On The Enhancement of Infrared Satellite Precipitation Estimates Using Genetic Algorithm Filter-Based Feature Selection

Majid Mahrooghy[1, 2], Valentine G. Anantharaj [2], Nicolas H. Younan[1, 2], and James Aanstoos[2]
[1]Department of Electrical and Computer Engineering, Mississippi State University, Mississippi State, MS 39762, USA- (younan@ece.msstate.edu)
[2]Geosystems Research Institute, Mississippi State University, Mississippi, Mississippi State, MS 39762, USA- (Majid, val, aanstoos@gri.msstate.edu)

Abstract- a methodology to enhance a satellite infrared – based high resolution rainfall retrieval algorithm is developed by intelligently selecting features based on a filter model. Our methodology for satellite-based rainfall estimation is similar to the PERSIANN-CCS approach. However, our algorithms are enriched by applying a filter-based feature selection using generic algorithm. The objective of using feature selection is to find the optimal set of features by removing the redundant and irrelevant features. Since we use unsupervised cloud classification technique, Self Organizing Map (SOM), an unsupervised feature selection method, is used. In our approach, first the redundant features are removed by using a feature similarity-based filter and then using Entropy Index along with genetic algorithm searching, the irrelevant features are eliminated. The result shows that using feature selection process can improve Rain/No Rain detection about 10 % at some threshold values and also decreases the RMSE about 2mm.

Keywords: Satellite precipitation estimation, feature extraction, unsupervised feature selection, clustering

1. INTRODUCTION

Precipitation estimation at high spatial and temporal resolutions is beneficial for research and applications in the areas of weather, precipitation forecasting, flood and flash flood forecasting, climate, hydrology, water resources management, soil moisture, evaporation, and agriculture (Anagnostou, 2004). Notwithstanding ground-based precipitation estimates facilitate routine monitoring of rainfall across much of the continental areas of the world. The ground-based observation systems are not uniformly covered in terms of spatial and temporal resolutions. For instance, radar coverage is sparse across mountain ranges and tropical rain forests. In addition, ground-based estimates cannot provide the precipitation estimates over the oceans. Alternatively, satellite-based observation systems can provide the routine monitoring of the earth's environment such as precipitation estimation at sufficient spatial and temporal resolutions.

Many different satellite precipitation estimation (SPE) algorithms have been developed. These algorithms are mainly classified based upon the sensors and platforms they employ. Active and passive radars, visible (VIS), infrared imagery (IR), in-situ measurements, and estimates from ground-based radars have been incorporated into these algorithms; however, each of these measurements have intrinsic limitations. Active and passive microwave sensors on satellites can provide physical information about clouds, but their temporal resolution is not appropriate for high temporal applications. Infrared sensors onboard geostationary (GEO) platforms can provide high temporal observation, but their cloud top information is not always physically related to the microphysical properties of precipitation (Anagnostou, 2004). Studies show that using infrared data with radar or passive microwave calibration can provide more accurate estimations at high temporal resolution (Huffman, 2007; Turk, 2005; Joyce, 2004). The preferred Popular High Resolution Satellite Precipitation (HSPE) algorithms include 1) the Precipitation Estimation from Remotely Sensed Imagery algorithm using an Artificial Neural Network (PERSIANN) (Hsu, 1997; Sorooshian, 2000) and PERSIANN-CCS (Hong, 2004), 2) the CPC morphing technique (CMORPH) algorithm developed by the National Oceanic and Atmospheric Administration (NOAA) Climate Prediction Center (CPC) (Joyce, 2004), 3) TMPA (Tropical Rainfall Monitoring Mission (TRMM) Multisatellite Precipitation Analysis) (Huffman, 2007), and 4) the Naval Research Laboratory (NRL) blended technique (Turk, 2005) algorithm.

The algorithm and approach used in the current study are similar to those employed by the PERSIANN-CCS and the previous work of Mahrooghy et al. (2010) , which use patch cloud classification to measure the precipitation, except the current study enhances precipitation estimation using the feature selection method. Feature selection is a process that selects an optimal subset of original features based on an evaluation criterion. It reduces the feature dimension by removing irrelevant and redundant information. This improves the accuracy and increases the speed of data processing (Liu, 2005). Feature selection has been studied in supervised (Liu, 2009) and unsupervised classification (Dash, 2000; Mitra, 2002).

2. DATA

The study region covers an area of the United States extending between 30N to 38N and -95E to - 85E during January and February 2008. The training data is obtained one month before the respective testing month. The IR brightness temperature observations are obtained from the GOES-12 satellite. The National Weather Service Next Generation Weather Radar (NEXRAD) Stage IV precipitation products are used for training and validation. Also, we use the PERSIANN-CCS precipitation estimates (obtained from the PERSIANN group) for comparing the results. The IR data from GOES-12 (Channel 4) has 30-minute interval images that cover the entire area of study.

3. METHODOLOGY

Figure 1 shows a diagram of the high resolution satellite precipitation estimation using cloud classification in the training and testing modes. In the training mode, the objective is to attain the parameters, such as classification weights and the temperature-rain rate relationship curve of each cluster.

First, the raw infrared images from GOES-12 are calibrated into cloud-top brightness temperature images. The images are segmented into patches using the region growing method. Figure 2a (top) shows the cloud-top brightness temperatures from the GOES-12 on February 4, 2008 at 0615 UTC; and the corresponding cloud patches segmented is depicted in Figure 2b.

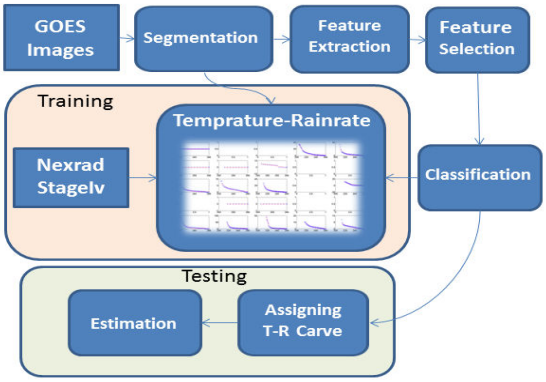

Figure 1. High resolution satellite precipitation estimation block diagram

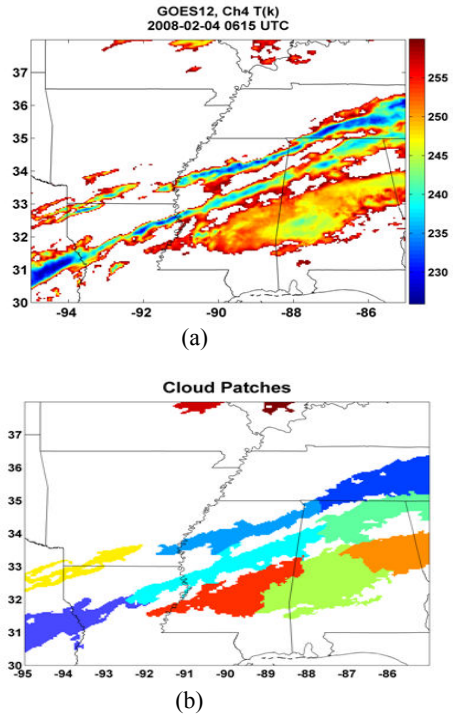

Figure 2. GOES-12 cloud-top brightness temperatures (a-top); and corresponding segmented patches of clouds (b-bottom)

The next step is feature extraction, in which the radiometric statistics (also referred to as the coldness features), geometry, and texture are extracted at the cloud patch temperature thresholds 220 K, 235 K, and 260 K. Coldness features include min and mean patch temperature at the thresholds. Texture features are the wavelet and occurrences matrix as well as the local statistic features (such as local mean and standard deviation). Geometry features are the area and shape index of each patch (Hong, 2004; Mahrooghy, 2010). Afterward, applying the feature selection to the features, the patches are classified into 100 clusters using a Self-Organizing Map (SOM) neural network. SOM projects patterns from the high dimensional space to a lower dimensional space. The projection enables the input patterns of many variables to be classified into a number of clusters and to be arranged in a two-dimensional coordinate. After clustering, a Temperature–Rain Rate (T-R) curve is assigned to each cluster. In order to obtain this T-R relationship, first T-R pixel pairs (obtained from GOES-12 observations and NEXRAD Stage IV rainfall) are redistributed by using the probability matching method (PMM) (Hong, 2004). The T-R transformation that is obtained from applying PMM are fitted by a polynomial curve fitting method (Mahrooghy, 2010).

In the testing mode, the operation is the same as in the training mode in terms of segmentation, feature selection, and feature reduction. However, in classification, the features of each patch are compared with the weights of each cluster and the most similar cluster is selected. The rain-rate estimation of the patch is computed based on the T-R curve of the cluster selected and the infrared pixel values of the patch.

3.1 Feature Selection

Most of the feature selection techniques employ a search procedure in order to generate a subset of features. However, there are some techniques that do not use search procedures (Mitra, 2002). A usual feature selection process with search criteria includes four steps: 1) subset generation, 2) subset evaluation ,3) stopping criterion, and 4) result validation (Liu, 2005). Figure 3 shows a block diagram of the process. The subset generation is a search procedure which produces candidate subsets based on a strategy. The searching process can be complete, sequential, or random. In the complete search, an exhaustive search is performed in order to find the optimal feature subset. In this type of search, it is guaranteed that the optimal feature subset is selected. The problem of this search is the complexity and long processing time. In the sequential search, not all possible feature subsets are considered and the optimal subset may be lost. The greedy search algorithms such as sequential forward selection, sequential backward elimination, and bidirectional selection are among the sequential searches. This kind of searching is fast, simple, and robust against over fitting. The random searching starts with a random selected subset and it then proceeds with a sequential search (Liu, 2005).

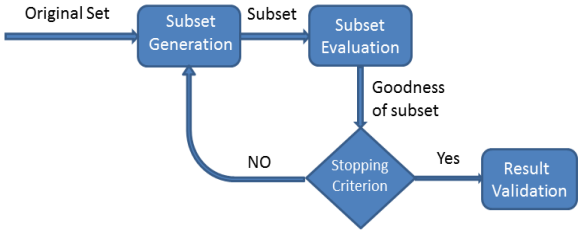

Figure 3. Block diagram of the search based feature selection

Subset evaluation is an important step in the feature selection techniques. There are two groups of evaluation criteria: independent criteria and dependent criteria. Independent criteria are usually used in filter-based feature selection. In fact, this criterion evaluates the characteristics of the training data without considering the classifier. These independent criteria can be distance, information, dependency, or consistency

measures. Dependent criteria are used in the wrapper feature selection models in which classifiers are taken into account to find the goodness of a feature subset. These criteria mostly use predictive accuracy as their primary measure. The stopping criteria decide when the algorithm of feature selection should stop. A programmed stop can be triggered by completing a search, achieving a sufficient subset, having no further addition or removal of features, or reaching a pre-specified boundary. Finally, if we have prior knowledge of the relevance or irrelevance of features, we can validate the results of the feature selection algorithm. In many cases, the prior knowledge of the relevance/irrelevance of features is not accessible.

3.2 Wrapper, Filter, and Hybrid-based feature selection

The existing feature subset algorithms can be sorted into three categories: 1) filter algorithms, 2) wrapper algorithm, or 3) Hybrid algorithms. In filter-based algorithms, one of the search strategies is selected (complete, sequential, or random search algorithm) and then the feature subset is evaluated by an independent measure. In wrapper algorithms, a classifier is exploited to evaluate the goodness of the current subset. Due to using a classifier to find an optimal feature subset, the wrapper performance is better than that of the filter-based at the cost of high computational expense. A hybrid algorithm uses the advantages of both filter and wrapper algorithms (Liu, 2005).

3.3 Genetic Algorithm (GA)

Genetic Algorithms are adaptive heuristic search algorithms based on the evolutionary ideas of natural selection and genetics. Similar to a natural system, the GA algorithm is designed to employ inheritance, mutation, selection, and crossover processes to find a candidate having the maximum fitness. GA is involved in selecting parents for reproduction, applying crossovers between the parents, and performing mutation operations on the bits representing the children. The GA randomly generates the initial population of limited size candidates and applies a selection process in which the members of the population having high fitness function survive while those having the least are eliminated (Goldberg, 1989).

3.4 Unsupervised Feature Evaluation Indices

There are different evaluation indices used in unsupervised feature selection such as Class Separability Index, Entropy Index, and Fuzzy Feature Evaluation Index (Mitra, 2002). The Entropy Index (E) can be obtained by computing the distance and similarity between two data points, p and q, as follows:

$$D_{pq} = \left[\sum_{j=1}^{M} \left(\frac{x_{pj} - x_{qj}}{\max_j - \min_j} \right)^2 \right]^{1/2} \quad (1)$$

where $\max_j$ and $\min_j$ are the maximum and minimum values along the j^{th} direction. X_{pj} and X_{qj} are feature values for p and q along the j^{th} axis, respectively, and M is the number of features. The similarity between p and q is given by

$$sim\ (p,q) = e^{-\alpha D_{pq}} \quad (2)$$

where α is equal to $-\ln 0.5 / \overline{D}$ with $\overline{D}$ being the average distance between data points. The entropy index is calculated by the following equation:

$$E = -\sum_{p=1}^{l} \sum_{q=1}^{l} (sim\ (p,q) \times \log\ sim\ (p,q) + (1 - sim(p,q)) \times \log(1 - sim(p,q))) \quad (3)$$

If the data is uniformly distributed the entropy is maximum. When the data is well-formed the uncertainty is low and the entropy is low (Dash, 2000).

3.5 Feature selection using clustering

The feature selection technique explained above uses a searching procedure to generate subset candidates. There are some other methods that are not using search operations such as feature selection using clustering. These techniques are generally used for removing redundant features (Mitra, 2002). These methods involve partitioning of the original feature set into clusters such that the features inside a cluster are highly similar while those in different clusters are disparate. Mitra et al (2002) developed an algorithm based on clustering and feature similarity to remove redundant features. In this algorithm, k nearest neighbors of each feature are computed, among them, the feature having the most compact subset is selected (the error threshold (ε) is also set to the distance of the k^{th} nearest neighbor of the feature selected) and then its k neighbors are discarded. The algorithm is repeated until all features are considered. The value k may decrease over iterations if the k^{th} nearest neighbors of the remaining features are greater than ε. In this algorithm, a similarity measure called Maximal Information Compression Index is used for clustering (Mitra, 2002). This similarity index is calculated as

$$\lambda = 0.5 \times (\mathrm{var}(x) + \mathrm{var}(y) - \sqrt{(\mathrm{var}(x) + \mathrm{var}(y))^2 - 4\,\mathrm{var}(x)\,\mathrm{var}(y)(1 - \rho(x,y)^2)}) \quad (4)$$

λ is symmetric, invariant to transform and rotate, and also sensitive to scaling of the variables. The λ value is zero when the features are linearly dependent and increases as the dependency decreases (Mitra, 2002).

4. RESULTS AND VALIDATION

Figure 4 shows the feature selection block diagram used in this study. Employing feature similarity clustering (redundant filter) based on Maximum Information Compression index (Mitra, 2002), K (= 10) redundant features are removed.

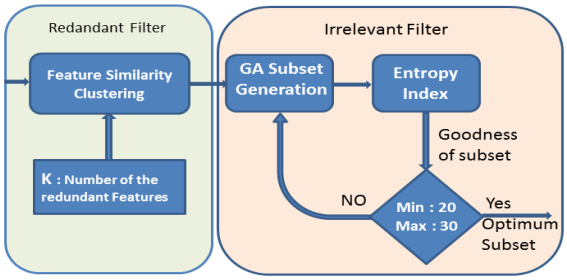

Figure 4. Feature selection used for the HSPE algorithm

The irrelevant features are also eliminated by utilizing a filter-based feature selection (irrelevant filter). The filter-based feature selection uses GA as subset generator and entropy as evaluation criteria. The optimal feature subset is obtained from the dimension between 20 and 30.

A set of 4 verification metrics, commonly used in the precipitation verification community (Ebert, 2007), are used to compare the performance of the two different algorithms,

'feature selection' and 'no feature selection'. Also the results are compared to the PERSIANN-CCS (simply referred to as CCS) product obtained from the PERSIANN group. The quantitative accuracy of the estimates is evaluated by using the Root Mean Squared Error (RMSE). The performance of rain/no-rain detection is evaluated by the Probability of Detection (POD), the False-Alarm Ratio (FAR), and the Heidke Skill Score (HSS).

The daily estimates from January and February 2008 are computed for both algorithms. The results of the evaluation metrics are depicted in Figure 5. As seen from this figure, feature selection can improve the FAR, especially at higher thresholds in which the HSS is improved about 10%. Generally speaking, using feature selection mends the HSS at all thresholds, if compared to the 'no feature selection' algorithm. In addition, using feature selection also decreases the RMSE by approximately 2 mm.

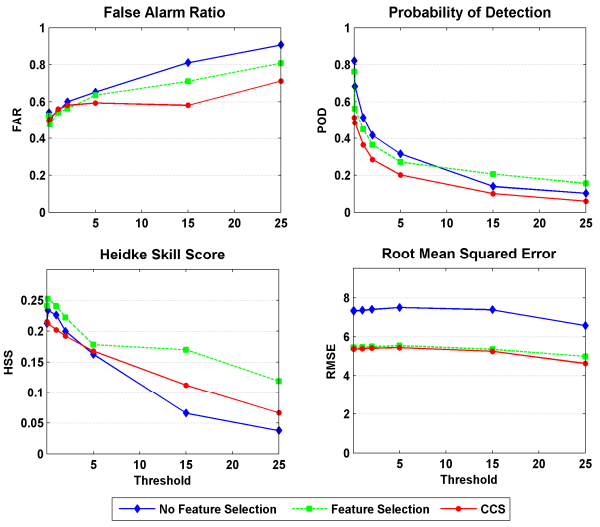

Figure 5. Validation results for Jan to Feb (winter) 2008 (daily estimate): (a) False Alarm Ratio; (b) Probability Of Detection; (c) Heidke Skill Score; (d) Root Mean Squared Error

5. CONCLUSION

A high resolution precipitation estimation algorithm based on cloud patch classification is developed using a feature selection method in which the redundant and irrelevant features are removed. The feature similarity clustering method using the Maximum Information Compression Index is used in order to eliminate the redundant features. In addition, a filter-based feature selection is utilized to find the optimal feature subset. The result shows that using feature selection improves the rain/no rain detection by about 10 % at some thresholds. It also decreases the RMSE by approximately 2 mm for all thresholds.

REFERENCES

E. N. Anagnostou, "Overview of overland satellite rainfall estimation for hydro-meteorological applications," *Surveys in Geophysics*, vol.25:5-6, pp. 511-537, 2004.

M. Dash and H. Liu, "Unsupervised feature selection," *Proc. Pacific Asia Conf. Knowledge Discovery and Data Mining*, pp. 110-121, 2000.

E.E. Ebert, J.E. Janowiak, and C. Kidd. "Comparison of near-real-time precipitation estimates from satellite observations and numerical models," *Bull. Amer. Meteor. Soc*., pp. 47-64, 2007.

D. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning, Readin*g, MA: Addison-Wesley, 1989.

Y. Hong, K. L. Hsu, S. Sorooshian, and X. G. Gao, "Precipitation estimation from remotely sensed imagery using an artificial neural network cloud classification system," *Journal of Applied Meteorology*, vol. 43, pp. 1834- 1852, 2004.7-503, 2004.

K. L. Hsu, X. G. Gao, S. Sorooshian, and H. V. Gupta, "Precipitation estimation from remotely sensed information using artificial neural networks," *Journal of Applied Meteorology*, vol. 36, pp. 1176-1190, 1997.

G. J. Huffman, R. F. Adler, D. T. Bolvin, G. J. Gu, E. J. Nelkin, K. P. Bowman, Y. Hong, E. F.Stocker, and D. B. Wolff, "The TRMM multisatellite precipitation analysis (TMPA): Quasi-global, multiyear, combined-sensor precipitation estimates at fine scales," *Journal of Hydrometeorology*, vol. 8, pp. 38-55, 2007.

R. J. Joyce, J. E. Janowiak, P. A. Arkin, and P. Xie, " CMORPH: A method that produces global precipitation estimates from passive microwave and infrared data at high spatial and temporal resolution," *Journal of Hydrometeorology*, vol. 5, pp. 48. 2004.

H. Liu, L. Yu, "Toward integrating feature selection algorithms for classification and clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol.17, no.4, pp. 491- 502, April 2005.

H. Liu, J. Sun, L. Liu, H. Zhang, "Feature selection with dynamic mutual information," *Pattern Recognition*, vol. 42, Issue 7, pp. 1330-1339, July 2009.

Majid Mahrooghy, Valentine G Anantharaj, Nicolas H. Younan, Walter A. Petersen, F. Joseph Turk, and James Aanstoos, James, "Infrared satellite precipitation estimate using wavelet-based cloud classification and radar calibration," *Proceedings of the IEEE Geoscience and Remote Sensing Symposium*, pp. 2345-2348, July 2010.

P. Mitra, C. A. Murthy, S. K. Pal, "Unsupervised feature selection using feature similarity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.24, no.3, pp. 301-312, Mar 2002.

S., Sorooshian, K. L. Hsu, X. Gao , H. V. Gupta , B. Imam, and D. Braithwaite, "Evaluation of PERSIANN system satellite based estimates of tropical rainfall," *Bull. Amer. Meteorol. Soc.,* 81, page 2035, 2000.

F. J. Turk and S. D. Miller, "Toward improved characterization of remotely sensed precipitation regimes with MODIS/AMSR-E blended data techniques," *IEEE Transactions on Geoscience and Remote Sensing,* vol. 43, pp. 1059–1069, 2005.