ISSN 1330-3651 (Print), ISSN 1848-6339 (Online) https://doi.org/10.17559/TV-20190903094335
Original scientific paper

Radar-based Hail-producing Storm Detection Using Positive Unlabeled Classification

Junzhi SHI, Ping WANG, Di WANG, Huizhen JIA

Abstract: Machine learning methods have been widely used in many fields of weather forecasting. However, some severe weather, such as hailstorm, is difficult to be completely and accurately recorded. These inaccurate data sets will affect the performance of machine-learning-based forecasting models. In this paper, a weather-radar-based hail-producing storm detection method is proposed. This method utilizes the bagging class-weighted support vector machine to learn from partly labeled hail case data and the other unlabeled data, with features extracted from radar and sounding data. The real case data from three radars of North China are used for evaluation. Results suggest that the proposed method could improve both the forecast accuracy and the forecast lead time comparing with the commonly used radar parameter methods. Besides, the proposed method works better than the method with the supervised learning model in any situation, especially when the number of positive samples contaminated in the unlabeled set is large.

Keywords: hailstorm; machine learning; positive unlabeled learning; weather forecasting; weather radar

1 INTRODUCTION

Hailstorms can bring severe damages to buildings, crops, vehicles, and other personal properties. As hailstorm has a short duration and small spatial scale, its detection and now forecasting are always challenging subjects. Currently, the most accurate hail forecasting methods rely on weather radars due to the fact that they could generate high-resolution volume data by scanning at multiple elevations.

Commonly in radar-based severe weather forecasting, corresponding radar parameters should be extracted from the radar images. In existing studies, various radar parameters have been used for hailstorm detection, such as maximum reflectivity [1], Waldvogel parameter [2], vertically integrated liquid water content (VIL) [3-5], VIL density [4], and severe hail index (SHI) [6]. These parameters can be used independently or in combination [7] with the combination methods include linear discriminant analysis [8], logistic regression [9], principal component analysis [10], and other machine learning methods [11-13].

Machine learning methods, which could learn from data, then make decisions without being explicitly programmed, have been proven to be effective in severe weather forecasting [14-19]. Some nonlinear models such as support vector machine (SVM) and random forest (RF) can find hyperplanes which can solve linear indivision problems. Therefore, in theory, machine learning models could be used to distinguish hail storms from no-hail storms by using radar parameters as features.

However, in practice, the performance of machine learning models is deeply influenced by data quality. Some weather data sets are difficult or expensive to be acquired, while some are not very accurate. Hail case data is such a kind of data. In many countries, the most reliable hail case data are from hail reports manually recorded by meteorological observation stations. Some areas make use of hailpads to automate or semi-automate the recording process [20-23]. Recently, some novel weather case collection methods are proposed. For example, the NOAA National Severe Storms Laboratory is using a mobile phone application named "mPing" to collect crowd-sourcing weather reports [24], and someone uses data mining techniques to crawl severe weather records from social networking sites like Twitter [25]. Either way, the time and place of hail occurrence have a massive impact on whether a case is recorded correctly. Another problem is that many hail reports are recorded in various text formats. They are not easy to be converted to structural data, which is needed for machine learning models.

Even if the hail case data are correct, labeling the radar data to the corresponding weather case is still very costly. Especially for the supervised learning, both the positive sample set and the negative sample set should be well labeled. However, in hail detection, if treating all the samples without hail record as negative samples in hail classification, due to hail cases being easy to miss, a large number of false-negative samples will contaminate the negative sample set. This may substantially affect the performance of classification.

Based on the above background, we find that supervised classifiers may have limited performance in hail storm detection. Training a classification model for a specific geographical area needs a lot of historical data collection and labeling work. Besides, the classification model is generally not universal due to the differences in climate and topography in different geographical areas. Applying the same model to other regions still requires much work. Therefore, reducing the cost of data labeling and processing is critical to applying machine-learning-based models to operational weather forecasting.

Weakly supervised learning refers to a class of models that attempt to learn from weakly supervised data [26]. Weak supervision can be divided into three categories: incomplete, inexact, and inaccurate supervision. Incomplete supervision means that only a subset of the training set is labeled. For inexact supervision, only coarse-grained labels are given. When the given labels are not always ground-truth, it is inaccurate supervision. Besides, Incomplete supervision includes semi-supervised learning [27-31], active learning [32-34], and transfer learning [35, 36].

In the task of hail storm classification, one can easily obtain a part of accurate case data from the hail reports, but accurately labeling the whole data is costly. A feasible solution is to train a weakly supervised classifier with the data set that only a subset of positive samples is labeled,

which is a typical positive unlabeled learning (PU learning) problem, one of the incomplete supervision methods. Unlike supervised learning using a totally labeled positive training set **P** and a negative training set **N**, PU learning requires only a positive training set **P**, which includes the partly labeled hail case data and an unlabeled set **U**, which includes all of the other unlabeled data.

There are various categories of approaches to solving PU learning problems [37]: (i) approaches that identify possible negative data in the unlabeled set using heuristic methods then perform supervised learning [38-42], (ii) approaches that regard the unlabeled set as negative set, but introduce a biased weight to classification models to penalize more misclassification of positive instances than misclassification of unlabeled instances [39, 43-45], (iii) approaches that treat the PU learning problem as one-class learning problems, which learn from positive samples only [46-49], and (iv) approaches that make use of bootstrap methods to build aggregate classifiers based on positive and unlabeled samples [50, 51].

In this paper, a machine-learning-based hail-producing storm detection method used for hail forecasting is presented. The machine learning model is designed based on the characteristics of the samples and the problem. The features are extracted from radar and sounding parameters based on operational forecasting experience and convective physical processes. One of the state-of-the-art PU classification models, the bagging class-weighted SVM, is used as the classification model in order to alleviate the problem that hail cases cannot be fully recorded. Then the method is compared with the classic radar parameters method and the method using supervised classification with real historical data for validation.

The paper is structured as follows: Section 2 gives a detailed description of the data used in this paper and the proposed method. The results of validation and discussion are provided in Section 3. The last Section 4 shortly draws the most important conclusions.

2 DATA AND METHODOLOGY

2.1 Data Sources

The data used in this paper include Doppler weather radar data, radiosonde sounding data, and severe weather observational data. Due to the type of severe weather varying with region, topography, and season, in order to avoid these effects on the model parameters, we focus the study on the convective seasons of Beijing-Tianjin-Hebei region, in North China. The radar data used in this paper are generated from three single-polarization S-Band radars, which are deployed in Tianjin, Beijing, and Shijiazhuang, respectively. The radars perform volume scans once every six minutes, and each volume scan includes nine elevations. The resolution of the generated plan position indicator (PPI) image is 1 × 1 km. The sounding data are from the nearby radiosonde stations, and are acquired twice a day at 0000 UTC and 1200 UTC. Each case uses the latest data before, and the data of each grid point are obtained by bilinear interpolation.

The hail case data are from the hail reports of the manual observation stations provided by the China Meteorological Administration. The manual observation stations record hail cases based on human eye-observations of hailstones of any size. Among the hail reports from 2011 to 2015, we extracted 146 hail cases that generate hailstones larger than 10 mm and are under the coverage of the radars. All of these cases have clear records of time and locations. If a hail case is detected by two or more radars at the same time, only the nearest radar is used. The geographical locations of radars and manual observation stations are shown in Fig. 1.

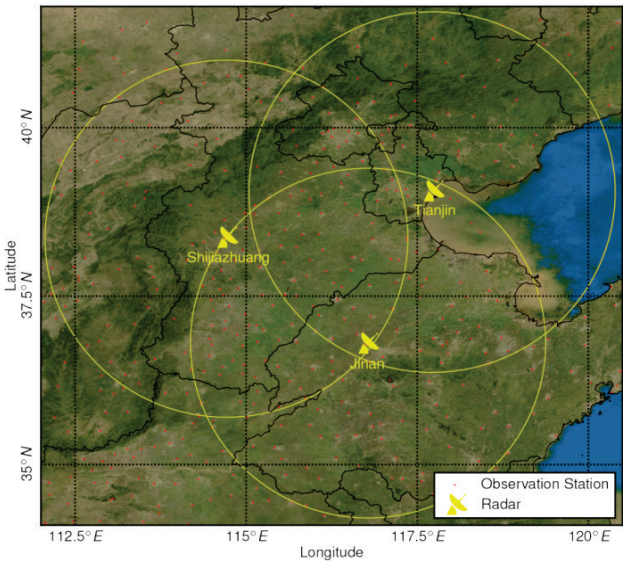

Figure 1 The geographical locations of the radar and observation stations. The yellow circle represents the scan range of the radars.

2.2 Data Handling

All the algorithms in this study are conducted on convective cells. To identify convective cells, we use a modified SCIT method [52], which utilizes a border following algorithm [53] to extract 2D components from PPI images instead of using radial images. After identifying the convective cells, a convective cell is labeled as a hail-producing cell if it is located above a manual observation station that reports hail during the recording period. Since the record time often lags behind the actual time of hail fall and the forecast lead time should be considered, we tracked backward until the time step when the severe convective cells appear and labeled the cells in the same hail process as positive samples. Finally, 1521 convective cells are labeled as hail-producing samples.

Although one can train the PU classifier with the positive and unlabeled set only, a refined negative set is still needed to evaluate the performance of the classifier. It is not feasible to regard all the convective cells that are not recorded in the hail reports as negative samples because the number of no-hail storms is too large, and there are missing cases on the hail reports. So, we prepared the negative set as follows. First, identify all the convective cells of the basedata between 0000 UTC and 1000 UTC. We chose these basedata because during this period, in the North China region is the daytime, so the hail cases are not easy to miss. Second, we only kept the convective cells whose maximum reflectivity is larger than 45 dBZ. According to the local historical cases, convective cells that do not meet this condition will hardly produce hail. Third, we removed all the convective cells that were related to any hail reports or were far away from any observation station.

After the above processing, we can obtain a negative sample set as clean as possible. However, this data set is still too large. Also, we should test the model trained by an unlabeled sample set with a high ratio of positive samples to see the performance under extreme conditions. So, we randomly extracted 13689 samples, nine times more than the number of hail samples, from it. When training the PU classifier, we randomly incorporate positive samples into the negative sample set to obtain an artificially generated unlabeled sample set. As the actual label of each sample is clear, we could verify whether each sample is correctly classified.

2.3 Features

Features are crucial for machine learning models. In this study, we divide the features used in the classification model into two groups: main features and auxiliary features. The main features are the classic radar parameters that can be used independently for hail detection, including maximum radar reflectivity in a vertical column (Z_{max}), Waldvogel parameter, *VIL* density, and SHI. As these parameters have been proven to be effective [7, 54-58], we do not need to verify their importance as features for classification. Therefore, the values of the main features are directly input into the classification model after standardization. The following is a brief introduction of the main features:

Z_{max} is the most straightforward criterion which predicts the presence of hail if the maximum reflectivity in a vertical column exceeds a certain threshold.

Waldvogel parameter is proposed by [2]. It predicts hail if the vertical distance between R_W dBZ echo top and the melting layer is greater than or equal to a threshold H_T:

$$WP = H_{ET-R_W} - H_{0\,°C} \geq H_T \tag{1}$$

Initially, the reflectivity threshold R_W is 45 dBZ and the height threshold H_T is 1.4 km.

VIL density is proposed by [4] to improve the warning of severe hail on the basis of *VIL*. It is defined as the *VIL* is divided by the radar echo top H_{ET}:

$$VILd = VIL / H_{ET} \tag{2}$$

One form of *VIL* is given by:

$$VIL = \sum 3.44 \times 10^{-6} \left[\left(Z_i + Z_{i+1} \right) / 2 \right]^{4/7} \Delta h \tag{3}$$

where Z_i and Z_{i+1} are radar reflectivity values at the lower and upper portions of the sampled layer, and Δh is the vertical thickness of the layer.

Table 1 Auxiliary features

	Abbreviation	Description	Unit
Radar	ET30	Height of 30 dBZ echo top	km
	ET45	Height of 45 dBZ echo top	km
	ET55	Height of 55 dBZ echo top	km
	H_c	Height of convective cell core centroid	km
	$Z_{max\ 0}$	Maximum reflectivity at the height of melting layer	dBZ
	$Z_{max\ m20}$	Maximum reflectivity at the height of −20 °C layer	dBZ
	A_l	Length of the convective cell major axis	km
	s^{-1}	Length of the convective cell minor axis	km
	V_{cell}	Volume of convective cell	km^3
	V_{core}	Volume of convective cell core (45 dBZ)	km^3
	V_{oh}	Volume of overhang echo	km^3
	v_h	Horizontal moving speed of convective cell centroid	km(6 min)$^{-1}$
	v_d	Rising speed of convective cell core	km(6 min)$^{-1}$
	v_{ET30}	Rising speed of 30 dBZ echo top	km(6 min)$^{-1}$
	v_{ET45}	Rising speed of 45 dBZ echo top	km(6 min)$^{-1}$
	ΔZ_{max}	Difference of maximum reflectivity between two volume scans	dBZ
	G_Z	Maximum reflectivity gradient near convective cell core	dBZ km^{-1}
Sounding	H_0	Height of the melting layer	m
	H_{m20}	Height of the −20 °C layer	m
	CAPE	Convective available potential energy	J kg^{-1}
	CIN	Convective inhibition	J kg^{-1}
	LIFT	Lifted index	°C
	ShowI	Showalter index	°C
	DT500	Difference of temperature at 500 hP	°C
	PWAT	Precipitable water for the entire sounding	mm
	EQLV	Equilibrium level	hPa
	WS6	Vertical wind shear (0-6 km)	s^{-1}

SHI is a thermally weighted vertical integration of reflectivity profile of a convective cell based on the semi-empirical relationship between the flux values of the hail kinetic energy and the radar reflectivity [6]:

$$SHI = 0.1 \int_{H_0}^{H_T} W_T \left(H \right) \dot{E} \mathrm{d}H \tag{4}$$

where H_0 is the height of the melting layer, H_T is the height of the storm top, $W_T(H)$ is the temperature-based weighting function, and $\dot{E}$ is the kinetic energy flux of the hailstones.

Note that Z_{max}, Waldvogel parameter, and *VIL* density are grid-based parameters, which calculate a value at a point or in its neighborhood, but our algorithm is cell-oriented. So, we should convert them into cell-based parameters. [52] has defined cell-based *VIL* by vertically

integrating a three-gate-averaged maximum reflectivity at each level through the depth of the storm.

Then based on it, [59] defines cell-based *VILd* by a ratio of the cell-based *VIL* to the storm top. The Z_{max} of a 3D convective cell is the 27-grid-averaged maximum reflectivity inside the convective cell, and the Waldvogel parameter of a convective cell is the vertical distance between the storm top and the melting layer.

The auxiliary features refer to a set of radar and sounding products that may have potential relationships with hail storms. These values cannot be used for hail detection independently but may improve the performances of machine learning models. The design and selection of auxiliary features are based on operational forecasting experience and convective physical processes, and also benefit from some previous studies [10, 11, 15, 60-65]. The radar products and sounding products introduced as auxiliary features are listed in Tab. 1.

However, more features do not mean that the classification results will get better. As the scale of the training sample set is not large, using a complicated model with too many features may make the model have a high variance. A model with high variance is overfitting to noisy or unrepresentative training data, resulting in a decline of performance [67, 68]. Since we do not have negative sample sets in the actual situation, it is hard to select useful features carefully. A feasible solution is employing an unsupervised dimensionality reduction method like principal component analysis (PCA), and using the principal components as features. We performed PCA on the dataset, and the contribution rates of the top 10 principal components are shown in Fig. 2. From it, we can see that the cumulative contribution rates of *PC*1 to *PC*7 have reached 89.85%. So, we choose the first seven principal components as features for the bagging CWSVM classifier. The overview of features is shown in Fig. 3.

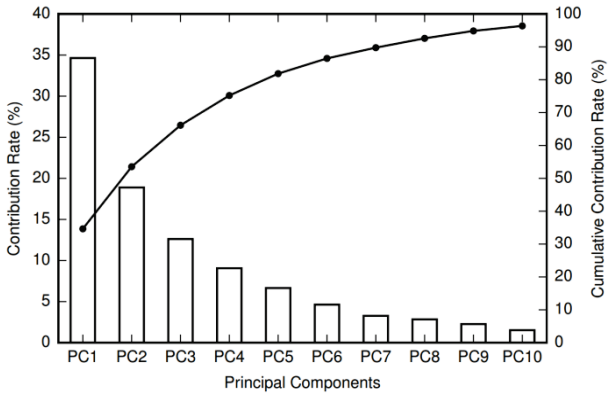

Figure 2 The contribution rates (bar plot) and the cumulative contribution rates (line plot) of the principal components. *PC*1, ..., *PC*10 refer to the principal components with top 10 contribution rates

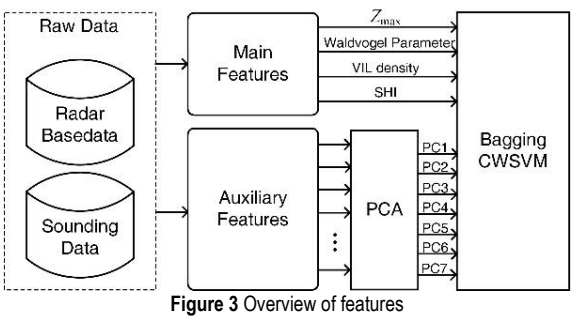

Figure 3 Overview of features

2.4 Model

The task of hail-producing storm detection can be transformed into a binary classification problem, with hail-producing cells as positive samples and no-hail cells as negative samples. However, considering hail cases cannot be completely recorded, constructing a PU classifier trained from partially labeled hail-producing cells and the other unlabeled cells is more suitable. In this study, we picked one of the state-of-the-art PU learning models, the bagging class-weighted SVM (CWSVM) [50], for this task, using the features described above. As mentioned before, four categories of approaches can be used to solve the PU learning problems. The bagging CWSVM combines two of them: the one is biased weight, and the other is bootstrap. In short, bagging CWSVM uses class-weighted SVM as base classifier then applies bootstrap aggregating (bagging) to further reduce the variances caused by the randomness in the negative samples.

Compared with the classic SVM, the CWSVM penalizes the misclassification of each class using an independent weight [40, 68]. In the context of PU learning, the penalty weight of misclassified positive samples **P** is larger than the penalty weight of misclassified unlabeled samples **U**, because the unlabeled set that is assumed to be negative also contains positive data. Then the optimization problem is:

$$\begin{aligned} \min_{\alpha,\,\xi,\,b} \quad & = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j\boldsymbol{y}_i\boldsymbol{y}_j\mathrm{K}\left(\boldsymbol{x}_i,\boldsymbol{x}_j\right)+ \\ & \quad + C_{\mathrm{P}}\sum_{i\in\mathrm{P}}\xi_i + C_{\mathrm{U}}\sum_{i\in\mathrm{U}}\xi_i \\ \text{s. t.} \quad & \boldsymbol{y}_i\left(\sum_{j=1}^{N}\alpha_j\boldsymbol{y}_j\mathrm{K}\left(\boldsymbol{x}_i,\boldsymbol{x}_j\right)+b\right)\geq 1-\xi,\ i=1,\ldots,N, \\ & \xi_i\geq 0,\ i=1,\ldots,N \end{aligned} \tag{5}$$

with $\alpha\in\mathbb{R}^N$ the support values, $\boldsymbol{y}\in\{-1,+1\}^N$ the label vector, $\mathrm{K}(\cdot,\cdot)$ the kernel function, b the bias term and $\xi\in\mathbb{R}^N$ the slack variables.

Bagging is an ensemble meta-algorithm to improve stability and accuracy, which is often applied to high-variance models [69]. Bagging constructs a sub-classifier using a subset drawn from the training set uniformly and with replacement, then combines their predictions. In PU learning, the bagging CWSVM draws a subset from **U**, and combines it with the whole **P**, as a training set to train the CWSVM. As the **U** is "contaminated" by positive samples, each subsampling will construct a subset with different portions of "contamination", which eventually will induce a large variability in the sub-classifiers. For this reason, bagging could improve the overall performance of PU learning.

The parameters needed to be tuned in bagging CWSVM include the number of samples drawn each time from the unlabeled set K, the number of classifiers for bagging T, and the penalty weights C_{P} and C_{U}. Commonly, in PU learning, the penalty weights C_{P} and C_{U} are set to make the total penalty equal for the two classes [70, 71]:

$$C_{\mathrm{P}}n_{\mathrm{P}} = C_{\mathrm{U}}K \tag{6}$$

where n_P is the size of **P**. Since the ratio n_P/K is fixed, only needs to tune C_P.

3 EXPERIMENTS AND RESULTS

We conducted a series of experiments to answer the following questions: (i) How much does the PU learning method improve comparing with the traditional radar parameter method? (ii) What are the performances when the ratio of hail samples contaminated in the unlabeled set is different? (iii) What if using the supervised classifier directly for positive and unlabeled classification? In other words, do we really need PU classification? (iv) What is the forecast lead time of the PU learning method?

3.1 Experiment Setup

The construction of datasets in this study is a little complicated compared with ones used for evaluating supervised learning, which is summarized in Fig. 4. As mentioned before, we prepared a refined no-hail sample set and should use a certain proportion of positive samples for contamination to construct the simulated unlabeled set. So the labeled hail sample set is divided into three: one for training, one for testing, and the third for contaminating. In this step, we split the hail samples in the unit of cases, considering that the features of hail samples in the same case may be similar, which can make the model easy to generalize. By random selection, 100 out of 146 hail cases are used for training and contaminating, and the rest is for testing. Accordingly, 68% of no-hail cases are added to the training set.

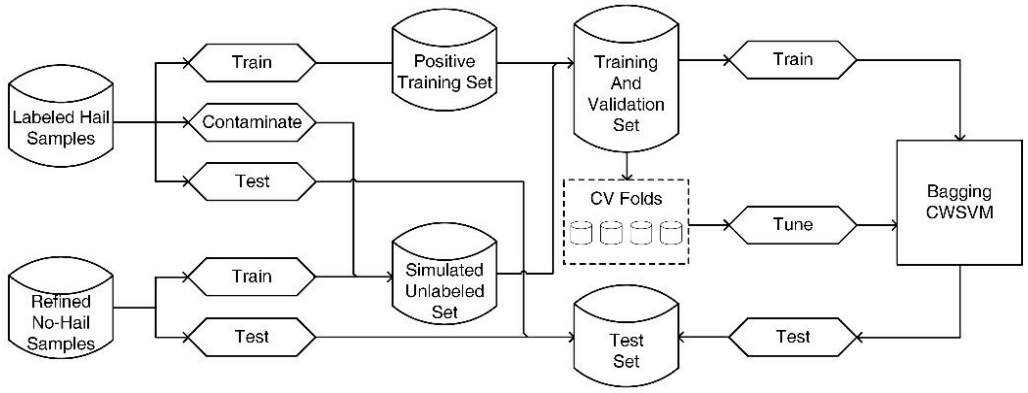

Figure 4 The schematic representation of dataset construction. CV refers to cross-validation

Part of the samples from the 100 cases was used to contaminate the unlabeled set, and this selection is in the unit of cells. Since we also want to test the performance with different contamination ratio, the number of cells for contaminating is different in each experiment. However, to ensure the comparability of the results in all experiments, samples in the positive set should remain the same. There are a total of 1027 convective cells in the 100 hail cases used for training, and 540 of them are taken as positive samples.

The hyperparameters of the bagging CWSVM model comprise the number of classifiers for bagging T, the number of resamples from the unlabeled set K, the positive class penalty weight C_P, and the other hyperparameters inherited from SVM. In theory, the performance is monotonically non-decreasing in T. Although the training time will increase with T, we set it to a large valve, 200, since we only focus on the performance. In addition, we found in our preliminary study that the gamma and the kernel types of SVM have little effects on the final results. Therefore, we assign gamma to the reciprocal of the number of features, which is a conventional treatment, and use the radial basis function as the kernel function. Consequently, there remain two hyperparameters that need to be determined. Due to the small number of samples, we also make use of the training set to tune them, instead of using an independent validation set. While tuning, the training set is fixed to the contamination ratio of 10%, and is split into four-folds. Then we conduct a grid search using 4-fold cross-validation to find the optimal parameter combination. Results of the grid search show that the optimal choice is $C_P = 100$ and $K = 2000$.

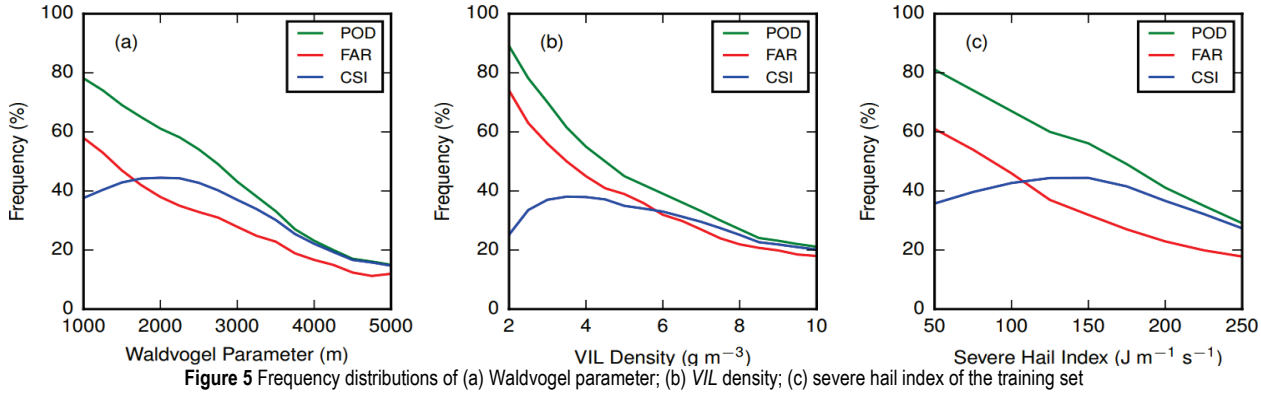

Figure 5 Frequency distributions of (a) Waldvogel parameter; (b) *VIL* density; (c) severe hail index of the training set

After obtaining the datasets and hyper parameters, a series of experiments with different contaminated rates are conducted. First of all, the proposed method is compared with three radar parameter methods, Waldvogel parameter, *VIL* density, and *SHI*. The warning thresholds used in the radar parameter methods are obtained by statistics on the same training set, as shown in Fig. 5. The threshold which could acquire the highest *CSI* is selected. Then, in order to demonstrate whether the PU learning is necessary, we compare it with the classic two-class SVM. At last, the forecast lead time is compared to see if the proposed method could forecast earlier than the traditional methods.

The metrics used for evaluation include area under the ROC curve (*AUC*), probability of detection (*POD*), false alarm rate (*FAR*), and critic success index (*CSI*). *AUC* measures the entire two-dimensional area underneath the entire receiver operating characteristic (ROC) curve, which provides an aggregate measure of performance across all possible classification thresholds. It is a commonly-used metric in evaluating classifiers. The *POD*, *FAR* and *CSI* are respectively defined as:

$$POD = \frac{TP}{TP + FN} \tag{7}$$

$$FAR = \frac{FP}{TP + FP} \tag{8}$$

and

$$CSI = \frac{TP}{TP + FN + FP} \tag{9}$$

where *TP* represents the number of true positives, that is the detected events, *FN* represents the number of false negatives, that is the miss-detected events, and *FP* represents the number of false positives, that is the false alarmed nonevents. These three metrics are commonly-used in evaluating weather forecasting methods.

3.2 Experiment Results

The performance diagram in Fig. 6 shows the *POD* and precision of traditional radar parameter methods and the proposed PU learning method trained by positive sets and unlabeled sets with different contamination ratios. As can be seen, among the three traditional radar parameters, the performances of the Waldvogel parameter and the SHI are similar, and they are better than the *VIL* density. The proposed bagging CWSVM methods trained by the unlabeled sets with up to 10% contamination rate significantly outperform the radar parameter methods. When the contamination rate reduces, the performance will improve. The *ROC* curve in Fig. 7 also demonstrates the same results.

The bar plots in Fig. 8 show the *POD*, *FAR*, *CSI* and *AUC* of different models. The warning thresholds of the three traditional radar parameter methods are selected using the same training set without contamination. From this figure, it is clear that the proposed model can both detect more positive samples and reduce false alarms compared with traditional methods. Unlike the contamination rate that has less influence on the *FAR*, contaminating the unlabeled set with more positive samples would lower the *POD*. When the contamination rate increases to 10%, the *POD* of the PU learning model drops to the same level as in the Waldvogel parameter method and the *VIL* density method.

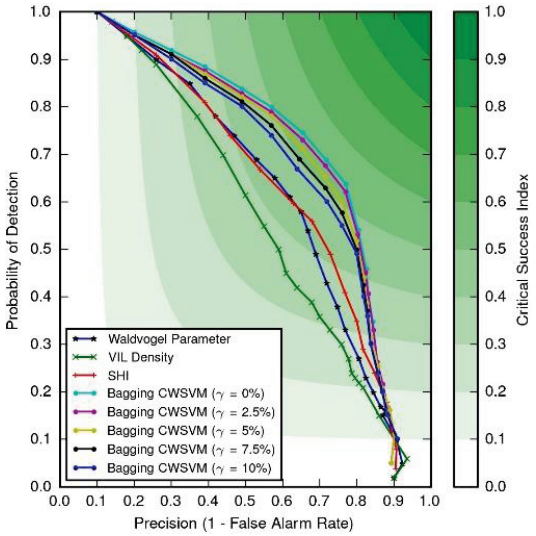

Figure 6 Performance curves for traditional radar parameter methods and the bagging CWSVM methods trained by the datasets with different contamination rates γ

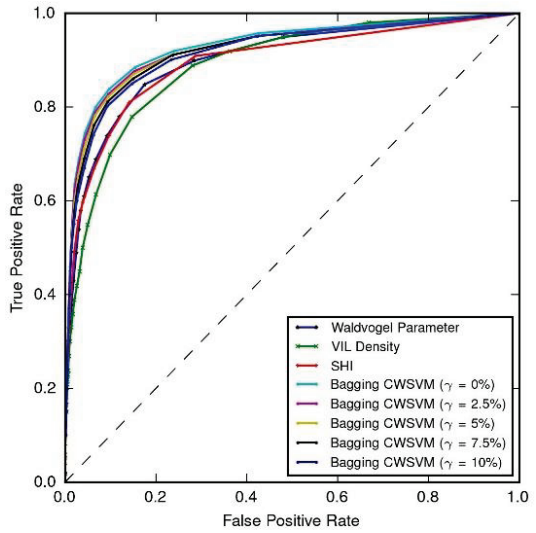

Figure 7 ROC curves for traditional radar parameter methods and the bagging CWSVM methods trained by the datasets with different contamination rates γ

The changes of *POD*, *FAR*, *CSI* and *AUC* of bagging CWSVM and SVM over different contamination rates are shown in Fig. 9. We can see from the figure that all the metrics except *FAR* of Bagging CWSVM are always equal to or better than the ones of SVM. The higher the contamination rate, the more significant the difference between the two methods. When the contamination rate is 0%, the bagging CWSVM can be treated as a binary supervised learning model, and its performance is at the same level as SVM because its meta-classifier is also SVM and bagging will not reduce the performance. Therefore, the bagging CWSVM can be used in any situation without worrying about whether the negative set is contaminated. Moreover, PU learning is necessary when the negative sample sets are not guaranteed to be clean.

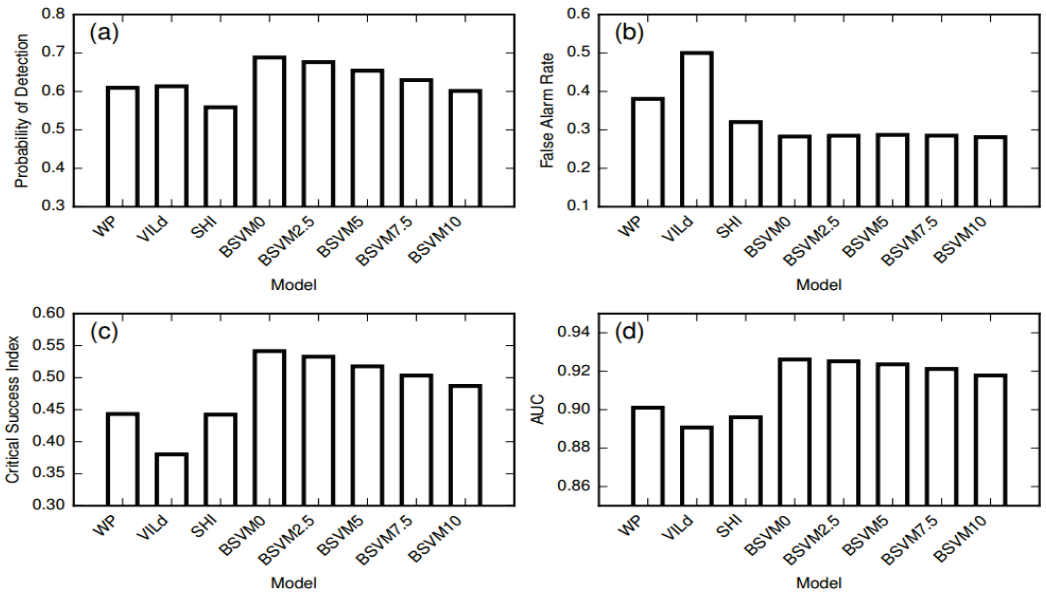

Figure 8 The (a) probability of detection; (b) false alarm rate; (c) critical success index; (d) area under ROC curve of different models. BSVM0, ..., BSVM10 refers to the bagging CWSVM models trained by the unlabeled sets with contamination rate 0%, ..., 10%

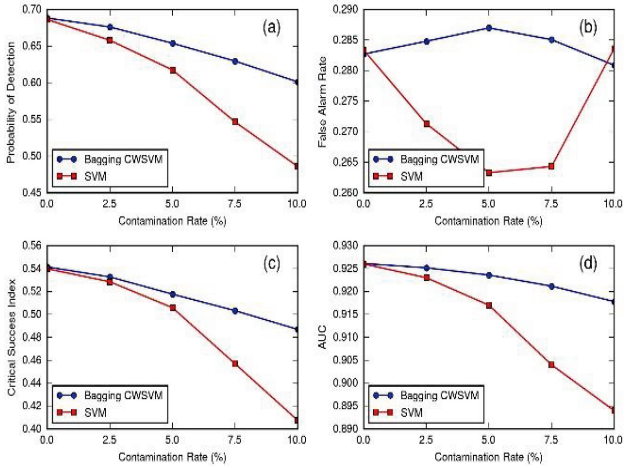

Figure 9 The change of (a) probability of detection; (b) false alarm rate; (c) critical success index; (d) area under *ROC* curve of bagging CWSVM and two-class SVM over different contamination rates

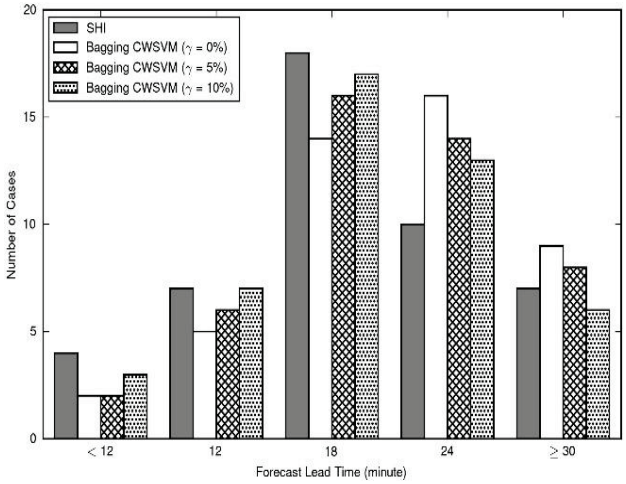

Figure 10 Forecast lead time of *SHI* and bagging CWSVM

The forecast lead time of *SHI* and Bagging CWSVM with different contamination rates for the 46 test cases is shown in Fig. 10. Since the forecast lead time of the three traditional radar parameters method has little differences, only *SHI* is used as a comparison. It can be seen that the proposed method can improve the forecast lead time to some extent, although a high contamination rate may influence the earlier forecast. When there are no contaminated samples in the unlabeled set, the bagging CWSVM method forecasts each case earlier by 6 to 12 minutes. When the contamination rate is 10%, the forecast lead time of the proposed method is at the same level as *SHI*.

4 CONCLUSION

In this paper, a radar-based hail-producing storm detection method based on positive unlabeled learning is proposed. Features used in the model are based on weather radar parameters and sounding parameters. Four radar parameters are directly input into the classifier, and the others are used after dimensionality reduction by PCA. The PU classifier model used in this study is bagging CWSVM, which iteratively trains many binary classifiers to discriminate the known positive examples from random subsamples of the unlabeled set, and averages their predictions.

Real weather radar data from three radars deployed in North China were used to evaluate the proposed method. Results show that the proposed method performs better forecast than any radar parameter method, and could improve the forecast lead time when the contamination rate in the unlabeled set is less than 10%. The comparison with SVM demonstrates that the proposed method is not inferior to supervised learning models at any time, and improvement of performance becomes more substantial when the contamination rate increases. Therefore, the proposed method is very suitable for hail-producing storm detection or other severe weather forecasting. It can significantly reduce the amount of work required for modeling and makes it possible to apply a unique model to each region.

The model can be further improved. On the one hand, in this work, we only used the radar parameters and radiosonde parameters as features. More values, such as the production from numerical weather prediction, can also be

made use of in the model. On the other hand, we did much work in data clean to make sure the positive samples are correct in this work, but hail reports are not always correct in practice. On that condition, using more robust PU learning models like the work of [51] is a better choice.

Acknowledgments

The authors would like to thank the Tianjin Meteorological Observatory for providing radar basedata and weather station data. This study is partially supported by the Natural Science Foundation of Tianjin, China under grant (14JCYBJC21800).

5 REFERENCES

[1] Kunz, M. & Puskeiler, M. (2010). High-resolution assessment of the hail hazard over complex terrain from radar and insurance data. *Meteorologische Zeitschrift*, *19*(5), 427-439. https://doi.org/10.1127/0941-2948/2010/0452

[2] Waldvogel, A., Federer, B., & Grimm, P. (1979). Criteria for the detection of hail cells. *Journal of Applied Meteorology*, *18*(12), 1521-1525. https://doi.org/10.1175/1520-0450(1979)018<1521:CFTDOH>2.0.CO;2

[3] Greene, D. R. & Clark, R. A. (1972). Vertically integrated liquid water - A new analysis tool. *Monthly Weather Review*, *100*(7), 548-552. https://doi.org/10.1175/1520-0493(1972)100<0548:VILWNA>2.3.CO;2

[4] Amburn, S. A. & Wolf, P. L. (1997). VIL density as a hail indicator. *Weather and forecasting*, *12*(3), 473-478. https://doi.org/10.1175/1520-0434(1997)012<0473:VDAAHI>2.0.CO;2

[5] Giaiotti, D., Nordio, S., & Stel, F. (2003). The climatology of hail in the plain of Friuli Venezia Giulia. *Atmospheric Research, 67*, 247-259. https://doi.org/10.1016/S0169-8095(03)00084-X

[6] Witt, A., Eilts, M. D., Stumpf, G. J., Johnson, J. T., Mitchell, E. D. W., & Thomas, K. W. (1998). An enhanced hail detection algorithm for the WSR-88D. *Weather and Forecasting*, *13*(2), 286-303. https://doi.org/10.1175/1520-0434(1998)013<0286:AEHDAF>2.0.CO;2

[7] Skripniková, K. & Řezáčová, D. (2014). Radar-based hail detection. *Atmospheric Research*, *144*, 175-185. https://doi.org/10.1016/j.atmosres.2013.06.002

[8] López, L. & Sánchez, J. L. (2009). Discriminant methods for radar detection of hail. *Atmospheric Research*, *93*(1-3), 358-368. https://doi.org/10.1016/j.atmosres.2008.09.028

[9] Stefan, S. & Barbu, N. (2018). Radar-derived parameters in hail-producing storms and the estimation of hail occurrence in Romania using a logistic regression approach. *Meteorological Applications*, *25*(4), 614-621. https://doi.org/10.1002/met.1726

[10] Mallafré, M. C., Ribas, T. R., Botija, M. D. C. L., & Sánchez, J. L. (2009). Improving hail identification in the Ebro Valley region using radar observations: Probability equations and warning thresholds. *Atmospheric Research*, *93*(1-3), 474-482. https://doi.org/10.1016/j.atmosres.2008.09.039

[11] Rigo, T. & Llasat, M. C. (2016). Forecasting hailfall using parameters for convective cells identified by radar. *Atmospheric research*, *169*, 366-376. https://doi.org/10.1016/j.atmosres.2015.10.021

[12] Besic, N., Grazioli, J., Gabella, M., Germann, U., & Berne, A. (2016). Hydrometeor classification through statistical clustering of polarimetric radar measurements: a semi-supervised approach. *Atmospheric Measurement Techniques, 9*(9). https://doi.org/10.5194/amt-9-4425-2016

[13] Roberto, N., Baldini, L., Adirosi, E., Facheris, L., Cuccoli, F., Lupidi, A., & Garzelli, A. (2017). A support vector machine hydrometeor classification algorithm for dual-polarization radar. *Atmosphere*, *8*(8), 134. https://doi.org/10.3390/atmos8080134

[14] McGovern, A., Elmore, K. L., Gagne, D. J., Haupt, S. E., Karstens, C. D., Lagerquist, R., & Williams, J. K. (2017). Using artificial intelligence to improve real-time decision-making for high-impact weather. *Bulletin of the American Meteorological Society*, *98*(10), 2073-2090. https://doi.org/10.1175/BAMS-D-16-0123.1

[15] Gagne, D. J., McGovern, A., Haupt, S. E., Sobash, R. A., Williams, J. K., & Xue, M. (2017). Storm-based probabilistic hail forecasting with machine learning applied to convection-allowing ensembles. *Weather and forecasting*, *32*(5), 1819-1840. https://doi.org/10.1175/WAF-D-17-0010.1

[16] Han, L., Sun, J., Zhang, W., Xiu, Y., Feng, H., & Lin, Y. (2017). A machine learning nowcasting method based on real-time reanalysis data. *Journal of Geophysical Research: Atmospheres*, *122*(7), 4038-4051. https://doi.org/10.1002/2016JD025783

[17] Herman, G. R. & Schumacher, R. S. (2018). Money doesn't grow on trees, but forecasts do: Forecasting extreme precipitation with random forests. *Monthly Weather Review*, *146*(5), 1571-1600. https://doi.org/10.1175/MWR-D-17-0250.1

[18] Haberlie, A. M. & Ashley, W. S. (2018). A method for identifying midlatitude mesoscale convective systems in radar mosaics. Part I: Segmentation and classification. *Journal of Applied Meteorology and Climatology*, *57*(7), 1575-1598. https://doi.org/10.1175/JAMC-D-17-0293.1

[19] Scher, S. & Messori, G. (2018). Predicting weather forecast uncertainty with machine learning. *Quarterly Journal of the Royal Meteorological Society*, *144*(717), 2830-2841. https://doi.org/10.1002/qj.3410

[20] Long, A. B., Matson, R. J., & Crow, E. L. (1980). The hailpad: Materials, data reduction and calibration. *Journal of Applied Meteorology*, *19*(11), 1300-1313. https://doi.org/10.1175/1520-0450(1980)019<1300:THMDRA>2.0.CO;2

[21] Fraile, R., Berthet, C., Dessens, J., & Sánchez, J. L. (2003). Return periods of severe hailfalls computed from hailpad data. *Atmospheric Research*, *67*, 189-202. https://doi.org/10.1016/S0169-8095(03)00051-6

[22] Dessens, J., Berthet, C., & Sanchez, J. L. (2007). A point hailfall classification based on hailpad measurements: The ANELFA scale. *Atmospheric research*, *83*(2-4), 132-139. https://doi.org/10.1016/j.atmosres.2006.02.029

[23] Sánchez, J. L., Gil-Robles, B., Dessens, J., Martin, E., Lopez, L., Marcos, J. L. et al. (2009). Characterization of hailstone size spectra in hailpad networks in France, Spain, and Argentina. *Atmospheric Research*, *93*(1-3), 641-654. https://doi.org/10.1016/j.atmosres.2008.09.033

[24] Elmore, K. L., Flamig, Z. L., Lakshmanan, V., Kaney, B. T., Farmer, V., Reeves, H. D., & Rothfusz, L. P. (2014). mPING: Crowd-sourcing weather reports for research. *Bulletin of the American Meteorological Society*, *95*(9), 1335-1342. https://doi.org/10.1175/BAMS-D-13-00014.1

[25] Xiao, Y., Li, B., & Gong, Z. (2018). Real-time identification of urban rainstorm waterlogging disasters based on Weibo big data. *Natural Hazards*, *94*(2), 833-842. https://doi.org/10.1007/s11069-018-3427-4

[26] Zhou, Z. H. (2017). A brief introduction to weakly supervised learning. *National Science Review*, *5*(1), 44-53. https://doi.org/10.1093/nsr/nwx106

[27] Fujino, A., Ueda, N., & Saito, K. (2005, July). A hybrid generative/discriminative approach to semi-supervised classifier design. *Proceedings of the National Conference on Artificial Intelligence, 20*(2), 764).

[28] Zhu, X. J. (2005). *Semi-supervised learning literature survey*. University of Wisconsin-Madison Department of Computer Sciences.
[29] Chapelle, O., Scholkopf, B., & Zien, A. (2009). Semi-supervised learning (chapelle, o. et al., eds.; 2006) [book reviews]. *IEEE Transactions on Neural Networks*, *20*(3), 542-542. https://doi.org/10.1109/TNN.2009.2015974
[30] Zhou, Z. H. (2009, June). When semi-supervised learning meets ensemble learning. In *International Workshop on Multiple Classifier Systems*, 529-538. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-02326-2_53
[31] Zhou, Z. H. & Li, M. (2010). Semi-supervised learning by disagreement. *Knowledge and Information Systems*, *24*(3), 415-439. https://doi.org/10.1007/s10115-009-0209-z
[32] Settles, B. (2009). *Active learning literature survey*. University of Wisconsin-Madison Department of Computer Sciences.
[33] Huang, S. J., Jin, R., & Zhou, Z. H. (2010). Active learning by querying informative and representative examples. In *Advances in neural information processing systems*, 892-900.
[34] Yan, Y., Rosales, R., Fung, G., & Dy, J. G. (2011, June). Active learning from crowds. *ICML, 11*, 1161-1168.
[35] Pan, S. J. & Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, *22*(10), 1345-1359. https://doi.org/10.1109/TKDE.2009.191
[36] Torrey, L. & Shavlik, J. (2010). Transfer learning. *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, 242-264. IGI Global. https://doi.org/10.4018/978-1-60566-766-9.ch011
[37] Yang, P., Liu, W., & Yang, J. (2017, August). Positive unlabeled learning via wrapper-based adaptive sampling. *IJCAI*, 3273-3279. https://doi.org/10.24963/ijcai.2017/457
[38] Nigam, K., McCallum, A., Thrun, S., & Mitchell, T. (1998). Learning to classify text from labeled and unlabeled documents. *AAAI/IAAI*, *792*, 6. https://doi.org/10.21236/ADA350490
[39] Liu, B., Lee, W. S., Yu, P. S., & Li, X. (2002, July). Partially supervised classification of text documents. *ICML, 2*, 387-394).
[40] Liu, B., Dai, Y., Li, X., Lee, W. S., & Philip, S. Y. (2003, November). Building Text Classifiers Using Positive and Unlabeled Examples. *ICDM, 3*, 179-188.
[41] Li, X. & Liu, B. (2003, August). Learning to classify texts using positive and unlabeled data. *IJCAI, 3*, 587-592.
[42] Yang, P., Li, X. L., Mei, J. P., Kwoh, C. K., & Ng, S. K. (2012). Positive-unlabeled learning for disease gene identification. *Bioinformatics*, *28*(20), 2640-2647. https://doi.org/10.1093/bioinformatics/bts504
[43] Nigam, K., McCallum, A. K., Thrun, S., & Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. *Machine learning*, *39*(2-3), 103-134. https://doi.org/10.1023/A:1007692713085
[44] Lee, W. S., & Liu, B. (2003, August). Learning with positive and unlabeled examples using weighted logistic regression. *ICML, 3*, 448-455.
[45] Elkan, C. & Noto, K. (2008, August). Learning classifiers from only positive and unlabeled data. *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 213-220. ACM. https://doi.org/10.1145/1401890.1401920
[46] Manevitz, L. M. & Yousef, M. (2001). One-class SVMs for document classification. *Journal of machine Learning research*, *2*(Dec), 139-154.
[47] Vert, R. & Vert, J. P. (2006). Consistency and convergence rates of one-class SVMs and related algorithms. *Journal of Machine Learning Research*, *7*(May), 817-854.
[48] Calvo, B., Larrañaga, P., & Lozano, J. A. (2007). Learning Bayesian classifiers from positive and unlabeled examples. *Pattern Recognition Letters*, *28*(16), 2375-2384. https://doi.org/10.1016/j.patrec.2007.08.003
[49] Geurts, P. (2011, June). Learning from positive and unlabeled examples by enforcing statistical significance. *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 305-314.
[50] Mordelet, F. & Vert, J. P. (2014). A bagging SVM to learn from positive and unlabeled examples. *Pattern Recognition Letters*, *37*, 201-209. https://doi.org/10.1016/j.patrec.2013.06.010
[51] Claesen, M., De Smet, F., Suykens, J. A., & De Moor, B. (2015). A robust ensemble approach to learn from positive and unlabeled data using SVM base models. *Neurocomputing*, *160*, 73-84. https://doi.org/10.1016/j.neucom.2014.10.081
[52] Johnson, J. T., MacKeen, P. L., Witt, A., Mitchell, E. D. W., Stumpf, G. J., Eilts, M. D., & Thomas, K. W. (1998). The storm cell identification and tracking algorithm: An enhanced WSR-88D algorithm. *Weather and forecasting*, *13*(2), 263-276. https://doi.org/10.1175/1520-0434(1998)013<0263:TSCIAT>2.0.CO;2
[53] Suzuki, S. (1985). Topological structural analysis of digitized binary images by border following. *Computer vision, graphics, and image processing*, *30*(1), 32-46. https://doi.org/10.1016/0734-189X(85)90016-7
[54] Holleman, I., Wessels, H. R. A., Onvlee, J. R. A., & Barlag, S. J. M. (2000). Development of a hail-detection-product: S10: Deep convection. *Physics and Chemistry of the Earth, Part B: Hydrology, Oceans and Atmosphere*, *25*(10-12), 1293-1297. https://doi.org/10.1016/S1464-1909(00)00197-0
[55] Delobbe, L. & Holleman, I. (2003). Radar-based hail detection: impact of height assignment errors on the measured vertical profiles of reflectivity. Preprints. *31st Conference on Radar Meteorology, Amer. Meteor. Soc., Seattle*, 475-478.
[56] Kunz, M. & Kugel, P. I. (2015). Detection of hail signatures from single-polarization C-band radar reflectivity. *Atmospheric Research*, *153*, 565-577. https://doi.org/10.1016/j.atmosres.2014.09.010
[57] Stržinar, G. & Skok, G. (2018). Comparison and optimization of radar-based hail detection algorithms in Slovenia. *Atmospheric research*, *203*, 275-285. https://doi.org/10.1016/j.atmosres.2018.01.005
[58] Murillo, E. M. & Homeyer, C. R. (2019). Severe Hail Fall and Hailstorm Detection Using Remote Sensing Observations. *Journal of Applied Meteorology and Climatology*, *58*(5), 947-970. https://doi.org/10.1175/JAMC-D-18-0247.1
[59] Belk, N. M. & Wilson, L. D. (1998). Using cell-based VIL density to identify severe-hail thunderstorms in the central Appalachians and middle Ohio Valley.
[60] Manzato, A. (2012). Hail in northeast Italy: Climatology and bivariate analysis with the sounding-derived indices. *Journal of Applied Meteorology and Climatology*, *51*(3), 449-467. https://doi.org/10.1175/JAMC-D-10-05012.1
[61] Manzato, A. (2013). Hail in northeast Italy: A neural network ensemble forecast using sounding-derived indices. *Weather and Forecasting*, *28*(1), 3-28. https://doi.org/10.1175/WAF-D-12-00034.1
[62] Wang, P. & Pan, Y. (2013). Severe hail identification model based on saliency characteristics.
[63] Gagne, D. J., McGovern, A., & Xue, M. (2014). Machine learning enhancement of storm-scale ensemble probabilistic quantitative precipitation forecasts. *Weather and Forecasting*, *29*(4), 1024-1043. https://doi.org/10.1175/WAF-D-13-00108.1
[64] Gagne II, D. J., McGovern, A., Brotzge, J., Coniglio, M., Correia Jr, J., & Xue, M. (2015, March). Day-ahead hail prediction integrating machine learning with storm-scale

numerical weather models. *Twenty-Seventh IAAI Conference*.
[65] Wang, P., Shi, J., Hou, J., & Hu, Y. (2018). The identification of hail storms in the early stage using time series analysis. *Journal of Geophysical Research: Atmospheres*, *123*(2), 929-947. https://doi.org/10.1002/2017JD027449
[66] Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning, 1*(10). New York: Springer series in statistics. https://doi.org/10.1007/978-0-387-21606-5_1
[67] Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.
[68] Hoi, C. H., Chan, C. H., Huang, K., Lyu, M. R., & King, I. (2004, July). Biased support vector machine for relevance feedback in image retrieval. *2004 IEEE International Joint Conference on Neural Networks, 4*, 3189-3194.
[69] Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123-140. https://doi.org/10.1007/BF00058655
[70] Cawley, G. C. (2006, July). Leave-one-out cross-validation based model selection criteria for weighted LS-SVMs. In *The 2006 IEEE international joint conference on neural network proceedings*, 1661-1668. https://doi.org/10.1109/IJCNN.2006.246634
[71] Daemen, A., Gevaert, O., Ojeda, F., Debucquoy, A., Suykens, J. A., Sempoux, C. et al. (2009). A kernel-based integration of genome-wide data for clinical decision support. *Genome medicine*, *1*(4), 39. https://doi.org/10.1186/gm39

Contact information:

Junzhi SHI
(Corresponding author)
School of Electrical and Information Engineering, Tianjin University,
Tianjin, 300072, China
E-mail: shijz@tju.edu.cn

Ping WANG
School of Electrical and Information Engineering, Tianjin University,
Tianjin, 300072, China

Di WANG
School of Electrical and Information Engineering, Tianjin University,
Tianjin, 300072, China

Huizhen JIA
Tianjin Bureau of Meteorology,
Tianjin 300074, China