P10.1 UTILIZING SITE-BASED DATA MINING IN NATIONAL CEILING AND VISIBILITY FORECASTING

G. Wiener*, P. Herzegh, R. Bateman and B. Chorbajian
National Center for Atmospheric Research, Boulder, CO

1 Introduction

The data mining field encompasses a wide variety of application domains including astronomy, atmospheric science, bioinformatics, business, computer vision, economics, high energy physics, medical imaging, molecular chemistry, robotics, security, surveillance and so on. Often the algorithms that have demonstrated capability in one domain have applicability in many of the other domains as well.

In this paper, we focus on the popular decision tree algorithm in the data mining field. We show how this algorithm can be applied to predicting ceiling, visibility and flight category for various forecast periods from a 2h forecast up through a 12h forecast.

Decision trees are useful when one is interested in predicting a categorical variable (a variable having a finite number of values) such as future flight category based on using a number of other variables such as current and past measurements of temperature, dew point, ceiling, visibility, wind speed, etc. The fact that the variable of interest or target variable must be categorical is not overly limiting in the ceiling and visibility case since both ceiling and visibility can be broken down into a finite set of intervals. For example, visibility can be broken down into ten categories of 1, 2, ..., 10 miles; a finer resolution would not be necessary for aviation.

There is a variant of the decision tree algorithm called the regression tree which is useful when the variable of interest or target variable has continuous values. For example, regression trees could be applied in predicting temperature based on other meteorological variables. In the ceiling and visibility case, regression trees can also be applied in predicting ceiling and visibility.

* Corresponding author address: Gerry Wiener, NCAR Research Applications Division, P.O. Box 3000, Boulder, CO 80307
Email: gerry@ucar.edu

In this particular study, we will discuss the application of the C5.0 decision tree software, http://www.rulequest.com.

2 Methodology

Our data mining process involved the following stages:

1. Preparing site-based data files
2. Organizing the data into training and test sets
3. Running the data mining software
4. Comparing the data mining results on the test set against persistence

2.1 Preparing site-based data files

We modified the standard Unidata METAR decoder in order to create a single daily file for each METAR site in the CONUS domain. The daily files contain all METAR observations for a particular day at a particular site. After the end of each day, we amalgamate the daily site file data with a corresponding history file for the site. These history files contain all METAR observations for CONUS sites from January 1, 1997 up to the current date.

2.2 Organizing the data into training and test sets

The C5.0 software expects the training and test set data to be in an ASCII tabular file format. Each line in the ASCII file has to contain all the predictor variables as well as the target variable for one training case. The totality of lines in the training file consists of all the training cases whereas the totality of lines in the test file consists of all the test cases. Note that the variables used in training must be identical to those used in testing. Thus, the tabular file format in the training and test files must be identical.

The first step in organizing the data into training and test sets involves determining what variables to include on each line in the training and test files. Basically, one needs to answer the question, "What data should be used for training?" For

example, should all training variables be from the same observation time or should training variables from different times be combined? Should trending information be included?

Before answering these questions, let's take a look at the variables in a METAR observation. Each METAR observation we utilize contains the following fields:

Year, month, day, hour, minute, sky cover, ceiling, visibility, obscuration, weather, severe weather, altimeter, temperature, dew point, max temperature, min temperature, max 24 hour temperature, min 24 hour temperature, wind speed, wind direction, wind gust, hourly precipitation, precipitation amount, 4 hour precipitation amount and snow depth.

All these fields are currently included in our training set except that year, day and minute are ignored. We also include derived fields such as dew point depression, flight category and u, v winds. In order to capture diurnal information we include 24 hour METAR observations prior to forecast valid time. Finally, to capture trending information we include 3 hour METAR observations prior to forecast initiation time.

More specifically, each line in our C5.0 training file consists of four data records that correspond to selected times at and before the forecast valid time. These are:

- Forecast valid time: The observed target parameter (e.g., ceiling, visibility or flight category) and any auxiliary forecast information used (as discussed in *Auxiliary Forecast Information* below).
- Forecast initiation time: METAR observations and associated derived parameters. These represent the latest (and most valuable) observations available for forecast generation.
- Forecast initiation time minus 3 hours: METAR observations and associated derived parameters. These data can be used to derive 3-hour tendencies in selected parameters.
- Forecast valid time minus 24 hours: METAR observations and associated derived parameters. These data are used to represent the role of diurnal effects as an aid to forecasting.

The second step in organizing the data into training and test sets is to determine which data records should go into the training set and which data records should go into the test set. Here one also has to take some care. The C5.0 software has the capability to randomly pull out data from the training set to use for testing. At first glance, this seems appealing since it's a feature of the software that obviates the need to create a test set. Unfortunately this approach leads to artificially high data mining scores. The reason for this is that the test data set produced by random selection is close in time to the data in the training set. In fact, training set examples will typically surround test set examples in time when using random test set selection. During standard operational usage, the training set will always be prior to the test set. Hence, it is important to model the data mining experiment in accord with standard operational use. Typically, we perform our data mining experiments on training data from 1997-2002 and use a test set consisting of data from 2003. Thus, we have taken special care to ensure that the training and test sets are properly formulated and accord with planned use.

2.3 Running the data mining software

An individual run of the C5.0 software has to be made for each target variable, for each forecast time and for each site. Thus, three target variables (flight category, ceiling and visibility), five forecast times (2h, 3h, 6h, 9h, 12h) and 10 research sites, represents 240 runs of the C5.0 software. C5.0 runs on our METAR data sets can take up to 10 minutes so 240 runs can take up to 40 hours. In order to complete the runs in an effective matter, the processing has to be automated. The Python programming language is an excellent choice for doing such processing. As a result, we created a collection of scripts to automatically prepare and organize the data, run the C5.0 software and then score the data mining results against persistence. When doing such batch processing, the configuration files used for the processing are saved so that each run is completely reproducible. We have designed the software in a modular fashion so that the configuration files encapsulate the data format of the input files. A change in the data format can then be handled entirely in the configuration files.

2.4 Comparing the data mining results against persistence

The last stage of the data mining process is one of evaluation. In this case, one is concerned with how a particular algorithm is performing against persistence. We implemented a variety of skill scores including PODY, PODN, FAR, CSI, bias, Pierce and Heidke scores for comparison purposes. Decision tree scores as well as persistence scores are generated for the same test set for comparison purposes.

3 Auxiliary Forecast Information

In our data mining experiments we have found that including in our training set additional information with regard to the future meteorological situation can boost performance significantly. In particular, suppose one has access to an independent forecast system which can generate a good estimate of future temperature or dew point and thus identify some key elements of the future weather regime. One could then incorporate such information in the data mining process discussed above for forecasting ceiling, visibility and flight category.

Along these lines, we have taken a perfect prog approach during training and have tacitly assumed that our external forecasting system is a perfect prognosticator. The next section presents an abbreviated listing of results. In real practice, there is no perfect forecasting system so we are in the process of assessing how sensitive the data mining is to the accuracy of the auxiliary forecast information.

4 Some Results

Tables I, II and III below present three different data mining experiments using C5.0 and compare the Peirce skill scores of C5.0 versus persistence. The data set consists of hourly METAR observations at Atlanta and Seattle from January 1997 through June 2004. The test set consists of all observations made in 2003. The training set is the complement of the test set. A small time gap was introduced between the training and test sets to guarantee that the observations in the test set were not close in time to those in the training set. Note that including 2004 in the training set was a matter of convenience; test results using 2004 to augment 1997-2002 are not significantly different from results where 2004 is excluded from training.

In Table I, we did not use any perfect prog forecast information. Even in this case, C5.0 shows improvement over persistence at Atlanta at 2h and 4h. The results at Seattle are not as good as persistence except at 6h.

In Table II, we include temperature and dew point perfect prog fields. Note the improvement in C5.0 over persistence at both Atlanta and Seattle in comparison with Table I.

Finally in Table III, we include temperature, dew point, sky cover and wind speed fields. Performance is enhanced even further over results in Table II.

Table I. Peirce skill score values comparing C5.0 against persistence for flight category (without perfect prog input fields)

Station	C5.0	Persistence
Atlanta		
2 hour	0.727	0.68
4 hour	0.625	0.56
6 hour	0.587	0.464
Seattle		
2 hour	0.571	0.579
4 hour	0.409	0.411
6 hour	0.373	0.297

Table II. As in Table I but including temperature and dew point perfect prog input fields

Station	C5.0	Persistence
Atlanta		
2 hour	0.731	0.68
4 hour	0.698	0.56
6 hour	0.679	0.464
Seattle		
2 hour	0.602	0.579
4 hour	0.493	0.411
6 hour	0.464	0.297

Table III. As in Table II but including sky cover and wind speed perfect prog input fields

Station	C5.0	Persistence
Atlanta		
2 hour	0.783	0.68
4 hour	0.732	0.56
6 hour	0.759	0.464
Seattle		
2 hour	0.626	0.579
4 hour	0.536	0.411
6 hour	0.541	0.297

5 Incorporating Decision Tree Data Mining in an Operational CV System

5.1 Deciding on the data mining approach for each site

The first step in constructing an operational ceiling and visibility system utilizing data mining is to determine which algorithmic approach is the most advantageous. One needs to consider performance, algorithm run time and general algorithm requirements. For example, even though an algorithm outperforms its competitors, its run time or disk/memory utilization may be prohibitive. In addition to algorithm selection, one must also determine the appropriate set of training variables.

5.2 Generating rule sets for each site

Once the algorithm and training sets are identified, one needs to generate rule sets for each variable (ceiling, visibility and flight category), each forecast hour (2h, 3h, 6h, 9h, 12h) and each site. This will typically involve a significant amount of processing so it's important to have a complete plan in place prior to beginning the processing.

5.3 Applying the rule sets to observation and forecast data

Once the rule sets have been generated, one needs to apply the rules to current observations in addition to forecasted temperature, dew point, etc. in order to generate the ceiling, visibility and flight category forecast for the different forecast times at all sites. We use the Local Data Manager (LDM) from Unidata to acquire the METAR observations and then decode the observations into daily site files as mentioned above. We also run an independent forecast system, http://www.rap.ucar.edu/projects/dicast, utilizing a number of forecast models including GFS, ETA, ECMWF, etc. in order to generate the forecast information for variables such as temperature, dew point and so forth.

5.4 Amalgamating the site forecasts into site forecast files

Once the site forecasts have been generated they are then recorded in site forecast files. This is useful for both analysis and verification. A gridded national ceiling and visibility product can be created by interpolating the forecasts over the CONUS.

6 Future Directions

There are a number of future directions which require further investigation. In particular, we are interested in combining information from multiple sites that are close in proximity before performing the data mining process. Next there are many data mining techniques that we have not explored which could be potentially beneficial. Finally, improving the output of the auxiliary forecasting system would contribute to an improved ceiling and visibility product.

Acknowledgements

This research is in response to requirements and funding by the Federal Aviation Administration (FAA). The views expressed are those of the authors and do not necessarily represent the official policy or position of the FAA.

References

Breiman, L., Friedman, J.H. Olshen, R.A., & Stone, P.J. 1984: Classification and regression trees. Belmont, CA: Wadsworth International Group.

Hansen, B.K. and Riordan, D., 2003: Fuzzy case-based prediction of ceiling and visibility, 3rd Conference on Aritifical Intelligence, American Meteorological Society**.**

Herzegh, P.H., R. Bankert, B. Hansen, M. Tryhane, G. Wiener 2004: Recent progress in the development of automated analysis and forecast products for ceiling and visibility conditions, *20th Conf. on IIPS*, AMS, Seattle.

Peirce, C. S., 1884. "The Numerical Measure of the Success of Predictions," *Science*, 4, pp. 454-454.

Quinlan, J. Ross 1996: Boosting, bagging, and C4.5. *Proceedings, Fourteenth National Conference on Artificial Intelligence*.