

A Comparative Study of Neural Network Based Feature Extraction Paradigms

Boaz Lerner*, Hugo Guterman#, Mayer Aladjem#, and Its'hak Dinstein#

*University of Cambridge Computer Laboratory, New Museums Site, Cambridge CB2 3QG, UK

#Department of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel

Published in *Pattern Recognition Letters*, vol. 20(1), pp. 7-14, 1999.

Abstract

The projection maps and derived classification accuracies of a neural network (NN) implementation of Sammon's mapping, an auto-associative NN (AANN) and a multilayer perceptron (MLP) feature extractor are compared with those of the conventional principal component analysis (PCA). Tested on five real-world databases, the MLP provides the highest classification accuracy at the cost of deforming the data structure, whereas the linear models preserve the structure but usually with inferior accuracy.

Keywords: Auto-associative neural network; Classification; Data projection; Feature extraction; Multilayer perceptron; Principal components; Sammon's mapping;

1. Introduction

The process of mapping original features (measurements) into fewer, more effective features is termed feature extraction. In each of the existing feature extraction methods (Fukunaga, 1990, Ch. 9-10), a mapping f transforms a d -dimensional pattern X to an m -dimensional pattern Y ($m < d$), $Y=f(X)$, such that a criterion J is optimised. Examples of such a criterion are the mean square error (used for example in PCA) and the inter-pattern distance error used in Sammon's mapping. The mapping f is determined from among all the transformations g as one that satisfies,

$$J\{f(X)\} = \min_g J\{g(X)\}. \quad (1)$$

Different mappings have different functional forms of g and different criteria to optimise. Mao and Jain (1995) group feature extraction methods into four categories: supervised versus unsupervised and linear versus non-linear. Examples of common feature extraction paradigms are (see Table 1) linear discriminant analysis (LDA)- supervised linear- (Fukunaga, 1990, Ch. 10); PCA- unsupervised linear-

*Corresponding author. Email: boaz.lerner@cl.cam.ac.uk

(Fukunaga, 1990, Ch. 9); and Sammon's mapping- unsupervised non-linear- (Sammon, 1969). It is also common to group feature extraction methods into exploratory data projection paradigms, which enable high-dimensional data visualisation for better understanding of the data structure and paradigms for classification, in which it is advantageous to reduce the number of features and therefore to decrease the computational complexity of the classification.

Recently, a large number of NN models and learning mechanisms for feature extraction have been proposed (Bishop, 1995, Ch. 8; Bourland and Kamp, 1988; Lowe and Tipping, 1996; Mao and Jain, 1995). The NN-based feature extraction paradigms adapt to changing environments and afford the possibility of relatively easy hardware implementation. They can even overcome the drawbacks of classical algorithms or enhance the performance of the classification (Lerner *et al.*, 1998; Mao and Jain, 1995). The MLP (when acting as a feature extractor) and the AANN respectively embed supervised and unsupervised (Table 1) mappings of the input feature space in their hidden layers. An NN implementation of Sammon's mapping which has recently been suggested by Mao and Jain (1995) and Kohonen's self-organizing map (SOM) (Bishop, 1995, Ch. 5) are other examples of NN-based feature extraction paradigms.

In this study, NN-based feature extraction paradigms, namely MLP, AANN and NN implementation of Sammon's mapping, are evaluated for both exploratory data projection and classification. The projection maps and derived classification accuracies of these methods are compared with those of the non-NN-based principal component (PC) feature extractor. The four paradigms are representatives of different families of models (Table 1) which are also usually used for different tasks (data projection or classification). To the best of our knowledge, only a few empirical comparative studies of NN-based feature extraction paradigms have been made (Lerner *et al.*, 1998; Mao and Jain, 1995). The paradigms in Mao and Jain (1995) are compared only for exploratory data projection and two-dimensional classification and in Lerner *et al.* (1998) only for one database. In the current study, however, we extract an arbitrary number of projections to enable the application of the mapping also to high-dimensional classification and extend the experiments of Lerner *et al.* (1998) by comparing the

paradigms using five real-world databases. We believe that empirical studies such as the one performed here are, and especially in real-world problems, the only way to find the “best” mapping for a data set. This is especially true where there is no *a priori* knowledge of the underlying distribution of the data and/or the preferable criterion J , and/or where randomly initialised paradigms are involved. Finally, this paper also suggests a methodology for experimenting with NN-based feature extraction paradigms.

Section 2 of the paper briefly introduces the three NN-based feature extraction paradigms. Sections 3 and 4 present the experiments and their results, respectively, while Section 5 summarises the work.

2. Feature Extraction Paradigms

The NN-based feature extraction paradigms accomplish feedforward projections. In these projections, the feature space (which is also the network input) is represented compactly by the AANN or MLP hidden units or the output units of the NN implementation of Sammon’s mapping. The methods are compared with the conventional PCA (Fukunaga, 1990, Ch. 9) which maps the feature space onto the space spanned by the eigenvectors corresponding to the largest eigenvalues of the covariance matrix of the mixture density. Only brief descriptions of the paradigms are given here and the interested reader is referred to one of the NN textbooks (*e.g.*, Bishop, 1995).

A. The AANN

Considering a two-layer AANN as an encoder-decoder mechanism, the network is forced to perform an identity mapping through a deliberately small hidden layer. Forcing the mapping to proceed through a small hidden layer ensures efficient encoding. Hence, an AANN has a configuration of $d:m:d$ with d units in both the input and output layers and $m < d$ hidden units in the hidden layer. This network is usually trained by minimising the sum-of-squares error (Bishop, 1995, Ch. 8):

$$E = \frac{1}{2} \sum_{n=1}^N \sum_{k=1}^d \{y_k(\mathbf{X}^n) - x_k^n\}^2, \quad (2)$$

where $y_k(\mathbf{X}^n)$ represents the k th output of the n th input vector $\mathbf{X}^n = (x_1^n, \dots, x_k^n, \dots, x_d^n)$ and N is the number of training patterns. At the global minimum of the error, the network performs a projection onto

the m -dimensional sub-space that is spanned by the first m PCs of the data (Bourland and Kamp, 1988). As Bourland and Kamp (1988) claimed and Cottrell *et al.* (1987) experimentally validated for image compression, non-linearity in the hidden units of the two-layer AANN is useless and the model implements PCA. Since it is interesting to test this conclusion in other domains as well (Kramer, 1991), a comparison of the two-layer AANN with PCA using various sources of data is made as part of this study. The performance based on the PCA also provides a reference for that based on the randomly initialised AANN. Finally, the restriction of the two-layer AANN to solve successfully only linear problems may be removed using an AANN of more than two layers (Kramer, 1991).

B. The MLP feature extractor

The MLP hidden units can be used as an implementation of a non-linear projection of the feature space. When their number is appropriately selected, the patterns represented in the projected space spanned by the hidden units are more easily separated by the network output layer and those hidden units may simultaneously supply means of data projection. The number of input units of the MLP feature extractor is specified to be the number of the original features, the number of output units to be the number of pattern classes and the hidden layer dimension is set according to the task- either exploratory data projection or classification. The criterion to minimise is generally the sum-of-squares error (Bishop, 1995, Ch. 4):

$$E = \frac{1}{2} \sum_{n=1}^N \sum_{k=1}^d \left\{ y_k(X^n) - t_k^n \right\}^2, \quad (3)$$

where t_k^n is the k th target value for the n th input pattern X^n .

C. Sammon's mapping

In Sammon's mapping (Sammon, 1969), the mapping error to minimise is the mean-square error between inter-pattern distances in the input (d_{ij}^*) and projected (d_{ij}) spaces, defined as:

$$E = \frac{1}{\sum_{i < j}^N d_{ij}^*} \sum_{i < j}^N \frac{[d_{ij}^* - d_{ij}]^2}{d_{ij}^*} \quad (4)$$

where i and j are indices of two out of N patterns. The distance between two patterns is commonly assessed with the Euclidean metric.

We use a two-layer perceptron implementation of Sammon's mapping (Lerner *et al.*, 1998) based on that of Mao and Jain (1995). The number of input units in the network is set to be the feature space dimension d and the number of output units is specified as the extracted feature space dimension m . Weights are updated using a gradient descent method that minimises the mapping error (Eq. 4). Following Lerner *et al.* (1998) we exploit the eigenvectors of the sample covariance matrix estimated from the training data set to establish the columns of the network's initial input-hidden weight matrix. By inspecting the dependence of the data variance on the number of eigenvalues, we choose *a priori* a sufficient number of weight vectors (eigenvectors) to set the number of hidden units and to initialise the mapping. Compared with the common random initialisation, fewer experiments are required and the network configuration can be set precisely without the commonly used trial-and-error experimentation. Finally, although Sammon's mapping is usually used for two-dimensional data projection (Sammon, 1969; Mao and Jain, 1995), we wish to confirm here the empirical result (Lerner *et al.*, 1998) that the input space can be mapped into an arbitrary number of projections to enable the use of the mapping in classification.

3. The Experiments

A. The methodology

We compare the NN-based feature extractors using data sets taken from five databases. The first database was derived by Lerner *et al.* (1995) from chromosome images, which were gathered at the Soroka Medical Center, Beer-Sheva, Israel. The chromosome patterns were represented by 64 density profile (d.p.) features (integral intensities along sections perpendicular to the medial axis of the chromosome) which were found by Lerner (1998) to be among the features that provide the best

discrimination in chromosome analysis. The second database is also obtained from chromosome images but since it consists of geometrical features, patterns of the database are independent of patterns of the former database, which consists of intensity features. The chromosome patterns are represented by four geometrical features, *i.e.*, the centromeric index (the ratio of the short arm length to the total length), length, perimeter and area of the chromosome (Lerner *et al.*, 1995).

The third database is extracted from satellite images of the STATLOG project (Michie *et al.*, 1994). Each pattern in the database corresponds to intensities measured in four spectral bands (from the green, red and two infra-red regions) of a 3x3 neighbourhood of pixels of a sub-scene image and hence consists of 36 features (nine pixels and four bands). The fourth database comes from the Research Assessment Exercise (RAE) of 72 subject areas in all higher education institutions in the UK. Variables such as the number of active researchers, postgraduate students, the values of grants awarded and number of publications formed a 79-dimensional database which is used to assess research on a scale of 1 to 5 (research ratings) in each subject area at each institute (Lowe and Tipping, 1996). The last database is the much-analysed iris data (Merz and Murphy, 1998) where the patterns are represented by four attributes (sepal and petal lengths and widths).

Since in three of the databases- chromosome (d.p.), chromosome (geometrical) and RAE- there are approximately one hundred patterns in each class, for comparison we also extract one hundred patterns per class from the satellite data and use all the fifty patterns per class which are available in the iris data. The experiments show (Section 4) that this choice does not lead to the “curse of dimensionality” even in those high-dimensional databases. The patterns of each database belong to one of three classes, which are chromosome types “13”, “19” and “x” in the first two databases, soil types in the third database, the subjects Physics, Chemistry and Biology in the fourth one and three iris types in the last database.

In summary, five databases are used to obtain data sets of three classes, each class consists of one hundred (fifty for the iris data) patterns with dimensions of 64 (chromosome (d.p.)), 4 (chromosome (geometrical)), 36 (satellite), 79 (RAE) and 4 (iris). Based on the methodology of Lerner *et al.* (1998),

the four paradigms map the patterns to create projection maps and to train and test a classifier. In the classification experiments, fifteen data sets are derived randomly from each database, each of which is partitioned into training (90%) and test (10%) sets (the holdout method (Fukunaga, 1990, Ch. 5)). Each feature extraction paradigm is applied to all of these data sets and the classification accuracy is averaged over the fifteen sets and ten classifier random initialisations (see Section 3C).

B. The paradigms- configurations and parameters

The two-layer AANN and MLP are trained by the backpropagation (BP) algorithm (Bishop, 1995, Ch. 4). Linear and logistic sigmoid hidden units are used for the AANN and MLP, respectively. The input vectors are 64, 4, 36, 79 or 4-dimensional for the five databases, respectively. The hidden layer dimension is kept lower than the input dimension during the classification experiments to ensure efficient mapping of the AANN (Section 2A), *i.e.*, in the range 1-10 for the chromosome (d.p.), satellite and RAE databases and in the range 1-3 for the chromosome (geometrical) and iris databases. It is set at two during the exploratory data projection experiments.

The two initial weight matrices of both the AANN and MLP paradigms are randomly selected and the classification accuracy is averaged over 100 experiments with ten input-hidden and ten hidden-output initial weight matrices. According to Lerner *et al.* (1998) we set the parameters of the AANN, MLP and NN implementation of Sammon's mapping to be: a learning rate of 0.1 (AANN, MLP) or 0.9 (Sammon's), a momentum constant of 0.95 (AANN, MLP) or 0.5 (Sammon's), and a training period of 500 (AANN, MLP) or 30 (Sammon's) epochs.

The number of hidden units in the NN implementation of Sammon's mapping is two to avoid overfitting. Eigenvectors corresponding to the largest eigenvalues define the initial input-hidden weight matrix of the implementation, whereas the initial hidden-output weight matrix is randomly selected (Section 2C) and the classification accuracy is averaged over four experiments. The network output dimension is set at two for exploratory data projection, whereas it is changed in the range 1-10 and 1-3 for classification. The input layer is 64, 4, 36, 79 or 4-dimensional as for the AANN and MLP.

Finally, the PC feature extractor employs the first two eigenvectors in the projection experiments and the eigenvectors corresponding to the first 1 to 10 (chromosome (d.p.), satellite and RAE) and 1 to 3 (chromosome (geometrical) and iris) eigenvalues in the classification experiments.

C. The classifier

More complex architectures than the two-layer perceptron are not considered here as candidates for the classifier since we are concerned only with a comparative study of feature extraction paradigms. The number m of input units of the classifier is set by the projected space dimension; the number of hidden units is set at two; and the number of output units is 3 (*i.e.*, the number of classes). For a fair comparison, the classifier parameters (learning rate, momentum constant and training period), that are the same as those of the MLP feature extractor (Section 3B), remain unchanged for all the feature extraction paradigms. For example, training of the classifier is always performed for a fixed duration of 500 epochs (Section 3B). These configuration and parameters are checked to avoid over-training and to provide sufficient accuracy. Finally, the classification accuracy is averaged over ten classifiers each of which having randomly chosen initial weight matrices.

4. Experimental Results

A. Projection maps

Fig. 1 shows two-dimensional projection maps based on the PC feature extractor, Sammon's mapping, MLP and AANN for the second database of chromosomes (geometrical). Random initialisation of Sammon's mapping is preferred here because the PC based initialisation (which is used in the classification experiments) is found to yield very similar maps to those of the PC feature extractor (Mao and Jain, 1995). The maps in Fig. 1 are obtained using fifty test patterns per class. Producing the same maps for the case that is tested in the classification experiment (10% of the data set used for testing) is of less interest since only ten test patterns per class are available for the experiment. The evaluation of the projection maps is based on visual judgement, which is, in our opinion, the best

qualitative way to evaluate these maps, except for complex psychophysical experiments. A quantitative evaluation of the projections may appear to be “biased” towards one of the paradigms, as in *e.g.*, Mao and Jain (1995) who, when using the error of Sammon’s mapping (Eq. 4) for a quantitative evaluation of projection methods, ranked Sammon’s mapping as the “best” projection method. Visually analysed, the maps of the PC, AANN and Sammon’s mapping are “easier” for interpretation compared to the map of the MLP since they better preserve the data structure and cluster shape. However, it is suggested, based on these maps, that the MLP may lead to a more accurate classification because the ratio of the *between*-class scatter to the *within*-class scatter in this map is larger compared to the other maps. Repeating these projections several times for random training sets and initialisations, we can conclude that the projection maps for each specific database are typical and only small variations between experiments exist due to the above randomness.

B. Classification

We use the probability of correct classification of the test set to evaluate the classification accuracy based on the four feature extraction paradigms for the five databases. This probability is plotted in Fig. 2 for 1 to 10 projections for the chromosome (d.p.), satellite and RAE databases and 1 to 3 features for the chromosome (geometrical) and iris databases. As is shown in Fig. 2, classification based on features extracted by the MLP usually outperforms classification based on the other features, whereas classification based on the PCA is usually second best. As the PCA and AANN implementing the same mapping (Section 2A), the small difference (except for the RAE) between the results based on these two paradigms is attributed to the random initialisation of the AANN. Moreover, these two linear mappings (and especially the PCA) are comparable with the non-linear mappings for the chromosome (geometrical), satellite and RAE databases. We find that the accuracy of the classification is improved, due to feature extraction, for the chromosome (d.p.), satellite and RAE databases for every number of extracted features and the chromosome (geometrical) and iris databases for most instances. In those successful cases, only a low percentage of the original features is necessary to achieve the ultimate

accuracy. Finally, we relate the variability shown in Fig. 2 to the limited amount of data and not to the paradigms themselves.

5. Conclusions

This work is a more complete evaluation, on additional databases, of Lerner *et al.* (1998). Although projection using the non-linear MLP feature extractor distorts the data structure and inter-pattern distances, it is found here that this paradigm yields the highest classification accuracy among four feature extraction methods of different families. Similar results using other databases support this conclusion (Mao and Jain, 1995). Linear models, however, besides preserving the data structure are found to provide generally a fair, “cheap” alternative to non-linear models in achieving high classification accuracy, as has been also concluded in Michie *et al.*, (1994, Ch. 11). These models are especially useful for problems which are known or believed to be linearly separable or when considering the application of an iterative “expensive” non-linear model. In other instances, more than one paradigm may be required if both data projection and feature extraction are needed. Therefore, we suspect that the desire for a generic NN-based feature extraction paradigm for both data projection and classification may never be achieved, especially when experimenting with real-world problems. In those cases, presumably only a careful comparative empirical study, such as the one described here, may help. It would therefore be of interest to extend this study to applications of more than three classes and to other domains. Then, the linear methods may fail and it will be necessary to replace them with non-linear paradigms, *e.g.*, non-linear PCA and non-linear AANN.

Acknowledgement: This work was supported in part by the Paul Ivanier Center for Robotics and Production Management, Ben-Gurion University, Beer-Sheva, Israel. The authors thank Mr. Peter Lambert for English corrections and improving the paper’s readability and the anonymous referees for their most helpful suggestions.

References

- Bishop, C.M. (1995). Neural networks for pattern recognition. Oxford Press.
- Bourland, H. and Y. Kamp. (1988). Auto-association by multilayer perceptrons and singular value decomposition. Biol. Cybern. 59, 291-294.

- Cottrell, G. W., P. Munro and D. Zipser. (1987). Learning internal representations from gray-scale images: An example of extensional programming, in 9th Ann. Conf. Cognitive Sci. Soc. Hillsdale: Erlbaum, Seattle, 462-473.
- Fukunaga, K. (1990). Introduction to Statistical Pattern Recognition 2nd ed. Academic Press, New York.
- Kramer, M. A. (1991). Nonlinear principal component analysis using autoassociative neural networks. AICHE Journal 37, 233-243.
- Lerner, B., H. Guterman, I. Dinstein and Y. Romem. (1995). Medial axis transform based features and a neural network for human chromosome classification. Pattern Recognition 28, 1673-1683.
- Lerner, B., H. Guterman, M. Aladjem, I. Dinstein, and Y. Romem. (1998). On pattern classification with Sammon's nonlinear mapping- an experimental study. Pattern Recognition 31, 371-381.
- Lerner, B. (1998). Toward a completely automatic neural network based human chromosome analysis. IEEE Trans. Syst. Man Cyber. 28, Part B, 544-552, special issue on Artificial Neural Networks.
- Lowe, D. and M. E. Tipping. (1996). Feed-forward neural networks and topographic mappings for exploratory data analysis. Neural Computing and Applications 4, 83-95.
- Mao, J. and A. K. Jain. (1995). Artificial neural networks for feature extraction and multivariate data projection. IEEE Trans. Neural Networks 6, 296-317.
- Merz, C. J. and P. M. Murphy. (1998). UCI repository of machine learning databases. Irvine, CA: University of California, Department of Information and Computer Science.
- Michie, D., D. J. Spiegelhalter and C. C. Taylor. (1994). Machine Learning, Neural and Statistical Classification. Ellis Horwood, New York.
- Sammon, J. W. Jr. (1969). A nonlinear mapping for data structure analysis. IEEE Trans. Computers 18, 401-409.

Figure caption

Fig. 1. Two-dimensional projection maps of the four feature extraction paradigms applied to the chromosome (geometrical) database.

Fig. 2. The probability of correct classification of the test set for an increasing number of features extracted by the four paradigms and for five databases: (a) chromosome (d.p.), (b) chromosome (geometrical), (c) satellite, (d) RAE and (e) iris. The accuracy is compared with the original (*) 1-10 or 1-3 first features of each database.

Table caption

Table 1. Examples for common feature extraction paradigms

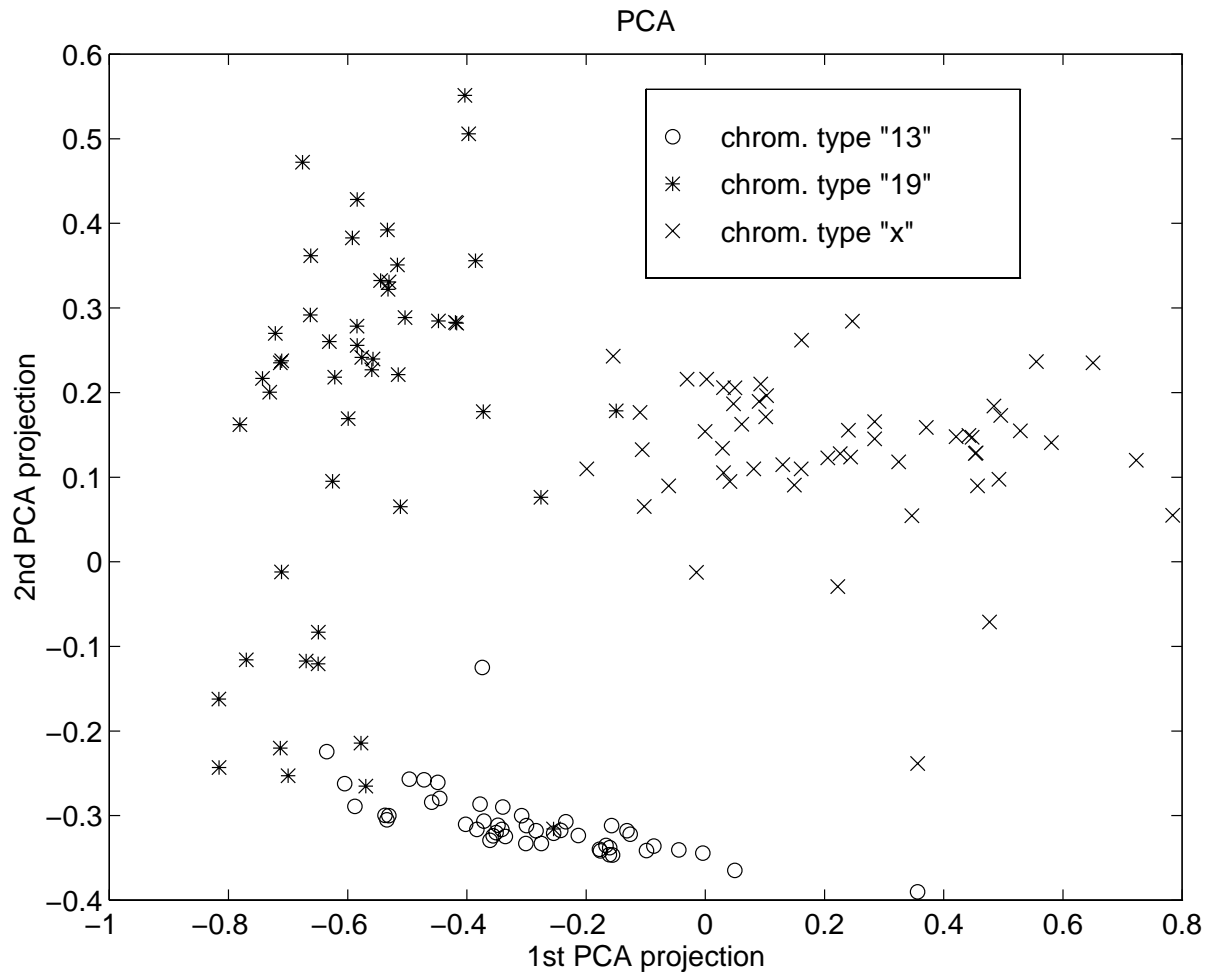


Fig. 1a

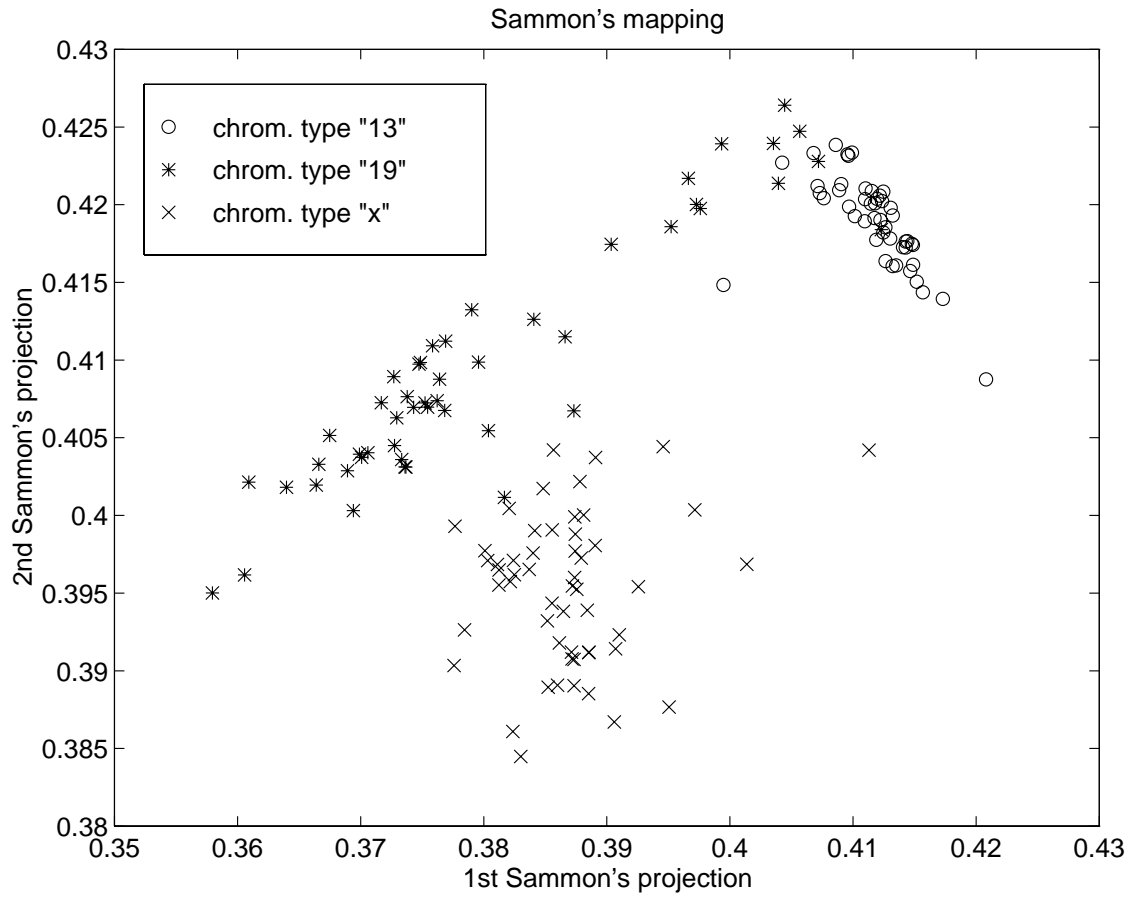


Fig. 1b

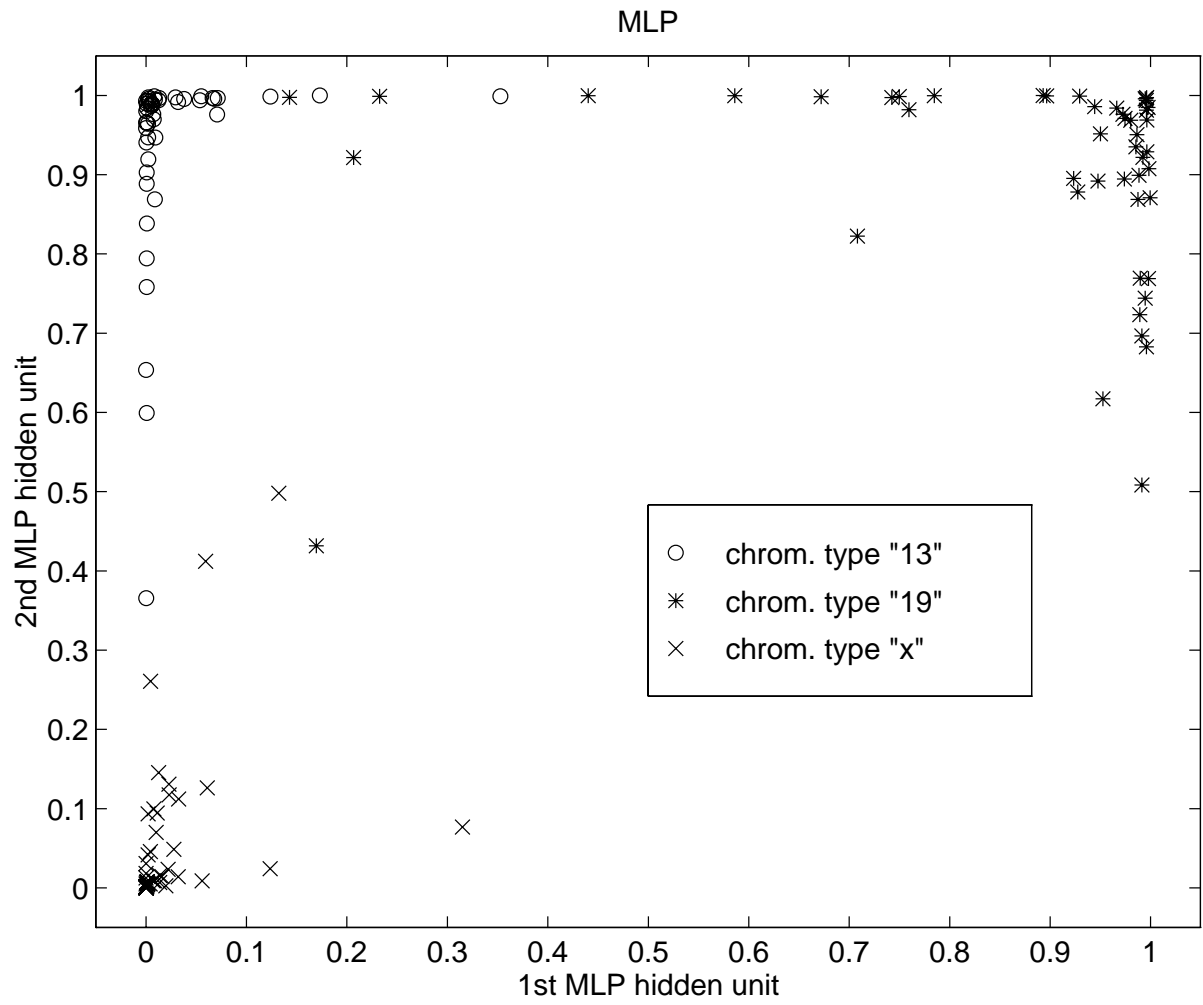


Fig. 1c

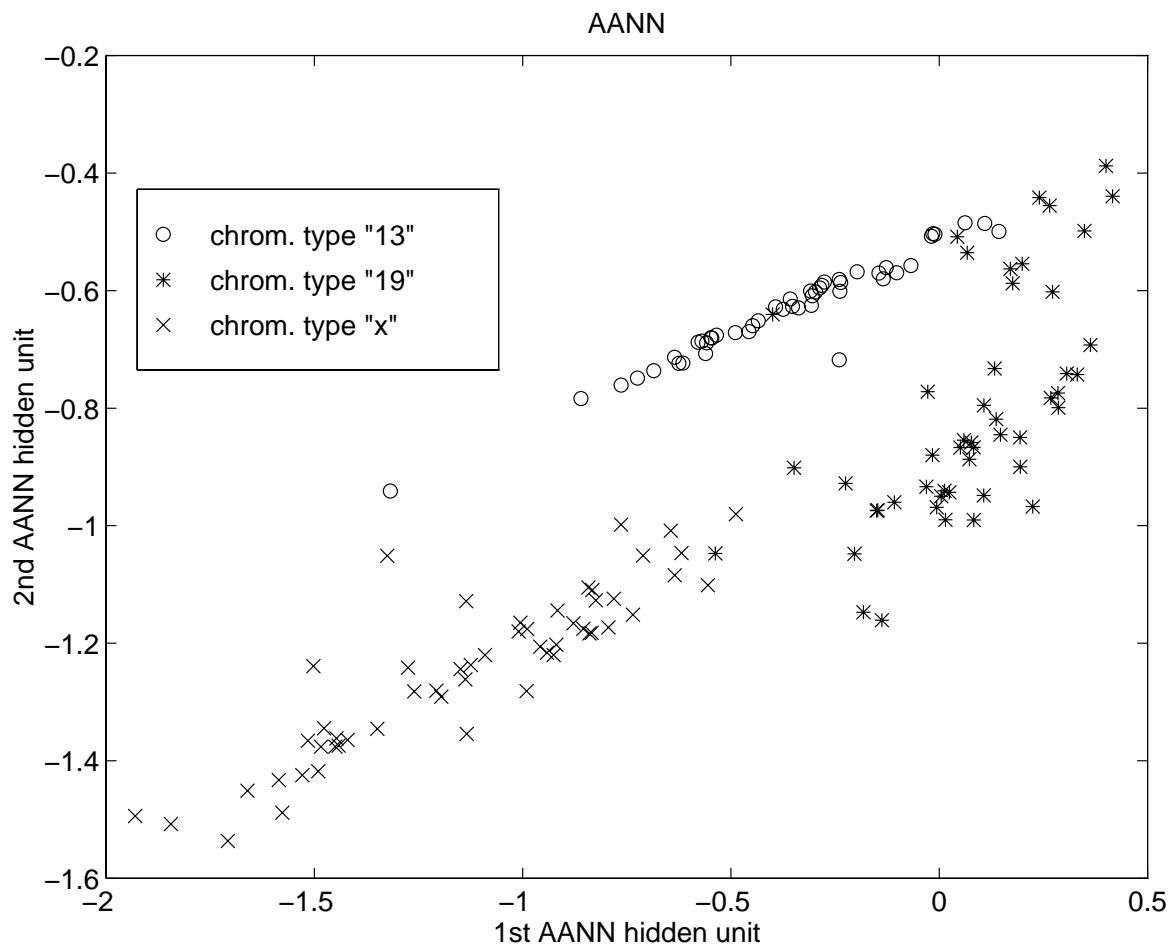


Fig. 1d

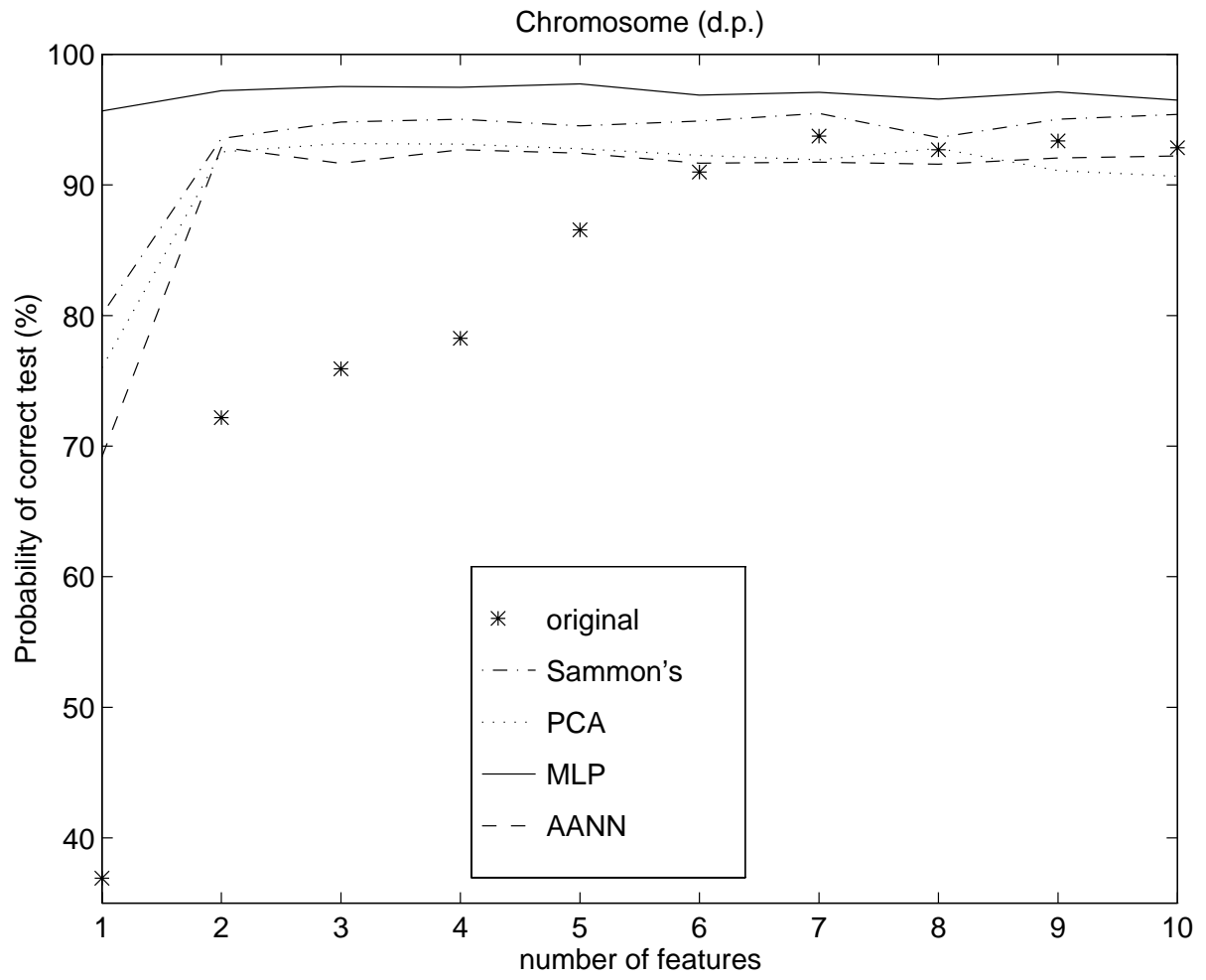


Fig. 2a

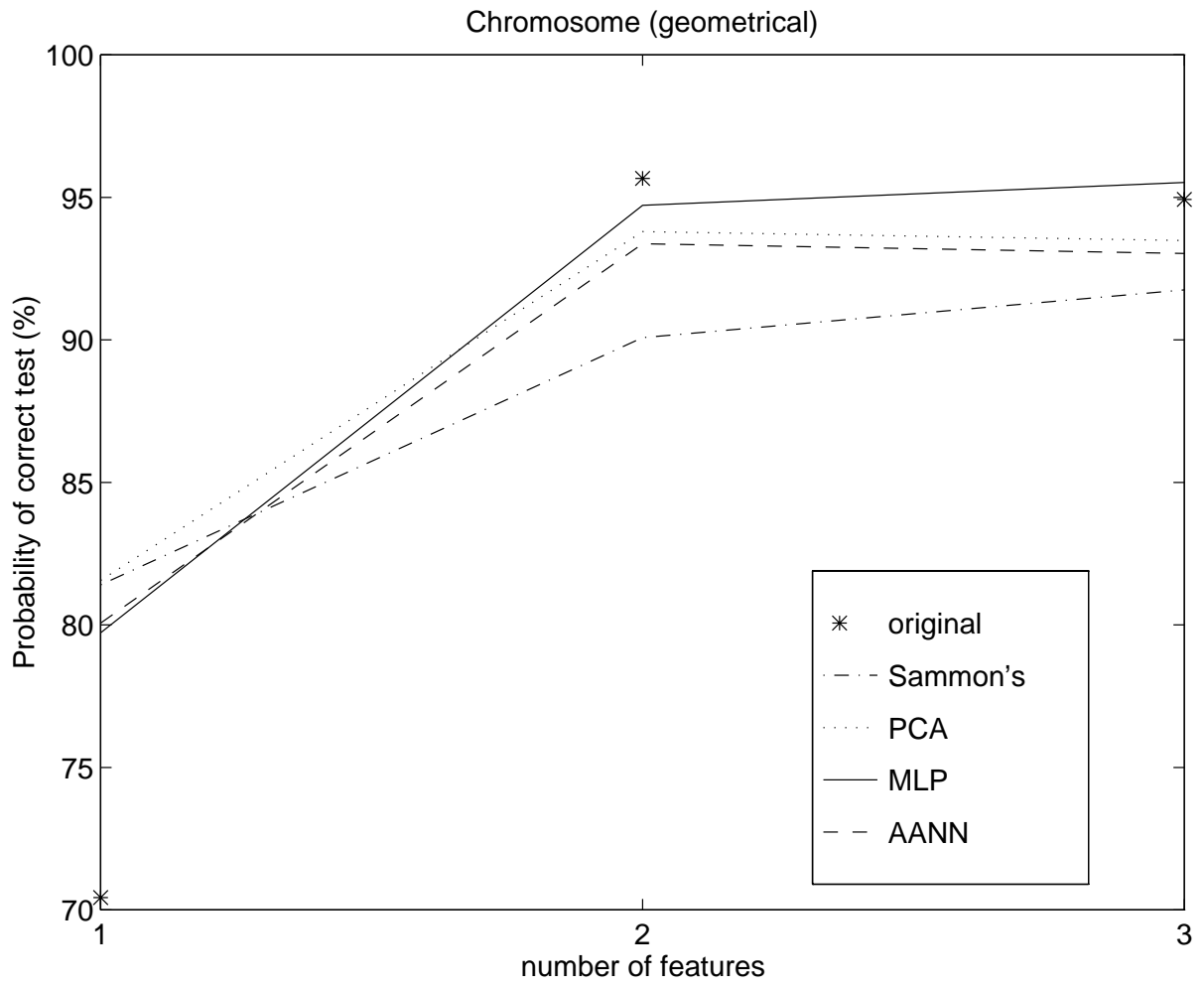


Fig. 2b

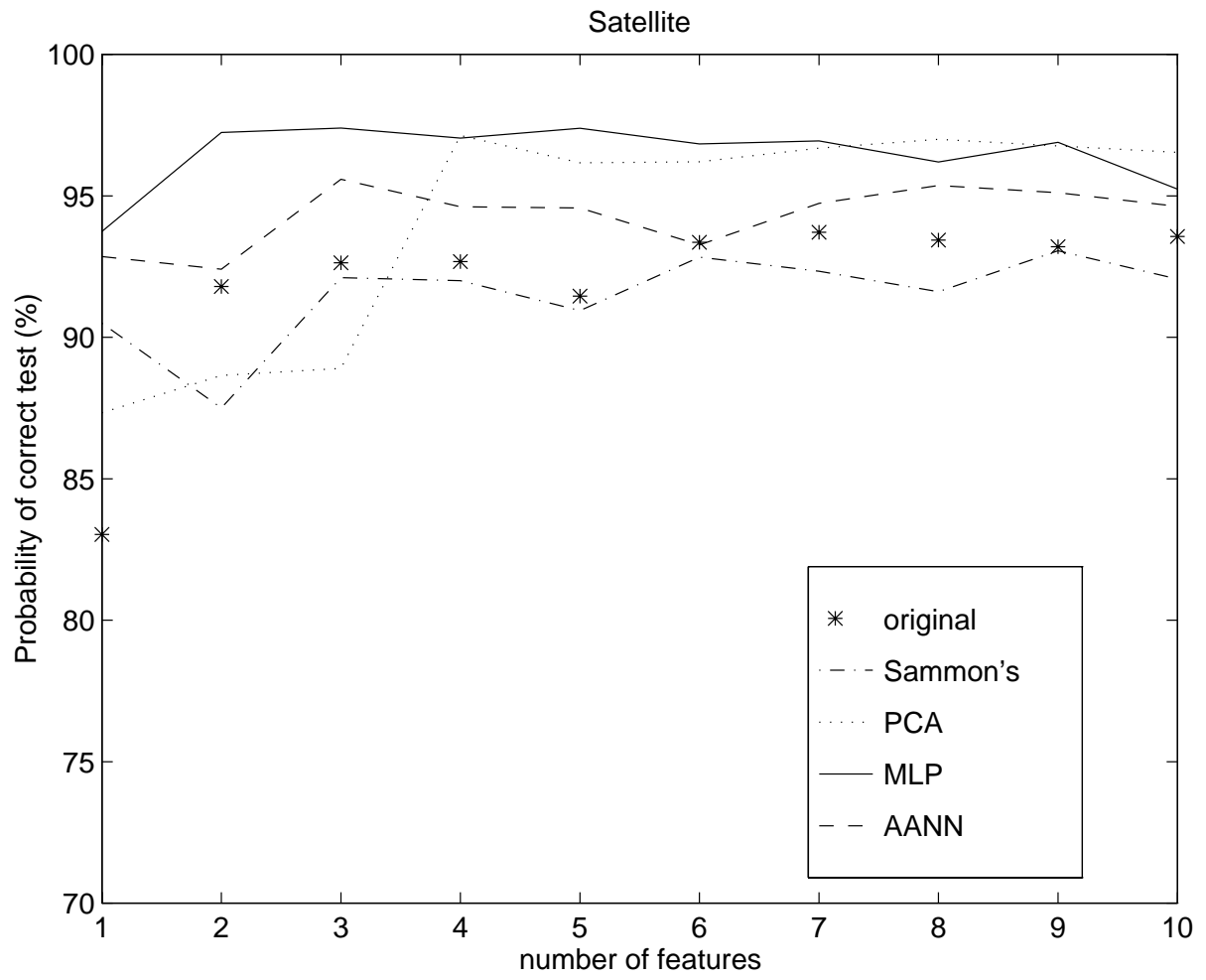


Fig. 2c

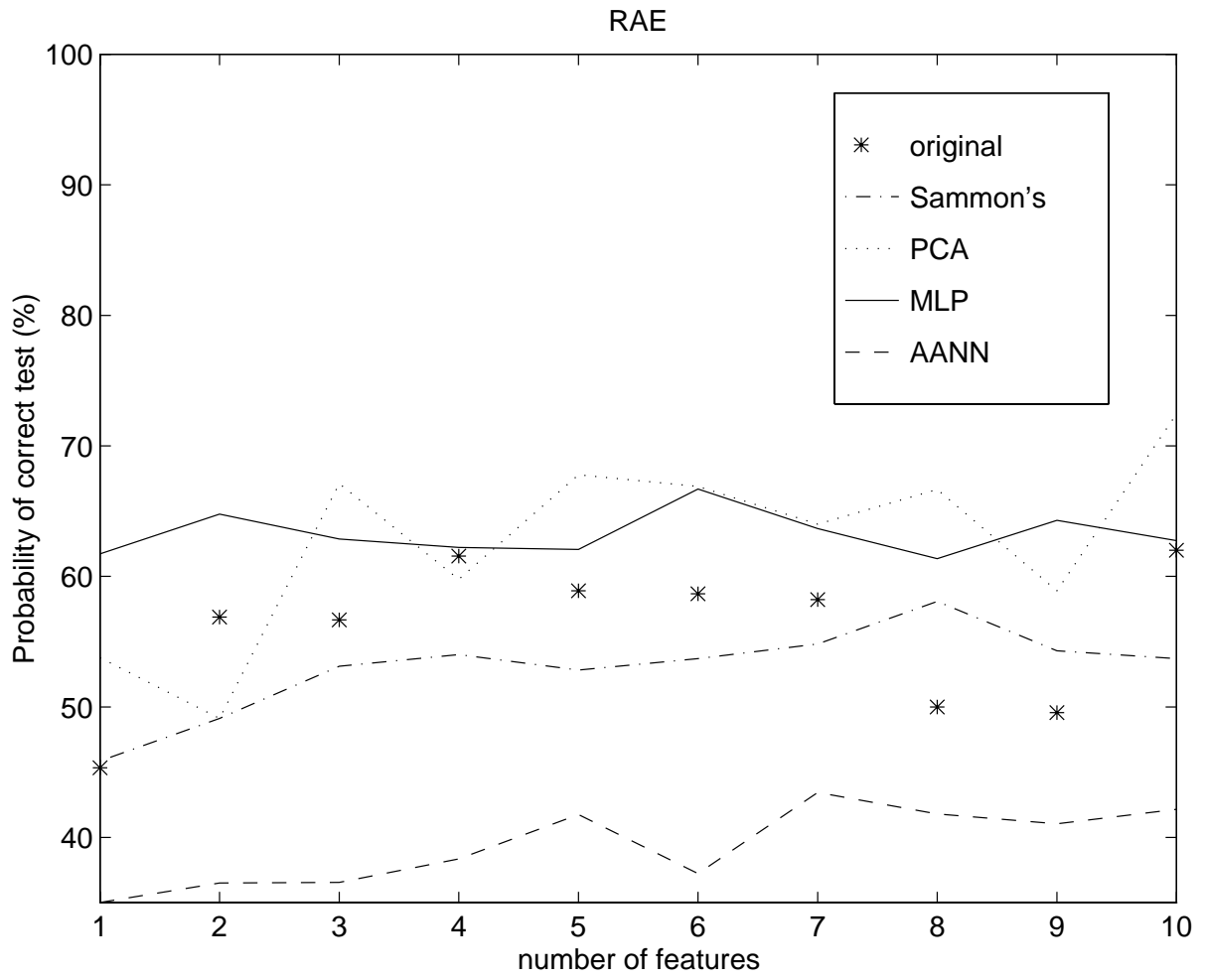


Fig. 2d

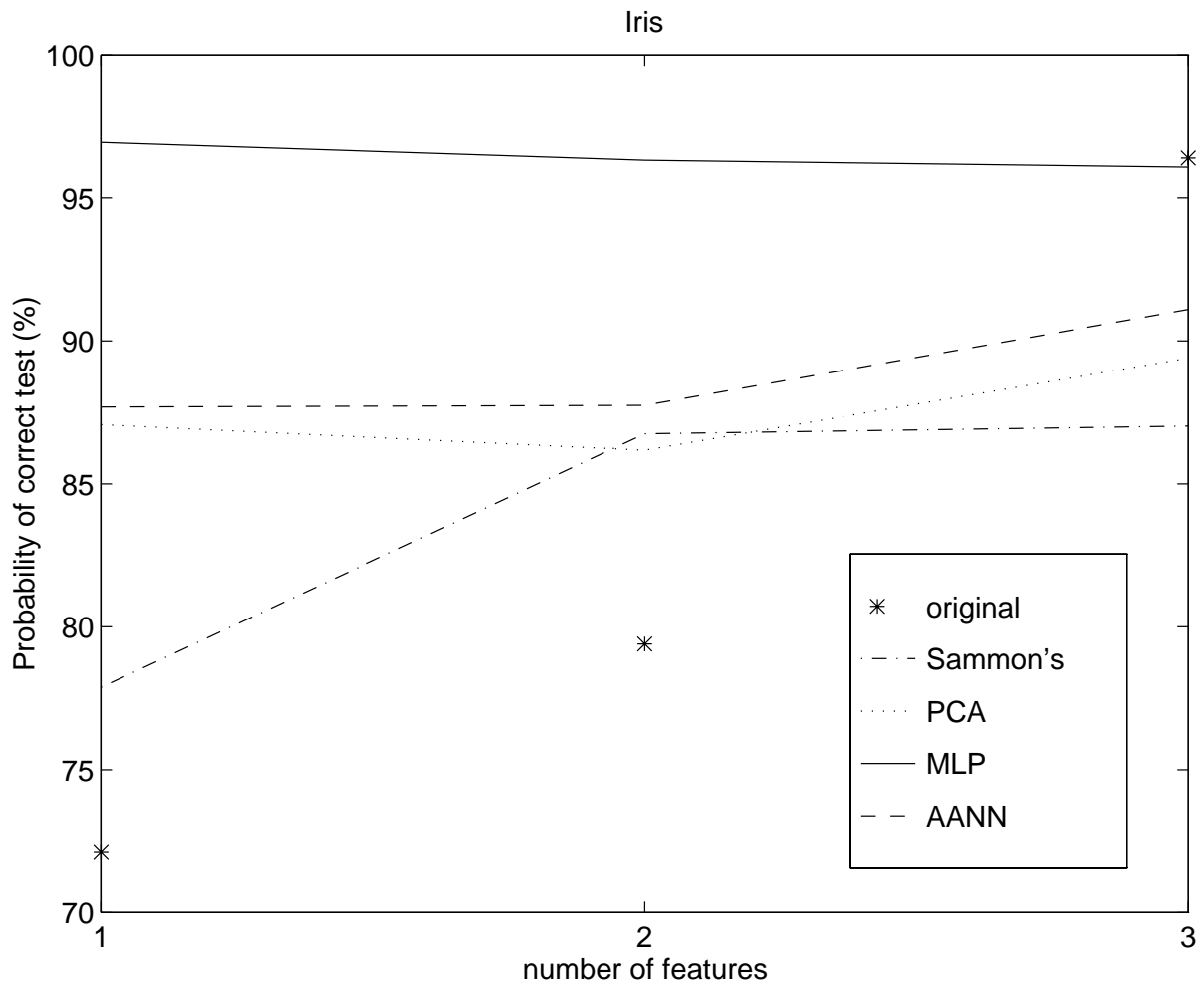


Fig. 2e

| | Supervised | Unsupervised |
|------------|------------|-----------------------|
| Linear | LDA | PCA, AANN |
| Non-linear | MLP | Sammon's mapping, SOM |

Table 1. Examples for common feature extraction paradigms