Application of Neural Net Models to classify and to forecast the observed precipitation type at the ground using the Artificial Intelligence Competition data set.

A. Pelliccioni (*), R. Cotroneo (*), F. Pungì (*)
(*)ISPESL-DIPIA, Via Fontana Candida 1, 00040, Monteporzio Catone (RM), Italy.

Introduction

In our work, neural networks (NN) methods have been developed in order to classify and forecast observed precipitation type at the ground. Neural networks have become a popular tool for solving complex problems in diverse domains, such as environmental field.
NN can work as *universal approximators* of non-linear functions[1];[2] and, consequently, can be used in assessing the dynamics of such systems. Among the complex systems, neural networks (NN) have become a useful tool either where correct phenomenological models are not available or when uncertainty in processes reproduction and input data complicates the application of a deterministic modelling. The NN parameters were obtained by a training procedure based on the use of an efficient unconstrained minimization algorithm.

For the NN simulations, the optimization of the input patterns is a strategic point. Data optimization processes are necessary in order to select patterns and variables having a high meaning for explaining data variability related to NN model performance in terms of Hanssen and Kuipers discriminant (true skill statistic, Peirces's skill score).
Therefore, data need to be judiciously prepared before they have to be supplied to a neural network for the training phase.
In particular, we applied a K-means clustering algorithm, used to cluster weather variables measured by the Polarmetric Radar, to select best patterns to improve the net performance in the training phase. The NN training patterns consist of a centroid related to the clusters. This is the real novelty of our approach. All people use cluster techniques to simplify the dataset. The number of clusters could be linked with the main information inside the patterns.

Dataset description

Data used in our simulations come from AI competition datasets to classify the observed precipitation type at the ground into one of three categories: liquid, frozen, or solid.
The design of input data consists of 16 variables.
We applied some transformations related to variables space to improve performance and speed training and in order to take into account the following questions:

- to reduce periodical variables (as for example tilt angles, azimuth angles) using sine and cosine components. We also used the Cartesian coordinates expressed in kilometres for latitude and longitude.
- original wind variable concerns the Cartesian components as measured at sites. We transformed these components in polar coordinates, considering in addition the absolute value of wind speed (Uspeed).

Further,

- we rescaled the following variables in the given dataset: relative humidity (in decimal), wimd components (divided for 10)

- standardized: latitude, longitude, temperature, absolute value of wind speed, freezing level, cos(tilt), sin(tilt), range from radar, cos(azimuth), sin(azimuth), height of radar data, specific differential phase, reflectivity.

In order to optimise the main variables, we performed 22 NN tests, considering a total of 22 variables (see table 1). Each test consists of input variables, as shown in table 1, and using as classifier a NN model. NN is used only as a preliminary analysis to evaluate the contribution of each variable to the explained variability.

Table 1: Contribution of each variable

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
lat (in km)_std	X	X																				
lon (in km)_std	X		X																			
tmpc_std				X																		
Relh/100					X																	
urel/10						X	X															
vrel/10						X	X															
absolute value of wind speed						X																
frzl_std								X														
tilt									X	X		X										
sen(tilt)_std									X		X		X									
cos(tilt)_std									X		X		X									
range_std									X	X	X			X								
azimuth									X	X	X				X	X						
sen(azimuth)									X	X	X				X		X					
cos(azimuth)									X	X	X				X		X					
hgt_std									X	X	X							X				
Zdr																			X			
RhoHV																				X		
Kdp_std																					X	
Z_std																						X
Correctly Classified Instances	61.9	61.39	58.3	62.8	60.2	63.7	63.9	63	63.5	61.3	62.8	58.32	59.5	59.7	59.7	58.3	58.3	58.3	58.3	61.8	58.3	60.2

The last row in the table indicates the percentage of correctly classified patterns for each input variable. We observed that freezing level, temperature, latitude and cross-correlation coefficient are very significant for explaining the relation with target variable (>0.60%).

After data pre-elaboration, our dataset is composed by 19 variables: latitude, longitude, temperature, relative humidity, wind components, absolute value of wind speed, freezing level, cos(tilt), sin(tilt), range from radar, cos(azimuth), sin(azimuth), height of radar data, differential reflectivity, cross-correlation coefficient, specific differential phase, reflectivity and at last the target variable.

Further, we examined bivariate frequency distributions of each selected variable versus target (observed precipitation type). The meaningful observed distributions concerned the reflectivity, temperature and the wind components (see Fig1-3).

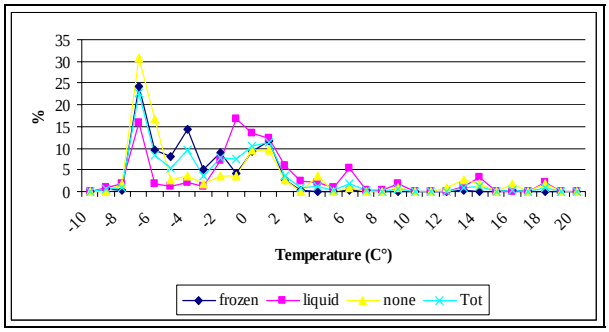

Fig. 1: Temperature (C°)

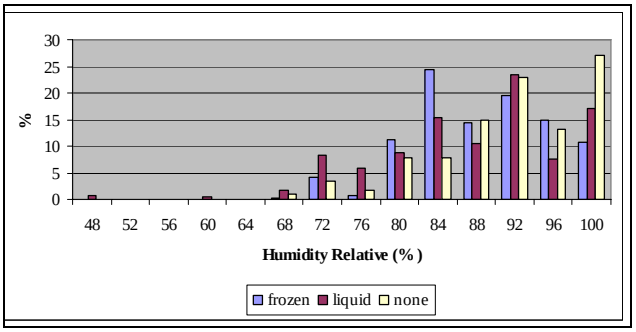

Fig. 2: Relative Humidity (%)

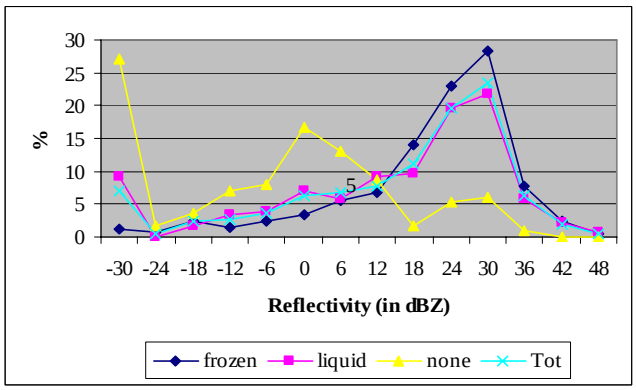

Fig. 3: Reflectivity

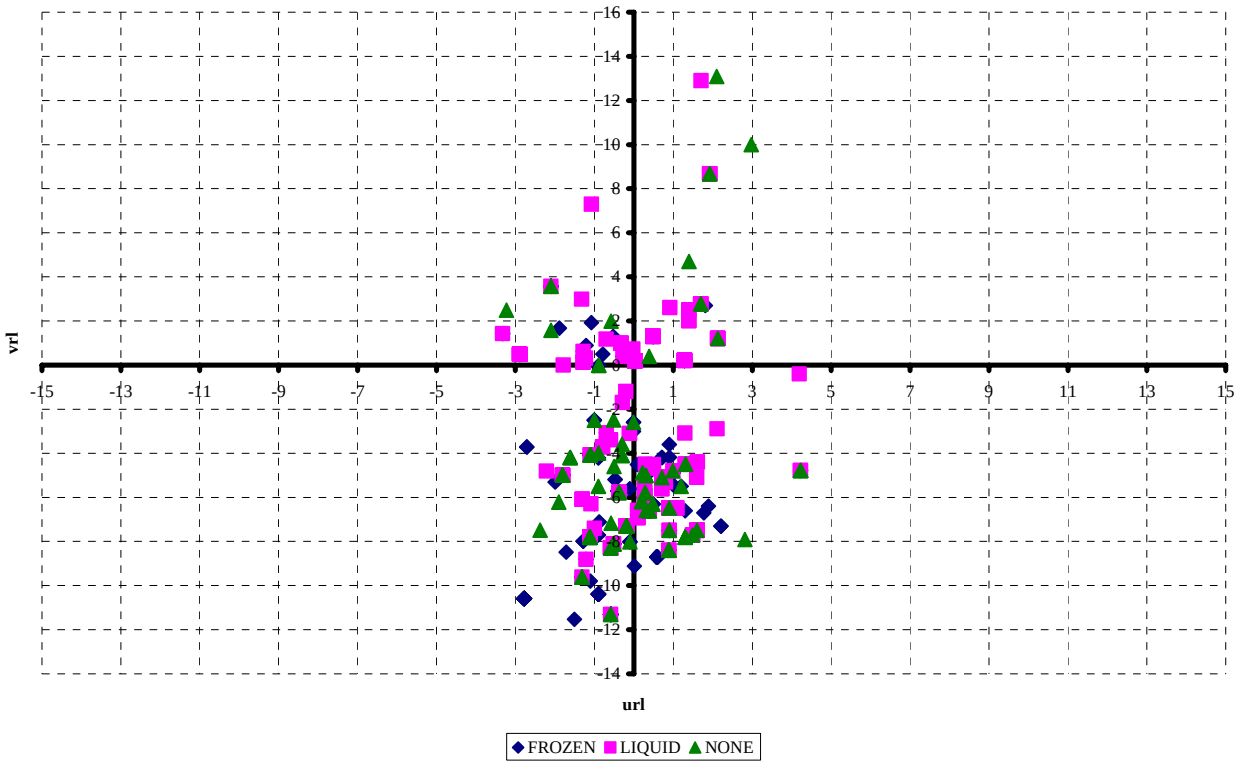

Fig. 4: Wind Components

Interesting information derived from the wind components versus target variable (fig. 4). In fact, when wind speed is high (3-12 m/s) the predominant direction is NNW and NNE; with lower wind speed value (up to 3 m/s) the predominant direction is SSE and NW.

Methodologies

We applied a novel approach that consists of an optimizing combination of Neural Net architecture in conjunction with cluster analysis[3] that is an important technique used in discovering some inherent structures present in data and does not require further assumptions or a priori knowledge.
Clustering of data is a method by which large sets of data are grouped into clusters of smaller sets of similar data. The purpose for the partitioning of a dataset of objects into k separate clusters is to find clusters whose members show an high degree of similarity among themselves but dissimilarity with the members of other clusters.
A clustering algorithm attempts to find natural groups of data based on some similarities and also finds the centroid of a group of data sets.

In our work, we utilize cluster analysis techniques in not conventional way. Usually, clustering is used to simplify and to select information from experimental dataset. Aim of this technique is obtaining only not redundant patterns. Here, we applied cluster analysis only to select the patterns that optimize the NN training phase. We aren't interested to the representative information typical of clustering, but in the percentage of well selected patterns to use during the training.

To determine clusters membership, we applied the K-means algorithm[4] that is one of the most non-hierarchical methods used for data clustering. The input patterns are clustered separately in such a way to produce a new dataset composed with the centroid of each cluster. Centroids are applied to the training dataset of NN to reduce the amount of patterns to be learned for the neural network, by automatically selecting an optimal set of patterns. Utilizing cluster analysis, it is possible to obtain more accurate and meaning results because the NN input data have an high reliability.

During NN training phase, a large number of input variables can provide an accurate description of the problem being considered, but yields over-parameterized model and requires more computational processing time and often more data for an effective understanding of the relationship between inputs and outputs.

For this work, the most suitable NN architectures at recognizing patterns are considered to be the Multi Layer Perceptron (MLP)[5];[6];[7];[8].

Results and Discussions

The WEKA software[9] was used to apply Multilayer Perceptron neural network, with a back propagation feed-forward learning algorithm, as classifier for our dataset (see table 2). We selected 14 hidden layers.

Table2: Neural Networks parameters

Hidden Layers	14
eta	0.01
momentum	0

NN technique is applied to validate the method by using 60% of clusters (centroids) from original training dataset (aicomp_training2008.csv) as the training set and the aicomp_testing2008.csv as the testing.

We experimented with various subsets of the input variables to achieve the best result concerning the classification and prediction of observed precipitation type at the ground. Table 3 and table 4 show the percentage of correctly classified instances using NN.

The adopted NN model shows good performances. Our results show, a multi-category form of the Pierce Skill Score index for observed precipitation type ranged from 0.83 to 0.88 in training phase.

Fig. 1: Pierce Skill Score index (PSS)

$$PSS = \frac{NC - E}{N - E^*}, \text{where:}$$

$$NC \text{ (number correct)} = \sum_{i=1}^{k} A_{ii} \qquad E = \sum_{i=1}^{k} \frac{C_i R_i}{N} \qquad E^* = \sum_{i=1}^{k} \frac{R_i R_i}{N}$$

The Peirce Skill Score is able to measure skill without being perturbed by the base rate. Thus, the PSS is ideally suited to measure the joint distribution, i.e. to display the potential forecast accuracy itself.

Further, the confusion matrix shows how many patterns of each class have been assigned to each class.

Table 3: PSS Neural Network

Classified	frozen	liquid	none	tot
frozen	483	8	3	494
liquid	21	215	3	239
none	18	14	82	114
tot	522	237	88	**847**

Correctly Classified Patterns	780	92,09 %
Incorrectly Classified Patterns	67	7,91 %
Mean absolute error	0.0747	
Root mean squared error	0.2109	
Relative absolute error		19.9103%
Root relative squared error		48.7166%

PSS	**0,834**
1-PSS	**0.166**

The goodness of the observed precipitation classified is strongly dependent by selection of the patterns during the training phase and the results are related with the statistical distribution of the Input/Output data set.

Using the centroids obtained form the clustering phase, we construct a new training set.
We choose different numbers of clusters needed to the NN learning: 20%; 50%, 60% and 70% of original training dataset. At the end, we selected 60% clusters of the original training dataset. The resulting set is presented to the neural network described above (table 4).

Table 4: PSS Neural Network – K-means (60%)

Classified	frozen	liquid	none	tot
frozen	290	5	1	296
liquid	7	135	1	143
none	8	7	53	68
tot	305	147	55	507

Correctly Classified Patterns	478	94.28 %
Incorrectly Classified Patterns	29	5.72 %
Mean absolute error	0.0584	
Root mean squared error	0.176	
Relative absolute error		15.5859 %
Root relative squared error		40.6827 %

PSS	**0.882**
1-PSS	**0.118**

Our results seem suggest the right way to optimize NN model by using cluster analysis in training phase. However, more sophisticate applications can be done.

Conclusion

Our research shows that the skill of the NN to capture information inside the data is highly dependent by the preliminary study of patterns.
The generalization capacity of the net to classify observed precipitation type at the ground, are connected by the essential information inside the dataset and this information is not necessarily regularly distribute inside all patterns.
Simulations based on cluster analysis results demonstrate that this method is feasible and effective, resulting in a substantial reduction of data input requirement.
Further, NN clustering revealed more subtle variations among patterns and an optimal AI technique for solving complex problems.

References

[1] Gardner, M.W., Dorling, S.R., 1999. Neural network modelling and prediction of hourly NOx and NO2 concentrations in urban air in London. Atmospheric Environment 33, 709-719.
Gardner, M.W., Dorling, S.R., 2000. Statistical surface ozone models: an improved methodology to account for non-linear behaviour. Atmospheric Environment 34, 21-34.

[2] Abdul-Wahab, S.A., Al-Alawi, S.M., 2002. Assessment and prediction of tropospheric ozone concentration levels using artificial neural networks. Environmental Modelling & Software 17, 219-228.

[3] E. Bradley "Bootstrap Methods,1997. Another Look at the Jackknife" Stanford University

[4] J. B. MacQueen, 1967. Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1:281-297

[5] Abdi, H., 1994. Les re´seaux de neurones. Presse Universitaire de Grenoble.

[6] Fausett, L.,1994. Fundamentals of Neural Networks. In: Architectures, Algorithms and Applications. Prentice Hall, Englewood Cliffs, NJ 07632.

[7] Bishop, C.M., 1995. Neural Networks for Pattern Recognition. Clarendon Press, Oxford.

[8] Ripley, B.D., 1996. Pattern Recognition an d Neural Networks. Cambridge University Press.

[9] I. H. Witten, E. Frank, June 2005. Data Mining: Practical Machine Learning Tools and Techniques.2nd Edition, Morgan Kaufmann Publishers.