New Feature Selection Method for Multi-class Data: Iteratively Weighted AUC (IWA)

Petr Honzik¹, Pavel Kučera¹, Ondřej Hynčica¹, Daniel Haupt¹
¹ Brno University of Technology, Faculty of Electrical Engineering and Communication, Department of Control and Instrumentation, Kolejní 2906/4, 61200, Brno, Czech Republic, honzikp@feec.vutbr.cz, kucera@feec.vutbr.cz, hyncica@feec.vutbr.cz, xhaupt00@stud.feec.vutbr.cz, http://www.uamt.feec.vutbr.cz

Abstract—This paper deals with the new filter feature selection method Iteratively Weighted Area under Receiver Operating Characteristic (IWA). It is aimed for the multi-class problems with quantitative inputs. The experiments prove its superior quality in comparison to the equivalent methods.

Keywords—feature selection; classification; multi-class; ROC; AUC; IWA

I. INTRODUCTION

With the increasing availability of the real data increases the need to identify and select preferably small subgroups of relevant and mutually independent features which hold the majority of the information. In machine learning, in the pre-processing phase, feature selection (FS) deals with this problem referred to as a curse of dimensionality [1], [2].

In the supervised learning the FS approaches fall into two groups: filter and wrapper methods [3], [4]. Wrapper methods usually use the same type of model during FS and afterward during modelling. Thanks to that the FS methods are suited exactly for the concrete model and can achieve better results. On the other hand this approach is followed with high computational requirements and risk of high bias inherited in the selected subgroup. Filter methods are faster, more general and don’t suffer with the risk of the biased estimate of the features subgroup [5], [2].

Furthermore the important limitation for the choice of the concrete FS method is the type of the output variable. It can be quantitative (regression models) or qualitative (classifiers). Especially in case of multi-class classification problems the computation of some measures is still unsolved. For instance the area under receiver operating characteristic (AUC) for binary classification can be interpreted in more ways (graphically, probabilistic, as a sorting problem). All the interpretations lead to the same algorithm in binary AUC but differ in the extension to the multi-class AUC [6]-[8].

The aim of this paper is to introduce a new filter FS method for multi-class problems based on AUC – the Iteratively Weighted AUC (IWA).

II. FEATURE SELECTION METHODS BASED ON AUC

There are two principles used for the filter FS that are based on the AUC. The first one is simple greedy algorithm that ranks and selects the features according to their AUC. The principle of this approach is described e.g. at [9]. The implementation itself was named FAST (Features Assessment by Sliding Threshold) [10].

The more sophisticated filter FS methods use the correlation to eliminate the same or similar features. The general principle [9] is defined as follows (1),

\[ i_k = \arg \max_j \left\{ \alpha C(j) - \frac{\alpha^2}{k-1} \sum_{r=1}^{k-1} \rho_{j,r} \right\} \]  (1)

where the \( i_k \) is the index of the \( k \)-th selected feature, \( C(j) \) is the selected measure of separability of the \( j \)-th class, \( \alpha_1 \) and \( \alpha_2 \) are the parameters and the \( \rho_{j,r} \) is the correlation between features \( i \) and \( j \).

Based on this principle, the concrete filter FS method using AUC as the measure of separability and nonparametric correlation (Spearman’s correlation) was suggested and was named ARCO (AUC and Rank Correlation coefficient Optimization) [11]. The selection criterion is based on the following equation (3),

\[ E_i = AUC(f_i) - \frac{\sum_{f_j \in S} \text{RCC}(f_i, f_j)}{|S|} \]  (2)

where \( f_i \) is a feature from the group of still unselected features, \( S \) is the group of already selected features and \( \text{RCC} \) is Spearman’s correlation coefficient.

To implement the above mentioned FS methods the multi-class AUC must be determined. There are more formulas for multi-class AUC computation [6] but the most extended and practically used is the method (3) published by Hand and Till [12].
\[ AUC = \frac{1}{c(c-1)} \sum_{i<j} A_{ij} \]  

(3)

where \( c \) is the number of classes and \( A_{ij} \) is the binary AUC between classes \( i \) and \( j \). For more details see [12].

III. NEW FEATURE SELECTION METHOD IWA

A. Weighting and Linear Transformation of AUC

Equation (3) describes the computation of the multi-class AUC by Hand and Till [12]. The same formula can be interpreted as weighted average where the matrix of binary AUCs is assigned as \( A \) and the weights are written into the weight-matrix \( W \) (see Fig. 1).

<table>
<thead>
<tr>
<th>Matrix of binary AUCs A</th>
<th>Weight-matrix W</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.6</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>

(1) A 2 3 4

Figure 1. Structure of the matrices \( A \) and \( W \); the indexes \( i \) and \( j \) assign the four output classes.

Then the equation for computation of the multi-class AUC can be modified as in (4) what is the formula for weighted average.

\[ AUC = \frac{\sum_{i<j} A_{ij} \cdot W_{ij}}{\sum_{i<j} W_{ij}} \]  

(4)

The same formula (4) can be written using the Hadamard product (5). This form will be used in rest of the paper.

\[ AUC = \frac{\text{sumAD}(A \circ W)}{\text{sumAD}(W)} \]  

(5)

The function \( \text{sumAD}() \) assigns the sum of elements above the main diagonal.

Furthermore for better interpretability of the algorithm the AUC, which is the value in the range \( 0,1 \) (0.5 assigns random feature, 0 and 1 assign maximal measure of separability), AUC is linearly transformed into the range \( (-1,1) \). Finally the absolute value of this transformation is used (see Eq. (6)) and will be assigned with the symbol \( D \) (0 assigns random feature and 1 assigns maximal measure of separability).

\[ D_{ij} = |A_{ij} \cdot 2 - 1| \]  

(6)

The advantage of this transformation is that the measure of separability expressed with \( D \) and the values of weight-matrix \( W \) are both in the same range \( (0,1) \) where 0 means none information and the value 1 means maximal information.

B. Information Profit

When having the matrix \( D^k \) describing separability measure of some feature \( F^k \) and the weight-matrix \( W \) describing knowledge in some subgroup of features, the potential information profit \( I \) from the new feature can be expressed as the weighted average of the sum of the elements in \( D^k_{ij} \) weighted each by \( W_{ij} \). This can be expressed using the Hadamard product (7).

\[ I^k = \frac{\text{sumAD}(D \circ (1-W))}{\text{sumAD}(1-W)} \]  

(7)

Finally the weight-matrix \( W \) should be recalculated after addition of feature \( F^k \) to the subgroup of features using the following formula (8).

\[ W = W + D \circ (1-W) \]  

(8)

C. IWA - the New Feature Selection Algorithm

IWA (Iteratively Weighted AUC) uses the previously mentioned principles to iteratively change the weight-matrix according to the quality of information hold by the already selected features. At the initial state the \( W(0) \) is matrix of zeros (no any feature is selected). The subgroup of selected features \( S_T \) in period \( T=0 \) is empty. In the first iteration the feature with highest AUC is selected and its transformed matrix \( D(1) \) is copied to the \( W(1) \). In the next iterations the information profit \( I \) from the remaining features in \( S_{T'} \) (subgroup of unselected features) is computed using the formula (8), the one with highest \( I \) is selected and \( W(T) \) is recalculated using the formula (8). The algorithm can be formally described as follows:

01 \( W(0) \) = matrix of zeros
02 \( S_0 = \{0\} \)
03 for \( T=1 \) to \( M \) do
04 for each \( k: F^k \in S_{T-1} \)
05 \( D^k(T) = D^k(0) \circ (1-W(T-1)) \)
06 end
07 for each \( k: F^k \in S_{T-1} \)
08 \( z = \text{argmax}(\text{sumAD}(D^k(T))/\text{sumAD}(1-W(T-1))) \)
09 end
10 \( S_T = S_{T-1} \cup F^z \)
11 \( W(T) = W(T-1) + D^z(T) \)
12 end
The interpretation of the pseudo-code:
01 all the weights are setup to zero
02 subgroup of selected features is empty
03 do M-times (M = number of features)
04 for all still unselected features
05 compute their weighted D matrix in iteration T
06 from all still unselected features
07 find index z of the feature with highest weighted sum
08 of elements above the main diagonal of D
10 the best feature F^z is added among already selected
11 the matrix of weights is updated using the D^z of the

D. Example

In the following example the principle of the new FS
algorithm IWA is shown. The matrices are drawn in
compressed way. The drawn parts are the white (non-zero)
parts of matrices at Fig. 1.

Let’s have the weight-matrix W(0) in its initial state
and 3 input features F^0, F^2 and F^3. Using equations (5)
and (6) their linearly transformed forms D^1(1), D^1(1) and
D^1(1) are computed. Using the formula (7) the information
profits I^1, I^2 and I^3 are computed. Because the I^1 has the
highest information profit, feature F^0 will be selected as
the first one and removed to the subgroup S^1. Until now
the computation is equivalent to almost all the filter FS
methods. But at this moment the new weight-matrix W(1)
is computed using the formula (8). In the first step the
W(1) is the same matrix as the matrix D of the best found
feature (see Fig. 2).

<table>
<thead>
<tr>
<th>0.3</th>
<th>0.6</th>
<th>0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.7</td>
<td>0.9</td>
<td>W(1)</td>
</tr>
</tbody>
</table>

F^5 \rightarrow D^1(2) F^2 \rightarrow D^1(2) F^3 \rightarrow D^1(2)

<table>
<thead>
<tr>
<th>0.3</th>
<th>0.6</th>
<th>0.9</th>
<th>0.8</th>
<th>0.4</th>
<th>0.3</th>
<th>0.2</th>
<th>0.5</th>
<th>0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.7</td>
<td>0.9</td>
<td>0.2</td>
<td>0.4</td>
<td>0.2</td>
<td>I^1=0.60</td>
<td>0.6</td>
<td>0.9</td>
<td>I^2=0.54</td>
</tr>
<tr>
<td>0.86</td>
<td>0.76</td>
<td>0.93</td>
<td>0.82</td>
<td>0.92</td>
<td>W(2)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.3</td>
<td>0.6</td>
<td>0.9</td>
<td>I^3=0.55</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.7</td>
<td>0.9</td>
<td>0.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 2. The first iteration step of the IWA algorithm. The indexes correspond to the indexes of white (non-zero) parts of matrices at Fig. 1.

What is important is the change in the information profit of the remaining features (compare Fig. 2 and Fig. 3). The first feature F^0 was already selected and will be no more used. The information profit I of the second feature F^2 seriously increased when compared to the first iteration at Fig. 2. It is caused by the best measures of separability for the pairs of classes A_12=0.8 and A_34=0.5. As can be seen from the weight-matrix W(1), the information about these two pairs is at lowest level in the subgroup of selected features and that is why they will have the biggest weight in computation of the information profit in this iteration. On the other hand the information profit I_3 of the third feature F^3 decreased. The feature F^2 has now the highest information profit I and will be selected. The F^3 will be selected as the last one.

IV. EXPERIMENTS

A. Datasets

The experiments are based on the dataset ISOLET5 (1560 records) for training and the dataset ISOLET1234 (6237 records) for testing [13]. The data consist of 617 quantitative input features and one qualitative output. The features were extracted from the names of letters spoken by 150 subjects. The number of output classes is equivalent to the number of letters in English alphabet what is 26.

B. Feature Selection Methods

The three filter FS methods are compared in the experiments. The first one is the FS method FAST [10] that is used as the baseline to which the other methods are compared.

The second FS method is ARCO [9], [11] that decreases the information of the newly selected feature proportionally to its correlation against the already selected features.
The third FS method is the newly introduced IWA. It computes the multi-class AUC as a weighted average where the weight–matrix changes according to the already selected features.

All the FS methods were programmed in MATLAB.

C. Feature Selection Evaluation

The performance of the FS methods is expressed as the accuracy achieved using the 1NN (one Nearest Neighbour) model. This model is trained using the training data and the accuracy, based on which the FS methods are compared, is computed using the test data. We prefer this approach against using Cross Validation over all the available data to eliminate the risk of misleading results [1]. 1NN is sensitive to the irrelevant, redundant and non-normalized features and is obviously used in studies aimed for comparing the filter FS methods [1], [10]. We emphasize that our experiments are organized with aim to compare the FS methods and not to achieve the highest performance. That is why the very sensitive model (1NN) was used.

The experiments were realized in the open-source software RapidMiner.

D. Results

Three FS methods were compared on the ISOLET data. Firstly the selection was made from 50 features (Fig. 4), secondly from 100 features (Fig. 5) and finally from 200 features (Fig. 6).

The most interesting and representative results from all three experiments are shown in Tab. 1.

<table>
<thead>
<tr>
<th>Number of Selected Features</th>
<th>Select from 50 features</th>
<th>Select from 100 features</th>
<th>Select from 200 features</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARCO</td>
<td>IWA</td>
<td>ARCO</td>
<td>IWA</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>61</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>17</td>
<td>36</td>
<td>-2</td>
</tr>
<tr>
<td>15</td>
<td>11</td>
<td>18</td>
<td>15</td>
</tr>
<tr>
<td>20</td>
<td>9</td>
<td>9</td>
<td>17</td>
</tr>
</tbody>
</table>

E. Discussion

From all the three graphs it is obvious that the new method IWA overperforms the last two methods. Especially the first few features selected by IWA hold significantly more information than the subgroups selected by FAST or ARCO. This can be useful in applications where the number of inputs is strictly limited.

The FAST method just ranks the features according to multi-class AUC and does not reflect e.g. the fact that two best features may be the same. The ARCO method tries to identify the same or similar features. It uses correlation and it is known to be better than the FAST method [9], [11]. The problem is that the correlation is computed without relation to the output variable. Two fully correlated features are interpreted as the same by ARCO but they can classify differently, hold different information and be complementary to each other [2].

The effectiveness of IWA is caused by the decomposition of the multi-class AUC to the set of the independent binary problems. The similarity between two features is not based on the correlation, which has not relation to the output variable, but on the similarity in the particular binary AUCs. It enables to focus on the problematic pairs of classes during the iterations.

The performance of ARCO may be changed by different setup of parameters \( \alpha_1 \) and \( \alpha_2 \) in the equation (1). But the need to setup these parameters with no guarantee that better results will be achieved makes the IWA method even more interesting.

What is necessary to do in next research is to test the IWA method on more different datasets, compare it to the
other FS methods including the wrapper methods. Furthermore it will be important to test different models to find out whether the effectiveness of IWA will be still so high when compared to other methods. Finally the different methods for modification of weight-matrix can be tested. The current approach suffers from the problem that after decades of selections all the weights become equal to one and the selection of next features is random. This could be the reason why the IWA loses its quality against the other methods after approximately 30 iterations (see Fig. 4, 5 and 6).

V. CONCLUSION

New feature selection method IWA (Iteratively Weighted AUC) specialized in multi-class problems was introduced. In three experiments it was shown that it significantly outperforms the two filter feature selection methods based on the AUC – the FAST (Feature Assessment by Sliding Threshold) and the ARCO (AUC and Rank Correlation coefficient Optimization) methods.

The next research will be focused on the comparison with more feature selection methods, testing on different types of models and more datasets.

REFERENCES


