# A GA-Based Wrapper Feature Selection for Animal Breeding Data Mining*

Olgierd Unold[1], Maciej Dobrowolski[2], Henryk Maciejewski[1],
Pawel Skrobanek[1], and Ewa Walkowicz[2]

[1] Institute of Computer Engineering, Control and Robotics
Wroclaw University of Technology, Wyb. Wyspianskiego 27, 50-370 Wroclaw, Poland
{olgierd.unold,henryk.maciejewski,pawel.skrobanek}@pwr.wroc.pl
[2] Department of Horse Breeding and Riding,
Wroclaw University of Environmental and Life Sciences
Kożuchowska 6, 51-631 Wrocaw, Poland
{maciej.dobrowolski,ewa.walkowicz}@up.wroc.pl

**Abstract.** Feature selection methods are used to tackle the problem of the curse of the dimensionality of data to be mined. This applies also to the area of animal breeding, in which datasets collect remarkably a large number of animal features. In this paper, we have conducted a comprehensive study of both 12 classification methods as well as 12 GA-based feature selection methods for classification of the Silesian horse data. To assess the performance of the wrappers and the classification methods over the animal dataset we used two metrics: a probability metric Area under the ROC curve (AUC), and a rank metric Root Mean Square Error (RMSE). All of the classifiers and the wrappers were taken from the Weka machine learning software. We find that most of the GA-based wrappers achieved results no worse than high-dimensional dataset. The statistical results obtained make the three classifiers: a decision tree ADT, a logistic regression Log and a bagging method Bag competitive method to be considered in the field of animal breeding data mining.

**Keywords:** Feature selection, Genetic algorithm, Data Mining, Breeding.

## 1 Introduction

The curse of the dimensionality (CofD) refers to the fact that the sample size needed to estimate a function of several variables to a given level of accuracy grows exponentially with the number of variables [1]. Another problem with the CofD, especially in data mining, is that it is one of tha main causes why classifiers over-fit the training data. A way to tackle the problem of CofD is to shrink the input dimension of the function to be estimated. Note that, in many situations, it is possible to reject the redundant information, because of

* This work was conducted as part of research project no. N516 415138 financed by the Ministry of Science and Higher Education.

E. Corchado et al. (Eds.): HAIS 2012, Part II, LNCS 7209, pp. 200–209, 2012.

(1) some of the variables may be correlated through linear combinations or other functional dependence with each other, and (2) some of the variables may have a variation smaller than the measurement noise and therefore may be irrelevant.

Scientists in many fields face the problem of simplifying high-dimensional data by finding low-dimensional structure in it. This also applies to the area of animal breeding, in which datasets collect very often a large number of animal features. There is few study on the use of data minig in phenotype driven animal breeding [6]. Although in [2] the performance of 6 classifiers over a horse breeding problem was investigated, the authors used genetic information (microsatellite markers) to perform a prediction of an individuals breed. To our knowledge an extensive evaluation of both feature selection as well as machine learning methods in the context of phenotype driven animal breeding data has not been conducted.

In this work, we carry an elaborate performance study of both different classification methods and different feature selection methods applied to Silesian horse dataset [10]. We investigate 12 classification methods, and for feature selection, we use and compare 12 GA-based wrapper approaches. We find that while most feature selection approaches give similar results, without impairing the results in comparison with an input, non-reduced dataset, there are three distinct classification methods outperforming all other chosen ones.

For the study, we used feature and classification methods taken from the standard Weka[1] software [11], and from the fuzzy-rough version of Weka[2] [8].

The remainder of this paper is organized as follows. Section 2 describes animal breeding data, a set of classifiers and GA-based wrapper feature selection methods mining the data, also Weka software used in this work. Section 3 shows the results obtained, and finally the conclusions are drawn in Section 4.

## 2 Data and Methods

### 2.1 Data

In our work, we used Silesian horse dataset [10]. This database consists of 18.980 records containing zoometric and breeding information of Silesian horses released in Poland in the last 50 years. The attributes depict such features of the horse as: the year of birth, the geographical location (the breeder), type of ownership (private and national), the assignment to the breeding program, origin (identification key of a father, identification key of a mother, family, race), relationship (offspring, inbreeding - connexion with the other records), sex, size, zoometric indexes, bonitation. Most of the entries in the Silesian horse database is information about the mares - 13.408 records, the remaining 5.572 records are the data for stallions.

A data mining goal was to predict the height at withers of the horse (or more precisely the height at withers of the mare) only on the basis of her parents.

[1] `http://www.cs.waikato.ac.nz/ml/weka/`, last accessed 13 November 2011.

[2] `http://users.aber.ac.uk/rkj/book/wekafull.jar`, last accessed 13 November 2011.

The height at the withers is a crucial factor that determines the usefulness of Silesian horses. After the necessary transformation of the dataset, we obtained the database divided up into 4.854 records and 69 attributes. The transformation was based on a joining all the attributes of the mother, father and offspring to one record, and then removing all data and string attributes as non-informative ones. Among these 69 attributes, 28 were nominal. The dataset contained missing values. The target feature was binary: 0 for the horse with the height less than the average of a population, and 1 for the height more than the average of the population. The dataset was balanced 2.521/2.333 records with the target feature of 0/1, respectively. Note that we are not interested in predicting the exact height of an offspring, and only the class of the height at withers of a child.

### 2.2 Classifiers

We conducted experiments comparing the classification accuracy and error of 12 classification methods implemented in the Weka software. The set of used classifiers can be grouped into the following categories:

- *Naive Bayes* classifiers based on the Bayesian Theorem in which it is assumed that the attributes have equal weight and are conditionally independent,
- *Support vector machines* trying to find a hypersurface in the space of possible inputs,
- *Decision trees* creating a hierarchy of nodes, each associated with a decision rule on one attribute,
- *Decision rules* generating rules, which can transformed from or in decision trees,
- *Nearest Neighbor* classifiers (known as instance-based classifiers) using the $k$ nearest neighbors in the feature space to decide which class an object belongs to,
- *Logistic Regression* classifier based on searching for a dependence of the target variable in the form of a logistic function,
- *Bagging* models trained on bootstrap replicates of the training data are combined by voting.

For more information on implemented in the Weka software classifiers see [11].

### 2.3 GA-Based Wrapper Feature Selection

Feature selection (FS) methods can be put into three categories from the point of view of a methods output. One category is about ranking feature according to the same evaluation criterion (filter approach); the other is about choosing a minimum subset of features that satisfies an evaluation criterion (the wrapper approach), and the last regularizes predictor estimation by constraining the dimension of the input space (the embedded approach).

The wrapper approach produces the best results out of the FS methods [12], although this is a time-consuming method since each feature subset considered must be evaluated with the classifier algorithm. In the wrapper method, the attribute subset selection algorithm exists as a wrapper around the data mining algorithm and outcome evaluation. The induction algorithm is used as a black box. The FS algorithm conducts a search for a proper subset using the induction algorithm itself as a part of the evaluation function. GA-based wrapper methods involve a genetic algorithm (GA) as a search method of subset features. GA is a random search method, effectively exploring large search spaces [7]. The basic idea of GA is to evolve a population of individuals (chromosemes), where individual is a possible solution to a given problem. In case of searching the appropriate subset of features, a population consists of different subsets evolved by a mutation, a crossover, and selection operations.

To perform GA-based wrapper feature selection we used a *Wrapper Subset Evaluator* toolbox included in the Weka software. This takes as a parameter the name of the classifier being used to evaluate attribute sets. To estimate the accuracy of a chosen classifier for an examined set of attributes the 5-fold cross validation is used (the Weka default parameter). The Weka *GeneticSearch* method performs a search for attribute subsets using a genetic algorithm. The evaluated population by GA consists of chromosomes, where each chromosome is a list of attribute indexes. After the initial population is chosen randomly, the algorithm evolves (using genetic operators) in such a way that the fitness of chromosomes (assessed by a classifier) increases over the generations. After reaching maximum generations, algorithms returns the chromosome with the highest fitness (in other words the subset of attributes with the highest accuracy).

### 2.4 Experimental Settings

For our evaluation of different classification methods we used 12 different classifiers from the Weka software: two Naive Bayes classifiers (NaiveBayes **NB** and BayesNet **BN**), one type of Support vector machine (**SMO**), one rule-based classifier (**JRip**), three types of decision trees (C4.5 **C45**, ADTree **ADT**, and Random Forest **RF**), three $k$-nearest neighbour algorithms (**IBk**, fuzzy nearest neighbour algorithm **FNN**, and fuzzy-rough nearest neighbour algorithm **FRNN**), one regression model (Logistic regression **Log**), and one bagging model **Bag**. All classifiers were trained using the Weka default parameters.

To investigate the impact of feature selection on the classification performance, 12 wrapper approaches were applied to the dataset. GA is used as random search method with 12 mentioned above different classifiers as induction methods.

Two metrics are used to assess the performance of wrappers and classification methods: RMSE and AUC. AUC is a probability metric, and RMSE a rank metric.

Root Mean Square Error (RMSE) is a metric corresponding to the expected value of the squared error loss or quadratic loss [3]. RMSE is a frequently used measurement of the differences between values predicted by a model or an estimator, and the values actually observed in what is being modelled or estimated.

Area under the ROC curve (AUC) measures the area under a plot of the fraction of positive examples misclassified on the $x$ axis against the fraction of positive examples correctly classified [4]. The AUC of a classifier is equivalent to the probability that the classifier has of ranking a randomly chosen positive instance higher than a randomly chosen negative instance. AUC was proved to be a better measure than accuracy when evaluating and comparing classifiers [9].

The data mining was conducted as follows: first, feature selection was performed on a given dataset. Each GA-based wrapper method used 5-fold cross validation protocol (the Weka default parameter). Then, the optimal feature subset indicated by each wrapper was used as an input data for different classifiers (induction methods). Each induction method split random the data into training and validation parts (in proportion 2/3 and 1/3, respectively), and the entire induction process was repeated 10 times.

The results of the Shapiro-Wilk tests rejected the hypothesis concerning the origination of every tested variable from a normal distribution. The non-parametric Friedman test [5] was used to test the null hypothesis between dependent groups (in each tested group we use the same list of the wrappers or the classifiers). For AUC and RMSE levels, we compared the values between groups using Friedman analysis. Friedman test is a multisample extension of the sign test, a non-parametric randomized block analysis of variance, free from the assumptions of normal distribution and equal variances. The R[3] software package was used for statistical computing.

## 3 Experimental Results

The performance of the 12 GA-based wrappers applied to the dataset is shown in Table 1. As a result, we obtained 12 datasets containing different subsets of features, and additionally the input dataset without dimension reduction (WRP.no). Most of the GA-based wrappers reduced the dimension ca. by a half. The minimum number of features contains the dataset performed by WRP.IBk (13 attributes).

The mean performance of the feature selection methods is shown in Figure 1 and Figure 2. We calculated boxplots of the AUC (Figure 1) and of RMSE (Figure 2) of each of the 12 wrappers plus the dataset without reducing over all classifiers (WRP.no). All boxplots in Figure 1 were ordered by decreasing AUC mean, and in Figure 2 by increasing RMSE mean. Mean AUC ranges between 0.74 (for WRP.C45) to 0.72 (for WRP.FRNN) .

Furthermore, all methods show extreme outliers caused mainly by Ibk induction method (for 8 wrappers), and next FNN (for 4 wrappers).

The AUC levels in wrapper WRP.ADT and WRP.C45 outperformed AUC levels of WRP.FRNN (P=0.05, Friedman test).

The RMSE boxplots indicate that most of the wrappers gained similar mean RMSE, varies from 0.483 for WRP.IBk to 0.497 for WRP.FRNN. The interquartile ranges (IQRs) of AUC and RMSE boxplots demonstrate the significant

[3] `http://www.r-project.org/`

**Table 1.** Results of feature selection of GA-based wrappers (incuding dataset without dimension reduction - WRP.no)

| Wrapper | Number of attributes |
|---|---|
| WRP.no | 69 |
| WRP.SMO | 40 |
| WRP.JRip | 36 |
| WRP.BN | 35 |
| WRP.C45 | 33 |
| WRP.FNN | 32 |
| WRP.RF | 31 |
| WRP.Log | 31 |
| WRP.NB | 29 |
| WRP.Bag | 29 |
| WRP.FRNN | 28 |
| WRP.ADT | 26 |
| WRP.IBk | 13 |

volatility among classifiers. As in the case of AUC, all extreme outliers are caused by two classifiers: Ibk (for 8 wrappers) and FNN (for 5 wrappers).

RMSE levels are not significantly different among wrappers, except wrapper WRP.ADT (P=0.5, Friedman test).

The mean performance of the classifier induction methods is shown in Figure 3 and Figure 4. The AUC boxplots, as well as RMSE boxplots, are no longer so aligned as in case of wrapper boxplots. The IQRs of AUC and RMSE boxplots are significant lower in comparison to wrapper boxplots indicating reduced variability among datasets. Mean AUC ranges between 0.79 (for WRP.Log) to 0.64 (for WRP.Ibk). All classifiers show moderate outliers (casued mainly by WRP.Ibk, which reduced the input dataset from 69 into 13 attributes).

The post-hoc test showed (P=0.05, Friedman test) that AUC levels in ADT and Log were significantly higher as compared with 5 other classifiers: FNN, FRNN, Ibk, Jrip, SMO, the AUC levels of Bag and BN higher than in 4 classifiers: FBB, FRNN, Ibk, SMO, the AUC levels in NB higher as compared with 3 classifiers: FNN, FRNN, Ibk, the levels of AUC in RF higher than in 2 classifiers: FNN and Ibk, and the AUC levels in C45 significant higher than in Log.

The mean RMSE ranges between 0.429 (for WRP.Log) to 0.596 (for WRP.Ibk). The RMSE levels were significantly lower in the classifiers ADT, Log, and Bag as compared with others classifiers: ADT classifier - BN, FNN, Ibk, NB, SMO; Log classifier - FNN, FRNN, Ibk, NB, and SMO; Bag classifier - BN, FNN, FRNN, Ibk, NB, and SMO (P=0.05, Friedman test). The levels of RMSE in Jrip classifier are significantly lower than in 4 other classifiers: FNN, Ibk, NB, and SMO. The RMSE levels in C45 classifier are lower as compared with 3 classifiers: FNN, Ibk, SMO. The RF classifier has the lower RMSE levels than 2 classifiers: FNN and SMO. THE RMSE levels in FRNN classifier are lower than in two classifiers: FNN, Ibk. After all, levels of RMSE in Ibk classifier are significantly lower as compared with RF. All post-hoc tests used Friedman test and P=0.05.

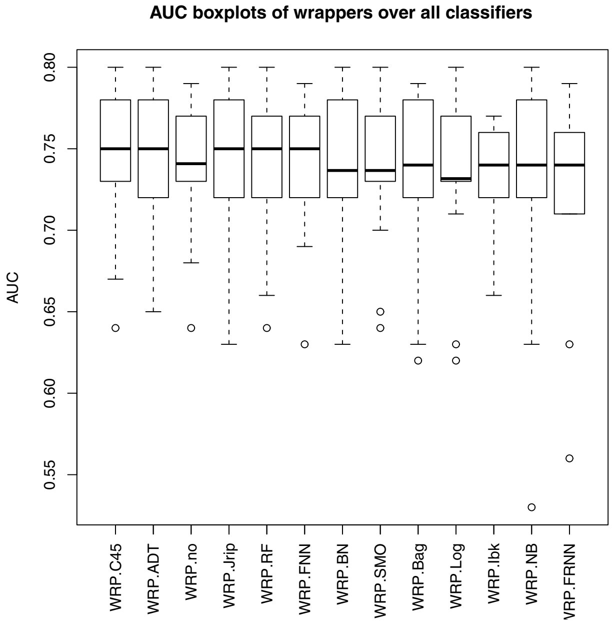

**Fig. 1.** Wrapper performance. AUC boxplots of the wrappers (including data without dimension reducing) over all classifiers ordered by decreasing AUC mean.

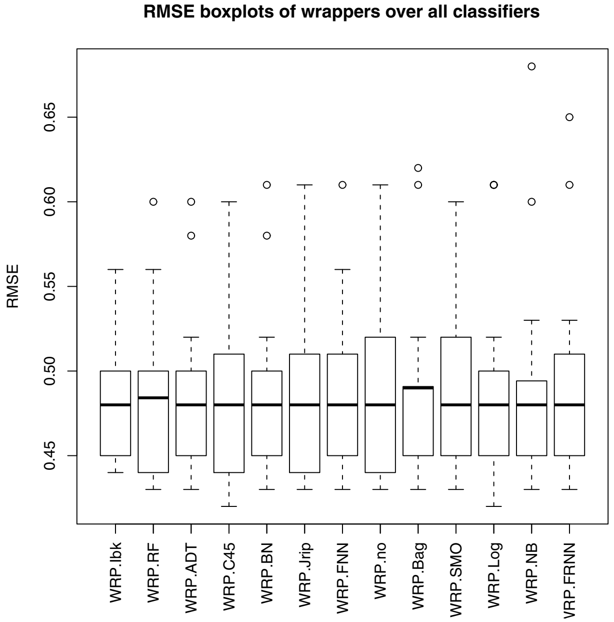

**Fig. 2.** Wrapper performance. RMSE boxplots of the wrappers (including data without dimension reducing) over all classifiers ordered by incresing RMSE mean.

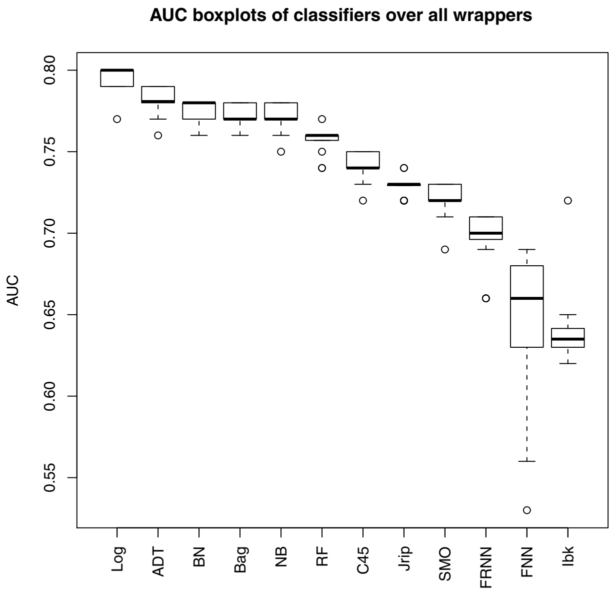

**Fig. 3.** Classifier performance. AUC boxplots of the classifiers over all wrappers (including data without dimension reducing) ordered by decreasing AUC mean.

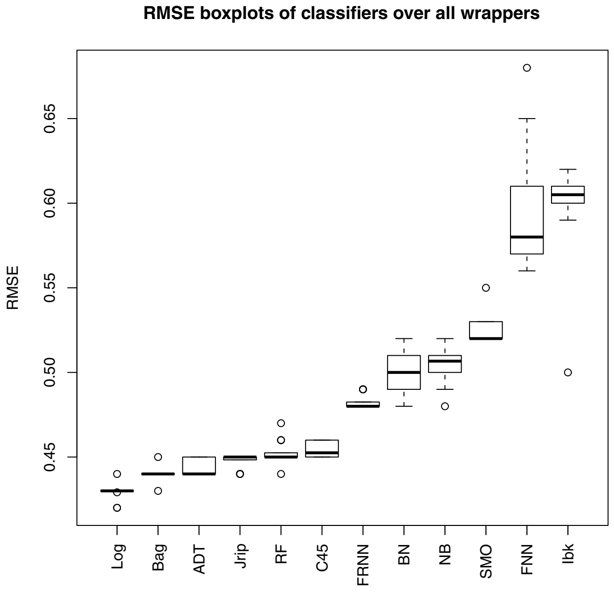

**Fig. 4.** Classifier performance. RMSE boxplots of the classifiers over all wrappers (including data without dimension reducing) ordered by decreasing RMSE mean.

In our study, we found that no significant difference exists between investigated feature selection methods. Most of the GA-based wrappers achieved results no worse than high-dimensional dataset (denoted by WRP.no).

The statistical results obtained and the multiple means comparison tests analyzed make the three classifiers ADT, Log and Bag competitive method to be considered in the field of animal breeding data mining, where there is a strong necessity of obtaining appropriate classification accuracy for dimension-reduced data. These classifiers obtained the highest mean rank when considering all the wrapper-based datasets and all the measures (AUC and RMSE). The statistical tests confirm that these differences are significant when these methods are compared to the other classifiers under study.

Particularly noteworthy is the fact that combination of a wrapper with the same classifier used as an induction method of the wrapper and next as a learning scheme over reduced set of attributes, gives average-good results in terms of AUC and RMSE.

## 4 Conclusions

In this work, we have conducted an extensive study of both classification methods as well as GA-based feature selection methods for classification of animal breeding data. The Weka software was used as a machine learning tool. Therefore, we can assume an almost equal quality of implementations and differences can be attributed to the methods themselves and not to implementations. The experiments focused on identifying the best combination of classifier and feature selection strategy.

In this study, we found that no significant difference exists between investigated feature selection methods. Most of the GA-based wrappers achieved results no worse than high-dimensional dataset (denoted by WRP.no).

The statistical results obtained and the multiple means comparison tests analyzed make the three classifiers ADT, Log and Bag competitive method to be considered in the field of animal breeding data mining, where there is a high necessity of obtaining appropriate classification accuracy for dimension-reduced data. These classifiers obtained the highest/lowest mean rank when considering all the wrapper-based datasets and all the measures (AUC and RMSE, respectively). The statistical tests confirm that these differences are significant when these methods are compared to the other classifiers under study. We are aware that the results drawn are biased by the selected dataset. You have to remember, that Silesian horse dataset is unique, both in terms of the horse species and attributes, and especially difficult would be to compare different wrappers over different - also in a term of dimensionality - datasets.

In a future work, we will investigate further classifiers also other feature selection methods (filters and embedded approaches) to perform more authoritative comparisons. It will be also necessary to explore more than one animal breeding dataset.

## References

1. Bellman, R.: Adaptive Control Processes: A Guided Tour. Princeton University Press, Princeton (1961)
2. Buróczіová, M., Řiha, J.: Horse breed discrimination using machine learning methods. J. Appl. Genet. 50(4), 375–377 (2009)
3. Caruana, R., Niculescu-Mizil, A.: Data mining in metric space: an empirical analysis of supervised learning performance criteria. In: Proceedings of the 10th Int. Conf. Knowl. Disc. Data Mining, pp. 69–78 (2004)
4. Fawcett, T.: An introduction to ROC analysis. Pattern Recognition Letters 27, 861–874 (2006)
5. Friedman, M.: The use of ranks to avoid the assumption of normality implicit in the analysis of variance. Journal of the American Statistical Association 32, 674–701 (1937)
6. Garner, S.R., Holmes, G., McQueen, R.J., Witten, I.H.: Machine learning from agricultural databases: practice and experience. J. Computing 6(1a), 69–73 (1997)
7. Goldberg, D.: Genetic Algorithms in Search, Optimization, and Machine Learning. Addison Wesley (1989)
8. Jensen, R., Shen, Q.: Computational Intelligence and Feature Selection: Rough and Fuzzy Approaches. IEEE Press, Wiley and Sons (2008)
9. Ling, C.X., Huang, J., Zhang, H.: AUC: A Better Measure than Accuracy in Comparing Learning Algorithms. In: Xiang, Y., Chaib-draa, B. (eds.) Canadian AI 2003. LNCS (LNAI), vol. 2671, pp. 329–341. Springer, Heidelberg (2003)
10. Walkowicz, E., Unold, O., Maciejewski, H., Skrobanek, P.: Zoometric indices in Silesian horses in the years 1945-2005. Ann. Anim. Sci. 11(4), 555–565 (2011)
11. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann (2005)
12. Zhiwei, X., Xinghua, W.: Research for Information Extraction Based on Wrapper Model Algorithm. In: 2010 Second International Conference on Computer Research and Development, Kuala Lumpur, Malaysia, pp. 652–655 (2010)