Comparison of Nearest Neighbor (ibk), Regression by Discretization and Isotonic Regression Classification Algorithms for Precipitation Classes Prediction

Solomon Mwanjele Mwagha
Taita Taveta University College
P.O.Box 308-80300
Voi Kenya

Masinde Muthoni
Central University of Technology
Bloemfontein
South Africa

Peter Ochieg
Taita Taveta University College
P.O.Box 635-80300

ABSTRACT
Selection of classifier for use in prediction is a challenge. To select the best classifier comparisons can be made on various aspects of the classifiers. The key objective of this paper was to compare performance of nearest neighbor (ibk), regression by discretization and isotonic regression classifiers for predicting predefined precipitation classes over Voi, Kenya. We sought to train, test and evaluate the performance of nearest neighbor (ibk), regression by discretization and isotonic regression classification algorithms in predicting precipitation classes. A period of 1979 to 2008 daily Kenya Meteorological Department historical dataset on minimum/maximum temperatures and precipitations for Voi station was obtained. Knowledge discovery and data mining method was applied. A preprocessing module was designed to produce training and testing sets for use with classifiers. Isotonic Regression, K-nearest neighbours classifier, and RegressionByDiscretization classifiers were used for training training and testing of the data sets. The error of the predicted values, root relative squared error and the time taken to train/build each classifier model were computed. Each classifier predicted output classes 12 months in advance. Classifiers performances were compared in terms of error of the predicted values, root relative squared error and the time taken to train/build each classifier model. The predicted output classes were also compared to actual year classes. Classifier performances to actual precipitation classes were compared. The study revealed that the nearest neighbor classifier is a suitable for training rainfall data for precipitation classes prediction.

General Terms
Clasification Algorithms, Data Mining, Knowledge Discovery

Keywords
Regression by discretization, isotonic regression, nearest neighbor(ibk), precipitation prediction, classification algorithms, classifier performance

1. INTRODUCTION
There has been a lot of research aimed at precipitation predictions over selected regions where solutions were based on traditional, statistical and modern computational methods or a combination. Though these precipitation predictions were useful in overall region rainfall picture depiction, challenges exist for the prediction of quantity classes of precipitation for fixed range durations for instance weekly or monthly quantities in a year. Successful prediction of precipitation in fixed range durations can aid in selection of activities during rainy seasons such as cropping where different crops require different water requirements, or selection of a grazing land for nomadic pastoralists for a particular duration of time.

Classification algorithms continue to play a big role in prediction of events based on historical data. In order to predict precipitation classes in advance algorithm performances must compared and the best one selected.

2. LITERATURE REVIEW
A study on drought forecasting [2] analyzed rainfall frequencies using data from 248 rain gauges (1938-2005). SPI was determined using ANN feed feedfoward and back propagation algorithm. The findings showed that the result of ANN is suitable for drought forecast. Another study aimed at comparing ANN and ANFIS in precipitation prediction [2] realized ANN efficient in rain prediction. predicting agricultural drought [12] was done using 1880-2005 rain data to analyze agricultural drought. By applying fuzzy sets analysis on the condition of crops and valid rain history, result of fuzzy clustering obtained. Drought years extracted from fuzzy clustering results. Time series used to predict next drought year. A study in [8] was aimed at translating seasonal forecast to agricultural terms using crop simulation model to translate seasonal forecast to agricultural terms. The results offered support to farmer's climate risk management. In China rainfall was predicted by direct determination of surface soil moisture using microwave observation [11] where data was acquired and analyzed over several test sites. The study was validated by conducted large field experiments. Agricultural drought was predicted in paddy fields using remotely sensed data [7] where NDVI was found to be reasonable in detecting agricultural drought. The study was limited by insufficient data as fuzzy was done in non cropping time. Metrological conditions causing drought were evaluated using the differentiate between precipitation & evapotranspiration to evaluate metrological conditions causing drought [4]. In USA historical patterns for drought were identified using VegOut Model that integrated Climate Ocean, satellite indicators[10]; regression trees were used to identify historical patterns for drought intensely and vegetation. SPI and PDSI were used to represent climate vulnerability. This study was evaluated using 2006 drought year. Unlike previous studies this paper contributes on prior work by considering crop production history and weather data history together with classification algorithms to come up with precipitation classes. Our work provides future classes projections with a limit of twelve months in advance predictions. By borrowing from previous studies this research

emphasis is on comparison of classification algorithms in rainfall prediction in order to select the best.

3. METHODOLOGY

A period of 1979 to 2008 daily KMD historical dataset on minimum/maximum temperatures and precipitations for Voi KMD station was obtained. Next the knowledge Discovery and Data mining (KDD) process steps were applied. A preprocessing module was designed to produce training and testing sets of files for use with the classifiers. Three classifiers (isotonic regression, k-nearest neighbours classifier, and regression by discretization) were used for training training and testing of the data sets. A knowledge flow was implemented for each of the three classifiers. The Waikato Environment for Knowledge Analysis (WEKA) and Java programming environment (JCreator and Net Beans) were used.

4. RESULTS

A preprocessing module was designed to produce two sets of files for use with the Weka Knowledge flow each with five attributes namely:-

Year,

Month,

Scaled precipitation values (range: 0 to 1),

Precipitation class values,

Index class (range: -2 to 2).

Three classifiers (Isotonic Regression, K-nearest neighbours classifier, and RegressionByDiscretization) were considered for training training and testing of the data sets. The classifiers produced output classes with the following attributes:-

Precipitation class values with removal filtered applied,

Index class identified as the class variable.

The output of each of the classifiers is as follows:-

Isotonic Regression

This algorithm learns an isotonic regression model to pick the attribute that result in the lowest squared error. It does not allow missing values and can only deal with numeric attributes. It considers the monotonically increasing case as well as the monotonically decreasing case.

The running information on using this classifier is as follows:-

Scheme: weka.classifiers.functions.IsotonicRegression

Relation: training-weka.filters.unsupervised.attribute.Remove-R4

Instances: 372

Attributes: 4

year

month

scaled_prec

index_class

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

Isotonic regression

Based on attribute: scaled_prec

prediction:	-2	cut point:	0
prediction:	-1.5	cut point:	0.08
prediction:	-1	cut point:	0.11
prediction:	-0.5	cut point:	0.19
prediction:	0	cut point:	0.37
prediction:	0.5	cut point:	0.46
prediction:	1	cut point:	0.56
prediction:	1.5	cut point:	0.65
prediction:	2		

Time taken to build model: 0.03 seconds

=== Cross-validation ===

=== Summary ===

Correlation coefficient	0.9993	
Class complexity | order 0 2.5057 bits/instance	932.1235	bits
Class complexity | scheme 0.0657 bits/instance	24.4368	bits
Complexity improvement (Sf) 2.44 bits/instance	907.6867	bits
Mean absolute error	0.0027	
Root mean squared error	0.0367	
Relative absolute error	0.328 %	
Root relative squared error	3.7127 %	
Total Number of Instances	372	

The error of the predicted values for numeric classes is 0.0027

The root relative squared error is 3.7127%

K-nearest neighbor classifier

K-nearest neighbour classifier can select appropriate value of K based on cross-validation. It can also do distance weighting using a simple distance measure to find the training instance closest to the given test instance, and predicts the same class as this training instance. If multiple instances are the same (smallest) distance to the test instance, the first one found is used.

The running information for this classifier is as follows:-

Scheme: weka.classifiers.lazy.IBk -K 1 -W 0 -X -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R 3,4\" -P"

Relation: training-weka.filters.unsupervised.attribute.Remove-R4

Instances: 372

Attributes: 4

year

month

scaled_prec

index_class

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

IB1 instance-based classifier

Using 1 nearest neighbour(s) for classification

Time taken to build model: 0 seconds

=== Cross-validation ===

=== Summary ===

Correlation coefficient	0.9993	
Class complexity | order 0 2.5178 bits/instance	936.637	bits
Class complexity | scheme 0.0618 bits/instance	22.9989	bits
Complexity improvement (Sf) 2.456 bits/instance	913.6381	bits
Mean absolute error	0.0027	
Root mean squared error	0.0367	
Relative absolute error	0.3275 %	
Root relative squared error	3.7048 %	
Total Number of Instances	372	

The error of the predicted values for numeric classes is 0.0.0027

The root relative squared error is 3.7048%

Regression By Discretization

Regression by discretization is a scheme that employs any classifier on a copy of the data that has the class attribute (equal-width) discretized. The predicted value is the expected value of the mean class value for each discretized interval (based on the predicted probabilities for each interval).

The base classifier used is J48 Class for generating a pruned or unpruned C4.5 decision trees.

The output of this classifier is as follows:-

Scheme: weka.classifiers.meta.RegressionByDiscretization -B 10 -W weka.classifiers.trees.J48 -- -C 0.25 -M 2

Confidence factor for pruning is 0.25, to use binary splits and restrict the minimum number of instances in a leaf to 2 (grow the tree fully).

Relation: training-weka.filters.unsupervised.attribute.Remove-R4

The name of the relation contains in it the name of data file used to build it, and the names of filters that removes the fourth attribute

Instances: 372

Attributes: 4

year

month

scaled_prec

index_class

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

Regression by discretization

Class attribute discretized into 10 values

Classifier spec: weka.classifiers.trees.J48 -C 0.25 -M 2

J48 pruned tree

---

scaled_prec <= 0.073766

| scaled_prec <= 0: '(-inf--1.6]' (45.0)

| scaled_prec > 0: '(-1.6--1.2]' (177.0)

scaled_prec > 0.073766

| scaled_prec <= 0.186199

| | scaled_prec <= 0.107971: '(-1.2--0.8]' (19.0)

| | scaled_prec > 0.107971: '(-0.8--0.4]' (40.0)

| scaled_prec > 0.186199

| | scaled_prec <= 0.364366: '(-0.4-0]' (47.0)

| | scaled_prec > 0.364366

| | | scaled_prec <= 0.555324

| | | | scaled_prec <= 0.447353: '(0.4-0.8]' (12.0)

| | | | scaled_prec > 0.447353: '(0.8-1.2]' (12.0)

| | | scaled_prec > 0.555324

| | | | scaled_prec <= 0.650506: '(1.2-1.6]' (14.0)

| | | | scaled_prec > 0.650506: '(1.6-inf)' (6.0)

Number of Leaves: 9

Size of the tree: 17

Time taken to build model: 0.05 seconds

=== Cross-validation ===

=== Summary ===

Correlation coefficient	0.9974	
Class complexity | order 0 2.5178 bits/instance	936.637	bits
Class complexity | scheme 0.1931 bits/instance	71.8388	bits
Complexity improvement (Sf) 2.3247 bits/instance	864.7982	bits
Mean absolute error	0.0108	
Root mean squared error	0.0733	
Relative absolute error	1.3101 %	
Root relative squared error	7.4096 %	
Total Number of Instances	372	

The error of the predicted values for numeric classes is 0.0108

The root relative squared error is 7.4096%

Evaluation

Comparison of classifiers 2009 predictions to 2009 actual precipitation classes

Month	Actual 2009	2009 Class predicted by classifier					
		Ibk		Isotonic			
		5 Step	10 Step	5 Step			
Jan	-0.5	2	-1	2			
Feb	-1.5	-1	-1.5	-0.5			
Mar	-1.5	-0.5	0	-0.5			
Apr	0.5	0	2	0.5			
May	-2	-1.5	-1.5	-2	-1.5	-2	-1.5
Jun	-1.5	-1.5	-1.5	-1.5	-1.5	-1.5	-1.5
Jul	-1.5	-2	-2	-1.5	-1.5	-1.5	-1.5
Aug	-1.5	-1.5	-1.5	-2	-1.5	-2	-1.5
Sep	-2	-1.5	-1.5	-2	-1.5	-2	-1.5
Oct	0	0	-1.5	-0.5	-2	0.5	-2
Nov	0	0	1.5	0	2	0	2
Dec	1.5	0.5	1.5	0.5	2	0.5	2
Error of the predicted values		0.0027		0.0027	0.0108		
Root relative squared error		3.7048%		3.7127%	7.4096%		
Time taken to train/build model		0 seconds		0.03 seconds	0.05 seconds		

Comparison of output classes from classifiers

Graph of classifiers 2009 predictions outputs compared to 2009 actual

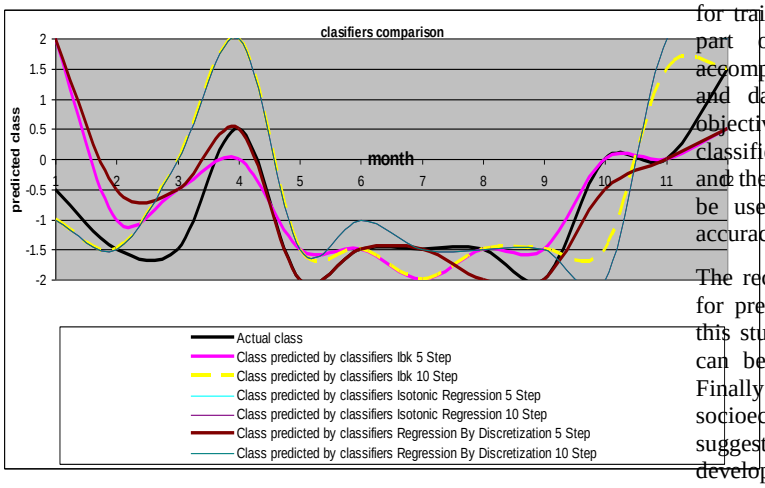

Figure 49: graph comparing classifiers outputs on 5 year and 10 year sample step with actual

In figure 55 below the predicted classes are compared to the actual computed for 2009. The predictions correspond to most of the monthly actual classes. The predictions were based on IBk classifier with 10 year step sampling.

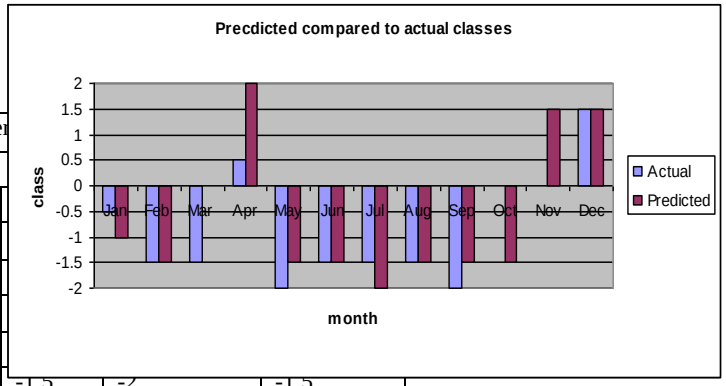

Figure 55: ibk classifier predictions for 2009 compared to actual

Source: study classifiers results analysis

The classifier results in prediction showed that the IBk (k-nearest neighbor) classifier was the best when applied in the weka knowledge flow model for training and prediction. Apart from generating the best results compared to other classifiers on performing 10 fold cross validation the IBk classifier took the least amount of time to build/train the model. The comparison of precipitation predictions for 2009 with the actual figures for 2009 results indicated that the predicted and actual results were comparable.

5. CONCLUSION

The precipitation class prediction output results obtained showed that the nearest neighbor classifier is a suitable tool for training meteorological data for precipitation classes. As part of machine learning the IBk classifier results accomplished intelligence through the knowledge discovery and data mining process as aimed in the study major objective. Evaluation of our study results shows that the Ibk classifier had the least error of the predicted values (0.0027) and the least root relative squared error (3.7048%) hence can be used to predict precipitation in advance with greater accuracy compared to the other two classifiers.

The recommendation for designing a solution that can cater for precipitation predictions in multiple regions is open for this study as future work. In Kenya for instance all districts can be represented and in precipitation classes predicted. Finally with adjustments of our prediction predictions socioeconomic measures on droughts anticipations can be suggested e.g. early warning allowing systems can be developed.

6. REFERENCES

[1] Ashok, M. et al, 2006. Linking Seasonal Climate Forecasts with Crop Simulation to Optimize Maize Management, CCSP Workshop: Climate Science in Support of Decision Making, 14-16 November 2005 Crystal Gateway Marriott Arlington, Virginia 14-16 November 2005

[2] Dostrani, M. et al (2010). Application of ANN and ANFIS Models on Dryland Precipitation Prediction (Case Study: Yazd in Central Iran). Journal of Applied Sciences, 10: 2387-2394.

[3] Gong, Z. et al, 2010. Risk Prediction of Agricultural Drought in China. 2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2010).Kenya Meteorological Department, Agrometeorological bulletin, Issue No. 27/2009.

[4] Kozyra, J. et al 2009. Institute Of Soil Science and Plant Cultivation National Research Institute, International Symposium, Climate change and Adaptation Options in Agriculture, Viena, June, 22-23 2009.

[5] Ladislaus B. et al, 2010. Indigenous knowledge in seasonal rainfall prediction in Tanzania: A case of theSouth-western Highland of Tanzania, Journal of Geography and Regional Planning Vol. 3(4), pp. 66-72,April 2010.

[6] Lin Zhu, Jing M. Chen, Qiming Qin, Mei Huang, Lianxi Wang, Jianping Li, Bao Cao: Assimilating Remote Sensing based Soil Moisture in an Ecosystem Model (BEPS) for Agricultural Drought Assessment.IGARSS (5) 2008: 437-440

[7] Niu Shulian; Susaki Junichi, 2006. Detection of Agricultural Drought in Paddy Fields Using NDVI from MODIS Data. A Case Study in Burirum Province, Thailand.

[8] Patrick O, 2006. Agricultural Policy in Kenya: Issues and Processes, A paper for the Future Agricultures Consortium workshop, Institute of Development Studies, 20-22 March 2006.

[9] Peter Reutemann, (2007). WEKA Knowledge Flow Tutorial for Version 3-5-7, University of Waikato 2007

[10] Tsegaye T. & Brian W. 2007. The Vegetation Outlook (VegOut): A New Tool for Providing Outlooks of General Vegetation Conditions Using Data Mining Techniques. ICDM Workshops 2007: 667-672

[11] Z. (Bob) Su, Y. Chen, M. Menenti, J. Sobrino, Z.-L. Li, W. Verhoef, L. Wang, Y. Ma, L. Wan, Y. He, Q.H.

[12] Liu, C. Li, J. WEN, R. van der Velde, M. van Helvoirt, W. Lin, X. Shan, 2007. Drought Monitoring and Prediction over China, In: Proceedings of the 2008 Dragon symposium, Dragon programme, final results, 2004-2007, Beijing, China 21-25 April 2008.