ABSTRACT

Evolution in data storage and large databases has generated an imperative need for new techniques and tools for data analysis and knowledge discovery. Statistical, computational, and machine learning tools have been used by scientific community and researchers in the area of scientific data analysis. Data mining is a data analysis technique which explores data and discovers meaningful information and knowledge. Predictive data mining will be a good aid for weather forecasting which predicts the future state of the atmosphere with the present state information.

Among weather elements, surface air temperature is the key determinant for vegetation, animals and human livelihood in a particular location of earth which also influences most atmospheric parameters like precipitation, humidity, pressure, wind speed and wind direction. Since deviations in surface air temperature claims many lives in the earth, timely prediction of minimum and maximum surface air temperature will help in planning and governing very hot and very cold climate. However, it is challenging because of the dynamic atmospheric parameters associated with the temperature event.

In the proposed work, data mining technique based ensemble forecast of lead seven days minimum and maximum surface air temperature is developed for the location Chennai, India. Ensemble forecasting has been adopted to reduce the uncertainty and the increase the reliability in the
forecasting and forewarning process. Multi-model ensemble forecast has been provided using forecast models generated using three statistical techniques, namely, Multiple Linear Regression (MLR), Support Vector Machine regression (SVMr), Multivariate Adaptive Regression Spline (MARS), and three soft computing techniques, namely, Artificial Neural Network (ANN), Adaptive Neuro Fuzzy Inference System (ANFIS) and Genetic Programming (GP). Surface temperature forecast models have been formulated using eight ground station recorded predictors. Dataset of nine years have been used in the analysis, out of which dataset of eight years have been used as training dataset for model creation and one year dataset has been used as testing dataset for model performance validation. The models have been created with the data analysis tools namely MATLAB and Statistica 8. To assess the model efficiency and performance, statistical error analysis by calculating Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE) and Coefficient of determination ($R^2$) have been carried out.

The result of the analysis states that all the six techniques used, provide reliable, precise and consistent performance for both minimum and maximum temperature forecast. By using the models created using these approaches as ensemble members, ensemble forecast has been done. Comparison of ensemble forecast with other forecast models suggests that, forecast produced by ensemble forecast is the best. The analysis also emphasizes that the prediction performance is good at smaller lead days and it decreases gradually to higher lead days for both minimum and maximum temperature.