Chapter 10
Application of Data Mining Techniques in Weather Forecasting

ThippaReddy Gadekallu
VIT University, India

Bushra Kidwai
VIT University, India

Saksham Sharma
VIT University, India

Rishabh Pareek
VIT University, India

Sudheer Karnam
VIT University, India

ABSTRACT

Weather forecasting is a vital application in meteorology and has been one of the most scientifically and technologically challenging problems around the world in the last century. In this chapter, the authors investigate the use of data mining techniques in forecasting maximum temperature, rainfall, evaporation, and wind speed. This was carried out using artificial decision tree, naive Bayes, random forest, K-nearest neighbors (IBk) algorithms, and meteorological data collected between 2013 and 2014 from the city of Delhi. The performances of these algorithms were compared using standard performance metrics, and the algorithm which gave the best results used to generate classification rules for the mean weather variables. The results show that given enough case data, data mining techniques can be used for weather forecasting and climate change studies.

INTRODUCTION

Weather forecasting has been one of the most scientifically and technologically challenging problems around the world in the last century. This is due mainly to two factors: first, it's used for many human activities and secondly, due to the opportunism created by the various technological advances that are

DOI: 10.4018/978-1-5225-4999-4.ch010

directly related to this concrete research field, like the evolution of computation and the improvement in measurement systems (Weather Underground, 2018).

Weather forecasting entails predicting how the present state of the atmosphere will change. Present weather conditions are obtained by ground observations, observations from ships and aircraft, radiosondes, Doppler radar, and satellites (India Meteorological Department, 2015). This information is sent to meteorological centers where the data are collected, analyzed, and made into a variety of charts, maps, and graphs. Modern high-speed computers transfer the many thousands of observations onto surface and upper-air maps. Computers draw the lines on the maps with help from meteorologists, who correct for any errors. A final map is called an analysis. Computers not only draw the maps but predict how the maps will look sometime in the future. The forecasting of weather by computer is known as numerical weather prediction (DePaul University, 2005; Elia, 2009).

To predict the weather by numerical means, meteorologists have developed atmospheric models that approximate the atmosphere by using mathematical equations to describe how atmospheric temperature, pressure, and moisture will change over time. The equations are programmed into a computer and data on the present atmospheric conditions are fed into the computer.

Data mining, also called Knowledge Discovery in Databases (KDD), is the field of discovering novel and potentially useful information from large amounts of data (Han & Micheline, 2007). In contrast to standard statistical methods, data mining techniques search for interesting information without demanding a priori hypotheses, the kind of patterns that can be discovered depend upon the data mining tasks employed (Brownlee, 2014; Gerardnico, 2016; Due, 2007).

A decision tree is a structure that includes a root node, branches, and leaf nodes (Accuweather, 2018; Dabberdt, 1981; Kumar et al., 2013; Casaset al., 2009). Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label. The topmost node in the tree is the root node. In machine learning, naive Bayes classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features (Technobium, 2016).

Naive Bayes has been studied extensively since the 1950s. It was introduced under a different name into the text retrieval community in the early 1960s, and remains a popular (baseline) method for text categorization, the problem of judging documents as belonging to one category or the other (such as spam or legitimate, sports or politics, etc.) with word frequencies as the features (Steingrimsson, n.d.). With appropriate pre-processing, it is competitive in this domain with more advanced methods including support vector machines. It also finds application in automatic medical diagnosis.

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of over fitting to their training set (Tribhuvan & Tribhuvan, 2014).

In pattern recognition, the k-nearest neighbours algorithm (k-NN) is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression.