An Effective Feature Selection Scheme via Genetic Algorithm Using Mutual Information[1]

Chunkai K. Zhang and Hong Hu

Member IEEE,
Department of Mechanical Engineering and Automation, Harbin Institute of Technology,
Shenzhen Graduate School, Shenzhen, China, 518055
ckzhang@hotmail.com

Abstract. In the artificial neural networks (ANNs), feature selection is a well-researched problem, which can improve the network performance and speed up the training of the network. The statistical-based methods and the artificial intelligence-based methods have been widely used to feature selection, and the latter are more attractive. In this paper, using genetic algorithm (GA) combining with mutual information (MI) to evolve a nearoptimal input feature subset for ANNs is proposed, in which mutual information between each input and each output of the data set is employed in mutation in evolutionary process to purposefully guide search direction based on some criterions. By examining the forecasting at the Australian Bureau of Meteorology, the simulation of three different methods of feature selection shows that the proposed method can reduce the dimensionality of inputs, speed up the training of the network and get better performance.

1 Introduction

In the artificial neural networks (ANNs), feature selection is a well-researched problem, aimed at reducing the dimensionality and noise in input set to improve the network performance and speed up the training of the network [1].

Many algorithms for feature selection have been proposed. Conventional methods are based on the statistical tools, such as the partial **F**-test, correlation coefficient, residual mean square [2,3,4,5] and mutual information (MI) [6,7,8,9]. Although the statistical-based feature selection techniques are widely used, they suffer from many limitations [10]. Firstly, most of them are computationally expensive, because the comparison of all feature subsets is equivalent to a combinatorial problem whose size exponentially increases with the growing number of features. Secondly, the selected feature subset cannot be guaranteed optimal. For example, in the mutual information method, selecting a fixed number of inputs from a ranked list consisting of combinations along with single entries is somewhat problematical, and once a feature is added at an early step, it cannot be removed although it may not constitute the best subset of features in conjunction with the later selected features. Finally, there are a number of parameters that need to be set a priori. For example, the number of features added or removed, the significance level for selecting features and the final feature size.

[1] This work is supported by High-tech Industrialization Special Research Project of China

L. Wang and Y. Jin (Eds.): FSKD 2005, LNAI 3614, pp. 73 – 80, 2005.

Because the problem of feature selection can be formulated as a search problem to find a nearoptimal input subset, so the artificial intelligence techniques, such as genetic algorithm (GA), is used to selects the optimal subset of features [11,12, 13]. In contrast with the statistical-based methods, the artificial intelligence-based methods are more attractive, as they can find nearoptimal feature subset in lower computational cost and the search process involves no user selectable parameters, such as the final feature size and the signification level etc.. In addition, they have the potential to simultaneous network evolution and feature selection. In most GA-based methods, only the correct recognition rate of a certain neural network is utilized to guide the search direction. In [14], Il-Seok Oh proposed the hybrid GAs for feature selection, which embeds local search operations into the simple GA, but useful information such as statistical information between inputs and outputs in data set don't be added in search process.

In this paper, we proposed a new feature selection scheme for ANNs via genetic algorithm using mutual information. In this method, mutual information (MI) between input and output is employed in mutation in GA to purposefully guide the evolutionary search direction based on some criterions, which can speed up the search process and get better performance. By examining the forecasting at the Australian Bureau of Meteorology [15], the simulation of three different methods of feature selection shows that the proposed method can reduce the dimensionality of inputs, speed up the training of the network and get better performance.

The rest of this paper is organized as follows. Section 2 describes mutual information (MI) and GA. Section 3 the hybrid of GA and MI is used to evolve an optimum input subset for an ANN. Section 4 presents experimental results in a real forecasting problem. The paper is concluded in Section 5.

2 Background

2.1 Definition of Mutual Information

In the information theory founded by Shannon [16], the uncertainty of a random variable C is measured by entropy $H(C)$. For two variables X and C, the conditional entropy $H(C\mid X)$ measures the uncertainty about C when X is known, and MI, $I(X;C)$, measures the certainty about C that is resolved by X. Apparently, the relation of $H(C)$, $H(C\mid X)$ and $I(X;C)$ is:

$$H(C)=H(C\mid X)+I(X;C) \tag{1}$$

or, equivalently,

$$I(X;C)=H(C)-H(C\mid X),$$

As we know, the goal of training classification model is to reduce the uncertainty about predictions on class labels C for the known observations X as much as possible. In terms of the mutual information, the purpose is just to increase MI $I(X;C)$

as much as possible, and the goal of feature selection is naturally to achieve the higher $I(X;C)$ with the fewer features.

With the entropy defined by Shannon, the prior entropy of class variable C is expressed as

$$H_s(C) = -\sum_{c \in C} P(c) \log P(c) \tag{2}$$

where $P(c)$ represents the probability of C, while the conditional entropy is $H(C \mid X)$ is

$$H_s(C \mid X) = -\int_x p(x)(\sum_{c \in C} p(c \mid x) \log p(c \mid x))dx \tag{3}$$

The MI between X and C is

$$I_s(X;C) = \sum_{c \in C} \int_x p(c,x) \log \frac{p(c,x)}{P(c)p(x)} dx \tag{4}$$

Mutual information can, in principle, be calculated exactly if the probability density function of the data is known. Exact calculations have been made for the Gaussian probability density function. However, in most cases the data is not distributed in a fixed pattern and the mutual information has to be estimated. In this study, the mutual information between each input and each output of the data set is estimated using Fraser & Swinney's method [9].

The mutual information of independent variables is zero, but is large between two strongly dependent variables with the maximum possible value depending on the size of the data set. And this assumes that all the inputs are independent and that no output is in fact a complex function of two or more of the input variables.

2.2 Genetic Algorithm

GA is an efficient search method due to its inherent parallelism and powerful capability of searching complex space based on the mechanics of natural selection and population genetics. The method of using GA to select input features in the neural network is straightforward. In GA, every candidate feature is mapped into individual (binary chromosomes) where a bit "1" (gene) denotes the corresponding feature is selected and a bit of "0" (gene) denotes the feature is eliminated. Successive populations are generated using a breeding process that favors fitter individuals. The fitness of an individual is considered a measure of the success of the input vector. Individuals with higher fitness will have a higher probability of contributing to the offspring in the next generation ('Survival of the Fittest').

There are three main operators that can interact to produce the next generation. In replication individual strings are copied directly into the next generation. The higher the fitness value of an individual, the higher the probability that that individual will be copied. New individuals are produced by mating existing individuals. The probability that a string will be chosen as a parent is fitness dependent. A number of crossover points are randomly chosen along the string. A child is produced by copying from one parent until a crossover point is reached, copying then switching to the other parent and repeating this process as often as required. An *N* bit string can have anything from 1 to *N*-1 crossover points. Strings produced by either reproduction or crossover may then be mutated. This involves randomly flipping the state of one or more bits. Mutation is needed so new generations are more than just a reorganization of existing genetic material. After a new generation is produced, each individual is evaluated and the process repeated until a satisfactory solution is reached. The procedure of GA for feature selection is expressed as follows:

```
Procedure of genetic algorithm for feature selection
Initialization
  $N \rightarrow$ Population size
  $P \rightarrow$ Initial population with $N$ subsets of Y
  $P_c \rightarrow$ Crossover probability
  $P_m \rightarrow$ Mutation probability
  $T \rightarrow$ Maximum number of generations
  $k \rightarrow$ 0
Evolution
  Evaluation of fitness of $P$
  while ( $k < T$ and $P$ does not converge) do
      Breeder Selection
      Crossover with $P_c$
      $\boxed{\text{Mutation with } P_m}$
      Evaluation of fitness of $P$ Replication
      Dispersal
      $k+1 \rightarrow k$
```

3 The Proposed Method for Feature Selection

In order to reduce time of calculating MI between single input and output in the whole data set, we randomly select some data from data set with probability 0.5 to construct a data set named *MI* set. Using Fraser & Swinney's method, the mutual information x_i between each candidate input and each output in *MI* set is estimated, which construct a data set $D = \{x_i, i = 1,...,N\}$, x_i represents the mutual information of i th candidate input, and N means there are N candidate inputs.

Then calculate the mathematical statistics of x_i: the mean $\bar{x}$ and standard deviation s_N

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{5}$$

$$s_N = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{6}$$

And define the three sets which satisfy $D = D_1 \bigcup D_2 \bigcup D_3$:

$$D_1 = \{x_i \mid x_i - \bar{x} > \frac{s_N}{2}\},$$

$$D_2 = \{x_i \mid -\frac{s_N}{2} \le x_i - \bar{x} \le \frac{s_N}{2}\}$$

$$D_3 = \{x_i \mid x_i - \bar{x} < -\frac{s_N}{2}\}$$

In GA, we use mutual information between each candidate input and each output to guide the mutation based on some criterions, as follows:

$$g_i = \begin{cases} 1 & x_i \in D_1 \\ 0 & x_i \in D_3 \\ rand & x_i \in D_2 \end{cases} \tag{7}$$

where g_i represents i th gene in a binary chromosome, it means i th candidate input. If the mutual information x_i of i th candidate input belongs to D_1, it means it is a highly correlated input for each output, so include it into input feature subset; if the mutual information x_i of i th candidate input belongs to D_2, it means it is a general correlated input for each output, so randomly include it into input feature subset; If the mutual information x_i of i th candidate input belongs to D_3, it means it is little correlated input for each output, so exclude it from input feature subset.

The procedure of the proposed method for feature selection is same as the procedure of GA for feature selection except the step of "mutation with P_m".

```
Mutation with $P_m$
        If $x_i$ of $i$ th candidate input belongs to $D_1$, include it into input feature subset;
        If $x_i$ of $i$ th candidate input belongs to $D_2$, randomly include it into input feature subset;
        If $x_i$ of $i$ th candidate input belongs to $D_3$, exclude it from input feature subset.
```

4 Experimental Studies

The temperature data for Australia was taken from the TOVS instrument equipped NOAA12 satellite in 1995 [13]. Infrared sounding of 30km horizontal resolution was supplemented with microwave soundings of 150 km horizontal resolution. This data set was used to evaluate the techniques for selecting the input subset. A number of single output networks were developed, each estimating the actual temperature at one of 4 pressure levels (1000, 700, 300 & 150 hPa) given the radiances measured by satellite. These are four of the standard pressure levels (levels 1, 3, 6 and 9) measured by satellite and radiosonde sounders. The input set of TOVS readings to be used by these networks was extracted using each of the three techniques: GA, MI [6] and the proposed method. The appropriate target output temperature was provided by collocated radiosonde measurement.

In MI method, a common input vector length of 8 was used as initial experimentation had proved this to be a suitable value. In GA and the proposed method, N =50, T =60, P_c =0.6 and P_m =0.02. And the *m*-12-1 network uses a learning rate of 0.1 and momentum of 0.8 for 10,000 iterations, where *m* represents the number of inputs. And the fitness function is defined to be $1/RMSE$, and the root mean square error (*RMSE*) is calculated by

$$RMSE = (\sum (Y - Y_r)^2)^{1/2} \tag{8}$$

where Y_r is the desired target value, and Y is the output of network.

After selecting an optimal input subset using one of the above techniques, these inputs were assessed by means of an evaluation neural network whose architecture was chosen based on initial experiments. The network used 12 hidden neurons and was trained using fixed parameters to facilitate comparison between the various techniques. It was trained for 2000 passes through the data set using a learning rate of 0.1 and a momentum of 0.8. The network was tested after each pass though the training data with the best result being recorded. The overall performance of this testing network was assumed to reflect the appropriateness of this particular selection of inputs.

The results reported are the mean *RMSE* values obtained from training the ten evaluation networks at each level and should be a reasonable reflection of the inherent worth of the input selection. The results using the full input set (all available inputs) are included in the table for comparison. Mean of *RMSE* (K) derived from all 3 techniques and using all inputs for levels 1,3, 6 & 9 is indicated in Table 1, and selected input subset is indicated in Table 2.

Table 1. Mean of *RMSE* (K) derived from all 3 techniques and using all inputs

	Full	GA	MI	the proposed method
Level 1	2.9	2.6	2.7	2.4
Level 3	2.7	2.9	3.6	2.8
Level 6	2.6	2.5	2.4	2.2
Level 9	3.9	3.6	3.4	3.3

Table 2. Selected input subset for various levels

	GA	MI	the proposed method
Level 1	1, 3, 7, 15, 17, 18, 19, 20, 21	22, 20, 14, 1, 2, 4, 13, 12	3, 8, 14, 20, 22, 4, 2
Level 3	0, 3, 6, 8, 10, 11, 14, 15, 17, 18, 19, 21	4, 21, 17, 15, 20, 3, 1, 9	1, 3, 4, 7, 10, 11, 14, 15, 17, 18, 19, 20, 22
Level 6	3, 6, 8, 14, 18, 21	14, 20, 4, 3, 15, 13, 22, 12	14, 20, 3, 4, 18, 13
Level 9	0, 4, 6, 8, 14, 16, 17, 18, 22	13, 14, 5, 4, 8, 6, 12, 20	4, 5, 6, 8, 12, 14, 20

Table 1 indicates that the proposed method exhibited better performance than the other techniques at all levels. GA was marginally better than MI, outperforming it in levels 1 and 3. Level 3 is interesting in that all three techniques produced networks with worse performance, especially MI. This seems to indicate that the predictive capability at this level is spread more across the inputs – there is less redundancy of information. It should be noted that at this difficult level GA and the proposed method outperformed MI.

Table 2 indicates that although there is considerable similarity between GA and the proposed method there are substantial differences between the inputs selected, and GA and the proposed method selected the different number of inputs for all level, especially in level 3, the number of inputs is large than MI, which explains the reason why the performance of them is better than MI. In contrast with GA, the proposed method can get more little number of inputs without loss of performance, and the content of input subset is the hybrid of that GA and MI. In addition, it was found that there was very little increase in performance after 43 generations for the proposed method, but 56 generations for GA.

5 Conclusion

We proposed an effective feature selection scheme using genetic algorithm (GA) combining with mutual information (MI), in which mutual information between each input and each output of the data set is employed in mutation in evolutionary process to purposefully guide search direction based on some criterions. By examining the forecasting at the Australian Bureau of Meteorology, the simulation of three different methods of feature selection shows that the proposed method can reduce the dimensionality of inputs, speed up the training of the network and get better performance.

References

1. Dash, M., Liu, H.: Feature selection for classification. Intelligent Data Analysis, vol. 1 (1997). 131–156.
2. D.C. Montgomery and E.A. Peck: Introduction to Linear Regression Analysis. John Wiley & Sons, New York (1982).
3. A. Sen and M. Serivastava: Regression Analysis: Theory, Methods, and Applications. Springer-Verlag, New York (1990).

4. Holz, H. J. and Loew, M. H.: Relative feature importance: A classifier-independent approach to feature selection. In: Gelsema, E. S. and Kanal, L. N. (eds.), Pattern Recognition in Practice IV. Amsterdam: Elsevier (1994) 473-487.
5. H. Wang, D. Bell, F. Murtagh: Automatic approach to feature subset selection based on relevance. IEEE Trans. PAMI 21 (3). (1999) 271–277.
6. Belinda Choi, Tim Hendtlass, and Kevin Bluff: A Comparison of Neural Network Input Vector Selection Techniques. LNAI 3029. Springer-Verlag Berlin Heidelberg (2004) 1-10.
7. N. Kwak, C-H. Choi: Input feature selection by mutual information based on parzen window. IEEE Trans. PAMI 24 (12). (2002) 1667–1671.
8. D. Huang,Tommy W.S. Chow: Effective feature selection scheme using mutual information. Neurocomputing vol. 63. (2005) 325 – 343.
9. A.M. Fraser & H.L. Swinney: Independent Coordinates for Strange Attractors from Mutual Information. Physical Review A, Vol. 33/2. (1986) 1134 – 1140.
10. T. Cibas, F.F. Soulie, P. Gallinari and S. Raudys: Variable selection with neural networks. Neurocomputing 12. (1996) 223–248.
11. W. Siedlechi and J. Sklansky: A note on genetic algorithms for large-scale feature selection. Pattern Recognition Letters, 10. (1989) 335–347.
12. C. Emmanouilidis, A. Hunter, J. Macintyre and C. Cox: Selecting features in neurofuzzy modeling by multiobjective genetic algorithms. Artificial Neural Networks, (1999) 749–754.
13. J.H. Yang and V. Honavar: Feature Subset Selection Using a Genetic Algorithm. IEEE Intelligent Systems, vol. 13, no. 2. (1998) 44-49.
14. Oh, I.-S., Lee, J.-S., Moon, B.-R.: Hybrid genetic algorithms for feature selection. IEEE Trans. Pattern Analysis and Machine Intelligence 26. (2004) 1424–1437.
15. J. LeMarshall: An Intercomparison of Temperature and Moisture Fields Derived from TIROS Operational Vertical Sounder Data by Different Retrieval Techniques. Part I: Basic Statistics. Journal of Applied Meteorology, Vol 27. (1988) 1282 – 1293.
16. T.M. Cover, J.A.: Thomas: Elements of Information Theory. Wiley, New York, (1991).