Modelling Geomagnetic Activity Data

ERNST D. SCHMITTER
University of Applied Sciences
Department of Engineering and Computer Sciences
Albrechtstr. 30, 49076 Osnabrueck
GERMANY
e.d.schmitter@fh-osnabrueck.de

Abstract: Strong geomagnetic activity is a hazard to electronics and electric power facilities. Assessment of the actual geomagnetic activity level from local magnetometer monitoring therefore is of importance for risk assessment but also in earth sciences and exploration. Wavelet based signal processing methods are applied to extract meaningful information from magnetic field time series in a noisy environment. Using a proper feature vector a local geomagnetic activity index can be derived under not ideal circumstances using computer intelligence methods. Locally linear radial basis function nets and self organizing maps are discussed in this context as data based process models.

Key–Words: geomagnetism, signal processing, wavelets, neuro fuzzy modelling, self organizing map

1 Introduction

Monitoring geomagnetic activity is a task of considerable interest for earth sciences but also for predicting hazards for electronics, communication, navigation and mains power failure. Along with global activity measurements averaged from a number of worldwide distributed magnetometer sites and satellites, local measurements are necessary for assessing local conditions - e.g. geomagnetically induced currents in electrical power grids - but also for the application of geophysical exploration methods [3] relying on magnetic field measurements (practical examples: directional drilling for oil and archeological prospection). Geomagnetic storm risk assessment, real-time activity monitoring and surface induced electric field modelling are topics of recent scientific studies of national geophysical institutions as for example the British Geological Survey (BGS, see: www.geomag.bgs.ac.uk).

Such recordings often cannot be done in a noise free environment and therefore call for advanced signal processing methods. After some remarks with respect to the magnetic field monitoring process we discuss statistical and transform based parameters that prove to be useful for characterising local deviations from a 'qiet' earth magnetic field condition and allow to quantify geomagnetic activity using neuro fuzzy and self organizing map classifiers.

2 Geomagnetic Activity

The space around earth that is influenced by the magnetic field of the earth is called magnetosphere. This domain interacts with the solar wind, i.e. a hot dilute plasma with frozen in magnetic fields constantly ejected by the sun. Solar activity phenomena as sunspots, prominences, flares, coronal holes and coronal mass ejections (CMEs) may strongly increase solar wind. These activities are controlled by the suns magnetic field and its instabilities. Especially coronal holes (corresponding to open solar magnetic field lines) and CMEs (massive, i.e. $10^{11} \ldots 10^{14}$ kg, bursts of plasma ejected from closed magnetic field regions) are responsible for solar plasma storms escaping into the interplanetary space. Geomagnetic storms lasting for hours up to several days are the reaction of the earths magnetic field to these invasions. The flow of injected ions and electrons within the magnetosphere and the ionosphere form current systems, causing variations in the intensity of the Earth's magnetic field. Strong geomagnetic activity, i.e. magnetic field variations can trigger transients on electrical power lines, communication channels and pipelines causing severe failures in these systems. The actual activity amplitudes however vary massively with location, especially with (geomagnetic) latitude, being strongest near the (geomagnetic) poles.

Magnetic indices are measures of geomagnetic activity. Often used indices are: the planetary range index Kp, the polar cap index PC, the auroral electrojet index AE, and the equatoial ring current index DST (Disturbance Storm Time).

In this paper we concentrate on the analysis of a local activity range index K related to the local variation of the E-W magnetic field component. The B_y or geomagnetic EW field component is zero under undisturbed conditions. Deviations from zero are activity

caused. Note that in this paper directions are defined with respect to the (local) geomagnetic and not the geographic coordinate system.

Figure 1: Strong geomagnetic activity at medium latitude (52N, 8.5E) at Dec. 15th 2006 caused by solar plasma ejection: 24 hour variation of the horizontal EW magnetic field component.

3 Magnetic Field Monitoring

Basically the local geomagnetic disturbance level is quantified by the 3 hour range of the horizontal magnetic field components B_x, B_y- with B_x as the N-S, B_y the E-W (and B_z the vertical) component. The range being understood as the difference between maximum and minimum values within this time span. The activity level a for a 3h time interval is defined by $a = max(range)/2$ (unit: nT, nano-Tesla), with 'max' taken over the horizontal field components. Usually not a, but a nearly logarithmic function of it, the K-index is used. Globally the K-indices of a number of geomagnetic observatories are combined to yield a global (planetary) index Kp with values from 0 to 9 for the 3h intervals starting at 0, 3, 6, 9, 12, 15, 18, 21 UTC (Universal Time Coordinated). In a not perfect environment the local 3h-range of magnetic field values is detoriated by temperature drift of the sensor and manmade field disturbances. They can be dealt with to a certain extent, but not completely removed. Therefore some additional features gained from the field time signal are proposed for a more secure local geomagnetic activity assessment.

The following discussion uses $B_y(t)$ time series gained with a fluxgate magnetometer with a resolution of $2nT$.

Signal processing is based on the magnetic field value relative to its value at 0 UTC: $B_{y,rel}(t) = B_y(t) - B_y(t = 0\ UTC)$.

3.1 Preprocessing

Despite heat isolation of the fluxgate temperature varies over the day (fig. 3). A temperature sensor therefore under the same isolation conditions is placed near the field sensor. A linear temperature drift correction is applied according to:

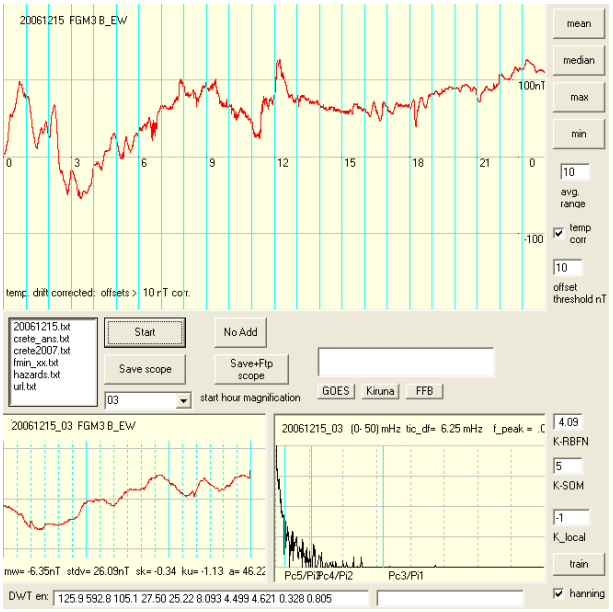

Figure 2: Monitor program with 24 hour display. Lower part: FFT and DWT analysed 3 hour range together with the RBFN and SOM classifications of the geomagnetic activity.

$$B_{y,rel,corrected} = B_{y,rel} + a * (T - T_0) \qquad (1)$$

With T_0: temperature at 0 UTC and a: sensor specific constant.

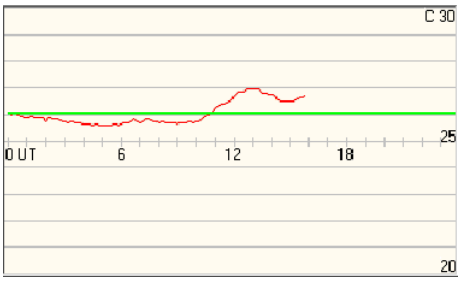

Figure 3: Typical temperature variation over the day relative to 0 UTC at the sensor.

As the activity level in the 3h interval basically is defined via the maximum field value difference within the time interval, artificial offsets have to be eliminated. A mass of magnetizable material, as for example cars, changes the local field at distances of several tens of meters, i.e. produce a constant offset as long as they are in place. The offset changes sign, if they are removed again. These kinds of offsets can be dealt with in by continuously logging step jumps in the signal with the appropriate sign (fig. 4).

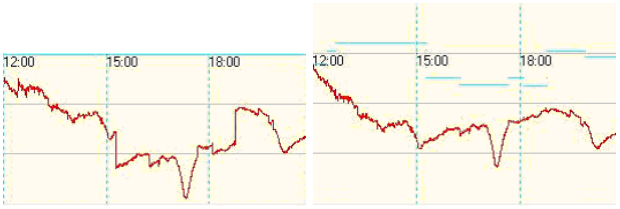

Figure 4: Offset compensation. Left: without, right: with compensation. The horizontal lines in the right figure indicate how the total offset level develops in time.

4 Data Features

The basic analysis interval is $T = 3$ hours with $n = 1024 = 2^{10}$ B_y-samples (this is a sample rate of nearly 0.1 Hz). Fig. 5 shows typical examples with a certain geomagnetic activity level in an undisturbed and - what is more typical - disturbed signal trace. We now want to extract features that are able to discriminate between the natural and noise part.

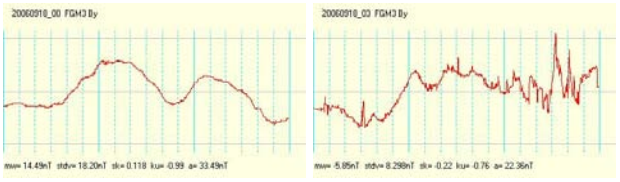

Figure 5: Geomagnetic activity within 3 hours intervals. Left: mostly undisturbed example right: signal with superimposed manmade disturbances.

Inspection of the wavelet decomposition (Daubechies4, [5], [14]) in combination with the Fourier transform of 3h-intervals shows that the wavelet energies in the scales $5, 6, 7$, i.e. $e(5), e(6), e(7)$ are most characteristic for the geomagnetic activity in this period. The energy $e(s)$ on scale s simply is the squared sum of the DWT coefficients of that scale. Using the equivalence $s = lb(4\ f\ T)$ (see fig. 6) a scale s corresponds to center frequencies $f_{center} = 1/T\ 2^{s-2}$ in the range $0.5 \cdots 5\ mHz$ (i.e. periods roughly from 3 to 30 minutes). In this frequency band geomagnetic pulsations of class Pc5 and Pi3 can be found. Geomagnetic field variations are categorized by their period and structure into classes Pc1 to Pc5 for continuous structured pulsations and Pi1 to Pi3 for irregular structures [4].

As a feature vector for the activity classification of a 3h interval we therefore choose the 4 components: $B_{y,max} - B_{y,min}, e(5), e(6), e(7)$.

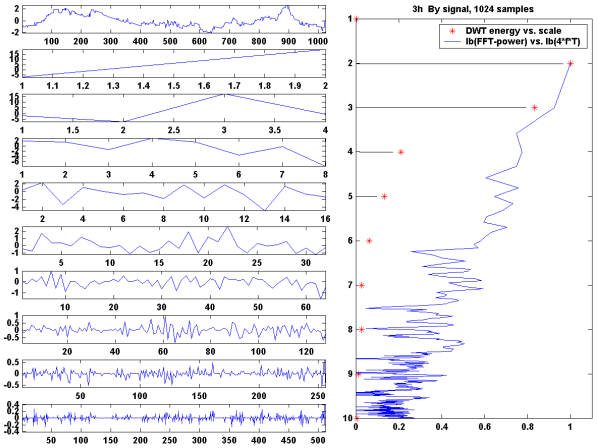

Figure 6: Example for a $T = 3h$ B_y signal ($n = 1024$ samples, normalized with respect to mean and standard deviation,top) and its DAUB4 wavelet decomposition ($scales = 2 \cdots 10$, below).In parallel (right) the DWT scale energies together with the Fourier decomposition are displayed. The information of the Fourier power spectrum $log(Re(FFT)^2 + Im(FFT)^2)$ around frequency f corresponds to a wavelet scale energy at scale $s = lb(4\ f\ T)$ in the ranges $f = 1/T \cdots (n/2)/T$ and $s = 1 \cdots lb(n)$.

5 Neuro Fuzzy Data Model

For the derivation of a local K-index quantifying geomagnetic activity we use a neuro fuzzy data model and alternatively a self organizing map (chapter 6). Radial Basis Function Neural Nets (RBFNs) and fuzzy logic extensions have been successfully used in very different applications ([15],[16],[17],[18]). In this chapter we describe a locally linear radial basis function network (LL-RBFN) with fuzzy interpretation possibility.

Each training vector consists of $p = 4$ features and a classification value y. The matrix of training vectors is normalised with respect to the mean and the standard deviation of each component. With normalised feature vector $\vec{x}$, weights w_{0j} and $\vec{w}_j$, N basis function centers $\vec{t}_j$ and width vectors $\vec{c}_j$ the normalised classification output $y^{(n)}$ for a (normalised) input $\vec{x}$ is

$$y^{(n)}(\vec{x}) = \frac{1}{s(\vec{x})} \sum_{j=1}^{N} (w_{0j} + \vec{w}_j \cdot (\vec{x} - \vec{t}_j))\ \phi_j(\vec{x}) \qquad (2)$$

with gaussian basis functions

$$\phi_j(\vec{x}) = e^{-\sum_{i=1}^{p}((x_i - t_{ji})/c_{ji})^2} \qquad (3)$$

and

$$s(\vec{x}) = \sum_{j=1}^{N} \phi_j(\vec{x}) \quad (4)$$

normalizing the basis functions. In total we have $N(3p + 1)$ parameters.

Training of a LL-RBFN can be done with gradient descent algorithms [6] or optimisation procedures that are simplex (Nelder-Mead [5]) or evolutionary based or are tree construction oriented like the LOLIMOT (LOcally LInear MOdel Tree) algorithm [10]. For our application a line search (numerical gradient descent) algorithm proved efficient.

As starter parameters for a training process we select N basis function centers $\vec{t}_j$ from the training set (input vectors). This can be done totally at random or better by assuring that each relevant index range y is represented by a center. We use a k-means cluster algorithm to this end. An upper limit for a meaningful number of centers N can be found by repeated k-means runs on the input vectors and looking for an about equally distributed number of input vectors in each cluster.

Width parameters are intitialised according to

$$c_{ji} := \frac{d_{max}}{N} \quad (5)$$

with the maximum center distance

$$d_{max} = max_{i,j} \left|\vec{t}_i - \vec{t}_j\right| \quad (6)$$

The initial weights we get from

$$w_{0j} := \sum_{i=1}^{m} g^{+}_{ji}\, y_i^{(n)} \quad (7)$$

with m training vectors $(\vec{x}_i, y_i^{(n)})$, and g^+ being the pseudoinverse matrix of $g_{ij} = \phi_j(\vec{x}_i)/s(\vec{x}_i)$.
The linear coefficient weights $\vec{w}_j$ are initialized to 0. So the LL-RBFN is initialized as a usual RBFN.

Weights, centers and width parameters are optimised (trained) using a numerical gradient descent (line search) algorithm with respect to the mean squared classification error. A training data set with 400 vectors was modelled by the trained net with a rms error of 0.4 with respect to an K-index ranging from 0 to 6. The rms error with respect to 100 test vectors was about 0.6. A difficulty with getting training and test data for a site is that bigger K-values occur exponentially less frequent.

The local linear RBFN approach allows to reduce effectively the number N of basis functions, i.e. hidden neurons, because of the additional free linear parameters. This is an important point with regard to execution speed, but especially with regard to interpretability. N typically is of the order of integer K-values, i.e. $N = 7$ for $K = 0 \cdots 6$ (the RBFN output being continuous however).

5.1 Fuzzy Rule Interpretation of the LL-RBFN

One of the reasons for choosing a LL-RBFN was, that it has a structure allowing a straightforward Takagi-Sugeno fuzzy rule interpretation [7], [8] It is therefore sometimes called Locally Linear Neuro Fuzzy Model (LLNFM).

Within the Takagi-Sugeno framework a rule has fuzzy input and crisp output and can be formulated as:

IF $\vec{x}$ is in the domain of basis function j THEN $y^{(n)} = w_{0j} + \vec{w}_j \cdot (\vec{x} - \vec{t}_j)$

So, the LL-RBFN output (equ. 2) can equivalently be looked at as the output of a system with N rules, each having fuzzy premises and crisp consequences. In this context $w_{0j} + \vec{w}_j \cdot (\vec{x} - \vec{t}_j)$ is the weight of rule j and $\phi_j(\vec{x})/s(\vec{x})$ the relevance of rule j for an input $\vec{x}$.

Because of the identity

$$e^{-\sum_{i=1}^{p}((x_i - t_{ji})/c_{ji})^2} = \prod_{i=1}^{p} e^{-((x_i - t_{ji})/c_{ji})^2} \quad (8)$$

for a p-dimensional input the premise part of rule j can be read as

IF $\vec{x}$ is in the domain of basis function j $\equiv$

IF x_1 is in d_{j1} AND .. AND x_p is in d_{jp}

where $d_{ji} = \frac{1}{s^{1/p}(\vec{x})}\, e^{-((x_i - t_{ji})/c_{ji})^2}$ is the gaussian membership function for input component i centered at t_{ji} with width parameter c_{ji}.

Fuzzy rule based interpretation of a LL-RBFN with a low number of basis functions allows for some more direct insight into the classification process than a pure RBFN (usually needing more basis functions for the same fit accuracy) or backpropagation networks. In this way domain analysis of the basis functions using the trained centers, widths and weights reveals correlations between feature combinations and signal characteristics and allows for rule extraction under certain prerequisites discussed in [9]. The RBFN training process generates for each input variable ($p = 4$ in our case) as many membership functions as there are rules, i.e. N. Figure 7 with graphic display in fig. 7 shows the centers and widths of the membership functions for 6 rules from an example training run. Having more than 5 or 6 membership

functions for a fuzzily interpreted variable usually hinders interpretability. Inspecting the table and especially the function plots we find that the membership functions are not equally distributed over the input value domain and some of the functions are quite similar. A retraining under the described constraints (not too many rules, more or less equally distributed and shaped membership functions) is usually necessary to gain some really interpretable and nontrivial rules. This however often diminishes the modelling accuracy of the retrained system.

a		e(5)		e(6)		e(7)		w
1.31	0.02	1.30	4.37	-3.72	4.61	-2.72	4.49	0.49
3.36	0.17	3.34	4.82	-0.57	0.03	-1.13	0.57	1.68
-2.04	0.22	-4.23	2.80	2.40	2.72	-6.68	1.19	-0.85
-2.75	0.27	2.30	3.28	-1.40	4.96	-2.12	0.54	3.19
10.60	1.04	0.95	0.10	1.15	0.26	-4.45	1.99	1.90
0.75	0.35	-5.20	0.88	1.02	0.13	6.21	0.58	-0.38

Figure 7: Centers and widths of the gaussian membership functions for the 4 features $a = B_{y,max} - B_{y,min}$ and DWT-energies $e(5)$, $e(6)$, $e(7)$ and the main weights w_{0j}. Each row corresponds to one rule. For each feature there are 6 membership functions according to the $N = 6$ rules.

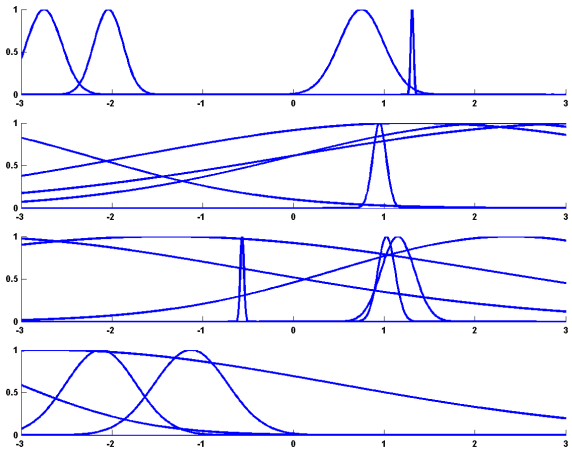

Figure 8: Plot of the gaussian membership functions according to fig. 7 for the $p = 4$ input variables (normalized with respect to mean and standard deviation). From top to bottom: $a = B_{y,max} - B_{y,min}$ and DWT-energies $e(5)$, $e(6)$, $e(7)$

With respect to these drawbacks we consider an alternative data modelling approach: the self organizing map.

6 Self Organizing Map (SOM)

A self organizing map (SOM) [11], [12] maps a high dimensional feature vector space into a one or two dimensional space in a topology preserving manner. This means it preserves mutual relationships in the feature space of input data by clustering mutually similar feature vectors in neighboring nodes. We use a two dimensional image space in this paper. Each node of the two dimensional topological feature map holds a codebook vector together with the output class defined for it. An input vector then is characterized by the output class of the nearest codebook vector, see figure 9. Neighbourhood (similarity) is usually defined with respect to the Euclidean norm.

Beyond pure classification of input vectors the two dimensional representation gives insight into the topological structure of the higher dimensional data basis and the relative position of each new input. The SOM even allows for rule extraction [13].

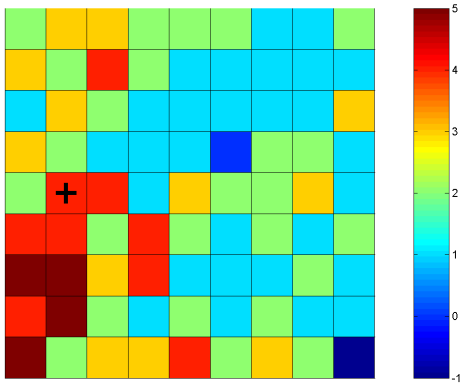

Figure 9: 10 x 10 self organizing map (SOM) with classification result (cross) with respect to an input vector identified as K-index = 3. Each node holds a four dimensional codebook vector. Their output class is indicated by the colorbar. By the nature of the self organizing construction process of the SOM some nodes might be empty, i.e. do not classify an input (here labeled as -1).

As the output of a SOM is a class and not a continuous value we loose some modelling accuracy by accepting outputs as natural numbers $K = 0, 1, 2, ...$

It is obvious from fig. 10 that a high output class K is related to high values of the input varable a and low values of $e(5)$, $e(6)$, $e(7)$. The DWT energies $e(5)$, $e(6)$, $e(7)$ show similar clusterings. Medium to high e-values are related to medium to low K-classes.

7 Conclusion

Signal processing and computational intelligence methods are discussed that proved successful in deriving a local geomagnetic activity index from magnetic field time series in a noisy environment. To this end a feature vector with mainly wavelet based

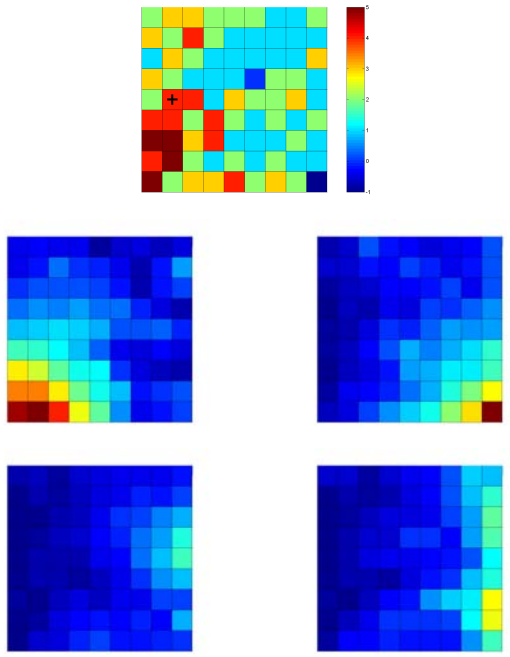

Figure 10: The SOM and its $p = 4$ feature planes. From upper left to lower right: $a = B_{y,max} - B_{y,min}$ and DWT-energies $e(5)$, $e(6)$, $e(7)$.

components is used with a locally linear radial basis function net (LL-RBFN) and a self organizing map (SOM). The LL-RBFN yields a superior modelling accuracy. However knowledge discovery by exploiting the interpretation of the LL-RBFN within a Takagi-Sugeno fuzzy rule framework is possible in principle but difficult in detail. Exploiting the cluster topology of a SOM proved to give some more insight into the data model. The extension of the data based process model from pure classification to predicting local geomagnetic activity is the subject of ongoing work.

References:

[1] E.D. Schmitter, Characterisation and Classification of Natural Transients, *Transactions on Engineering, Computing and Technology* 13, May 2006, pp. 30–33

[2] E.D. Schmitter, Analysing and Classifying VLF Transients, *International Journal of Signal Processing (IJSP)* 3, 2006, pp. 238–242

[3] Redford M.S., in Lanzerotti L.J., Kennel C.F. and Parker E.N., Eds., Problems of magnetic fluctuations in geophysical exploration, *Solar system plasma physics* 3, North Holland Publ. Co., 1979, p365

[4] Orr D., Magnetic pulsations within the magnetosphere: A review, *Journal Atmos. Terr. Physics* 35, 1973, p1

[5] Press W.H., Teukolsky S.A., Vetterling W.T., Flannery B.P., *Numerical Recipes in C*, Cambridge University Press, 1992

[6] Haykin S., *Neural networks*, Prentice Hall, 1999

[7] Takagi T., Sugeno M., Fuzzy identification of systems and its applications to modeling and control, *IEEE Transactions on Systems, Man and Cybernetics* 15, 1985, pp. 116–132

[8] Jang, S.R., Sun, C.T.,Functional equivalence between radial basis function networks and fuzzy inference, *IEEE Transactions on neural networks* 4(1), 1993, pp. 156–159

[9] Jin Y., von Seelen W., Sendhoff B., Extracting interpretable fuzzy rules from RBF Neural Networks, *Internal Report, IR-INI 2000-02, ISSN 0943-2752*, Institut fuer Neuroinformatik, Ruhr-Universitaet Bochum, Bochum, January 2000

[10] Nelles, O., *Nonlinear System Identification*, Springer, Berlin, 2001

[11] Kohonen T., Self Organizing Maps, *Springer, Berlin* 2nd edition, 1997

[12] Haykin, S., Neural Networks, *Prentice Hall* 2nd edition, 1999

[13] Malone J., McGarry K., Wermter S. Bowerman C., Data Mining using Rule Extraction from Kohonen Self-Organising Maps, *Neural Computing and Applications* 15(1), 2006.

[14] Graham J.L., Goodwin C., Frequency-domain Constructed Redundant Bases for Denoising *Proceedings of the 7th WSEAS International Conference on Multimedia Systems & Signal Processing*, Hangzhou, China, April 15-17, 2007

[15] Coufal D., Classification of EEG signals by radial neuro-fuzzy system *WSEAS TRANSACTIONS on SYSTEMS*, Issue 2, Volume 5, February 2006, ISSN 1109-2777

[16] Cpalka K., Rutkowski L., A New Method for Complexity Reduction of Neuro-Fuzzy Systems *WSEAS TRANSACTIONS on SYSTEMS*, Issue 11, Volume 5, November 2006, ISSN 1109-2777

[17] Chandrakar V. K.,Kothari A. G., RBFN Based Unified Power Flow Controller (UPFC) for Improving Transient Stability Performance *WSEAS TRANSACTIONS on POWER SYSTEMS*, Issue 1, Volume 2, January 2007, ISSN 1790-5060

[18] Rodriguez A., Carracajo I., Dafonte C., Arcay B., Manteiga M., Hybrid Approach to MK Classification of Stars Neural Networks and Knowledge-based Systems *Proceedings of the 6th WSEAS Int. Conf. on Artificial Intelligence, Knowledge Engineering and Data Bases*, Corfu Island, Greece, February 16-19, 2007