ENHANCED CONFIDENCE INTERPRETATIONS OF GP BASED ENSEMBLE MODELING RESULTS

Michael Affenzeller[(a)], Stephan M. Winkler[(b)], Stefan Forstenlechner[(c)], Gabriel Kronberger[(d)], Michael Kommenda[(e)], Stefan Wagner[(f)], Herbert Stekel[(g)]

[(a – f)] University of Applied Sciences Upper Austria
Heuristic and Evolutionary Algorithms Laboratory
Softwarepark 11, 4232 Hagenberg, Austria

[(a, b)] University of Applied Sciences Upper Austria
Bioinformatics Research Group
Softwarepark 11, 4232 Hagenberg, Austria

[(g)] General Hospital Linz
Central Laboratory
Krankenhausstraße 9, 4021 Linz, Austria

[(a)] michael.affenzeller@fh-hagenberg.at, [(b)] stephan.winkler@fh-hagenberg.at, [(c)] stefan.forstenlechner@fh-hagenberg.at
[(d)] gabriel.kronberger@fh-hagenberg.at, [(e)] michael.kommenda@fh-hagenberg.at, [(f)]stefan.wagner@fh-hagenberg.at,
[(g)] herbert.stekel@akh.linz.at

ABSTRACT
In this paper we describe the integration of ensemble modeling into genetic programming based classification and discuss concepts how to use genetic programming specific features for achieving new confidence indicators that estimate the trustworthiness of predictions. These new concepts are tested on a real world dataset from the field of medical diagnosis for cancer prediction where the trustworthiness of modeling results is of highest importance.

Keywords: data mining, genetic programming, ensemble modeling, medical data analysis

1. INTRODUCTION, RESEARCH GOALS

Genetic Programming (GP) plays an outstanding role among the data-mining techniques from the field of machine learning and computational intelligence: Due to its model representation, GP is able to produce human interpretable models without any assumptions about the nature of the analyzed relationships. Furthermore, GP-based data analysis has been shown to have good generalization; GP is also able to simultaneously evolve the structure and the parameters of a model with implicit feature selection. Due to the combination of these aspects GP is considered a very powerful and also robust method for various data analysis tasks.

Apart from these general aspects of genetic programming based modeling, the hybridization of symbolic regression and ensemble modeling is still rarely considered. A good overview article about genetic programing and ensemble modeling has been presented by Keijzer and Babovic (Keijzer, Babovic 2000).

In this paper we introduce new methods for the generation and interpretation of model ensembles consisting of symbolic regression / classification models; concretely, we introduce new confidence measures for assessing the trustworthiness of genetic programming ensemble models. Furthermore, we discuss and interpret symbolic classification ensemble modeling results achieved for medical data mining tasks.

The experimental part of the paper discusses the value of trust of correct classifications opposed to the value of trust of incorrect classifications, and analyzes if the confidence of the correctly classified instances is significantly higher than that of the incorrectly classified samples.

2. THEORECICAL FOUNDATIONS

2.1. Genetic Programming Based Symbolic Classification

Symbolic classification with genetic programming uses the general concept of a genetic algorithm in order to search the space of hypotheses which are represented as mathematical formulae in structure tree representation.
We have also applied a classification algorithm based on genetic programming (GP, Koza (1992)) using a structure identification framework described in Winkler (2008) and Affenzeller et al. (2009).

We have used genetic programming with gender specific parents selection (Wagner 2005) (combining random and roulette selection) as well as strict offspring

selection (Affenzeller et al. 2009) (OS, with success ratio as well as comparison factor set to 1.0) as this variant has proven to be especially suited for symbolic classification. The function set described in (Winkler 2008) (including arithmetic as well as logical ones) was used for building composite function expressions.

In addition to splitting the given data into training and test data, the GP based training algorithm used in our research project has been designed in such a way that a part of the given training data is not used for training models and serves as validation set; in the end, when it comes to returning classifiers, the algorithm returns those models that perform best on validation data.

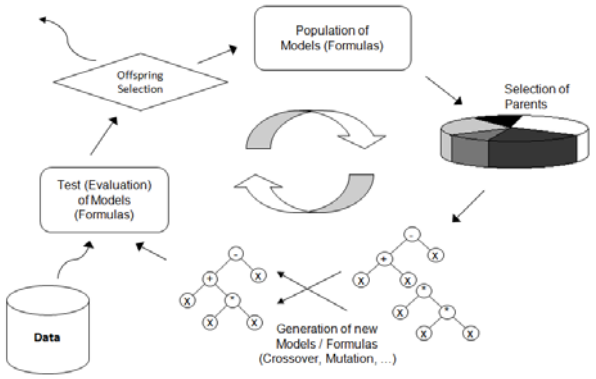

Figure 1: Genetic programming including offspring selection.

The used fitness function is the mean squared error of the training samples using a threshold value that gives optimal accuracy on the training data set. The distance between a certain prediction and the threshold will be used in our ensemble interpretation to achieve additional estimates about the trustworthiness of the ensemble prediction.

2.2. Ensemble Modeling

Ensemble modeling techniques may be used as an extension to various regression and classification techniques. Supported by the increasing availability of hardware resources and parallel computing infrastructures, the general concept of ensemble modeling is used in order to increase the reliability and robustness of predictive models.

The general idea is to apply not only a single hypothesis for predictive modeling but a large number of stochastically independent models. The result of an ensemble predictor is typically given by the average or mean value of the ensemble results for regression modeling or by some kind of majority voting for classification tasks. The reliability of an ensemble prediction may be indicated by the variance of the ensemble results for regression or by the clearness of the majority voting in classification applications (Keijzer, Babovic 2000).

Basically the idea of ensemble modeling can be coupled with any stochastic hypothesis search technique. A well-known example is given by the concept of random forests (Breiman, 2001) where ensemble modeling is integrated into the general concept of tree learning strategies.

In this paper we discuss the integration of ensemble modeling into the concept of genetic programming based symbolic classification under the special focus of developing and analyzing extended confidence estimates of the ensemble modeling results.

3. NEW CONFIDENCE MEASURES OF SYMBOLIC CLASSIFICATION ENSEMBLE MODELS

The idea is to combine the degree of majority of a majority voting with the average clearness of the ensemble predictors for a certain sample.

On the one hand, a certain prediction will be considered the more reliable the clearer the result of the majority voting is. If we evaluate, for example, 99 ensemble models we will have more trust in a 91 to 8 majority than in a 50 to 49 majority. This kind of interpretation may be applied to any ensemble model as it does not use genetic programing specific aspects.

Assuming a two-class classification problem the corresponding first confidence measure cm_1 is defined as

$$cm_1 := 2\left(\frac{|votes(winning\ class)|}{|votes|} - 0{,}5\right) \in [0,1]$$

The fraction of votes for the winner class in relation to the total number of ensemble votes $((|votes(winning\ class)|)/(|votes|))$ will range between 0.5 and 1 for the majority vote. In order to normalize this confidence measure between 0 and 1 we have adapted the formula accordingly.

On the other hand we will discuss a second indicator of trustfulness which is unique to genetic programming which optimizes the mean squared error of residuals in the training data set. Supposing a two-class classification problem with the classes 0 and 1 we will rather trust an ensemble predictor where most of the prognoses are clustered around 0 compared to a predictor where the ensemble results will be around 0.45. The according second confidence measure cm_2 is therefore designed to consider both, the majority of the voting as well as the uniqueness of the certain ensemble predictors for a certain sample. Concretely cm_2 is defined as

$$cm_2 = \min\left(\frac{\Delta(m(t),m(e))}{\Delta(m(t),class)}, 1\right) \in [0,1] \qquad (3)$$

where Δ denotes the difference between the mean value of the ensemble thresholds and the median or mean value of the ensemble predictions in the nominator of the formula. In the denominator of the formula the distance Δ between the median or mean value of the thresholds and the class value predicted by the ensemble is represented (which is basically the range for the predictors). The confidence cm_2 will therefore on the

one hand be rather low ($cm_2 \approx 0$) if the majority of ensemble predictors are near the threshold; on the other hand, if the majority of ensemble votes is around the actual class this confidence will rather high ($cm_2 \approx 1$). By considering some kind of average value (median or mean), confidence measure cm_2 is expected to implicitly consider both, the majority of the voting result and the uniqueness of the single ensemble predictors.

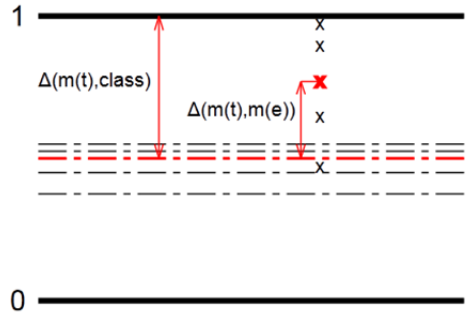

Figure 1: Illustration of different GP ensemble estimators and their corresponding thresholds.

As illustrated in Fig. 1 for the second confidence measure cm_2 the difference between the mean ensemble predictor m(e) to the mean threshold m(t) is considered with respect to the difference between the mean threshold m(t) and the corresponding class.

The effectiveness of the both confidence measures cm_1 and cm_2 will be analyzed by discussing the average confidence of correct predictions compared to the average confidence of incorrect predictions. Of course, confidence is expected to be higher for correct predictions and the delta between confidence in correct predictions and incorrect predictions may be used in order to interpret the suitability of the concrete confidence measure for a certain application. Due to its usage of additional information cm_2 is expected to be more suited than cm_1 in general for binary classification problems.

4. THE MEDICAL DATASET

In this section we describe results achieved within the Josef Ressel Centre for Heuristic Optimization *Heureka!*: Data of thousands of patients of the General Hospital (AKH) Linz, Austria, have been preprocessed and analyzed in order to identify mathematical models for cancer diagnoses. We have used a medical database compiled at the central laboratory of AKH in the years 2005 – 2008: 28 routinely measured blood values of thousands of patients are available as well as several tumor markers (TMs, substances found in humans that can be used as indicators for certain types of cancer). Not all values are measured for all patients, especially tumor marker values are determined and documented only if there are indications for the presence of cancer. The results of empirical research work done on the data based identification of estimation models for standard blood parameters as well as tumor markers. The main goal is to generate mathematical models for cancer prediction based on blood parameters.

The blood data measured at the AKH in the years 2005-2008 have been compiled in a database storing each set of measurements (belonging to one patient): Each sample in this database contains an unique ID number of the respective patient, the date of the measurement series, the ID number of the measurement, standard blood parameters, tumor marker values, and cancer diagnosis information. Patients' personal data were at no time available for the authors except for the head of the laboratory.

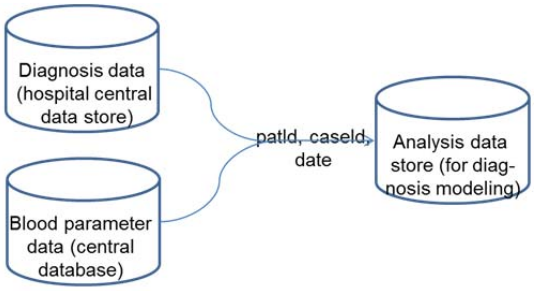

Figure 2: Integration of the relevant data from different hospital databases

In total, information about 20,819 patients is stored in 48,580 samples. Please note that of course not all values are available in all samples; there are many missing values simply because not all blood values are measured during each examination. Further details about the data set and necessary data preprocessing steps can for example be found in Winkler et al. (2010) and Winkler et al. (2011), e.g.

Standard blood parameters include for example the patients' sex and age, information about the amount of cholesterol and iron found in the blood, the amount of hemoglobin, and the amount of red and white blood cells; in total, 29 routinely available patient parameters are available.

An important aspect in the data preprocessing stage was to identification of relevant blood parameters for a positive cancer diagnosis. In order to identify characteristic blood parameters for a positive cancer prediction we have collected those blood parameters which have been measured in a time window of two weeks before the first cancer diagnosis based in ICD 10 classification of diseases. Blood parameters taken earlier may not be characteristic as the patient may still have been healthy at that time and measurements taken after the diagnoses may be diluted due to medical treatment. The blood parameters for the class of healthy persons have been taken only from patients with no positive cancer diagnosis in their case history.

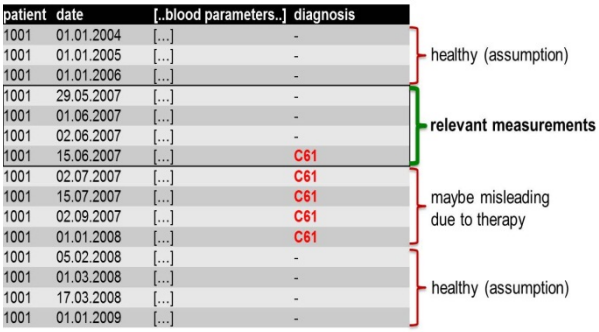

Figure 3: Identification of appropriate periods of relevant measurements for cancer diagnosis

5. GENETIC PROGRAMMING

For generating the ensemble models we have used gender specific strict offspring selection genetic programming which is available in the HeuristicLab 3.3.6 framework http://dev.heuristiclab.com).

We have used the following parameter settings for our GP test series: The mutation rate was set to 20%, gender specific parents selection (Wagner 2005) (combining random and roulette selection) was applied as well as strict offspring selection (Affenzeller et al. 2009) (OS, with success ratio as well as comparison factor set to 1.0). The functions set described in (Winkler 2008) (including arithmetic as well as logical ones) has been used for building composite function expressions.

In addition to splitting the given data into training and test data, the GP based training algorithm used in our research project has been designed in such a way that a part of the given training data is not used for training models and serves as validation set; in the end, when it comes to returning classifiers, the algorithm returns those models that perform best on validation data.

6. RESULTS

For the results presented in this section a set of GP-based models of different model sizes have been generated for the prediction of breast cancer based solely on blood parameters and also based on blood parameters together with its corresponding tumor markers (C125, C15-3, CEA). The main GP parameters have been adjusted according to the settings shown in Table 1.

Algorithm	Offspring Selection Genetic Programming
Runs	100 per tree size (with three different tree sizes)
TreeCreator	Probabilistic Tree Creator
Symbols	+, -, *, /, sin, cos, tan, exp, log, IfThenElse, <, >, and, or, not
Fitness function	MSE (mean squared error)
Selector	GenederSpecific (Random, Proportional)
Mutator	ChangeNodeTypeManipulation, FullTreeShaker, OnePointTreeShaker, ReplaceBranchManipulation
Crossover	SubtreeCrossover
Elites	1
Population Size	700
Mutation Rate	20%
Maximal Generations	1000
Maximal Selection Pressure	100
Cross Validation Folds	5
Tree size (Length/Depth)	20/7, 35/8, 50/10

Table 1: Algorithmic settings

Overall 300 (100 per tree size) independent runs have been performed for predicting breast cancer and the 75 best models concerning their training quality have been used for ensemble modeling as well as for standard evaluations as shown in Table 2. Table 2 shows the average training and test qualities of the best 75 training models (the same ones which are used also for the ensemble models and confidence analyses). Moreover the training and test accuracies of the overall best training model are shown in Table 2.

Tables 3 and 4 show the results in ensemble interpretation for modeling breast cancer with standard blood parameters alone (Table 3) as well as together with tumor markers as additional input features. Together with the accuracies the confidence measures described in section 3 are shown for the correctly as well as for the incorrectly classified instances.

There are several conclusions that can be drawn from the following result tables:

- The results – especially the test results – of the ensemble predictions (Tables 3 and 4) are better than the results achieved with conventional GP models.
- The confidence in the predictions measured by the two confidence measures cm_1 and cm_2 is significantly higher for the correctly classified instances than in those of the incorrectly classified instances. The delta is even higher if tumor markers can be used for the models.
- The distribution of the two confidence values for the correctly and incorrectly classified instances with cm_1 (left boxplots) and cm_2 (right boxplots) indicate high potential for future research based on these confidence measures: By introducing a confidence threshold below which the results are categorized as untrustworthy, the trusted results are expected to have significantly higher quality.
- Contrary to the theoretical considerations of Section 3 the statistical properties of the more complex cm_2 do not seem to lead to more

significant differentiation between the correctly and incorrectly classified instances.

	Avg. training accuracy of 100 best models	Avg. test accuracy of 100 best models	Training accuracy of best model	Test accuracy of best training model
Breast with TM	83.17%	77.89%	84.74%	79.33%
Breast without TM	76.79%	72.61%	78.15%	73.08%

Table 2: Results with standard GP-based predictive modeling (without ensemble modeling)

	Majority Vote	Average Threshold
Accuracy training	78.47%	78.47%
Accuracy test	76.49%	76.49%
Average Confidence Correct Classified	cm_1 = 0.8202	cm_2 = 0.4835
Average Confidence Incorrect Classified	cm_1 =0.5543	cm_2 = 0.2525
Confidence Delta	0.2659	0.2310

Table 3: Breast cancer without tumor marker

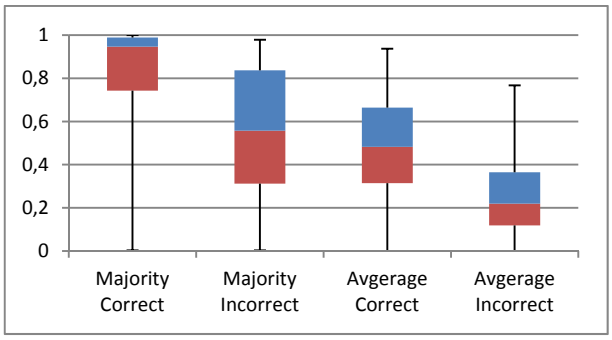

Figure 4: Boxplot breast cancer without tumor marker

	Majority Vote	Average Threshold
Accuracy training	84.99%	84.99%
Accuracy test	81.44%	81.44%
Average Confidence Correctly Classified	cm_1 = 0.8500	cm_2 = 0.6182
Average Confidence Incorrectly Classified	cm_1 = 0.4806	cm_2 = 0.2449
Confidence Delta	0.3694	0.3733

Table 4: Breast cancer with tumor marker

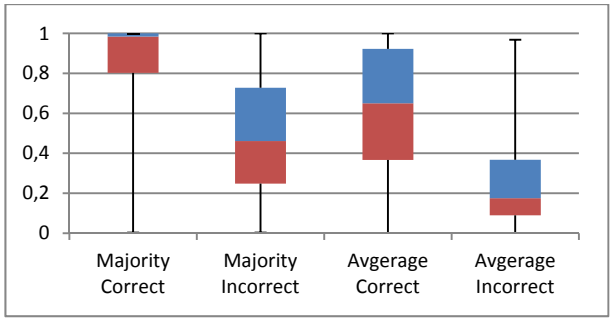

Figure 5: Boxplot breast cancer with tumor marker

7. CONCLUSION

In this paper two confidence measures have been discussed for predictive models generated by GP-based ensemble modeling. The results for a breast cancer data set have shown that the confidence for the correctly classified instances is significantly higher than the confidence for incorrectly classified instances which is considered as the main result of this contribution and which opens new areas of application especially for predictive modeling in the medical field where the confidence in a prediction is very important.

Obvious aspects for future investigations in this field are analyses for further cancer types as well as the introduction of a new class for the uncertain prediction (below a certain confidence threshold) and corresponding analyses of the statistical properties of the so achieved predictions.

ACKNOWLEDGMENTS

The work described in this chapter was done within the Josef Ressel-Centre *Heureka!* for Heuristic Optimization sponsored by the Austrian Research Promotion Agency (FFG).

REFERENCES

Affenzeller, M., Winkler, S., Wagner, S., A. Beham, 2009. *Genetic Algorithms and Genetic Programming - Modern Concepts and Practical Applications*. Chapman & Hall/CRC. ISBN 978-1584886297. 2009.

Breiman, L. 2001. Random Forests. *Machine Learning 45*, pp. 5 – 32.

Keijzer, M., Babovic, V., 2000. Genetic Programming, Ensemble Methods and the Bias/Variance Tradeoff – Introductory Investigations. *Lecture Notes in Computer Science, 2000, Volume 1802/2000*, pp. 76-90.

Wagner, S., 2009. *Heuristic Optimization Software Systems - Modeling of Heuristic Optimization Algorithms in the HeuristicLab Software Environment*. PhD Thesis, Institute for Formal Models and Verification, Johannes Kepler University Linz, Austria.

Winkler, S., Affenzeller, M., Jacak, W., Stekel, H., 2010. Classification of Tumor Marker Values Using Heuristic Data Mining Methods. *Proceedings of Genetic and Evolutionary Computation Conference 2010, Workshop on*

Medical Applications of Genetic and Evolutionary Computation, pp. 1915–1922.

Winkler, S., Affenzeller, M., Jacak, W., Stekel, H., 2011. Identification of Cancer Diagnosis Estimation Models Using Evolutionary Algorithms – A Case Study for Breast Cancer, Melanoma, and Cancer in the Respiratory System. *Proceedings of Genetic and Evolutionary Computation Conference 2011, Workshop on Medical Applications of Genetic and Evolutionary Computation.*

AUTHORS BIOGRAPHIES

MICHAEL AFFENZELLER has published several papers, journal articles and books dealing with theoretical and practical aspects of evolutionary computation, genetic algorithms, and meta-heuristics in general. In 2001 he received his PhD in engineering sciences and in 2004 he received his habilitation in applied systems engineering, both from the Johannes Kepler University of Linz, Austria. Michael Affenzeller is professor at UAS, Campus Hagenberg, and head of the Josef Ressel Center *Heureka!* at Hagenberg.

STEPHAN M. WINKLER received his PhD in engineering sciences in 2008 from Johannes Kepler University (JKU) Linz, Austria. His research interests include genetic programming, nonlinear model identification and machine learning. Since 2009, Dr. Winkler is professor at the Department for Medical and Bioinformatics at the University of Applied Sciences (UAS) Upper Austria at Hagenberg Campus; since 2010, Dr. Winkler is head of the BioinformaticsResearch Group at UAS, Hagenberg.

Stefan Forstenlechner received his BSc in software engineering in 2011 from the Upper Austria University of Applied Sciences, Campus Hagenberg. He is currently pursuing studies for his master's degree and working at UAS Research Center Hagenberg within *Heureka!*.

GABRIEL KRONBERGER received his PhD in engineering sciences in 2010 from JKU Linz, Austria, and is a research associate at the UAS Research Center Hagenberg. His research interests include genetic programming, machine learning, and data mining and knowledge discovery.

MICHAEL KOMMENDA finished his studies in bioinformatics at Upper Austria University of Applied Sciences in 2007. Currently he is a research associate at the UAS Research Center Hagenberg working on data-based modeling algorithms for complex systems within *Heureka!*.

STEFAN WAGNER received his PhD in engineering sciences in 2009 from JKU Linz, Austria; he is professor at the Upper Austrian University of Applied Sciences (Campus Hagenberg). Dr. Wagner's research interests include evolutionary computation and heuristic optimization, theory and application of genetic algorithms, and software development.

WITOLD JACAK received his PhD in electric engineering in 1977 from the Technical University Wroclaw, Poland, where he was appointed Professor for Intelligent Systems in 1990. Since 1994 Prof. Jacak is head of the Department for Software Engineering at the Upper Austrian University of Applied Sciences (Campus Hagenberg) where he currently also serves as Dean of the School of Informatics, Communications and Media.

HERBERT STEKEL received his MD from the University of Vienna in 1985. Since 1997 Dr. Stekel is chief physician at the General Hospital Linz, Austria, where Dr. Stekel serves as head of the central laboratory.