Predictive Modeling of Forecast Uncertainty in the Route Availability Planning Tool (RAPT)*†

John Hayward‡, Ngaire Underhill and Richard DeLaura
Massachusetts Institute of Technology - Lincoln Laboratory

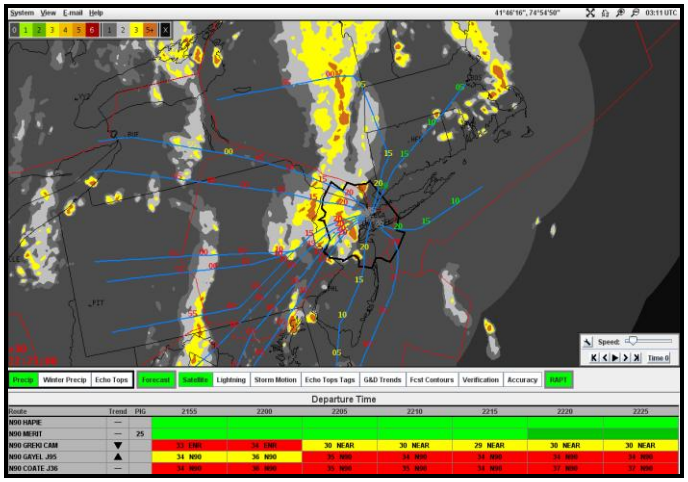
MIT Lincoln Laboratory has developed the Route Availability Planning Tool (RAPT), which provides automated convective weather guidance to air traffic managers of the NYC metro region. Prior studies of RAPT have shown high-accuracy guidance from forecast weather, but further refinements to prevent forecast misclassification is still desirable. An attribute set of highly correlated predictors for forecast misclassification is identified. Using this attribute set, a variety of prediction models for forecast misclassification are generated and evaluated. Rule-based models, decision trees, multi-layer perceptrons, and Bayesian prediction model techniques are used. Filtering, resampling, and attribute selection methods are applied to refine model generation. Our results show promising accuracy rates for multi-layer perceptrons trained on full attribute sets.

I. Introduction

THE Route Availability Planning Tool (RAPT) was developed at MIT Lincoln Laboratory to help air traffic managers of the NYC metro region determine the operational impact of convective weather on airspace departure routes [1]. RAPT creates route status timelines using weather forecasts from the Corridor Integrated Weather System (CIWS) [2]. Research conducted in [3] compared RAPT air traffic guidance using forecast weather to guidance based on true weather inputs. The study showed RAPT guidance was highly accurate. However, further refinements to status prediction and enhanced operational feedback are still desirable. The goal of this research is to create and evaluate different prediction modeling methods to anticipate the possibility of forecast misclassifications, given certain weather and airspace conditions.

II. RAPT Algorithm and CSI Scoring

RAPT creates route status timelines using CIWS forecasts of Vertically Integrated Liquid (VIL) and echo top heights. Each route timeline contains seven status tuples, representing forecasted route status for departures from time t in 5 minute binned intervals, ranging from (t + 5) to (t + 35). The route status tuple provides a discretized assessment of first-worst convective weather blockage on a route, binned into GREEN (route clear), YELLOW (route impacted), or RED (route blocked). A screenshot of the RAPT Display is shown in Figure 1. These RAPT route timelines are generated through four sequential algorithm processes.



Departure Time

Route	Trend	PIG	2155	2200	2205	2210	2215	2220	2225
N90 HAPIE	—		🟩	🟩	🟩	🟩	🟩	🟩	🟩
N90 MERIT	—	25	🟩	🟩	🟩	🟩	🟩	🟩	🟩
N90 GREKI CAM	▼		33 ENR	34 ENR	30 NEAR	30 NEAR	29 NEAR	30 NEAR	30 NEAR
N90 GAYEL J95	▲		34 N90	36 N90	35 N90	34 N90	34 N90	34 N90	34 N90
N90 COATE J36	—		34 N90	36 N90	35 N90	35 N90	34 N90	37 N90	37 N90

Figure 1 - RAPT display with route status timelines.

Forecasted VIL and echo top heights are combined using a Gaussian classifier to produce a regional Weather Avoidance Field (WAF) [4]. Each pixel in the WAF grid represents a pilot's likelihood of deviating around the weather at that point. At each point along a departure route, a heuristic scoring of airspace blockage of 0-100 is calculated from local WAF values. A discretized route status is derived from an algorithmic combination of route blockages and a regional airspace sensitivity field. Finally, an inertia algorithm is applied to smooth status variability from small blockage changes, and operational confidence statistics are generated.

Evaluating and refining forecast accuracy is crucial for automated air traffic management tools. Evaluations of RAPT operational accuracy may be performed by creating dual sets of departure status timelines: an operational "forecast" timelines set calculated from CIWS weather forecasts, and a "true"

[*] This work was sponsored by the Federal Aviation Administration under Air Force Contract No. FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

[†] CSC'10 - The 2010 International Conference on Scientific Computing, Las Vegas, NV. Topic key words: "Data mining", "air traffic control", "meteorology", "real-time computing", "prediction models".

[‡] C*orresponding author address*: John Hayward, MIT Lincoln Laboratory, 244 Wood Street, Lexington, MA 02420-9185.

timelines set using actual weather. The true weather data set is created retrospectively by replaying RAPT using CIWS weather archives. The comparison of forecast and true weather blockages can be used to evaluate the ability of RAPT to provide operationally valid impact data.

The evaluation of RAPT conducted in [3] used forecast and true weather CIWS inputs from the 2009 convective weather season. The study revealed a high rate of accurate classifications from forecast weather, shown in Figure 2. The number of levels that a forecast miscast a true weather status is represented as FC |Δ|. FC |Δ| = 0 means that the forecast predicted the true weather correctly. One-level errors (e.g., forecast GREEN that verified as true YELLOW) were infrequent and two-level errors (forecast GREEN that verified as RED and vice-versa) extremely rare.

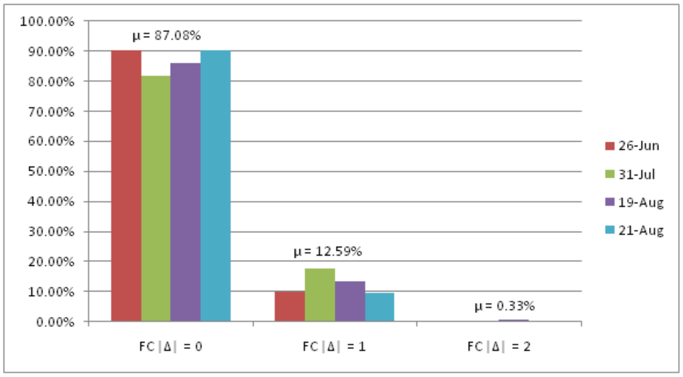


Figure 2 – RAPT forecast status accuracy.

Our research seeks to create a prediction model which can predict forecast misclassifications. Potential inputs to these models include measurements of forecast uncertainty. RAPT generates two measurements of forecast uncertainty in an operational airspace: a Modified Critical Success Index (mod-CSI) score [5] and a True Critical Success Index (true-CSI) score [3]. Both scores range from 0-100 and are calculated by comparing blockage values from the forecast and true weather grids.

The mod-CSI uses a region-wide calculation of blockage scores with a 40 minute parameterizable lag. In the ZNY airspace, scores are calculated for regions that encompass the NORTH, WEST, and SOUTH departure gates. The true-CSI model applies to RAPT route departure regions only, and requires an 80 minute time lag between forecast and verification.

III. Prediction modeling methods

The data set of N = 43008 tuple comparisons from the summer 2009 convective weather season was used for our prediction model research. These tuple comparisons represent a total of 6144 RAPT route timelines produced from CIWS forecasts generated at five-minute intervals. Eight major departure routes out of LaGuardia Airport (LGA) were used for the study. These departure routes are depicted in Figure 3.

The classification target of our models was Forecast Delta, a discretization of status misclassifications into "Overwarned" (e.g. true GREEN predicted as RED), "Underwarned", and "Correct". The six attributes initially used to predict Forecast Delta were:

- Date-Stamp – Date/time marker for tuple.
- Route ID – Name of departure route.
- Tuple Level – Departure time of tuple. Nominal divisions of t+5 to t+35 in 5 minute increments.
- Forecast Status – GREEN, YELLOW, or RED forecasted route blockage.
- Mod-CSI Score – CSI score of forecast uncertainty for departure gate region. Scores range 0-100, with lower scores indicting more uncertainty.
- True-CSI Score - CSI score of forecast uncertainty for departure route region. Scores range 0-100, with lower scores indicting more uncertainty.

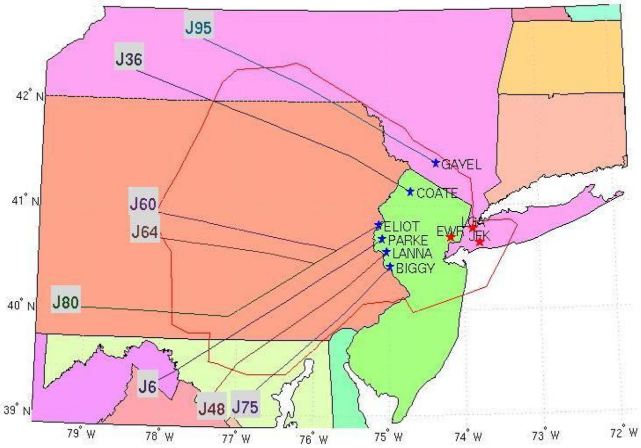


Figure 3 - LGA departure routes.

The following machine learning methods were utilized in our study. All model generation was done using the Waikato Environment for Knowledge Analysis (WEKA), version 3.7.1 [6]. Each prediction model was generated using 10-fold cross-validation. Comparisons of accuracy were performed using t-testing over 10 iterations, utilizing random seeding for sampling and probabilistic attributes.

Supervised Discretization – Filter to discretize numeric attributes into nominal values. Attributes are binned relative to changes in the target classification using the Minimum Descriptive Length (MDL) principle [7].

Correlation-based Feature Selection (CFS) – Attribute selection filter used to eliminate noisy and redundant features in data sets. Algorithm pares attributes down to subsets exhibiting high correlation to target class and low cross-correlation to remaining features [8].

Resampling – Instance filter to produce a randomized subset of a dataset. Sampling can preserve class distribution of sample, or create a sample with uniformly distributed class values. Uniform bias class resamples allow prediction models to train with equal weight to majority and minority target values [6].

OneR – Rudimentary prediction algorithm which uses single-attribute models to predict target classification. Also known as 1R or Learn-One-Rule [9].

J48 – Java implementation of C4.5 decision tree learning algorithm. Decision tree algorithms use information gain metrics to create logical conjunctions which represent classification values. C4.5 is an evolution of the basic ID3 decision tree algorithm, and accounts for missing values, decision tree pruning, and rule deviation [6, 10].

Bayes Net – Bayesian networks are directed acyclic graphs which represent conditional statistical relations for attributes of an entity. Bayes net predictors construct a networked probability model for classification using a specified network evaluator and network-space search function [6].

Naïve Bayes – Variant on Bayes Net predictors. Assumes statistical independence amongst attributes in predicting a target classification. Known for high characterization accuracy despite comparative algorithmic simplicity [9].

Multi-layer Perceptron (MLP) – A neural network classifier which uses backpropagation to train weights of network connections. The number of layers for each model varies by experiment. Attributes and numeric classes are normalized during execution [6].

IV. Initial results and analysis

Date-stamp and Route ID attributes were removed during the initial experiments, as they were found to "serialize" the target classes. The combination of Date-Stamp and Route ID created a nominative label of each instance. Prediction models built from these data sets overtrained to the specific contents of each label.

The remaining attributes were Tuple Level, Forecast Status, Mod-CSI Score, and True-CSI Score. Prior research in [3] provided some initial knowledge about these attributes. Time series analysis showed that both Mod-CSI and True-CSI are internally stable processes with high rates of autocorrelation. CSI scores also have low rates of cross-correlation and are mutually poor predictors of each other. A forecast status of RED or YELLOW is associated with higher rates of misclassification. The rate of underwarning is known to grow linearly with Tuple Level: status underwarning for the t+35 tuples occurs at about twice the rate of t+5 tuples.

The majority class value for the data set was Correct (p = 0.871). The Overwarned class value (p = 0.071) occurs at a slightly higher rate than Underwarned (p = 0.058). The distribution of Forecast Delta values is shown in Figure 4. Black denotes Correct, dark-gray denotes Underwarned, and light-gray denotes Overwarned.

The attribute distributions are shown in Figure 5 through Figure 8. The Forecast Delta categorical colors are incorporated into the attribute histograms. Each histogram bar is colored by the distribution of the target class relative to the attribute value (e.g.: a histogram bar for Forecast Status = GREEN colored one-quarter light gray implies that a quarter of GREEN forecasts have an Underwarned classification).

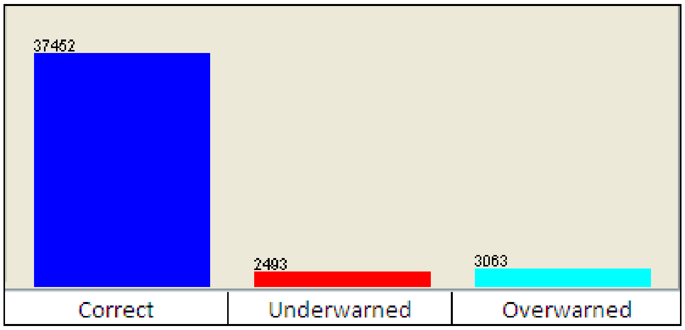


Figure 4 - Forecast delta class distribution.

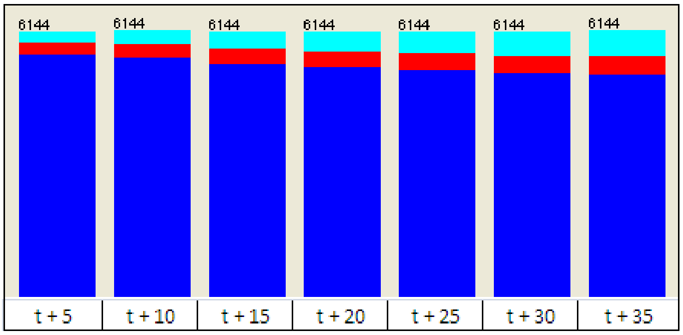


Figure 5 - Tuple level attribute distribution.

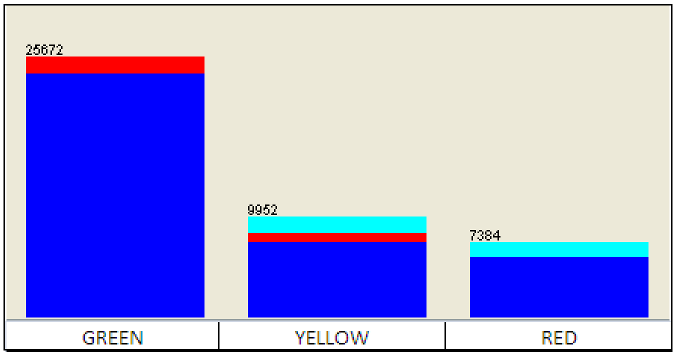


Figure 6 - Forecast status attribute distribution.

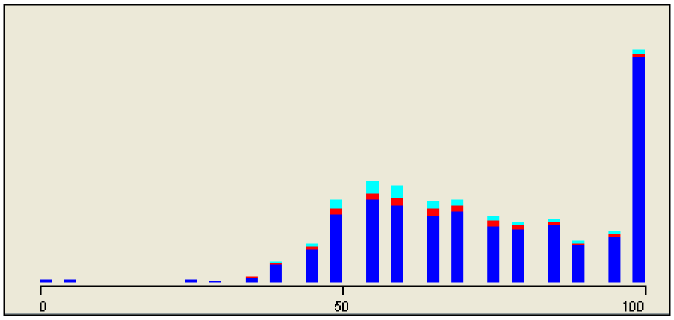


Figure 7 - Mod-CSI score attribute distribution.

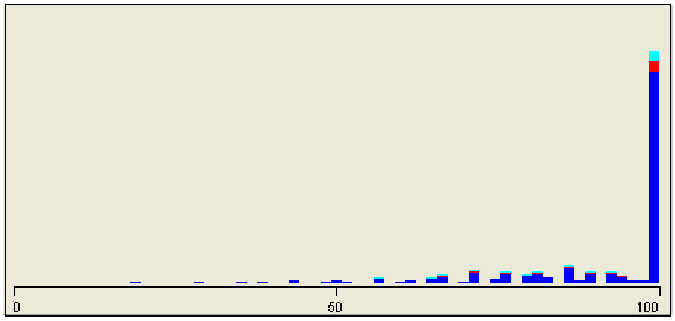


Figure 8 - True-CSI score attribute distribution.

V. Initial prediction model results

Supervised Discretization – Supervised discretization was applied to the two numeric attributes: Mod-CSI Score and True-CSI Score. The MDL-based binning of these attributes is illustrated in Figure 9 and Figure 10. Eight distinct bins were assigned to Mod-CSI, and six were assigned to True-CSI. Training on discretized attribute sets generally did not affect rates of prediction accuracy, although it did contribute to simplification of decision tree models.

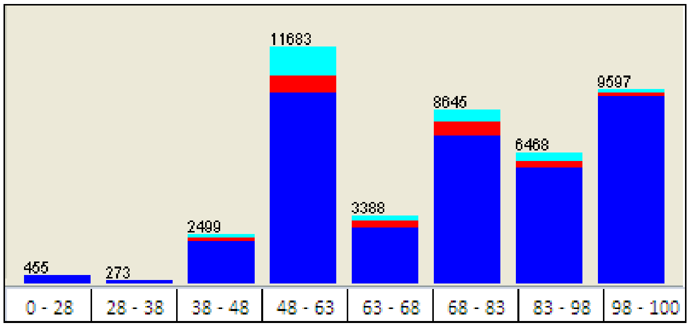


Figure 9 - Supervised discretization of mod-CSI scores.

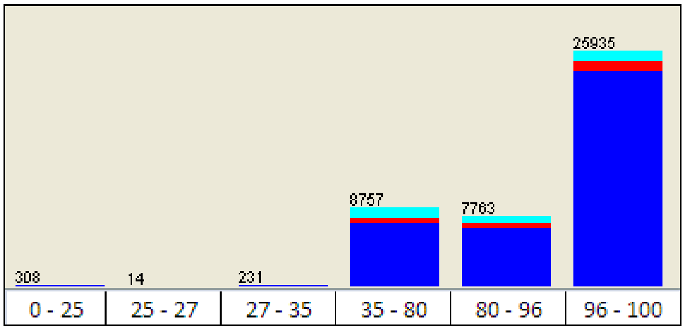


Figure 10 - Supervised discretization of true-CSI scores.

Correlation-based Feature Selection Filter Results – Feature selection chooses the attribute subset with the lowest internal cross-correlation and highest correlation to the target class. CFS filtering chose Forecast Status and Tuple Level for this subset. Prediction model generation was conducted on both filtered and original attribute sets. Exceptional differences in the resultant accuracy are noted.

Resampling Filter Results – To create prediction models that trained equally on majority and minority cases, a uniform bias resampling was applied to the target class. The resampled distribution of Forecast Delta is shown in Figure 11.

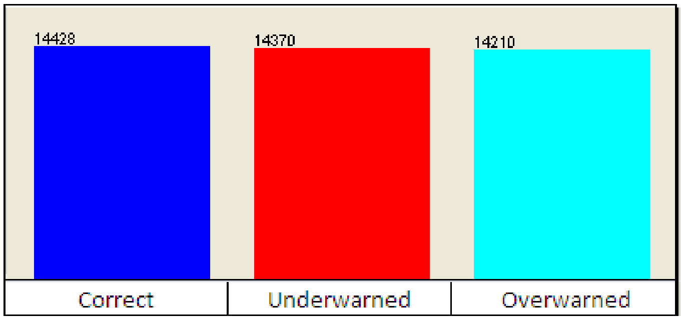


Figure 11 - Resampled forecast delta distribution with uniform class bias.

OneR Prediction Model – The OneR algorithm builds a classification model based on a single attribute. OneR identified Forecast Status as the best individual predictor of Forecast Delta. Accuracy rate for model was p = 0.546. This model's predictions of Forecast Delta based on Forecast Status are:

Forecast Status = G -> Correct
Forecast Status = Y -> Overwarned
Forecast Status = R -> Overwarned

J48 Prediction Model – The J48 decision tree model scored a high accuracy rate at p = .736. However, the decision tree was overtrained to the specific contents of each data instance. The J48 algorithm produced a model that essentially contained serialized outputs, similar to models trained on Date-Stamp and

Route ID. The decision tree was also structurally cumbersome, with 12 layers, 1335 nodes, and 771 leaves.

A more concise decision tree was generated from the attribute filtered set. This decision tree had an accuracy rate of p = 0.562. The predictions of Forecast Delta based on Forecast Status and Tuple Level for this decision tree model are:

Forecast Status = G ->
| Tuple Level = t + 1 -> Correct
| Tuple Level = t + 2 -> Correct
| Tuple Level = t + 3 -> Correct
| Tuple Level = t + 4 -> Underwarned
| Tuple Level = t + 5 -> Underwarned
| Tuple Level = t + 6 -> Underwarned
| Tuple Level = t + 7 -> Underwarned
Forecast Status = Y -> Overwarned
Forecast Status = R -> Overwarned

Bayes Net/Naïve Bayes Model – The Bayes Net network-space search concluded statistical independence amongst attributes produced models with the most accurate target classifications. This statistical relationship between attributes produced Bayes Nets which structurally match the Naïve Bayes model. The accuracy rate for this Bayes Net is p = 0.599, and p = 0.561 using the attribute filtered set.

The full attribute Bayes Net is illustrated in Figure 12. Note that the edges in this network depict conditional probability relationships, not the logical progression of a data point classification. Bayes Nets classify new data points using the *maximum a posteriori*, or most likely hypothesis, from the attribute probability distributions. Sample probability distributions for Tuple Level and Forecast Status are shown in Table 1 and Table 2.

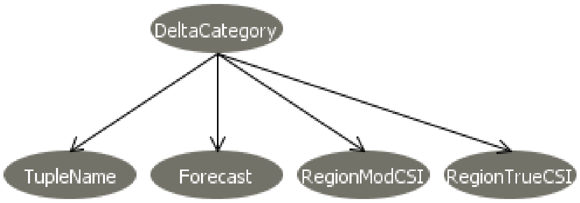
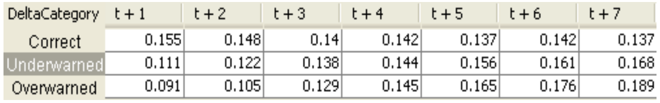


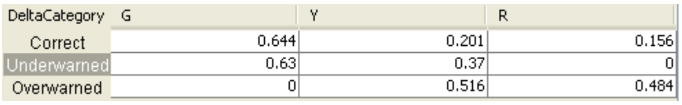
Figure 12 - Bayes Net/Naïve Bayes prediction model.

Table 1- Bayes Net - Probability Distribution for Tuple Level



DeltaCategory	t + 1	t + 2	t + 3	t + 4	t + 5	t + 6	t + 7
Correct	0.155	0.148	0.14	0.142	0.137	0.142	0.137
Underwarned	0.111	0.122	0.138	0.144	0.156	0.161	0.168
Overwarned	0.091	0.105	0.129	0.145	0.165	0.176	0.189

Table 2 - Bayes Net - Probability Distribution for Forecast Status



DeltaCategory	G	Y	R
Correct	0.644	0.201	0.156
Underwarned	0.63	0.37	0
Overwarned	0	0.516	0.484

Multi-layer Perceptron Model – The Forecast Delta MLP was trained on ten folds of 500 epochs using a learning rate = 0.3 and momentum = 0.2. The generated network contained 12 input nodes, 7 sigmoid nodes, and 3 output nodes. A visualization of the MLP is shown in Figure 13. The final predictor scored a high accuracy rate with p = .653.

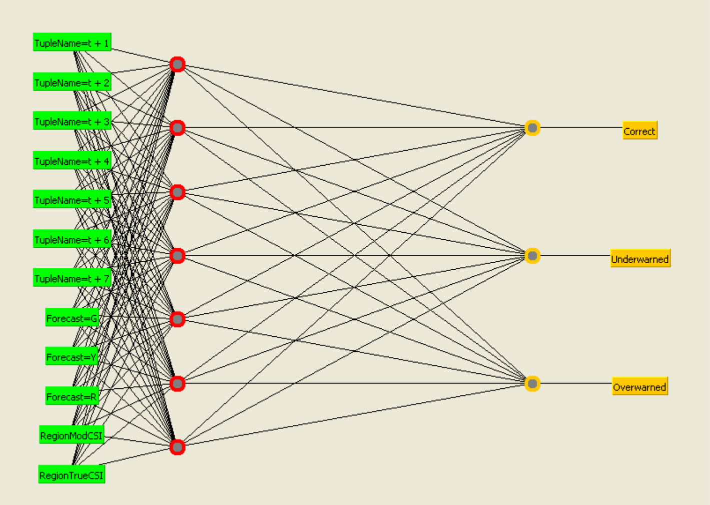


Figure 13 - Forecast delta MLP predictor.

The performance of this MLP was compared the other non-serialized prediction models. The higher performance of the MLP was confirmed as statistically significant via 10-fold t-testing with α = 0.05. This makes the MLP the best performing non-serialized prediction model in the study.

High performance is not the only promising aspect of the Forecast Delta MLP. A software implementation of the MLP can also continually train through backpropagation. This allows the MLP to adapt dynamically to larger data sets and varied domains. MLP retraining is also comparatively efficient. Many prediction models must be rebuilt from the full training sets with the receipt of new data. MLPs can incorporate new information into its connection weights iteratively, without requiring a full network retraining.

VI. Future work

Additional classification methods may be examined in future work. Refinements of promising high-accuracy methods, including the current multi-layer perceptrons, will be performed. These refinement techniques may include an expanded use of attribute selection, discretization, and cost matrix specification. Within the RAPT software, the application of implemented Forecast Delta prediction models will be examined. These applications may include refined status prediction and various forms of enhanced operational feedback.

References

[1] DeLaura, R., M. Robinson, R. Todd, and K. MacKenzie: "Evaluation of Weather Impact Models in Departure Management Decision Support: Operational Performance of the Route Availability Planning Tool (RAPT) Prototype", 13th Conference on Aviation, Range, and Aerospace Meteorology, AMS, New Orleans, LA, 2008.

[2] Klingle-Wilson, D., J. Evans: "Description of the Corridor Integrated Weather System (CIWS) Weather Products", MIT Lincoln Laboratory Project Report ATC-317, August, 2005.

[3] Hayward, J., N. Underhill, M. Matthews, R. DeLaura: "Operational Scoring of Forecasts of Convective Weather Impacts in the Route Availability Planning Tool (RAPT)", 15th Conference on Aviation, Range, and Aerospace Meteorology, AMS, Atlanta, GA, 2010.

[4] DeLaura, R., M. Robinson, M. Pawlak, and J. Evans: "Modeling Convective Weather Avoidance in Enroute Airspace", 13th Conference on Aviation, Range, and Aerospace Meteorology, AMS, New Orleans, LA, 2008.

[5] Matthews, M.: "Measuring the Uncertainty of Weather Forecast Specific to Air Traffic Management Operations", PD1.4, Aviation, Range and Aerospace Meteorology Special Symposium on Weather-Air Traffic Management Integration, AMS, Phoenix, AZ, January, 2009.

[6] Frank, E., I. Witten: "Data Mining: Practical Machine Learning Tools and Techniques." Morgan Kaufmann, 2nd edition, 2005.

[7] Fayyad, U., K. Irani: "Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning", In IJCAI, pages 1022–1029, 1993.

[8] M. Hall: "Correlation-Based Feature Selection for Machine Learning", PhD thesis, Waikato University, Department of Computer Science, 1998.

[9] T. Mitchell: "Machine Learning", McGraw-Hill Science/Engineering/Math, 1st edition, 1997.

[10] J. Quinlan: "C4.5: Programs for Machine Learning", Morgan Kaufmann, Los Altos, California, 1993.