Seasonal to Inter-annual Climate Prediction Using Data Mining KNN Technique

Zahoor Jan[1], M. Abrar[2], Shariq Bashir[3], and Anwar M. Mirza[1]

[1] FAST-National University of Computer and Emerging Sciences, A. K. Brohi Road, H-11/4, Islamabad, Pakistan
zahoor_jan2003@yahoo.com, anwar.m.mirza@nu.edu.pk
[2] NWFP Agricultural University Peshawar, Pakistan
gulabson2@gmail.com
[3] Vienna University of Technology, Austria
shariqadel@yahoo.com

Abstract. The impact of seasonal to inter-annual climate prediction on society, business, agriculture and almost all aspects of human life, force the scientist to give proper attention to the matter. The last few years show tremendous achievements in this field. All systems and techniques developed so far, use the Sea Surface Temperature (SST) as the main factor, among other seasonal climatic attributes. Statistical and mathematical models are then used for further climate predictions. In this paper, we develop a system that uses the historical weather data of a region (rain, wind speed, dew point, temperature, etc.), and apply the data-mining algorithm "**K-Nearest Neighbor (KNN)**" for classification of these historical data into a specific time span. The k nearest time spans (k nearest neighbors) are then taken to predict the weather. Our experiments show that the system generates accurate results within reasonable time for months in advance.

Keywords: climate prediction, weather prediction, data mining, k-Nearest Neighbor (KNN).

1 Introduction

Seasonal to inter annual (S2I) climate prediction is the recent development of meteorology with the collaboration of oceanography and climatology all over the world. Weather and climate affects human society in all dimensions. In agriculture, it increases or decreases crop production [1], [2]. In water management [3] rain, the most important factor for water resources, an element of weather. Energy sources, e.g. natural gas and electricity are greatly depends on weather conditions. The day-to-day weather prediction is used for decades to forecast few days in advance, but recent developments move the trend from few days to inter annual forecast [4]. The S2I forecast is to forecast climate from months to year in advance. Climate is changing from year to year , e.g. rain/ dry, cold/warm seasons significantly influence society as well as economy. Technological improvements increase the understanding in meteorology that how the different cycles, ENSO (El Niño Southern Oscillation, i.e. the warm and cold and vice versa phenomena of ocean) over the Pacific Ocean and Sea Surface Temperature (SST), affects the

D.M.A. Hussain et al. (Eds.): IMTIC 2008, CCIS 20, pp. 40–51, 2008.

climate of regions world widely. Many countries United State of America (National Ocean and Atmospheric Administration – NOAA), England (MetOffice – Meteorology Department and London Weather Center), Sri Lanka (Department of Meteorology, Sri Lanka), India (India Meteorology Department and National Center for Medium Range Weather Forecasting – NCMRWF), and Bangladesh (Bangladesh Meteorological Department) etc have started the utilization of Seasonal Climate forecast Systems. In 1995, the International Research Institute for climate prediction (IRI) was formed by a group of scientists to overcome the shortcomings of previous work. The main goal and purpose of this institute was to improve performance of existing systems, develop new accurate and sophisticated systems for S2I. The IRI with the collaboration of other international institute for climate prediction developed new models for forecasting. IRI developed dynamical models based on the Atmosphere General Circulation Model (AGCM), e.g. ECHAM3 (Max Planck Institute), MRF9 (National Center for Environmental Prediction – NECP), CCA3 (National Center for Atmospheric Research – NCAR). Other models are Canonical Correlation Analysis (CCA) [5], [6], Nonlinear Canonical Correlation Analysis (NCCA) [7], Tropical Atmosphere Ocean Array TAOA [8], Global Forecast System, Climate Forecast Model[1] are statistical, Numerical and dynamical (Two-tiered). These models use SST as main attribute for forecasting among other climatic attributes. The sources of these attributes [9] are ENSO teleconnections (effects the global climate), Indian and Atlantic Ocean (effect the regional climate). None of these models is accurate for all situations and regions. These systems also use the geographical (longitude and latitude) location to identify the different regions instead of specific country or city, e.g. The Tropical region (the area between 23.5° N and 23.5° S along the equator).

The main purpose of this paper is how to use a data mining technique, "K-Nearest Neighbor (KNN)", how to develop a system that uses numeric historical data (instead of geographical Location) to forecast the climate of a specific region, city or country months in advance.

Data Mining is the recent development in the field of VLDB/ Data warehousing. It is used to discover the hidden interesting patterns in huge databases that are impossible otherwise. Data Mining can be classified by its techniques into three basic types, i.e. Association Rules Mining, Cluster analysis and Classification/Prediction [10]. KNN is classification algorithm that is based on Euclidean distance formula, which is used to find out the closeness between unknown samples with the known classes. The unknown sample is then mapped to the most common class in its k nearest neighbors. Rest of the paper is organized as follows, section two narrate the related work, section three describes the general aspect of KNN algorithm section 4 shows the KNN application in forecast system, section 5 is about the experiment and result, section 6 concludes the work and for the future extension.

2 Related Work

A number of tools are available for climate Prediction. All the initial efforts use statistical models. Most of these techniques predict the SST based on ENSO phenomena

[1] GFS, CFS and large collection of regional and global climate forecasting system are the developed and run by NOAA. http://www.noaa.gov, http://www.nco.ncep.noaa.gov/pmb

[6]. NINO3 uses simple average for a specific latitude and longitude (5°S-5°N; 150°E-90°W), etc. Canonical correlation analysis [15, 16] is another statistical model that takes data from different oceans, i.e. Indian, Atlantic, Pacific etc) and forecast the SST monthly anomalies (notable changes from routine measurement, very high or very low SST).

Data Mining is recently applied that how the climate effects variation in vegetation [17]. The *Regression Tree* technique is used to find this relation and predicts the future effects of climate on vegetation variability. The Independent Component Analysis is incorporated in Data Mining [18] to find the independent component match in spatio-temporal data specifically for North Atlantic Oscillation (NAO). The Neural Networks using nonlinear canonical correlation analysis [7] are used to find the relationship between Sea Level Pressure (SLP) and Sea Surface Temperature (SST) that how SST is effected by SLP and changes the climate of specific regions.

3 K Nearest Neighbor (KNN)

KNN [11] is widely and extensively used for supervised classification, estimation and prediction [10], [12]. It classify the unknown sample **s** to a predefine class $c_i \in C$, $1 < i \leq n$, based on previously classified samples (training data). It is called *Lazy Learner* because it performs the learning process at the time when new sample is to be classified, instead of its counterpart *Eager Learner*, which pre-classifies the training data before the new sample is to be classified. The KNN therefore requires more computation than eager learner techniques. The KNN is however beneficial to dynamic data, the data that changes/updates rapidly.

When new sample **s** is to be classified, the KNN measures its distance with all samples in training data. The *Euclidean distance* is the most common technique for distance measurement. All the distance values are then arranged such that $d_i \leq d_{i+1}$, $i = 1,2,3,\ldots n$. The k samples with the smallest distance to the new sample are known k-nearest neighbors and are used to classify the new sample **s** to the existing class $c_i \in C$, $1 < i \leq n$. The decision of classification depends on the nature of the data. If the data is of categorical nature then simple voting or weighted voting is used for classification. In case of continuous/quantitative data, the average, median or geometric mean is used. The new sample **s** is then classified to $c_i \in C$, $1 < i \leq n$. The process is simplified below.

KNN Algorithms:

Step 1: Measure the distance between the new sample s and training data.

$$\left(\textit{Euclidean Distance}, D(x_s, y_s) = \sqrt{\sum_{i=0}^{n} (x_i - y_i)^2} \right)$$

Step 2: Sort the distance values as $d_i \leq d_{i+1}$, select k smallest samples

Step 3: Apply voting or means according to the application

Table 1. The training dataset with class labels

x	y	Label
2	3	RED
7	8	RED
5	7	BLUE
5	5	RED
4	4	BLUE
1	8	BLUE

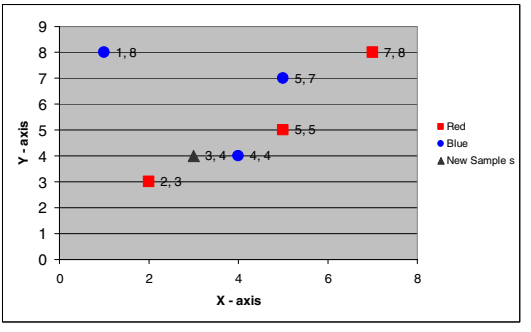

Fig. 1(a). The Plotted data in XY plane

Example: A typical example of KNN is classification. Suppose we have the following training data with class C = {RED, BLUE}.

The new test sample is s = (3,4), what will be the class of s for k = 3, k = 5? The data is plotted on XY plane in fig 1(a).

The KNN show the nearest neighbors with k = 3 And k = 5 with respect to new sample *s*. The distance calculation between the training Dataset and new sample s is shown in Table 2.

Table 2. The distance calculation between the training Dataset and new sample **s**

X	Y	Euclidean Simplest form	Distance	Label
2	3	$\sqrt{(3-2)^2+(4-3)^2}$	1.41	RED
7	8	$\sqrt{(3-7)^2+(4-8)^2}$	5.66	RED
5	7	$\sqrt{(3-5)^2+(4-7)^2}$	3.61	BLUE
5	5	$\sqrt{(3-5)^2+(4-5)^2}$	2.24	RED
4	4	$\sqrt{(3-4)^2+(4-4)^2}$	1.00	BLUE
1	8	$\sqrt{(3-1)^2+(4-8)^2}$	4.47	BLUE

Table 3. The sorted list based on distance from Table 2

S No	Distance	Label
5	1.00	BLUE
1	1.41	RED
4	2.24	RED
3	3.61	BLUE
6	4.47	BLUE
2	5.66	RED

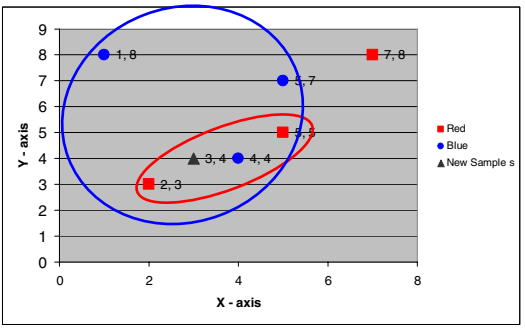

Fig. 1. (b) The classification for k = 3 and k = 5

In simple voting based on KNN where k = 3 the number of RED label is greater than the number of BLUE label so the new sample s will be classified as RED while for k = 5, the number of BLUE label is greater than the number of RED label so the s will be classified as BLUE. The process is shown in fig. 1(b)

4 Forecasting Using KNN

The dataset for the system was collected from the Pakistan Meteorological Department (PMD) and the National Climatic Data Center (NCDC), which consist of 10 years of historic data with a rich collection of attributes, i.e. temperature, Max Temp, Min Temp, Wind Speed, Max Wind Speed, Min Wind Speed, Dew Point, Sea Level, Snow Depth, Fog, etc. The dataset was noisy with some missing values. The data mining techniques (Means, bin means) were used for data cleansing. The attributes with textual values were replaced with numeric values, i.e. (YES/NO with 1/0, etc.). The final dataset (consisting of 10 years, 17 attributes and 40000 records for 10 cities) was brought in a consistent state and stored in MS ACCESS format.

Initialization and Prediction range setting: four basic values need to be initialized. *Current date*, the date from which onward the prediction is sought; *No-of-days-to-forecast*; *k – value* required by KNN; and *Attribute selection*, for which the prediction should be made. Based on this information, the data is extracted from the database and loaded into main memory in appropriate data structures (arrays). We have the following information in main memory.

Actual Record, for comparison with predicted values
Previous Records, these records are used as a base sequence during Distance Measurement (explained later in this section)
All Records, all records in a database for selected attributes and selected regions/City.

Applying Euclidean Distance (ED): Once the data is ready in memory, the data is divided into sequences. Each sequence size is *No of days to forecast* x *selected Attribute*. Thus, the *previous records* become a sequence by itself and will be used as a *base sequence* in ED. The dataset consists of all record for selected attributes are divided into sequences. The number of total sequences is *Total values in all records / total elements in base sequence.* Each sequence is taken and ED is applied element by element with base sequence. The process is summarized

```
Algorithm ED(No_of_Sequences, Size_of_Sequence)
While i < Total_element Do
  For j = 0 to Size_of_Sequence
    Sum = Sum + (Power(BaseSequence[j] - AllSequence[i]),2)
    i++
  End For
Distance = Sqrt(Sum)
End While.
```

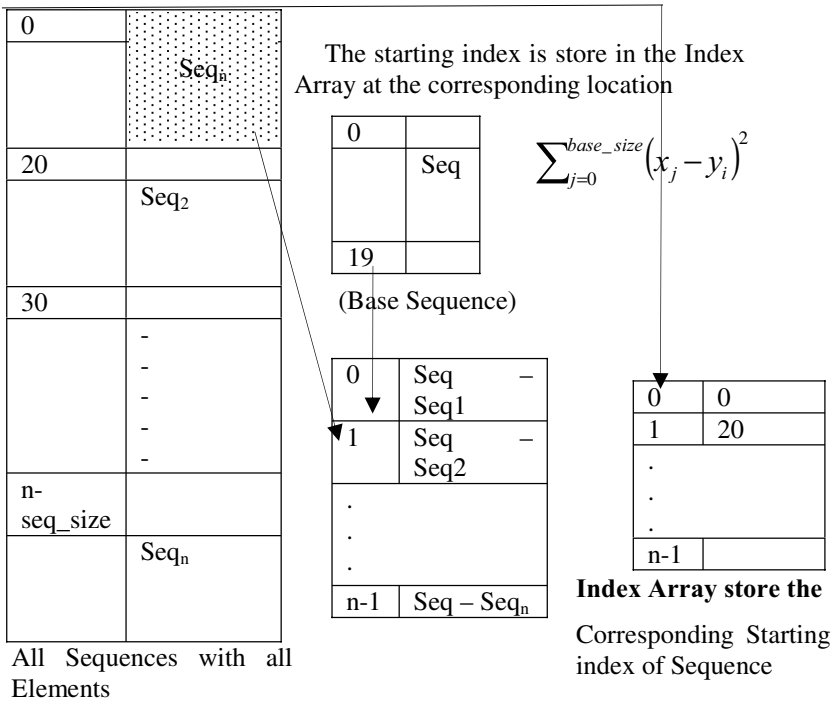

Fig. 2. Sequence wise Distance Calculation array management

Example: if *current date* is 20-Jan-1995 and *days to forecast* is 10 with 2 attribute selected for forecast and we have 507 records, then the previous record consist of 10 days back from the current date, i.e. 11-Jan-1995 to 20-Jan-1995, and the size of each sequence will be (10x2) = 20 elements. The total number of sequences will be 507/20 = 25.

The outer loop in the algorithm will take care of the sequences while the inner loop will calculate the formula

$$\sum_{j=0}^{total\_elements} \left(x_j - y_i\right)^2, 0 \le i < size\_of\_Sequence$$

where x_j is element in *base sequence* and y_i is an element in all sequences. The square root of the sum is taken sequences wise for all sequences and store in *Distance Array* along with starting index in *index Array* of each sequence at the corresponding element. (Fig. 2).

Table 4. The k nearest sequences

Sequences	Days	Attributes	Values
Seq 1	1	Max	52.2546
		Min	35.2457
	2	Max	50.3658
		Min	37.2657
	3	Max	48.2689
		Min	36.9882
	4	Max	49.8754
		Min	32.3655
Seq 2	1	Max	4.5
		Min	34.4587
	2	Max	55.6589
		Min	38.4556
	3	Max	51.2154
		Min	36.3636
	4	Max	48.1458
		Min	32.1554

Table 5. Final Predicted values

Days	Attributes	Values
1	Max	51.50313
	Min	33.41967
2	Max	50.86453
	Min	37.01423
3	Max	49.20323
	Min	36.53553
4	Max	49.4933
	Min	34.1263

Day1.Max = Seq1.Day1.Max + Seq2.Day1.Max;

Day1.Min = Seq1.Day1.Min + Seq2.Day1.Min;

Fig. 3. The top k sequences that are used for prediction; the mean is applied attribute wise and day wise

Sorting of Distance array and index array: The Distance array and Index array is arrange into such order that $d_i \le d_{i+1}$ the following algorithm is used to sort the both arrays.

```
Algorithm Sort(DistanceArray, IndexArray, Size)
For i = 0 to size - 1
  For j = 0 to (size - i) - 1
    If Distance[i] > DistanceArray[i+1] then
     Swap DistanceArray[i] and DistanceArray[i+1]
```

```
      Swap IndexArray[i] and IndexArray[i+1]
    End If
  End For
End For.
```

Applying KNN mean: Once the arrays are sorted the top – k sequences (the sequences with the smallest differences) are chosen for KNN means and their corresponding values are retrieved using the indices store in IndexArray. The sum is taken, attributes wise and day wise, of all sequences and then its mean is calculated, that is the predicted value for the specific day and attribute.

Example: The final prediction for 2 selected attributes (Max Temperature, min temperature), days to forecast = 4 and k = 2, is shown in Fig. 3.

5 Experiment

The experiments were performed on two datasets. One mentioned in section 3 had over 40000 records, and a second had 80000+ records. Each had the same number of attributes and belongs to same regions. The accuracy of the results was checked for deferent values of k and different time spans for both dataset. The performance of the system is evaluated on Windows XP, RAM 256 with 733 MHz. The execution time for KNN and prediction was measure from one to seventeen attributes. Fig. 4 show the KNN and Prediction for all attribute. The average time is less than 1 second when seventeen attributes are predicted at the same time.

Fig. 5 shows the load time for database into main memory. Although MS Access provides less overhead on loading the database, still it is long for attribute greater than 5. Fig. 6 show the accuracy of prediction with actual record for Sea Level Pressure (SLP) for 40K dataset with k = 5. The predicted records were checked against the actual records in the database for accuracy, i.e. prediction for 20 days from 15-Feb-2000, and the results were compared to the actual record of the specified time span in the database. Fig. 7 compares the database load time and prediction time for different values of k. the experiment shows that the load time is not effected by deferent values of k, if number of attribute remain constant. The prediction time although varies but remain less than 1 second for all values of k.

The value of K is important and should be selected carefully. It depends on the nature and volume of dataset. In our experiment k = 5 to k = 10 gives accurate results for small (40k) dataset while for large (80k) dataset the results are accurate for k = 35 to k = 40. Figure 8 shows the graphs for different values of k. The attributes are same in all figures to show the difference for both datasets.

The reason for these different values of K is, for large dataset we take the sequences with larger difference that in turn affects the predicted value which is mean of these selected sequences. Thus for small dataset we select small no of nearest neighbors while in large dataset, the small no of sequences are not enough to predict the accurate result and we require to select larger value for k to get accurate results. All odd figures (a, c, e, etc.) belong to 40k dataset while the even figures (b, d, f, etc.) are for 80k. The value of k is same for two dataset in consecutive figures.

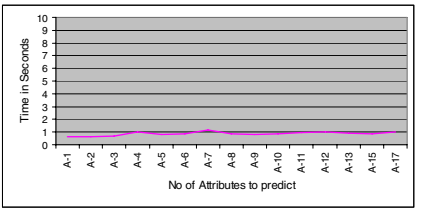

Fig. 4. Shows the prediction time for various values of k

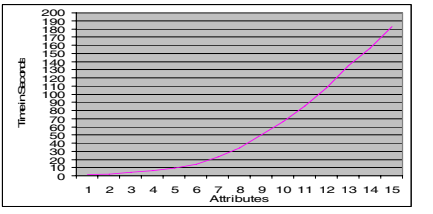

Fig. 5. Database load time for all attributes

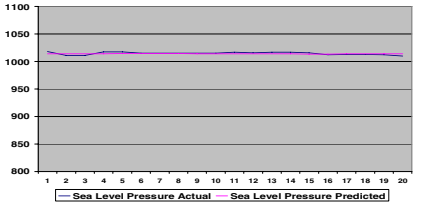

Fig. 6. Compare the actual and predicted record for SPL

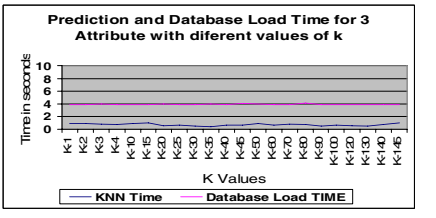

Fig. 7. Database load time and prediction time All Attributes simultaneously

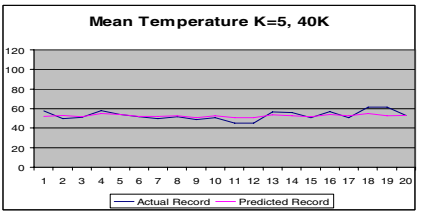

Fig. 8a. Mean Temperature for K=5

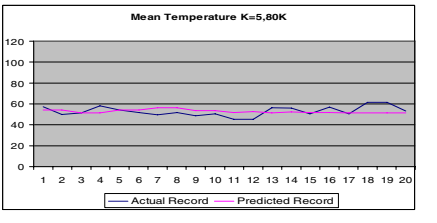

Fig. 8b. Mean Temperature for K=5

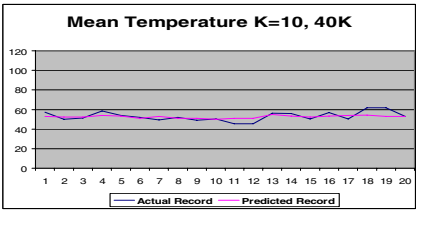

Fig. 8c. Mean Temperature for K=10

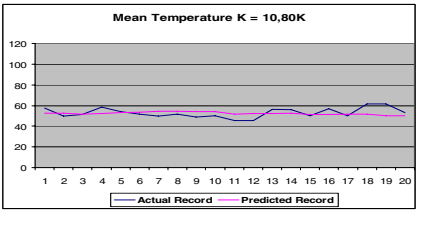

Fig. 8d. Mean Temperature for K=10

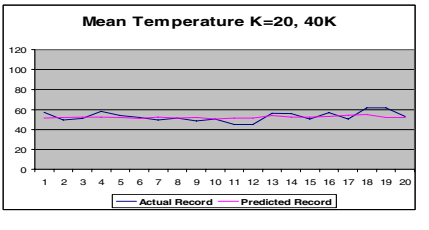

Fig. 8e. Mean Temperature for K=80

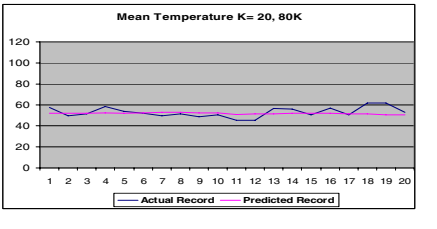

Fig. 8f. Mean Temperature for K=20

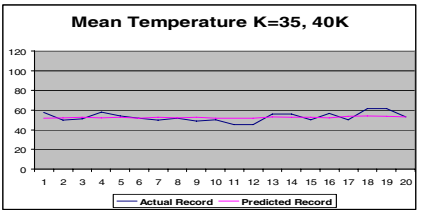

Fig. 8g. Mean Temperature for K=35

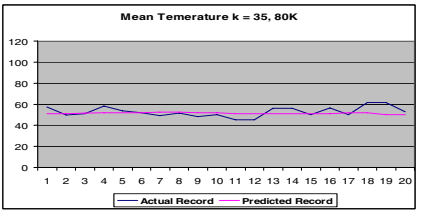

Fig. 8h. Mean Temperature for K=35

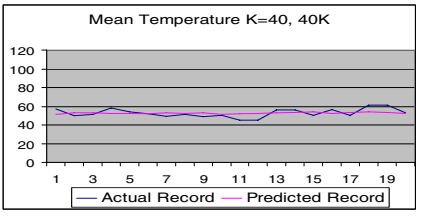

Fig. 8k. Mean Temperature for K=40

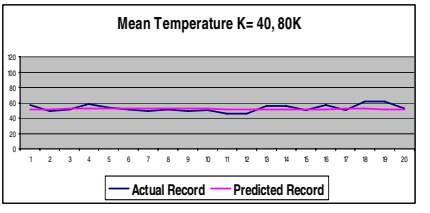

Fig. 8j. Mean Temperature for K=40

The accuracy is 96.66% for attributes having Boolean values, i.e. Fog, Hail, snow ice, Thunder etc. and for larger dataset the accuracy is even above 98.33%. Fig. 9a. shows prediction for small (40K) dataset and Fig. 9b. shows the same results for large (80K) dataset. The Figures show the result is 100% for hail and snow ice in both data-set. While for fog the results of 80K is better than 40k. The reason is discrete values of attributes. The results generated by the system are real numbers and were rounded to convert it to yes/no. the values 0.0 to 0.49 were rounded to Zero (No) and 0.5 and above are considered 1 (Yes).

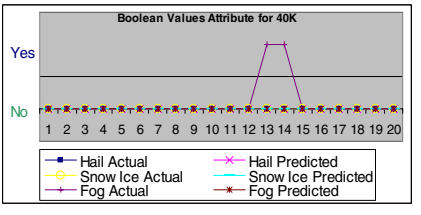

Fig. 9a. Boolean values attribute k=10

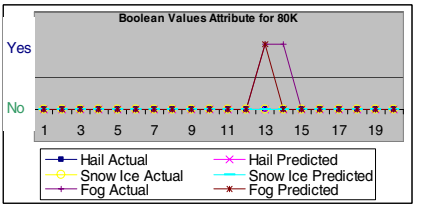

Fig. 9b. Boolean values attribute k=40

6 Conclusion and Future Work

The CP-KNN can predict up to seventeen climatic attributes, i.e. Mean Temperature, Max Temperature, Min Temperature, SST, SLP, Gust etc. at the same time. None of the previous developed systems can predict such a huge set of attribute at the same time with such level of accuracy. Recently Climate Prediction Tool (CPT) [from NOAA] has developed that works with multiple attributes but this new version is only

used in research labs and not publicly available. The predicted result of CP-KNN is easier to understand as these results are in form YES/NO (for Boolean Attributes) and numeric values. Thus this new system can be used by non-professionals related to any field, e.g. Agriculture, Irrigation, retailers and those specially related to weather sensitive businesses [13].

CP-KNN cannot incorporate to reflect the global changes (ENSO events) but will work correctly with the areas not prone to these global effects. However as these events has some known pattern is advance, e.g. SLP, SST and wind speed etc [6, 14]. It can be modeled using data mining pattern recognition techniques.

References

1. Hansen, J.W., Sivakumar, M.V.K.: Advances in applying climate prediction to agriculture. Climate Research 33, 1–2 (2006)
2. Sayuti, R., Karyadi, W., Yasin, I., Abawi, Y.: Factors affecting the use of climate forecasts in agriculture: a case study of Lombok Island, Indonesia. ACIAR Technical Reports Series, No. 59, pp. 15-21 (2004)
3. Murray-Ruest, H., Lashari, B., Memon, Y.: Water distribution equity in Sindh Province, Pakistan, Pakistan Country Series No. 1, Working Paper 9, International Water Management Institute, Lahore, Pakistan (2000)
4. Stern, P.C., Easterling, W.E.: Making Climate Forecasts Matter. National Academy Press (1999)
5. Landman, W.A., Mason, S.J.: Change in the association between Indian Ocean sea-surface temperatures and summer rainfall over South Africa and Namibia. International Journal of Climatology 19, 1477–1492 (1999)
6. Landman, W.A.: A canonical correlation analysis model to predict South African summer rainfall. NOAA Experimental Long-Lead Forecast Bulletin 4(4), 23–24 (1995)
7. Hsieh, W.W.: Nonlinear Canonical Correlation Analysis of the Tropical Pacific Climate Variability Using a Neural Network Approach. Journal of Climate 14(12), 2528–2539 (2001)
8. Hays, S.P., Mangum, L.J., Picaut, J., Sumi, A., Takeuchi, K.: TOGA-TAO: A moored array for real time measurement in the tropical Pacific ocean. Bulletin of the American Meteorological Society 72(3), 339–347 (1991)
9. Mason, S.E., Goddard, L., Zebiak, S.J., Ropelewski, C.F., Basher, R., Cane, M.A.: Current Approaches to Seasonal to Interannual Climate Predictions. International Journal of Climatology 21, 1111–1152 (2001)
10. Han, J., Kamber, M.: Data Mining Concepts and Techniques. Elsevier Science and Technology, Amsterdam (2006)
11. Fix, E., Hodges, J.L.: Discriminatory Analysis - Nonparametric Discrimination: Consistency Properties. USAF school of Aviation Medicine, Randolph Field Texas (1951)
12. Larose, D.T.: Discovering Knowledge in Data: An Introduction to Data Mining. Wiley, Chichester (2005)
13. Lettre, J.: Business Planning, Decisionmaking and Ethical Aspects of Seasonal Climate Forecasting (1999), http://members.aol.com/gml1000/busclim.html
14. Mason, S.J., Goddard, L., Graham, N.E., Yulaeva, E., Sun, L., Arkin, P.A.: The IRI seasonal climate prediction system and the 1997/1998 El Niño event. Bulletin of the American Meteorological Society 80, 1853–1873 (1999)

15. Landman, W.A., Mason, S.J.: Forecasts of Near-Global Sea Surface Temperatures Using Canonical Correlation Analysis. Journal of Climate 14(18), 3819–3833 (2001)
16. Rogel, P., Maisonnave, E.: Using Jason-1 and Topex/Poseidon data for seasonal climate prediction studies. AVISO Altimetry Newsletter 8, 115–116 (2002)
17. White, A.B., Kumar, P., Tcheng, D.: A data mining approach for understanding control on climate induced inter-annual vegetation variability over the United State. Remote sensing of Environments 98, 1–20 (2005)
18. Basak, J., Sudarshan, A., Trivedi, D., Santhanam, M.S.: Weather Data Mining using Component Analysis. Journal of Machine Learning Research 5, 239–253 (2004)