

# 봄철서리발생예보를 위한 지지벡토르기와 론리회귀모형

함영식, 엄금철

농작물의 생육활동에 엄청난 피해를 주는 서리현상의 발생예보를 정확히 하는것은 농업부문에서 서리피해를 줄이기 위한 중요한 요구로 나선다.

론문에서는 봄철서리발생을 예보하기 위한 지지벡토르기와 론리회귀모형의 연구정형에 대하여 서술하였다.

## 1. 지지벡토르기와 론리회귀에 기초한 서리에보모형과 검증

### 1) 지지벡토르기와 론리회귀에 기초한 서리에보모형

1960년대에 통계적학습리론에 기초한 지지벡토르기리론이 제기되었다.[1, 2] 그러나 그 실제적인 적용알고리즘은 1990년대 중엽에 와서야 컴퓨터의 광범한 적용과 함께 실현 가능해졌다. 지지벡토르기의 기본원리는 입력자료를 리용하여 클라스들사이의 최대여백을 가지는 지지벡토르들에 의하여 결정되는 최량초평면을 찾는것이다.[1]

자료모임이 선형적으로 분리가능한 경우에 최량초평면은 라그랑주승수법을 리용하면 쉽게 구할수 있다. 초평면을 결정하는 지지벡토르들은 2차계획문제의 풀이로 되는데 이때 정의라그랑주승수에 대응하는 벡토르들은 지지벡토르들이며 판별함수는 이 지지벡토르들에 의하여 결정된다.

자료모임이 선형적으로 분리불가능한 경우에는 입력자료를 매우 높은 차원의 특징공간으로 사영함으로써 문제는 선형분리가능한 문제로 취급되게 된다. 이 경우 분류에서의 오류를 줄이기 위하여 완화변수를 추가한 연간격지지벡토르기의 개념이 적용되는데 라그랑주쌍대형식을 리용하여 2차계획문제로 넘겨서 지지벡토르를 구한다. 이때 핵함수라고 하는 고차원공간에서의 스칼라적합수를 리용하면 고차원공간으로의 넘기기함수를 명백히 정의하지 않아도 무방하다.

지지벡토르기에 기초한 모형의 식별능력을 높이기 위해서는 핵함수를 잘 선택하는것이 중요하다.[2]

론문에서는 흔히 리용되는 핵함수들인 선형핵함수(LSVM), 2차(QSVM) 및 3차다항식(CSVM)핵함수,  $K(X_i, X_j) = \exp(-\gamma \|X_i - X_j\|^2)$  로 정의되는 방사토대핵함수(GSVM)들을 리용하였다.

다음으로 론리회귀(LR)모형에 의한 다음의 식으로 사건발생의 확률을 계산한다.[3, 4]

$$p = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_n x_n)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_n x_n)} \quad (1)$$

여기서  $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_n$  은 회귀결수,  $x_1, x_2, \dots, x_n$  은 예보자료에 기초한 예보인자들이며  $p$  는 0부터 1사이의 실수로서 일기현상의 발생확률을 나타낸다.

만일 예보된  $p$ 가 어떤 턱값보다 크게 되면 사건발생을 예보한다.

## 2) 서리에보인자와 모형검증

서리에보인자로서 서리발생에 영향을 주는 다음의 기상요소들을 리용하였다.(표 1)

표 1. 서리에보인자와 그에 대한 설명

번호	서리에보인자	서리에보인자에 대한 설명
1	$RH_{03}$	03시 지면상대습도예보값
2	$T_{03}$	03시 2m지면기온예보값
3	$T_{d03}$	03시 이슬점온도예보값
4	$U_{03}$	03시 10m위권바람속도성분예보값
5	$V_{03}$	03시 10m자오선바람속도성분예보값
6	$T_{03} - T_{925,03}$	지면과 925hPa등압면기온과의 차
7	$T_{925,03} - T_{850,03}$	925hPa등압면과 850hPa등압면기온과의 차

수치예보모형이 09시와 21시에 실행되므로 09시 기초시간인 경우에 다음날 03시까지 18h예견기, 21시 기초시간인 경우에는 다음날 03시까지 예견기가 6h, 2일후 03시까지는 30h이며 연구에서는 이 3개의 예견기에 대하여 모형을 개발하였다.

모형의 검증은 이미 잘 알려진 성능평가지표들인 비판적성공지수(CSI), 공보률(FAR), 편차성적(BIAS), 정확한 백분률(PC)을 리용하여 진행한다.[4]

## 2. 계산결과와 모형들사이의 비교

연구에서는 봄철서리에보를 위한 계산실험지점으로서 평양지방의 5개의 관측지점(교, ㄱ, ㄴ, ㄷ, ㄹ)들을 선정하였다.

또한 지지벡토르기와 논리회귀모형에 기초한 서리에보모형을 구성하기 위하여 2012년—2018년까지의 봄철(3월과 4월)기간의 수치예보자료와 우에서 선정된 기상관측지점들에서 3h에 한번씩 진행한 서리관측자료를 리용하였다.

2012년—2016년까지의 자료는 모형개발에 리용되었으며 2017년—2018년기간의 자료는 모형검증에 리용되었다.

검증자료(2017년—2018년)에 대한 지점별예견기에 따르는 예보성적을 보면 표 2와 같다.

표 2에 제시된바와 같이 검증기간 제일 좋은 73.3%의 CSI, 15.4%의 FAR, 1.00의 BIAS, 90.5%의 PC지수들은 ㄱ지점에서 30h예견기를 가진 CSVM모형에서 나타났으며 반면에 가장 낮은 30.3%의 CSI는 교지점에서 30h예견기를 가진 QSVM에서, 가장 큰 53.5%의 FAR와 가장 작은 PC는 ㄱ지점에서 9h예견기를 가진 논리회귀모형에서, 가장 큰 1.64의 BIAS는 6, 18, 30h예견기를 가진 논리회귀모형에서 나타났다.

표 3은 각이한 모형들에 대한 연구된 5개 지점들에서의 예견기평균 CSI, FAR, BIAS예보성적을 보여준다.

표 3에서 보는바와 같이 매 모형들에 대한 지점별CSI는 각이하며 교지점과 ㄱ지점에서의 성적이 뚜렷하게 낮고 SVM모형들은 ㄴ, ㄷ, ㄹ지점들에서 거의 동일한 성적을 가진다.

표 2. 검증자료(2017년-2018년)에 대한 지점별예견기에 따르는 예보성적

지점	모형	예견기											
		6h				18h				30h			
		CSI/%	FAR/%	BIAS	PC/%	CSI/%	FAR/%	BIAS	PC/%	CSI/%	FAR/%	BIAS	PC/%
교	LR	36.8	50.0	1.17	79.8	37.8	48.1	1.13	81.1	43.6	46.9	1.33	80.7
	LSVM	43.8	36.4	0.92	80.6	35.5	38.9	0.75	79.2	33.3	45.0	0.83	75.0
	QSVM	44.8	36.4	0.75	82.8	38.7	38.9	0.79	80.2	30.3	47.4	0.79	73.8
	CSVM	44.8	31.6	0.75	82.7	35.5	38.9	0.75	79.2	34.4	42.1	0.79	76.1
	GSVM	44.9	27.8	0.75	82.8	34.4	42.1	0.79	78.1	32.4	47.6	0.88	73.9
코	LR	59.1	36.6	1.41	84.9	56.5	39.5	1.48	83.6	61.9	33.3	1.34	86.0
	LSVM	59.5	26.7	1.03	87.4	56.4	31.3	1.10	86.1	58.3	25.0	0.97	86.8
	QSVM	61.5	29.4	1.17	87.4	60.5	28.1	1.10	87.7	67.7	17.8	0.97	90.4
	CSVM	60.0	31.4	1.2	86.6	60.5	32.4	1.10	87.7	62.2	25.8	1.07	87.7
	GSVM	63.1	31.3	1.19	88.3	62.2	32.3	1.08	85.3	64.9	25.0	1.10	88.6
나	LR	60.5	34.3	1.35	83.1	55.3	36.4	1.27	81.5	59.5	33.3	1.27	82.1
	LSVM	64.9	31.4	1.34	85.4	63.9	30.3	1.26	85.9	60.0	30.0	1.15	83.3
	QSVM	67.6	25.8	1.19	87.6	66.7	24.1	1.12	88.0	65.6	22.2	1.04	86.9
	CSVM	67.6	25.8	1.19	87.6	63.6	25.0	1.08	87.0	73.3	15.4	1.00	90.5
	GSVM	72.7	25.0	1.19	89.9	66.8	24.1	1.11	88.0	65.6	22.2	1.04	86.9
다	LR	56.6	36.2	1.31	75.5	57.7	34.8	1.28	77.3	56.6	36.2	1.30	74.2
	LSVM	60.0	30.2	1.16	79.4	58.3	28.2	1.05	80.0	63.8	25.0	1.08	81.5
	QSVM	62.0	29.5	1.19	80.4	58.3	32.5	1.05	80.0	65.9	24.4	1.11	82.7
	CSVM	62.0	29.5	1.18	80.4	60.0	22.8	0.94	82.0	66.0	24.4	1.10	82.6
	GSVM	64.0	28.9	1.22	81.4	63.3	27.9	1.16	82.0	63.8	25.0	1.08	81.5
나	LR	35.9	53.5	1.30	73.4	37.8	50.0	1.22	76.3	41.0	50.0	1.39	74.2
	LSVM	45.0	45.5	1.32	81.5	40.5	45.5	1.36	79.5	41.5	45.5	1.32	78.9
	QSVM	58.5	40.0	1.60	85.7	51.3	40.0	1.36	84.0	55.0	40.0	1.48	84.2
	CSVM	46.3	45.7	1.42	81.5	43.2	45.7	1.52	79.5	45.5	45.7	1.56	78.9
	GSVM	57.1	41.5	1.64	84.9	57.1	41.5	1.64	85.2	57.1	41.5	1.64	84.2

LR모형에서는 모든 지점들의 CSI성적이 60%보다 작으며 반면에 코지점에서 LSVM모형을 제외하고는 3개 지점들에서 성적이 60%이상을 기록한다. 특히 나지점에서 지지벡토르모형의 성능이 가장 높고 다, 코, 나, 고지점순서로 낮아진다.

FAR지수는 모든 모형들에서 나지점이 가장 큰값을 가지며 고지점과 나지점은 40%를 넘는다. 즉 이 지점들에서 모형의 허의경보률이 매우 높다는것을 알수 있다.

그러나 코, 나, 다지점들에서는 비교적 안정하게 나타나고 30%보다 큰 LR모형을 제외한 SVM모형들에서는 30%이하이며 나지점에서는 CSVM모형이 가장 작은 값을 나타낸다.

제안된 모형들의 BIAS성적으로부터 론리회귀모형과 지지벡토르기모형은 고지점을 제외한 모든 지점들에서 과대평가하며 고지점에서는 론리회귀모형은 과대예보하지만 다른 나머지 모형들은 과소예보한다는것을 알수 있다.

특히 CSI와 FAR성적이 좋은 코, 나, 다지점들에서 론리회귀모형을 제외한 모형들에서 1.2보다 작은 값을 가지고 비교적 1에 가까운 BIAS를 가진다.

우의 결과는 연구된 5개 지점들이 평양지방안의 조밀한 관측지점들임에도 불구하고 모형성능에서 큰 차이를 보여준다. 나, 다, 코지점들은 비교적 좋은 결과를 보여주지만 고, 나지점들은 그렇지 못하다.

이것은 매 지점들의 기후학적 및 자연지리적조건에서의 차이 그리고 서리발생시간을 03시로 제한한것에 의하여 산생된 결과일수 있다.

표 3. 각이한 모형들에서 예견기평균성적

지점	모형	CSI/%	FAR/%	BIAS
교	LR	39.4	48.3	1.2
	LSVM	37.5	40.1	0.8
	QSVM	37.9	40.9	0.8
	CSVM	38.2	37.5	0.8
	GSVM	37.2	39.2	0.8
□	LR	59.2	36.5	1.4
	LSVM	58.1	27.7	1.0
	QSVM	63.2	25.1	1.1
	CSVM	60.9	29.9	1.1
	GSVM	63.4	29.5	1.1
入	LR	58.4	34.7	1.3
	LSVM	62.9	30.6	1.3
	QSVM	66.6	24.0	1.1
	CSVM	68.2	22.1	1.1
	GSVM	68.4	23.8	1.1
ㄷ	LR	56.9	35.7	1.3
	LSVM	60.7	27.8	1.1
	QSVM	62.1	28.8	1.1
	CSVM	62.7	25.6	1.1
	GSVM	63.7	27.3	1.1
ㄱ	LR	38.2	51.2	1.3
	LSVM	42.3	45.5	1.3
	QSVM	54.9	40	1.5
	CSVM	45	45.7	1.5
	GSVM	57.1	41.5	1.6

## 맺 는 말

1) 지지벡토르기모형은 모든 지점들에서 모든 예견기에 대하여 논리회귀모형보다 우월하다. 선형, 2차, 3차, 가우스지지벡토르기의 예보성능은 크게 차이나지 않지만 3차와 가우스핵지지벡토르기들이 비교적 높은 성적을 가지었다.

2) 入, ㄷ, □관측지점들에서의 지지벡토르기에 기초한 서리예보성적은 CSI가 60%이상으로서 비교적 좋은 결과를 보여주었지만 교, ㄱ지점들에서는 결과가 좋지 못하였다. 그 원인은 도시열섬효과와 지역의 기후학적특징으로 산생된 결과라고 볼수 있다.

따라서 예보성적을 높이자면 도시지역에 따르는 보다 더 적중한 예보인자를 선택하여야 한다.

3) 지지벡토르기에 기초한 모형들은 서리현상을 과대예보하는 경향성을 가진다. 교지점을 제외한 모든 지점들에서 평균 BIAS성적은 1보다 큰 값을 가지고 과대예보되었다.

이것은 5년간의 짧은 기초학습자료와 관련된다고 볼수 있다.

## 참 고 문 헌

- [1] V. N. Vapnik; Transactions on Neural Networks, 8, 6, 1564, 1999.
- [2] V. N. Vapnik; Statistical Learning Theory, Wiley-Interscience, 401~440, 1998.
- [3] J. Aaron et al.; Monthly Weather Review, 140, 3054, 2012.
- [4] D. S. Wilks; Statistical Methods in the Atmospheric Sciences, Academic Press, 201~299, 2006.

주체108(2019)년 4월 5일 원고접수

## **Support Vector Machine and Logistic Regression for Spring Frost Prediction**

*Ham Yong Sik, Om Kum Chol*

In this paper, Support Vector Machine(SVM) and Logistic Regression (LR)-based models were suggested to make a prediction model by using NWP model output for spring frost in the Pyongyang region.

Key words: frost, Support Vector Machine(SVM), Logistic Regression (LR)