Two-level Clustering Approach to Training Data Instance Selection: a Case Study For the Steel Industry

Heli Koskimäki, Ilmari Juutilainen, Perttu Laurinen, Juha Röning, *Member, IEEE*

Abstract— Nowadays, huge amounts of information from different industrial processes are stored into databases and companies can improve their production efficiency by mining some new knowledge from this information. However, when these databases becomes too large, it is not efficient to process all the available data with practical data mining applications. As a solution, different approaches for intelligent selection of training data for model fitting have to be developed. In this article, training instances are selected to fit predictive regression models developed for optimization of the steel manufacturing process settings beforehand, and the selection is approached from a clustering point of view. Because basic k-means clustering was found to consume too much time and memory for the purpose, a new algorithm was developed to divide the data coarsely, after which k-means clustering could be performed. The instances were selected using the cluster structure by weighting more the observations from scattered and separated clusters. The study shows that by using this kind of approach to data set selection, the prediction accuracy of the models will get even better. It was noticed that only a quarter of the data, selected with our approach, could be used to achieve results comparable with a reference case, while the procedure can be easily developed for an actual industrial environment.

I. INTRODUCTION

In recent years the amount of digital information around us has exploded. In industry, huge databases have been introduced for storing the information gathered from manufacturing processes, consisting of hundreds of thousands different numerical or descriptive values. Naturally, some interesting knowledge can be assumed to lie in the data, but finding it is not a trivial matter or even possible with the human eye. Thus, a computer-driven approach called data mining (also called knowledge discovery) has become an attractive research area. The term can be defined as nontrivial extraction of implicit, previously unknown, and potentially useful information from data [1].

However, while these data mining methods are becoming more common, also more interest is focused on the background of these methods. Data selection is considered essential for successful mining applications, but as the amount of data available increases it can be overwhelming to use all of the data. In many cases the problem is solved by using different feature selection methods where meaningless columns of the data set are removed, while another approach called instance selection performs a reduction of the insignificant rows of the data set [2].

Data set selection also has an important role when fitting predictive industrial data-based models. At Ruukki's steel

Heli Koskimäki is with the Department of Electrical and Information Engineering, University of Oulu, Oulu, Finland (phone: +358 85532818; fax: +358 85532612; email: heli.koskimaki@ee.oulu.fi).

works in Raahe, Finland, development engineers control mechanical properties such as yield strength, tensile strength, and elongation of the metal plates beforehand on the basis of planned production settings. The solution is achieved using regression models introduced in [3]. However, acquirement of new data and the passing of time decrease the reliability of the models, which can bring economic losses to the plant. Thus, maintenance of the models emerges as an important step in improving modelling in the long run. In [4] the exact time when the performance of the model has decreased too much and updating of the model is needed was studied, while this study aims to find a reliable method that would select a suitable training data set for the models also in a long run when use of all the data available is not computationally reasonable. The input variables of the model are already fixed, thus the data set reduction has to be performed by selecting only significant instances as training data.

Nonetheless, it is not a trivial matter to select suitable training data instances when the data set consists of unevenly distributed observations (see Figure 1 for an example). For example, if there is a limited amount of data that can be used to train the model, how is it assured that when selecting training instances from a bigger data set, all the different
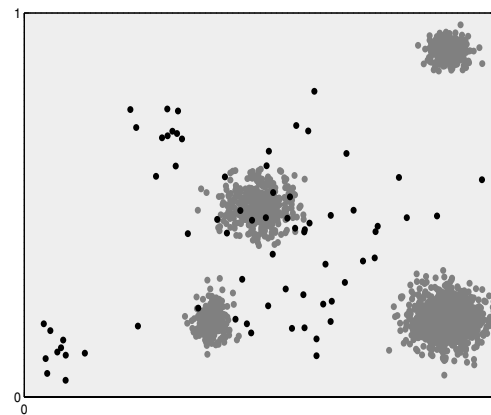

Fig. 1. Simplified example of data: common observations (light grey) form their own large clusters, while rare observations (black) are scattered around, also empty areas exist. Rare observations are influential and thus most valuable in model fitting. An instance selection algorithm should select the rare, influential observations, otherwise there will be significant loss in the information contained by the selected training data set.

cases are included in the training data. It is probable that by using basic sampling the common cases would dominate the model and some of the rare cases would be completely left out. In other hand, the rare cases are the ones that are the most necessary. For example, in the steel industry common products are made daily and the model to plan the production settings is not used, but in the production of rare cases the model is generally needed. Thus, leaving out the rare cases from the model training data would in the worst case lead to a weak model.

In addition, while it is not only necessary that selection of the training data could be performed automatically, it should also be reliable in the long run and calculation of the selection should be performed within a sensible time. These aspects that need to be considered make the problem more challenging. Thus, as an approach to training instance selection, a clustering structure was utilized and a procedure for selecting the instances using this structure sufficiently was studied. In addition, the cluster structure was formed in a way that new clusters can be added to the structure online, enabling it to be used to reveal areas where the reliability of the model is decreased. For example, the system can be trained to tell when a new data point creates its own cluster, meaning that at the time the model was trained no similar data were available and the reliability of model can be weak at this new data point.

Although questions of instance selection are tackled in several articles from the classification point of view ([5], [6], [7]), very few articles considering instance selection on other kinds of problems were found, regardless an extensive study of the author. In addition, most of the methods introduced in the classification-oriented articles are not applicable when class labels are not available.

The few instance selection studies found that were comparable with or adaptable to our problem have solved the problem with genetic algorithms [8]. There also are articles where the approach is based on tuning random sampling. For example, a data set is selected using random sampling and tuned by replacing random observations of the selected data with observations left out if the replacement improves the model's accuracy [9]. However, in this case this approach would be impractical and highly time-consuming, because the models would have to be refitted after each tuning step. Also approaches based on k-means clustering have given promising results in the classification area ([10], [11]) and they can also be adapted to a case where class labels are not available. Thus, the k-means algorithm-based solution was chosen for this study.

In this case, the instances are selected to fit predictive industrial data-based models. Thus, the effectiveness of the developed clustering is afterwards verified using the goodness of the method as an indirect measure (in this case the model prediction accuracy when clustering is used to select the training data instances). Indirect measures are also widely used in feature selection articles to select the best feature set ([12], [13]). Because it is noticed that there is not yet any universal selection method that would perform well in every situation, and since the best universal method for instance selection is still random sampling [2], in this article the results are compared with random sampling.

II. CLUSTERING ALGORITHMS

In this paper data set selection was approached by dividing the data into clusters containing similar observations. The most straightforward approach to clustering would have been to use the well known k-means algorithm directly with the data. The k-means clustering method aims to minimize the distance inside the clusters and maximize the distance between clusters by minimizing the squared error function

$$V = \sum_{i=1}^{k} \sum_{x_j \in S_i} |x_j - \mu_i|^2 \qquad (1)$$

where $S_i, i = 1, \ldots, k$ are the k clusters and μ_i are the cluster centroids and also the means of points $x_j \in S_i$. This is done using an iterative training sequence. In the algorithm the amount of clusters, $(=k)$, is fixed beforehand. The k cluster centroids are chosen randomly, after which every data point is attached to a cluster with a closest centroid point. The new centres for the clusters are then calculated and the centroid value is updated. The process is carried on iteratively until a point is reached where no changes to cluster centroids happen or a certain amount of iterations have been done [14].

However, the memory and calculation capacities needed to perform k-means clustering were found to be too large for this application. Thus, a new clustering method, called a coarse cluster algorithm, was developed to decrease the need for these capacities.

In coarse clustering the whole data set is divided into coarse clusters using a certain predefined distance limit so that two different clusters cannot contain any observations within this limit. This means the amount of clusters does not have to be decided beforehand. The structure also makes it possible to update the clustering online. Whenever the distance from a new data point to existing data points exceeds this limit, the new data point forms its own cluster. The pseudocode for coarse clustering is introduced as Algorithm 1.

Algorithm 1 Coarse clustering algorithm
**for** each x_i in data **do**
**if** $\exists$ one cluster: dist$(x_j, x_i) \leq$ d $\& x_j \in$ cluster **then**
cluster$\leftarrow x_i$
**else**
**if** $\exists$ several clusters: dist$(x_j, x_i) \leq$ d $\& x_j \in$ cluster
**then**
merge these clusters$\leftarrow x_i$
**else**
create new cluster$\leftarrow x_i$
**end if**
**end if**
**end for**

The advantage of this method is that when the distance limit, d, stays unchanged, the results of the coarse clustering are saved in the database, and the algorithm has to be performed only for a new observation, reducing the calculation time significantly. However, if the distance limit, d, is changed, the calculation time is still on a reasonable scale, giving the possibility to tune the algorithm.

After the coarse clustering is performed k-means clustering can be carried out for these clusters separately. The size of the coarse clusters depends on the limit and data set used. In a special case when the limit is high, all the data points are situated in a single cluster. Thus, applying k-means clustering directly to the whole data set is a special case of the proposed algorithm. On the other hand, when the data set is very scattered, as seen in Figure 1, some of these coarse clusters can be very small, although the chosen limit was not small. In that case, it is not reasonable to re-cluster all the coarse clusters, but to decide an amount, V, which tells how many observations there have to be in a coarse cluster for it to be further partitioned with k-means.

III. Instance selection algorithm

Although computers are getting more efficient, there are still problems in training data mining methods with all the data gathered. The memory or calculation capacities are not sufficient enough because the sizes of data sets are also growing. To tackle this problem, suitable methods have to be developed for selecting suitable instances from the whole data set as training data.

In this study clustering was formed to help in selecting suitable training instances from the original data. It was determined that by using clustering, similar observations would be situated in the same clusters and it would be sufficient to select a reduced amount of data from each of them to represent the actual data.

The study for finding the most suitable clustering method was started by recognizing the background requirements. The method will be implemented in software with the main purpose of maintaining the prediction models in a long run. The models are automatically retrained using the approach introduced in [4], and clustering is mainly used to select a suitable training data for the models. It is required that the user of the algorithm can decide beforehand the amount of observations to be included in the selected training data. In addition, easy implementation and reasonable time for calculations were considered important aspects.

Compactly stated, the whole instance selection algorithm developed in this article constitutes five steps:

1) Perform coarse clustering
2) Split coarse clusters whose size is greater than V using k-means clustering
3) Select all the instances from non-partitioned coarse clusters
4) Select instances from k-means clusters using equation 3
5) Select the rest of the instances using random sampling

IV. Selecting instances from clusters

After the clustering has been formed, the actual data selection phase follows. A sufficient amount of instances is selected from each cluster to guarantee that the selected training data set contains enough observations from all the regions of available data.

It seems reasonable that the number of observations from each cluster should reflect the information value that the cluster can offer for model fitting. Thus, it is proposed that the information value for model fitting is calculated for each cluster and then the number of selected observations is determined in the proportion of the cluster's information values. It is proposed that the information value of a cluster is measured on the basis of the location and spread of the cluster: Isolated clusters are more valuable because there are no other observations nearby. Clusters that are spread in a large volume in the input space are more valuable because also the internal variation of clusters helps the modelling algorithms to learn dependencies. Thus, it is proposed that the information contents of the ith k-means cluster is based on these measures.

Let $C_1, C_2, \ldots, C_q$ denote the clusters resulting from the application of coarse clustering and k-means algorithms and let $c_1, c_2, \ldots, c_q$ be the corresponding cluster centroids. Let $n_i = \#\{x_j | x_j \in C_i\}$ denote the number of observations included in the ith cluster. Then the information contents of the ith k-means cluster can be measured by

$$J_i = \exp(a_1 \sigma_i + a_2 m_i) \qquad (2)$$

where $\sigma_i = \sqrt{(1/n_i) \sum_{x_j \in C_i} ||c_i - x_j||}$ is the within-cluster deviation, $m_i = \max_j ||c_i - c_j||$ is the distance to the closest cluster centroid and a_1, a_2 are tuning constants. Small coarse clusters are valuable for model fitting because they are isolated from other input space regions that contain data. Thus, all observations from coarse clusters that were not partitioned using k-means are included in the selected training data. The amount of selected instances, S_i, of the partitioned clusters is achieved using the equation:

$$S_i = \frac{N * J_i}{\sum_{j=1}^{q} J_i}, \qquad (3)$$

where q is the amount of k-means clusters and N is the amount of observation that still need to be selected after the observations of the coarse clusters have been chosen.

The use of the exp function to define the information contents of a cluster makes it possible that the amount of observations that need to be selected from a cluster can be significantly more than there are observations in a cluster. This means if the amount of observations desired from a cluster is, for example 3000, but there are only 1000 observations in a cluster, 2000 observations will be missing from the final training data set. Thus, the amount of observations selected for training data after this selection phase is smaller than the actual desired data set size. However, use of the exp function makes it possible to select enough observations from the isolated and widely spread clusters and random

Metaparameter	value
scaling of inputs	gradient based
size of training data	50 000
distance limit, d	1.8
amount of observation for re-clustering, V	2000
k for kmeans	$\left\lfloor \frac{\#\text{observations in coarse cluster}}{1000} \right\rfloor$
$(a1, a2)$	$(8.2,\ 1.8)$

sampling can be used to select the remaining observations. This approach made it possible to include the interesting rare observations in the training data and at the same time also include enough variation to represent the actual data.

V. DATA SET AND REGRESSION MODELS

At Ruukki's steel works in Raahe, Finland, liquid steel is cast into steel slabs that are then rolled into steel plates. Many different process variables and mechanisms affect the mechanical properties of the final steel plates (for example, the element concentrations, the heat in the furnace and the thicknesses of the final plates). The production settings that fulfill certain mechanical properties are planned using a combination of two regression models. The first one tells the predicted property value while the second model gives the deviation around the property value, thus giving the actual working limits.

The data for this study were collected from Ruukki's steel works production database between July 2001 and April 2006. The whole data set consisted of approximately 250,000 observations. Information was gathered from element concentrations of actual ladle analysis, normalization indicators, rolling variables, steel plate thicknesses, and other process-related variables [15]. The observations were gathered during actual product manufacturing. This means there are plenty of observations of some daily made products, but also rare observations of products made less than once in a year. In addition, although there are plenty of observations of common products, they are not identical; the same product can be manufactured differently every time, forming a cluster structure shown in Figure 1. On the other hand, it is usually most important to predict the rare products scattered around the data space well from the application's point of view.

In the studied prediction model, the response variable used in the regression modelling was the Box-Cox-transformed yield strength of the steel plates. The Box-Cox transformation was selected to produce a Gaussian-distributed error term. Also the deviation in yield strength depended strongly on the input variables. Thus, the studied prediction model included separate link-linear models for both mean and variance

$$
\begin{aligned}
y_i &\sim N(\mu_i, \sigma_i^2) \\
\mu_i &= f(x_i'\beta) \\
\sigma_i &= g(z_i'\tau).
\end{aligned} \tag{4}
$$

The length of the parameter vector of mean model β was 130 and the length of the parameter vector of variance model τ was 30. The input vectors x_i and z_i included 100 carefully chosen non-linear transformations of the 30 original input variables. For example, many of these transformations were products of two or three original inputs. The link functions f and g were power transformations selected to maximize the fit with the data. The results are presented in the original (nontransformed) scale of the response variable [16].

VI. STUDY

The study was started by finding suitable metaparameters for the procedure (see Table I). The data was weighted using gradient-based scaling (weighting relative to the importance of the variables in the regression model) and the amount of training data was decided to be 50,000 observations (approximately one year of data). The rest of the parameters were chosen using prior knowledge of the data and the results of test runs. The distance limit, d, was selected using prior information from the data set. The effect of varying material concentrations was known beforehand and the limit was chosen to correlate with this information. The amount of observations for re-clustering, V, was chosen using a few test runs and the k for k-means was related to the value of V. Naturally, the selection of the metaparameters is highly related to the previous selections. For example, the metaparameters of clustering affect the σ_i and m_i values of equation 2, meaning that also a_1 and a_2 have to be re-scaled. In this study, the values a_1 and a_2 were chosen so that the effect of σ_i and m_i is approximately the same (the distances to the closest cluster centres are ca. 4 times larger compared with the deviations in the clusters).

The goodness of the training data selection was studied by comparing the prediction accuracies of the models trained with the limited data chosen with our approach, with limited data chosen randomly and with the whole data set (reference). The accuracy of the models was measured in the property prediction case using the weighted mean absolute prediction error

$$
\text{MAE} = \frac{1}{\sum_{i=1}^{N} w(i)} \sum_{i=1}^{N} w(i)|y_i - \widehat{\mu}_i|. \tag{5}
$$

In the variation model case, a robustified negative log-

TABLE II

GOODNESS VALUES FOR PROPERTY AND DEVIATION MODELS TRAINED USING WHOLE DATA, DATA SET SELECTED USING CLUSTERING STRUCTURE (CLUST 1-3) AND DATA SET SELECTED USING RANDOM SAMPLING (RS). THE SMALLER VALUES MEAN BETTER RESULTS.

Observations in training data	Goodness 1, MAE	Goodness 2, MAE	Goodness 1, RobL	Goodness 2, RobL
all data	12.90	22.64	6.52	7.49
clust 1	12.90	21.83	6.53	7.45
clust 2	12.87	21.84	6.52	7.48
clust 3	12.92	22.24	6.54	7.59
rs	12.93	22.92	6.53	7.70

likelihood was employed to take into account the variance

$$\text{robL} = \frac{1}{\sum_{i=1}^{N} w(i)} \sum_i w(i) \left(\log(\widehat{\sigma}_i^2) + \rho \left[\frac{(y_i - \widehat{\mu}_i)^2}{\widehat{\sigma}_i^2} \right] \right). \tag{6}$$

Here, the function $\rho(\cdot)$ is a robust function; this study employed

$$\rho(t) = \left\{ \begin{array}{l} t, \text{ when } t \leq 25 \\ b^2, \text{ when } t > 25 \end{array} \right. \tag{7}$$

which truncates the squared standardized residuals if the standardized residual is below -5 or above +5.

Two different methods were used to define the weights, $w(i)$. They were chosen to reflect the usability value of the models. In the first goodness criterion (Goodness 1) the weights $w(i)$ were defined productwise, meaning the weight of the observations of a product could be at most as much as the weight of T observations. Let T_i be the number of observations that belong to the same product as the ith observation. Then the weight of observation i is

$$w(i) = \left\{ \begin{array}{l} 1, \text{ when } T_i \leq T \\ T/T_i, \text{ when } T_i > T. \end{array} \right. \tag{8}$$

Here the value of $T = 50$, meaning if there is more than 50 observation of a product, the weight is scaled down. The second goodness criterion (Goodness 2) was formed to take only rare observations into account. Thus, only the cases for which there were only less than 30 previous observations within a distance 0.9 or a distance 1.8 were included, but the weight of the latter was dual (equation 9).

$$w(i) = \left\{ \begin{array}{ll} 1, & \text{when } \{\#x_j |||x_i - x_j|| < 0.9 \& j < i\} < 30 \\ 2, & \text{when } \{\#x_j |||x_i - x_j|| < 1.8 \& j < i\} < 30 \\ 0, & \text{else} \end{array} \right. \tag{9}$$

Table II shows the result of the training data selection. The models were trained using all the observations (1-200,000, all), and using the reduced amount of observation (50,000, clust and rs), while the prediction accuracy of the models was studied for the remaining observations (ca. 45,000). Because the results varied a little bit during different instance selection runs, the table shows the averages of ten different runs (although the minimum, maximum and standard deviation of the results for goodness 2 are presented in Table III). In addition, the table presents the results of our method for three different cluster structures (clust 1, 2 and 3). They are introduced to show that although k-means clustering does not necessarily find the optimal solution in every run, the use of our approach for instance selection minimizes the problem.

From Table II it can be seen that even when using the means of ten different runs of random sampling, the property prediction ability of the model is weaker compared with the reference model. The difference of this mean is less than two percent weaker than the reference model compared with the rarest cases (goodness 2, MAE). Nevertheless, when looking at the maximum values of Table III, the weakness of this approach becomes clear. In the worst case the prediction accuracy achieved using random sampling can be almost eight percent smaller than in the reference case. Since the results should be consistent in every run, this kind of variation in the results is not acceptable from the application's point of view.

On the other hand, when comparing the prediction accuracies achieved using our approach with the accuracies of the reference model, better results can be seen from Table II. Although for the third k-means clustering the results are little bit weaker measured with other goodness criteria, the result of the goodness criteria 2, MAE is also in this case significantly better using our approach. For the first and second clustering the results are at least as good as the reference, and also with these clusterings our approach clearly shows an improvement for the rarest cases.

From Table III it can be seen that in every individual run the results were better for this goodness measure, meaning the approach is also consistent. Naturally, the results achieved in clust 3 are weaker than in other clusterings, and extra attention is drawn to the weaker results also in rare deviations. However, they are still significantly better than in

TABLE III

BEST AND WORST GOODNESS 2 VALUES AND STANDARD DEVIATION FOR THE PROPERTY MODEL OF DIFFERENT RUNS USING OUR METHOD AND WITH RANDOM SAMPLING

	min min	max max	std std	min min	max max	std std
clust 1	21.48	22.35	0.28	7.37	7.55	0.05
clust 2	21.55	22.15	0.20	7.41	7.54	0.04
clust 3	21.72	22.59	0.28	7.38	7.72	0.10
rs	22.00	24.52	0.79	7.36	8.55	0.37

the random sampling case. To be concrete, the application-specific working limits are a combination of the property and deviation measures, and with two runs of clust 3 there was an actual drop in this limit, but the drop was 0.5 percent while in the best case the improvement was more than 4 percent.

An extra advantage of the method is the ability to generalize at the boundaries of the training data, which is considered an important property of a prediction model in many occasions. The model's ability to predict well for rare steel grades is emphasized also by the engineers who work with the prediction model studied in this application. Therefore, the criteria based only on isolated observations (equation 9)are used to analyze the results. The results measured using these criteria indicate that the proposed instance selection algorithm improves the model's ability to predict the rare steel grades. Hence, it seems the approach can improve prediction at the boundaries of the data. This advantage can be strengthened by decreasing the amount of observations sampled from large clusters, at least in this data set.

VII. CONCLUSIONS

In this article, an approach to selecting training data instances from actual industrial data was introduced. The challenging part of the study was to ensure that the rare observations of the data would be included in the reduced training data set, since the data from the steel manufacturing process are very unevenly distributed. Thus, two different clustering algorithms were utilized and a procedure for selecting the instances using the cluster structure was discovered. The results showed that with this approach the models will be as accurate as in a reference case, and also that the results can be improved a little in the interesting border areas. This was considered beforehand as a very important property when the training data selection is implemented as a part of the maintenance software aiming to automatically keep the prediction models operational in a long run. It should be noted that although in this study the results are compared with the model trained with all the data, in the actual maintenance system the whole data set cannot be used, making this study a necessary step.

The next step of the study is to examine some of the still remaining open questions. The performance of the approach depends highly on the decided metaparameters, although there is no universal way to choose them. This is naturally the case in many existing methods (like selection of the amount of nodes, hidden layers, etc. in neural networks), but some directional values could be found while applying the method to other data sets or to generated data. Another question is if there is a need for selection of k-means clustering. Although

with high probability the results are better than the reference with every clustering, the results could be improved when choosing the best clustering of three different k-means runs, for example. However, also in this case the practical aspects should be recognized, thus needing more consideration.

REFERENCES

[1] W. J. Frawley, G. Piatetsky-Shapiro, and C. J. Matheus, "Knowledge discovery in databases - an overview," *Ai Magazine*, vol. 13, pp. 57–70, 1992.
[2] H. Liu and H. Motoda, "On issues of instance selection," *Data Mining and Knowledge Discovery*, vol. 6, pp. 115–130, 2002.
[3] I. Juutilainen and J. Röning, "Planning of strength margins using joint modelling of mean and dispersion," *Materials and Manufacturing Processes*, vol. 21, pp. 367–373, 2006.
[4] H. Koskimäki, I. Juutilainen, P. Laurinen, and J. Röning, "Detection of the need for a model update in steel manufacturing," *Proceedings of 4th International Conference on Informatics in Control, Automation and Robotics*, pp. 55–59, 2007.
[5] N. Jankowski and M. Grochowski, "Comparison of instances seletion algorithms i. algorithms survey," *Lecture Notes in Computer Science, Artificial Intelligence and Soft Computing*, pp. 598–603, 2004.
[6] H. Brighton and C. Mellish, "Advances in instance selection for instance-based learning algorithms," *Data Mining and Knowledge Discovery*, vol. 6, pp. 153–172, 2002.
[7] ——, "On the consistency of information filters for lazy learning algorithms," *Proceedings of the Third European Conference on Principles of Data Mining and Knowledge Discovery*, pp. 283–288, 1999.
[8] K. jae Kim, "Artificial neural networks with evolutionary instance selection for financial forecasting," *Expert Systems with Applications*, vol. 30, pp. 519–526, 2006.
[9] D. R.Wilson and T. R. Martinez, "Reduction techniques for instance-based learning algorithms," *Machine Learning*, vol. 38, p. 257268, 2000.
[10] H. Altinçay and C. Ergün, "Clustering based under-sampling for improving speaker verification decisions using adaboost," *SSPR/SPR*, pp. 698–706, 2004.
[11] L. Rokach, O. Maimon, and I. Lavi, "Space decomposition in data mining: A clustering approach," *Lecture Notes in Computer Science, Foundations of Intelligent Systems*, vol. 2871, pp. 24–31, 2003.
[12] E. Haapalainen, P. Laurinen, H. Junno, L. Tuovinen, and J. Röning, "Feature selection for identification of spot welding processes." *Proceedings of the 3rd International Conference on Informatics in Control, Automation and Robotics*, pp. 40–46, 2006.
[13] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *The Journal of Machine Learning Research*, vol. 5, pp. 1205–1224, 2004.
[14] C. Bishop, *Pattern Recognition and Machine Learning.* Springer, 2006.
[15] I. Juutilainen, J. Röning, and L. Myllykoski, "Modelling the strength of steel plates using regression analysis and neural networks," *Proceedings of International Conference on Computational Intelligence for Modelling, Control and Automation*, pp. 681–691, 2003.
[16] I. Juutilainen and J. Röning, "A method for measuring distance from a training data set," *Communications in Statistics- Theory and Methods*, vol. 36, no. 14, pp. 2625–2639, 2007.