The 2008 Artificial Intelligence Competition Data: Source and Characteristics

Kimberly L. Elmore and Michael B. Richman

1. Introduction

The training and testing data used in the 2008 AMS AI competition come form the WInter Hydrometeor Ground Truth Experiment (WHCGT), recently renamed to the Winter Precipitation Identification Near the Ground, or W-PING, Project.

This experiment began in the Winter of 2006-2007 as a way to approach the problem of developing a Winter Hydrometeor Classification Algorithm (HCA) for the KOUN Polarimetric testbed radar (Scharfenberg et al. 2005).

Prior to this, the existing HCA had been developed with only warm season convection in mind. There was some question as to how well this HCA would perform in cold season applications. Thus the WHCGT (now W-PING) project was conceived.

2. The Data

To determine how well the NSSL HCA (Scharfenberg, et al. 2005) performed in cold season precipitation, an experiment was launched that uses the public as observers of precipitation type. Based solely upon press releases and newscasts, the WHCGT Experiment was initiated in November 2006. The idea was to collect public observations of precipitation type and enter those observations into a data based using a secure web form.

Fortunately, the Winter of 2006-2007 proved very active, with an unusually high frequency of ice and snow events. In Oklahoma. The Oklahoma public is usually attuned to weather events and while Winter weather is no novelty in Oklahoma, it is rare enough that it sparks a great deal of public interest when it occurs. The WHCGT Experiment was launched on the eve of a well-forecast and heavily publicized winter storm and the public response was exceptionally high. Competition data come from three major events: 29 Nov through 30 Nov 2006, 11 Jan through 14 Jan 2007, and 19 Jan through 20 Jan 2007.

A web site contains information about the experiment, guidance about how to distinguish various winter precipitation types, a status message, and a web form that could be filled in to provide observations of precipitation type. That page may be found at http://www.nssl.noaa.gov/projects/winter/. There was no need for the public observers to “sign up” and, in fact, all of the provided information was purposely kept anonymous.

The public was asked to distinguish between the following categories: rain, drizzle, freezing rain, freezing drizzle, ice pellets (sleet), graupel, snow, hail, and none, all within a 150 km radius from the KOUN radar. As a practical matter, a cold-season HCA must be able to distinguish between frozen, liquid, and no precipitation, so the above categories were amalgamated into the three used in the competition. Freezing rain and freezing drizzle were combined with rain and drizzle, and classed as “liquid.” Snow, ice pellets (sleet), graupel and hail were all combined into “frozen,” while “none” was retained as is.

The observed precipitation type data are quality controlled using rather broad criteria. If an observation is clearly inconsistent with nearby observation in time and space, e.g., observations of “hail” in the midst of “snow” are removed. Observations well outside of the project area have been removed as have been obvious duplicate entries.

The KOUN polarimetric testbed radar operated during most events. The KOUN radar differs from standard weather radar in that it transmits in both horizontal and vertical polarization; standard weather radars use only horizontal polarization. KOUN collects the familiar standard radar parameters – horizontal reflectivity, Z_h, and radial velocity V_r – along with differential reflectivity, Z_{dr}, differential phase shift, ϕ_{dp}, specific differential phase shift, k_{dp} (the radial derivative of ϕ_{dp} and so independent of the initial phase shift), and correlation coefficient between horizontal and vertical polarization reflectivity, ρ_{hv}. Each of these parameters are affected id different ways by the nature of the hydrometeors that scatter the radiation back to the radar receiver.

Among the things that affect the returned signal are the shapes of the hydrometeors and their composition (whether liquid or ice) and their density. Thus the composition and 2-dimensional size distribution, along with number concentration all define the polarimetric variables observed by the radar.

Around each ground observation, radar data for each parameter is averaged over a 5 x 5 (range by azimuth) kernel centered on each ground observation. Only observations associated with radar data between 0.3 km and 1.2 km AGL are used. Within that height range, only the lowest scan is chosen. All data are filtered to remove observations within ground clutter.

For the three main events, about 2650 observations were logged. After the rudimentary QC,

about 2500 remain. Of these 2500, 1573 meet all the other criteria stated earlier. It is important to note that these data are unique in that no additional such data exist from any source. A few other minor events occurred, but the 2007-2008 season was marred by radar problems, and little data was gathered.

In 2008, the project was renamed Winter Precipitation Identification Near the Ground (W-PING) project, and very little suitable weather has occurred so far.

The testing data is generated by sampling, without replacement, from the full data set. The testing data constitutes 30% of the full data set, leaving the other 70% for training. No attempt was made to "balance" the proportion of the various categories. The training data contain 58.3% frozen, 28.2% liquid, and 13.5% none, while the testing data contain 56.7% frozen, 32.9% liquid, and 11.3% none."

3. Statistics and Skill Scores

We choose Peirce's Skill Score (PSS, Peirce 1884) for determining the winners. The PSS is equitable and so not subject to hedging or gaming (creating forecasts that do not the true beliefs of the developer). However, late in the competition (after classifiers has been developed and scores returned to participants) an error was discovered on the web page used as a reference for the PSS. The correct score is:

$$PSS = \frac{\sum_{i=1}^{I} p(y_i, o_i) - \sum_{i=1}^{I} p(y_i)p(o_i)}{1 - \sum_{j=1}^{I} [p(o_i)]^2},$$

where $I = J$ = number of categories. The erroneous score had f_i in the denominator instead of o_i.

The erroneous formulation leads to a score that is easily hedged and typically provide higher values that does the correct PSS formulation. All scores for ranking are computed using the correct formulation of the PSS.

4. Conclusions and Comments

Unlike last year, bootstrap resampling (using 1000 replicates) is performed on the submissions to ascertain a measure of uncertainty in the PSS generated by the various classifiers. Here, 95% confidence bounds are placed on the PSS based on bootstrap resampling (Efron and Tibshirani 1993). Based on these resampling estimates, no classifier is statistically different from any other, except for the Gordon Strategy 3 entry. These results are summarized in Table 1.

Table 1: PSS confidence Estimates

	2.5%	Mean	97.5%	PSS
Sullivan	**0.271**	**0.354**	**0.440**	**0.355**
Goosseart	**0.243**	**0.321**	**0.421**	**0.321**
Pocernic	**0.236**	**0.312**	**0.404**	**0.312**
Lak	**0.236**	**0.298**	**0.384**	**0.298**
Pocernic 1	**0.220**	**0.297**	**0.387**	**0.2947**
Gordon	**0.217**	**0.295**	**0.386**	**0.2945**
Gordon 1	**0.203**	**0.291**	**0.377**	**0.291**
McCandless	**0.204**	**0.284**	**0.373**	**0.285**
Lak 1	**0.181**	**0.245**	**0.345**	**0.266**
Armando	**0.154**	**0.234**	**0.326**	**0.236**
Gordon 3	**0.000**	**0.007**	**0.038**	**0.007**

Note that for all classifiers save for the Gordon 3, all confidence intervals overlap the mean. This is a good indication in itself that there is no significant difference between the classifiers (Ramsey and Schafer 2002). However, a more sensitive and quantitative test may be had with a permutation test (Efron and Tibshirani 1992).

A pairwise permutation test between all possible classifier pairs, using 5000 permutations, shows that there is indeed no statistical difference between the different classifiers at the 95% level.

Do these results mean that there is really no difference between the various methods employed by the contestants? No, but it does mean that there is enough variability within the available sample of 363 cases that differences cannot be discerned at the 95% level. Thus, while rankings have been declared, these rankings are not *statistically* supportable.

5. References

Efron, B. and R. J. Tibshirani, 1993: *An Introduction to the Bootstrap,* Chapman and Hall, 436 pp.

Peirce, C.S., 1884: The numerical measure of the success of predictions. *Science,* **4,** 453–454.

Ramsey, F. L. and D. W. Schafer, 2002: *The Statistical Sleuth.* Duxbury Press, 742 pp.

Scharfenberg, K. A., D. J. Miller, T. J. Schuur, P. T. Schlatter, S. E. Giagrande, V. M. Melnikov, D. W. Burgess, D. L. Andra, M. P. Foster, and J. M. Krause, 2005: The Joint Polarization Experiment: Polarimetric radar in forecasting and warning decision making. Bull. Amer Meteor. Soc, 20, 775–788.