Volume 1 No. 3, July 2011 ISSN-2223-4985
International Journal of Information and Communication Technology Research

http://www.esjournals.org

Implementation of Data Mining Techniques for Meteorological Data Analysis
(A case study for Gaza Strip)

Sarah N. Kohail, Alaa M. El-Halees
Faculty of Information Technology
The Islamic University of Gaza

ABSTRACT

Meteorological data mining is a form of data mining concerned with finding hidden patterns inside largely available meteorological data, so that the information retrieved can be transformed into usable knowledge. Weather is one of the meteorological data that is rich by important knowledge. In this paper we try to extract useful knowledge from weather daily historical data collected locally at Gaza Strip city. The data include nine years period [1977-1985]. After data preprocessing, we apply outlier analysis, clustering, prediction, classification and association rules mining techniques. For each mining technique, we present the extracted knowledge and describe its importance in meteorological field.

Keywords: *Data Mining ,Meteorology, Weather Prediction.*

I. INTRODUCTION

The increasing availability of climate data during the last decades (observational records, radar and satellite maps, proxy data, etc.) makes it important to find an effective and accurate tools to analyze and extract hidden knowledge from this huge data. Meteorological data mining is a form of Data mining concerned with finding hidden patterns inside largely available meteorological data, so that the information retrieved can be transformed into usable knowledge [1]. Useful knowledge can play important role in understanding the climate variability and climate prediction. In turn, this understanding can be used to support many important sectors that are affected by climate like agriculture, vegetation, water resources and tourism.

The study area, Gaza Strip lies on the Eastern coast of the Mediterranean Sea. Gaza locates at 31 25 N latitude and 34 20 E longitude and characterized by a Mediterranean climate. At winter, climate is temperate and wet resulted from a penetration of mid-latitude depressions accompanied by westerly wind moving eastward over the Mediterranean basin. Summer is dry and warm to hot summer caused by eastward extension of the Azores high pressure [5].

In this paper we try to extract useful knowledge from daily historical weather data collected locally at Gaza Strip city. The data include nine years period [1977-1985]. After data preprocessing, we apply outlier analysis, clustering, prediction, classification and association rules mining techniques. After each mining technique, we present the extracted knowledge and describe its importance in meteorological field.

The rest of this paper is organized as follows: Section 2 presents some related works. Section 3 summarize data collection process. Section 4 describes the data preprocessing and preparation step. Section 5 present and discuss our experimentation. Finally conclusions are presented in the last section.

II. RELATED WORKS

Many researchers have tried to use data mining technologies in areas related to meteorology and weather prediction. Kotsiantis et al. [12] predict daily average, maximum and minimum temperature for Patras city in Greek by using six different data mining methods: Feed-Forward Back Propagation (BP), k-Nearest Neighbor (KNN), M5rules algorithm, linear least-squares regression (LR), Decision tree and instance based learning (IB3). They use four years period data [2002-2005] of temperature, relative humidity and rainfall. The results they obtained in this study were accurate in terms of Correlation Coefficient and Root Mean Square. The emphasis in [4] is on using DBSCAN (Density Based Spatial Clustering of Applications with Noise) clustering algorithm to categorize Turkey into regions according to climatic characteristics. They use the daily maximum and minimum temperature records between 1930 and 1996 from 258 stations. They draw that this type of data mining application can help meteorological to create faster forecast and decisions and provide more performance and reliability than any other methods.

Data mining have been employed successfully to build a very important applications in the field of meteorology like predicting abnormal events like hurricanes, storms and river flood prediction [2][15]. These applications can maintain public safety and welfare. In this context, Zhang and Huang [22] propose a new framework to discover dynamic interdimension association rules for local-scale weather prediction of Dallas City. The usefulness of applying association mining is to find a strong relation between severe conditions and the change tendencies of the measurements of the weather. The authors conclude with some predicates extracted from

the obtained rules. Another contribution to detect severe events using data mining is by [14] and [18]. Peters at el. [18] used the volumetric radar data to detect storm events and classify them into four types: hail, heavy rain, tornadoes, and wind.

Using data mining in meteorological application is not limited to prediction, but it also extend to participate in many important fields like water resource management [11] and air pollution management [13].

Mining techniques also can be applied to various types of data like weather images and radar maps extend to characteristic features extracted from this weather images can be used to represent various weather patterns [21].

III. DATA COLLECTION

To achieve this study, we use historical data records for nine years period [1977-1985] recorded for one weather station at Gaza Strip. The obtained record include the daily average relative humidity (%), average temperature (Celsius), wind speed (KM/H), wind direction, the time of the highest wind speed and rainfall observation.

IV. DATA PREPROCESSING

An important step in the data mining process is data preprocessing. One of the challenges that face the knowledge discovery process in meteorological data is poor data quality. For this reason we try to prepare our data carefully to obtain accurate and correct results. First we choose the most related attributes to our mining task. For this purpose we neglect the year, wind direction and time of the highest wind speed attributes. Then we try to fill the missing with appropriate values. In our data we have little missing (no more than 11 value). Because we are working with weather data that is a form of time series, we must preserve the series smoothness and consistency. So we use linear interpolation method. This method is effective method to fill missing values in the case of time series where the missed value is strongly related to its previous and next values. After filling the missing values we apply windowing operation on temperature attribute to create three lags (time frame) of temperature lag_{t-1}, lag_{t-2}, lag_{t-3}, where lag is a past observations (days before) and t represent the day of the current class label (in the case of classification and prediction).

V. EXPERIMENTS AND RESULTS

This section presents our experimentation, discusses the extracted knowledge and describes its importance in meteorological field.

A. Outlier Analysis

Outliers in weather data can occur due to data entry problem and faulty data collection instruments, or it can represent abnormal change, sudden natural events such as tornadoes, hurricane, and forest fires [12]. Outlier detection can be considered as preprocessing process. For this reason, we apply it first to remove outliers early to avoid its impact on other mining methods.

We use distance base approach and utilize visualization to find the relationship between the weather elements and detect the out of relation points. We detect 12 outliers. Table 1 present the analysis of detected outliers. Figure 1 illustrate outlier example detected at wind speed attribute. The outlier represents very high speed wind (strong hurricane).

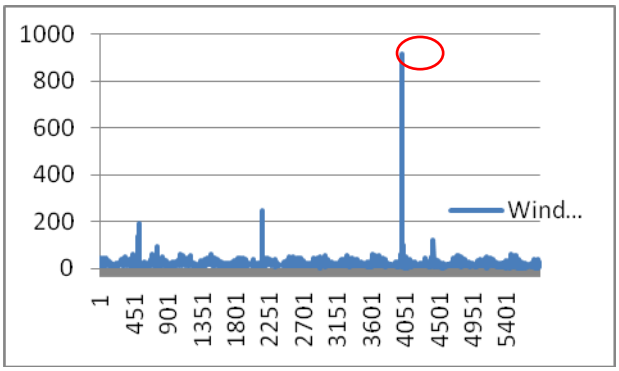

Figure 1: Example of outlier in wind speed attribute

Table 1: Outlier analysis for Gaza city weather data

Number of outliers	Analysis
4	Entry problem (error or unrealistic data)
4	Rainy days in non-rainy seasons (hot weather)
3	Very high speed wind (strong hurricane)
1	Sudden change in temperature

To avoid the negative impact of unrealistic data on our results we remove them.

B. Clustering

In the field of meteorology and climate monitoring, highly sophisticated measurement

technologies have been elaborated over the last few years, producing a huge amount of data. This huge raw data is difficult to analyze and understand. In this case clustering aim to improve the understanding of natural climate processes, to assess the quality of climate model results and to identify prevailing system features and their typical scales for specific atmospheric regimes [17]. Clustering have been applied successfully in many meteorological application like determinate the precipitation weather type by finding the similarity between satellite cloud images [19], seasonal clustering [10] [11] and climatology [4].

In our experiments we use k-means clustering algorithm using k=4. K-means algorithm is the most popular clustering tool used in scientific and industrial applications. The name comes from representing each of k clusters by the mean (or weighted average) of its points, the so-called centroid. The centroid of a cluster is a point whose coordinates are the mean of the coordinates of all the points in the clusters [3]. Figure 2, show the clusters distribution and Figure 3 show the clusters centroid. From these two figures we can recognize the characteristics of Gaza city seasons. Cluster 1 show the largest amount of rain, lower temperature, moderate humidity and faster wind speed, so we can say that it represent winter season period and its characteristics. The distributions of this cluster include: December, January, February, March and April.

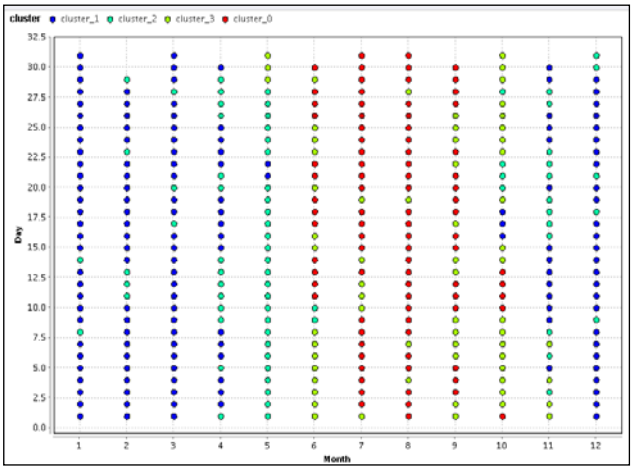

Figure 2: Clusters distribution for Gaza city weather data (k=4)

Cluster 0 represent the least amount of rain, higher temperature, higher humidity and slower wind speed, so we can say that it represent summer season period. The distributions of this cluster include: the end of June, July, August and September. In this way we can consider cluster 2 as autumn (the period to navigate from summer to winter) and cluster 3 as spring (the period to navigate from winter to summer). Figure 2 show clearly the navigation between seasons.

This understanding of seasons is very important to many sectors as well as many industries which largely dependent on the weather conditions like agriculture, vegetation, water resources and tourism.

Attribute	cluster_0	cluster_1	cluster_2	cluster_3
Rain	0.061	9.402	0.517	0.194
Month	7.735	4.455	5.019	6.869
Wind	14.434	36.898	15.419	18.903
Day	15.953	14.875	16.236	14.803
Temp-1	23.634	14.008	15.198	19.513
Temp-0	23.720	13.654	15.552	18.868
Temp-2	23.825	14.785	15.389	18.303
Temp-3	23.836	14.926	15.484	18.082
RH	75.935	70.807	73.713	55.470

Figure 3: Clusters centroid

C. Prediction

Prediction is the most used data mining task in the field of meteorology. Data mining techniques provides with a level of confidence about the predicted solutions in terms of the consistency of prediction and in terms of the frequency of correct predictions [1]. Also it applied successfully to predict different weather elements like wind speed [7], rainfall [6], cloud [9] and temperature [12].

In our paper we use two prediction methods to predict the daily average temperature of Gaza city. The first method is artificial neural networks (ANN) with 8 input layer, 6 hidden layer and one output layer. The second method is least median squares linear regression by Rousseeuw [20]. We use day, month, three lags temperature (days before) humidity and wind speed as inputs. We use 70% of data for training (as continuous series) and reminder for testing. The best obtained results are recorded in Table 2. ANN provides better correlation coefficient between the actual and predicted temperature, and lower Root Mean Square Error (RMSE). When performing the t-test statistical evaluation, it indicates that the two models are probably different; therefore we can say that neural network is better than least median squares linear regression because it provide better accuracy and higher correlation coefficient between the actual and predicted temperature.

Table 2: Prediction results for two prediction methods applied for Gaza city weather data

Method	**Correlation Coefficient**	**RMSE**
Least Median Squares Linear Regression	0.924	1.691
Neural Networks Learning rate: 0.3 Momentum: 0.2	0.933	1.726

The prediction models after that can be used to predict the daily temperature for Gaza city. Agricultural sector can benefit from these predictions, especially that income for many peoples in Gaza city depend highly on agriculture.

D. Classification

Classification has been utilized in many meteorological applications; for example classifying to predict the weather on a particular day will be "sunny", "rainy" or "cloudy". Also it used widely in classifying geographical location based on its climate and classify weather conditions based on the agricultural crops suitable to cultivate on each climate. Classification include methods that can produce useful rules like decision tree. These rules can be utilized as prediction statements.

We apply four classification techniques on our data and record the best results in Table 3. The same dataset used in section C, but we change the class label into categorical. Classification task try to classify the data records into three classes hot (temperature is higher than 23 °C), warm (between 16 °C and 23 °C) or cold weather (bellow 16 °C).

Table 3: Classification results for two prediction methods applied for Gaza city weather data

Method	Accuracy	RMSE
Naive Bayes	82.11%	0.389
KNN (*k=30*)	81.81%	0.384
Decision Tree	81.40%	0.365
Neural Networks	85.77%	0.333

Neutral networks provide the best results in term of classification accuracy and Root Mean Square Error (RMSE). In the future we can utilize the produced classification model to predict new instance membership.

E. Association

Association rule mining finds interesting relationships in data. The goal of associative rule data mining is to find all associative rules that have high confidence (Strong Rules) in the data set. In meteorological application, association mining used to find the relationship between the weather elements and natural events, weather and disaster prediction [8], and multi-station atmospheric data analysis [16].

Table 4 illustrates some useful rules extracted from Gaza weather data ordered by confidence (higher confidence represent more general and effective rule).

Table 4: Associations rules for Gaza city weather data

#	Rule	Conf.
1	[RH=mid Temp=warm Wind=Moderate] ==> [Rain=no rain]	0.99
2	[RH=high Temp=warm] ==> [Rain=no rain]	0.99
3	[Temp=warm Wind=Moderate] ==> [Rain=no rain]	0.99
4	[month = 2]==> [temp = cold]	0.96
5	[month = 1] ==> [temp = cold]	0.96
6	[month = 12] ==> [temp = cold]	0.95
7	[Wind=Light] ==> [Rain=no rain]	0.91
8	[Wind=light Temp=cold rain]==>[RH=moderate]	0.91
9	[Rain= Heavy Rain]==>[Temp = cold]	0.88
10	[Temp = cold]==>[Wind = Moderate]	0.74
11	[RH= low Wind = Moderate Temp=warm] ==> [Rain=Light Rain]	0.65
12	[Wind = Moderate] --> [RH = mid]	0.60

Rules #1, #2, #3, #7 and #11, can be used to predict rainfall. For example from rule #1 we understand that there is no rain tomorrow if today is warm (temperature between 16 °C and 23 °C), wind speed is moderate (13-30 km/h) and relative humidity is mid (between 56.5 - 76.0). Also Rule #11 could be used for rain prediction, it means that if the relative humidity today is low (below 36), wind speed is moderate and temperature is warm then, rain tomorrow maybe light (< 2.5 millimeters per hour). Rules #4, #5 and #6 provide with better understanding for Gaza city weather. These rules give us an indication that cold season includes December, January and February.

VI. CONCLUSION AND FUTURE WORK

In this paper we applied knowledge discovery process to extract knowledge from Gaza city weather dataset. The dataset include nine years period [1977-1985] of daily weather observation. We went through all knowledge discovery process and applied many data mining techniques like outlier analysis, prediction, classification, association mining and clustering. Data mining tasks provide a very useful and accurate knowledge in a form of rules, models, and visual graphs. This knowledge can be used to obtain useful prediction and support the decision making for different sectors.

Our future work include building adaptive and dynamic data mining methods that can learn dynamically to match the nature of rapidly changeable weather nature and sudden events.

REFERENCES

[1] Baboo S., and Shereef K., "Applicability of Data Mining Techniques for Climate Prediction – A Survey Approach," International Journal of Computer

Science and Information Security, Vol. 8, No. 1, April 2010.

[2] Bartok J., Habala O., Bednar P., Gazak M., and Hluch L., "Data mining and integration for predicting significant meteorological phenomena," Procedia Computer Science, p.37 – 46. 2010.

[3] Berkhin P., "Survey of clustering data mining techniques, Accrue Software, San Jose", CA, Tech. Rep., 2002.

[4] Bilgin T., and Çamurcu Y., "A Data Mining Application on Air Temperature Database," Advances in Information Systems, Springer Berlin, Heidelberg, pp.68-76 .2004.

[5] Central Intelligence Agency, "CIA World Factbook 2011," MobileReference. 2010.

[6] Dong-Jun S., Breidenbach, J. P., "Real-Time Correction of Spatially Nonuniform Bias in Radar Rainfall Data Using Rain Gauge Measurements," Journal of Hydrometeorology, vol. 3, no. 2, pp. 93-111. 2002.

[7] G. Li, and J. Shi, "On comparing three artificial neural networks for wind speed forecasting," Applied Energy, vol. 87, no. 7, pp. 2313-2320, Jul. 2010.

[8] Guo Z., Dai X., Lin H.,"Application of Association Rule in Disaster Weather Forecasting, "Annals of GIS, Volume 10, pp. 68 – 7201. June 2004.

[9] Hluchý L., Habala O., Bartok J., Bednár P., Gažák M., "Prediction of significant meteorological phenomena using advanced data mining and integration methods," Fuzzy Systems and Knowledge Discovery (FSKD), vol.6, no., pp.2998-3002, 10-12 Aug. 2010.

[10] Inniss, T. R., "Seasonal clustering technique for time series data". European Journal of Operational Research, 175, 376–384. 2006.

[11] Jan Z., Abrar M., Bashir S., Mirza A., "Seasonal to Inter-annual Climate Prediction Using Data Mining KNN Technique," Wireless Networks, Information Processing and Systems Communications in Computer and Information Science, pp. 40-51.2009.

[12] Kotsiantis S., Kostoulas A., Lykoudis S., Argiriou A., and Menagias K., "Using Data Mining Techniques for Estimating Minimum, Maximum and Average Daily Temperature Values, International Journal of Mathematical, Physical and Engineering Sciences", pp. 16-20. 2007.

[13] Li S., and Shue L., "Data mining to aid policy making in air pollution management," Expert Systems with Applications, vol. 27, pp. 331-340, 2004.

[14] Li X., Plale N., Vijayakumar R., Ramachandran S., Graves H., "Conover. Real-time storm detection and weather forecast activation through data mining and events processing, "To appear Earth Science Informatics, H.A. Babaie, Ed., Springer. 2008.

[15] Mohammadi K., Eslami H. R., Kahawita R., "Parameter estimation of an ARMA model for river flow forecasting using goal programming. "Journal of Hydrology, 331, 293–299. 2006.

[16] Nandagopal S., Karthik S., Arunachalam V.,"Mining of Meteorological Data Using Modified Apriori Algorithm," European Journal of Scientific Research, No.2, pp.295-308. 2010.

[17] Nocke T., Schumann H., Böhm U., "Methods for the Visualization of Clustered Climate Data," Computational Statistics 19(1), pp. 75–94, 2004.

[18] Peters J., Suraj Z., Shan S., Ramanna S., Pedrycz W., Pizzi N., "Classification of meteorological volumetric radar data using rough set methods," Pattern Recognition Letters, pp.911–920. 2003.

[19] Qin K., Xu M., Du Y., Yue S., "Cloud Model and Hierarchical Clustering Based Spatial Data Mining Method and Application". The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. 2008.

[20] Rousseeuw P., "Least Median of Squares Regression," Journal of the American Statistical Association Vol. 79, No. 388, pp. 871-880. Dec, 1984.

[21] Siddiqui, K.J. and Nugen, S.M., "Knowledge based system for weather information processing and forecasting," Geoscience and Remote Sensing Symposium, pp.1099-1101. 27-31 May 1996.

[22] Zhang Z., Wu W., Huang Y., "Mining dynamic interdimension association rules for local-scale weather prediction," In the Proceedings of the 28th Annual International Computer Software and Applications Conference, 2004.