A New Supervised Classification of Credit Approval Data via the Hybridized RBF Neural Network Model Using Information Complexity

Oguz Akbilgic and Hamparsum Bozdogan

Abstract In this paper, we introduce a new approach for supervised classification to handle mixed-data (i.e., categorical, binary, and continuous) data structures using a hybrid radial basis function neural networks (HRBF-NN). HRBF-NN supervised classification combines regression trees, ridge regression, and the genetic algorithm (GA) with radial basis function (RBF) neural networks (NN) along with information complexity (ICOMP) criterion as the fitness function to carry out both classification and subset selection of best predictors which discriminate between the classes. In this manner, we reduce the dimensionality of the data and at the same time improve classification accuracy of the fitted predictive model. We apply HRBF-NN supervised classification to a real benchmark credit approval mixed-data set to classify the customers into good/bad classes for credit approval. Our results show the excellent performance of HRBF-NN method in supervised classification tasks.

1 Introduction

Credit approval is one of the most critical decisions of banking requiring solid risk analysis. Credit scoring systems are introduced to evaluate the customers' eligibility for credit approval based on historical and current information about the customers. This information can be numeric such as income, age, volume of previous credit history as well as nominal-categorical such as sex, race, type of criminal record, and so on. Although processing such nominal-categorical variables

O. Akbilgic
Department of Business Analytics and Statistics, University of Tennessee, Knoxville, TN 37996, USA

Department of Quantitative Methods, Istanbul University School of Business, Istanbul, Turkey
e-mail: oguzakbilgic@gmail.com

H. Bozdogan (✉)
Department of Business Analytics and Statistics, University of Tennessee, Knoxville, TN 37996, USA
e-mail: bozdogan@utk.edu

B. Lausen et al. (eds.), *Data Science, Learning by Latent Structures, and Knowledge Discovery*, Studies in Classification, Data Analysis, and Knowledge Organization, DOI 10.1007/978-3-662-44983-7_2

can be easy by simple credit scoring systems, it can be difficult to handle them in more sophisticated statistical methods for credit approval decision making.

Traditional techniques such as discriminant analysis and logistics regression suffer in the presence of nominal-categorical data. When the variables are nominal (categorical) definitions of the similarity (dissimilarity) measures become difficult and it requires a new metric. In this paper, our objective is to introduce a new approach for supervised classification using a hybrid radial basis function neural networks (HRBF-NN) with continuity justification on dependent variable so as to handle mixture of nominal-categorical and continuous predictors without using dummy variables for classification. We illustrate the practical utility and the importance of our approach by providing a real example on a benchmark credit approval data from the banking industry to classify good and bad customers. Most of the technical details of this paper can be found in Akbilgic et al. (2013), Akbilgic (2011), Akbilgic and Bozdogan (2011). Here, we only recapitulate the necessary parts from these papers to set up the background of this current paper.

The paper is organized as follows. In Section 2, we briefly explain HRBF-NN and what radial basis function neural network (RBF-NN) model is. In Section 3, we discuss classification trees (CT) and its usage in HRBF-NN model; transforming tree nodes into RBFs. Estimation of the weight parameters is presented in Section 4 using the least-squares method. Later, we explain how to make classification problem look like non-parametric regression by adding a threshold function into output neuron of RBF-NN model. Our threshold function turns out to be a non-linear function of the predictive model. For other threshold selection methods, we refer the readers to Flach et al. (2013) in this volume. In Section 5, for model selection, we develop and use information-theoretic measure of complexity (ICOMP) criterion as our fitness function and show its derived form under both correctly and misspecified HRBF-NN models. We also give the forms of Akaike's information criterion (AIC) (Akaike 1973; Bozdogan 1987) and Rissanen/Schwarz (MDL/SBC) (Rissanen 1978; Schwarz 1978). In Section 6, we briefly explain the background of the genetic algorithm (GA) and the implementation of GA for the subset selection of the best predictors which discriminate between the classes. In Section 7, we give a numerical example to illustrate the performance of the proposed new supervised classification approach via the HRBF-NN model on a real credit approval data set to classify the customers into good/bad credit card customers or classes. Later, in Section 8, we conclude the paper with a discussion.

2 Hybrid Radial Basis Function Neural Networks: HRBF-NN Model

In this section, we briefly introduce the structure of HRBF-NN model as a combination of RBF-NNs, classification trees, ridge regression, information complexity ICOMP, and the genetic algorithm (GA).

2.1 RBF-NN Model

RBF-NNs model is a technique that transforms *non-linearly separable features* to *linearly separable features* using *radial basis functions* (*RBFs*). *RBF-NN model* is a *nonparametric regression* technique (Bishop 1995) defined as

$$y = f(w, x) = \sum_{j=1}^{m} w_j h_j(x) = w_1 h_1 + w_2 h_2 + \cdots + w_m h_m. \tag{1}$$

In equation (1), y is the dependent variable, $x_1, x_2, \ldots, x_m$ are independent variables, $\{h_j(x)\}_{j=1}^{m}$, and $\{w_j\}_{j=1}^{m}$ are the unknown adaptable coefficients, or weights. Equation (1) is represented in matrix form in equation (2) where H is the $(n \times m)$ *design matrix*, and ε is an $(n \times 1)$ vector of random noise term, such that

$$y = Hw + \epsilon. \tag{2}$$

2.2 Radial Basis Functions

RBF-NN gains its flexibility from RBFs. We shall consider four most common RBFs in this work although there are many others. These are Gaussian (GS), Cauchy (CH), multi-quadratic (MQ), and inverse multi-quadratic (IMQ) which are given in Table 1.

The RBF-NN non-linearly transforms n-dimensional inputs to m-dimensional space by m basis functions, each characterized by their centers c_j in the (original) input space and a width or radius vector r_j , $j \in \{1, 2, \ldots, m\}$ (Orr 2000).

Table 1 The most common radial basis functions

RBF kernels	Functional form
Gaussian (GS)	$h_j(x) = \exp\left(-\sum_{k=1}^{p} \frac{(x_k - c_{jk})^2}{r_{jk}^2}\right)$
Cauchy (CH)	$h_j(x) = \sqrt[2]{1 + \exp\left(-\sum_{k=1}^{p} \frac{(x_k - c_{jk})^2}{r_{jk}^2}\right)}$
Multiquadric (MQ)	$h_j(x) = \sqrt[2]{1 + \exp\left(-\sum_{k=1}^{p} \frac{(x_k - c_{jk})^2}{r_{jk}^2}\right)}$
Inverse multiquadric (IMQ)	$h_j(x) = \frac{1}{\sqrt[2]{1+\exp\left(-\sum_{k=1}^{p} \frac{(x_k - c_{jk})^2}{r_{jk}^2}\right)}}$

3 Classification Trees and its Use in HRBF-NN Model

3.1 Classification Trees

Classification and regression trees (or CART in short) models are used for both prediction and classification. Classification trees algorithm is based on recursively partitioning of the input space into two parallel hyper-rectangles. The hyper-rectangles with no more splits are called terminal nodes of the tree and a class label is assigned for each terminal nodes. The class assignment rule for a terminal node is simply to correspond the class label having the largest number of members in the terminal node (Sutton 2005).

During the process of recursive partitioning of input space, each split is parallel to one of the axes and can be expressed as an inequality involving of the input components (e.g. $x_k > b$). The input space is divided into hyper-rectangles organized into a binary tree where each branch is determined by the dimension (k) and boundary (b) which together minimize the misclassification error (Orr 2000). The root node of the classification tree is the smallest hyper-rectangle that will include all of the training data $\{x_i\}_{i=1}^{p}$. Its size s_k (half-width) and center c_k in each dimension k are

$$s_k = \frac{1}{2}(\max_{i \in S}(x_{ik}) - \min_{i \in S}(x_{ik}) \tag{3}$$

$$c_k = \frac{1}{2}(\max_{i \in S}(x_{ik}) + \min_{i \in S}(x_{ik})) \tag{4}$$

where $k \in K$ is the set of predictor indices, and $S = \{1, 2, \ldots, p\}$ is the set of training set indices. A split of the root node divides the training samples into left and right subsets, S_L and S_R, on either side of a boundary b in one of the dimensions k such that

$$s_L = \{i : x_{ik} \leq b\}, \tag{5}$$

$$s_R = \{i : x_{ik} > b\}, \tag{6}$$

In classification trees, for a given set of class labels $\{A_1, A_2, , A_3 \ldots\}$, the output values of each side of the bifurcations are

$$\hat{y}_L = A_{\text{argmax}_{i \epsilon s_L}\{a_i\}} \tag{7}$$

$$\hat{y}_R = A_{\text{argmax}_{i \epsilon s_R}\{a_i\}} \tag{8}$$

where the number of members of class label in each subset is defined with the set $a = \{a_1, a_2, a_3 \ldots\}$. The misclassification error (MCE) rate is then

$$\text{MCE}(k, b) = \frac{\sum_{i \epsilon s_L} M(y_i, \hat{y}_L) + \sum_{i \epsilon s_R} M(y_i, \hat{y}_L)}{n}, \tag{9}$$

where n is the total sample size, and $M(y_i, \hat{y}_L)$ is a function equal to 0 if $y_i = \hat{y}$, and 1 otherwise.

The split which minimizes MCE (k, b) over all possible choices of k and b is used to create the children of the root node and is found by simple discrete search over m dimensions and p observations. The children of the root node are split recursively in the same manner and the process terminates when every remaining split creates children containing fewer than $p_{\min}$ samples, which is another parameter of the method. The children are shifted with respect to their parent nodes and their sizes reduced in the k-th dimension (Akbilgic et al. 2013; Akbilgic 2011; Akbilgic and Bozdogan 2011).

3.2 Transforming Tree Nodes into RBFs

The classification trees contain a root node, some non-terminal nodes (having children) and some terminal nodes (having no children). Each node is associated with a hyper-rectangle of input space having a center c and size s as described above. The node corresponding to the largest hyper-rectangle is the root node and it is divided up into smaller and smaller pieces progressing down the tree (Breiman et al. 1984; Orr 2000). To transform the hyper-rectangle into different basis kernel RBFs we use its center c as the RBF center and its size s, scaled by a parameter α as the RBF radius given by

$$r = \alpha s. \tag{10}$$

The scalar α has the same value for all nodes (Kubat 1998), and it is another parameter of the method. In this study we set $\alpha = \sqrt{2}\alpha_K^{-1}$ where α_K is the Kubat's parameter (Kubat 1998; Orr 2000).

4 Estimation of Weight Parameters

4.1 Least-Squares Estimation

Given a network model in equation (1) consisting of m RBFs with centers $\{c_j\}_{j=1}^m$ and radii $\{r_j\}_{j=1}^m$ and a training set with p patterns, $\{(x_i, y_i)\}_{i=1}^p$, the optimal network weights can be found by minimizing the sum of squared errors:

$$\text{SSE} = \sum_{i=1}^{p} (f(x_i) - y_i)^2 \tag{11}$$

and is given by

$$\hat{w} = \left(H'H\right)^{-1} H'y \tag{12}$$

the so-called least squares estimation. Here H is the design or model matrix, with its elements $H_{ij} = h_j(x_i)$, and $y = \left(y_1, y_2, \ldots, y_p\right)'$ is the p-dimensional vector of training set of output values.

In RBF-NN, one of the most common problems is singularity of the $(H'H)$ matrix. At this point, to overcome possible singularity problem in the model matrix, we use global ridge regression (Tikhonov and Arsenin 1977; Bishop 1991) to regularize HRBF-NN model with the cost function given by

$$C(w, \lambda) = \sum_{i=1}^{p} \left(f(x_i) - y_i\right)^2 + \lambda \sum_{i=1}^{m} w_j^2 = \varepsilon'\varepsilon + w'w. \tag{13}$$

$C(w, \lambda)$ is minimized to find a weight vector which is more robust to noise in the training set. The optimal weight vector for global ridge regression is given in equation (14), where I_m is the m dimensional identity matrix, and λ is the regularization parameter.

$$\hat{w} = \left(H'H + \lambda I_m\right)^{-1} H'y. \tag{14}$$

We use *Hoerl, Kennard, and Baldwin (HKB)* (Hoerl et al. 1975) approach to data adaptively determine optimal λ that is given by

$$\hat{\lambda}_{\text{HKB}} = \frac{ms^2}{\hat{\mathbf{w}}'_{LS}\hat{\mathbf{w}}_{LS}}, \tag{15}$$

where $m = k$, the number of predictors not including the intercept term, n is the number of observations, s^2 is the estimated error variance using k predictors so that

$$s^2 = \frac{1}{(n-k+1)} \left(y - H\hat{\mathbf{w}}_{LS}\right)' \left(y - H\hat{\mathbf{w}}_{LS}\right), \tag{16}$$

where $\hat{\mathbf{w}}_{LS}$ is the estimated coefficient vector obtained from a no-constant model given in matrix form by

$$\hat{\mathbf{w}}_{LS} = \left(H'H\right)^{-1} H'y. \tag{17}$$

4.2 RBF Neural Networks for Classification

The goal of classification is to assign observations into target categories or classes based on their characteristics in some optimal way. Thus, in classification case, outcomes are one of discrete set of possible classes rather than of a continuous function as in non-parametric regression (Bishop 1995). However, we can make the classification problem look like a non-parametric regression by incorporating a threshold function into the output of the neuron of the RBF-NN model.

For a binary dependent variable case, we can assign HRBF-NN predictions to class labels by substituting equation (1) in the threshold function $t\left(f(w,H);t_0\right)$ given by

$$t\left(f(w,H)\right)=\begin{cases}0 & f(w,x)<t_0\\ 1 & f(w,x)>t_0\end{cases} \tag{18}$$

where t_0 is the value separating two classes.

When two clusters have equal number of observations, then $t_0=0.5$.

Assuming that the classes are represented with 0, and 1 and having n_1, and n_2, the number of observations in each class, the calculation of threshold value is given by

$$t_0=\frac{n_1}{n_1+n_2}. \tag{19}$$

Threshold value can be considered as a prior probability of the first group which is equal to 0.5 when two of the groups have equal number of observations.

5 Information Theoretic Model Selection Criteria

In HRBF-NN, we use ICOMP criterion of Bozdogan (1994, 2000, 2004) and Liu and Bozdogan (2004) as the fitness function to carry out variable selection with GA. The complexity of a nonparametric regression model increases with the number of independent and adjustable parameters, which is also termed effective degrees of freedom in the model. According to the qualitative principle of Occam's Razor, the simplest model that fits the observed data is the best model. Following this principle, we aim to provide a trade-off between how well the model fits the data and the model complexity (Akbilgic et al. 2013).

The derived forms of information criteria are used to evaluate and compare different horizontal and vertical subset selection in the genetic algorithm (GA) for the regularized regression and classification trees and RBF networks model given in equation (1) under the assumption, $\varepsilon\sim N\left(0,\sigma^2 I\right)$ or equivalently $\varepsilon_i\sim N\left(0,\sigma^2\right)$ for $i=1,2,\ldots,n$.

General form of ICOMP is an approximation to the sum of two Kullback–Leibler (KL) (Kullback and Leibler 1951) distances. For general multivariate normal

linear or nonlinear structural model suppose $C_1\left(\hat{\Sigma}_{\text{model}}\right)$ is approximated by the complexity of the IFIM $C_1\left(\hat{\mathscr{F}}^{-1}\left(\hat{\theta}\right)\right)$. Then, we define ICOMP(IFIM) as

$$\text{ICOMP(IFIM)} = -2\log L\left(\hat{\theta}\right) + 2C_1\left(\hat{\mathscr{F}}^{-1}\left(\hat{\theta}\right)\right), \tag{20}$$

where $C_1(.)$ is a maximal information theoretic measure of complexity of the estimated inverse Fisher information matrix (IFIM) of a multivariate normal distribution given by

$$C_1\left(\hat{\mathscr{F}}^{-1}\left(\hat{\theta}\right)\right) = \frac{s}{2}\log\left(\frac{\text{tr}\left(\hat{\mathscr{F}}^{-1}\left(\hat{\theta}\right)\right)}{s}\right) - \frac{1}{2}\log \mid \hat{\mathscr{F}}^{-1}\left(\hat{\theta}\right) \mid, \tag{21}$$

and where $s = \dim\left(\hat{\mathscr{F}}^{-1}\right) = \text{rank}\left(\hat{\mathscr{F}}^{-1}\right)$. The estimated IFIM for the HRBF-NN model is given by

$$\widehat{\text{Cov}}\left(\hat{w}, \hat{\sigma}^2\right) = \hat{\mathscr{F}}^{-1} = \begin{bmatrix} \hat{\sigma}^2 \left(H'H\right)^{-1} & 0 \\ 0 & \frac{2\hat{\sigma}^4}{4} \end{bmatrix}, \tag{22}$$

where

$$\hat{\sigma}^2 = \frac{(y - H\hat{w})'(y - H\hat{w})}{n}. \tag{23}$$

Then, the definition of ICOMP(IFIM) in equation (20) becomes:

$$\text{ICOMP(IFIM)} = n\log(2\pi) + n\log\left(\hat{\sigma}^2\right) + n + 2C_1\left(\hat{\mathscr{F}}^{-1}\left(\hat{\theta}\right)\right), \tag{24}$$

where the entropic complexity is

$$C_1\left(\hat{\mathscr{F}}^{-1}\left(\hat{\theta}_m\right)\right) = (m+1)\log\left[\frac{\text{tr}\hat{\sigma}^2\left(H'H\right)^{-1} + \frac{2\hat{\theta}^4}{4}}{m+1}\right] - \frac{1}{2}\log \mid \hat{\sigma}^2\left(H'H\right)^{-1} \mid + \log\left(\frac{2\hat{\sigma}^4}{4}\right). \tag{25}$$

We can also define ICOMP for misspecified models given as follows:

$$\begin{aligned} \text{ICOMP(IFIM)}_{\text{Misspec}} &= -2\log L\left(\hat{\theta}\right) + 2C_1\left(\widehat{\text{Cov}}\left(\hat{\theta}\right)_{\text{Misspec}}\right) \\ &= n\log(2\pi) + n\log\left(\hat{\sigma}^2\right) + n + 2C_1\left(\widehat{\text{Cov}}\left(\hat{\theta}\right)_{\text{Misspec}}\right), \end{aligned} \tag{26}$$

where

$$\widehat{\mathrm{Cov}}\left(\hat{\theta}\right)_{\mathrm{Misspec}} = \hat{\mathscr{F}}^{-1}\hat{\mathscr{R}}\hat{\mathscr{F}}^{-1} \tag{27}$$

$$\begin{bmatrix} \hat{\sigma}^2(H'H)^{-1} & 0 \\ 0 & \frac{2\hat{\sigma}^4}{n} \end{bmatrix} \begin{bmatrix} \frac{1}{\hat{\sigma}^4}H'D^2H & H'1\frac{Sk}{2\hat{\sigma}^3} \\ \left(H'1\frac{Sk}{2\hat{\sigma}^3}\right)' & \frac{(n-m)(Kt-1)}{4\hat{\sigma}^4} \end{bmatrix} \begin{bmatrix} \hat{\sigma}^2(H'H)^{-1} & 0 \\ 0 & \frac{2\hat{\sigma}^4}{n} \end{bmatrix}$$

is a consistent estimator of the covariance matrix $\mathrm{Cov}(\theta_k^*)$, which is often called the sandwich covariance or robust covariance estimator, since it is a correct covariance regardless whether the assumed model is correct or not. When the model is correct we get $\hat{\mathscr{F}} = \hat{\mathscr{R}}$. Hence, the sandwich covariance reduces to the usual IFIM $\hat{\mathscr{F}}^{-1}$ (White 1982). Note that this covariance matrix takes into account the presence of skewness and kurtosis, which is not possible with AIC (Akaike 1973) and other Akaike-type criteria such as Rissanen/Schwarz (MDL/SBC) (Rissanen 1978; Schwarz 1978). The derived forms of these criteria for the HRBF-NN model are:

$$\mathrm{AIC}(m) = n\log(2\pi) + n\log\left(\frac{(y - H\hat{w})'(y - H\hat{w})}{n}\right) + n + 2(m+1), \tag{28}$$

$$\mathrm{MDL/SBC}(m) = n\log(2\pi) + n\log\left(\frac{(y - H\hat{w})'(y - H\hat{w})}{n}\right) + n + m\log(n). \tag{29}$$

6 Genetic Algorithm for Subset Selection

There are several standard techniques available for variable selection such as forward selection, backward elimination, a combination of the two, or all possible subset selection. Both forward and backward procedures cannot deal with the collinearity in the predictor variables. Major criticisms on the forward, backward, and stepwise selection are that, little or no theoretical justification exists for the order in which variables enter or exit the algorithm. Stepwise searching rarely finds the overall best model or even the best subsets of a particular size. Stepwise selection, at the very best, can only produce an *"adequate"* model.

All possible subset selection is a fail proof method, but it is not computationally feasible. It takes too much time to compute and it is costly. For 20 predictor variables, for the usual subset regression model, total number of possible models we need to evaluate is: $2^{20} = 1{,}048{,}576$. At this point, we use genetic algorithm to carry out variable selection in HRBF-NN with *ICOMP* as the fitness function.

Genetic algorithm is a robust evolutionary optimization search technique with very few restrictions (David and Alice 1996). GA treats information as a series of codes on a binary string, where each string represents a different solution for a

given problem. It follows the principles of survival of the fittest, which is introduced by Charles Darwin. The algorithm searches for optimum solution within a defined search space to solve a problem (Eiben and Smith 2010). It has outstanding performance in finding the optimal solution for problems in many different fields (Akbilgic et al. 2013; Akbilgic 2011; Akbilgic and Bozdogan 2011).

7 A Numerical Example: Analysis of Credit Approval Data

In this section, we report our computational results on a credit approval data sets to classify the customers into good/bad classes using our hybrid RBF-NN approach with regularization, GA, and ICOMP(IFIM) as the fitness function.

Our modern world depends upon credit. Entire economies are driven by people's ability to *"buy-now, pay later"* (Anderson 2007).

Therefore, credit approval is one of the most critical decisions of banking industry requiring solid risk analysis.

Credit scoring systems are introduced almost 50 years ago to evaluate the customers' eligibility for credit approval based on historic and current information about the customers.

This information can be numeric such as *income*, *age*, *volume of previous credit history* as well as nominal-categorical such as *sex*, *race*, *type of criminal record*, and so on with high dimensions.

Our *credit approval data set* is obtained from UCI Machine Learning Repository (2013). Original version of credit approval data set is consisted of 690 observations including fifteen independent variables; six continuous and nine categorical, and one binary dependent variable. However, by excluding the observation with missing attributes, we reduced the data size to 654 representing 296 positive, and 358 negative credit ratings. Because all of the nine categorical independent variables were coded by meaningless letters to protect confidentiality of the data, we transformed them into numbers, $1, 2, 3, \ldots$, based on the number of categories in each variable. The representation of the original data and the usage of them in our study are given in Table 2.

We first analyzed credit approval data via HRBF-NN model separately for four different RBFs: Gaussian, Cauchy, Multi-Quadratic, and Inverse Multi-Quadratic using saturated model. Confusion matrix for different RBFs are reported in the Tables 3, 4, and 5 where ICOMP(IFIM)miss values are reported in the last column of Table 8. For simplicity in text, we will use ICOMP for ICOMP(IFIM)miss in our report in this study. Note that calculation of classification accuracy is carried out using equation (30). The reason we run HRBF-NN model for saturated model is to compare the results after variable selection. The classification accuracy is defined by

$$\text{Classification accuracy} = 100\frac{\text{number of correctly classified observations}}{\text{total number of observations}}. \quad (30)$$

Table 2 Usage of credit approval data in our analysis

Variables	Original presentation	Usage in our analysis
A1	b, a	1, 2
A2	Continuous	Continuous
A3	Continuous	Continuous
A4	u, y, l, t	1, 2, 3, 4
A5	g, p, gg	1, 2, 3
A6	c, d, cc, i, j, k, m, r, q, w, x, e, aa, ff	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14
A7	v, h, bb, j, n, z, dd, ff, o	1, 2, 3, 4, 5, 6, 7, 8, 9
A8	Continuous	Continuous
A9	t, f	1, 2
A10	t, f	1, 2
A11	Continuous	Continuous
A12	t, f	1, 2
A13	g, p, s	1, 2, 3
A14	Continuous	Continuous
A15	Continuous	Continuous
A16	+, − (class attributes)	1, 2

Table 3 Gaussian

Classes	C1	C2	Total	Accuracy (%)
C1	274	22	296	92.57
C2	39	319	358	89.11
Overall			654	90.67

Table 4 Cauchy

Classes	C1	C2	Total	Accuracy (%)
C1	270	26	296	91.22
C2	36	322	358	89.94
Overall			654	90.52

Table 5 MQ

Classes	C1	C2	Total	Accuracy (%)
C1	275	21	296	92.91
C2	52	306	358	85.48
Overall			654	88.84

Tables 3, 4, 5, and 6 show the high performance of HRBF-NN model for classification of credit data which is approximately 90 %. At this point we run variable selection on credit data using GA with ICOMP as the fitness function. Parameter setting of GA is based on our previous studies on HRBF-NN model (Akbilgic et al. 2013). Thus, we set our GA parameters as given in Table 7.

After finishing the first stage of analysis for saturated model and setting the GA parameters, next, we carried out variable selection for credit data using GA separately for four different RBFs. Table 8 shows the selected variable subsets and

Table 6 Inverse MQ RBF

Classes	C1	C2	Total	Accuracy (%)
C1	271	25	296	91.55
C2	33	325	358	91.34
Overall			654	91.44

Table 7 Parameter setting of GA for variable selection

Parameter	Setting
Number of generations	35
Number of populations	20
Mutation probability	0.01
Crossover probability	0.65
Crossover type	Single point
Elitism rule	Yes

Table 8 Variable selection under different RBFs

RBF type	Best subset	ICOMP: Best subset	ICOMP: Saturated model
Gaussian	3-6-9-10-14	191.45	461.56
Cauchy	3-5-6-9-10-11-14	191.28	400.49
Multi-quadratic	1-3-4-7-9-10-11-13-14	248.00	570.77
Inverse multi-quadratic	3-4-5-6-9-10-13-14	214.28	491.45

Table 9 Gaussian RBF

Classes	C1	C2	Accuracy (%)
C1	269	27	90.88
C2	33	325	90.78
Overall			90.83

minimized ICOMP values under selected variable subsets for different RBFs. We also showed the ICOMP values we calculated before for saturated model in Table 8 to give a better comparison.

It is noted from Table 8 that ICOMP values for selected subsets are significantly lower than the ICOMP values calculated for saturated model. At this point, it is important to see if obtained lower ICOMP values correspond to a simple model giving good classification accuracy. To show this, we run HRBF-NN model for all four of the RBFs with corresponding selected best subsets given in Table 8. Confusion matrix and classification accuracy is calculated for each case and the results are reported in Tables 9, 10, 11, 12.

The important results appearing in Tables 9, 10, 11, and 12 show that variable selection within HRBF-NN allows us to reduce dimension of input variables without any loss in classification accuracy. Comparing the classification accuracy results

Table 10 Cauchy RBF

Classes	C1	C2	Accuracy (%)
C1	264	32	89.19
C2	37	321	89.66
Overall			89.45

Table 11 MQ RBF

Classes	C1	C2	Accuracy (%)
C1	271	25	91.55
C2	57	321	89.66
Overall			90.52

Table 12 IMQ RBF

Classes	C1	C2	Accuracy (%)
C1	268	28	90.54
C2	32	326	91.06
Overall			90.83

between saturated model and best subsets shows the similarity of classification performance while the dimensionality is significantly reduced for best subsets. According to Table 8, by carrying out variable selection with Gaussian RBF has resulted in selecting a subset with only five variables out of fifteen where ICOMP value is minimized. Note that, there is even slightly better classification accuracy for best subset selected for Gaussian RBF in comparison with classification accuracy for the saturated model.

Finally, for comparison purposes, we carried out the usual logistic regression analysis, although the assumptions are violated here for this data set, we obtained a classification accuracy of 87.1 % using stepwise variable selection which gave nine predictors as the best predictors including the constant term. These nine predictors are: 0, 4, 5, 7, 8, 9, 10, 11, and 15. Note that this subset does not include variables 3, 6, 9, 10, and 14 obtained from our results.

8 Conclusions and Discussion

In this paper, we introduced a novel approach for supervised classification using a HRBF-NN model with ICOMP. Our study shows that HRBF-NN model is a highly clever technique to handle hard classification problems even if the data is mixture of continuous and categorical variables. We demonstrated that the GA is a powerful optimization tool for selecting the best subset of predictors that discriminate between the classes or groups. HRBF-NN using ICOMP with GA provides us a flexible variable selection and at the same time a classification tool which gives better results than the full saturated model. With our approach we can now provide a practical method for choosing the best kernel basis RBF for a given

data set which was not possible before in the literature of RBF based-methods. In real-world applications, we frequently encounter data sets with 100 and 1,000 of variables. Our results show that HRBF-NN model is a very flexible procedure that can handle dimensionality reduction drastically without losing information in classification accuracy. In our example, we reduced the number of input variables from fifteen to five with even slightly better classification accuracy which is around 91%. As is well known, recently, kernel-based supervised classification techniques such as the support vector machines (SVMs) and multi-class SVMs have become popular. One problem that has not been addressed in the literature is that kernelization and supervised classification takes place in the high dimensional reproducing Hilbert kernel space (RHKS) and not in the original data space. The transformed kernel space mapping is not one-to-one and onto, and not invertible to the original data space due to the dot product operations in using the kernel trick. This makes the practical interpretation of the results difficult even though one can get good classification error rates.

The new HRBF-NN approach proposed in this paper overcomes the difficulties encountered in the RHKS type supervised classification and provides us a flexible technique in the original data space that combines regression trees, regularized regression, and the genetic algorithm (GA) with radial basis function (RBF) neural networks (NN) along with information complexity ICOMP criterion as the fitness function to carry out both classification and at the same time subset selection of best predictors which discriminate between the classes.

Therefore, we believe our approach is a viable means of data mining and knowledge discovery via the HRBF-NN method.

Acknowledgements This paper was invited as a keynote presentation by Prof. Bozdogan at the European Conference on Data Analysis (ECDA-2013) at the University of Luxembourg in Luxembourg during July 10–12, 2013. Prof. Bozdogan extents his gratitude to the conference organizers: Professors Sabine Krolak-Schwerdt, Matthias Bömer, and Berthold Lausen.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. H. Petrox & F. Csaki, (Eds.), *Second International Symposium on Information Theory* (pp. 267–281). Budapest: Academiai Kiado.

Akbilgic, O. (2011). *Variable selection and prediction using hybrid radial basis function neural networks: A case study on stock markets*. PhD thesis, Istanbul University.

Akbilgic, O., & Bozdogan, H. (2011). Predictive subset selection using regression trees and rbf neural networks hybridized with the genetic algorithm. *European Journal of Pure and Applied Mathematics, 4*(4), 467–485.

Akbilgic, O., Bozdogan, H., & Balaban, M. E. (2013). A novel hybrid RBF neural network model as a forecaster. *Statistics and Computing*. doi:10.1007/s11222-013-9375-7.

Anderson, R. (2007). *The credit scoring toolkit*. Oxford: Oxford University Press.

Bishop, C. M. (1991). Improving the generalization properties of radial basis function neural networks. *Neural Computation, 3*(4), 579–588.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford: Oxford University Press.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and it's analytical extension. *Journal of Mathematical Psychology, 52*(3), 345–370.

Bozdogan, H. (1994). Mixture-model cluster analysis using a new informational complexity and model selection criteria. In H. Bozdogan (Ed.), *Multivariate Statistical Modeling, Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach* (Vol. 2, pp. 69–113). North-Holland: Springer

Bozdogan, H. (2000). Akaike's information criterion and recent developments in informational complexity. *Journal of Mathematical Psychology, 44*, 62–91.

Bozdogan, H. (2004) Intelligent statistical data mining with information complexity and genetic algorithms. In H. Bozdogan (Ed.) *Statistical data mining and knowledge discovery* (pp. 15–56). Boca Raton: Chapman and Hall/CRC

Breiman, L., Freidman, J., Stone, J. C., & Olsen, R. A. (1984). *Classification and regression trees*. Boca Raton: Chapman and Hall.

Credit Approval Data Set by UCI MAchine Learning Repository. http://archive.ics.uci.edu/ml/datasets/Credit+Approval. Cited April 26, 2013

David, W. C., & Alice, E. S. (1996). Reliability optimization of series-parallel systems using a genetic algorithm. *IEEE Transactions on Reliability, 45*(2), 254–266.

Eiben, A. E., & Smith, J. E. (2010). *Introduction to evolutionary computing*. New York: Springer.

Flach, P. A., Hernandez-Orallo, J., & Ferri, C. (2013). *Comparing apples and oranges: Towards commensurate evaluation metrics in classification*. Keynote lecture presented in the European Conference on Data Analysis (ECDA-2013), Luxembourg.

Hoerl, A. E., Kennard, R. W., & Baldwin, K. F. (1975). Ridge regression: Some simulations. *Communications in Statistics, 4*, 105–123.

Kubat, M. (1998). Decision trees can initialize radial basis function networks. *Transactions on Neural Networks, 9*(5), 813–821.

Kullback, A., & Leibler, R. (1951). On information and sufficiency. *Annals of Mathematical Statistics, 22*, 79–86.

Liu, Z., & Bozdogan, H. (2004) Improving the performance of radial basis function classification using information criteria. In H. Bozdogan (Ed.), *Statistical data mining and knowledge discovery* (pp. 193–216). Boca Raton: Chapman and Hall/CRC.

Orr, M. (2000). Combining regression trees and RBFs. *International Journal of Neural Systems, 10*(6), 453–465.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica, 14*(5), 465–471.

Schwarz, G. (1978). Estimating the dimension of model. *Annals of Statistics, 6*, 461–464.

Sutton, C. D. (2005). Classification and regression trees, bagging, and boosting. In *Handbook of statistics* Vol. 24, pp. 303–329. Elsevier B.V. doi: 10.1016/s0169-716(04)24004-4.

Tikhonov, A. H., & Arsenin, V. Y. (1977). *Solutions of ill-posed problems*. New York: Wiley.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica, 50*, 1–25.