Predicting the Type of Nanostructure Using Data Mining Techniques and Multinomial Logistic Regression

Mahmoud Shehadeh*, Nader Ebrahimi, and Abel Ochigbo

*Division of Statistics, Northern Illinois University, DeKalb, IL 60115, USA

Abstract

Nanotechnology and nanomaterials have a promised future in different aspects of modern life that involve medicine, environment, space, energy, electronics, security, and many others. While the applications of nanomaterials seem to be limitless, new challenges are also being posed. With regard to the type of one-dimensional nanostructure of Cadmium Selenide (CdSe), there are three possible morphologies presented: nanosaws, nanowires, and nanobelts. Since the synthesis of these morphologies are by trial and error, our goal in this paper is to use statistical and data mining techniques to predict the type of CdSe nanostructure. The methods used for prediction are: a multinomial logistic regression, a support vector machine, and a random forest. The results are compared using two statistical indices: sensitivity and specificity, and the factors that influence the possible nanostructure are identified. Based on the results, data mining techniques showed to be a better fit for prediction comparing to the multinomial logistic regression model. We also identify the levels of these factors that maximize the proportions of nanosaws, nanowires, and nanobelts.

Keywords: Nanotechnology; CdSe nanostructure; Prediction; Data Mining; Multinomial Logistic Regression.

1. Introduction

The prefix "nano", meaning dwarf in ancient Greek, refers to a reduction of size or time by $10^{-9}$ which is a billionth of a meter. By comparison, the width across a DNA molecule is two nanometers (nms) and the wavelength of Ultraviolet light is 300 nms [1]. The inspiration of nanoscience and technology began in 1959 when the Nobel laureate and physicist Richard Feynman challenged the scientific community of the idea of manipulating matter at the atomic scale to create systems, devices, and materials with fundamentally new properties and functions [2]. The study of matter at dimensions of roughly 1 to 100 nms is broadly referred to nanotechnology. Because of the novel physical, chemical, and biological properties that nanomaterials exhibit [3], they have a promised future in different aspects of modern life that involve medicine, environment, space, energy, and many others [4-7].
As a relatively new multidisciplinary research field, nanotechnology is facing many challenges, among many others, is in the area of designing and fabricating nanostructures with fundamentally new properties [8]. To achieve the future goals, it is essential to synthesize nanomaterials in large quantities with reproducible structures, sizes, and compositions by shifting from laboratory-scale research to industrial production. In recent years, the field of statistics has been helpful to the rapid development in nanotechnology by meeting some of the challenges especially in data collection, quality control, achieving desired nanostructures by systematically investigating the process variables, modeling the yields of various types of nanostructures as functions of the experimental conditions, and producing nanostructures with high yield and minimal variation [9-12].

An example of nanomaterials is the one-dimensional cadmium selenide (CdSe) which has been investigated by researchers to develop controlled syntheses of CdSe nanoparticles for applications in biomedical imaging, lasing materials, and optoelectronics [13,14]. The CdSe was found to exhibit the one dimensional morphologies of nanowires, nanobelts, and nanosaws, and the synthesis of these morphologies are by trial and error [15]. Figure 1 shows the images of these three nanostructures. This type of nanostructure is considered one-dimensional because of the way atoms are arranged in chains and the way they interact. Atoms are positioned in chains and mostly interact with other atoms in one chain, along a single dimension of space, and little with atoms in other chains. In this article, three statistical modeling techniques are used for predicting the type and describing the growth of one-dimensional CdSe nanostructures based on the experimental data published by Ma and Wang [15]. Then the process factors that influence the possible outcome are identified. The modeling techniques are: a multinomial logistic regression, a support vector machine, and a random forest. The response variable in this study is the type nanostructure with four nominal scale levels: nanosaws, nanowires, nanobelts, and no growth, and explanatory variables are the temperature (TEMP), pressure (PRES), and distance (DIST). We also provide the levels of TEMP, PRES, and DIST that maximize the proportions of nanosaws, nanowires, and nanobelts.

Fig. 1. SEM images of the CdSe morphologies structures and corresponding transmission electron microscopy (TEM) images (lower right) and electron diffraction patterns (lower left). a) Nanosaws, b) nanobelts, and c) nanowires [15].

2. Materials and methods

2.1 Multinomial logistic regression

Let $X$ be the matrix of the explanatory variables, and $Y$ be the vector of the response variable with $J$ possible outcomes. In the case of $J > 2$, then the response variable $Y$ follows a multinomial distribution; hence $\sum_{i=1}^{J} p_i = 1$, where $p_i$ is the probability of obtaining the $i^{th}$ outcome, $p_i = P(Y = i)$, $i = 1, 2, ..., J$. The multinomial logits can be expressed as $\eta_i = \log \frac{p_i}{\sum_{j=1}^{J} p_j} = X_i \beta$, $i = 1, 2, ..., J-1$. In this setup, $\eta_i$ denotes the log-odds ratio of obtaining the $i^{th}$ category as compared to the reference category, $J$, and it can be expressed as a function of the independent variables. The regression coefficients, $\beta$, in the multinomial logistic model are estimated using the maximum likelihood method (MLE), and can be interpreted as the increase in log-odds of falling into category $i$ versus the reference category $J$. For a given input, if $p_k = \text{Max}(p_1, p_2, ..., p_J)$ then the input will be classified into the $k^{th}$ category of $Y$ [16].
2.2 Support vector machine

A support vector machine (SVM) is a state-of-the-art supervised learning algorithm used widely for regression and classification analysis due to its high accuracy, and has its root from statistical learning theory [17]. The idea of classification SVM is to predict the possible outcome for each given input by mapping the input vectors into a higher-dimensional feature space defined by a kernel function, and this can be done by constructing optimal separating hyperplanes that maximize the margin (i.e., the space between regions) between different classes[18].

2.3 Random forest

Similar to SVM, random forest is a supervised learning algorithm also used for regression and classification due to its robustness to noise and overfitting. It is considered a method of ensemble learning which based on the idea of combining multiple models into a single model [19]. In a random forest, each decision tree is built to its maximal depth without pruning [20]. Each node is then split using the best among a subset of predictors at that node. It is worth noting that random forest has only two parameters (the number of variables at each node and the number of trees in the forest), and is not very sensitive to their values [21]. The error rate estimate can be obtained, based on the training data, by predicting the “out-of-bag” data (OOB) and aggregating the OOB predictions.

2.4 Predicting the type of nanostructure

2.4.1 Description of data set

Temperature and pressure are the two main process variables that affect the morphology of cadmium selenide (CdSe) nanostructures, which are created through a thermal evaporation process in a single zone horizontal tube furnace (Thermolyne 79300). During a thermal evaporation process, the source material is evaporated in a vacuum; the vacuum allows the vapor particles to travel directly to substrate, where they condense back to a solid state. Commercial grade CdSe was used as a source material for the creation of CdSe nanosaws, nanowires, and nanobelts. During each experiment, the CdSe was placed in the furnace in the center of a polycrystalline Al2O3 tube of length 30 inches with an inner diameter of 1.5 inches. Between 4 and 6 single-crystal silicon substrates with a 2-nanometer thermally evaporated non-continuous layer of gold were placed downstream from the source material to collect the deposition of nanostructures. The furnace was set at specified combinations of temperature and pressure for 60 minutes and then cooled to room temperature. The resulting CdSe nanostructures were counted and classified by scanning electron microscopy (SEM) (LEO 1530 FEG) and transmission electron microscopy (TEM) (Hitachi HF-2000 FEG at 200 kV). As many as 180 individual nanostructures occurred on each substrate.

A total of 415 substrates were obtained from 135 runs of the furnace. A $5 \times 9$ full factorial experiment was conducted with five levels of source temperature (630, 700, 750, 800, 850° C) and nine levels of pressure (4, 100, 200, 300, 400, 500, 600, 700, 800 mbar), and each of the 45 combinations of temperature and pressure were run 3 times. Since the number of substrates used and the location of the substrates were not the same for each run, the runs cannot be considered replications, and a third variable, distance from the midpoint of the substrate to the source, must be treated as a process variable as well. To clarify, consider the temperature and pressure combination (630° C, 4 mb) for example. Six substrates were used during the first run at distances of 12.4, 14.7, 15.4, 16.9, 18.6, and 20.7 cm; four substrates were used during the second run at distances of 12.2, 15.1, 17.6, and 19.4 cm; and seven substrates were used during the third run at distances of 12.5, 14.8, 15.4, 16.9, 19, 21.1, 23.5 cm. So, seventeen unique substrates were used with this specific temperature and pressure combination. Each substrate constitutes a row of the dataset which can be seen in Table 1. The complete data can be downloaded from www.isye.gatech.edu/~roshan [15, 22].

Now if each substrate is treated as an experimental unit, then the design matrix will be $415 \times 3$, where the three columns correspond to temperature (TEMP), pressure (PRES), and distance (DIST). The response variable for each unit is a vector of the form $y = (y_1, y_2, y_3, y_4)$, where $y_1, y_2, y_3,$ and $y_4$ denote the number of nanosaws, nanowires, nanobelts, and no morphology, respectively. Note that $\sum_{j=1}^{4} y_j = 180$. However, when the temperature was 850° C, it was observed that almost no morphology occurred. So the 67 resulting observations when temperature was 850° C are excluded from the analysis, leaving $= 348$ observations for the analysis.
Table 1 Partial data obtained from the experiment

<table>
<thead>
<tr>
<th>Temperature (°C)</th>
<th>Pressure (mbar)</th>
<th>Distance (cm)</th>
<th>Nanosaws</th>
<th>Nanowires</th>
<th>Nanobelts</th>
<th>No growth</th>
</tr>
</thead>
<tbody>
<tr>
<td>630</td>
<td>4</td>
<td>10.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>180</td>
</tr>
<tr>
<td>630</td>
<td>4</td>
<td>14.3</td>
<td>74</td>
<td>106</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>630</td>
<td>4</td>
<td>15.1</td>
<td>59</td>
<td>121</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>630</td>
<td>200</td>
<td>10.5</td>
<td>180</td>
<td>0</td>
<td>0</td>
<td>180</td>
</tr>
<tr>
<td>630</td>
<td>200</td>
<td>17.2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>180</td>
</tr>
<tr>
<td>630</td>
<td>200</td>
<td>10.5</td>
<td>108</td>
<td>0</td>
<td>72</td>
<td>0</td>
</tr>
<tr>
<td>630</td>
<td>200</td>
<td>13.3</td>
<td>117</td>
<td>0</td>
<td>63</td>
<td>0</td>
</tr>
<tr>
<td>630</td>
<td>400</td>
<td>15.8</td>
<td>180</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>630</td>
<td>400</td>
<td>18</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>180</td>
</tr>
<tr>
<td>630</td>
<td>400</td>
<td>10.5</td>
<td>137</td>
<td>0</td>
<td>43</td>
<td>0</td>
</tr>
<tr>
<td>630</td>
<td>400</td>
<td>13.2</td>
<td>124</td>
<td>0</td>
<td>56</td>
<td>0</td>
</tr>
<tr>
<td>750</td>
<td>4</td>
<td>12.9</td>
<td>43</td>
<td>88</td>
<td>49</td>
<td>0</td>
</tr>
<tr>
<td>750</td>
<td>4</td>
<td>16.6</td>
<td>22</td>
<td>158</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>750</td>
<td>4</td>
<td>19.8</td>
<td>0</td>
<td>0</td>
<td>180</td>
<td>0</td>
</tr>
</tbody>
</table>

2.4.2 Data preparation

In order to the process variables, TEMP, PRES, and DIST, to be scaled free. Each one of them is scaled to [-1,1] by appropriate transformation [12]. Let \( T \), \( P \), and \( D \) be the scaled variables obtained by transforming TEMP, PRES, and DIST, respectively. Then the following transformations were used [12] to obtain \( T \), \( P \), and \( D \):

\[
T = 2\left(\frac{TEMP - \text{Min}(TEMP)}{\text{Max}(TEMP) - \text{Min}(TEMP)}\right) - 1, \quad P = 2\left(\frac{PRES - \text{Min}(PRES)}{\text{Max}(PRES) - \text{Min}(PRES)}\right) - 1, \quad D = 2\left(\frac{DIST - \text{Min}(DIST)}{\text{Max}(DIST) - \text{Min}(DIST)}\right) - 1.
\]

2.4.3 Multinomial logistic regression

Using zelig [23,24] command in R Statistical Software [25], a multinomial logistic model was fitted using the training data to describe the growth of nanosaws, nanowires, and nanobelts. All the regression coefficients, including the interaction terms, were highly significant with p-values < 0.05. Then the coefficient estimates were used to predict type of the nanostructure for each given input of the testing data.

2.4.4 Support vector machine

To make up the SVM model, ksvm command in R [26] was used with a Gaussian Radial Basis kernel function. The training was terminated at the tolerance criterion of 0.001.

2.4.5 Random forest

The random forest model was built using randomForest package in R [27]. Number of variables that considered for partitioning the dataset at any given time was one, and number of trees that was built is 500. As was mentioned in Section 2.3, the model is not very sensitive to the values of these two parameters.

2.4.6 Model assessment

To assess how the different models will accurately classify the type of nanostructure for a given input. The data was split randomly into two subsamples, training and testing, using sample function in R [25]. This command can be used to select specific index number in the dataset randomly and without replacement, and then use these selected index numbers to split the dataset into training and testing sets. Users can specify the proportion of each set. The first subsample is used as training data to obtain appropriate models, and the second subsample is retained as testing data. Throughout this paper we use 75% of the original data as our training data and the remaining 25% as the testing data. Then two statistical indices are used: sensitivity and specificity. For this experiment, with four possible
outcomes, we can compute these indices using the confusion matrix as described in Table 2.

### Table 2 Confusion matrix for a given model

<table>
<thead>
<tr>
<th></th>
<th>Predicted</th>
<th>Actual</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nanosaws</td>
<td>$n_{11}$</td>
<td>$n_{12}$</td>
</tr>
<tr>
<td>Nanowires</td>
<td>$n_{21}$</td>
<td>$n_{22}$</td>
</tr>
<tr>
<td>Nanobelts</td>
<td>$n_{31}$</td>
<td>$n_{32}$</td>
</tr>
<tr>
<td>No Growth</td>
<td>$n_{41}$</td>
<td>$n_{42}$</td>
</tr>
</tbody>
</table>

As an example, sensitivity and specificity for category 1 are:

- **Sensitivity of category 1** = \( \frac{n_{22} + n_{32} + n_{42} + n_{23} + n_{33} + n_{43} + n_{24} + n_{34} + n_{44}}{n_{11}} \), and
- **Specificity of category 1** = \( \frac{n_{11} + n_{21} + n_{31} + n_{41}}{n_{11}} \).

### 3. Results

#### 3.1 Sensitivity and specificity

After building the models using the training data, and predicting the nanostructure category for each given input in the testing data, we compute \( n_{11}, n_{12}, ..., n_{44} \) to obtain the confusion matrix for each model, and then compute the sensitivity and specificity as described in Section 2.4.6. Table 3 shows the sensitivity and specificity for each of the three models’ categories, namely nanosaws, nanowires, nanobelts, and no growth.

### Table 3 Sensitivity and specificity

<table>
<thead>
<tr>
<th>Category</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Multinomial logistic</td>
<td>SVM</td>
</tr>
<tr>
<td>1</td>
<td>0.897</td>
<td>0.929</td>
</tr>
<tr>
<td>2</td>
<td>0.823</td>
<td>0.875</td>
</tr>
<tr>
<td>3</td>
<td>0.921</td>
<td>0.936</td>
</tr>
<tr>
<td>4</td>
<td>0.85</td>
<td>0.916</td>
</tr>
</tbody>
</table>

From Table 3, it seems that the random forest model does a better job in prediction the type of nanostructure compared to SVM and multinomial logistic regression approaches.

#### 3.2 Maximization the yield of the CdSe Nanostructures

Since the random forest model does a better job in terms of prediction, we will use it to identify the conditions that maximize the yield of nanosaws, nanowires, and nanobelts. For example, from Table 4, it is clear that to achieve maximum proportion of, say, nanowires, the **TEMP** should be about 700°C, **PRES** should be about 100 mbar, and **DIST** should be about 13.846 cm.

### Table 4 The Conditions for Maximizing the Proportions of Nanosaws, Nanowires, and Nanobelts.

<table>
<thead>
<tr>
<th>Nanostructures</th>
<th>Temperature</th>
<th>Pressure</th>
<th>Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nanosaws</td>
<td>630</td>
<td>300</td>
<td>14.154</td>
</tr>
<tr>
<td>Nanowires</td>
<td>700</td>
<td>100</td>
<td>13.846</td>
</tr>
<tr>
<td>Nanobelts</td>
<td>700</td>
<td>4</td>
<td>16.738</td>
</tr>
</tbody>
</table>
4. Conclusion

Capabilities of three models to predict the type of one-dimensional CdSe nanostructure were outlined in detail. To evaluate the performance of these models for prediction, sensitivity and specificity statistical indices were used. According to the results, the random forest showed to be the best performed approach. The results obtained by this model were used to identify the values of temperature, pressure, and distance that maximize the yield of each type of nanostructure.

Acknowledgements

The research of the second author was partially supported by a grant from the National Security Agency under grant number H98230-11-1-0138. We also would like to thank two anonymous referees for their comments that led to improvements in the paper.

References