation of the normal distribution over year state (see Table 3.2). It is possible to know the distribution of Matsutake mushroom production related to different year state.

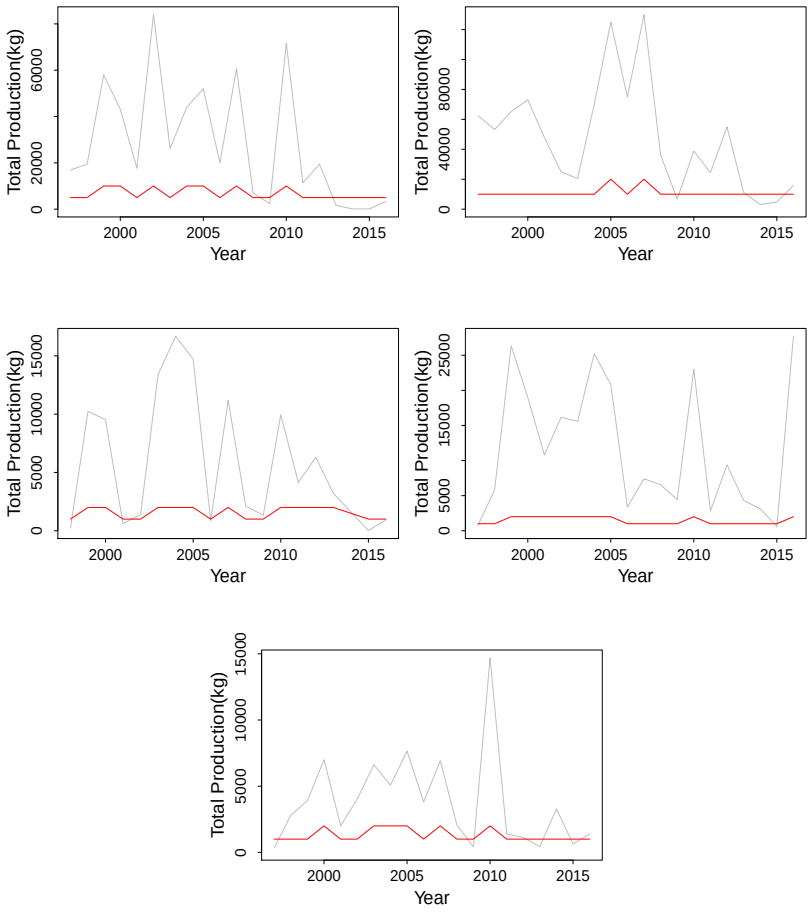

Figure 3.1 Optimal global decoding index (red lines) for total Matsutake mushroom production for five study regions

In order to identify the relevance between the production of the Matsutake mushroom and the meteorological factors, we consider the log-normal regression model using the the yield of the Matsutake mushroom as dependent variable (see Table 3.3). As a result, there are 4 significant weather variables. But, coefficient of determination is very low (R^2=0.27).

Therefore, we now consider the (hidden) year state as a dependent variable instead of the yield of the Matsutake mushroom. Then a variety of classification analyzes such as logistic regression can be conducted. For example, we consider a model with p meteorological predictors, $x_1, \ldots, x_p$, and one binary response variable C_t, which we denote $\pi = P(C_t = 1)$.

Table 3.2 The state transition probability matrix (left) and the mean and the standard deviation of distribution (right) for the year state

Bonghwa

	lean	bumper
lean	0.5845	0.4155
bumper	0.6969	0.3031

	lean	bumper
mean	11136.43	58555.87
sd	8829.524	14328.174

Yeong-ju

	lean	bumper
lean	0.3843	0.6155
bumper	0.3744	0.6256

	lean	bumper
mean	895.1527	9476.8595
sd	609.8546	4489.3181

Andong

	lean	bumper
lean	0.7146	0.2854
bumper	0.2655	0.7345

	lean	bumper
mean	4432.791	20447.27
sd	2608.135	5413.074

Yeongyang

	lean	bumper
lean	0.6634	0.3366
bumper	0.2428	0.7572

	lean	bumper
mean	1355.310	5770.481
sd	971.7514	3431.6677

Uljin

	lean	bumper
lean	0.8826	0.1174
bumper	1.0000	0.0000

	lean	bumper
mean	38266.58	127501.37
sd	24459.27	2500.00

Table 3.3 Estimated coefficient (Est.) and standard error (S.E.) values for Log-Nomal model

Variables	Intercept	hhour_precip8	lwat_pre7	hwat_pre7	mground_temp5
Est.	15.203	0.031	0.197	-0.165	-0.248
S.E.	0.723	0.014	0.065	0.072	0.099

A linear relationship between the predictor variables and the log-odds of the event that $C_t = 1$ is assumed, which can be written in the following form:

$$log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \tag{3.1}$$

In order to compare the performance of various classification models including logistic regression, decision tree, bagging and random forest, 5-fold cross-validation was used. The result shows that accuracy of logistic regression is the highest at 78%.

Table 3.4 Results of classification accuracy: 5-kold cross-validation approach

Model	Accuracy
Logistic regression	0.7847
Decision tree	0.6642
Bagging	0.7031
Random Forest	0.7442

Based on the optimal model which is the multiple logistic regression analysis (see Table 3.5), we identify the meteorological factors which explain the variation of the production of the Matsutake mushroom indirectly. It turns out that total of 9 variables were significant. More specifically, the average temperature in June, the maximum temperature in June, the relative humidity in June, and the water pressure in May were found to be additional variables affecting Matsutake mushroom production compared to the previous model.

Table 3.5 Estimated coefficient (Est.) and standard error (S.E.) values for GLM model

Variables	Intercept	hmin_precip8	precip8	hwat_pre7	m_humid6
Est.	284.375	0.167	0.007	-0.480	0.182
S.E.	101.356	0.088	0.003	0.193	0.087

Variables	lwat_pre6	h_temp6	m_temp6	hwat_pre5	lsea_pre5
Est.	-0.425	0.355	1.681	-0.513	-0.321
S.E.	0.225	0.201	0.525	0.171	0.105

4. Conclusion

In this paper, we propose an classification approach using hidden Markov model to generate binary hidden state variable through the HMMs and then analyze the state variables using various classification methods. It differs from the existing analysis method by using the state variable and the mixture distribution of the state rather than using the observed value directly in the analysis. This method can be used to identify the relevance between response variable and explanatory variables indirectly especially when it is difficult to show the direct relevance in the existing model. By application, we used the annual production of Matsutake mushroom(kg) data observed in five regions of Bonghwa-gun, Uljin-gun, Yeongju-si, Andong-si and Yeongyang-gun from 1997 to 2016. As a result, we could find additional meteorological factors affecting the annual production of Matsutake mushroom compared to existing methods.

References

Baum, L. and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics*, **37**, 1554-1563.

Bicego, M., Murino, V. and Figueiredo, M. (2004). Similarity-based classification of sequences using hidden Markov models. *Pattern Recognition*, **37**, 2281-2291.

Campbell, M., Macdonald, I. and Zucchini, W. (1998). Hidden Markov and other models for discrete-valued time series. *Biometrics*, **54**, 394-395.

Chib, S. (1996). Calculating posterior distributions and modal estimates in Markov mixture models. *Journal of Econometrics*, **75**, 79-97.

Jackson, J. C., Albert, P. S. and Zhang, Z. (2015). A two-state mixed hidden Markov model for risky teenage driving behavior. *Annals of Applied Statistics*, **9**, 849-865.

Kang, J. (2019). Comparison of false discovery rate procedures in microarray studies. *Journal of the Korean Data & Information Science Society*, **30**, 455-468.

Keroglou, C. and Hadjicostis, C. N. (2014). Hidden Markov model classification based on empirical frequencies of observed symbols. *IFAC Proceedings Volumes*, **47**, 7-12.

Kim, E. C. and Oh, K. J. (2019). Asset allocation strategy using hidden Markov model and genetic algorithm. *Journal of the Korean Data & Information Science Society*, **30**, 33-44.

MacDonald, I. and Zucchini, W. (1997). *Hidden markov and other models for discrete-valued time series*, Chapman and Hall, London.

Rabiner, L. and Juang, B. (1986). An introduction to hidden Markov models. *IEEE Acoustics, Speech, and Signal Processing Newsletter*, **3**, 4-16.

Robert, C., Celeux, G. and Diebolt, J. (1993). Bayesian estimation of hidden Markov chains: A stochastic implementation. *Statistics and Probability Letters*, **16**, 77-83.

Titterington, D., Smith, A. and Makov, U. (1985). *Statistical analysis of finite mixture distributions*, Wiley and Sons, Chichester.