2008

Data-driven metallurgical design for high strength low alloy (HSLA) steel

Wei Hu
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/etd

Part of the Materials Science and Engineering Commons

Recommended Citation
Hu, Wei, "Data-driven metallurgical design for high strength low alloy (HSLA) steel" (2008). Graduate Theses and Dissertations. 10904. https://lib.dr.iastate.edu/etd/10904

This Thesis is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Data-driven metallurgical design for high strength low alloy (HSLA) steel

By

Wei Hu

A thesis submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of
MASTER OF SCIENCE

Major: Materials Science and Engineering

Program of Study Committee:
Krishna Rajan, Major Professor
James H. Oliver
Alan M. Russell

Iowa State University
Ames, Iowa
2008

Copyright © Wei Hu, 2008. All rights reserved
# Tables of Contents

Abstract................................................................. iii

Introduction................................................................................. 1

Chapter 1 Metallurgical Design of High Strength Low Alloyed Steel.......................... 3
  1.1 High Strength Low Alloyed Steel............................................ 3
  1.2 Multivariate Interaction in HSLA Steel.................................... 3
  1.3 Information Collection for HSLA Steel Design......................... 5

Chapter 2 Data-driven Models for Steel Design................................................. 10
  2.1 Linear Regression Models....................................................... 10
  2.2 Nonlinear Model: Artificial Neural Network............................... 12
  2.3 Data Mining Technology.......................................................... 14

Chapter 3 Methodology: Recursive Partitioning............................................... 16
  3.1 Introduction ........................................................................... 16
  3.2 Recursive Partitioning of Input Space and Tree Growing............... 18
  3.3 Evaluation of Models and Optimization...................................... 21
  3.4 Performance of Recursive Partitioning Models............................ 25

Chapter 4 Results and Discussion.................................................................... 27
  4.1 Yield Strength Model................................................................. 27
  4.2 Ultimate Tensile Strength Model............................................... 37
  4.3 Elongation Rate Model............................................................... 43

Chapter 5 Conclusion and Future Work......................................................... 52

Reference ..................................................................................... 54

Acknowledgement ........................................................................... 59
Abstract

Metallurgical design of microalloyed steel used to be a challenge due to its multivariate nature. Over ten kinds of microalloying elements and multi-step processing routes have complex interactions and different contributions to the final mechanical properties. Data-driven model is able to throw a rapid insight into the composition-processing-property correlation of steel metallurgy in a systematical and efficient way. In this study, a data mining technology, Recursive Partitioning is applied to model the tensile properties of high strength low alloyed (HSLA) steel. The results show that recursive partitioning is able to reveal the complex nonlinear dependence of tensile properties of HSLA steel upon the composition and hot rolling processing parameters. With a relatively simple mathematical structure, Recursive Partitioning can achieve effective performance in predicting the yield strength, ultimate tensile strength, and elongation of steel. In addition, the tree-graph representation of the results provides a powerful multi-dimensional screening tool for searching interesting regions in the composition-processing space, which can be used as a guideline for metallurgical design and further experimental and computational investigation.

Key words: Metallurgical design, HSLA steel, Recursive Partitioning, Tensile properties, Composition-processing-property correlation
Introduction

Metallurgical design of microalloyed steel used to be difficult due to its multivariate nature. Over ten kinds of alloying elements and various processing parameters have complex interactions and different contributions to the final mechanical properties. Although lots of experimental and computational research has been done on HSLA steel, a systematical and efficient strategy to understand the dependence of property upon composition and processing conditions still remains a challenge.

The traditional alloy development strategy consists of producing numerous samples with systematic composition and processing variations to discover the alloy with the optimal properties [1]. This kind of trial-and-error strategy resulted in high cost of experimentation. Instead of performing real experiments, the computational design approach combines physical models of known composition/processing/property interactions to predict the property for a combination of composition and processing [2]. The possible preferential alloys predicted by physical models are then validated through experimental findings. The aid of computational approach, such as thermodynamic, kinetic and first principle calculations significantly reduced the amount of experimental cost [3] [4]. However, there was no well-established physical model that was able to integrate the metallurgical mechanisms with different length scales [5]. Also the increasing of the number of variables considered usually leads to a drastic increase in the complexity of physic models, which makes calculation much more time-consuming and
demanding for computation capacity.

An alternative approach is to develop statistic-based data-driven models to describe the phenomenological patterns in large amount of materials data. Compared to physical models, the advantage of statistical models lies in its ability to survey complex, multi-scale information in a timely and effective manner, even when well-established physic theories and models do not exist [6][7]. Although data-driven models cannot solely provide a satisfactory explanation about the physical mechanism behind the pattern of data, their results can provide a guideline for further experimental and computational investigation [8]. Data-driven model can be used as a screening tool, the output of which becomes the input of the next step of experiment and computational models [9].

This thesis begins with a review of HSLA steel to identify the composition and processing parameters that are influential to the tensile properties. Correspondingly, a set of data was generated and collected for modeling the composition-processing-property relationship of HSLA steel. In chapter 2, two data-driven models that have been applied to steel design will be reviewed and the possibility of applying emerging data mining technology to steel design will be discussed. Chapter 3 will introduce the mechanism of Recursive Partitioning, a data mining technology used in this study. Its performance on modeling data will be evaluated and compared with previous methods. Chapter 4 will present the results of applying recursive partitioning to model tensile properties of HSLA steel, and their significance as guidelines for steel design will be discussed.
Chapter 1 Metallurgical Design of High Strength Low Alloyed Steel

1.1 High Strength Low Alloyed Steel

High strength low alloyed (HSLA) steel specifies a group of steels that contain a low content of carbon and have many more benefits than conventional carbon steels. Generally, it is much higher in strength, along with improved toughness, ductility, weldability, formability, and additional resistance to atmospheric corrosion [10]. Due to these excellent properties, HSLA steels are widely used as structure materials for automobiles, bridges, oil storage and structure which are designed to carry high stress [11][12].

HSLA steel usually contains a much lower content of carbon than regular carbon steels, typically below 0.15%. It is often referred to microalloyed steel since its excellent properties usually come from various alloying elements in extremely small amounts compared to conventional steels [10]. Besides carbon, manganese and silicon, over ten kinds of other elements, such as chromium, nickel, niobium, boron, molybdenum, titanium and copper, can be added to steels to achieve extra strength through various strengthening mechanisms.

1.2 Multivariate Interaction in HSLA Steel.

When alloying elements are added into the steel, their interactions with iron and carbon result in various strengthening effects. Solved in iron lattice as substitutional or interstitial
atoms, they strengthen the steel by solid-solution hardening. Combining with carbon to form various fine and disperse carbide precipitates, alloying elements strengthen the steel by precipitation hardening. Complex interactions also exist between alloying elements. For example, adding one element may influence the diffusion of existing elements and therefore influence the precipitation or segregation of existing elements [13]. Several elements combine together to form carbide compound and then enhance the precipitation [14].

HSLA steels are typically strengthened by precipitation strengthening, grain refinement, and, to a less extent, solid-solution strengthening due to the small amounts of alloying elements [14]. All of these strengthening effects are sensitive to the processing environment, such as temperature, time, deformation extent and cooling rate. For example, temperature controls the solubility and diffusions of microalloying elements, and therefore greatly influences the formation of precipitates. Since grain growth is a thermal activated process, the grain size is strongly dependent on the both temperature and time during steel processing.

The challenge to design HSLA is to systematically understand the complex composition-processing-property relationship. Not only does each element strengthen steel through more than one mechanism, but sometimes two or three elements can combine together to influence the strength. In addition, the realization of strengthening effects of these elements strongly relies on the processing conditions. Coupled effects and
intensive interaction of alloying elements and processing routes make metallurgy of HSLA steel very complicated to understand. Microalloyed steel metallurgy is a typical multidimensional problem, as shown in Figure 1.1.

Figure 1.1 Multidimensional nature of microalloyed steel metallurgy. Various composition and processing parameters have complex interactions and different contributions to the final mechanical property.

1.3 Information Collection for HSLA Steel Design

The mechanical properties of HSLA steel depend on the composition and processing conditions. Therefore it is of practical importance to understand the composition-processing-property correlation, which would allow a quick evaluation for the alternation of composition and processing route. It also helps to optimize the composition and processing parameters to achieve desired properties. The tensile properties are the essential properties of HSLA steel. Therefore, this study aims at modeling the tensile properties of HSLA steels as a function of composition and processing variables.
As the first step, proper selection of descriptors is a key issue before model building. Three major measurements of tensile property, yield strength, ultimate tensile strength and percentage of elongation, are taken into consideration in this study. The information to describe the composition of alloy is relatively simple, as it can be represented by the weight percentage of elements in the steel. In this study, carbon and nine common microalloying elements in HSLA are considered (manganese, silicon, nickel, copper, molybdenum, niobium, chromium, titanium, boron). On the other hand, the selection of information to describe the processing routes of HSLA steel is not trivial, since the manufacture of steels is a complex process and consists of a large amount of steps. Including the most influential parameters is necessary for reducing the noise and simplification of models.

Most of HSLA steels are manufactured by hot rolling processing. As shown in Figure 1.2, a typical hot rolling process includes four stages [14]:

![Typical hot rolling processes](image)

**Figure 1.2 Typical hot rolling processes**

**A. Reheating**

Steel slabs are reheated up to 1100-1200°C homogeneously before hot rolling so that the
austenite is fully annealed with a low defect density. The slab reheating temperature (SRT) usually affects the dissolution of microalloying elements and particles precipitated in previous processing. If some precipitates remain undissolved during reheating, they can inhibit the austenite grain growth by Zener Pinning.

**B. Hot Rolling**

Usually the hot rolling process consists of two steps: rough rolling to roll the steel to intermediate thickness and finishing rolling to roll the steel to the final thickness. The deformation rate during rolling and the finish rolling temperature (FRT) control the complex recrystallization behavior and microstructure evolution during hot rolling.

**C. Accelerated cooling**

Accelerated cooling after hot rolling help to maintain the austenite structure developed in the last process by drastically limiting the time in the austenite temperature range. In this way, the decomposition of austenite can be controlled to occur after coiling. Thus cooling speed, that is, the cooling rate (CR) is the control factor in this step.

In sum, processing parameters that greatly influence the microstructure evolution and final properties of HSLA steel should be included. Four processing parameters consisting of slab reheating temperature (SRT), deformation percentage during the recrystallization temperature region (D1), finish rolling temperature (FRT), and cooling rate of the accelerated cooling, are included in this study.
The data used in this study were mostly generated in the laboratory, which came from the group of Prof. Datta at Bengal Science and Engineering University [15]. The chemical analysis was performed in atomic spectrometer. Hot rolling was carried out in a laboratory scale rolling mill. The tensile testing has been carried out in INSTRON 4204 machine. Some data from the published literature were also taken into consideration to get a wide range of variation [15]. Totally 121 samples were collected. The statistical summery of given data including the range, mean and standard deviation of each input and output variable can be found in Table 1.1. Since cooling rate (CR) only contains two values 5 and 95, it was treated as a categorical variable.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Symbol</th>
<th>Min</th>
<th>Max</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carbon (wt %)</td>
<td>C</td>
<td>0.024</td>
<td>0.077</td>
<td>0.046</td>
<td>0.011</td>
</tr>
<tr>
<td>Manganese (wt %)</td>
<td>Mn</td>
<td>0.5</td>
<td>1.72</td>
<td>1.38</td>
<td>0.25</td>
</tr>
<tr>
<td>Silicon (wt %)</td>
<td>Si</td>
<td>0.019</td>
<td>0.58</td>
<td>0.405</td>
<td>0.107</td>
</tr>
<tr>
<td>Chromium (wt %)</td>
<td>Cr</td>
<td>0</td>
<td>0.85</td>
<td>0.32</td>
<td>0.37</td>
</tr>
<tr>
<td>Nickel (wt %)</td>
<td>Ni</td>
<td>0</td>
<td>1.9</td>
<td>0.68</td>
<td>0.65</td>
</tr>
<tr>
<td>Molybdenum (wt %)</td>
<td>Mo</td>
<td>0</td>
<td>0.61</td>
<td>0.24</td>
<td>0.25</td>
</tr>
<tr>
<td>Titanium (wt %)</td>
<td>Ti</td>
<td>0</td>
<td>0.08</td>
<td>0.02</td>
<td>0.020</td>
</tr>
<tr>
<td>Boron (wt %)</td>
<td>B</td>
<td>0</td>
<td>0.0025</td>
<td>0.0013</td>
<td>0.0008</td>
</tr>
<tr>
<td>Niobium (wt %)</td>
<td>Nb</td>
<td>0</td>
<td>0.06</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>Copper (wt%)</td>
<td>Cu</td>
<td>0</td>
<td>2.17</td>
<td>1.00</td>
<td>0.70</td>
</tr>
<tr>
<td>Deformation above recrystallization temperature (%)</td>
<td>D1</td>
<td>48</td>
<td>76</td>
<td>57.8</td>
<td>9.6</td>
</tr>
<tr>
<td>Cooling Rate (°C/s)</td>
<td>CR</td>
<td>5</td>
<td>95</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Variable</td>
<td>Symbol</td>
<td>Min</td>
<td>Max</td>
<td>Mean</td>
<td>Standard Deviation</td>
</tr>
<tr>
<td>----------------------------------------</td>
<td>--------</td>
<td>-----</td>
<td>------</td>
<td>--------</td>
<td>--------------------</td>
</tr>
<tr>
<td>Slab Reheating Temperature (°C)</td>
<td>SRT</td>
<td>1050</td>
<td>1200</td>
<td>1164.5</td>
<td>50.4</td>
</tr>
<tr>
<td>Finish Rolling Temperature (°C)</td>
<td>FRT</td>
<td>725</td>
<td>954</td>
<td>768.0</td>
<td>37.9</td>
</tr>
<tr>
<td>Yield Strength (MPa)</td>
<td>YS</td>
<td>247</td>
<td>1067</td>
<td>662.0</td>
<td>195.3</td>
</tr>
<tr>
<td>Ultimate Tensile Strength (MPa)</td>
<td>UTS</td>
<td>430</td>
<td>1193</td>
<td>830.5</td>
<td>176.8</td>
</tr>
<tr>
<td>Elongation Rate (%)</td>
<td>EL</td>
<td>11</td>
<td>41.3</td>
<td>23.1</td>
<td>5.9</td>
</tr>
</tbody>
</table>
Chapter 2 Data-driven Models for Steel Design

During the last several decades, a number of scientists tried to develop statistical models to predict the mechanical properties of steels as a function of composition and processing. Linear Regression Models have been successfully used to model mechanical properties of steels, providing a simple mathematical description of steels metallurgy [14][16]. However, the dependence of mechanical properties upon composition and processing usually are complex and nonlinear. To solve this problem, nonlinear models, such as artificial neural networks (ANN), have been widely applied to estimate the mechanical properties and microstructural evolution of steels [17].

In the following sections, the linear regression and artificial neural networks models will be reviewed in terms of their advantages and disadvantages. Then the objective of this study, to use a data mining technology Recursive Partitioning (RP) to model the composition-processing-property relationship, will be addressed.

2.1 Linear Regression Models

Linear regression is a form of regression analysis in which data are modeled as a linear combination of model parameters. Model parameters are determined by minimizing the sum of squared residuals of the model when fitting data. As a basic and widely used statistical method, linear regression has been widely used to model multivariable relationship in steel design for several decades [18] [19] [20].
For example, for high strength low alloyed steel, there are four empirical equations [15] to predict the yield stress $\sigma_y$, uniform strain $\varepsilon_u$, work hardening rate $d\sigma / d\varepsilon$, and strain at fracture $\varepsilon_{fr}$ as a linear function of composition and microstructure:

$$\sigma_y = 15.4(16 + 0.27 \text{perl} + 2.9Mn + 9Si + 60P + 11Sn + 244N + 0.97d^{-1/2})$$  \hspace{1cm} (2-1)

$$\varepsilon_u = 0.27 - 0.016 \text{perl} - 0.015Mn - 0.040Si - 0.043Sn - 1.1N$$  \hspace{1cm} (2-2)

$$d\sigma / d\varepsilon = 385 + 1.39 \text{perl} + 111Si + 462P + 152Sn + 1369N + 15.4d^{-1/2}$$  \hspace{1cm} (2-3)

$$\varepsilon_{fr} = 1.30 - 0.02 \text{perl} + 0.3Mn + 0.2Si - 4.4P + 2.9Sn + 0.015d^{-1/2} - 3.4S$$  \hspace{1cm} (2-4)

Where $\text{perl}$ is the volume fraction of pearlite, $d$ is the ferrite grain size in $\mu$m, and alloy content is in weight percentage.

Based on linear equations, the effect of a particular element on the tensile properties of steel can be evaluated based on the corresponding parameter. Figure 2.1 shows the effect of element content on the changing of yield stress of low carbon steels, which were calculated based on Equation (2-1). It should be noticed here, the solubility of elements such as N, P are much small so that the strengthening from these elements cannot be as significant as shown in the figure 2.1.

The advantage of linear regression is that the model can be represented as an equation with simple structure, which can be used for qualitative or quantitative evaluations. However, sometimes linear models are too simple to describe the nonlinear relationship, which is common in the complex system of steel metallurgy [17]. Sometimes it leads to ultra high error and uncertainty in prediction. Consequently, nonlinear models are needed.
Fig. 2.1 Effect of element content on yield strength based on the equation (2-1)

2.2 Nonlinear Model: Artificial Neural Network

Artificial Neural Network (ANN) is a learning algorithm, which was originally developed by psychologists and neurobiologists who sought to model and understand biological neural networks [21]. ANN model is made up of interconnected input/output units (neurons), in which each connection is assigned a weight. Usually a network includes one layer of input nodes, one layer of output nodes and several layers of hidden nodes, as shown in Figure 2.2. During learning from training data, the network adjusts the weights so that the model can perform better in predicting the correct class or values of output of training data[22][23][24].
During the last decade, artificial neural networks have been widely applied to model the correlation of composition, processing, structure, properties and performance of steel and related products [25]. It has been proven that this method can efficiently simulate the intricate nonlinear relationship in the field of material science [17].

In order to model the nonlinear relationship, neural network usually use a flexible nonlinear function. For example, the final output could be [21]:

\[ y = \sum_{i} w_i O_i + \theta \]  \hspace{1cm} (2-5)

\[ O_i = \tanh(\sum_{j} w_{ij} x_j + \theta_j) \]  \hspace{1cm} (2-6)

where \( y \) is the final output, \( O_i \) is the \( i \)th output node, \( w_i \) is the weight of \( i \)th output node for final output, \( w_{ij} \) is the weight of \( j \)th hidden node for \( i \)th output node, and \( \theta, \theta_j \) are the biases for output layer and hidden layer respectively.
Flexible nonlinear functions allow neural network models to achieve excellent performance on learning data. On the other hand, complex models are hard to interpret and use compared to linear equations. Neural networks are often referred to as a “black box” approach of modeling, since the output value is generated without gaining much understanding of the interactions between the variables and therefore providing no insight into the underlying mechanism behind the established model [26]. Also using artificial neural networks involves a long training time and is demanding for computation capacity with the large amount of data.

2.3 Data Mining Technology

As a nonlinear method, artificial neural networks can achieve high accuracy in fitting the given data. However, the complexity of the model impedes the further interpretation and application of neural networks model.

An important criterion for a good prediction method is that it not only produces accurate predictions within the limits of the given data, but it also provides insight and understanding into the predictive structure of the data [27]. A statistical model which is not only able to describe the nonlinear relationship in the microalloying steel metallurgy, but also reflect comprehensive information within the data, would be the preferred method for the data driven modeling of steel metallurgy.

By merging ideas of statistics and computer science, data mining technologies provide a
variety of possible solutions to analyze the materials data [6][7]. Among these technologies, recursive partitioning is a good algorithm that is able to describe the nonlinear relationship with a relatively simple mathematical structure. What is more important, recursive partitioning provides a tree-graph representation of models which is easy to understand and use as a guideline.

While recursive partitioning has been extensively applied to many areas, its application to materials science is limited. Sturrock et al. compared recursive partitioning with four other data analysis technique in studying a two-classification problem on stress corrosion cracking of austenitic stainless steels and concluded that recursive partitioning method provided the best performance in terms of both classification accuracy and intelligibility of the output [28]. Yi used recursive partitioning as a regression method to predict the creep rupture life of austenitic stainless steels and to evaluate the relative importance of composition, microstructure and environment parameter [29].

In this study, the recursive partitioning method was used to model the correlation between composition, processing, and tensile properties of HSLA steel on the experimental data. How to evaluate and optimize the recursive partitioning model were discussed. The comprehensive information about composition-processing-property interaction revealed by tree graph was also discussed and could be used as guidelines for HSLA steel design.
Chapter 3 Methodology: Recursive Partitioning

3.1 Introduction

Recursive partitioning is a data analysis technique which can be applied to explore large amount of datasets with complex structure in order to uncover underlying patterns within the data and to divide the data into statistically significant sub-groups [27]. This technique is especially suitable for the situations in which conventional statistical techniques, such as linear regression, can not function well. Recursive Partitioning has become one of the most flexible, intuitive and powerful data analytical tools, which has been widely used in many application areas, including business [30], manufacturing [31], marketing [32], clinical medicine [33] and scientific research [34].

Generally speaking, recursive partitioning is a tool for relating a dependent variable ($Y$) to a number of independent variables ($X_i$) in order to uncover the trend of $Y$ as a function of $X_i$ [21]. The basic idea is to recursively partition the data, usually dichotomously, into different subsets based on their similarity. The essential result of recursive partitioning is a tree-like structure, which is usually called a “decision tree”. Data are partitioned into nodes (leaves of the tree) along the branches of the tree. Data which are more similar according to some criteria tend to be assigned to the same branch or leaf, while more dissimilar data tend to occupy different branches or leaves, as shown in Figure 3.1. The statistical significance of splitting the data into a tree, which can be designed and flexibly chosen according to the requirement of users, lies on some quantitative mathematical
principles [27].

Figure 3.1 An illustration of tree growing through recursive partitioning dealing with classification problem [35]

The class or value of each node is determined statistically based on all the samples the node contains. Usually, for categorical output, the class of a node is the same as the most frequent class in the subset; for numerical output, the value of a node is the average of output of the subset. After the model is built on the training data, it can be used to predict the class or value of output-variable for new data. When a sample is assigned to a leaf along the branches of the tree, the output is predicted as the class or value of that leaf.

Compared to other statistical algorithm, the advantages of recursive partitioning are:

A. The tree representation of results is highly interpretable [36]. The set of criteria from root to leaf along branches provides an intuitive guideline which is generally easy to assimilate by human. For example, Figure 3.2 is a recursive partitioning model that
explores the dependence of stress corrosion cracking upon the aqueous environment. From the lowest leaf, we can learn that: if the concentration of oxygen $> 0.3 \text{mg/L}$ and concentration of chlorides $> 1.5 \text{mg/L}$, stress corrosion cracking would take place [28].

B. Powerful. Able to handle large and high dimensional data and predict nonlinear relationship.

C. Flexible with data type. It can deal with both categorical and numerical data. Many data are not complete and include missing values; and recursive partitioning has many strategies to deal with these values [37].

D. Efficient. The learning process of tree induction is fast and simple

E. Accurate. Research shows that with a simple structure recursive partitioning can be much accurate than other complex nonlinear method. However, successful use may depend on the data [21].

![Decision tree](image)

Figure 3.2 Decision tree about dependence of stress corrosion cracking (SCC) of austenitic stainless steels upon the aqueous environments [28].

3.2 Recursive Partitioning of Input Space and Tree Growing

Depending upon the output variable types, recursive partitioning can be divided into two
kinds of tree: classification trees for categorical dependent variables and regression tree for numerical dependent variables. At each step of tree splitting, the algorithm will search and investigate all the split possibilities for the best split point. To point out the split, the algorithm calculate a kind of measurement of error for the current tree and potential partitioning, and all the possible splits are ranked based on their contribution to reduction of error to fit data. Generally, this kind of measurement, in the classification is information entropy, and in regression tree is the sum of deviation [21]. Then the best split will be chosen for the current step of tree growing.

### 3.2.1 Classification Rules Based on Reduction of Information entropy

During the classification processing, a specific variable and value are selected to maximize the reduction of information entropy (H), defined as information gain. The entropy concept in the information theory was introduced by C.E. Shannon [38], as a measurement of uncertainty associated with a given system.

For example, as shown in figure 2.3, node $D$ can be partitioned into two branches $D_1$ and $D_2$ by criterion A. The Shannon information entropy of node D is defined as [21]:

$$Info(D) = - \sum_{i=1}^{m} p_i \log_2 (p_i)$$

(3-1)

where $p_i$ is the probability that a data point in dataset $D$ belongs to class $C_i$. Then the information gain of splitting $D$ into $D_1$ and $D_2$ by criterion $A$ is defined as:

$$Gain(A) = Info(D) - (Info(D_1) + Info(D_2))$$

(3-2)
Information gains of all the potential partitioning are calculated and then ranked. The best split that leads to maximum information gain will be selected for tree growing.

### 3.2.2 Regression Rules Based on Deviation Reduction

The basic process of tree splitting in regression tree is almost the same as classification, except the splitting rules. While classification trees select the best split to maximize information gain, regression trees choose to minimize the deviation between predicted and actual values.

Usually the deviation between prediction and actual data is measured by residual sum of squares (RSS), which is used in this study. The $RSS$ for a node $D$ can be expressed as[21]

$$RSS(D) = \sum_{i} (y_i - y'_i)^2$$  \hspace{1cm} (3-3)
where $y'_i$ is the predictions of the output value by current node and $y_i$ is the actual value.

Therefore the deviation reduction by splitting node $D$ into two nodes $D1$ and $D2$ according to a criterion $A$ is,

$$\Delta RSS(A) = RSS(D) - (RSS(D1) + RSS(D2)) \quad (3-4)$$

In this way, the tree keeps growing in a way that the overall deviation of the prediction from the actual data keeps decreasing.

### 3.3 Evaluation of Models and Optimization

When the node contains only samples with uniform class or has a deviation lower than the algorithm threshold, it stops splitting. When recursive partitioning is complete and a decision tree fully grows, many of the branches may reflect the anomalies in the training data due to noise or outliers, which leads to the problem of overfitting the data [21].

Overfitting is a statistical concept that a model is too complex and has too many parameters. This kind of model may fit perfectly to the training data due to its high complexity. However, it may turn out to be false or absurd when applying to new data and lead to high error rate. As shown in Figure, the optimal solution can be found when the model complexity matches the system complexity.

In order to avoid overfitting, it is necessary to use additional techniques to evaluate the performance upon overfitting and to simplify the model to optimize its function.
Figure 3.4 Overfitting occurs when the model is too complex. A point need to be found where the model complexity matches the system complexity [37]

In this study, accuracy estimation method cross-validation is used to judge the performance of a decision tree, which can effectively detect overfitting. Also pruning of the tree is used for simplifying the built recursive partitioning model.

3.3.1 Cross-validation to Prevent Overfitting

In order to obtain a reliable estimate of prediction accuracy, many techniques are used for accessing accuracy based on randomly sampled partitions of the given data [21]. Generally the given data are randomly divided into two independent datasets, a training set and a test set. The learning model is built on the training set and then evaluated on the test set. The use of such techniques increases the overall computation time yet is useful for model selection.

There are two main kinds of the cross-validation method: $k$-fold cross validation and leave-one-out.
In $k$-fold cross-validation, the initial data are randomly divided into $k$ mutually exclusive subsets or “folds” with approximately equal size, $D_1, D_2 \ldots D_k$. The learning and testing is performed $k$ times. In the iteration $i$, partition $D_i$ is reserved as the test data, and the remaining folds are used to train the model. After $k$ iterations, the overall accuracy of the model can be evaluated based on the average of the error rates of all the $k$ folds [37]. The process of a typical 10-fold cross-validation is shown in Fig. 3.5.

![Diagram of 10-fold cross-validation](image)

**Figure 3.5 Illustration of 10 folder cross-validation.** The original data is randomly divided into ten folds with approximately equal size. Each time one fold is reserved as test data. Models are built on 90% of data and then tested on 10% of data. After ten iterations, the overall accuracy is evaluated by the average of accuracies tested on all the iterations.
Leave-one-out is a special case of cross validation where only one sample is used as validation data and the remaining samples are used for training set. The process is repeated until each sample is used once for the validation. That is, it is a $k$-fold cross validation where $k$ is equal to the number of samples in the data [37].

In General, 10-fold cross validation is recommended to evaluate the model due to its relative low bias and variance. In this thesis, 10-fold cross-validation is used for evaluating the recursive partitioning model.

3.3.2 Optimization of Recursive Partitioning Model

In order to address the problem of overfitting, decision tree model need to be simplified, that is, to be pruned. Tree pruning methods usually use statistical measures to remove the least reliable branches. Usually after pruning, the simpler tree become faster and better at correctly predicting independent test data than the unpruned tree.

There are two approaches for tree pruning: prepruning and postpruning [37]. In the prepruning approach, a tree is pruned by halting its construction early according to some statistical measures. The rules used in splitting nodes can be used as measures to estimate the goodness of a split. For example, in the classification tree, if the information gain of a split falls below a pre-set threshold, further partitioning of the given subset is ceased. An appropriate threshold needs to be carefully chosen. High threshold could lead to oversimplified trees, while a low threshold could result in very little simplification.
The other approach is postpruning, which removes unreliable sub trees from a fully grown tree. A subtree is pruned at a given node by removing its branches and replacing it with a leaf.

3.4 Performance of Recursive Partitioning Models

In this thesis, recursive partitioning is used for predicting the tensile properties of microalloyed steels. The performance of recursive partition regression, linear regression and artificial neural network are evaluated using 10-folds cross-validation on the data of ultimate tensile strength of transformation induced plasticity (TRIP) steel. The accuracies in terms of $R^2$ are shown in the Table 3.1 and the corresponding predicted-actual plots are shown in the Figure 3.6.

<table>
<thead>
<tr>
<th></th>
<th>Linear Regression</th>
<th>Regression Tree</th>
<th>ANN</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$</td>
<td>0.8415</td>
<td>0.9174</td>
<td>0.9273</td>
</tr>
</tbody>
</table>

Based on the comparison, it can be concluded that for given tensile strength data, the recursive partitioning is much better than linear regression and possible to achieve the competing accuracy as artificial neural network by a much simpler structure. In addition, the tree model is much more interpretable than the “black box” core of neural network.

With a good combination of accuracy and interpretability, recursive partitioning could be
a good candidate to model the complex composition-processing-property interaction of HSLA steel.

Figure 3.6 Actual-Predicted Plots of the results from different regression models

a) Linear Regression   b) Regression Tree   c) Neural Network
Chapter 4 Results and Discussion

In this section, recursive partitioning is applied to model the quantitative relationships between compositions, processing and tensile properties of high strength low alloy steels. The yield strength, ultimate tensile strength and elongation rate will be modeled separately on the collected data described in Chapter 1.

4.1 Yield Strength Model

All of 121 samples are used to train the regression model by recursive partitioning. The built regression tree is evaluated using 10 folds cross-validation and then pruned to obtain the optimized tree. The minimum number of samples in a leaf is set as five, since a smaller number would lead to a sharp decrease in accuracy during tree growing, which indicated the model was unstable.

The accuracy of the tree model is measured by the coefficient of determination $R^2$, which is expressed in Equation 4-1.[29]

$$ R^2 = 1 - \frac{\sum (Y_i - \hat{Y}_i)^2}{\sum (Y_i - \bar{Y})^2} $$

(4-1)

where $Y_i$ is the measured value of output variable of $i$th sample, $\hat{Y}_i$ is the predicted value, $\bar{Y}$ is the average of the output variable of whole samples. The value of $R^2$ ranges from [0, 1] and a value of 1 indicates a perfect prediction.
Figure 4.1 shows that the overall $R^2$ of the yield strength model, which was calculated including all the data for training, keeps increasing while splitting. It indicates that the regression tree model grows in a way so that the model keeps improving to have better performance to fit the given data. If the tree grows fully, there are 18 total splits and the highest $R^2$ is 0.9416.

Figure 4.1 Split history of coefficient of determination of yield strength model

However, overall $R^2$ can not be used directly for evaluating the model, which may lead to overfitting. Using 10-folds cross-validation, the performance of models with different split number was evaluated and the optimal number of splits was determined to be 15, as shown in Figure 4.2. In this way, the full grown tree is pruned and the optimal tree is selected. The $R^2$ of the optimal model is 0.9281. The actual-predicted plot of the optimal pruned tree is shown in figure 4.3.

Figure 4.4 shows a snapshot of a portion of the optimal regression tree. In each node box,
it displays four kinds of information: (1) the split criteria by which samples in the parent node are partitioned (2) the number of samples assigned to this node (3) the mean value of yield stress.

Figure 4.2 Change of coefficient of determination with increase in number of splits in YS model evaluated by 10 fold cross-validation

Figure 4.3 Actual-predicted plots of yield strength model

The decision tree model is very easy to interpret. For each leaf, a set of predictive rules
can be obtained by tracing the pathway from the root to the leaf and combining all the split criteria along it. For example, the pathway for the far right leaf in Fig. 4.4 can be interpreted as a set of if-then rules, as shown in Table 4.1. All the rules deduced from the regression tree can make up a design map to predict the yield strength of steels. A list of rules and corresponding support samples are shown in Table 4.2. It should be noticed that these rules would be valid only within the particular input space described in Table 1.1. Apparently, if the targeted design objective is high strength, the pathway “Mo>=0.32&C>=0.057&Nb>=0.06” should be taken to achieve a desired yield strength over 1000 MPa. It also can be found that tree model greatly reduce the dimensionality of original data by selecting at most 7 out of 14 attributes for prediction.

Figure 4.4 A portion of the regression tree for the yield strength of HSLA steel
Table 4.1 An example of the predictive rules for leaf node.

If

\[(\text{Content of Mo}<0.32\% \&\& \text{Cooling Rate } = 5^\circ\text{C}/\text{s } \&\& \text{Content of Ni}<0.79\% \&\& \text{Content of Cu}<1.19\% \&\& \text{Finish Rolling Temperature} \geq 775^\circ\text{C})]\)

Then

The average yield strength is 306.3MPa

Table 4.2 Brief description of deduced rules from YS model

<table>
<thead>
<tr>
<th>Pathway from root to leaf</th>
<th>Predicted YS (MPa)</th>
<th>Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mo&lt;0.32&amp;CR(5)&amp;Ni&lt;0.79&amp;Cu&lt;1.19&amp;FRT&gt;=775</td>
<td>306.3</td>
<td>6</td>
</tr>
<tr>
<td>Mo&lt;0.32&amp;CR(5)&amp;Ni&lt;0.79&amp;Cu&lt;1.19&amp;FRT&lt;775</td>
<td>343.8</td>
<td>9</td>
</tr>
<tr>
<td>Mo&lt;0.32&amp;CR(5)&amp;Ni&lt;0.79&amp;Cu&gt;=1.19</td>
<td>411.8</td>
<td>14</td>
</tr>
<tr>
<td>Mo&lt;0.32&amp;CR(5)&amp;Ni&gt;=0.79</td>
<td>539.0</td>
<td>6</td>
</tr>
<tr>
<td>Mo&lt;0.32&amp;CR(95)&amp;B&lt;0.0011</td>
<td>682.0</td>
<td>9</td>
</tr>
<tr>
<td>Mo&lt;0.32&amp;CR(95)&amp;B&gt;=0.0011&amp;Cu&lt;1.65</td>
<td>786.1</td>
<td>14</td>
</tr>
<tr>
<td>Mo&lt;0.32&amp;CR(95)&amp;B&gt;=0.0011&amp;Cu&gt;=1.65</td>
<td>845.4</td>
<td>6</td>
</tr>
<tr>
<td>Mo&gt;=0.32&amp;C&lt;0.057&amp;D1&gt;=70</td>
<td>621.2</td>
<td>6</td>
</tr>
<tr>
<td>Mo&gt;=0.32&amp;C&lt;0.057&amp;D1&lt;70&amp;Cu&lt;1.59&amp;D1&lt;67&amp;S</td>
<td>RT&lt;1100</td>
<td>695.8</td>
</tr>
<tr>
<td>Mo&gt;=0.32&amp;C&lt;0.057&amp;D1&lt;70&amp;Cu&lt;1.59&amp;D1&lt;67&amp;S</td>
<td>RT&gt;=1100&amp;D1&lt;65</td>
<td>703.5</td>
</tr>
<tr>
<td>Mo&gt;=0.32&amp;C&lt;0.057&amp;D1&lt;70&amp;Cu&lt;1.59&amp;D1&lt;67&amp;S</td>
<td>RT&gt;=1100&amp;D1&gt;=65</td>
<td>735.4</td>
</tr>
<tr>
<td>Mo&gt;=0.32&amp;C&lt;0.057&amp;D1&lt;70&amp;Cu&lt;1.59&amp;D1&gt;=67</td>
<td>759.8</td>
<td>7</td>
</tr>
<tr>
<td>Mo&gt;=0.32&amp;C&lt;0.057&amp;D1&lt;70&amp;Cu&gt;=1.59</td>
<td>839.0</td>
<td>5</td>
</tr>
<tr>
<td>Mo&gt;=0.32&amp;C&lt;0.057&amp;Nb&lt;0.06&amp;D1&gt;=67</td>
<td>808.2</td>
<td>5</td>
</tr>
<tr>
<td>Mo&gt;=0.32&amp;C&lt;0.057&amp;Nb&lt;0.06&amp;D1&lt;67</td>
<td>862.5</td>
<td>6</td>
</tr>
<tr>
<td>Mo&gt;=0.32&amp;C&lt;0.057&amp;Nb&gt;=0.06</td>
<td>1006.2</td>
<td>5</td>
</tr>
</tbody>
</table>
An important way to interpret the prediction model is to evaluate the relative importance of the input variables. For recursive partitioning regression, there are two technological parameters that can be used to rank the importance of variables. The first is the contribution of a variable to the reduction of deviation when it serves as splitting criteria, as shown in Figure 4.5; the second is the time that a variable is used as the splitting criteria, as shown in Figure 4.6.

![Figure 4.5 Contribution of input variables to error reduction in YS model](image1)

![Figure 4.6 Time of input variables as split criteria in YS model](image2)
In the given data, there are 14 input variables. However, in the recursive partitioning model, not all the variables are used. That is, after the training process, the model turned out to reduce the dimension of data. As shown in Figure 4.6, elements Mn, Si, Cr and Ti were not used to split the tree, which indicates that the model automatically decide that the contribution of these four elements to the change of yield stress are so small in given data that they can be neglected.

Figure 4.5 indicates that elements Mo and C and processing parameter cooling rate (CR) are the three input variables which make the largest contributions to the error reduction of model. An important reason is that Mo, C, and CR were selected as split criteria from the very first stage of tree growing, as shown in the Figure 4.7.

The content of Mo (>0.32 or <=0.32) is used in the highest branch, and made the second largest contribution to the error reduction of model. This criterion divided the whole
samples into higher yield strength group (average: 771.0 MPa) and lower yield strength group (average: 564.8 MPa). From the data distribution of molybdenum, it can be found that most samples which fulfill the criterion Mo<0.32% contain no molybdenum. It indicates that addition of molybdenum will great increase the yield strength and plays an important role in the initial development of HSLA steel within the range of the given data.

It is well-known that among the substitutional elements, the molybdenum exerts the largest strengthening effect upon HSLA steel through solid-solution strengthening and microstructure modifications. Molybdenum produces a fine grain structure of acicular ferrite and substantially enhances the precipitation hardening effects achieved with the other alloying elements [39].

It is shown in the Table 4.2 that the group of steel with highest yield strength was only assigned by combination of three elements Mo, C and Nb. This deducted rule is consistent with the strong interaction between Mo, Nb and C and their important roles in strengthening. In HSLA steels the influence of niobium microalloying is very important for formation of phase structure and a fine microstructure of steels during controlled rolling and subsequent accelerated cooling of steel [40]. Both Mo and Nb are strong carbide former, forming strong precipitation strengthening to the steel. Meanwhile, it was reported that HSLA steel containing both of Mo and Nb exhibited superior strength to the conventional HSLA steels containing Nb and V. [41][42]. It was reported that Mo decreased the diffusivity of carbide forming species including Nb and C, and thus delay the precipitation of MC carbide [43][44]. With less precipitation in austenite, more
precipitates could form in ferrite resulting in enhanced strength. As the content of Mo and Nb increases, volume fraction of precipitates increases, resulting in higher strength [45]. The regression tree model indicates that with the given group of steel, the precipitation hardening of Mo and Nb carbide and the interaction between these elements make the most influential contribution to the strength of HSLA steel. It is possible to reduce the content of other elements to the average level in order to reduce the cost while maintaining the strength of steel to the high level.

Processing condition CR (Cooling Rate) (95 or 5) made the largest contribution to the error reduction of the tree regression model, though it was only used once in the second layer of tree. This criterion further divides the lower yield strength group (average: 564.8 MPa) into low yield strength group (average: 398.0MPa) and middle yield strength group (average: 766.1MPa). Higher cooling rate leads to a greater supercooling and results in fine ferrite structure [46].

Similarly, the content of C distinguished middle yield strength steel (average: 724.4MPa) and very high yield strength group (average: 890.4MPa). Carbon is the primary strengthening element in steel. Increase in Carbon content will lead to improvement of yield strength.

Processing condition D1 (Deformation extent at Recrystallization Temperature) and Cu content were used more than once in the tree split. However their contributions to error
reduction were much lower compared to Mo, CR and C. They were used in the lower branches of tree. These parameters are important factors to enhance yield strength of steel, but to a relatively low content, within the yield strength range which is determined by Mo, C and cooling rate.

Content of Ni, B, Nb and the processing parameters SRT and FRT were used only one time, and their contributions to the model are relatively small.

Basically, we can found that as the content of microalloying element increases, the strength of steel increase. It is consistent with the fact that the effect of solid-solution precipitation strengthening enhance with the higher content of elements. However, the trend for the processing parameters is a little more complicate. For example, as shown in the Figure 4.8, the choice of suitable extent of deformation during hot rolling (D1) seems delicate and strongly depends on other composition and processing parameters, such as C, Cu and SRT. These rules generated by regression can be good start points to optimize the deformation content in given composition region.

In sum, Mo, C and CR are determinative factors in predicting yield strength that primarily classify the steels in given data into three strength regions: the higher strength, mild strength and lower strength steel. Within these regions, content of Cu and deformation rate above recrystallization temperature are the control factors which are frequently used to improve the yield strength. Content of Ni, B and Nb and processing parameters SRT
and FRT also can be used though less frequently.

Figure 4.8 The influence of deformation rate to the yield strength.

4.2 Ultimate Tensile Strength Model

The ultimate tensile strength model is built by the approach same as the yields strength model. The minimum number of samples in a leaf is set to five. When the tree was full grown, $R^2$ reached 0.8999. Evaluated by 10 folds cross-validation, the optimal tree has 15 splits with an accuracy of 0.8689. The increase of coefficient of determination during tree growth is shown in Figure 4.9, and the accuracy of different size of trees, evaluated by 10
folds cross-validation, is shown in Figure 4.10. The performance of the model is displayed by the experimental-predicted plot, as shown in Figure 4.11.

Figure 4.9 split history of coefficient of determination of UTS model

Figure 4.10 Change of coefficient of determination with increase in number of splits in UTS model evaluated by 10 fold cross-validation
A list of rules deduced by UTS regression tree mode and corresponding support samples are shown in Table 4.3. Following the pathway \( \text{CR}=95\&B>=0.001\&\text{Cu}=1.18\&\text{Ti}=0.044 \), the tensile strength may achieve a maximum of 1087MPa.

Table 4.3 Deduced rules from UTS regression tree model

<table>
<thead>
<tr>
<th>Pathway from root to leaf</th>
<th>Predicted UTS(MPa)</th>
<th>Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>CR(5)&amp;Ni&lt;1.15&amp;Ti&lt;0.032&amp;Cu&lt;1.22&amp;Ti&lt;0.03</td>
<td>481.1</td>
<td>9</td>
</tr>
<tr>
<td>CR(5)&amp;Ni&lt;1.15&amp;Ti&lt;0.032&amp;Cu&lt;1.22&amp;Ti&gt;=0.03</td>
<td>557.6</td>
<td>5</td>
</tr>
<tr>
<td>CR(5)&amp;Ni&lt;1.15&amp;Ti&lt;0.032&amp;Cu&gt;=1.22</td>
<td>589.0</td>
<td>5</td>
</tr>
<tr>
<td>CR(5)&amp;Ni&lt;1.15&amp;Ti&gt;=0.032&amp;Cu&lt;1.65</td>
<td>673.4</td>
<td>9</td>
</tr>
<tr>
<td>CR(5)&amp;Ni&lt;1.15&amp;Ti&gt;=0.032&amp;Cu&gt;=1.65</td>
<td>829.9</td>
<td>5</td>
</tr>
<tr>
<td>CR(5)&amp;Ni&gt;=1.15&amp;C&lt;0.057&amp;Ni&lt;1.26&amp;Mn&gt;=1.32</td>
<td>780.1</td>
<td>14</td>
</tr>
<tr>
<td>CR(5)&amp;Ni&gt;=1.15&amp;C&lt;0.057&amp;Ni&lt;1.26&amp;Mn&lt;1.32&amp;C&lt;0.055</td>
<td>818.2</td>
<td>6</td>
</tr>
<tr>
<td>CR(5)&amp;Ni&gt;=1.15&amp;C&lt;0.057&amp;Ni&lt;1.26&amp;Mn&lt;1.32&amp;C&gt;=0.055</td>
<td>830.5</td>
<td>6</td>
</tr>
<tr>
<td>CR(5)&amp;Ni&gt;=1.15&amp;C&lt;0.057&amp;Ni&gt;=1.26&amp;Ni&gt;=1.31</td>
<td>841.0</td>
<td>8</td>
</tr>
<tr>
<td>CR(5)&amp;Ni&gt;=1.15&amp;C&lt;0.057&amp;Ni&gt;=1.26&amp;Ni&lt;1.31</td>
<td>918.9</td>
<td>7</td>
</tr>
<tr>
<td>CR(5)&amp;Ni&gt;=1.15&amp;C&gt;=0.057</td>
<td>968.8</td>
<td>8</td>
</tr>
<tr>
<td>CR(95)&amp;B&lt;0.001</td>
<td>902.3</td>
<td>11</td>
</tr>
<tr>
<td>CR(95)&amp;B&gt;=0.001&amp;Cu&lt;1.18&amp;FRT&lt;750</td>
<td>944.2</td>
<td>5</td>
</tr>
</tbody>
</table>
Table 4.3 (Continued)

<table>
<thead>
<tr>
<th>Pathway from root to leaf</th>
<th>Predicted UTS(MPa)</th>
<th>Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>CR(95)&amp;B&gt;=0.001&amp;Cu&lt;1.18&amp;FRT&gt;=750</td>
<td>1007.6</td>
<td>11</td>
</tr>
<tr>
<td>CR(95)&amp;B&gt;=0.001&amp;Cu&gt;=1.18&amp;Ti&lt;0.044</td>
<td>1041.2</td>
<td>6</td>
</tr>
<tr>
<td>CR(95)&amp;B&gt;=0.001&amp;Cu&gt;=1.18&amp;Ti&gt;=0.044</td>
<td>1087.0</td>
<td>6</td>
</tr>
</tbody>
</table>

As mentioned in previous section, the relative influence of input variables to UTS can be evaluated from their contribution to deviation reduction and split number, as shown in Figure 4.12 and 4.13.

![Figure 4.12 Contribution of input variables to error reduction in UTS Model](image-url)
In the UTS model, the most influential parameter is cooling rate (CR), which served as the split criterion on the top of tree and made the largest contribution. Meanwhile, the most influential element in yield strength model Mo has no influence. It indicates that the dependence of yield strength and ultimate tensile strength upon composition and processing is quite different in given data.

Nickel was determined be the most influential composition factor, which made the second largest contribution. In some nodes, the content of nickel served as criteria for three times in the pathway from the root to the nodes. It was found that the content of nickel is the important criterion in the branch with lower cooling rate (CR=5), as shown in Figure 4.14. It indicates that under the lower cooling rate, the content of nickel need to be carefully adjusted.
For the branch with higher cooling rate and higher UTS samples, the nonmetallic element boron plays an important role to further improve the tensile strength. If the content of Boron is higher than 0.001%, it can make the steels have an average tensile strength over 1000 MPa, as shown in Figure 4.15.

For the lower branch cooling rate branch, a strong interactions between titanium, niobium and copper, which contributes to the variation of tensile strength, was detected by the regression tree, as shown in Figure 4.16.
4.3 Elongation Rate Model

A model to predict the elongation rate was built by the same approach mentioned previously. The $R^2$ of the fully grown tree was 0.8508, and the optimal-size tree with an accuracy of 0.8012 was found after 15 splits. The increase of coefficient of determination
during tree growth is shown in Figure 4.17, and the accuracy of different sizes of trees, evaluated by 10 folds cross-validation is shown in Figure 4.18. The performance of the model is displayed by experimental-predicted plot, as shown in Figure 4.19.

Figure 4.17 Split history of coefficient of determination of EL model

Figure 4.18 Change of coefficient of determination with increase in number of splits in EL model evaluated by 10 fold cross-validation
Figure 4.19 Actual-predicted plots of EL model

A list of rules deduced by EL regression tree mode and corresponding support samples are shown in Table 4.4.

Table 4.4 Deduced rules from UTS regression tree model

<table>
<thead>
<tr>
<th>Pathway from root to leaf</th>
<th>Predicted EL (%)</th>
<th>Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mo&gt;=0.43&amp;Si&gt;=0.491</td>
<td>15.6</td>
<td>9</td>
</tr>
<tr>
<td>Mo&gt;=0.43&amp;Si&lt;0.491&amp;Ti&lt;0.021&amp;Nb&gt;=0.035&amp;Mn&lt;1.32</td>
<td>18.8</td>
<td>12</td>
</tr>
<tr>
<td>Mo&gt;=0.43&amp;Si&lt;0.491&amp;Ti&lt;0.021&amp;Nb&gt;=0.035&amp;Mn&gt;=1.32</td>
<td>20.2</td>
<td>6</td>
</tr>
<tr>
<td>Mo&gt;=0.43&amp;Si&lt;0.491&amp;Ti&lt;0.021&amp;Nb&lt;0.035&amp;Mn&gt;=1.26</td>
<td>20.5</td>
<td>18</td>
</tr>
<tr>
<td>Mo&gt;=0.43&amp;Si&lt;0.491&amp;Ti&lt;0.021&amp;Nb&lt;0.035&amp;Mn&lt;1.26</td>
<td>21.5</td>
<td>6</td>
</tr>
<tr>
<td>Mo&gt;=0.43&amp;Si&lt;0.491&amp;Ti&gt;=0.021</td>
<td>25.8</td>
<td>5</td>
</tr>
<tr>
<td>Mo&lt;0.43&amp;CR(95)&amp;B&gt;=0.0009&amp;FRT&gt;=775</td>
<td>18.9</td>
<td>10</td>
</tr>
<tr>
<td>Mo&lt;0.43&amp;CR(95)&amp;B&gt;=0.0009&amp;FRT&lt;775&amp;FRT&lt;750</td>
<td>20.1</td>
<td>5</td>
</tr>
<tr>
<td>Mo&lt;0.43&amp;CR(95)&amp;B&gt;=0.0009&amp;FRT&lt;755&amp;FRT&gt;750</td>
<td>21.9</td>
<td>9</td>
</tr>
<tr>
<td>Mo&lt;0.43&amp;CR(95)&amp;B&lt;0.0009</td>
<td>23.9</td>
<td>5</td>
</tr>
</tbody>
</table>
The contribution to deviation reduction and the number of splits of input variables are shown in Figure 4.20 and 4.21.

![Figure 4.20 Contribution of input variables to error reduction in EL model](image_url)
It is very interesting that both in the yield strength and elongation model, the content of molybdenum is one of the most influential factors and serve as the split criterion during the initial growing of the tree. In the yield strength model, the criterion Mo>=0.32% or not divide the sample into a lower yield strength group and a higher yield strength group; in elongation model, as shown in Figure 4.22, the criterion Mo <=0.43% or not assigned samples into a higher elongation group or a lower elongation group. Molybdenum has been found to be helpful for a fine grain structure and has strong interaction with the precipitation effect of other elements [45]. By comparing these two models, it can be found that in the given data and corresponding input space, the molybdenum can both increase the yield strength and percentage of elongation of steel when its content is carefully adjusted within some region, possibly 0.32−0.43%.
Like yield strength and UTS models, cooling rate (CR) still plays an important role in the elongation model, which distinguishes further the higher elongation group. Though CR was not used in the first split, its contribution to error reduction is the greatest. As expected, higher cooling rate decreases the ductility of steel while increasing the strength.

It can be found that in the elongation model, most branches with lower element content have higher average percentage of elongation, while most of higher strength groups are assigned to branches with higher element content in the yield strength and the UTS models. It is possible that most of the strengthening effects of these elements come from the solid solution strengthening and precipitation strengthening which increase the strength but reduce the ductility. However, in the elongation model, there are some
branches where higher element content leads to higher elongation. Take a portion of regression tree in Figure 4.23 for example. When the content of Nb>=0.035%, the higher content of Mn leads to higher elongation, while it is opposite in the other branch Nb<0.035%. It is interesting that there seems to be some interaction between the microalloying effect of nickel and manganese. Figure 4.24 shows a part of middle level of tree where the content of silicon obviously has different contribution to the elongation. When the content of silicon is between 0.491% and 0.4%, there is a sudden increase in the ductility which may be worthy of investigation.

![Figure 4.23 A portion of the regression tree of EL model where higher element content leads to higher elongation](image)

The processing variable Finish Rolling Temperature (FRT) plays an active role in the elongation model. FRT was used as split criteria four times and contributed to error reduction to a relatively large extent. And three of four splits involving FRT used the criterion whether FRT >775°C or not, and the other is 750°C, as shown in Figure 4.25. It
indicates that 750~775°C could be a critical finish rolling temperature which greatly influenced the phase transformation and precipitates formation in the given sample. It was reported that in the steel with similar composition, 750~775°C was the edge of γ austenite area and γ + α austenite and ferrite mixed area [40]. Cooling in a γ + α mixed area may lead to a reduction of share of acicular ferrite and increase in a share of polygonal ferrite, which results in decrease in strength and increase in ductility.

![Figure 4.24 A portion of regression tree of EL model where the content of silicon has different effect on the elongation](image)

Figure 4.24 A portion of regression tree of EL model where the content of silicon has different effect on the elongation
Figure 4.25 FRT serve as split criterion in several parts in the EL regression tree; The value of FRT frequently used is 750°C or 755°C.
Chapter 5 Conclusion and Future Work

In this study a data mining technology, Recursive Partitioning was applied to model the tensile properties of HSLA steel as a function of composition and processing parameters. Several models were established for yield strength, ultimate tensile strength and elongation rate of high strength low alloyed steel. A model evaluation method, 10 folds cross-validation was used to assess the accuracy and detect the overfitting of the established models. Parameters of recursive partitioning were varied and selected to achieve best performance. The optimal tree size was determined by pruning to achieve highest accuracy and prevent overfitting. Compared to linear regression and neural networks, recursive portioning demonstrates good performance in predicting nonlinear behavior with a relatively simple structure.

The recursive partitioning method was successful in reducing the dimensionality of training data. In the models, usually five to seven variables were selected from the original fourteen variables to make prediction. Predictive rules were deduced from the models and can be used as a guideline for further experimental design. The relative influences of input variables were ranked by their contributions to the error reduction and times as split criteria. The tree graph generated by recursive partitioning proved to be very interpretable. Some interesting regions where different interactions exist between elements have been identified for further experimental investigation.
The performance of recursive partitioning depends strongly on the training data; more data will make the model more accurate and robust. In this study, the time during processing is assumed as invariants and less important. In order to examine the complex mechanism during hot rolling processing, more data about the processing parameters, such as the time between rolling stands, and strain rate during the rolling should be included in the further study. It is also very important to examine the microstructure evolution affected by interaction of the composition and processing. More data about the microstructural feature such as grain size, volume fraction of ferrite and volume fraction of precipitates will be helpful for establishing more comprehensive models.

While the present model can provide comprehensive interpretable information, the accuracy and capability of describing nonlinearity need to be improved. In the present model, the prediction is simply made by the average value of the leaf which the sample is assigned to. It is possible to incorporate the tree structure with more complex prediction model, which is called “model tree”. Application of model tree should provide more accurate rules for steel design based on the given data.
Reference


[34] L. Harmon, "Experiment planning for combinatorial materials discovery", *Journal of Materials Science* 2003, 38:4479–4485


Acknowledgement

I would like to give my most sincere gratitude to my major professor Dr. Krishna Rajan, who has been consistently helping me and giving me invaluable advice on the research project. I am especially grateful to Dr. Shubhabrata Datta, who provided experimental data used in this study. Many thanks go to Dr. Tao Wang and Dr. Changwon Suh for many stimulating discussions throughout this work. I would also like to express appreciation to my colleague and friend Li Xue, who gave me lots of advice on the data mining issue and encouraged me in the hardest time. I want to thank other group members Chang Sun Kong, Scott Broderick for their assistance and suggestions on the project.

This study is supported by the National Science Foundation International Materials Institute Program for Combinatorial Sciences and Materials Informatics Collaboratory (CoSMIC IMI), Grant no. DMR-0603644.

Thank all the committee members for donating their time to make my thesis a better document. Thank Prof. Alan M. Russell for his careful revision of my thesis.

I want to express my thankfulness to my parents. Without them I wouldn’t have made it this far and they continue to be my strong support.