Wind Power Forecasting With Entropy-Based Criteria Algorithms

Ricardo Bessa, Vladimiro Miranda, *Fellow, IEEE* and João Gama

Abstract-- This paper reports new results in adopting entropy concepts to the training of mappers such as neural networks to perform wind power prediction as a function of wind characteristics (mainly speed and direction) in wind parks connected to a power grid. Renyi's Entropy is combined with a Parzen Windows estimation of the error pdf to form the basis of three criteria (MEE, MCC and MEEF) under which neural networks are trained. The results are favourably compared with the traditional minimum square error (MSE) criterion. Real case examples for two distinct wind parks are presented.

***Index Terms*—Wind power forecasting, neural networks, correntropy, entropy, Parzen windows.**

I. Introduction

THE benefits of accurate wind power forecasting in electrical power systems are being increasingly recognised. Nowadays power systems with renewable energy generation must live with uncertainty both on supply and on demand, but a high uncertainty in wind generation for systems with large penetration of this source (the case in several European Union countries) may cause difficulties in system operation and in operation planning or conventional generation scheduling. The potential for reduced operation costs (via better scheduling decisions) of accepting high penetrations of wind energy [1] has become an attractive factor for the development of new models and tools.

Wind park owners also benefit from better wind power prediction. Short-term wind power forecasting tools are necessary to support a competitive participation of wind power in electricity markets against more predictable energy sources [2].

Short-term wind power forecasting employs mathematical models to map a function whose analytic form is unknown. The search for the best model parameters is usually made in a supervised training mode and we may witness a generalized adoption of a criterion for measuring the performance quality in terms of the prediction error: the MSE (Minimum Square Error).

Wind power prediction at a wind park level is contaminated with uncertainty from two origins: one inherited from the process of wind prediction and the other related with the complex wind park structure and terrain characteristics. Some issues must be underlined:

- The wind turbine conceptually has a non-linear transfer function. The power curve has a linear relation with to the cube of wind speed for small values, a rather steep slope for medium wind speeds and saturation for large wind speed.

  The wind speed must be above a cut-in value to get the turbine into operation. Small deviations of the prediction to the real value in the steep part of the curve are transferred to large differences between power prediction and real power. In contrast, a small deviation in the almost-linear and flat parts of the curve originates a small error in power prediction.

  The transformation of wind speed to wind power changes the statistical properties of the errors, the non-linearity of the power curve amplifies (increases by a factor 2 to 2.6) initial errors in the wind speed prediction according to its local derivative at a certain wind speed. Gaussian errors in wind speed prediction are no longer gaussian when going through this "filter".
- Measurements of wind speed and are usually taken in 10 minute intervals in many facilities. However, the wind speed signal includes components with much shorter wavelengths. Wind speed may fluctuate up or down within the interval of 10 minutes and therefore, one faces a problem of undersampling.
- There is usually no instant power measurement but average power in 10 minute periods (energy generated divided by the duration of the interval). Wind speed is measured every 10 minutes or calculated in terms of average speed within the 10 minute period. This combined with the non-linearity of the generator model is a cause for the introduction of additional uncertainty of non-gaussian nature.

We believe that mapping methods such as neural networks or fuzzy inference systems are well suited to emulate a speed/power relation. Accuracy is of the utmost importance and because average errors are significant (2 digits in percentage) any improvement will be economically valuable. This paper is devoted to reporting further advancement in adopting entropy related concepts in training mappers for wind power prediction, rejecting the traditional approach of tuning systems based on the variance of the error distribution, i.e., based on the Mean Square Error (MSE). The rationale behind this approach lies in the fact that an entropy evaluation may allow one to extract more information from data that just

---

R. J. Bessa is with INESC Porto, Instituto de Engenharia de Sistemas e Computadores do Porto, Portugal, and also with FEP, Faculty of Economy of the University of Porto, Portugal (email: rbessa@inescporto.pt).

V. Miranda is with INESC Porto and also with FEUP, Faculty of Engineering of the University of Porto, Portugal (email: vmiranda@inescporto.pt).

J. Gama is with INESC Porto LA/LIAAD, Laboratory of Artificial Intelligence and Decision Aid, and with FEP, Porto, Portugal (email: jgama@liaad.up.pt).

using a variance-based method, because entropy takes in account all the moments of a probability distribution. Therefore, if errors are not Gaussian (as it seems to be the case in wind power prediction), one may hope to obtain error distributions with higher frequencies close to zero and approximating a Dirac function, which has minimum entropy.

Furthermore, this approach will in principle be less sensitive to outliers and noise.

II. Entropy And Parzen Window Pdf Estimation

Renyi's entropy [3] of a discrete probability distribution P = (p_1, p_2,…, p_n) is defined as

$$H_{R\alpha} = \frac{1}{1-\alpha}\log\sum_{k=1}^{N} p_k^{\alpha} \quad \text{with} \quad \alpha > 0, \alpha \neq 1 \qquad (1)$$

Renyi's entropy is a family of functions $H_{R\alpha}$ depending on a real parameter α. When α = 2, we have what is called quadratic entropy

$$H_{R2} = -\log\sum_{k=1}^{N} p_k^2 \qquad (2)$$

This definition can be generalized for a continuous random variable Y with pdf $f_Y(z)$:

$$H_{R2} = -\log\int_{-\infty}^{+\infty} f_Y(z)^2\,dz \qquad (3)$$

The estimation of the pdf of data from a sample constituted by discrete points $y_i \in R^M$, i=1,…,N in a M-dimensional space, may be done by the Parzen Window method[4]. This technique uses a kernel function centered on each point; it looks at a point as being locally described by a probability density Dirac function, which is replaced or approximated by a continuous set whose density is represented by the kernel. If a Gaussian kernel is used, the expression of the estimation $\hat{f}_Y$ for the real pdf f_Y of a set of N points is a summation of individual contributions:

$$\hat{f}_Y(z) = \frac{1}{N}\sum_{i=1}^{N} G\left(z-y_i,\sigma^2 I\right) \qquad (4)$$

Where G(.,.) is the Gaussian kernel and $\sigma^2 I$ is the covariance matrix (here assumed with independent and equal variances in all dimensions). In each dimension we have

$$G\left(z_k - y_{ik},\sigma^2\right) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(z_k-y_{ik})^2} \qquad (5)$$

It is easy to understand that the "size" of the window, here defined by the value of σ, is important in obtaining a smoother or more "spiky" estimate for f_Y .

III. Information Theoretic Learning

A breakthrough has been achieved by signal processing researchers when they proposed combining Renyi's entropy definition with an estimate of a pdf by the Parzen window method [5][6] – this has been called Information Theoretic Learning. An entropy estimator for a discrete set of data points {y} is

$$H_{R2}(y) = -\log\int_{-\infty}^{+\infty} \hat{f}_Y(z)^2\,dz = -\log V(y) \qquad (6)$$

Where, using (6)

$$V(y) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\int_{-\infty}^{+\infty} G\left(z-y_i,\sigma^2 I\right)G\left(z-y_j,\sigma^2 I\right)dz \qquad (7)$$

In this expression we recognize the convolution of Gaussian functions, and the integral of two Gaussians with equal standard deviations is a Gaussian with twice the standard deviation. Then we have the following result:

$$V(y) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} G\left(yi-yj, 2\sigma^2 I\right) \qquad (8)$$

which allows the practical evaluation of entropy by simply calculating the Gaussian function values of the vector distances between pairs of samples. V(y) is called the information potential (IP) of the data set.

As the objective in supervised training is to minimize H of the errors *e*, one can instead maximize the information potential V. So, Max V(e) becomes the cost function for optimizing a mapper with minimum output entropy [7] – the MEE criterion, for Minimum Entropy Error.

The discovery of weights in a mapper may be done by applying a suitable optimization method that will discover the weights *w* that minimize the objective function

$$\min_{w} H_{R2}(e) \qquad (9)$$

This can be achieved by the classical back-propagation algorithm [8] but in [9] we have applied an evolutionary algorithm to minimize entropy: EPSO, Evolutionary Particle Swarm Optimization. The entropy concept is independent of the training method but several avenues must be explored (backpropagation with a gradient search or other methods) in order to build efficient algorithms to optimize mappers because different objective functions may lead to distinct performance in different algorithms.

IV. Correntropy

Correntropy is a generalized similarity measure between two arbitrary scalar random variables X and Y defined by:

$$V_\sigma(X,Y) = E\left[k_\sigma(X-Y)\right] \qquad (10)$$

where k_σ is the kernel function (usually gaussian).

In real problems, the joint pdf is unknown and we only have available a finite number of points. The estimator is:

$$\hat{V}_\sigma(X,Y) = \frac{1}{N}\sum_{i=1}^{N} k_\sigma(x_i - y_i) \qquad (11)$$

Correntropy is directly related to the probability of how similar two random variables are in a neighbourhood of the joint space defined by the kernel bandwidth, and provides the probability density of the event p(X=Y). Using Parzen windows, the bandwidth controls the observation window in which the similarity is assessed but makes one unable to assess similarity in the whole joint space.

In [11] one may find a discussion on the properties of correntropy. It is proved that a measure CIM(X,Y) = (k(0)-V(X,Y) related with correntropy satisfies all the properties of a metric. CIM may be divided in three different regions: when the error is close to zero CIM is equivalent to L2 norm; when the error grows CIM becomes a L1 norm; when the error is very large CIM becomes a L0 norm, the metric saturates and becomes very insensitive to large errors. This property shows the robustness of CIM and the importance of kernel bandwidth. A small kernel size leads to a small Euclidean zone while a large kernel size will increase the Euclidean region where the metric behaves like the MSE criterion.

When we use correntropy to train adaptive systems, we actually make the system output close to the desired response in the CIM sense if we maximize the correntropy of the error distribution – MCC criterion. We can use the MCC as a new performance function, with the advantage over MSE of being a local criterion of similarity and very useful for cases with non-zero mean, non-Gaussian, with large outliers. It does not require the computing effort of MEE but tends to minimize entropy because it tends to maximize the pdf value at the origin.

The MCC criterion becomes

$$\max_{w} J(e) = \frac{1}{N}\sum_{i=1}^{N} K_{\sigma}\left(g(w, x_i) - T\right) \qquad (12)$$

where g(w,x) represents the mapper producing y = g(w,x) responses from input x as a function of weights w, and T represents the target values. This criterion may be differentiated and a backpropagation algorithm used to train a neural network (NN).

V. Hybrid Approach MEEF

MEE is an exact criterion in terms of entropy concept while MCC is only an approximation. However, MEE is much more consuming in computing effort. Also, MEE has degenerate minima because it is insensitive to the mean of the error. There are methods to deal with the problem. The first method is to correct the result by properly modifying the output bias of the neural network to yield zero mean error over the training data set just after training ends. The other way is to add a so-called MCC term as a function of the errors to the MEE cost function (a function of the difference of errors) as

$$J(e) = \gamma\sum_{i=1}^{N} K_{\sigma\sqrt{2}}\left(e_i\right) + (1-\gamma)\sum_{j=1}^{N}\sum_{i=1}^{N} K_{\sigma\sqrt{2}}\left(e_j - e_i\right) \qquad (13)$$

where γ is a weighting constant between 0 and 1 and K is the kernel function. This cost function is called Minimum Error Entropy with Fiducial points (MEEF) [10]. The MEE term minimizes the error entropy and the MCC term anchors the mean of error at zero.

VI. Training Procedure Optimization

In this paper we are training neural networks in order to produce wind power predictions. For this purpose, the sample entropy estimation is very heavy computationally. We are usually working with very large data sets and using an algorithm with complexity $O(N^2)$ by adopting MEE or MEEF we could easily have processing times of hours in a modern fast PC. Therefore, following the analysis performed in [8] we used instead of the classic backpropagation, the iRPROP[12] algorithm, which with a variable learning rate can achieve a faster convergence and maintaining a good performance. Other strategy that can lead to improved results is using an adaptive kernel size during training [13]. In each test case we start with a large kernel size and during the adaptation gradually and slowly decrease it toward 10% of the initial size with the following empirical rule:

$$\begin{array}{l} \text{IF } \left|Vnorm_t - Vnorm_{t-1}\right| < 0.002 \text{ and } Vnorm_t > Vnorm_{t-1} \\ \text{THEN } \sigma_t = 0.95 \cdot \sigma_{t-1} \end{array} \qquad (14)$$

where V_{norm} is the normalized information potential relative to $V_{\sigma}(0)$.

For the MCC criteria the same logic was followed but instead of annealing the kernel size, we first train with MSE criterion during the first epochs, and switch the criterion to MCC during the remaining epochs. Finally, to reduce the large processing time of batch training we use the batch-sequential approach tested in [14]: the algorithm combines the two methods and takes all the advantages of each. The batch-sequential algorithm consists in randomly dividing the training set in several groups with an equal number of samples (size L) that are presented to the algorithm in a sequential way. In each group we apply the batch training mode and a new subset division is performed in each epoch of the training process. The results of using this method is the MEE complexity reduction from $O(N^2)$ to $O([N/L]/^2)$, the fast convergence and reduction of the probability of the algorithm getting trapped in local minima.

VII. Comparing ITL Criteria With MSE

In this section we present results for training of neural networks in real wind power forecasting problems, comparing the performance of MSE and three ITL inspired criteria (MCC, MEE and MEEF).

The real test cases refer to two wind parks. Wind park A has a rated power of 21600 kW and comprises 12 equal wind turbines of 1.8 MW. The wind park is situated in the north of Portugal, in a complex mountainous region 65 km far from the Atlantic Ocean. Wind park B has a rated power of 16200 kW and comprises 17 equal wind turbines of 0.6 MW and 3 turbines of 2 MW. The wind park is situated in the center of Portugal, in a complex mountainous region 68 km far from the Atlantic Ocean and 200 km far from the wind park A.

The data collected from the wind park included SCADA registers with mean electric power delivered by the wind park into the substation connecting it to the electric power network.

Other model's input variables include forecasts generated from a NWP (numerical weather prediction) model including the well known MM5 mesoscale model (see http://www.mmm.ucar.edu/mm5/) for mean wind speed and wind direction, for a reference point in the wind park with forecasting horizons ranging from 0 to 24 hours.

Fig. 1 displays one year of the NWP/MM5 prediction errors against real wind speed values measured by the metering station at wind park A.

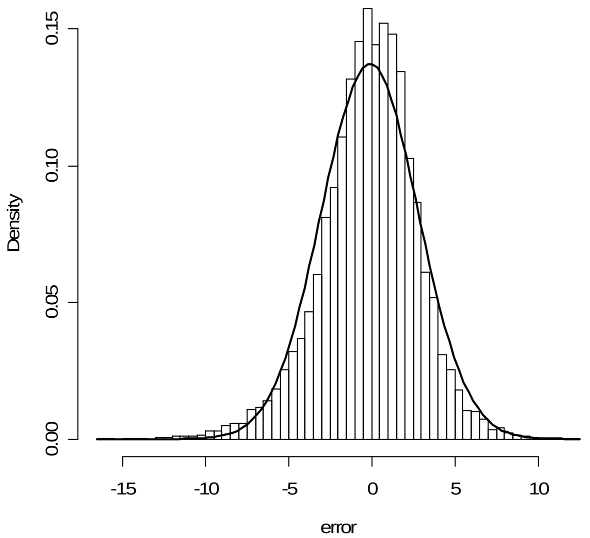

Fig. 1. MM5 wind speed prediction error in wind park A and the normal distribution (curve) with mean of -0.10 and standard deviation of 2.90

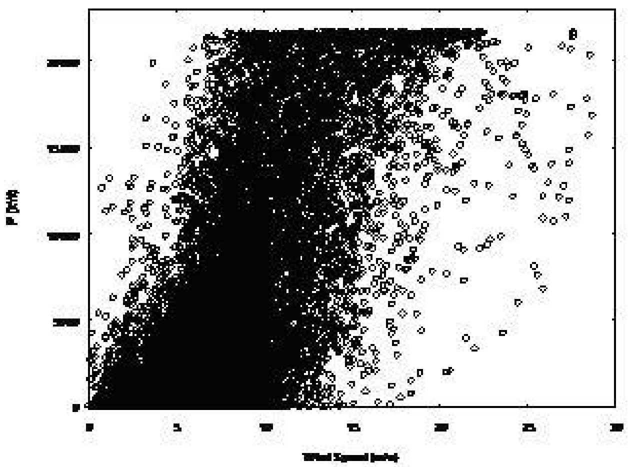

Fig. 2. Wind speed prediction(x axis) against power output measured (y axis) in 30 minute intervals.

It might seem that the NWP error could be approximated by a Gaussian distribution, but if we apply the *Kolmogorov-Smirnov* test to the error data then the null hypothesis (error with normal distribution) is rejected. Even if the NWP error is Gaussian, the non-linearity of the power curve changes the statistical properties of the distribution of deviations between forecasts and measurement.

Fig. 2 displays one year of measurements of power output against wind speed predictions by the MM5 model with an interleave period of 30 minutes. From Fig. 2 it's clear the utility of a local criterion like MCC and the noisy robustness of MEE. We have variables with noisy and a very large number of outliers, some duo to NWP forecast errors and others duo to wind park maintenance.

For example, we have points where the wind speed is above wind speed cut-in but the wind form production is zero or a low value, the cause is some wind park turbine maintenance or malfunction.

These elements show that the application of the ITL criteria have the potential to produce a better mapper than the application of the classical MSE. To demonstrate this, we have trained a feed forward MLP neural network with only one hidden layer comprising 7 neurons, using a hyperbolic tangent activation function.

For the iRPROP training algorithm we chose the standard parameter values for the increase and decrease factor η^+=1.2, η^-=0.5 because they actually produce the best results independent of the learning problem [12]. The remaining parameters of iRPROP are the range of the weight update values Δ_{min}=0, Δ_{max}=50; and the initial weight update value Δ_0=0.0125. The tuning of these parameters is not critical.

The initial kernel size for MEE is 0.3, and the kernel window size for MCC is 0.02. Also the value of these parameters is not critical. In the case of the MEEF criterion the same values for kernel size and same strategy were used. The value of the weighting constant γ in MEEF was set to 0.3.

This training methodology is very useful for wind forecasting as we can train a neural network for any wind park without concerning about tuning parameters. The stopping criterion for each cost function is the early stopping and the maximal number of iterations (we set this value to 150). We stopped the training when the validation error started to increase. The batch-sequential subsets length was 300 points.

To validate the results we adopted a Monte Carlo procedure running 25 simulations, generating randomly weights in each case by a uniform distribution in the interval [-1,1]. The final conclusions derive from the average of all simulations. When comparing criteria, the same weights were used in all cases.

The inputs of the MLP are NWP meteorological forecasted values: mean wind speed values, mean wind direction values. Duo to the cyclic character of the wind direction, this variable comprised two components, i.e. the sine and cosine components. The model totalled 3 input variables, and they were standardized using the min-max method.

The available data of the two wind parks were divided into three data sets. Table I and II shows the statistical characteristics of the mean electric power series data of the two wind parks. The training set is constituted of data from January to March on 2005; the validation set of data from April; the test set of data from May to June on 2005.

TABLE I

CHARACTERISTICS OF THE WIND PARK A DATA

	Training	**Validation**	**Test**
Mean (kW)	7947.06	2567.81	3986.83
Variance (kW^2)	53959535	12806109	23704424
Skewness	0.576	2.154	1.527
Minimum (kW)	-74.33	-68.17	-73.50
Maximum (kW)	21662.70	19177.50	21439.00
Number	5184	1267	2204

TABLE I
CHARACTERISTICS OF THE WIND PARK B DATA

	Training	Validation	Test
Mean (kW)	6184.00	2567.81	3986.83
Variance (kW^2)	27276937	29329712	18641596
Skewness	0.459	2.154	1.527
Minimum (kW)	-60.50	-68.17	-73.50
Maximum (kW)	16142.67	19177.50	21439.00
Number	4893	1267	2204

The training set is composed of data from July to September 2005; the validation set, from October 2005; the test set, from November to December 2005.The negative values indicate electric power consumption of the wind park.

Fig. 3 and 4 compare the pdf of errors obtained with MEE and MEEF in batch-sequential training for wind parks A and B. We see that for case A the pdf of errors is more centred in zero error when the NN is trained with MEEF. It's clear that for wind power forecasting the use of fiducial points as reference for the pfd peak is a better method than the zero mean error. One of the drawbacks in the MEE criteria is the insensitivity to the mean and in this case we have error distributions with large skewness and kurtosis. For case B the results for each criterion are similar.

One drawback of these models was the long processing time of the training of the algorithm but with the training process explained in section VI we have reduced the complexity of the algorithm and the number of training epochs until convergence.

With the classical batch backpropagation training, the processing time for the MEE model in the case of wind park A was about 3 hours for only one simulation, with a Pentium IV 3.0 GHz and 1 GB of RAM. With the optimization of the training, the processing time of one simulation became 10 minutes, with less computational complexity and number of epochs.

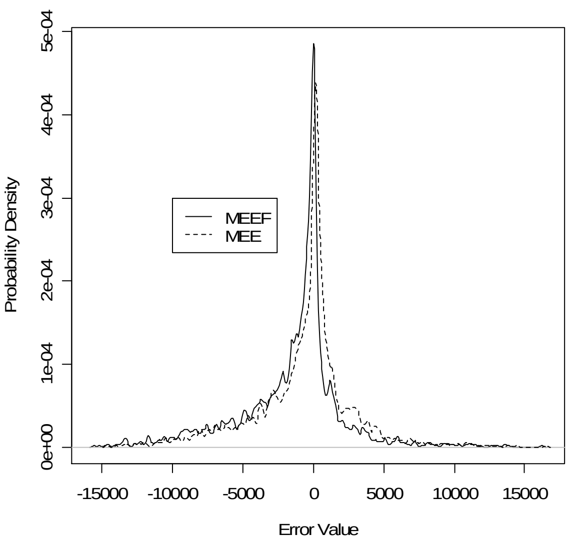

Fig. 3. Comparison of the error pdf generated by NN trained with MEEF (dotted) and MEE (solid) criteria for wind park A

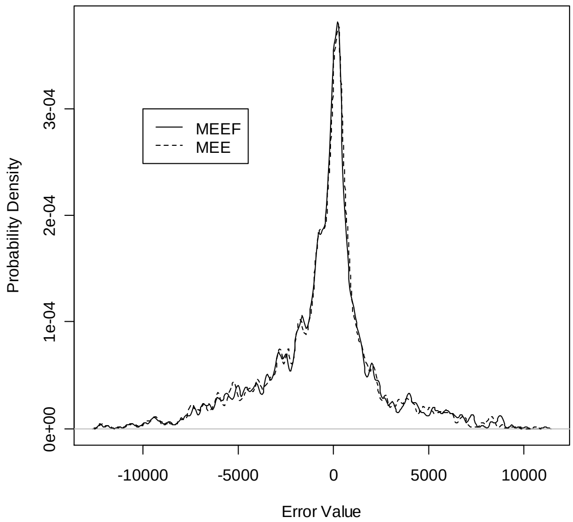

Fig. 4. Comparison of the error pdf generated by NN trained with MEEF (dotted) and MEE (solid) criteria for wind park B

Fig. 5 and 6 compare the pdf of errors obtained with MEEF, MCC and MSE in batch-sequential training for wind park A and B respectively.

From these figures we confirm that the prediction errors are not Gaussian: it was possible to obtain narrower pdf with entropy and correntropy criteria than with a variance-based criterion. In fact, if the error pdf were gaussian, the MSE criterion would perform as well as an entropy-based criterion, but this was not the case.

Therefore, in agreement with the theory, it was possible to design a mapper that would produce a predictor with a higher frequency of errors close to zero, a characteristic associated with smaller entropy of the pdf. In general, this is desirable. For wind park A the MCC presents higher probability density near zero than MEEF, but we have larger errors in the negative part of the PDF.

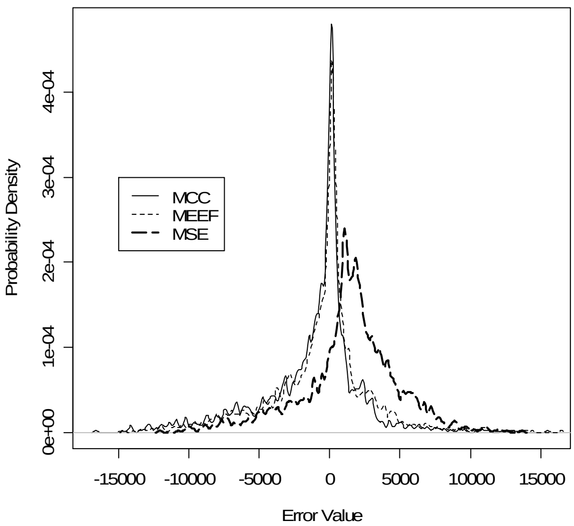

Fig. 5. Comparison of the error pdf generated by NN trained with MEE

(dotted), MCC (solid) and MSE (dotted line width) criteria for wind park A

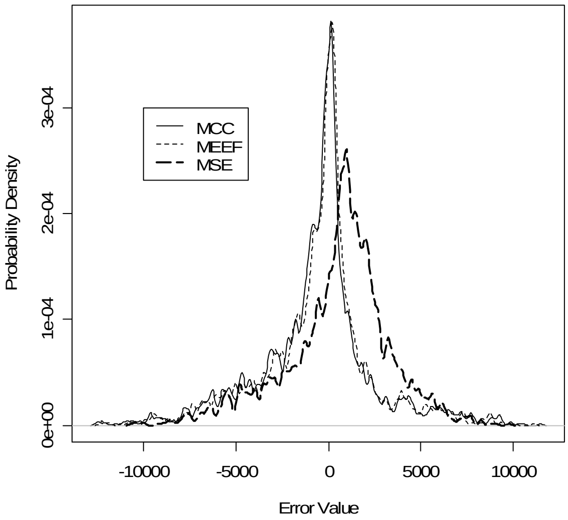

Fig. 6. Comparison of the error pdf generated by NN trained with MEE (dotted), MCC (solid) and MSE (dotted line width) criteria for wind park B

Table III shows the Renyi's Entropy estimated with kernel size of 100, for the error distributions of each criteria and wind park.

TABLE III
RENYI'S ENTROPY OF THE ERROR PDF PER CRITERION

	Wind Park A	Wind Park B
MSE	3.931	3.873
MCC	3.751	3.783
MEE	3.747	3.779
MEEF	3.767	3.789

Table III shows, as expected, that the MEE criterion produces a solution with minimal error entropy, but it's interesting to note that the MCC also minimizes the error entropy with excellent performance.

Fig. 7 shows the forecasted values for the mean 30-minute electric power from the wind park obtained with three criteria (2 ITL criteria and MSE), when compared to the real value registered in the park SCADA. It's clear that the ITL criteria made a better fit and followed better the SCADA measured. It's logical that in some points the MSE produced small errors, but as we see in Fig. 7 the ITL produces a more number of errors close to zero.

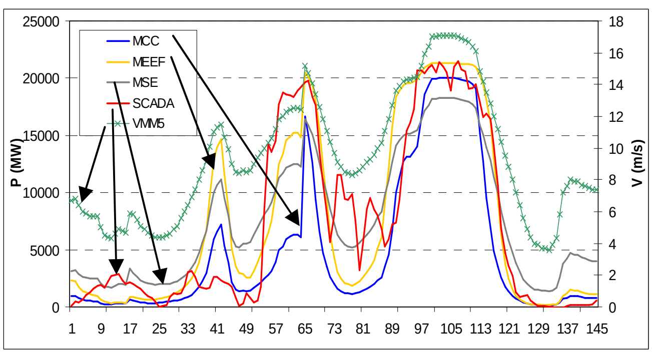

Fig. 7. Comparison among predictions by three criteria on 2005 May 26, 27 and 28 for wind park A. The graph also displays the wind speed prediction provided by the MM5 model (wind speed scale on the right)

It is important to emphasize the fact that we used explanatory variables that are predictions from a NWP model. So, the divergence between the predictions and the real value isn't upsetting. In some cases we can see that the wind speed has a peak which the neural network follows, but the wind park is producing much less; this may due to some turbines being in maintenance, and in these cases the MCC criterion lead to a smaller prediction value because of its robustness to outliers.

VIII. Conclusions

Wind power prediction for wind parks displays obvious non-Gaussian characteristics in the error probability distributions and therefore training input-output systems (mappers) to perform wind power prediction gains from moving away from criteria based on Variance (such as MSE) and instead adopting criteria based on Entropy concepts. For this purpose, a direct use of an entropy measure (the MEE criterion) is possible, but an approximation such as the MCC criterion measuring correntropy has clear computing advantages. The combination of these two criteria is a robust alternative; we can get all the advantages of the two criteria. Also optimization procedure techniques are needed to deal with the large computing effort of calculating entropy for large data sets.

This paper also contributes with evidence from real cases supporting that Entropy-based criteria perform better than a variance-based criterion, based on studies performed with real data from wind parks in Europe and comparisons between NN trained under the three competing criteria – MSE, MEEF and MCC.

VI. References

[1] Mark L. Ahlstrom, Robert M. Zavadil, "The Role of Wind Forecasting in Grid Operations & Reliability" *IEEE/PES Transmission and Distribution Conference & Exhibition: Asia and Pacific*, China, pp. 1-5, 2005.

[2] P. Pinson, C. Chevallier, G. Kariniotakis, "Optimizing Benefits from Wind Power Participation in Electricity Markets using Wind Power Forecasting Embedded with Uncertainty Management Tools", *European Wind Energy Conference EWEC 2004*, London, UK, 22-25 Nov. 2004.

[3] A. Renyi, "Some Fundamental Questions of Information Theory", *Selected Papers of Alfred Renyi*, vol 2, pp. 526-552, Akademia Kiado, Budapest, 1976.

[4] [10] E. Parzen, "On the estimation of a probability density function an the mode", Annals Math Statistics, vol 33, 1962

[5] J. C. Principe and Dongxin Xu, "Introduction to information theoretic learning", *Proc. International Joint Conference on Neural Networks (IJCNN'99)*, Washington DC, USA, 10-16 July 1999, pp. 1783-1787.

[6] J. C. Principe and D. Xu "Information-theoretic learning using Renyi's quadratic entropy", *in J.-F. Cardoso, C. Jutten, and P. Loubaton, editors, Proceedings of the First International Workshop on Independent Component Analysis and Signal Separation*, Aussois, France, pages 407-412, 1999.

[7] D. Erdogmus and J. C. Principe, "Generalized Information Potential Criterion for Adaptive System Training", *IEEE Transactions on Neural Networks*, vol. 13, no. 5, September 2002, pp. 1035-1044.

[8] R.A. Morejon, J.C. Principe, "Advanced search algorithms for information-theoretic learning with kernel-based estimators", *IEEE Transactions on Neural Networks*, vol. 15(4), pp. 874 - 884, July 2004

[9] V. Miranda, C. Cerqueira, C. Monteiro, "Training a FIS with EPSO under an Entropy Criterion for Wind Power prediction", *International*

Conference on Probabilistic Methods Applied to Power Systems, PMAPS 2006, Stockholm, 11-15 June 2006.

[10] Liu Weifeng, P.P. Pokharel, J.C. Principe, "Error Entropy, Correntropy and M-Estimation", *16th IEEE Signal Processing Society Workshop on Machine learning for Signal Processing*, pp. 179–184, 2006

[11] W. Liu, P. Pokharel, J. Principe, "Correntropy: properties and applications in non-Gaussian signal processing", *Accepted in IEEE Transactions on Signal Processing*, 2007.

[12] C. Igel, M. Husken, "Improving the Rprop learning algorithm", *Proceedings of the Second International ICSC Symposium on Neural Computation (NC 2000)*, ICSC Academic Press, pp.115-121, 2000

[13] D. Erdogmus and J. C. Principe, "Generalized information potential criterion for adaptive system training," *IEEE Transactions on Neural Networks*, vol. 13, pp. 1035–1044, Sept. 2002.

[14] Jorge M. Santos, J. Marques de Sá, Luís A. Alexandre, "Batch-sequential algorithm for neural networks trained with entropic criteria", *Proceedings of the 15th Int. Conf. on Artificial Neural Networks*, pp. 91–96, 2005

VII. Biographies

Ricardo Bessa received his Licenciado degree from the Faculty of Engineering of the University of Porto, Portugal (FEUP) in 2005 in Electrical and Computer Engineering. He then joined INESC Porto and is a researcher in its Power Systems Unit while completing a M.Sc. degree in the Faculty of Economy of the University of Porto (FEP).

Vladimiro Miranda (M'90, SM'04, F'05) received his Licenciado, Ph.D. and Agregado degrees from the Faculty of Engineering of the University of Porto, Portugal (FEUP) in 1977, 1982 and 1991, all in Electrical Engineering. In 1981 he joined FEUP and currently holds the position of Professor Catedrático. He is also currently a Director of INESC Porto, an advanced research institute in Portugal. He has authored many papers and been responsible for many projects in areas related with the application of Computational Intelligence to Power Systems.

João Gama received his Licenciado degree from the Faculty of Engineering of the University of Porto, Portugal (FEUP) in Electrical and Computer Engineering.. In 2000 he received his Ph.D. degree in Computer Science from the Faculty of Sciences of the same University. He joined the Faculty of Economy (FEP) where he holds the position of Assistant Professor, He is also a senior researcher at LIAAD (Laboratory for Artificial Intelligence and Decision Aid), a group belonging to INESC Porto LA. He has worked in projects and authored papers in areas related to machine learning, data streams and adaptive learning systems and is a member of the editorial board of international journals in his area of expertise.