Developing Machine Learning Tools for Long-Lead Heavy Precipitation Prediction with Multi-sensor Data

Yahui Di, Wei Ding, Yang Mu
Department of Computer Science
University of Massachusetts Boston
Boston, USA
{yahuidi, ding, yangmu}@cs.umb.edu

Ni-Bin Chang
Department of Civil, Environmental and Construction Engineering
University of Central Florida
Orlando, FL
nchang@mail.ucf.edu

David L. Small, Shafiqul Islam
Department of Civil and Environmental Engineering
Tufts University
Medford, MA
{David.Small, Shafiqul.Islam}@tufts.edu

***Abstract*— A large number of extreme floods were closely related to heavy precipitation which lasted for several days or weeks. Long-lead prediction of extreme precipitation, i.e., prediction of 6-15 days ahead of time, is important for understanding the prognostic forecasting potential of many natural disasters, such as floods. Yet, long-lead flood forecasting is a challenging task due to the cascaded uncertainty with prediction errors from measurements to modeling, which makes the current physics-based numerical simulation models extremely complex and inaccurate. In this paper, we formulate the modeling work as a machine learning problem and introduce a complementary data mining framework for heavy precipitation prediction. Heavy precipitation that may lead to extreme floods is a rare event. Long-lead prediction requires the corresponding feature space to be sampled from extremely high spatio-temporal dimensions. Such a complexity makes long-lead heavy precipitation prediction a high dimensional and imbalanced machine learning problem. In this work, we firstly define the extreme precipitation and non-extreme precipitation clusters and then design the Nearest-Sample Choosing method to handle the imbalanced data sets. We introduce streaming feature selection and subspace learning to extract the most relevant features from high dimensional data. We evaluate the machine learning tools using historical flood data collected in the State of Iowa, the United States and associated hydrometeorological variables from 1948 to 2010.**

Keywords—Heavy Precipitation Predicition, Nearest Sample Choosing, Fast Online Streaming Feature Selection, Machine Learning

I. INTRODUCTION

Many countries have suffered from catastrophic floods, like New Zealand, Australia, Pakistan, the United States and some European countries [2][4][13]. An extreme flood is often closely related to an extended period of heavy precipitation [7]. A flood happens if the volume of input water is greater than the output water in a watershed, and the excess water is not being discharged promptly at regional outfalls. A large number of extreme floods were closely related to heavy precipitation which lasted for several days or weeks. Providing an early warning of extreme floods has a significant societal impact [8].

Long-lead flood forecasting is a challenging task due to the cascaded uncertainty with prediction errors, especially in precipitation forecasting, which cannot be accurately solved by the current physics-based numerical simulation models. Heavy precipitation that may lead to extreme floods is a rare event. Long-lead prediction requires the corresponding feature space to be sampled from extremely high spatio-temporal dimensions. The study objective of this paper is to define the long-lead flood forecasting process as a machine learning process and develop a data mining framework for heavy precipitation prediction.

In our data mining framework, we distinguish the extreme precipitation cluster (EPC) and the non-extreme precipitation cluster (NEPC) based on the amount of precipitation and elapsed time. Precipitation forecasting can be defined as a machine learning classification problem—classify whether a cluster is a NEPC or EPC. If a cluster belongs to an EPC, the event raises the alarm of a potential flood. Given that the meteorological system at the global scale is a closed system, local precipitation can be linked with global hydrological cycle. For example, summer rainfalls in Iowa are mostly caused by the prevailing moist southerly flow from the Gulf of Mexico with a few lead days, which might trigger a flood [1]. With this philosophy of pattern recognition, we construct features for each cluster based on a suite of hydrometeorological parameters across uniformly sampled locations (i.e., a wealth of grid points globally) in the north hemisphere associated with 6-15 days lead time. This endeavor results in a feature space in high dimensions, which requires further classification.

In this classification problem, the heavy precipitation events are often much less frequently occurred. Most clusters are the NEPCs. Such an imbalanced data set, if not handled carefully, will result in biased results toward the majority class [15], which are the NEPCs. In this study, we propose a new method, denoted as Nearest-Sample Choosing (NSC), to build a training data set for classification by choosing the NEPCs

which are most close to the boundary with the aid of data mining algorithms. Then we apply the Online Fast Streaming Feature Selection and Discriminative Feature Selection (DFS) to handle the high dimensional problem. Fast Online Streaming Feature Selection (Fast-OSFS) had been adopted from our previous work to deal with this high dimensional problem [6]. We choose the strongly relevant and non-redundant features one by one and discard the remaining redundant and irrelevant features. We further utilize these features to extract discriminative information. DFS is a subspace learning and linear dimension reduction method. The rationale hidden behind is that the intrinsic structure of the data can be found if we transform the data from a high dimensional space to a low dimensional space [16]. In our previous study, discriminative subspace learning-based algorithms have shown great success in gait recognition [22] and crater detection [23]. Consequently, the Fast-OSFS and DFS empower us to downscale the complexity toward developing a friendly big data analytics that is tied to the NSC algorithm as a whole.

In summary, the main contributions of this paper are as follows:

- We formulate the precipitation prediction as a machine learning problem with the classification of the EPCs and NEPCs.
- We design NSC to construct a balanced training data set for handling the big data.
- We apply Fast-OSFS and subspace learning to build the most relevant features for processing the big data.

II. Relate Work

Our work is mostly related with flood prediction, imbalanced data processing, streaming feature selection, and discriminative feature selection.

Recently, scientists have designed and improved the physics-based flood forecasting models [2][25]. Ensembles of numerical weather prediction have been widely adopted for medium range (2-15 days ahead) flood forecasting with relative low accuracy [2]. A Global Forecast System model was proposed to predict the flood in in Waikato River basin of New Zealand. However, the amount of precipitation was significantly underestimated [4]. A European Flood Forecasting System was developed to determine the skill for flood forecast [9]. Pappenberger et al. [13] suggested that deep understanding of cascaded uncertainties could help improve the accuracy of the prediction. However, the high dimensionality and non-linearity in weather forecasting have collectively constrained the use of physics-based flood forecasting models. In this paper, we propose precipitation prediction as a machine learning problem from a complementary perspective in order to breakthrough this barrier.

The problem of imbalanced data often exists in many real world applications when the interesting class is rare such as floods, cancer, or droughts. Many solutions were proposed to deal with imbalanced data through imbalanced learning [18]. The up-sampling approach increases the number of minority samples through repeated sampling [19]. The down-sampling (DS) approach underestimates the majority samples [20]. Our proposed NSC algorithm is related to DS. We form a balanced training data at first by selecting NEPCs that are closest to EPCs along the boundary from infinite precipitation clusters.

Feature subset selection methods select the most relevant and non-redundant subset of features and discard irrelevant and redundant features [5]. In our previous work, Wu et al. [6] proposed a method called Fast-OSFS to process features one by one which is applicable for precipitation forecasting because of its capability in dealing with high-dimensional data. . Dimension reduction algorithms, like principal components analysis (PCA) [21] and Fisher's linear discriminant analysis (LDA) [17], have been successfully applied to many fields. PCA is only functional for the Gaussian distributed data and LDA assumes that all samples, including marginal samples, have equivalent contributions [16]. The climate data cannot be assumed to follow Gaussian distribution in most cases and each sample may have a great impact on its local neighborhoods. This requires a local learning approach such as DFS [24]. We use Discriminative Locality Alignment (DLA) [16] as our dimension feature selection method to build discriminative features.

III. Overview

Figure 1 presents an overview of our data mining framework that has the following components:

- **Precipitation Cluster Identification**: Group the historical precipitation data into contiguous clusters, and label the precipitation clusters as EPCs and NEPCs.
- **Feature Construction**: Construct the features for each cluster in spatial and temporal domains.
- **Handling the Imbalanced Data**: Use the NSC algorithm to form a balanced training set.
- **Handling the High Dimensional Data**: Apply Fast-OSFS, and use DLA to build the final feature set.
- **Building the Forecasting Model**: Use the K-Nearest Neighbor classifier to sort out mostly related features for precipitation prediction.

IV. Method

A. Identification of Precipitation Cluster

The prediction of heavy precipitation events starts with a classification effort for each sample to examine whether it belongs to an EPC or NEPC. Whereas these EPCs can be treated as positive samples, the NEPCs can be treated as negative samples in this context. With this distinction the precipitation prediction can be handled under a binary classification context.

We use two factors, including the amount of precipitation and elapsed time, to differentiate EPCs from NEPCs. Historical data are grouped into contiguous clusters. In our case study, if rain spans several days but the average precipitation of a cluster decreases in two contiguous days, then a cluster is formed and the next day is the start day of a new cluster.

Fig. 1. The proposed machine learning framework for long-lead heavy precipitation prediction.

Another case is that if the current day is a non-precipitation day but the previous day is a precipitation day, then a new cluster starts with this non-precipitation day. The detailed steps of this classification are as follows:

For a given i^{th} day, $(i-2)^{th}$, $(i-1)^{th}$, and $(i+1)^{th}$ represent the 2nd day, 1st day before and 1st day after the i^{th} day. Let us assume the precipitation in $(i-2)^{th}$, $(i-1)^{th}$, i^{th} and $(i+1)^{th}$ days are $P_{i-2}, P_{i-1}, P_i, P_{i+1}$. The P_i means that the amount of precipitation in i^{th} day. The cluster index in $(i-1)^{th}$ day is j. As a consequence, $avgP_i$ represents the average precipitation of a cluster from the start day to i^{th} day.

1) If $P_i = 0$ and if $P_{i-1} > 0$:
It means that it rained previously on day $(i-1)^{th}$ and does not rain on day i^{th}. Then i^{th} day belongs to a new cluster, and the cluster index of i^{th} is $j+1$.

2) If $P_i > 0$ and if $P_{i-1} = 0$:
It means that the i^{th} day rained while the $(i-1)^{th}$ day did not. Then the i^{th} day belongs to a precipitation cluster while the $(i-1)^{th}$ day is a non–precipitation cluster. So the cluster index of i^{th} day is $j+1$.

3) If $P_i > 0$, $(i-2)^{th}$, $(i-1)^{th}$ days belong to the same cluster j, and $avgP_i < avgP_{i-1} < avgP_{i-2}$:
It means that the average of precipitation in a cluster has continuously decreased in two days. Then the i^{th} day still belongs to cluster j, while the $(i+1)^{th}$ day belongs to $j+1$.

4) Otherwise:
The cluster index for the i^{th} day is j.

Then we label the clusters to be either an EPC or an NEPC, using a threshold below:

1) Calculate the sum of precipitation in each cluster.

2) Sort the sum of precipitations in each cluster from low to high.

3) Label a cluster as an EPC, if the sum of precipitations this cluster is high (e.g. 95% percentile of the sorted clusters), otherwise, label it as a NEPC.

B. Feature Construction

We take a suite of explanatory variables in the hydrometeorological regime into account associated with each cluster across all potential global locations over the time horizon of interest. A gridded map (see Figure 2) that may help visualize the sampling effort from the big data pool can be retrieved for a scenario of planning in this study. The hydrometeorological variables in the case study are showed in Table I. In Figure 2, the black small dots illustrate the 5,328 locations which are sampled in the north hemisphere. We sample all the hydrometeorological variables as mentioned above at these locations as a set of one-day features and construct a time horizon of 6 to 15 lead days for forecasting analysis. This 10 days' time frame will contribute the features from No. (t-15) day to No. (t-5) day to support the precipitation forecasting of the weather in No. t day.

In doing so, our first scenario ends up with building a group of 479,520 features in each possible cluster (5,328 locations ×9 variables ×10 days). Because the earth is round, as shown in Figure 2, the values of the first and last columns are the same. We thus remove the overlapping column. With this adjustment, the number of features for each possible cluster is 466,650 in our case study. It is a challenging task to directly deal with such a large feature space as a whole while all imbalanced data have to be sorted out stepwise.

C. Handling the imbalanced data

The training data set is imbalanced for the EPC analysis because the extreme precipitation events rarely occurred in a year. As a result, the total number of EPCs is small while the total number of NEPCs is pretty large. If we directly use the imbalanced data to train a classifier, most of the EPCs could be misclassified as NEPCs, while we are mostly interested in accurate classification of EPCs. In this case, the accuracy is still high because the majority samples of NEPC are still labeled correctly. However, almost all of the minority samples (EPCs) are misclassified. In order to deal with this problem, we design the NSC algorithm to form a balanced training dataset.

With this endeavor, we may reduce the number of NEPCs in training data set by choosing the NEPCs that are mostly similar to EPCs. If a new classifier can correctly separate those mostly similar negative samples from positive samples, the negative samples far away from the positive samples can be correctly classified distinctively. We choose the negative samples that have similar feature values to the positive samples to construct a balanced data set with a blanket algorithm and call this method as NSC. We also choose non-precipitation clusters to help improve the potential of a classifier to recognize those non-precipitation days.

D. Handling High Dimensional Data

We use Fast-OSFS to select the most-related features and DFS to weight the select features. The most-related features are the strongly relevant and non-redundant features [6]. The

algorithm defines steaming features as features which are no longer static but flow in one by one over time. Fast-OSFS uses a local Baysian structure to select strongly relevant and non-redundant features:

1) Online Relevance Analysis:
If a new feature Y and the targe class attribute T are conditionally dependent, add Y to BCF(the best candidate features), go back to Online Redundancy Analysis; otherwise, discard Y, and streams in a new feature.

2) Online Redundancy Analysis:
For the new feature Y included in BCF, denoting M a subset of features for BCF, if for the given M the following property $\forall Y \in F - M$ such that $P(T \mid M, Y) = P(T \mid M)$ holds, then M is a Markov blanket for T (MB(T) for short), where F is the set of original features. For any existing feature X in MB(T), if there exists a subset $S \subseteq MB(T) - X$ such that $P(T \mid X, S) = P(T \mid S)$, then X is redundant and should be excluded from MB(T).

DFS can help discover the intrinsic structure of a data set [16] by finding the features which contribute mostly with respect to a classifier [22]. In this paper, we apply DLA [9] to reduce the dimension. The generic problem of linear dimension reduction is defined as follows: given a set of samples $X = [x_1, ..., x_n]$, each sample belongs to one of the two classes (positive or negative) and is represented as an m dimension vector. The goal is to find a projection matrix W mapping $X \in R^{mxn}$ to $Y \in R^{dxn}$, i.e., $Y = W^T X$, where $d < m$. The samples in low dimension representation are $Y = [y_1, ..., y_n]$.

DLA finds the projection matrix by moving the similar samples together and separating the samples from different classes in two steps:

1) Minimize the distance of samples with same class label and maximizes the distance of samples from different classes in a small local patch by using Euclidean distance.
2) Put the small patches together to generate the global view point.

Finally, we attain the proper training features constructed for the forecasting practice.

V. EMPRICIAL STUDIES

A. Data Descrpition

The historical precipitation data in Iowa are used in our case study because the state of Iowa experienced floods which were closely related to heavy precipitation in years 2007, 2008, 2010 [10] [11] [12]. All of the hydrometeorological variables are chosen from the NCEP-NCAR Reanalysis dataset [26], which are commonly used by meteorologists for weather forecasting.

We use the spatially averaged historical precipitation of every day from 1/1/1948 to 12/30/2010 (63 years) and nine hydrometeorological variables of 63 years from 5,328 locations in the north hemisphere. These locations are spread from the North Pole and the Equator (144 longitudes and 37 latitudes) [25]. In Figure 2, the black dots illustrate the 5,328 locations. The nine hydrometeorological variables are described in Table I. These variables are deemed influential to the weather in Iowa. As a summary of this study, the final number of spatio-temporal features associated with each grid point is 466,650.

B. Data Processing

In order to evaluate our prediction scheme, two types of experiments are conducted.

- EPC prediction in the year of 2010. 2010 was a flooding year in Iowa which was caused by several heavy precipitation events [12]. Data from 2010 are treated as the testing data set and all the data from previous years form the training data set.
- EPC prediction in important historical flooding years. We want to demonstrate that the proposed method can predict precipitation under different situations in dealing with real word complexity.

For all constructed training and testing data sets, we utilize data collected from wet season (April to October) in each year, which is highly related to floods caused by precipitation [1]. However, the extreme precipitation events are so rare, which provides a highly imbalanced training data set. In this case, samples in minority class could be affected by this poorly-conditioned data base. Therefore, we design NSC to reconstruct a balanced training data set. Let the ratio of NEPCs and EPCs to be θ in the training set, and θ would be closed to 1 to ensure a balanced data set. The original θ is 21 in the first experiment setting. Then we set θ to 1.1 to construct a balanced data via the NSC mechanism. This mechanism is operated in two steps: 1) Based on the methods in section IV.A, the total number of EPCs is 339 from year 1948 to 2009; therefore, 339 NEPCs are chosen to form a balanced training data set. 2) 40 non-precipitation clusters are chosen to determine the non-precipitation days in the testing data set.

The features are then constructed by considering the spatial and temporal information simultaneously. Finally, the total number of features for each sample is 466,650. This data is too large and it is impractical to run algorithms based on this data Nine hydrometeorological Variables in the north hemisphere set directly. By considering most features are redundant set directly. By considering most features are redundant for the

Hydrometeorological Variables		Data source
PW	Precipitable Water	NCEP-NCAR Reanalysis Dataset
T850	850hPa Temperature	
Z300	300hPa Geo-Potential Height	
Z500	500hPa Geo-Potential Height	
Z1000	1000hPa Geo-Potential Height	
U300	300hPa Zonal Wind	
U850	850hPa Zonal Wind	
V300	300hPa Meridional Wind	
V850	850hPa Meridional Wind	

precipitation prediction, we apply Fast-OSFS to select related features and employ the DLA to determine appropriate sub feature space.

Finally, we use the K-Nearest Neighbor (KNN) classifier to predict the precipitation based on the test samples because the classifier is able to deal with multi-modal distributions exhibited in climate data. Accuracy and recall for the predictions are utilized to evaluate the forecasting outcome. Accuracy tells us about the number of samples that are correctly classified. Recall shows that how many positive samples we have correctly captured.

C. Results

To evaluate the efficacy of our sampling methods, we compare our NSC method with the DS [20]. Both DS and NSC reduce the number of majority samples in the training data set to form a balanced data set. In general, DS minimizes the number of majority samples with an integer factor. In this DS experiment, we use the data of year 2010 as a testing data set, and the data of previous years as a training data set. For this type of DS experiment, the number of NEPCs should be reduced to 339 to form a balanced training data set, since the number of EPCs is 339 in the training set. So, the integer factor we choose for the DS is 11. It means that for every group of 11 NEPCs, the algorithm selects one NEPC. For our comparative study, we also choose the NEPCs via using the NSC method. After running the selected data set with the aid of the NSC the result is shown in Table II. In this table, we can see that the accuracy for choosing the DS method followed by the KNN method is up to 66.49%; however, the recall is null. It means that none of the EPCs are classified correctly, and all of them are classified as NEPCs. In our NSC approach, the recall is 39.47%, which is significantly higher than its counterpart. This result confirms our assumption that the NSC method can provide a better performance because it targets those grey boundaries between EPCs and NEPCs while maintaining a balanced dataset.

Because we carry out the prediction for a 10-day period, the Fast-OSFS to select the strongly relevant and non-redundant features after constructing the balanced training data the total number of spatio-temporal features for each sample is 466,650. Most of them are redundant and weak relevant to the weather prediction in Iowa. To reduce the dimension, we use set from year 1948 to year 2009. In addition to the sampling location across the globe, Figure 2 also illustrates the 15 features we have identified after using the Fast-OSFS algorithm. All of them contribute to the prediction of the EPCs in Iowa in the year of 2010. For example, the red star one represents that the 850hPa Meridional Wind (V850) has significant effects for predicting the EPC in Iowa with a lead time of 8 days. From the figure, we also found that there is no feature shown for a lead time of the 6th day and 11th day. It is all because that the Fast-OSFS algorithm only selects the strongly relevant and non-redundant features, some of the features in these days may have significant effect but redundant for the whole selected feature set.

To show the holistic improvement of our approach, we collectively compare the original data, the balanced data with NSC, the balanced data with NSC and Fast-OSFS, and the balance data with NSC, Fast-OSFS and DLA as a whole. These data sets are all tested with the common classifer – the KNN method. The comparative results are shown in Table III. From this table, the data processed by the KNN with the aid of the NSC gain 39.47% recall that is better than the original data processed by the KNN only which has 10.53% recall. In this case, most EPCs are misclassified when using the original data processed by KNN only, which is mainly caused by the irrelevant features which disturb the classifier. With the inclusion of the Fast-OSFS algorithm, we can confirm that the recall is improved to be 47.37% since the most-relevant features can be chosen. The feature space is further tuned under the discriminative subspace learning, and the recall is improved up to 60.63%, which is significantly better than all of the rest methods. With this finding, we can conclude that recall can be increased from 10.53% to 60.63% due to the inherent power of imbalanced learning algorithms in the context of big data analytics.

TABLE I. THE RESULT OF PREDICTION BY USING DIFFERENT SAMPLING METHOD. DS+KNN MEANS THAT WE APPLY DS APPROACH TO GET THE BALANCED DATA AT FIRST, THEN RUN THEM WITH KNN. NSC+KNN MEANS THAT GET THE BALANCED DATASET BY APPLYING NSC, THEN RUN WITH KNN.

	DS+KNN	***NSC+KNN***
Accuracy	66.49%	52.80%
Recall	NaN	39.47%

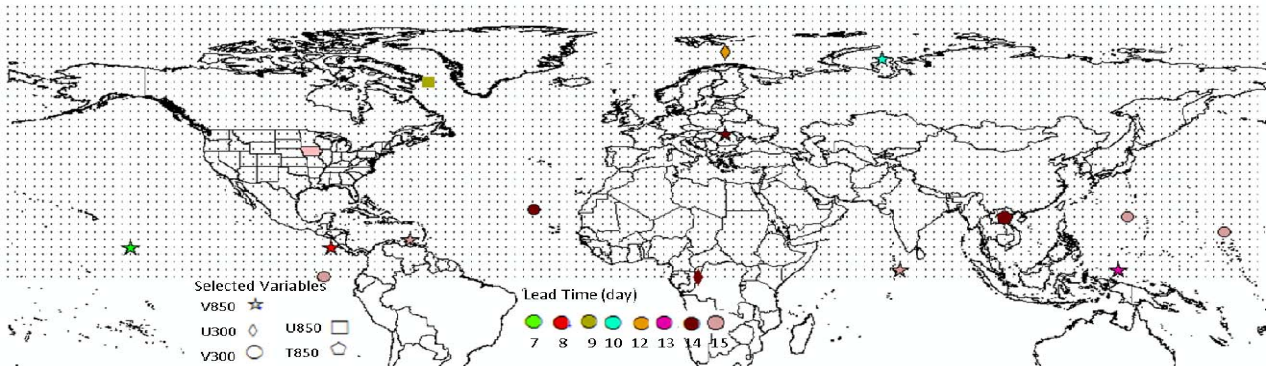

Fig. 2. The map of the features which contribute to the prediction of EPCs in Iowa (the pink rectangle in the map) 2010. The selected variables represents that the remaining most-related and non-redundant features after applying the Fsat-OSFS. The lead time means the feature has a special effect for predicting the EPC in Iowa with such a lead time.

Finally, we carry out the precipitation prediction in several word situations by using 2010, 2008, 2007 as testing data sets, and 2005~2009, 2003~2007, and 2002~2006 as training data sets, respectively. In these datasets, we first use the NSC to balance the training data, then we use the Fast-OSFS and the DLA to screen out the most related and proper features; finally, we use the KNN for performing the final prediction. In Table IV, we can find that the recalls are 78.95%, 70.77%, and 87.27% for these three years. It confirms that we can capture most of the EPCs smoothly which have significantly chances to predict the occurrence of floods with a high accuracy.

VI. CONCLUSION

In this paper, we introduce the two new concepts including the NEPCs and the EPCs. We design the NSC to form the balanced data set, and apply the Fast-OSFS and DFS to search for the most-relevant features. Such a machine learning framework enables us to analyze a mix of semi-structured and unstructured data in search of valuable information and insights for precipitation forecasting with a credible accuracy.

TABLE II. THE PREDICTION EXPERIMENT RESULTS WITH DIFFERENT METHODS. KNN MEANS THE CLASSIFIER WE USED IS KNN. NSC MEANS WE USE NSC TO FORM THE BALANCED DATA. FAST-OSFS MEANS WE USE FAST-OSFS TO SELECT USEFUL FEATURES. DLA MEANS WE USE DLA TO REDUCE THE DIMENSION.

	KNN	NSC + KNN	NSC + FAST-OSFS + KNN	NSC + FAST-OSFS + DLA +KNN
Accuracy	77.53%	52.80%	55.14%	60.28%
Recall	10.53%	39.47%	47.37%	60.53%

TABLE III. THE PREDICTION RESULTS OF DIFFERENT YEARS. ALL THE TRAINING DATA ARE FORMED AS BALANCED DATASET AND FAST-OSFS AND DLA ARE APPLIED TO GET THE RELATED FEATURES WHICH CONTRIBUTE TO THE PRECIPITATION PREDICTION, THEN KNN IS USED.

Train data (Year)	**2005-2009**	**2003 - 2007**	**2002 - 2006**
Test data (Year)	**2010**	**2008**	**2007**
Accuracy	55.61%	49.07%	52.81 %
Recall	78.95%	70.77%	87.27%

REFERENCES

[1] Climateof Iowa. Retrieved December 3, 2014, from . http://www.crh.noaa.gov/Image/dvn/downloads/Clim_IA_01.pdf.

[2] H. L. Cloke and F. Pappenberger. "Ensemble flood forecasting: a review." Journal of Hydrology 375, no. 3 (2009): 613-626.

[3] J. Thielen, K. Bogner, F. Pappenberger, M. Kalas, M. Del Medico, and A. de Roo. "Monthly‐, medium‐, and short‐range flood warning: testing the limits of predictability." Meteorological Applications 16, no. 1 (2009): 77-90.

[4] S. Dravitzki and J. McGregor. "Predictability of heavy precipitation in the waikato river basin of new zealand." Monthly Weather Review 139, no. 7 (2011): 2184-2197.

[5] R. Kohavi and G. H. John. "Wrappers for feature subset selection." Artificial intelligence 97, no. 1 (1997): 273-324.

[6] X. Wu, K. Yu, W. Ding, H. Wang, and X. Zhu. "Online feature selection with streaming features." Pattern Analysis and Machine Intelligence, IEEE Transactions on 35, no. 5 (2013): 1178-1192.

[7] J. H. Christensen and O. B. Christensen. "Climate modelling: severe summertime flooding in Europe." Nature 421, no. 6925 (2003): 805-806.

[8] T. M. Hopson and P. J. Webster. "A 1-10-day ensemble forecasting scheme for the major river basins of bangladesh: Forecasting severe floods of 2003-07*." Journal of Hydrometeorology 11, no. 3 (2010): 618-641.

[9] A. P. de Roo, B. Gouweleeuw, J. Thielen, J. Bartholmes, P. Bongioannini‐Cerlini, E. Todini, P. D. Bates et al. "Development of a European flood forecasting system." International Journal of River Basin Management 1, no. 1 (2003): 49-59.

[10] E. P. Fillmore, M. Ramirez, L. Roth, M. Robertson, C. G. Atchison, and C. Peek-Asa. "After the waters receded: A qualitative study of university official's disaster experiences during the Great Iowa Flood of 2008." Journal of community health 36, no. 2 (2011): 307-315.

[11] 2007 Midwest flooding. (2014, November 29). Retrieved December 3, 2014, from http://en.wikipedia.org/wiki/2007_Midwest_flooding

[12] M. Skopec. "Iowa floods: the 'new normal'?." Iowa Natural Heritage (2010).

[13] F. Pappenberger, K. J. Beven, N. M. Hunter, P. D. Bates, B. T. Gouweleeuw, J. Thielen, and A. P. J. De Roo. "Cascading model uncertainty from medium range weather forecasts (10 days) through a rainfall-runoff model to flood inundation predictions within the European Flood Forecasting System (EFFS)." Hydrology and Earth System Sciences Discussions 9, no. 4 (2005): 381-393.

[14] C. Cortes, and V. Vapnik. "Support-vector networks." Machine learning 20, no. 3 (1995): 273-297.

[15] Z. Zheng, X. Wu, and R. Srihari. "Feature selection for text categorization on imbalanced data." ACM SIGKDD Explorations Newsletter 6, no. 1 (2004): 80-89.

[16] T. Zhang, D. Tao, X. Li, and J. Yang. "Patch alignment for dimensionality reduction." Knowledge and Data Engineering, IEEE Transactions on 21, no. 9 (2009): 1299-1313.

[17] B. Scholkopft and K. R. Mullert. "Fisher discriminant analysis with kernels." In Proceedings of the 1999 IEEE Signal Processing Society Workshop Neural Networks for Signal Processing IX, Madison, WI, USA, pp. 23-25. 1999.

[18] S. Visa and A. Ralescu. "Issues in mining imbalanced data sets-a review paper." In Proceedings of the sixteen midwest artificial intelligence and cognitive science conference, pp. 67-73. sn, 2005.

[19] C. X. Ling, and C. Li. "Data Mining for Direct Marketing: Problems and Solutions." In KDD-98, vol. 98, pp. 73-79. 1998.

[20] M. Kubat and S. Matwin. "Addressing the curse of imbalanced data sets: One sided sampling." In Proc. of the Int'l Conf. on Machine Learning. 1997.

[21] H. Hotelling. "Analysis of a complex of statistical variables into principal components." Journal of educational psychology 24, no. 6 (1933): 417.

[22] Y. Mu and D. Tao. "Biologically inspired feature manifold for gait recognition." Neurocomputing 73, no. 4 (2010): 895-902.

[23] Y. Mu, W. Ding, D. Tao, and T. F. Stepinski. "Biologically inspired model for crater detection." In Neural Networks (IJCNN), The 2011 International Joint Conference on, pp. 2487-2494. IEEE, 2011.

[24] Y. Mu, W. Ding, and D. Tao. "Local discriminative distance metrics ensemble learning." Pattern Recognition 46, no. 8 (2013): 2337-2349.

[25] D. Wang, W. Ding, K. Yu, X. Wu, P. Chen, D. L. Small, and S. Islam. "Towards long-lead forecasting of extreme flood events: a data mining framework for precipitation cluster precursors identification." In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1285-1293. ACM, 2013.

[26] E. Kalnay, M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell et al. "The NCEP/NCAR 40-year reanalysis project." Bulletin of the American meteorological Society 77, no. 3 (1996): 437-471.