Computational Materials Science 154 (2018) 346–354

Contents lists available at ScienceDirect

Computational Materials Science

journal homepage: www.elsevier.com/locate/commatsci

A data-driven statistical model for predicting the critical temperature of a superconductor

Kam Hamidieh

Statistics Department, The Wharton School, University of Pennsylvania, 400 Jon M. Huntsman Hall, 3730 Walnut Street, Philadelphia, PA 19104-6340, United States

ARTICLE INFO

Keywords:
Superconductivity
Superconductor
Machine learning
Statistical learning
Data mining
Critical temperature

ABSTRACT

We estimate a statistical model to predict the superconducting critical temperature based on the features extracted from the superconductor's chemical formula. The statistical model gives reasonable out-of-sample predictions: ± 9.5 K based on root-mean-squared-error. Features extracted based on thermal conductivity, atomic radius, valence, electron affinity, and atomic mass contribute the most to the model's predictive accuracy. It is crucial to note that our model does not predict whether a material is a superconductor or not; it only gives predictions for superconductors.

1. Introduction

Superconducting materials - materials that conduct current with zero resistance - have significant practical applications. Perhaps the best known application is in the Magnetic Resonance Imaging (MRI) systems widely employed by health care professionals for detailed internal body imaging. Other prominent applications include the superconducting coils used to maintain high magnetic fields in the Large Hadron Collider at CERN, where the existence of Higgs Boson was recently confirmed, and the extremely sensitive magnetic field measuring devices called SQUIDs (Superconducting Quantum Interference Devices). Furthermore, superconductors could revolutionize the energy industry as frictionless (zero resistance) superconducting wires and electrical system may transport and deliver electricity with no energy loss; see Hassenzahl [9].

However, the wide spread applications of superconductors have been held back by two major issues: (1) A superconductor conducts current with zero resistance only at or below its superconducting critical temperature (T_c). Often impractically, a superconductor must be cooled to extremely low temperatures near or below the boiling temperature of nitrogen (77 K) before exhibiting the zero resistance property. (2) The scientific model and theory that *predicts* T_c is an open problem which has been baffling the scientific community since the discovery of superconductivity in 1911 by Heike Kamerlingh Onnes, in Leiden.

In the absence of any theory-based prediction models, simple empirical rules based on experimental results have guided researchers in synthesizing superconducting materials for many years. For example, the eminent experimental physicist Matthias [12] concluded that T_c is related to the number of available valence electrons per atom. (A few of these rules came to be known as the Matthias's rules.) It is now well known that many of the simple empirical rules are violated; see Conder [4].

In this study, we take an entirely data-driven approach to create a statistical model that predicts T_c based on its chemical formula. The superconductor data comes from the Superconducting Material Database maintained by Japan's National Institute for Materials Science (NIMS) at http://supercon.nims.go.jp/index_en.html. After some data preprocessing, 21,263 superconductors are used.

To our knowledge, Valentin et al. [19] and our work are the only papers that focus on statistical models to *predict* T_c for a *broad class* of materials. However, Owolabi et al. [15], Owolabi and Olatunji [14] focus on predicting T_c for Fe and MgB_2 based superconductors respectively.

We derive features (or predictors) based on the superconductor's elemental properties that could be helpful in predicting T_c. For example, consider $Nb_{0.8}Pd_{0.2}$ with $T_c = 1.98$ K. We can derive a feature based on the average thermal conductivities of the elements. Niobium and palladium's thermal conductivity coefficients are 54 and 71 W/(m×K) respectively. The mean thermal conductivity is $(54 + 71)/2 = 62.5$ W/(m×K). We can treat the mean thermal conductivity variable as a feature to predict T_c. In total, we define and extract 81 features from each superconductor.

We tried various statistical models but we eventually settled on two: A multiple regression model which serves as a benchmark model, and a gradient boosted model as the main prediction model which is implemented in our software.

Our software tool to predict T_c and the associated data are available

E-mail address: hkam@wharton.upenn.edu.

https://doi.org/10.1016/j.commatsci.2018.07.052
Received 2 April 2018; Received in revised form 27 June 2018; Accepted 28 July 2018
Available online 10 August 2018

Table 1
This table shows the properties of an element which are used for creating features to predict T_c.

Variable	Units	Description
Atomic Mass	Atomic mass units (AMU)	Total proton and neutron rest masses
First Ionization Energy	Kilo-Joules per mole (kJ/mol)	Energy required to remove a valence electron
Atomic Radius	Picometer (pm)	Calculated atomic radius
Density	Kilograms per meters cubed (kg/m^3)	Density at standard temperature and pressure
Electron Affinity	Kilo-Joules per mole (kJ/mol)	Energy required to add an electron to a neutral atom
Fusion Heat	Kilo-Joules per mole (kJ/mol)	Energy to change from solid to liquid without temperature change
Thermal Conductivity	Watts per meter-Kelvin (W/(m K))	Thermal conductivity coefficient κ
Valence	No units	Typical number of chemical bonds formed by the element

at https://github.com/khamidieh/predict_tc and will also be available at the publisher's complementary site. We have done our best to make the software use and access to the data as easy as possible.

Gradient boosted models create an ensemble of trees to predict a response. The trees are added in a sequential manner to improve the model by accounting for the points which are difficult to predict. Once a gradient boosted model is fitted, the weighted average of all the trees is used to give a final prediction. Gradient boosted models predict well because they are able to account for the complex interactions and correlations among the features.

The boosted models were first developed by Schapire [17], Freund [6]. The boosted models were generalized to *gradient* boosting by Friedman [7]. We use the latest improvement called XGBoost (eXtreme Gradient Boosting) by Chen and Guestrin [1], and the associated open-source R implementation of XGBoost by Chen et al. [2]. XGBoost is also available in other popular programming languages such as python and Julia. The full source code is at https://github.com/dmlc/xgboost.

Anthony Goldbloom, CEO of Kaggle (now a Google company), the premier data competition site, stated: "It used to be random forest that was the big winner, but over the last six months a new algorithm called XGBoost has cropped up, and it's winning practically every competition in the structured data category." You can see the talk at https://www.youtube.com/watch?v=GTs5ZQ6XwUM. Outside the competition realm, XGBoost has been successfully applied in disease prediction by Chen et al. [3], and in quantitative structure activity relationships studies by Sheridan et al. [18].

Our XGBoost model gives reasonable predictions: an out-of-sample error of about 9.5 K based on root-mean-squared-error (rmse), and an out-of-sample R^2 values of about 0.92. The numbers for the multiple regression model are about 17.6 K and 0.74 for the out-of-sample rmse and R^2 respectively. The multiple regression serves as a benchmark model.

We are able to assess the importance of the features in prediction accuracy. Features defined based on thermal conductivity, atomic radius, valence, electron affinity, and atomic mass are the most important features in predicting T_c. On the downside, simple conclusions such as the exact nature of the relationship between the features and T_c can't be inferred from the XGBoost model.

Valentin et al. [19] also create a model to predict T_c. Our approach is different than Valentin et al. [19] in the following ways: (1) We use XGBoost versus random forests, (2) we use a larger data set, (3) we use a single large model to obtain predictions rather than a cascade of models, (4) we create a larger number features *only* from the elemental properties, and (5) most importantly, we quantify the out-of-sample prediction error.

2. Data preparation

This section describes the detailed steps for the data preparation and feature extraction. Section 2.1 describes how the element data is obtained and processed. Section 2.2 describes the data preparation from NIMS Superconducting Material Database. Section 2.3 details how the features are extracted.

2.1. Element data preparation

The element data with 46 variables and 86 rows (corresponding to 86 elements) are obtained by using the `ElementData` function from Mathematica Version 11.1 by Wolfram and Research [20]. Appendix A lists the information sources for the element properties used by `ElementData`. The first ionization energy data came from http://www.ptable.com/ and is merged with the Mathematica data. About 12% of the entries out of the 3956 (= 46 × 86) entries are missing.

In choosing the properties, we are guided by Conder [4] but we also use our judgement to pick certain properties. For example, we drop the boiling point variable, and instead use the fusion heat variable which has no missing values, and is highly correlated with the boiling point variable. We had also gained some experience and insight creating some initial models for predicting T_c of elements only. We settle on 8 properties shown in Table 1.

With the choice of the above variables, we are only missing the atomic radii of La and Ce; we replace them with their covalent radii since atomic radii and covalent radii have very high correlation (≈ 0.95) and approximately on the same scale and range. Some bias may be introduced into our data with this minor imputation. We add a small constant of 1.5 to the electron affinity values of all the elements to prevent issues when taking logarithm of 0.

2.2. Superconducting material data preparation

Superconducting Material Database is supported by the NIMS, a public institution based in Japan. The database contains a large list of superconductors, their critical temperatures, and the source references mostly from journal articles. To our knowledge, this is the most comprehensive database of superconductors. Access to the database requires a login id and password but this is provided with a simple registration process.

We accessed the data on July 24, 2017 at http://supercon.nims.go.jp/supercon/material_menu. Once logged in, we chose "OXIDE & METALLIC" material. Fig. 1 shows a screen shot of the menu. We clicked on the "search" button to get *all* the data. We obtained 31,611 rows of data in a comma separated file format. The key columns (variables) were "element", the chemical formula of the material, and "Tc", the critical temperature. Variable "num" was a unique identifier for each row. Column "refno" contained links to the referenced source. The next few steps describe the manual clean up process:

1. We remove columns "ma1" to "mj2".
2. We sort the data by "Tc" from the highest to lowest.
3. The critical temperature for the following "num" variables are mistakenly shifted by one column to the right. We fix these by recording them under the "Tc" column: 31,020, 31,021, 31,022, 31,023, 31,024, 31,025, 153,150, 153,149, 42,170, 42,171, 30,716, 30,717, 30,718, 30,719, 150,001, 150,002, 150,003, 150,004, 150,005, 150,006, 150,007, 30,712, 30,713, 30,714, 30,715.
4. The following are removed since the critical temperatures seemed to have been misrecorded; They have critical temperatures over 203 K