Abstract—We present a neural network based approach for identifying salient features for classification in neural network committees. Our approach involves neural network training with an augmented cross-entropy error function. The augmented error function forces the neural network to keep low derivatives of the transfer functions of neurons of the network when learning a classification task. Such an approach reduces output sensitivity to the input changes. Feature selection is based on the reaction of the cross-validation data set classification error due to the removal of the individual features. We compared the approach with two other neural network based feature selection methods. The algorithm developed outperformed the methods by achieving a higher classification accuracy on three real world problems tested.

I. INTRODUCTION

The pattern recognition problem is traditionally divided into the stages of feature extraction and classification. Feature extraction aims to find a mapping that reduces the dimensionality of the patterns being classified. The mapping found projects the N-dimensional data onto the M-dimensional space, where $M < N$. Feature selection is a special case of feature extraction. Employing feature extraction all $N$ measurements are used for obtaining the $M$-dimensional data. Therefore, all $N$ measurements need to be obtained. Feature selection, by contrast, enables us to discard $(N - M)$ irrelevant features. Hence, by collecting only relevant attributes, the cost of future data collecting may be reduced.

A large number of features can be usually measured in many pattern recognition applications. Not all of the features, however, are equally important for a specific task. Some of the variables may be redundant or even irrelevant. Usually better performance may be achieved by discarding such variables [13]. Moreover, as the number of features used grows, the number of training samples required grows exponentially [8]. Therefore, in many practical applications we need to reduce the dimensionality of the data.

Feature selection in general is a difficult problem. In a general case, only an exhaustive search can guarantee an optimal solution. The branch and bound algorithm [22] can also guarantee an optimal solution, if the monotonicity constraint imposed on a criterion function is fulfilled. The branch and bound based optimization has been used for feature selection by several authors [10], [14]. A large variety of feature selection techniques that result in a suboptimal feature set have been proposed [15], [17], [19], ranging from the sequential forward and backward selection [21] to the sequential forward floating selection characterized by a dynamically changing number of features included or eliminated at each step [26]. Though not numerous, techniques for feature selection based on the fuzzy set theory have also been proposed [2], [24], [25].

Feature selection with neural nets can be thought of as a special case of architecture pruning [27], where input features are pruned, rather than hidden neurons or weights. Pruning procedures extended to the removal of input features have been proposed in [4], [18], [20], where the feature selection process is usually based on some saliency measure aiming to remove less relevant features. Zurada et al. have recently proposed a saliency measure based feature selection method for regression [37]. The authors assume that the trained network provides a continuous differentiable mapping. In [7], a technique for backward feature selection is proposed that iteratively removes features one by one in a trained network and adjusts the remaining weights in such a way that the overall input-output behaviour learnt by the network remains approximately unchanged. The selection criterion is based on the policy of least adjustments required. An approach based on a formal hypothesis test for testing the statistical significance of a $q$-dimensional subset of weights has also been proposed for feature selection [30].

It is well known that a combination of many different neural networks can improve classification accuracy. A variety of schemes have been proposed for combining multiple classifiers. The approaches used most often include the majority vote, averaging, weighted averaging, the Bayesian approach, the fuzzy integral, the Dempster-Shafer theory, the Borda count, aggregation through order statistics, probabilistic aggregation, the fuzzy templates, and aggregation by a neural network [32], [34]. Numerous previous works on neural network committees have shown that an efficient committee should consist of networks that are not only very accurate, but also diverse in the sense that the network errors occur in different regions of the input space [23], [36]. Bootstrapping [5], [9], Boosting [1], [36], and AdaBoosting [12], [28] are the most often used approaches for data sampling when training members of neural network committees. It has been recently shown that half & half bagging [6] is capable of creating very
accurate committees of decision trees [6] and neural networks [33], [34], [35].

Despite the considerable interest of researchers in neural network based feature selection, to our knowledge, there were no attempts to select features for neural network committees. In this paper, we propose a technique for identifying salient features for classification in neural network committees. We use the half & half sampling technique when training committee members. To reduce committee’s output sensitivity to the input changes a term constraining the derivatives of the transfer functions of the neural network output and hidden nodes is added to the cross-entropy error cost function. Committee members are trained by minimizing such an extended cost function. Feature selection is based on the reaction of the cross-validation data set classification error due to the removal of the individual features.

II. HALF & HALF BAGGING

It has been demonstrated that the AdaBoost algorithm [12] generates classifiers with a low generalization error [11]. AdaBoost is a complex algorithm. Breiman has recently proposed a very simple algorithm, the so called half & half bagging approach [6]. When tested on decision trees the approach was competitive with the AdaBoost algorithm [6] and neural network committees. We use the half & half sampling technique when training committee members. To reduce committee’s output sensitivity to the input changes a term constraining the derivatives of the transfer functions of the neural network output and hidden nodes is added to the cross-entropy error cost function. Committee members are trained by minimizing such an extended cost function. Feature selection is based on the reaction of the cross-validation data set classification error due to the removal of the individual features.

The basic idea of half & half bagging is very simple. It is assumed that the training set contains $N$ data points. Suppose that $k$ classifiers have been already constructed. To obtain the next training set, randomly select a data point $x$. Present $x$ to that subset of $k$ classifiers which did not use $x$ in their training sets. Use the majority vote to predict the classification result of $x$ by that subset of classifiers. If $x$ is misclassified, put it in set MC. Otherwise, put $x$ in set CC. Stop when the sizes of both MC and CC are equal to $M$, where $2M \leq N$. Usually, CC is filled first but the sampling continues until MC reaches the same size. In [6], $M = N/4$ has been used. The next training set is given by a union of the sets MC and CC.

III. THE NEURAL NETWORKS USED

We use fully connected feedforward neural networks. Let $o_{ij}^{(q)}$ denote the output signal of the $j$th neuron in the $q$th layer and $w_{ij}^{(q)}$ - the connection weight coming from the $i$th neuron in the $(q-1)$ layer to the $j$th neuron in the $q$th layer. Then $o_{ij}^{(q)} = f(\text{net}_{ij}^{(q)})$ and $\text{net}_{ij}^{(q)} = \sum_{l=0}^{n_{q-1}} w_{ij}^{(l)} o_{il}^{(l-1)}$, where $f(\text{net})$ is the sigmoid activation function.

When given an augmented input vector $x = [1, x_1, x_2, \ldots, x_N]^T$, the output signal of the $j$th neuron in the output $(L)$th layer is given by:

$$o_{j}^{(L)} = f\left(\sum_{m} w_{mj}^{(L)} f\left(\sum_{i} w_{iq}^{(1)} x_i \right)\right) \quad (1)$$

In our tests, we used half & half bagged committees made of one hidden layer neural networks.

IV. SHORTCOMINGS OF THE WEIGHTS-BASED SALIENCY MEASURES

A significant number of feature saliency measures used for neural network based feature selection are weights-based, for instance $\Lambda_1 = \sum_{l=1}^{n_{L}} (w_{ij})^2$, where $w_{ij}$ is the weight between the $i$th input feature and the $j$th hidden node, and $n_L$ is the number of the hidden nodes [3], [31], or neural network’s output sensitivity based [4], [37].

The weights-based feature saliency measures bank on the idea that weights connected to important features attain large absolute values while weights connected to unimportant features would probably attain values somewhere near zero.

To force the training process to result in weights manifesting larger differences between the values of weights connected to the relevant features and the useless ones, a kind of the weight-decay regularization term, as $R_1(w) = \sum_{l=1}^{N} \sum_{j=1}^{n_{h}} (w_{ij})^2$ or that given by Eq. 2 [29], is often added to the error function being minimized.

$$R_2(w) = \varepsilon_1 \left\{ \sum_{i=1}^{N} \sum_{j=1}^{n_h} (\beta (w_{ij}))^2 \right\} + \varepsilon_2 \left\{ \sum_{i=1}^{N} \sum_{j=1}^{n_h} (w_{ij})^2 \right\} \quad (2)$$

where the constants $\varepsilon_1$, $\varepsilon_2$ and $\beta$ have to be chosen experimentally. The term $R_1(w)$ and the second part of the term $R_2(w)$ are exactly the weight-decay terms, except that only input to hidden weights are constrained. The first term of the function $R_2(w)$ can be considered as a measure of the total number of nonzero input weights in the network.

However, concerning feature selection, weight-decay possesses the following drawback. A simple weight decay algorithm tries to get smaller weights. Smaller weights usually result in smaller inputs to neurons and larger sigmoid derivatives in general. Therefore, output sensitivity to the input increases.

For the purpose of classification, by contrast, we need low sensitivity of output to the input. Hence, it seems reasonable to constrain the derivatives of the transfer functions of neurons instead of input layer weights during training. By constraining the derivatives we can force neurons to work in the saturation region.

V. THE TECHNIQUE PROPOSED

Using the error backpropagation rule from Eq. 1 we can get:

$$\frac{\partial o_{j}^{(L)}}{\partial x_i} = o_{j}^{(L)} \sum_{m} w_{mj}^{(L)} f^{(L-1)} \cdots \sum_{l} w_{il}^{(3)} \delta_{j}^{(2)} \times \sum_{q} w_{iq}^{(2)} f^{(1)} w_{iq}^{(1)} \quad (3)$$
where $\delta$ is the derivative of the neuron’s transfer function. For the sigmoid function $\sigma_j^{(L)} = \sigma_j^{(L)}(1 - \sigma_j^{(L)})$.

From Eq. 3 it can be seen that output sensitivity to the input depends on both weight values and derivatives of the transfer functions of the hidden and output layer nodes. To obtain the low sensitivity we have chosen to constrain the derivatives. We train a neural network by minimizing the cross-entropy error function augmented with two additional terms:

$$E = \frac{E_0}{n_L} + \alpha_1 \frac{1}{P_{nh}} \sum_{p=1}^{P} \sum_{k=1}^{n_h} f'(net_{kp}) +$$
$$+ \alpha_2 \frac{1}{P_{nl}} \sum_{j=1}^{n_L} \sum_{j'=1}^{n_L} f'(net_{jp})^{(L)}$$

where $\alpha_1$ and $\alpha_2$ are parameters to be chosen experimentally, $P$ is the number of training samples, $n_h$ is the number of the output layer nodes, $f'(net_{kp})$ and $f'(net_{jp})^{(L)}$ are derivatives of the transfer functions of the $k$th hidden and $j$th output node, respectively, and

$$E_0 = \frac{1}{2P} \left[ \sum_{p=1}^{P} \sum_{j=1}^{Q} (d_{jp} \log o_{jp}^{(L)} +$$
$$+ (1 - d_{jp}) \log (1 - o_{jp}^{(L)})) \right]$$

where $d_{jp}$ is the desired output for the $p$th data point at the $j$th output node and $Q$ is the number of classes.

The second and third terms of the cost function constrain the derivatives and forces the neurons of the hidden and output layers to work in the saturation region. In [16], it was demonstrated that neural networks regularized by constraining derivatives of the transfer functions of the hidden layer nodes possess good generalization properties. We can expect that different sensitivity of the hidden and output nodes could be required for solving a task with the lowest generalization error. Therefore, two hyper-parameters $\alpha_1$ and $\alpha_2$ are used in the error function. The feature selection procedure is summarized in the following steps.

A. The Feature Selection Procedure

1. Choose the number of neural network committee members $L$, and the number of initialisations $I$.

2. Randomly divide the data set available into Training, Cross-Validation, and Test data sets. Use half of the Training data set when training the first committee member. Set the committee member index $j = 1$. Set $A_{CMAX} = 0$ - the maximum Cross-Validation data set classification accuracy achieved by the committee.

3. Set the initialisation index $i = 1$.

4. Randomly initialize the weights of the committee member. Set the actual number of features $k = N$.

5. Train the committee member by minimizing the error function given by Eq. 4 and validate the network at each epoch on the Cross-Validation data set. Equip the network with the weights yielding the minimum Cross-Validation error.

6. Compute the Cross-Validation data set classification accuracy $A_{jik}$ for the committee. The committee consists of $j$ members including the one being trained. If $A_{jik} > A_{CMAX}$ set $A_{CMAX} = A_{jik}$.

7. Identify the feature yielding the smallest drop of the committee classification accuracy for the Cross-Validation data set when eliminating the feature. Elimination is implemented by setting the value of the feature to zero. Eliminate the feature.

8. Set $k := k - 1$. If the actual number of features $k > 1$ goto Step 5.

9. The selected number of features $M$ for the committee member being trained is given by the minimum value of $k$ satisfying the conditions: $A_{CMAX} - A_{jik} < \Delta A_0$, $k_1 \geq k_1, j = 2, ..., L$, where $\Delta A_0$ is the acceptable drop in the classification accuracy and $k_j$ is the number of features selected for the $j$th member. Memorize the classification accuracy $A_{jim}$, the set of selected features $F_{ji} = \{f_{j1}, ..., f_{jim}\}$, and the weight matrix $W_{jim}$. The feature set contains the remaining and the $M - 1$ last eliminated features.

10. Set $i := i + 1$. If $i \leq I$ goto Step 4.

11. Amongst the $I$ feature sets $F_{ji}, i = 1, ..., I$ select the feature set $F_{jn}$ for the $j$th committee member according to the rule: $n = \arg \max_{n=1,...,I} A_{jim}$. Equip the $j$th member with the $W_{jnM}$ weight matrix.

12. Set $j := j + 1$. If $j \leq L$ select, according to the half & half sampling procedure, the Training data set for training the next committee member and goto Step 3.

13. Stop. The neural network committee is defined by the weight matrices $W_{jnM}, j = 1, ..., L$.

We compare the proposed approach with two other neural network based feature selection techniques when only a single network is used for data classification. Next, we briefly describe the techniques used for the comparisons.

VI. TECHNIQUES USED FOR THE COMPARISON

A. Neural-network feature selector

The neural-network feature selector (NNFS) is trained by minimizing the cross-entropy error function augmented with the additional term given by Eq. 2. Feature selection is based on the reaction of the cross-validation data
set classification error due to the removal of the individual features [29]. For our comparisons we use the results presented in [29].

B. Signal-to-noise ratio based technique

The signal-to-noise ratio (SNR) saliency measure based feature ranking and selection technique has been recently proposed in [3]. The saliency of a feature is determined by comparing it to that of an injected noise feature. The SNR saliency measure for feature $i$ is given by:

$$\text{SNR}_i = 10 \log_{10} \left( \frac{\sum_{j=1}^{n_h} (w_{ij})^2}{\sum_{j=1}^{n_m} (w_{ij})^2} \right)$$

with $w_{ij}$ being the weight between the $i$th input feature and the $j$th hidden node, $w_{ij}$ is the weight from the injected noise feature $I$ to the $j$th hidden node, and $n_h$ is the number of the hidden nodes.

The authors have demonstrated that the technique is competitive with the method proposed by Setiono and Liu [29]. The number of features to be chosen is identified by the significant decrease of the classification accuracy of the test data set when eliminating a feature.

C. Discriminant analysis based feature ranking

Let $\mathbf{m}_j$ denote the sample mean vector of the $j$th class:

$$\mathbf{m}_j = \frac{1}{N_j} \sum_{k=1}^{N_j} \mathbf{x}_{jk},$$

where $N_j$ is the number of samples in the $j$th class. Similarly, $\mathbf{m}$ denotes the mixture sample mean: $\mathbf{m} = \sum_{j=1}^{Q} P_j \mathbf{m}_j$, with $P_j$ being a priori probability of the $j$th class. We can now define a within class-covariance and between class-covariance matrices $\mathbf{S}_w$ and $\mathbf{S}_b$, respectively. Now the following criterion function $J_i(\mathbf{x})$ can be formed for feature ranking:

$$J_i(\mathbf{x}) = \frac{\text{tr}(\mathbf{S}_b)}{\text{tr}(\mathbf{S}_w)} - \frac{\text{tr}(\mathbf{S}_b)}{\text{tr}(\mathbf{S}_w)}$$

where $\text{tr}(\mathbf{S}_b)$ stands for the trace of $\mathbf{S}_b$ with the $i$th diagonal element excluded. The larger the value of $J_i(\mathbf{x})$ the higher individual discrimination power the $i$th feature possesses. We used this feature ranking approach as a basic technique to compare individual features.

VII. EXPERIMENTAL INVESTIGATIONS

In all the tests, we run an experiment 30 times with different initial values of weights and different partitioning of the data set into <Training>, <Cross-Validation>, and <Test> sets. The mean values and standard deviations of the correct classification rate presented in this paper were calculated from these 30 trials.

A. Training parameters

There are three parameters to be chosen, namely the regularization constants $\alpha_1$ and $\alpha_2$ and the parameter of the acceptable drop in classification accuracy $\Delta A_0$ when eliminating a feature. The parameter affects the number of features included in the feature subset sought. The values of the parameters $\alpha_1$ and $\alpha_2$ have been found by cross validation. The values of the parameters ranged: $\alpha_1 \in [0.001, 0.02]$ and $\alpha_2 \in [0.001, 0.2]$. The value of the parameter $\Delta A_0$ has been set to 3%. All the committees consisted of $L = 5$ one hidden layer perceptrons. As in [29] and [3], we also used 12 nodes in the hidden layer of the committee members when learning the problems.

B. Data used

To test the approach proposed we used three real-world problems. The data used in the experiments are available at: www.ics.uci.edu/~mlearn/MLRepository.html. We randomly assign available data exemplars into learning $D_l$, validation $D_v$, and testing $D_t$ data sets.

US Congressional voting records problem. The United States Congressional voting records data set consists of the voting records of 435 congressmen on 16 major issues in the 98th Congress. The votes are categorized into one of the three types of votes: (1) Yea, (2) Nay, and (3) Unknown. The task is to predict the correct political party affiliation of each congressmen. We used the same learning and testing conditions as in [3] and [29], namely 197 samples were randomly selected for training, 21 samples were selected for cross-validation, and 217 for testing.

The diabetes diagnosis problem. The Pima Indians Diabetes Data Set contains 768 samples taken from patients who may show signs of diabetes. Each sample is described by eight features. There are 500 samples from patients who do not have diabetes and 268 samples from patients who are known to have diabetes. From the data set, we have randomly selected 345 samples for training, 39 samples for cross-validation, and 384 samples for testing.

The breast cancer diagnosis problem. The University of Wisconsin Breast Cancer Data Set consists of 699 patterns. Amongst them there are 458 benign samples and 241 malignant ones. Each of these patterns consists of nine measurements taken from fine needle aspirates from a patient’s breast. To test the approaches we randomly selected 315 samples for training, 35 samples for cross-validation, and 349 for testing.

C. Results of the tests

The US Congressional Voting Records problem is an easy task from the feature selection point of view, since only one feature <4> exhibits almost the same discrimination power as the whole feature set. Fig. 1 plots the criterion $J_i(\mathbf{x})$ values for the features of the Voting data set. All the techniques tested deemed the feature <4> as the most salient one. Table I presents the test data set correct classification rate obtained from the method proposed. In the table, we also provide the results taken from the references [3] and [29]. In the parentheses, the standard deviations of the correct classification rate are provided. As can be seen from Table I, the method pro-
The proposed achieved the highest classification accuracy on the test data set. In all the 30 trials, the method proposed selected two features for each member of a committee. Features \( <4, 14, 15, 11, 16, 3> \) are the six most often selected features in the different trials. On average, 6 different features were used by one committee. Since each committee consisted of five members, some features, for example feature \(<4>\), have been used by several committee members. According to the discriminant analysis based feature ranking technique the six most salient features are \( <4, 5, 3, 8, 14, 12> \). Observe that only three of the six features are amongst the six most often selected features by the method proposed. It is interesting to note that several features have been never chosen in all the 30 trials.

Thus, four of the five most salient features, as deemed by the \((DA)\) technique, are amongst the five most often utilized features. Table II provides the test data set correct classification rate achieved by the different techniques. The superiority of the approach proposed should be obvious from the table. As can be seen from the table, on average, committees, the members of which utilize only two features, are more accurate than those exploiting the whole feature set.

![Fig. 1. Criterion \( J_i(x) \) values for the features of the Voting data set.](image)

**TABLE I**

<table>
<thead>
<tr>
<th>Case</th>
<th>Proposed</th>
<th>SNR [3]</th>
<th>NNFS [29]</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Feat.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Train. Set</td>
<td>99.06(0.53)</td>
<td>98.92(0.22)</td>
<td>100.00(0.00)</td>
</tr>
<tr>
<td>Test Set</td>
<td>95.70(0.57)</td>
<td>95.42(0.18)</td>
<td>92.00(0.18)</td>
</tr>
</tbody>
</table>

| Sel. Feat. |          |         |           |
| # ofFeat.  |          |         |           |
| Train. Set | 2(0.00)  | 1(0.00)  | 2.03(0.18) |
| Test Set   | 96.92(0.25) | 96.62(0.30) | 95.63(0.08) |

| The Pima Indians Diabetes problem. To solve the task, the method proposed selected two features for each member of a committee. On average, 5 different features were used by neural networks of one committee. According to the discriminant analysis \((DA)\) based technique, five the most salient features, starting from the most salient one, are \( <2, 6, 8, 1, 7> \). Five the most often selected features in the different trials are as follows \( <2, 5, 7, 8, 3> \). |

<table>
<thead>
<tr>
<th>Case</th>
<th>Proposed</th>
<th>SNR [3]</th>
<th>NNFS [29]</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Feat.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Train. Set</td>
<td>97.57(0.41)</td>
<td>97.66(0.18)</td>
<td>100.00(0.00)</td>
</tr>
<tr>
<td>Test Set</td>
<td>96.90(0.50)</td>
<td>96.49(0.15)</td>
<td>93.94(0.17)</td>
</tr>
</tbody>
</table>

| Sel. Feat. |          |         |           |
| # ofFeat.  |          |         |           |
| Train. Set | 2(0.00)  | 1(0.00)  | 2.7(1.02)  |
| Test Set   | 98.15(0.62) | 94.03(0.97) | 98.05(0.24) |

The University of Wisconsin Breast Cancer problem. In all the 30 runs performed, our technique suggested that 2 features should be selected for each committee member to solve the task. Feature \(<6>\) was the most often selected and also the most salient one when using the criterion \( J_i(x) \) for the evaluation. On average seven features out of the nine available have been utilized by one committee.

**TABLE III**

<table>
<thead>
<tr>
<th>Case</th>
<th>Proposed</th>
<th>SNR [3]</th>
<th>NNFS [29]</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Feat.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Train. Set</td>
<td>98.15(0.62)</td>
<td>94.03(0.97)</td>
<td>98.05(0.24)</td>
</tr>
<tr>
<td>Test Set</td>
<td>97.65(0.57)</td>
<td>92.53(0.77)</td>
<td>94.15(0.18)</td>
</tr>
</tbody>
</table>

Table III presents the test data set correct classification rate obtained from the different techniques. As can be seen from the table, again, committees utilizing the whole feature set perform worse than those exploiting only the...
selected features. The small difference between the correct classification rates obtained on the training and test data sets witnesses the good generalization properties of the neural network committees trained according to the approach proposed.

VIII. CONCLUSIONS

In this paper, we presented a neural network based feature selection technique for neural network committees. Committee members are trained on half & half sampled training data sets by minimizing an augmented cross-entropy error function. The augmented error function forces the neural network to keep low derivatives of the transfer functions of neurons when learning a classification task. Such an approach reduces output sensitivity to the input changes. The feature selection is based on the reaction of the cross-validation data set classification error due to the removal of the individual features.

We have tested the technique proposed on three real-world problems and demonstrated the ability of the technique to create accurate neural network committees exhibiting good generalization properties. The algorithm developed removed a large number of features from the original data sets without reducing the classification accuracy of the committees. We compared the proposed approach with two other methods. The technique developed outperformed the methods by achieving a higher test data set classification accuracy on all the problems tested.

REFERENCES