On the Integration of Ensembles of Neural Networks: Application to Seismic Signal Classification

Yair Shimshoni and Nathan Intrator
School of Mathematical Sciences, Tel-Aviv University
Tel Aviv 69978, Israel
shimsh@math.tau.ac.il

Abstract

This paper proposes a classification scheme based on the integration of multiple Ensembles of ANNs. It is demonstrated on a classification problem, in which Seismic recordings of *Natural* Earthquakes must be distinguished from the recordings of *Artificial* Explosions. A Redundant Classification Environment consists of several Ensembles of Neural Networks is created and trained on Bootstrap Sample Sets, using various data representations and architectures. The ANNs within the Ensembles are aggregated (as in Bagging) while the Ensembles are integrated non-linearly, in a *signal adaptive* manner, using a posterior confidence measure based on the agreement (variance) within the Ensembles. The proposed *Integrated Classification Machine* achieved 92.1% correct classification on the seismic test data. Cross Validation evaluations and comparisons indicate that such integration of a collection of ANN's Ensembles is a robust way for handling high dimensional problems with a complex non-stationary signal space as in the current Seismic Classification problem.

Introduction

The lack of a-priori knowledge about the true underlying model of the data in many real-life problems, leads the practitioner to examine various sub-optimal estimators. Different estimators can exploit various types and sets of features, hence combining multiple estimators might yield better performance than the best single candidate.

In this paper we propose a classification scheme based on the integration of multiple Ensembles of Artificial Neural Networks ANN. It will be demonstrated it on a two class classification problem, in which Seismic recordings of *Natural* Earthquakes must be distinguished from the recordings of *Artificial* Explosions. Although being implemented here on a Seismic Classification problem, the presented scheme is general purpose and can be applied directly on any two class classification, while problems with more than two classes require only slight changes. We will first give some background on Seismic Classification and elaborate on the subject of Combining Multiple Estimators. Next, we will present the framework of our approach and formulate the Integrated Classification Machine (ICM). A posterior classification confidence measure will be discussed and a method for integrating the Ensembles of ANNs will be suggested. Then will follow the implementation details, namely, the pre-process of the seismic signals, the training and evaluation scheme and the results.

Seismic Signal Classification

In the recent years researchers have addressed the problems of Seismic Classification using various disciplines other than classic seismologic methods including Artificial Neural Networks (Dowla, Taylor, & Anderson 1990; Dysart & Pulli 1990). The vast majority of the recorded seismicity in most countries is artificial (i.e., man-made events like quarry blasts, mine explosions, etc.). The amount of recorded events have grown rapidly in the last decade (especially for weak events) due to the improvements of the seismometers, thus the work load in the seismic observatories have grown accordingly. As natural events are much more important for the analysis of the regional seismicity and for seismic hazard assessment, constructing a reliable automatic method for selecting the signals of the natural events is crucial for efficient seismic research.

Most of the works done so far on this classification problem concern with regional events vs. nuclear tests, rather than local events vs. conventional explosions at near distances. Many of the methods appear in the literature are based on geophysical parametric models (Dysart & Pulli 1990) which need explicit information to be extracted or estimated by the analyst, thus are very difficult to be fully automated. The signal space formed by the seismic waveforms is very high dimensional and when dealing with weak local events, one has to face also a low Signal-to-Noise ratio due to the weak signal energy. The non-homogeneous earth crust and the various source mechanisms of the seismic events cause the signal space to have a considerable complexity and variability, while the waveforms' distribution is usually non-stationary.

This work presents a model which can be applied to automated classification of seismic signals, including weak local events. The model is data-driven and does not require prior information or any intervention of a human analyst. It is designed to handle the non-stationary nature of the data and uses the frequency spectrum of the waveform recorded by a single seismometer to produce the class label of a *Natural* or *Artificial* event.

Combining Multiple Estimators

In the statistical and machine learning literature there are several methods for combining estimators and the questions involved with this topic mainly *what* types of estimators to combine and *how* to combine them, are recently getting considerable attention.

Mixture models like Adaptive Mixture of Experts (Jacobs *et al.* 1991) and Hierarchical Mixture of Experts (Jordan & Jacobs 1994), are based on the *divide and conquer* approach in which a mixture of experts compete to gain responsibility in modeling the output in a given input region. The system's output is obtained as a linear combination of the experts' output, where the weights are computed as a parametric function of the input. The underlying probabilistic model is based on the assumption of mutual exclusivity, i.e., a single expert is responsible for each data point. Usually in such models, the different experts are trained on a single data set simultaneously by minimizing a combined cost function and the final combination of the experts is determined by a gating module which is constructed in the same training session as well.

A more general framework for combining multiple estimators is "Stacked Generalization" (Wolpert 1992; Breiman 1992), where each estimator is trained with different sub-set of the data and the optimal combination is estimated using cross validation methods. It was shown that in order for the combination of experts to be optimal, the experts should be made as independent as possible (Krogh & Vedelsby 1995; Jacobs 1995; Meir 1994) and the optimal combination of the experts should be estimated in a robust way i.e., by averaging or cross validation techniques rather than by parametric estimation based on the same training data (LeBlanc & Tibshirani 1993). Combining multiple classifiers can eliminate the need to regularize over-fitted models as it reduces the variance of the combined estimator (Krogh & Vedelsby 1995).

A new method which produces an aggregated estimator using Bootstrap replicas of the training data is known as Bagging (Breiman 1994). It is reported to be useful whenever the estimator is un-stable i.e., when perturbing the training set can cause significant changes in the constructed classifier.

A Redundant Classification Environment

Since the search for an optimal classifier is strongly bound with the search for a suitable data representation. It is advisable to examine (and possibly use) more than a single representation, especially when there is no solid assumption about a unique underlying model.

As the desire is to maximize the information extracted from the data, we suggest to construct a Redundant Classification Environment, which allows for different data representations to supply a wide coverage of the effective feature space, using several modeling architectures. For example, when the input signals are waveforms, one can represent the data by multiple Time-Frequency or Atomic Dictionary decompositions, apply various smoothening methods, as well as use several estimating methods.

In the hierarchical scheme we will propose shortly, the Experts are Ensembles of ANNs, where each Ensemble is associated with a specific setting of data representation and network architecture. The ANNs in an Ensemble have the same number of hidden units and use one specific data representation from the Classification Environment. In order to enhance the collective contribution of the Ensembles, the ANNs are trained on different sub-sets of the data, yielding relatively independent realizations, that are aggregated to give the Ensemble's opinion. Finally, all the Ensembles are integrated to produce a robust prediction of the whole Classification Environment. This scheme will be now formulated, followed by a suggested method to implement the integration.

The Integrated Classification Machine

The *Integrated Classification Machine* (ICM) is constructed of a hierarchy of classifiers shown in Figure 1. It's smallest building block is a Neural Network (Feed-Forward MLP with Sigmoidal units) which are trained to predict the class label of the given signals. The input dimension (N) of the ANNs is according to the input representation used and the hidden layer can contain various numbers (H), of sigmoidal units. The desired output of the ANNs is {1,0} or {0,1} according to the class membership, thus the output layer contains two sigmoidal units as follows:

$$O^l = \sigma(\sum_{i=1}^{H} W_{li}\ \sigma(\sum_{j=1}^{N} w_{ij}x_j + w_{i0}) + W_{l0}) \quad l = 1, 2. \quad (1)$$

Let us define the *prediction-value* of an ANN as the difference of its two output units: $y(x) = (O^1 - O^2)$. As we use sigmoidal outputs, $y \in [-1, 1]$, the predicted class label is given by thresholding y at zero. Each ANN can be trained on T repeated trials, changing only the initial random weights. Hence, the *prediction-value* of the Network components in the ICM is the average *prediction-value* $y(x)$ over the T training trials: $y^{\mathrm{NET}}(x) = \frac{1}{T}\sum_{t=1}^{T} y_t(x)$.

Each Ensemble is a collection of B Networks, where every Network is trained on one of B Bootstrap Sample Sets (Efron & Tibshirani 1993), re-sampled from the original data: $D_r^b \quad b = 1, \ldots, B$ (for some data

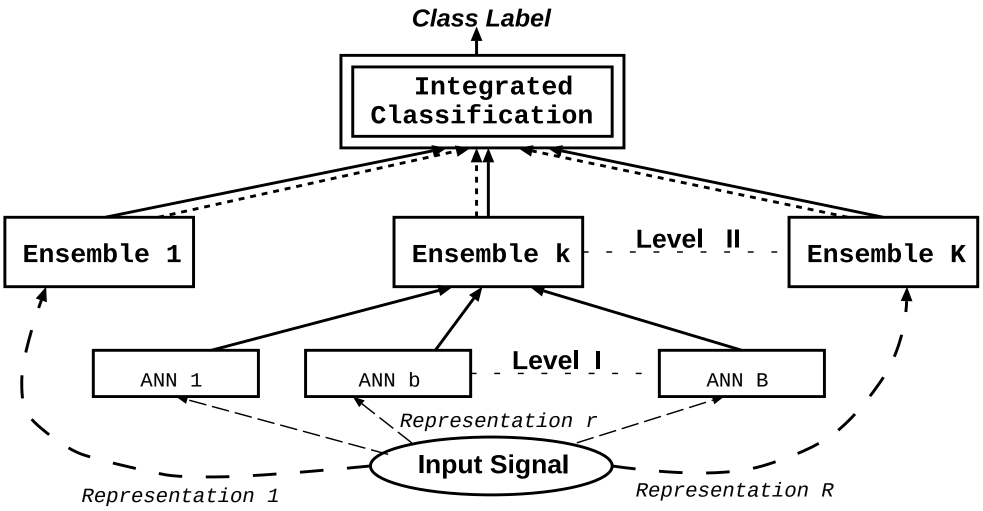

Figure 1: The Integrated Classification Machine (ICM). Several representations of the input signals are fed into different Ensembles in level I, while the integration of those Ensembles is made in level II. (The regular arrows are *prediction values* and the dotted arrows are confidence values).

representation r). All the Networks in an Ensemble share the same data representation and the same architecture. The Ensemble's *prediction-value* is defined as the average *prediction-values* over all the participating Networks, as in 'Bagging' (Breiman 1994):

$$y^{\mathrm{ENS}}(x|D_r) = \frac{1}{B}\sum_{b=1}^{B} y_b^{\mathrm{NET}}(x|D_r^b) \tag{2}$$

A collection of Ensembles ($\mathcal{K}$), which use different input representations and architectures, form the *Integrated Classification Machine* (ICM). The integrated *prediction-value* is defined as follows:

$$y^{\star}(x) = \sum_{k\in\mathcal{K}} \alpha_k\ \beta_k(x)\ y_k^{\mathrm{ENS}}(x) \tag{3}$$

Where α_k is a prior reliability measure of the k'th Ensemble, which can be determined from the training data or assumed using prior knowledge (otherwise, $\alpha_k = \frac{1}{K}$). The second value, $\beta_k(x)$ is a posterior classification confidence measure (specific to each signal), on which we discuss in the next section. Both measures are normalized and determine together the strength of the vote of Ensemble k in the Classification Committee of signal x.

Measuring the Classification Confidence

High dimensional signals from different classes are often separated along the input space in a very unsmooth manner. The difficulty in classifying a given signal can be measured by its distance from the decision boundaries. However, in high dimensional, nonlinear input spaces, it is not realistic to estimate these boundaries explicitly.In order to detect the signals with an ambiguous class membership, or to monitor the classification accuracy, we have constructed a confidence measure for classification of signal x by Ensemble k. This posterior confidence is based on the variance of the Networks' *prediction-value*,

$$CONF^{\mathrm{ENS}}(x) = [\mathrm{VAR}(y^{\mathrm{NET}}(x))]^{-1} \tag{4}$$

This value represents the amount of 'agreement' among all the participating Networks in the Ensemble (Hansen & Salamon 1990; Krogh & Vedelsby 1995).

Combining the Hierarchy of Classifiers

The ICM shown in Figure 1, is a hierarchy of classifiers which are combined in two levels. In Level I, for every Ensemble, B Networks are combined using simple averaging to construct an aggregated ('bagged') Ensemble classification. These Networks are in fact multiple realizations of the same classifier trained on different Bootstrap samples. In the second level the Ensembles are integrated into the final classification (3).

When integrating the Ensembles, decision must be made regarding the strength of the 'vote' of each Ensemble in the classification of a given signal x. In this level (II), it is less likely that a fixed weighting will produce better classification than *all* single Ensembles. Considering our 'pluralistic' classification environment, there might be some Ensembles which use inferior data representations or less suitable model architectures, which will have a disturbing effect on a fixed weighted result as well as on a non-fixed linear weighted result.

Given a low signal-to-noise ratio and the non-stationarity of the signal space, along with a shortage of training data, it is furthermore desirable not to decide upon the integration coefficients based on the characteristics of the current available data set. Therefore, we suggest using the signal x in order to find the optimal integration of its 'own' Classification Committee. Assuming our confidence measure $CONF^{\mathrm{ENS}}(x)$ is effective (i.e., that the bias of the classifier is small enough), we have implemented a non-linear weighting, which is an adaptive version of *'Winner Takes All'* strategy, namely, for *each* signal x, one optimal Ensemble is chosen. The algorithm that is presented in the next section, finds this optimal Ensemble for each signal x separately, using the classification confidence $CONF^{\mathrm{ENS}}(x)$. This approach is different from the *linear* non-fixed weighting that was suggested by (Tresp & Taniguchi 1995) for combining ANNs, where all modules participate in the Classification Committee, with weights inversely proportional to the variance (which is similar to the $CONF$ values used here, but estimated in a different way as they use single ANN modules rather than ensembles of ANNs).

Competing Rejection Algorithm (CRA)

A signal is said to be rejected by a classifier if some measure representing the quality of classification, does not exceed a pre-defined threshold. Obviously, the higher is the threshold, the more signals will be rejected thus the smaller will be the miss classification rate over the remaining un-rejected signals. We present an algorithm (see Figure 2) which performs a sequence of classifications by polling the group of K classifiers w.r.t. the signal at hand. Each classifier (Ensemble) in turn, can either classify or reject the signal. The main motivation is based on the observation that some classifiers perform globally better than others. Nevertheless, classifiers can outperform 'superior' classifiers on a local basis, thus should be given the opportunity to compete and possibly 'steal' a classification whenever the signal was rejected by those 'superior' classifiers. To implement this idea, a prior ranking of the classifiers has to be set and a rejection criterion has to be defined. The rejection is done by thresholding the confidence measure $CONF^{\mathrm{ENS}}(x)$, while the algorithm is set to reject a classification when it's confidence value is lower than the *Reject-Threshold* $\theta_k, \;\; k \in \mathcal{K}$. Each Ensemble k has its own threshold θ_k depending on it's accuracy, that fixes the minimum level of confidence 'allowed' for it's classifications. The θ_k's are calculated from un-labeled data, as a function of the classifier's variability: $\theta_k = \mathrm{percentile}\{CONF^k(x), \varphi_k\}$. The *Reject-Rate* ($\varphi_k$) of a classifier, represents its global credibility. One can determine the φ_k's from the performance on training data or by using subjective information. Otherwise, all ensembles are considered to have the same credibility and a uniform *Reject-Rate* is set.

When all classifiers reject a signal, it can be either 'globally rejected' or classified by the *Ultimate Classifier* which is optionally pre-defined by the user.

program CRA

Order the Ensembles according to their reliability.
Determine the *Reject-Thresholds*
Choose an *Ultimate-Classifier*, $E^\star$.

$LABEL = NULL$
FOR $k = 1$ to K
  $[y^{\mathrm{E_k}}, CONF^{\mathrm{E_k}}] = \mathrm{Classify}_k(x)$
  IF $CONF^{\mathrm{E_k}} >$ *Reject-Thresholds* THEN
    $LABEL = Sign(y^{\mathrm{E_k}})$
    BREAK
  END-IF
END-FOR

IF $LABEL = NULL$ THEN
  $LABEL = Sign(y^{\mathrm{E}^\star})$
END-IF

end CRA

Figure 2: The Competing Rejection Algorithm (CRA)

Notice that the ICM is easily scalable and no retraining is required when new Ensembles are added. It is also flexible in the sense that it is straight forward to incorporate other types of Experts (including human ones) as long as they produce suitable *prediction-value* and $CONF$ value.

Pre-Processing the Seismic Signals

To test the above classification scheme on the seismic classification problem, we have collected a comprehensive seismic data set[1] which consists of 380 events that occurred from January 1990 to June 1993 at a restricted area of 22500 km^2. The events were band-pass filtered to frequency band (0.2-12.5 Hz) and digitized with sampling rate of 50 Hz, using 12 bits A/D converter. All events have magnitude $M_L < 2.7$, 77% of them are below 2.0 and the mean magnitude is 1.53. A given recorded event, includes 8000 samples on average, about half of which consist the actual seismic event, while the rest are recorded before and after the event for synchronization with other seismometers (see Figure 3). We have detected (automatically) the signals' on-set and extracted a fixed size window of $\sim$ 2000 samples, starting at the on-set. This window is about 45 seconds of recording (Shimshoni & Intrator 1995). We have used three spectral decompositions of the waveforms as our basic input representations (denoted: W1,W3,W10). The first decomposition is made of one window of 2048 samples. The second is based on three windows of 1024 samples each, with 384 overlapping samples. For the third representation we used ten windows of 256 samples and 32 overlapping. The size of those windows and their overlapping define the *Time-Frequency* Resolution of the representation.

[1] `ftp.math.tau.ac.il/pub/shimsh/seismic-data/`

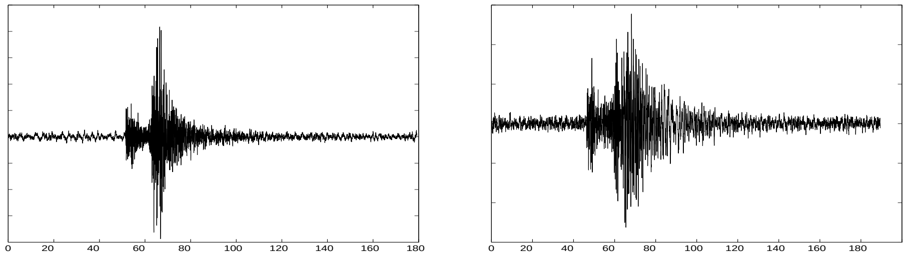

Figure 3: Examples of Seismic Waveforms. Left: Natural Earthquake, Right: Artificial Explosion. (The horizontal axis is recording time in seconds, the Sampling Rate is $50H_z$).

After the *FFT* was applied on a window of samples, it consists of frequency coefficients representing the amplitudes of the respective frequency components in this specific window. For the W3 and W10 representations, the transformed windows were concatenated after scaling to form the final input representation. As the seismic power spectra show a fast decay of energy, it was more appropriate to use a Logarithmic scale, which has also a smoothening effect. The signals after the Log scaling were standardized making each dimension (frequency coefficient) zero-meaned with variance equal one unit.

The pre-processed signals contain more than 1000 dimensions, and are hardly smooth. Before training any classifier on it, we smoothened by taking the averages of a sliding window (thus also reducing its dimension). We applied four levels of smoothening, to achieve the final dimensionalities of: $333, 127, 50, 31$.

The pre-process procedure done here (FFT, scaling, standardization and smoothening) are straight forward and common in many other domains of problems. One might use sophisticated pre-process of the seismic signals to equalize the effect of several factors on the recorded waveform i.e., the distance from the source to the seismometer or the event's magnitude (Dowla, Taylor, & Anderson 1990). We have decided to stay within the common pre-process in order to avoid additional bias and to keep our model as general as possible.

Constructing the Bootstrap Sample Sets

Eight input representations were selected from the combinations of the three T-F resolutions and the four smoothening levels. For each one we have constructed a data set $D_r, r = 1, \ldots, 8$, containing 191 *Earthquakes* and 189 *Explosions*. We followed the *"Bootstrap Pairs"* technique (LeBlanc & Tibshirani 1993), by which every data set D_r was pseudo replicated $B = 30$ times by sampling uniformly with repetitions from the two sets of signals respectively (to preserve the class probability in the resulting replicated samples).

The result is 30 different TRAIN-SETs of 380 signals, each set does *not* contain about 36% of the signals on average. The remaining (un-sampled) signals form 30 corresponding TEST-SETs. These sets are *not* used in any part of the training process, they are used only to evaluate the generalization performance of the model as described in the next section. In conclusion, for each input representation, we have the data sets D_r replicated into 30 couples of disjoint sets such that, $D_r^b = \mathrm{TRAIN}_r^b \cup \mathrm{TEST}_r^b \;\; b = 1, \ldots, 30 \; ; \; r = 1, \ldots, 8$.

Training and Evaluating the Ensembles

All the Networks were trained on 5 trials using Back-Propagation (MATLAB'S TRAINBP), with a fixed learning rate $\eta = 0.001$, initial weights $\in U[-0.1, 0.1]$ and 1000 epochs. The performance estimates we used for comparison and evaluation of the classifiers, were based on Cross-Validation techniques (Efron & Tibshirani 1993).

A single Network's performance (averaged over 5 trials) is estimated via the misclassification rate (MCR) over the TEST-SET corresponds to the TRAIN-SET on which it was trained. The performance of an Ensemble could have been evaluated by averaging the performance of it's 30 Networks and the final integrated classification could have been evaluated accordingly by averaging over the Ensembles. However, this simple bottom-up evaluation does not take into consideration the *combined* classification (Breiman 1994; Perrone 1993), thus yields a pessimistic performance estimate.

In order to present a more realistic cross validated error estimation, we classify every signal x, by combining only a subset of the Ensemble, namely only those Networks that were trained on data sets which did *not* contain the signal x. Denote this subset of Networks as the *Cross-Validated Subset* $CV(x)$ corresponding to signal x. It consists of $B^\star < B$ Networks (each signal was excluded from 11 TRAIN-SETs on average). Hence, for all signals, we have produced a combined classification which can be cross-validated, based on approximately 11 Networks in the corresponding subset $CV(x)$ as follows:

$$\hat{y}^{\mathrm{ENS}}(x) = \frac{1}{B^\star} \sum_{b \in CV(x)} y_b^{\mathrm{NET}}(x) \qquad (5)$$

Representation		Single Classifier			Aggregated Classifier		
T-F	DIM	*k-NN*	*LDA*	*ANN*	*k-NN*	*LDA*	*ANN*
W1	333	0.141	0.462	0.099	0.116	0.437	0.082
W1	127	0.137	0.215	0.100	0.116	0.174	0.087
W1	50	0.129	0.132	0.102	0.113	0.121	0.089
W1	31	0.142	0.117	0.110	0.118	0.105	0.095
W3	375	0.156	0.467	0.104	0.137	0.392	0.097
W3	75	0.155	0.149	0.111	0.134	0.139	0.103
W10	300	0.191	0.484	0.125	0.176	0.471	0.111
W10	120	0.213	0.238	0.134	0.168	0.195	0.113

Table 1: Miss Classification Rates (MCR) of single and aggregated classifiers, for eight different representations.

This estimate is similar to the ϵ_0 in (Efron & Tibshirani 1993), where it is used in a weighted combination with the Apparent error (calculated over the original data set) to construct the .632 estimator: $Err_{.632} = 0.632\ \epsilon_0 + 0.368\ \mathrm{Err}_{app}$. We think that in the framework of trained Neural Networks the .632 estimator is still quite optimistic, therefore we have used the pure ϵ_0. Notice that in a later work the .632 was refined to enable more freedom in choosing the above weights as a function of the *relative overfitting rate* (Efron & Tibshirani 1995). As we combine only one third of the Networks in each Ensemble and use possibly pessimistic estimate ϵ_0, we suggest to consider our evaluation of the Ensembles' performance, thus of the whole ICM, as a lower bound of it's true performance.

Results

The ICM model we presented in this work, was tested on a comprehensive seismic data set consists of 380 events. Fifteen Ensembles of 30 Networks each, were trained 5 trials, on 8 different input representation with various numbers of hidden units (2250 ANNs in total). The data set along with the complete results are available by `ftp` (Shimshoni 1995). Table 1 shows the performance of the ANN classifiers, along with two other classical classifiers for comparison, *LDA* and *k-NN* (applied on the same Bootstrap Sample Sets, under the exact scheme of aggregating multiple classifiers of the ICM). The table shows the Miss Classification Rates (MCR) for single and aggregated (30 realizations) classifiers. The results for *ANN* and *k-NN* are for the best among several values of H and k respectively.

The results show that the most relevant factor is the data representation, namely, the Time-Frequency resolution and also the smoothening level. For all three T-F resolutions, the larger models (higher input dimensionalities) yielded better results than models with more smoothened data. This have probably occurred because of the lower bias of the large models due to their greater capacity (Geman, Bienenstock, & Doursat 1992). The exact Bias-Variance calculation which appear in the complete results, showed that the massive averaging decreases the variance of the classifiers, thus eliminating the need for regularization. It was also evident that there was hardly any influence of the number of hidden units within the tested range of 3 to 12. The preferred model is of the single spectral window (W1) using the higher input dimensionality (dim=333). Comparing the ANN with the classical methods showed that the ANN were preferable addressing the problem at hand, than *LDA* or *k-NN*, for single as well as for aggregated classification.

To test level II of the ICM, we have grouped together the large models of the three T-F representations, providing a wide Time-Frequency coverage. The Ensembles were ordered [W1,W3,W10] (for the rejection algorithm) and the *Ultimate Classifier* was set to be W1. The Rejection Rate was set to 20% for all three Ensembles and the *Rejection Thresholds* θ_k, were extracted from the unlabeled data as defined earlier.

Table 2 shows MCR of the ICM, using three (level II) integration methods: (i) the Competing Rejection Algorithm; (ii) non-fixed, linear weighting based on the variance; (iii) uniform averaging. The table summarizes the MCR for the different levels in the ICM, starting from a single ANN (right), until the full integrated machine (left). All models were trained on the same data and were evaluated as we have described in the previous section. The results show that the best Ensemble's MCR (based on W1, dim=333) was further improved by the integration with the other two Ensembles of the three & ten spectral windows (W3,W10) only when the *CRA* was used. It is also shown that the result of the best single Network is im-

Integrated Machine			Ensemble	Network
CRA	VAR	AVG	(W1-333)	
0.079	0.095	0.095	0.082	0.099

Table 2: Integrated and non-integrated Miss Classification Rates. The 1st column corresponds to non-linear integration with the Competing Rejection Algorithm, the 2nd column shows the result with linear integration based on variance. The 3rd column refers to fixed uniform averaging. The last two columns show the best MCR for un-integrated Ensemble and un-aggregated ANN.

proved by the Ensemble (level I), proving the benefit of averaging multiple ANNs. On average the combined Ensemble classification results, improved the single Network classification by 10-20%.

Conclusions

An Integrated Classification Machine, based on Ensembles of ANNs, was presented and used to address the problem of Seismic Signal Classification. We have created a Redundant Classification Environment that consists of over 2000 ANNs of several architectures, trained with different input representations. By integrating such a large number of classifiers we tried to achieve a robust classification for the noisy and non-stationary seismic waveforms.

Since the massive averaging of the ICM reduces the classification variance, it is advisable in such aggregated models, to use classifiers with larger capacity thus greater ability to handle high-dimensional signals with lower bias, and then apply an efficient combination methods, in order to achieve better generalization.

Examining the classification results corresponding to the different input representations, it appears that there was no gain in using more than a single spectral window (W1), unless using the Competing Rejection Algorithm, which succeeded in extracting a complementary contribution to the final integrated classification. The algorithm, chooses the optimal Ensemble per signal, using the posterior classification confidence, which is based on the classification variance of the Ensemble. By applying such method, we aim to exploit both the robustness of the aggregated classification (level I) and the adaptiveness of the integration strategy (level II). The integrated classification by this method outperformed the non-integrated models as well as other integration methods like linear non-fixed weighting (also based on the variance) and uniform averaging of the Ensembles.

Achieving a result of 92.1% correct classification on the testing data, while using waveform recordings from a single monitoring station only, motivates an implementation of the proposed method for classification of seismic events, a task generally done by humans.

References

Breiman, L. 1992. Stacked regression. Technical report TR-367, Univ. of Cal, Berkely.

Breiman, L. 1994. Bagging predictors. Technical report TR-421, Univ. of Cal, Berkeley.

Dowla, F. U.; Taylor, S.; and Anderson, R. 1990. Seismic discrimination with artificial neural networks: Preliminary results with regional spectral data. *Bull. of the Seis. Soc. of Am.* 80:1346–1373.

Dysart, P. S., and Pulli, J. 1990. Regional seismic event classification at the NORESS array: Seismological measurements and the use of trained neural networks. *Bull. of the Seis. Soc. of Am.* 80:1910–1933.

Efron, B., and Tibshirani, R. J. 1993. *An Introduction to the Bootstrap.* New York: Chapman And Hall.

Efron, B., and Tibshirani, R. 1995. Cross-validation and the bootstrap: Estimating the error rate of a prediction rule. Technical Report (TR-477), Dept. of Statistics, Stanford University.

Geman, S.; Bienenstock, E.; and Doursat, R. 1992. Neural networks and the bias-variance dilemma. *Neural Computation* 4:1–58.

Hansen, L. K., and Salamon, P. 1990. Neural networks ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12:993–1001.

Jacobs, R. A.; Jordan, M. I.; Nowlan, S. J.; and Hinton, G. E. 1991. Adaptive mixtures of local experts. *Neural Computation* 3(1):79–87.

Jacobs, R. A. 1995. Methods for combining experts' probability assessments. *Neural Computation* 7:867–888.

Jordan, M. I., and Jacobs, R. A. 1994. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation* 6:181–214.

Krogh, A., and Vedelsby, J. 1995. Neural network ensembles, cross validation, and active learning. *Advances in Neural Information Processing Systems 7.*

LeBlanc, M., and Tibshirani, R. 1993. Combining estimates in regression and classification. *NeuroProse.*

Meir, R. 1994. Bias, variance and the combination of estimators: The case of linear least squares. Technical report TR-922, Dept of Electrical Eng., Technion, Haifa, Israel.

Perrone, M. P. 1993. *Improving Regression Estimation: Averaging Methods for Variance Reduction with Extensions to General Convex Measure Optimization.* Ph.D. Dissertation, Brown University.

Shimshoni, Y., and Intrator, N. 1995. Automatic discrimination between local earthquakes and quarry blasts by integrating ensembles of neural networks. Proc. of the 2'nd Workshop for AI in Seismology, Lxembourg 1995.

Shimshoni, Y. 1995. Classification of seismic signals using ensembles of neural networks - results appendix. (`ftp.math.tau.ac.il/pub/shimsh/results.ps.Z`).

Tresp, V., and Taniguchi, M. 1995. Combining estimators using non-constant weighting functions. *Advances in Neural Information Processing Systems 7.*

Wolpert, D. H. 1992. Stacked generalization. *Neural Networks* 5(2):241–259.