

FEATURE SELECTION FOR SMALL-SIGNAL STABILITY ASSESSMENT

S.P. Teeuwsen
University of Duisburg
teeuwsen@uni-duisburg.de

Abstract

This paper introduces different feature selection techniques for neural network based small-signal stability assessment. Large-scale power systems like the European interconnected network may experience low frequency oscillations between remote parts of the system. These oscillations are caused by large power transits in the network.

In dynamic security assessment, a fast and accurate artificial intelligence technique can be applied. Hereby, the state of the system is predicted by the use of a neural network (NN), which provides information about the system eigenvalues and therefore the damping of the oscillations.

Because NN cannot be trained with the complete power system data, a reduction technique needs to be implemented.

Therefore, this paper introduces different feature selection techniques and their applications.

1 INTRODUCTION

The system under study is the European interconnected power system, also known as UCTE/CENTREL. The system consists of the western European *Union for the Coordination of Transmission of Electricity* (UCTE) and the *central European power system* (CENTREL), which includes the central European countries Poland, Hungary, the Czech and Slovak Republic. Due to the recent integration of the CENTREL power system, the European network has grown rapidly. Further extensions, e.g. in the Balkan area, are under investigations.

The integration of the two large power systems (UCTE and CENTREL) led to a different stability behavior. Although the European network is strongly meshed, it includes parts with high power concentration, which could swing against each other. Interarea oscillations are observed particularly when two or more net groups in the power system (i.e. power

supply companies) exchange energy. These so-called interarea oscillations are slow damped oscillations with quite low frequencies.

In the European system, small-signal stability is largely a problem of insufficient damping of these oscillations [4], [5].

With the deregulation of the electricity market in Europe, the utilities are allowed to sell their generated power outside their traditional borders and compete directly for customers.

For economical reasons, the operators are often forced to steer the system closer to the stability limits. Thus, the operators need different computational tools for system stability. These tools must be accurate and fast to allow on-line stability assessment.

The small-signal stability method, the modal analysis, is based on the computation of eigenvalues and eigenvectors [1]. The inter-area modes are associated with the swinging of many machines in one part of the system against machines in other parts. In the European case, three global modes (eigenvalues) are of particular importance because when they lack damping the whole system starts to oscillate. For example, load flow situations including large power transits between Spain, Portugal, Poland or some

Balkan States lead very often to a weakly damped power system.

Figure 1 shows 3 dominant eigenvalues in the complex plain. These eigenvalues are interesting because they show low frequencies, which identify them as inter-area eigenvalues. The slant lines in the figure characterize constant damping in the range of 0% to 20%. For many different load flow situations, these eigenvalues remain in the stable region, but in some cases they shift to the low damping region and can cause system instability.

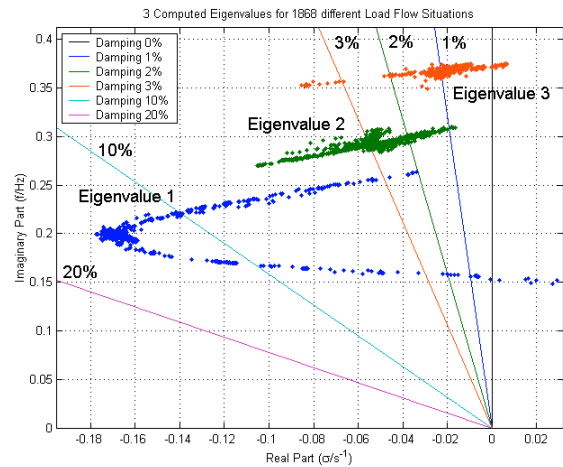


Figure 1: Changes of Dominant Eigenvalues under 1,868 Different Load Flow Situations

2 APPROACH

The computation of the small signal stability is a time consuming process for large networks because it includes the load flow computation, the linearization at the operating point, and the eigenvalue

computation. Moreover, it requires the knowledge of the complete system data. Thus, it is not suitable for on-line applications.

An alternative method is to use a neural network (NN) trained with off-line data for different load flow conditions. By using NN, a fast computation of the eigenvalues is possible, providing that the network is properly designed. For on-line applications, the NN predicts the dominant eigenvalues based on the current operating conditions.

The off-line data can be generated by simulating various load flow situations using a model of the UCTE/CENTREL power system. Hereby, the generation of net groups in the power system is changed to create diverse load flows between the different net groups.

Each new load flow situation in the network provides a new pattern for NN training and the basic challenge is to simulate load flow cases that are highly correlated with the system stability.

Another advantage of the NN is that it can be properly trained with few input features. This is also important considering that due to increasing competition utilities may not share essential information. Only very few features are commonly available such as the transmitted power or the generation of

complete net groups. Information about single generators or transmission lines is usually not available.

Once it is trained, the NN can predict the eigenvalues within milliseconds. However, the key issue is to find the best-input features that describe the system under study. These input features have to be measurable and need to contain as much information as possible about the small-signal stability.

The principal applicability of NN for stability prediction has been proven in previous works about the large-scale dynamic model of the UCTE/CENTREL power system [6], [7].

3 FEATURE REDUCTION

The entire data for the UCTE/CENTREL system include features for power equipment such as the transmission lines, transformers, generators, and loads. Hence, there is a large number of features in such an extensive power system. The size of this feature set creates the bottleneck problem for NN training. Therefore, feature extraction or selection techniques are indispensable for NN based small-signal stability assessment.

First, a pre-selection is performed by engineering judgment, whereby only the available and measurable features are used. After pre-selection, the size of these feature sets can be reduced using a reduction technique.

In this study, the selected features are:

- Total generated real and reactive power in each net group
- Real and reactive power transmitted between neighboring net groups
- Voltage and voltage angle on generators, loads, and transmission lines between neighboring net groups

The total power generated in one net group is the sum of all generator power within this net group, and the power flow between two neighboring net groups is the sum of power over all transmission lines between them.

However, the total number of all pre-selected features is 4,379, which is still too extensive for NN training. Therefore, the next sections will introduce some feature selection techniques for further reduction.

In contrast to feature selection methods, feature extraction methods are not applied in this study. On one side, they lead to highly accurate reduction results, but on the other side the physical meaning of the features is lost after reduction. Both techniques have already been applied and

compared previously and for this reason, only more advanced selection techniques are investigated in this study [7].

4 APPLIED SELECTION TECHNIQUES

In literature, one can find many different techniques suggested for feature selection. But as a matter of fact, there are only few methods applicable due to special constraints for the given problem.

The following example using correlation as measure of goodness for selection may help to understand this statement: If the correlation between eigenvalues and the total generated power in one net group is relative low, the correlation between eigenvalues and a set of some total generated power features might be much higher.

This is reasonable considering the fact that eigenvalues do not depend on the generated power of a single net group but on the load flow scenario in the entire power system including more than one net group.

Therefore, the selection using correlation as measure of goodness is focused on the canonical correlation, which computes the correlation between two groups of features.

Another applicable technique for feature selection are clustering methods [7]. Good results can be obtained by a technique called principal feature analysis (PFA), which is a selection by clustering, but on the base of transformed feature vectors reduced in dimension. The reduction is carried out with the help of principal component analysis (PCA) [8].

However, these techniques can be expanded to a multiple step selection (MSS).

Hereby, only small homogenous subsets of features are reduced in the first step using the PFA method. Then, the results are added to a new set, which is reduced in a second step.

Advantages of this technique are combinations of different selection methods for step 1 and step 2.

5 DATA PREPROCESSING

The first step in data reduction is the preprocessing of the data. Data preprocessing can be necessary to equal the data set. The data will be normalized, which is a linear transformation of each feature in the data set to obtain a zero mean and unit variance.

The initial feature matrix \mathbf{X} is defined as

$$\mathbf{X} = \begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{p1} & \cdots & x_{pn} \end{bmatrix} \quad (1)$$

whereby p is the number of patterns and n is the number of features. Thus, the column vectors \mathbf{x}_j are given by

$$\mathbf{x}_j = (x_1, x_2, \dots, x_p)^T \quad (1 \leq j \leq n) \quad (2)$$

However, the standardized features \mathbf{f} are computed by

$$\mathbf{f}_j = \frac{\mathbf{x}_j - \bar{x}_j}{\sigma_{x_j}} \quad (3)$$

Hereby, the standard deviation of \mathbf{x}_j is given by

$$\sigma_{x_j} = \sqrt{\frac{1}{p-1} \cdot \sum_{i=1}^p (\mathbf{x}_j(i) - \bar{x}_j)^2} \quad (4)$$

and the mean value of \mathbf{x}_j is

$$\bar{x}_j = \frac{1}{p} \cdot \sum_{i=1}^p \mathbf{x}_j(i) \quad (5)$$

Therefore, let \mathbf{F} be the standardized feature matrix of dimension $p \times n$, whereby n is the number of the original feature vectors and p is the number of patterns.

$$\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_n] \quad (6)$$

$$\mathbf{f}_i^T = [f_{i1}, f_{i2}, \dots, f_{in}] \quad i = 1, \dots, n \quad (7)$$

6 CANONICAL CORRELATION

In some applications it might be useful to compute the correlation not between two given features, but between two groups of features. In this case, the well-known correlation coefficient is not applicable since it does not regard the correlation between groups.

A solution to this problem is provided by the canonical correlation method (CCM) [2]. Hereby, a canonical correlation coefficient is computed, which determines the correlation between two groups of features.

Given are the standardized feature matrices \mathbf{F}_x and \mathbf{F}_y including n and m features, respectively. The matrices include p pattern and the empirical covariance matrices can be computed by

$$\mathbf{C}_x = \frac{1}{p-1} \cdot \mathbf{F}_x^T \cdot \mathbf{F}_x \quad (8)$$

$$\text{and} \quad \mathbf{C}_y = \frac{1}{p-1} \cdot \mathbf{F}_y^T \cdot \mathbf{F}_y \quad (9)$$

The matrix \mathbf{C}_{xy} determines the covariance between the features in \mathbf{F}_x and \mathbf{F}_y , the matrix is given by

$$\mathbf{C}_{xy} = \frac{1}{p-1} \cdot \mathbf{F}_x^T \cdot \mathbf{F}_y \quad (10)$$

In the next step, the matrix \mathbf{Q} can be obtained

$$\mathbf{Q} = \mathbf{C}_x^{-1} \cdot \mathbf{C}_{xy} \cdot \mathbf{C}_y^{-1} \cdot \mathbf{C}_{xy}^T \quad (11)$$

Then, the n eigenvalues of \mathbf{Q} are given by λ_i , whereby the largest eigenvalue is denoted λ_G . This eigenvalue provides an estimation for the maximal canonical correlation r .

$$r = \sqrt{\lambda_G} \quad (12)$$

However, in contrast to the correlation coefficient, the computation of canonical correlation coefficients r for different combinations of input feature sets allows a much better investigation of the impact of input features to the eigenvalues. Therefore, this feature selection method can be applied on the given problem.

While the merit of this technique is proven by the high reduction rate and the accurate results after NN training, the algorithm can only be implemented as a Monte-Carlo-Method. The features in the original set are not ordered in any way but added number by number. Therefore, a systematic search algorithm or strategy cannot be used to find a combination of features with a high canonical correlation.

Considering that about 60 features need to be selected from a total set of more than 4000 features, only very few combinations can be computed. Therefore, the solution found by the Monte-Carlo-Algorithm is only a first approximation of a set of selected features.

7 PRINCIPAL FEATURE ANALYSIS

One solution for feature reduction by the canonical correlation method is given in Figure 2 and Figure 3.

These figures show the training and testing results of a NN trained with the selected feature set.

The eigenvalues marked with crosses are the ones used as targets. The circles are the NN outputs. The targets and the NN outputs are connected by lines.

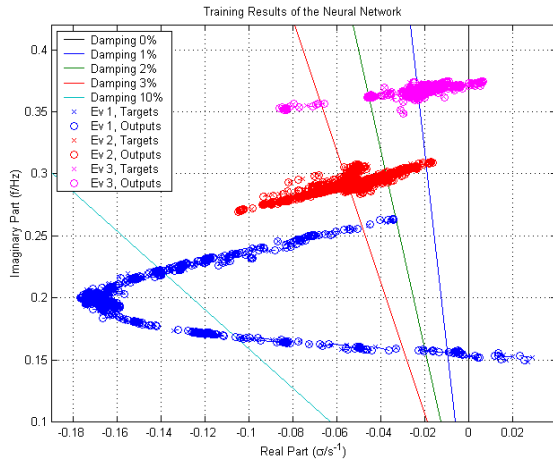


Figure 2: Training Results of the NN after Training with Features selected by CCM

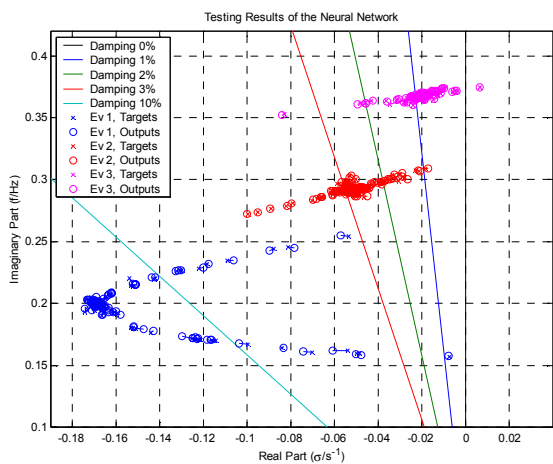


Figure 3: Testing Results of the NN after Training with Features selected by CCM

The PCA technique, a feature extraction method, leads to the best possible reduction results, which are a good representation of the original data [8]. This is obvious, when PCA is analyzed in detail. The projection onto a smaller number of orthonormal axes leads to a coordinate system with axes of largest spread.

The PCA technique is characterized by a high reduction rate and a minimal loss of information. Moreover, the technique is fast and can be applied on large data sets. It's only lack is the fact, that the projection to the lower dimensional space transforms the original features into new ones under loss of their physical meaning.

Because feature selection techniques do not have this disadvantage, the combination of PCA and feature selection will have both benefits.

One way of combining these methods is described in this section.

First, the empirical covariance matrix \mathbf{C} of the normalized feature matrix \mathbf{F} , given in equation (6), is computed

$$\mathbf{C} = \frac{1}{p-1} \cdot \mathbf{F}^T \cdot \mathbf{F} \quad (13)$$

If \mathbf{T} is a $n \times n$ matrix including the eigenvectors of the covariance matrix \mathbf{C} ,

the diagonal variance matrix Σ^2 is given by

$$\Sigma^2 = \mathbf{T}^T \cdot \mathbf{C} \cdot \mathbf{T} \quad (14)$$

Σ^2 includes the variances σ_x^2 . Noticeable is hereby, that the eigenvalues λ_k of the covariance matrix \mathbf{C} are equal to the elements of the variance matrix Σ^2 , whereby the standard deviation σ_k is also called singular value of \mathbf{F} :

$$\sigma_k^2 = \lambda_k \quad (1 \leq k \leq n) \quad (15)$$

The n eigenvalues of \mathbf{C} can be determined and sorted in descending order $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$. While \mathbf{T} is a n -dimensional matrix whose columns are the eigenvectors of \mathbf{C} , \mathbf{T}_q is a $n \times q$ matrix including q eigenvectors of \mathbf{C} corresponding to the q largest eigenvalues of \mathbf{C} .

The value of q determines the size of the new dimension and is smaller than n . It also determines the retained variability of the features, which is the ratio between the first q eigenvalues and the sum of all n eigenvalues.

$$v = \frac{\sum_{i=1}^q \lambda_i}{\sum_{i=1}^n \lambda_i} \quad (16)$$

The n rows of \mathbf{T}_q can be assumed as vectors \mathbf{v}_i^T , which represent the projection

of the i -th feature of \mathbf{F} onto the lower q -dimensional space. However, the q elements of \mathbf{v}_i^T correspond to the weights of the i -th feature of \mathbf{F} on the q axes of the subspace. From [3] and [8] follows, that original features, which are highly correlated, have similar absolute value weight vectors \mathbf{v}_i^T .

In other words, two independent features will show a high divergence of their corresponding weight vectors \mathbf{v}_i^T . On the other side, two identical features will lead to identical absolute weight vectors.

Once the weight vectors \mathbf{v}_i^T are computed, they can be clustered to $p > q$ groups. The number of cluster needs to be greater than the number of weight vectors \mathbf{v}_i^T to obtain the same variability as the PCA. Usually, 1-5 additional dimensions are needed.

Because of the similarity between the features within a cluster, one of them can be selected and the others can be treated as redundant information. The feature in one cluster, which is closest to the centroid of this cluster, will be chosen as a principal feature. Thus, a group of p features will remain as principal features.

Hereby, the centroid \mathbf{c} of a cluster including n vectors \mathbf{a} is defined as

$$\mathbf{c} = \frac{1}{n} \cdot \sum_{i=1}^n \mathbf{a}_i \quad (17)$$

and the distance d between centroid \mathbf{c} and vector \mathbf{a} is computed by

$$d(\mathbf{a}, \mathbf{c}) = \|\mathbf{c} - \mathbf{a}\| \quad (18)$$

The PFA method was applied to the given set of more than 4,000 features. First, the set was reduced to a subset using only the 55 largest principal components. Considering the retained variability in the set, which is 98.3 %, the number of 55 principal components is sufficient for the given problem. Then, the corresponding eigenvectors were clustered into 60 groups. These groups have been used for NN training and the results are shown in Figure 4 and Figure 5.

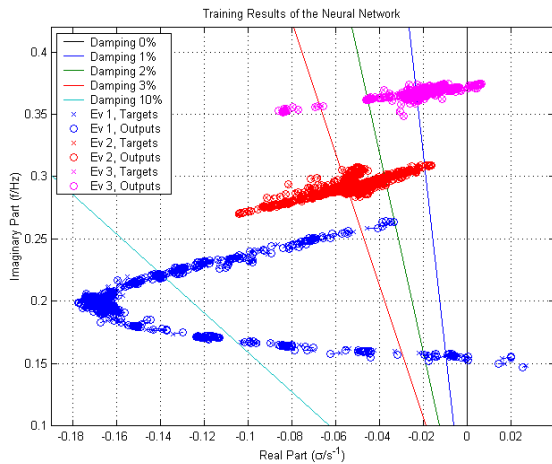


Figure 4: Training Results of the NN after Training with Features selected by PFA

However, it is even possible to skip the PCA computation in PFA, which means, that the features are clustered directly from the beginning. In fact, the computation of

the principal components is not necessary to obtain accurate reduction results [7].

But in case of large feature sets, the k-means cluster algorithm results in increasing inaccuracy. This is why the PFA technique is used which results in transformed and reduced dimensionality feature vectors. However, these vectors correspond to the original feature vectors. Therefore they are more suitable for a fast and accurate clustering.

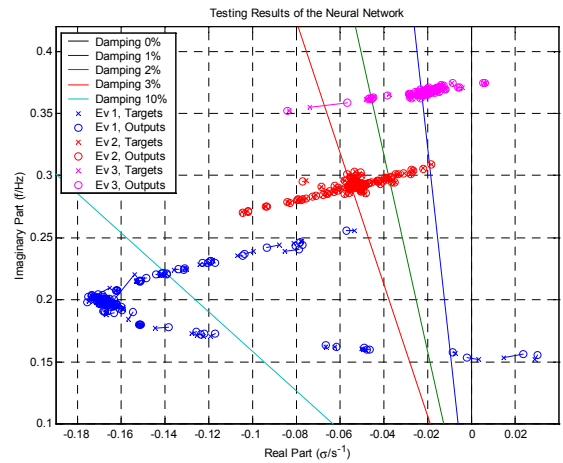


Figure 5: Testing Results of the NN after Training with Features selected by PFA

8 MULTIPLE STEP SELECTION

Multiple step selection (MSS) is the continuation of the PFA technique. In MSS, the PFA method is applied more than one time or different techniques are combined following each other.

This is obvious comparing to CCM and PFA. These techniques reduce the total set in one selection step, but this can be time consuming (CCM) or problematic regarding the used method (PFA), where the cluster algorithm works fine only with small sets.

However, since the total set of features is highly inhomogeneous including features with different physical meaning, the split into two or even more selection steps might be helpful. A first selection is done only within homogenous groups. Then, a second selection step built on the first one is performed to obtain the final set of features.

In this study, the selection was made in 3 steps. Because the total set of features is extremely inhomogeneous as mentioned before, the set was split into 3 homogeneous subsets including power features (total generated power in each net group and power transmitted between net groups), voltage features (voltages on generators, loads, and transmission lines between neighboring net groups), and the corresponding voltage angle features.

In the first step, which is similar to a pre-selection, the correlation between input features and outputs is separately for the 3 subsets computed. About 10% of the

features in any subset, which show least input/output correlation, are sorted out.

The second step of the selection concentrates on the redundancy in the subsets. Therefore, PFA is applied to cluster the subsets to a smaller size. In this study, the power features were clustered into 50 groups, the voltage features into 10 groups, and the voltage angle features into 100 groups. The number of clusters is variable and can be defined by the user. Different combinations lead to different results and an optimum needs to be found.

From any group, one feature closest to the centroid is selected and then these features are added to a set of 160 features.

In the third step, these sets were reduced again by the PFA method to obtain the desired number of 60 remaining features for NN training.

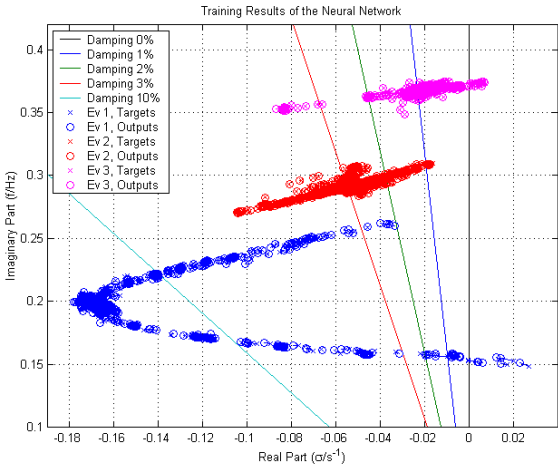


Figure 6: Training Results of the NN after Training with Features selected by MSS

The results of the NN training and testing are shown in Figure 6 and Figure 7, respectively.

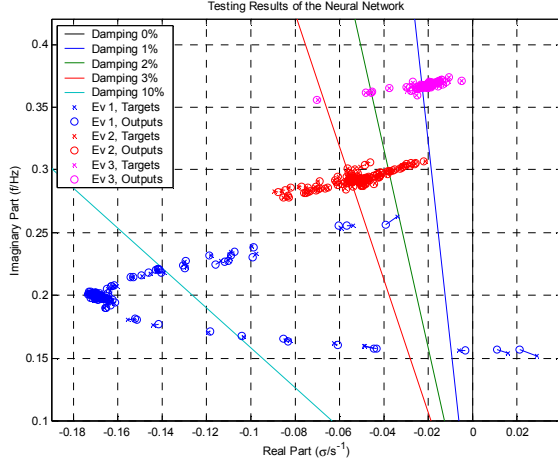


Figure 7: Testing Results of the NN after Training with Features selected by MSS

9 CONCLUSIONS

Any of the introduced techniques show good results as shown in the figures before. However, the data include much redundancy and so they are reducible up to a high ratio. This fact leads to a reduction ratio of 98.6 %, which results from a set of 60 features selected from a total set of 4,379 features. For more detailed comparison of the different techniques, an error function needs to be defined. In this study, the error E_λ for a given eigenvalue $\lambda = \sigma + j\omega$ determines the normalized distance between targets and outputs for

one pattern. It is defined by the following equation:

$$E_\lambda = \frac{\sqrt{(\sigma_{\text{output}} - \sigma_{\text{target}})^2 + (\omega_{\text{output}} - \omega_{\text{target}})^2}}{\sqrt{\sigma_{\text{target}}^2 + \omega_{\text{target}}^2}} \quad (19)$$

This error can be computed for all testing patterns. Table 1 shows the mean and the standard deviation for 3 eigenvalues and all patterns.

	Mean Error	Std. Deviation
CCM	0.11%	0.38%
PFA	0.12 %	0.47 %
MSS	0.07 %	0.24 %

Table 1: Comparison of the Applied Techniques using Mean Error and Standard Deviation of the Error Function defined by Equation (19)

Table 1 allows to compare the different techniques, but the main criterion for the stability assessment is the prediction of the damping coefficient. Therefore, a second error function E_ξ for the damping coefficient can be defined as follows by equation (20) and (21):

$$E_\xi = \frac{|\xi_{\text{output}} - \xi_{\text{target}}|}{\xi_{\text{target}}} \quad (20)$$

$$\xi = \frac{-\sigma}{\sqrt{\sigma^2 + \omega^2}} \quad (21)$$

Table 2 shows the mean and the standard deviation for 3 eigenvalues and all patterns using an error function regards the damping coefficient defined by equation (20).

	Mean Error	Std. Deviation
CCM	1.10 %	3.99 %
PFA	1.31 %	7.52 %
MSS	0.69 %	2.87 %

Table 2: Comparison of the Applied Techniques using Mean Error and Standard Deviation of the Error Function defined by Equation (20)

Table 1 and 2 allow to compare errors of the 3 applied techniques. The results of CCM and PFA are accurate and applicable for on-line stability assessment and fast eigenvalue prediction, but the results of MSS are much more accurate.

The reason can be found in the application of more than one selection step, which is recommendable for selection problems with a large number of inhomogeneous features.

Another way of comparison is the time used for selection. Therefore, Table 3 shows the time comparison of the applied techniques.

	Reduction Time
CCM	> 4 h
PFA	19 min.
MSS	14 min.

Table 3: Comparison of the Reduction Time of the Applied Techniques (Pentium 4, 1.7 GHz)

CCM needs much time to find an appropriate selection of features. This is

due to the use of a Monte Carlo-Algorithm. But the better the required solution, the more time is needed to compute. For small data sets, the CCM can be used within a maintainable duration. But with increasing number of features, the time required to find an appropriate solution will increase superproportional.

PFA uses as well much time for reduction of large feature sets because of the eigenvalue computation of a large feature matrix. For small feature sets, the reduction time will be very fast.

The reduction time of MSS compared to PFA is almost the same. The eigenvalue computation in MSS is much faster than in PFA because the eigenvalues are separately computed for the 3 subsets. In fact, the reduction time depends mostly on the cluster algorithm, which clusters much larger feature sets than in PFA.

However, even if there is not much difference in computation time between PFA and MSS, the accuracy of MSS is much higher. Moreover, this technique is absolutely variable and allows a high number of different combinations and possibilities.

Therefore, MSS techniques are highly recommended for feature selection applications for NN based small-signal stability assessment.

10 REFERENCES

- [1] P. Kundur, Power System Stability and Control, McGraw-Hill, New York, 1994
- [2] Hartung, Joachim, Multivariate Statistik : Lehr- und Handbuch der angewandten Statistik, München, 1992
- [3] I.T. Jolliffe, Principal Component Analysis, Springer-Verlag, New York, 1986
- [4] U. Bachmann, I. Erlich and E. Grebe, "Analysis of interarea oscillations in the European electric power system in synchronous parallel operation with the Central-European networks", IEEE PowerTech, Budapest 1999
- [5] H. Breulmann, E. Grebe, M. Lösing, W. Winter, R. Witzmann, P. Dupuis, M.P. Houry, T. Margotin, J. Zerényi, J. Dudzik, J. Machowski, L. Martín, J.M. Rodríguez, E. Urretavizcaya, "Analysis and Damping of Inter-Area Oscillations in the UCTE/CENTREL Power System", CIGRE 38-113, Session 2000
- [6] S.P. Teeuwsen, A. Fischer, I. Erlich, M.A. El-Sharkawi, "Assessment of the Small Signal Stability of the European Interconnected Electric Power System Using Neural Networks", LESCOPE 2001, Halifax, Canada, June 2001
- [7] S.P. Teeuwsen, I. Erlich "Feature Reduction for Neural Network based Small-Signal Stability Assessment", PSSC 2002, Sevilla, Spain, June 2002
- [8] I. Cohen, Q. Tian, X. Zhou, T. S. Huang, "Feature Selection and Dimensionality Reduction Using Principal Feature Analysis", submitted to IEEE International Conference on Image Processing (ICIP'02), Rochester, New York, Sep. 22-25, 2002