Exploration and Prediction of Supercouncting Materials

Jie Yong

June 20, 2016

This purpose of this article is to explore the superconductor database and build models to predict whether superconducting Transition temperature Tc is high or low from three “golden coordinates” given by Rabe, Villars, Philips and Brown:

- **dX**: metallic electronegativity
- **dR**: orbital radii difference
- **Nv**: average valence electron number

The detailed definition of these features is from this article: article (http://journals.aps.org/prb/abstract/10.1103/PhysRevB.45.7650)

PHYSICAL REVIEW B VOLUME 45, NUMBER 14 1 APRIL 1992-II

Global multinary structural chemistry of stable quasicrystals, high-T_C ferroelectrics, and high-T_c superconductors

K. M. Rabe
Department of Applied Physics, Yale University, New Haven, Connecticut 06520

J. C. Phillips
AT&T Bell Laboratories, Murray Hill, New Jersey 07974

P. Villars
Villars Intermetallic Phases Databank, Schwanden Postfach, 6354 Vitznau, Switzerland

I. D. Brown
Department of Physics, McMaster University, Hamilton, Ontario, Canada L8S 4M1
(Received 2 May 1991; revised manuscript received 1 November 1991)

A common feature of stable quasicrystals, high-T_C ferroelectrics, and high-T_c superconductors is that their unusual physical properties are correlated with characteristic crystal stuctures that are often vicinal to structural instabilities. We describe a statistically based diagrammatic scheme, due to Villars, for classifying the full database of crystal structures of binary, ternary, and quaternary compounds and tendency to compound formation in binary- and ternary-alloy systems. Principles of an expert system using this global organization to study small sets of compounds of special interest are formulated. Application to quasicrystals, ferroelectrics, and superconductors results in the identification of diagrammatic regularities which enable us to recognize phenomenological trends and to develop computerized search strategies for the prediction of new materials.

The work is done with collobaration of:

- Gilad Kusne, Research Scientist, National Institute of Science and Technology
- Ichiro Takeuchi, Professor of Materials Science and Technology in University of Maryland.

Outline

1. Importing, Cleaning and Saving Data

2. Visualization and Prelimarary exploration
3. High Tc Predictions With Different Models
4. Sweet spots reidentification
5. summary

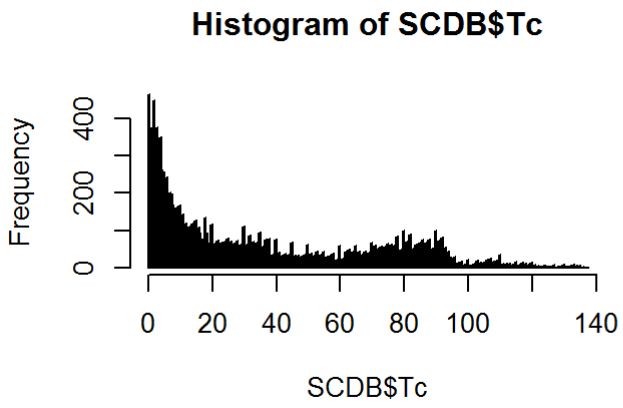
Importing, Cleaning and Saving Data

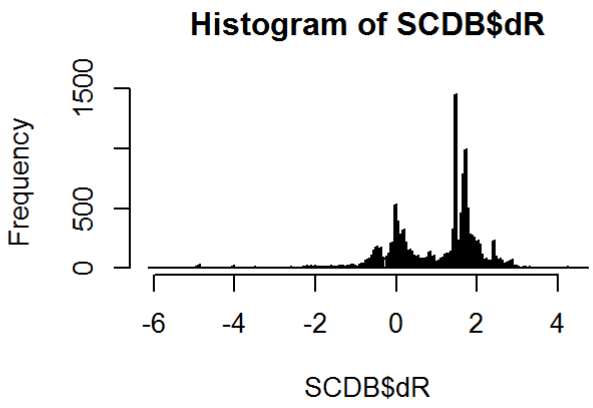
```
library("R.matlab")
# read raw matlab file
a <- readMat("data/SCDB.mat")
# build data frame
name <-as.character()
SCDB <-data.frame(dX=a$dX,Nv=a$Nv,dR=a$dR,Tc=a$Tc)
#extract names
for (i in 1:nrow(SCDB)){
  value <- (a$grouped.str)[[i]][[1]][1,1]
 name[i] <- ifelse(is.null(value), NA,value)
}
SCDB <-cbind(name=name,SCDB)
#remove NAs
na <-sum(is.na(SCDB))
totalnum <- nrow(SCDB)
SCDB <- SCDB[!(is.na(SCDB$name)) & !(is.na(SCDB$dX)),]
#save as txt file
write.table(SCDB, "data/neatdata.txt", sep="\t")
```

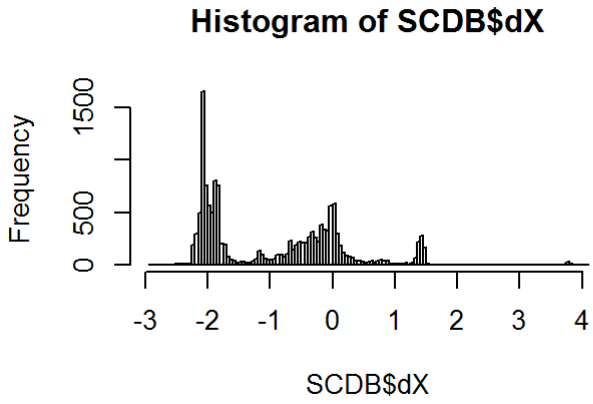
There are 1210 NAs from a total number of 14540 entries. We decide to remove them for our analysis. The data is saved in data/neatdata.txt.

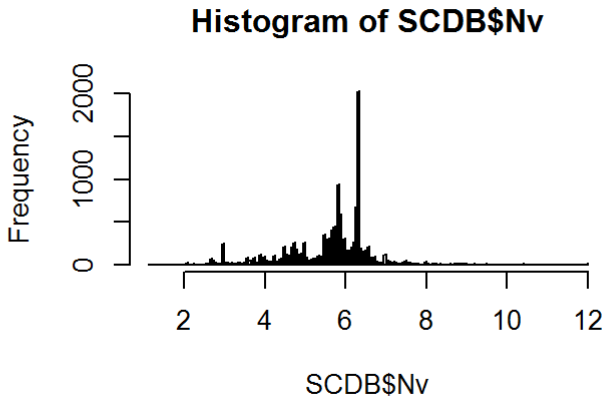
Data Exploration

The histogram of the features and Tc are in the following:









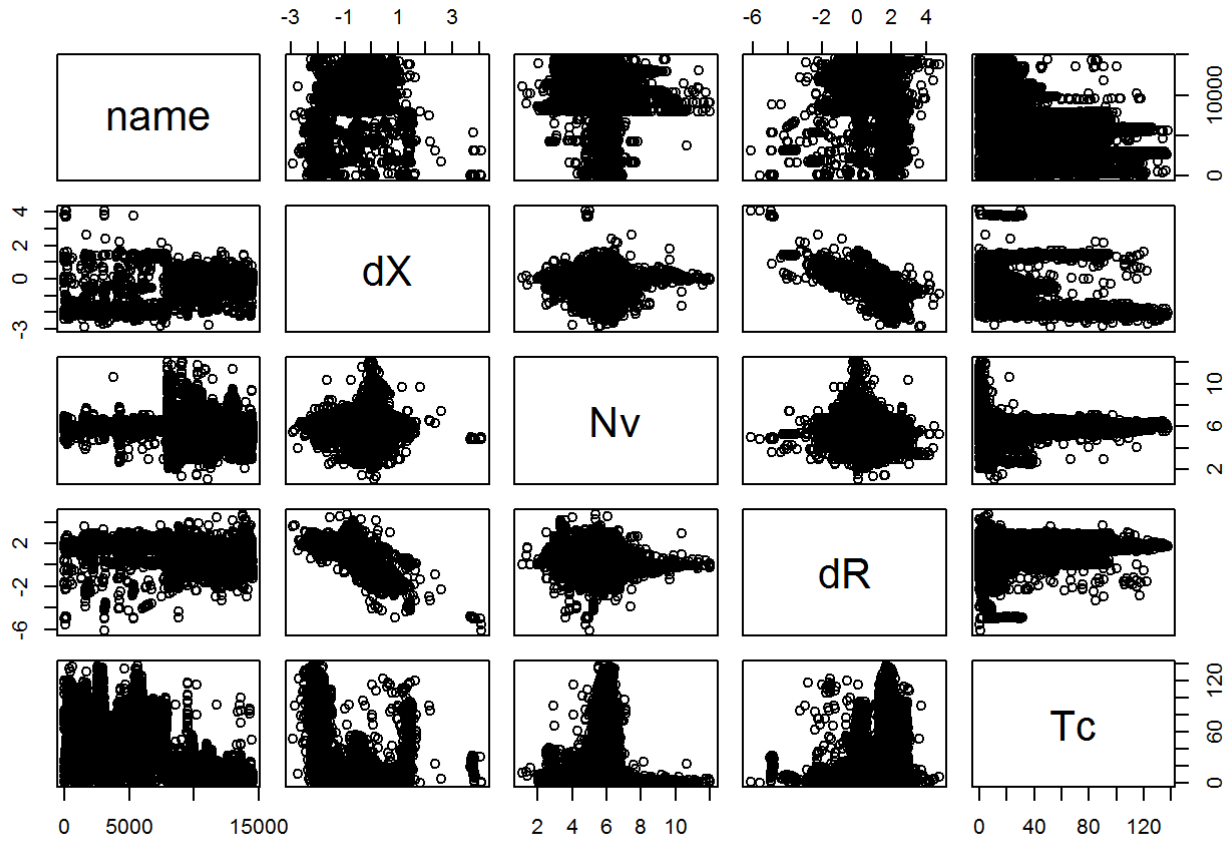

```
## png
##   3
```

```
## png
##   2
```

We noticed that all three feature ranges are similar(between -10 to 10). So we will skip the feature normalization before future analysis. Also notice that there are some peaks in the histogram data.

Feature ranges and pair plot of the features:

```
## 'data.frame':    14137 obs. of  5 variables:
##  $ name: Factor w/ 14528 levels "[Ag0.25Cu0.75Cu4Ba2Ca3]O11",..: 42 35 36 33 47 91 222 5286 5
285 5297 ...
##  $ dX  : num  -1.85 -1.85 -1.85 -1.85 -1.85 ...
##  $ Nv  : num  5.83 5.84 5.84 5.84 5.81 ...
##  $ dR  : num  1.66 1.67 1.65 1.65 1.66 ...
##  $ Tc  : num  29 26.5 19 26.7 23 ...
```




```
## png
##   3
```

```
## png
##   2
```

Model Building

First, a simple linear model is used to predict Tc:

```
library(caret)
model_lm <- train(Tc ~dX+Nv+dR, data=SCDB,method="lm")
finMod <-model_lm$finalModel
print(finMod)
```

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Coefficients:
## (Intercept)           dX           Nv           dR
##     -8.9490     -12.7270       5.2966       0.7141
```

```
rmse <- sqrt(mean((finMod$fitted.values-SCDB$Tc)^2))
```

The rmse is as large as 29.5196579 K. This means the relationship between Tc and the features are highly nonlinear.

```
m<- median(SCDB$Tc)
SCDB$label <-as.factor(ifelse(SCDB$Tc>m,"high","low"))
```

The median of the Tc is 18.7K. We split these superconductors into two parts with the same number: high Tc ones with label “high’ and low Tc with label”low“.

Next we will use three models to predict the labels in SCDB: generalized linear model(glm), decision tree and k-nearest neighbours. We will use 10-fold cross validation for all the models we use below. Since the prediction label is perfectly balanced, we will use the prediction precision to compare models.

Generalized Linear Modeling (glm)

First we will use Generalized Linear Modeling as our model.

```
train_control <- trainControl(method="cv", number=10)
model_glm <- train(label~dX+Nv+dR,data=SCDB,trControl=train_control, method="glm",tuneLength =
9)
pred_glm<- predict(model_glm , SCDB)
Pglm <- postResample(pred_glm, SCDB$label)[1]
```

The accuracy of the glm is 0.7934498.

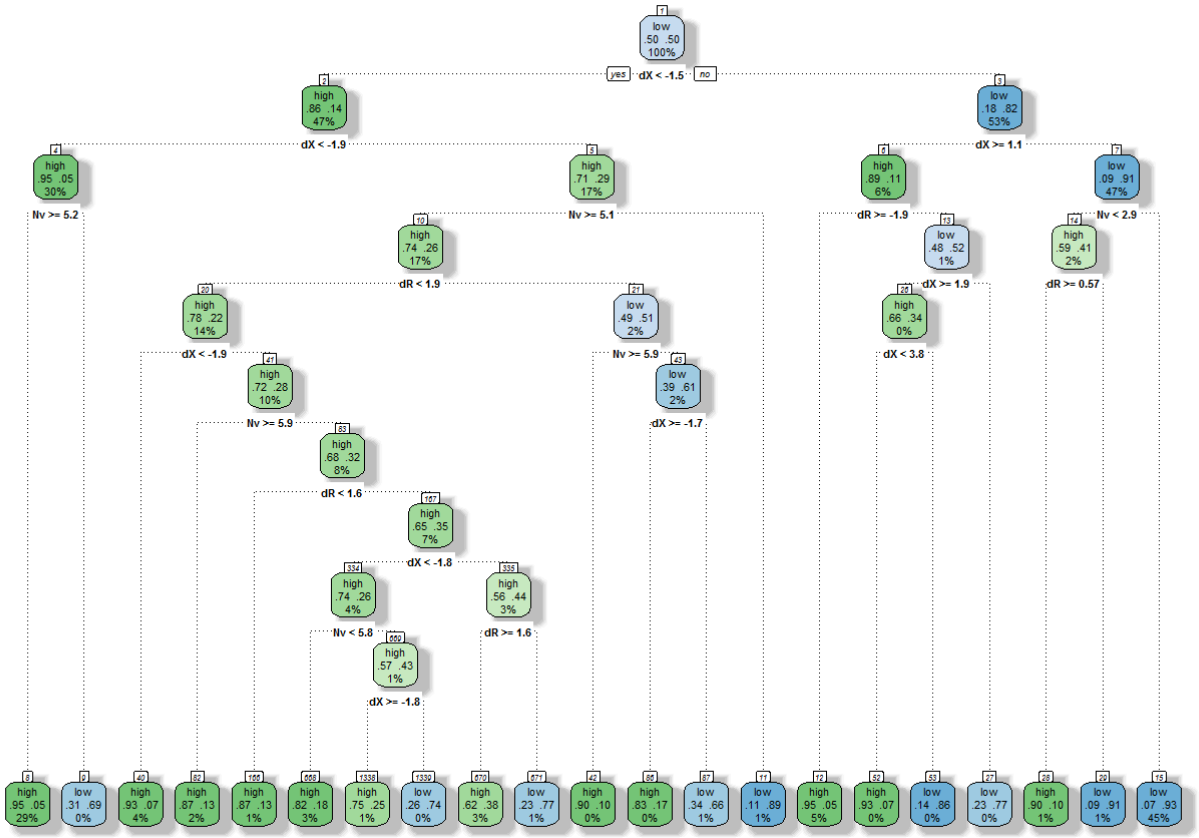
Decision Tree

Second we will use decision tree as our model.

```
train_control <- trainControl(method="cv", number=10)
model_tree <- train(label~dX+Nv+dR,data=SCDB,trControl=train_control, method="rpart",tuneLength
= 9)
pred_tree<- predict(model_tree , SCDB)
Ptree <- postResample(pred_tree, SCDB$label)[1]
```

The accuracy of the decision tree modeling is 0.913348.

This is how our decision tree looks like:

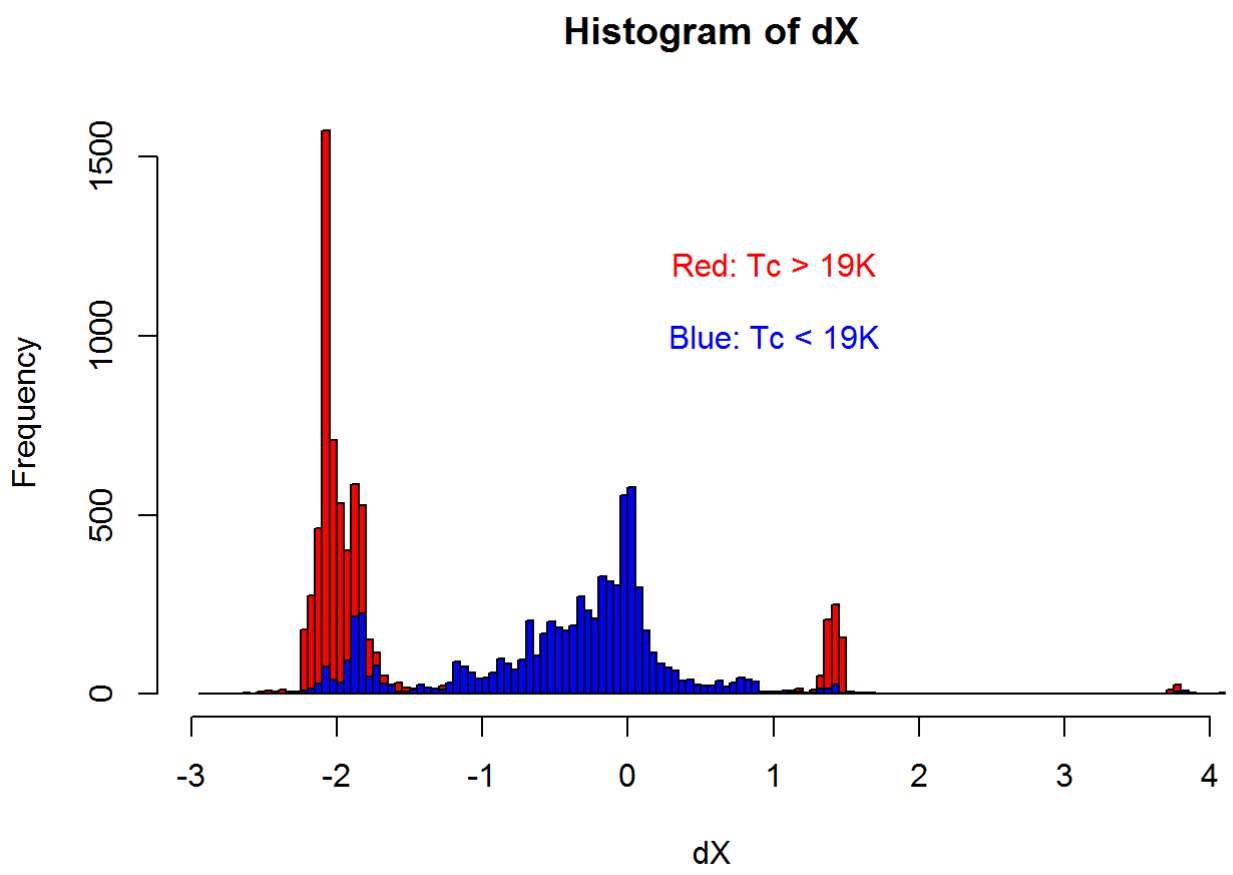


Rattle 2016-Jun-21 14:35:35 Yong

```
## png
##   3
```

```
## png
##   2
```

The decision tree uses dX as the first decision. This can be confirmed by the hisogram plot of dX labeled with high Tc or low Tc.




```
## png
##   3
```

```
## png
##   2
```

k-nearest neighbours modeling

Finally, knn is used:

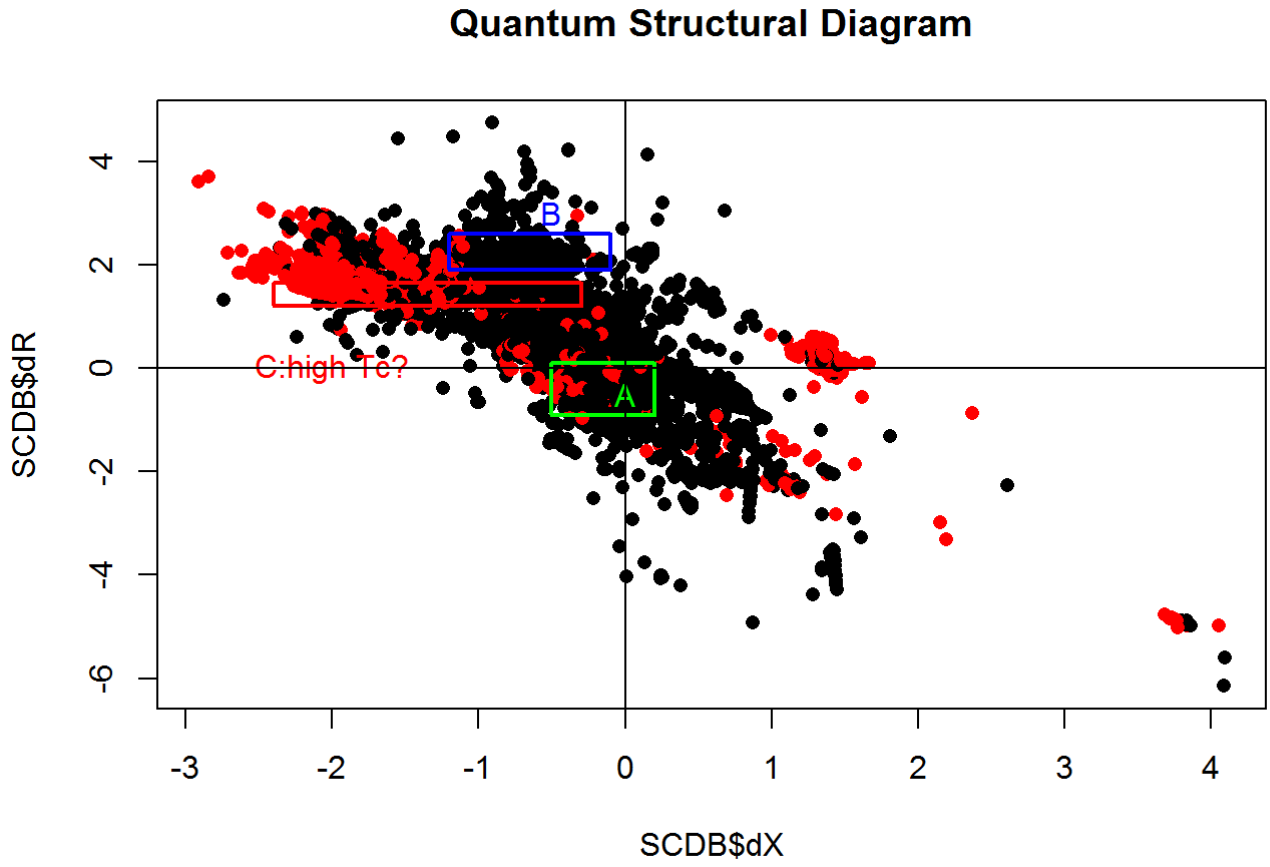
```
modelknn <- train(label~dX+Nv+dR, data=SCDB,method="knn",trControl=train_control,tuneLength = 9)
predknn <- predict(modelknn , SCDB)
Pknn <- postResample(predknn, SCDB$label)[1]
```

The accuracy of the knn modeling is 0.9405107. n = 5 is used in the final model. Recall that our glm precision is 0.7934498 and decision tree precision is 0.913348. Thus the best model in our prediction task is knn.

Sweet spots reidentification

In the original article, the authors' identified three regions in the feature space where it is more likely to find high Tc materials. Are these regions still the sweet spots after almost 30 years of discovery of new materials? Let us find it out.

The three rectangles A, B and C are the replicates from the original paper.Red dots indicates high Tc regions.




```
## png
##   3
```

```
## png
##   2
```

Bases on the this figure, it seems these three regions are not the sweet regions anymore. There might be some new regions emerging according to this figure.

Summary

Based on our analysis above, we drew the following conclusion from the data:

- k-nearest neighbours(KNN) with k = 5 gives the best model prediction on the high Tc superconductors.
- The three sweet regins discovered in 80s are no longer "sweet".
- dX, the ionic displacement, is the most important variable in determining high Tc properties.

If you have furthur questions or comments, please email: jieyong0731@gmail.com (mailto:jieyong0731@gmail.com).