Abstract—The main goal of this paper is introduction of fast heterogeneous boosting algorithm. ‘Heterogeneous’ means that boosting is based not on single-type learning machine, but may use machines of several types coherently. The main idea behind the construction of heterogeneous boostings was to use it with learning machines of low complexity ($O(nd)$). Thanks to that, the heterogeneous boosting is still a fast algorithm of linear learning (and usage) complexity.

The paper presents a comparison of homogeneous boostings of a few types of fast learning machines with introduced heterogeneous boosting, which base on a small group of fast learning machines. The presented comparison proves that heterogeneous boosting is efficient and accurate.

Index Terms—Computational intelligence, adaptive boosting, classifier, machine learning

I. INTRODUCTION

Learning machine ensembles (or committees) play an important role in the construction of trustful classifiers. In most of the cases, the usage of ensembles leads to construction of more stable models1 with higher accuracy [1], [2], [3], [4], [5], [6], [7].

There are several types of ensembles. The main source of the difference between ensembles is the strategy of using submachines. Some of the committees combine submodels (like weighting committees) while other select one submodel (like voting committees). The second discriminative feature of committees is whether they base on a fixed set of submodels or they are able to generate new ones. To see more about different ensemble methods see [8]. In almost all cases, the final decision of a committee can be described by:

$$ M(x) = \sum_{i=1}^{T} w_i M_i(x) $$ (1)

where $w_i$ are weights (in case of voting only one weight is equal to 1 and the rest are equal to 0), and the function $M_i(x)$ is the decision taken by $i$-th submachine. $T$ defines the number of base machines (the committee members). One of very popular techniques of constructing committees is to construct members of the same type but based on different distributions of the same, original learning data. One of the foremost concepts of such kind was presented in Breiman’s Bagging in [9]. Another, and the most popular kind of such committee is Adaboosting [1], [10], [11], [12] (shortly Boosting). The main difference between those two methods is that in bagging, the data for learning of submachines is drawn independently, while in boosting each data distribution depends directly on the previous steps. The base machines are learned sequentially in boosting. After each learning of the base machine, a classification test is performed. And the construction of the new learning data depends directly on performance of the previous submachine and the previous data distribution. In time, boostings were modified in several ways or adopted to specific task types (see http://www.boosting.org/publications for more).

If boosting has to be used with many submachines then (almost) only the weak classifiers should be considered as submachines. In other case the learning (and using) becomes too complex. Note that in case of weak classifiers the complexity may be as small as $O(nd)$ (where $n$ is the number of instances and $d$ is the number of features of learning data) while methods like the support vector machine ([13]) are at least $O(n^2d)$.

The differences between the data distribution for each $M_i(\cdot)$ submachines are necessary because in other case each of the submachines would be of the same type, of the same configuration and learned on the same data. It would mean that all submachines would be equivalent to each other and, in consequence, redundant. The changing distribution is the source of diversity in boosting and bagging ensembles. It is known that this diversity is the source of success in learning of committees [14], [15], [16], [17], [18], [19]. However, the diversity may be obtained in a natural way by using heterogeneous members of committee. Such investigations were presented in [20], [21]. Diversity in such case is a derivative of the difference in behaviour of heterogeneous members of committee.

This is the reason to present analysis of construction of heterogeneous boosting, which will base on two sources of diversity: the data distribution and the heterogeneous set of base learning machines.

II. HETEROGENEOUS BOOSTING

Let’s assume learning data set consists of a sequence of pairs: $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$.

Adaboosting during each iteration of learning repeats sequentially below steps:

1) Create learning data $D_i$ for $i$-th classifier according to distribution $D_i$.
2) Learn $L_i$ on $D_i$.
3) Test $L_i$ on original learning data $D$ and calculate misclassification for each instance ($miss_i(x_j)$).
4) Calculate boosting error $\epsilon_i$.
5) Estimate a new distribution $D_{i+1}$ if necessary.
The initial data distribution $D_1$ is set according to:

$$D_1(x_j) = \frac{1}{n}. \quad (2)$$

Further iteration of boosting uses the adjusted distribution:

$$D_{i+1}(x_j) = D_i(x_j) \cdot \left\{ \begin{array}{ll} \frac{1}{2} & \text{if } miss_i(x_j) = 1 \\ \frac{1}{2} & \text{otherwise} \end{array} \right., \quad (3)$$

where $miss_i(x_j) = 1$ if $i$-th classifier fails on the instance $x_j$ and is 0 otherwise. The $\epsilon_i$ is defined by

$$\epsilon_i = \sum_{j=1}^{n} D_i(x_j) \cdot miss_i(x_j). \quad (4)$$

The decision module of boosting classifies given instance $x$ according to rule:

$$M(x) = \max_{c=1,\ldots,C} \sum_{i|M_i(x) = c} \log \frac{1 - \epsilon_i}{\epsilon_i}, \quad (5)$$

where $M_i(x)$ is the decision of $i$-th submachine, and $C$ is the number of classes.

The above boosting bases on machines $M_i$ of the same type. Now this algorithm will be extended in a simple way to use several types of machines. Instead of learning a single base machine $M_i$ on the new data distribution $D_i$ at the $i$-th iteration of boosting, the series of machines $M_i^k$ ($k = 1, \ldots, K$) will be started to learn. After the learning of each machine in the series is finished, the tests are computed:

$$\epsilon_i^k = \sum_{j=1}^{n} D_i(x_j) \cdot miss_i(x_j, M_i^k) \quad (6)$$

independently on each machine in the series.

Now the best machine is chosen via:

$$M_i^* = \arg \min_{M_i^k \in \{M_i^k : k = 1, \ldots, K\}} \epsilon_i^k. \quad (7)$$

And all machines $M_i^1, \ldots, M_i^K$ except $M_i^*$ are removed. In the end the boosting is composed from series of machines: $M_1^1, \ldots, M_m^m$ ($m$ is the number of boosting iterations). It means that at each iteration according to changes in the distribution $D_i$ the most adequate machine is extracted. In the same way as the $M_i^*$ where chosen, $\epsilon_i^*$ are selected:

$$\epsilon_i^* = \min_{\{k = 1, \ldots, K\}} \epsilon_i^k. \quad (8)$$

And finally boosting decision module uses $\epsilon_i^*$ instead of $\epsilon_i$ in \((5)\).

The types of base machines used to learn heterogeneous boosting should exhibit natural diversity and additionally it is highly recommended to use weak learners to keep the overall complexity small. Note that the complexity of heterogeneous boosting used with a few base machines is still smaller than complexity of boosting of SVM \([13]\) or even than complexity of the SVM itself. The only assumption is that the number of boosting iterations is significantly smaller than $O(n)$—in experiments presented in further part a fixed number of iteration was used.

Homogeneity of the type of the base learner is indeed the biggest disadvantage of boosting at all. It is known that if a given kind of a weak learner (and not only weak learner) does not provide a sufficiently good generalization for a given data set, the boosting with such machine does not guarantee that the generalization will be much higher (however it is expected). Heterogeneity is characterized by natural diversity, while the diversity is the main source of success in ensemble composition (no diversity is equivalent with usage of single machine without any ensemble). As it will be seen in further parts, even a few types of weak learners brings interesting results.

### III. Experiment design

**a) Data sets:** All experimental results were collected on over 40 benchmarks from the UCI Machine learning repository \([22]\). Base properties of data sets can be compared in Table I. Each dataset was first standardized (except for symbolic features, those features remain unchanged). All tests were computed using 10 times repeated 10-fold stratified cross-validation test. Information about accuracies was computed only from test parts.

**b) Base learners and boosting configurations:** To prepare valuable analysis of several boosting configurations the following base learners were chosen: Naive Bayes (NB) \([23]\), Learning Vector Quantization (LVQ) \([24]\), Maximum Likelihood Classifier (MLC) \([25]\) and Decision Stump Tree (DS) \([26]\). All methods except for DS are linear time learning classifiers $O(nd)$ ($d$ is the number of data attributes). Decision stump has $O(dn \log n)$ complexity (and $O(nd)$ if used only with discrete attributes).

LVQ is used with the number of neurons equal to the number of classes (to keep it very simple and non-complex). The number of epoch is constant, equal to 100. Maximum Likelihood Classifier bases on the assumption of Gaussian distributions and calculations of maximum likelihood. First, dispersion $\sigma^k_i$ for each dimension $i = 1, \ldots, d$ and class $k = 1, \ldots, C$ is calculated, and then the class probability is estimated according to:

$$P(x|C_k) = \frac{G(x|C_k)}{\sum_{c=1}^{C} G(x|C_c)} \quad (9)$$

where

$$G(x|C_k) = \prod_{i=1}^{d} g \left( x_i; r^k_i, \sigma^k_i \right) \quad (10)$$

and

$$g(x_i; r^k_i, \sigma^k_i) = \exp \left[ \frac{- (r^k_i - x_i)^2}{2(\sigma^k_i)^2} \right] \quad (11)$$

$g(\cdot)$ is a Gaussian function, $C_k$ is the class for which the classification probability of vector $x$ is estimated, and $r^k_i$ is the $i$-th coordinate of the $K$-th center of $k$-th class.

Performance of the heterogeneous boosting will be compared to boosting with each of the weak learners. The heterogeneous boosting was configured to use only the (very) fast weak learners: Naive Bayes, LVQ and MLC. And this
is the reason for the title of the article to contain the word fast (because it is dedicated for fast submachines) in contrary to use 10 or more weak learners. What’s more, the results presented in the next section confirm that chosen three weak learners are enough to keep the heterogeneous boosting at the top of performance ranks.

To make the tests trustful, boostings were tested with different numbers of base classifiers. Heterogenous and homogeneous boosting configurations were tested with $2^0$, $2^1$, \ldots, $2^8$ submachines. The case $2^0$ is equivalent with direct usage of the base learner.

Table I presents information about the benchmarks used in tests.

<table>
<thead>
<tr>
<th>Data set</th>
<th># classes</th>
<th># inst.</th>
<th># feat.</th>
<th># ord. f.</th>
</tr>
</thead>
<tbody>
<tr>
<td>arrhythmia</td>
<td>11</td>
<td>63</td>
<td>279</td>
<td>206</td>
</tr>
<tr>
<td>autos</td>
<td>6</td>
<td>159</td>
<td>25</td>
<td>15</td>
</tr>
<tr>
<td>balance-scale</td>
<td>3</td>
<td>625</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>blood-transfusion-service-center</td>
<td>2</td>
<td>748</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>breast-cancer-wisconsin-diagnostic</td>
<td>2</td>
<td>569</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>breast-cancer-wisconsin-original</td>
<td>2</td>
<td>683</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>breast-tissue</td>
<td>6</td>
<td>106</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>car-evaluation</td>
<td>4</td>
<td>1728</td>
<td>21</td>
<td>21</td>
</tr>
<tr>
<td>cardiotocography-1</td>
<td>10</td>
<td>2126</td>
<td>21</td>
<td>21</td>
</tr>
<tr>
<td>cardiotocography-2</td>
<td>3</td>
<td>2126</td>
<td>21</td>
<td>21</td>
</tr>
<tr>
<td>chess-king-rook-vs-king-pawn</td>
<td>2</td>
<td>3196</td>
<td>38</td>
<td>38</td>
</tr>
<tr>
<td>cme</td>
<td>3</td>
<td>1473</td>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>connectionist-bench-sonar-mines-vs-rocks</td>
<td>2</td>
<td>208</td>
<td>60</td>
<td>60</td>
</tr>
<tr>
<td>connectionist-bench-vowel-recognition-detector</td>
<td>11</td>
<td>528</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>cylinder-bands</td>
<td>2</td>
<td>277</td>
<td>39</td>
<td>18</td>
</tr>
<tr>
<td>dermatology</td>
<td>6</td>
<td>355</td>
<td>64</td>
<td>1</td>
</tr>
<tr>
<td>ecoli</td>
<td>8</td>
<td>336</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>glass</td>
<td>6</td>
<td>214</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>habermans-survival</td>
<td>2</td>
<td>306</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>hepatitis</td>
<td>2</td>
<td>80</td>
<td>19</td>
<td>6</td>
</tr>
<tr>
<td>ionosphere</td>
<td>2</td>
<td>351</td>
<td>34</td>
<td>34</td>
</tr>
<tr>
<td>iris</td>
<td>3</td>
<td>150</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>libras-movement</td>
<td>15</td>
<td>360</td>
<td>90</td>
<td>90</td>
</tr>
<tr>
<td>liver-disorders</td>
<td>2</td>
<td>345</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>lymph</td>
<td>4</td>
<td>148</td>
<td>18</td>
<td>3</td>
</tr>
<tr>
<td>monks-problems-1</td>
<td>2</td>
<td>556</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>monks-problems-2</td>
<td>2</td>
<td>601</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>monks-problems-3</td>
<td>2</td>
<td>554</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>parkinsons</td>
<td>2</td>
<td>195</td>
<td>22</td>
<td>22</td>
</tr>
<tr>
<td>pima-indians-diabetes</td>
<td>2</td>
<td>768</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>sonar</td>
<td>2</td>
<td>208</td>
<td>60</td>
<td>60</td>
</tr>
<tr>
<td>spambase</td>
<td>2</td>
<td>4601</td>
<td>57</td>
<td>57</td>
</tr>
<tr>
<td>spec-heart</td>
<td>2</td>
<td>267</td>
<td>22</td>
<td>22</td>
</tr>
<tr>
<td>specif-heart</td>
<td>2</td>
<td>267</td>
<td>44</td>
<td>44</td>
</tr>
<tr>
<td>statlog-australian-credit</td>
<td>2</td>
<td>690</td>
<td>38</td>
<td>38</td>
</tr>
<tr>
<td>statlog-german-credit</td>
<td>2</td>
<td>1000</td>
<td>24</td>
<td>24</td>
</tr>
<tr>
<td>statlog-heart</td>
<td>2</td>
<td>270</td>
<td>20</td>
<td>7</td>
</tr>
<tr>
<td>statlog-vehicle-silhouettes</td>
<td>4</td>
<td>846</td>
<td>18</td>
<td>18</td>
</tr>
<tr>
<td>teaching-assistant-evaluation</td>
<td>3</td>
<td>151</td>
<td>54</td>
<td>54</td>
</tr>
<tr>
<td>thyroid-disease</td>
<td>3</td>
<td>7200</td>
<td>21</td>
<td>21</td>
</tr>
<tr>
<td>vote</td>
<td>2</td>
<td>232</td>
<td>16</td>
<td>0</td>
</tr>
<tr>
<td>wine</td>
<td>3</td>
<td>178</td>
<td>13</td>
<td>13</td>
</tr>
</tbody>
</table>

is the reason for the title of the article to contain the word fast (because it is dedicated for fast submachines) in contrary to use 10 or more weak learners. What’s more, the results presented in the next section confirm that chosen three weak learners are enough to keep the heterogeneous boosting at the top of performance ranks.

To make the tests trustful, boostings were tested with different numbers of base classifiers. Heterogenous and homogeneous boosting configurations were tested with $2^0$, $2^1$, \ldots, $2^8$ submachines. The case $2^0$ is equivalent with direct usage of the base learner.

IV. EXPERIMENT RESULTS AND ANALYSIS

Figures [1][4] present the application of the above-described boostings’ configurations to several data benchmarks. Each subfigure is devoted to the presentation of computations on a single benchmark and displays averaged accuracies for mentioned boosting configurations with changing number of submachines. The OX axis (labeled by $T$) is the log$_2$ of number of submachines in boosting, and OY axis presents the accuracy of given configuration. Heterogenous boosting is marked as ‘hetero’ and boosting with a single base learner are marked adequately to the short name of base learners: ‘NB’, ‘LVQ’, ‘MLC’, ‘DS’ (see the previous section for details).

The most characteristic feature, which can be observed in Figures [1][4] is that the heterogeneous boosting is almost always at the top (the top means best performance), while the performance of homogeneous boostings change from benchmark to benchmark. And no one type of a weak learner is enough to keep homogeneous boosting so frequently at the top as the heterogeneous boosting can. In very rare cases the performance of all configurations of homogeneous boostings are similar. What proves that it matters what weak learner is used in a boosting ensemble.

In several cases at the top can be found heterogeneous boosting together with boosting of Naive Bayes or decision stump—see for example following benchmarks ‘arythmia’, ‘autos’, ‘cardiotocography-1’ or ‘cardiotocography-2’, etc. However it happens that heterogeneous boosting and boosting of MLC can be found at the top, like, for example in case of ‘car-evaluation’, ‘statlog-vehicle-silhouettes’ or ‘glass’. And in several cases at the top there are three or more configurations of boosting.

The averaged accuracy of boostings used with one of the weak learners may be significantly bigger than with others, see for example ‘dermatology’ (see the difference between ‘NB’ and rest), ‘glass’ (see the difference between ‘MLC’ and ‘DS’), ‘ionosphere’, ‘libras-movement’, ‘parkinsons’, ‘statlog-vehicle-silhouettes’. Some significant losses of one boosting configuration over the other configurations can also be found. Compare results for ‘chess-king-rook-vs-king-pawn’, ‘cardiotocography-1’, ‘ionosphere’, ‘libras-movement’, ‘statlog-heart’, ‘thyroid-disease’ and ‘vote’.

Another important observation is that for most of the benchmarks the $2^5$ submachines are enough to obtain good results for most boosting configurations.

Basing on accuracies for each boosting configuration, for a given benchmark and given number of submachines the ranking of boosting configurations can be constructed. Using such ranks, the average ranking over all benchmarks can be computed. Table [I][4] presents averaged ranking for different number of submachines in boosting and for different configurations. It is very clear that even when averaging over more than 40 benchmarks the best average rank always belongs to the heterogeneous boosting—with average rank from 1.52 to 1.88. The second position belongs to the boosting with Naive Bayes with average ranking between 2.64 and 2.86.
is quite far from average ranks of heterogeneous boosting. The distance between the second and the third position is smaller than between the first and the second (for all sizes of boosting).

V. CONCLUSIONS

Basing on the results presented above it was proved that the presented heterogeneous boosting is more stable than homogeneous boostings. What’s more heterogeneous boosting gained the best average rank, regardless of the size of boosting (the total number of weak learners used in boosting). Proposed heterogeneous boosting is not difficult in contruction—rules used for the learning of heterogeneous boosting are relatively simple while results become significantly better. A very important feature of the presented method is that heterogeneous boosting was not used with sophisticated learners of high complexity but with simple ones. This results in the same complexity of learning as the complexity of homogeneous boosting (also with weak learners). This means that learning of heterogeneous boosting is linear time $O(nd)$ with respect to the size of data (assuming fixed number of boosting iterations as it was used in all experiments). However, before using homo- or heterogeneous boosting, the performance of simpler algorithms (that of smaller complexity of learning algorithm) should be validated—at least boosted and non-boosted weak learners should be validated.

REFERENCES


<table>
<thead>
<tr>
<th>T</th>
<th>hetero</th>
<th>NB</th>
<th>LVQ</th>
<th>MLC</th>
<th>DS</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1.71</td>
<td>2.71</td>
<td>3.29</td>
<td>3.07</td>
<td>3.57</td>
</tr>
<tr>
<td>1</td>
<td>1.81</td>
<td>2.67</td>
<td>3.45</td>
<td>3.12</td>
<td>3.67</td>
</tr>
<tr>
<td>2</td>
<td>1.67</td>
<td>2.64</td>
<td>3.5</td>
<td>3.14</td>
<td>3.55</td>
</tr>
<tr>
<td>3</td>
<td>1.52</td>
<td>2.74</td>
<td>3.71</td>
<td>3.36</td>
<td>3.19</td>
</tr>
<tr>
<td>4</td>
<td>1.67</td>
<td>2.71</td>
<td>3.6</td>
<td>3.38</td>
<td>3.07</td>
</tr>
<tr>
<td>5</td>
<td>1.71</td>
<td>2.76</td>
<td>3.62</td>
<td>3.36</td>
<td>3.05</td>
</tr>
<tr>
<td>6</td>
<td>1.71</td>
<td>2.76</td>
<td>3.62</td>
<td>3.36</td>
<td>3.05</td>
</tr>
<tr>
<td>7</td>
<td>1.79</td>
<td>2.71</td>
<td>3.64</td>
<td>3.29</td>
<td>3.07</td>
</tr>
<tr>
<td>8</td>
<td>1.88</td>
<td>2.86</td>
<td>3.55</td>
<td>3.21</td>
<td>2.95</td>
</tr>
</tbody>
</table>
Fig. 1. Average accuracies of heterogeneous and homogeneous boostings.

- (a) arrhythmia
- (b) autos
- (c) balance-scale
- (d) blood-transfusion-service-center
- (e) breast-cancer-wisconsin-diagnostic
- (f) breast-cancer-wisconsin-original
- (g) breast-tissue
- (h) car-evaluation
- (i) cardiotocography-1
- (j) cardiotocography-2
- (k) chess-king-rook-vs-king-pawn
- (l) cmc
Fig. 2. Average accuracies of heterogeneous and homogeneous boostings.

- **(a)** connectionist-bench-sonar-mines-vs-rocks
- **(b)** connectionist-bench-vowel-recognition-deterding
- **(c)** cylinder-bands
- **(d)** dermatology
- **(e)** ecoli
- **(f)** glass
- **(g)** habermans-survival
- **(h)** hepatitis
- **(i)** ionosphere
- **(j)** iris
- **(k)** libras-movement
- **(l)** liver-disorders
Fig. 3. Average accuracies of heterogeneous and homogeneous boostings.

(a) lymph
(b) monks-problems-1
(c) monks-problems-2
(d) monks-problems-3
(e) parkinsons
(f) pima-indians-diabetes
(g) sonar
(h) spambase
(i) spect-heart
(j) spectf-heart
(k) statlog-australian-credit
(l) statlog-german-credit
Fig. 4. Average accuracies of heterogeneous and homogeneous boostings.

(a) statlog-heart
(b) statlog-vehicle-silhouettes
(c) teaching-assistant-evaluation
(d) thyroid-disease
(e) vote
(f) wine