Information Theory vs Correlation Based Feature Ranking Methods in Application to Metallurgical Problem Solving

Marcin Blachnik[1], Adam Bukowiec[1], Mirosław Kordos[2], Jacek Biesiada[1]

[1] Silesian University of Technology, Electrotechnology Department, Katowice, Krasinskiego 8, Poland;
[2] University of Bielsko-Biała, Department of Mathematics and Computer Science, Bielsko-Biała, Willowa 2, Poland;

Abstract. Feature selection is a typical stage of building any classification or regression model. There are several approaches to it, however one of the fastest is based on determining the relevance of each feature independently by calculating ranking values. In this paper we provide empirical comparison of four different ranking criteria that belong to two different groups *information theory* and *correlation metrics*. The comparison is performed on the empirical datasets obtained while building a model used for predicting mass of chemical compounds necessary to obtain steel of predefined quality.

1 Introduction

One of the most popular metallurgical process of steel production base on melting steel scraps in an electric arc furnace (EAF) [1]. This process can be divided into three stages, where in the first stage steel scraps are melted in EAF, then the other elements are added in order to fine-tune the composition of the steel to meet the specification and the customers requirements. This is done in the ladle arc furnace (LHF), and the last stage is casting. Aluminum, silicon and manganese are the most common deoxidizers used in LHF stage. Therefore during the tapping some elements are added into the metal stream casting from EAF to LHF furnace, other are placed on the bottom of the ladle, and the rest of compounds are added during LHF procedure.

In this process one of the challenging problems is the prediction of the amount of compounds (aluminum, silicon and manganese and others) that are necessary to fulfill steel specification. In practice the refining process is often suspended while verifying chemical properties of the steel. This part of that process is very expensive requiring turning on and off the furnace and the arc. Therefore reducing the number of chemical tests may decrease the costs by reducing time and energy consumption. Another goal facing steel making engineers is the reduction of the amount of steel add-ins which often are very expensive.

Currently this problem is solved by some simple theoretically defined linear equation, and the experience of the engineer controlling that process. The existing solution calculates the amount of chemical compounds just considering the information of the given steel specification. In practice these parameters depend on melting time of the

EAF process (longer melting determines higher reduction of chemical elements), time of melting during LHF process, the steel specification, steel scrap types used for melting, etc. To consider all these parameters a new model has to be built. Our preliminary estimation showed that linear model is not suitable for that problem, so a more advanced nonlinear model had to be considered. Our choice was one of the most popular and widely used Support Vector Regression (SVR) algorithm, with gaussian kernel function. Another important stage of data mining process is feature selection which may simplify the model, and often increase model accuracy. In our experiments we decided to test quality of different ranking based feature selection methods. It's very simple and fast feature selection approach. The choice of this method is motivated by the results of NIPS'2003 Feature Selection Challenge [5] where one of the best results were obtained using mentioned ranking based approach. However we face the problem of selecting the appropriate ranking coefficient. For that purpose we have compared two families of ranking coefficients based on information theory - two metrics such as *Information Gain* (IGR-index), and *Mantaras distance* (D_{ML}), and two correlation matrics: simple Pearsons linear correlation coefficient, and Spearmans rang correlation.

Both of these problems - building the model for predicting amount of chemical elements necessary to obtain the correct steel grade, and the comparison of both feature ranking families will be covered in this paper.

Next section briefly describes feature selection, in details describing ranking based methods, with four different ranking criterions. Section (3) provides phases of metallurgical problem modeling, while section (4) presents results of feature selection for all listed ranking methods. Last section concludes the paper.

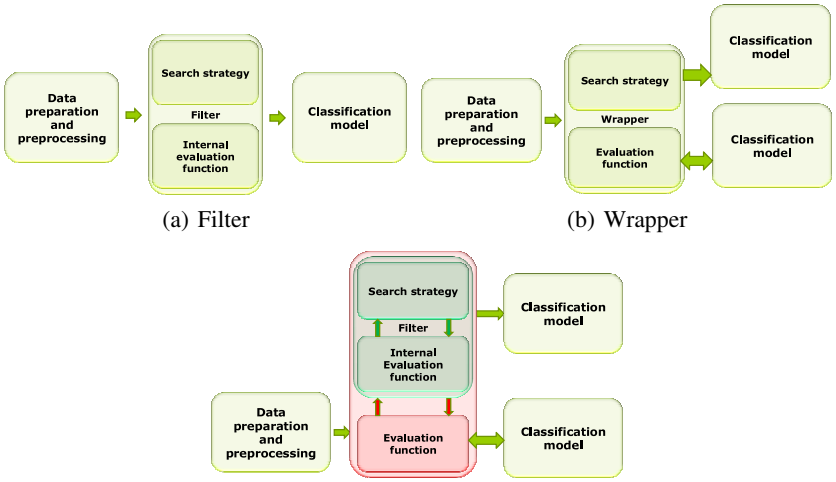
2 Feature selection methods

Feature selection is an important problem in data mining tasks. Selecting best subset of n features out of full set of m is an NP-hard problem. Already many different search strategies have been developed for solving that problem [5], however in real application most of them can not be applied because of computational complexity. There are four type of methods used in feature selection:

- Embedded methods - where feature selection is embedded into the prediction algorithm, like in decision trees
- Filter method - where feature selection is done independently to the classification model. As presented in fig. (1) the filter selects a feature subset using evaluation function defined as some statistical criteria like *Kullback–Leibler divergence* or *mutual information*. The advantage of this approach is feature selection speed, being one of the fastest methods, and taking feature subset independently to the classifier. However, its drawback is that the feature subset is not optimally selected for a particular prediction model.
- Wrapper approach - [7] where the evaluation function used to determine quality of the feature subset is defined as accuracy of a specific classifier, which is also used as a predictive model. Block diagram of the that approach is presented in fig. (1). The most important advantage of that approach is the prediction accuracy of the

final model, unfortunately it is atone by higher computational complexity needed to train the classifier for each searching step.

– Frappers approach - is a combination of filters and wrappers, where hyperparameters of the feature filter are tuned by the accuracy of the classifier. A scheme of this approach is represented in fig. (1). This approach is a compromise between filters and wrappers caring advantages and disadvantages of both approaches.



(a) Filter (b) Wrapper

(c) Frapper

Fig. 1. Three types of feature selection methods

2.1 Ranking based feature selection

One of the simplest and fastest group of methods dedicated to feature selection are ranking based algorithms. In this approach, which belongs to feature filters, each feature is independently assigned a coefficient $J_i = f(\mathbf{f}_i, \mathbf{y})$ that describes it's abilities to predict values of output variable $\mathbf{y}$ given just single input variable $\mathbf{f}_i$. Having such a coefficient features are sorted from the most relevant (with the highest coefficient) to the least relevant (with the lowest coefficients value). Then from initial m features, first n features are selected, and for this subset the prediction model is build. The sketch of this algorithm is presented in fig. (1)

As described earlier this is the fastest algorithm while its computational complexity is linear with respect to the number of features $O(m)$. The only problem that appears using this algorithm is determining correct number of the selected features (n). This part can be done using wrappers approach, where the quality of selected number of features is determined using prediction algorithm. In ranking based methods, one problem still remains; what kind of relevance index $J(\cdot)$ should be used. In classification

Algorithm 1 Feature selection based on feature ranking

Require: $\mathbf{f}$ {Initial feature set}
Require: $\mathbf{y}$ {Output variable}
Require: $J(\cdot)$ {Ranking function}
  for $i = 1 \ldots m$ **do**
    $a_i \leftarrow J(f_i, \mathbf{y})$ {Calculate rank value}
  end for
  $\mathbf{f} \leftarrow \mathrm{sort}(\mathbf{f}, \mathbf{a})$ {Sort features according to $J()$}
  $acc \leftarrow 0$
  $\mathbf{f}^a \leftarrow \emptyset$
  for $j = 1 \ldots m$ **do**
    $\mathbf{f}^a = \mathbf{f}^a \cup f_i$ {Add new feature $\mathbf{f}^a$ to current subset}
    $tacc \leftarrow$oceń$(\mathbf{f}^a)$ { estimate subset quality $\mathbf{f}^a$ }
    if $tacc > acc$ **then**
      $acc \leftarrow tacc$ {If new subset is better then the previous one }
      $\mathbf{f}' \leftarrow \mathbf{f}^a$ {store current subset}
    end if
  end for
  return $\mathbf{f}'$

problems a comparison between different ranking algorithms can be found (ex. in [2]) where authors where unable to determine the best relevance index, and they concluded that the quality of the filter strongly depends on the analyzed dataset. Based on such conclusions while building a model predicting amount of necessary steel compounds we did a comparison between different rankers comparing two families of indexes - correlation based [9] and based on information theory [].

Pearson Correlation The simplest method determining J_i value for regression problems is Pearsons linear correlation coefficient. The value of J is calculated according to formula

$$J_P(\mathbf{f}_j, \mathbf{y}) = \left| \frac{\sum_{i=1}^{k} (\mathbf{y}_i - mean(\mathbf{y}))\,(\mathbf{f}_{i,j} - mean(\mathbf{f}_j))^2}{\sigma_{\mathbf{f}_j}\sigma_{\mathbf{y}}} \right| \tag{1}$$

where:

- $\sigma_{\mathbf{f}_j}$ and $\sigma_{\mathbf{y}}$ are standard deviation of variables $\mathbf{f}_i$ and $\mathbf{y}$.
- $mean(\mathbf{x})$ is a function returning mean value of given vector $\mathbf{x}$

This well known and popular coefficient is able to find only linear dependence between two random variables.

Spearman's rang correlation coefficient [] More advanced correlation coefficient is Spearman's rang coefficient, which is able to find nonlinear correlations between

two random variables. This metric is based on replacing original values of variables by associated rang values $\mathbf{r}$, and calculating earlier defined Pearsons linear correlation coefficient for variables replaced by their ranges.

$$J_S(\mathbf{f}_j, \mathbf{y}) = J_P(\mathbf{r}_{\mathbf{f}_j}, \mathbf{r}_{\mathbf{y}}) \tag{2}$$

Where:

- $\mathbf{r}_{\mathbf{f}_j}$ - rang values associated to $\mathbf{f}_j$ variable
- $\mathbf{r}_{\mathbf{y}}$ - rang values associated to the $\mathbf{y}$ variable

Appropriate ranges are given by ascending sorting variable values and assigning to each value number equal to the position in the sorted list. In case of ties, when certain value (a) appear q-times, where $q > 1$ (more then ones) $a = [v_i, v_{i+1}, v_{i+q}]$, the rang assigned with each of $v_i \cdot v_{i+q}$ values is equal mean rang

$$\underset{e=i,\ldots,i+q}{\forall} v_e = \frac{1}{q}\sum_{z=i}^{i+q}(r_z) \tag{3}$$

This solution is able to find monotonic nonlinear correlation between features.

Information Gain Ratio *Information Gain Ratio* (IGR) is a metric that belongs to information theory coefficients. It is based on Shanon entropy where:

$$\begin{aligned} H(\mathbf{x}) &= -\sum_{i=1}^{c} p(x_i)\lg_2 p(x_i) \\ H(\mathbf{x},\mathbf{y}) &= -\sum_{i,j=1}^{n,m} p(x_i, y_j)\lg_2 p(x_i, p_j) \\ H(\mathbf{x}|\mathbf{y}) &= H(\mathbf{x},\mathbf{y}) - H(\mathbf{x}) \end{aligned} \tag{4}$$

- $p(x_i)$ - is the probability of x_i

used to define *mutual information*:

$$MI(\mathbf{x},\mathbf{y}) = -H(\mathbf{x},\mathbf{y}) + H(\mathbf{x}) + H(\mathbf{y}) \tag{5}$$

The final formula determining is defined as:

$$IGR(\mathbf{x},\mathbf{y}) = \frac{MI\,(\mathbf{x},\mathbf{y})}{H\,(\mathbf{x})} \tag{6}$$

The IGR metric describes amount of information about variable $\mathbf{y}$ providing variable $\mathbf{x}$. The IGR value is equal to the *Information Gain* normalized by the entropy of variable $\mathbf{x}$.

Mantaras Distance Ranking Mantaras distance D_{ML} is a matric defined as

$$D_{ML}(\mathbf{x}, \mathbf{y}) \; = \; H(\mathbf{x}|\mathbf{y}) \; + \; H(\mathbf{y}|\mathbf{x}) \tag{7}$$

the advantage of D_{ML} metric is the ability to preserve all distance axioms.

3 Dataset and preprocessing steps

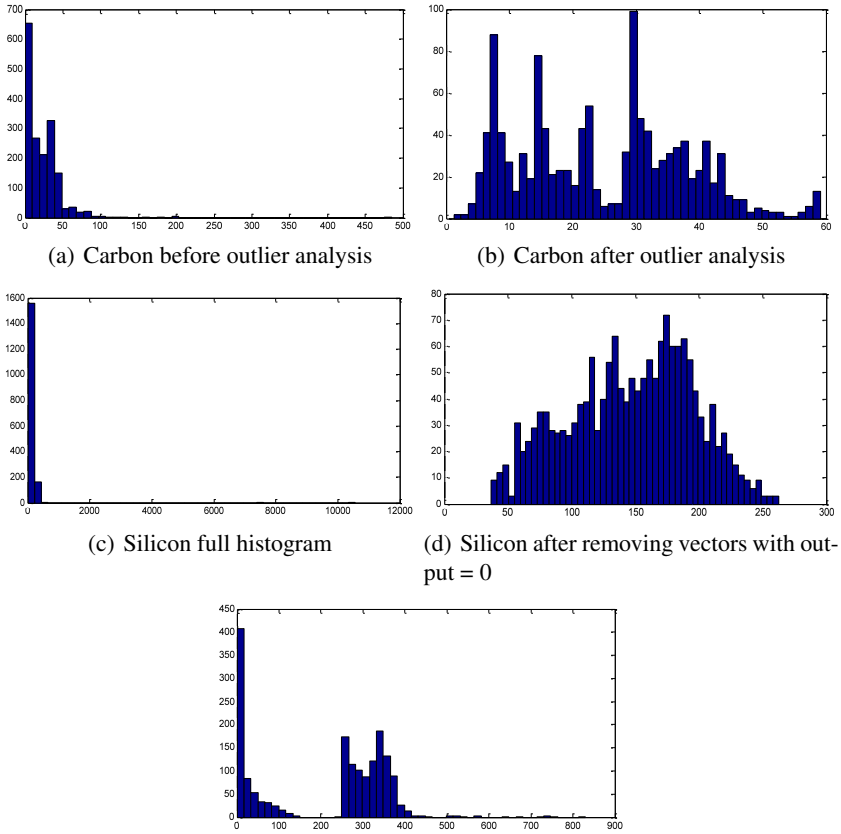
The dataset used to build the prediction model consists of historical data obtained from one of polish steel mills. The data describes refining and fine tuning steel parameters during the ladle arc furnace stage of melting. The problem is described as predicting amount of three most important chemical compounds necessary to assure appropriate steel quality based on 41 features. This three compounds are amount of carbon, manganese and silicon, which were obtained from the real chemicals put to the steel knowing its percentage composition. The input variables are weight of steel while flushing the EAF furnace, amount of energy needed to melt the steel, amount of oxygen utilized during the EAF stage, results of chemical analysis before flushing the EAF furnace, and results of chemical analysis of the steel during the LHF stage. Presented below histograms are just for single steel grade called S235JRG2.

3.1 Amount of Carbon analysis

The histogram of output variable before any preprocessing is presented in fig (2). Analysis of this histogram point out that predicting amount of Carbon over 100kg is impossible due to the low number of training samples. To avoid the problem of unstable behavior of the training process caused by the small number of training samples over 100kg outlier analysis was provided based on interquartil range, also to avoid the problem of dominating 0 value, when no Carbon was added, all samples with carbon = 0 were removed, the histogram of remaining output variable is presented on fig. (2), and for that data regression model was built. The process of predicting 0 carbon was further realized based on binary classification problem, where class C_{-1} was related to "no carbon", and class C_1 to "add carbon" where the mass of necessary carbon was delivered by the regression model. In this paper we will present results only for regression problems.

3.2 Amount of Manganese analysis

The analysis of histogram of output variable for Manganese (fig.(2)) suggests existence of two clusters. One when amount of Manganese was below 200kg, and second, when Manganese was over 200 kg, which after outlier analysis was shrank to the period between 200 and 500 kg.



(a) Carbon before outlier analysis

(b) Carbon after outlier analysis

(c) Silicon full histogram

(d) Silicon after removing vectors with output = 0

(e) Manganese

Fig. 2. Histograms of the output variable for chemical elements

3.3 Amount of Silicon analysis

Similar problem of dominating histogram bar for 0 value that we face during Carbon analysis appeared also during Silicon analysis. The histogram before and after removing dominating 0 are presented in figures (2.c,d) respectively. Also in this problem classification model was built to predict either "no Silicon" or "add Silicon", and also only the results of the regression algorithm are presented

4 Comparison of ranking criterions

As suggested in the introduction SVM for regression (SVR) [3] with Gaussian kernel and ϵ-insensitive cost function was selected to be used for solving all regression problems. Unfortunately training SVR algorithm require searching for optimal hyperparameters in the cubic space. The hyperparameters of the SVR model are margin

softness (C - value), kernel parameter (γ), as well as ϵ in the ϵ-insensitive cost function. This problem was solved by greed search algorithm, where $C = [12832128]$, $\gamma = [0.50.711.31.5]$, and $\epsilon = [0.10.010.001]$.

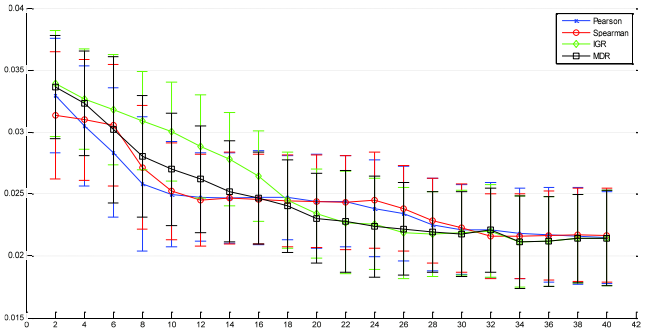
For each number of features ordered by the ranking, SVR with all possible set of parameters was trained. Results presented in the figures were obtained with a 10-fold cross validation (CV) that was repeated 3 times (3x10 CV). The best of all possible results for given number of features were selected and presented in figure (3) (smallest MSE). All calculations were obtained using Matlab toolbox for data mining called Spider [8], extended by the Infosel++ feature selection library created by our group [6].

5 Conclusions

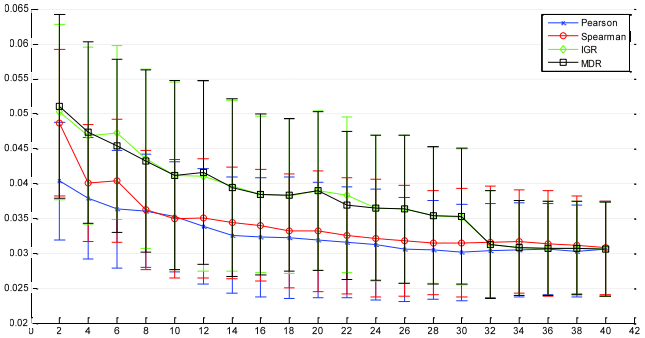
As suggested in the introduction, this paper covered two type of problems: building a statistical model for metallurgical problems and comparing two groups of ranking criteria used for feature selection. After outlining the problem, and after discussing different types of feature selection in details analyzing ranking based feature selection we have considered steps necessary to take before building regression models. While building the model we have found the problem of small number of samples for some ranges of output variable. Because of that we decided to remove such ranges performing outlier analysis. This helped us obtain better accuracy of our model. Similarly much better results were obtained splitting the output variable into two groups like for Manganese problem.

We have also separately considered the problem of feature selection based on two types of ranking criteria. The first of them were information theory metrics such as *Information Gain Ratio* and *Mantaras distance*. The second group consists of two correlation metrics: Pearsons linear correlation and Spearmans rang correlation. The obtained results were surprisingly since, in almost all cases the number of all relevant features was almost equal to the whole number of features in the dataset. For Manganese (B) and Carbon both *correlation based criteria* were dominating *information theory metrics*. However, the best average results for Carbon dataset were obtained for 32 features using both information theory metics, while for Manganese all features were required. In case of silicon dataset it can be seen that in the beginning the best results are provided by the linear correlation matric while in the second part for higher number of features information theory metrics started to dominate, allowing to obtain the best results for 34 features. Analysis of Silicon problem shows that both *information theory* indexes are dominating in the whole analyzed feature space. Concluding *information theory* based ranking sims to be better then the correlation measures, however not for all cases (ex. Manganese (B)). Obtained results allow for another interesting observation that the results obtained by Spearmans rang correlation were usually dominated by the simple linear correlation index.

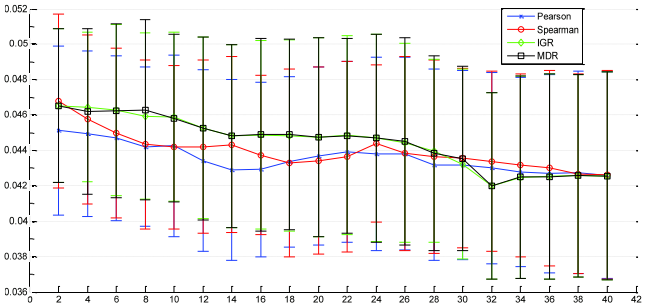
Acknowledgement: This work was funded by the Polish Committee for Scientific Research grant 2007-2010 No.: N N519 1506 33.



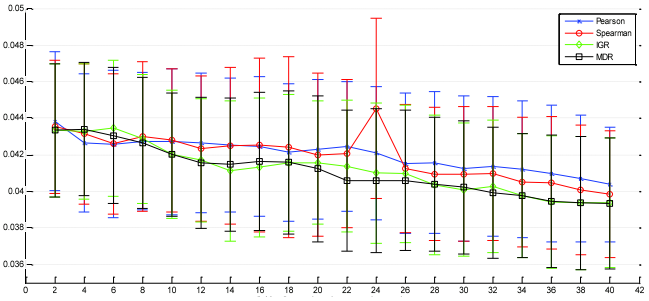
(a) Manganese Type A



(b) Manganese Type B



(c) Carbon



(d) Silicon

Fig. 3. MSE and its standard deviation in the function of selected features

References

1. Lis T., Współczesne metody otrzymywania stali. Wydawnictwo Politechniki Śląskiej, Gliwice, 2000.
2. Duch W, Wieczorek T, Biesiada J, Blachnik M, Comparison of feature ranking methods based on information entropy. Proc. of International Joint Conference on Neural Networks (IJCNN), IEEE Press, Budapest 2004,
3. Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm
4. Duch, W.: Filter methods. In Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L., eds.: Feature extraction, foundations and applications. Springer (2006) 89–118
5. Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L. (eds): Feature extraction, foundations and applications. Springer, Berlin (2006)
6. Kachel, A., Blachnik, B., Duch, W., Biesiada, J., Infosel++: Information Based Feature Selection C++ Library, Software avaiable at http://metet.polsl.pl/~jbiesiada/infosel - in preparation.
7. Kohavi R. and John G., Wrappers for Feature Subset Selection. In Artificial Intelligence journal, special issue on relevance, 1997, Vol. 97, Nos 1-2, pp. 273-324.
8. Weston J., Elisseeff A., BakIr G. and Sinz F., Spider 1.71 2006, Software available at http://www.kyb.tuebingen.mpg.de/bs/people/spider/
9. Hollander M., Wolfe D.A., Nonparametric statistical methods, Wiley, 1973