Magpie tutorial

May 27, 2018

References

- A general-purpose machine learning framework for predicting properties of inorganic materials, npj Computational Materials, vol 2: 16028 (2016) doi:10.1038/npjcompumats.2016.28
- Paper: https://www.nature.com/articles/npjcompumats201628#supplementary-information

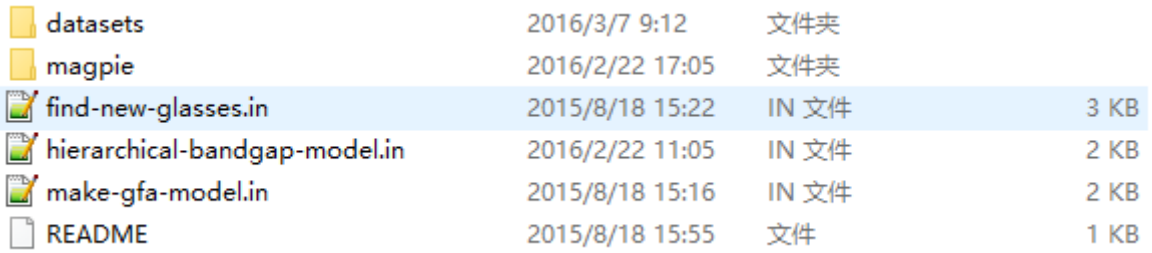
- Zipfiles: https://media.nature.com/original/nature-assets/npjcompumats/2016/npjcompumats201628/extref/npjcompumats201628-s2.zip

Java is a pre-request

- Check your Java: java –version (command line, Win: cmd or PowerShell; Mac: terminal)
- For Win, **Java SE Runtime Environment 8u171** is needed. Visit http://www.oracle.com/technetwork/java/javase/downloads/jre8-downloads-2133155.html
- For mac, **Java SE Development Kit 8u171** is needed. Visit http://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html
  don't download and install **Java SE 10.0.1**
- In case a wrong version is installed, to uninstall: https://docs.oracle.com/javase/10/install/installation-jdk-and-jre-macos.htm#JSJIG-GUID-F9183C70-2E96-40F4-9104-F3814A5A331F

Unzip the provided *.zip to a folder, eg. C:\

In the Code folder

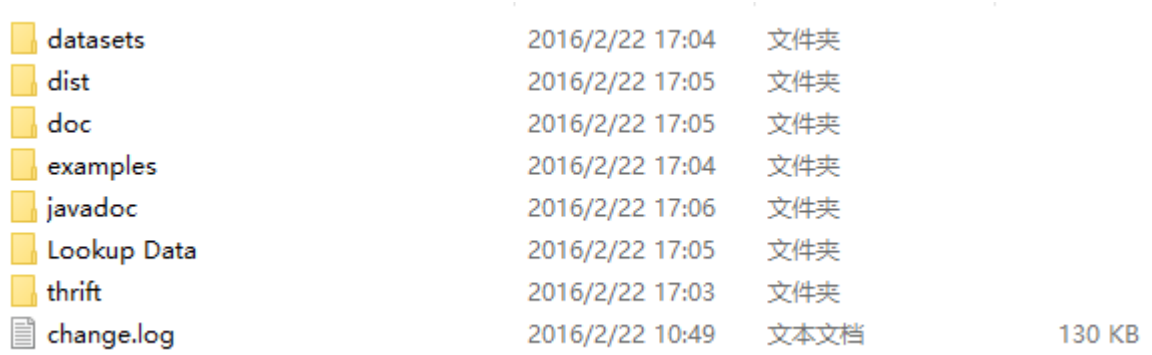


The Magpie folder:

Core code：code\magpie\dist\Magpie.jar
Examples：a couple of cases
Lookup data：element attributes tables
Doc：documentation

In the Datasets folder:

There are three *.data file. The “glass.data” is our training set, which can be edited by a text editor.



Comp	gfa{AM,AC,CR}
Ag20Al25La55	AM
Ag15Al10Mg75	AM
Ag25Al10Mg65	AM
Ag25Al20Mg55	AM
Ag35Al10Mg55	AM
Ag35Al20Mg45	AM
Ag45Al20Mg35	AM

Partial of the “glass.data”
Two columns

Running the code

运行软件-command line，Win: cmd or PowerShell; Mac: terminal
软件为命令行界面，需要通过命令行（shell）打开，同时，为了方便后续的操作，先将工作目录切换到code目录下。

Windows PowerShell

```
Windows PowerShell
版权所有 (C) Microsoft Corporation。保留所有权利。

PS C:\Users\15902>
```

（以win下的PowerShell举例）
打开后 PowerShell的当前目录为c盘user中
先切换到D盘中 （code的解压位置）

Windows PowerShell

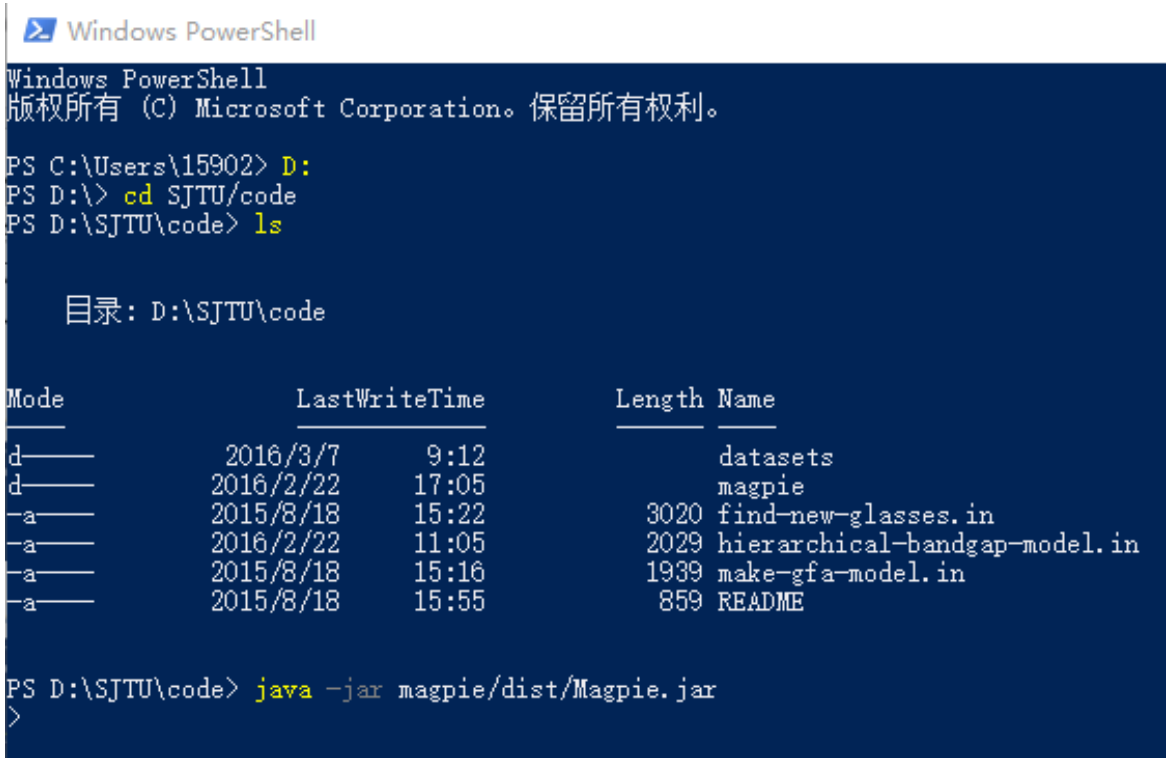
```
Windows PowerShell
版权所有 (C) Microsoft Corporation。保留所有权利。

PS C:\Users\15902> D:
PS D:\> cd SJTU/code
PS D:\SJTU\code> ls


    目录: D:\SJTU\code


Mode                 LastWriteTime         Length Name
----                 -------------         ------ ----
d-----         2016/3/7      9:12                datasets
d-----        2016/2/22     17:05                magpie
-a----        2015/8/18     15:22           3020 find-new-glasses.in
-a----        2016/2/22     11:05           2029 hierarchical-bandgap-model.in
-a----        2015/8/18     15:16           1939 make-gfa-model.in
-a----        2015/8/18     15:55            859 README


PS D:\SJTU\code>
```

进入D盘后，输入 cd （directory）进入code文件夹
进入code文件夹后，可以输入ls指令查看当前目录中的文件

使用的命令行指令：
cd +目录： 切换工作目录
ls： 展示文件夹下的所有文件

** win cmd中文件夹分割为\ 而非/
目录的方式有相对路径与绝对路径两种形式

Running the code



Go to the “code” folder

In the “code” folder, 打开程序的指令为：
java -jar magpie/dist/Magpie.jar

回车后，指令行变成”>“后 就表示已经进入了Magpie的软件

A simple case: running the code using a *.in file

Tip: 推荐使用*.in 文件编写再read的方式，这样方便反复使用以及记录。

程序的运行方式有两种：

1）直接在 > 符号后输入语句，回车后程序会执行相应的命令

2）通过写输入文件的方式

输入文件为 .in 文件，可以通过文本编辑器创建txt后，再修改后缀名为in

输入文件写完之后，通过**在 “>” 后输入 read 文件名.in 的形式执行**输入文件中的指令

（read命令读取文件为当前地址的文件，注意文件保存的位置，）

程序安装完成打开后可以用例子中的 simple-mode.in 实例文件作为例子，**初步跑通计算程序**。

01-simple-mode.in

```
data = new data.materials.CompositionDataset
data import magpie/datasets/small_set.txt

data attributes properties directory magpie/Lookup Data/

data attributes properties add set general
data attributes generate

data target delta_e

model = new models.regression.WekaRegression trees.RandomForest
model train $data
model crossvalidate $data 10

print model validation stats
```

该例子使用了small set这个示范数据集
数据中一共有630个示例
每个示例有化合物名称，以及7个计算得到的性质

该例子根据化合物名称计算特征值，并以delta_e 为目标进行回归，并通过10-fold cross-validation 来对模型进行判断。

该输入文件先读取这个数据集，并通过算法生成data同时建立model为使用的模型，并且用data对model进行训练，最终进行10-fold cross-validation

Code explanation

01-simple-mode.in

```
data = new data.materials.CompositionDataset
data import magpie/datasets/small_set.txt

data attributes properties directory magpie/Lookup Data/

data attributes properties add set general
data attributes generate

data target delta_e

model = new models.regression.WekaRegression trees.RandomForest
model train $data
model crossvalidate $data 10

print model validation stats
```

代码简析：

data = new data.materials.CompositionDataset
创建了一个新的变量
类似于编程语言中的（a=new int）

data import magpie/datasets/small_set.txt
将数据集导入到data这个变量中
形式:data import 文件位置
(注：输入命令中需要以空格分割)

data attributes properties directory magpie/Lookup Data/
指定生产变量需要使用的参考文件位置
形式:data attributes properties directory 文件位置

data attributes properties add set general
data attributes generate
这两天指令的作用是使data变量中产生文章中提及的特征值作为样本的特征值

data target delta_e
这条指令将机器学习的目标设定为delta_e这个值

至此数据的准备工作以及完成

Code explanation

01-simple-mode.in

```
data = new data.materials.CompositionDataset
data import magpie/datasets/small_set.txt

data attributes properties directory magpie/Lookup Data/

data attributes properties add set general
data attributes generate

data target delta_e

model = new models.regression.WekaRegression trees.RandomForest
model train $data
model crossvalidate $data 10

print model validation stats
```

model = new models.regression.WekaRegression trees.RandomForest
建立一个新的变量 model 它使用了RandomForest这个机器学习模型

model train $data
该命令用data 对模型进行训练
形式 model train 数据
$字符表示对变量内容进行访问

model crossvalidate $data 10
该命令是对model再data这个数据集合上进行
10重交叉验证

print model validation stats
输出刚刚验证的结果

Output

```
PS D:\SJTU\code> java -jar magpie/dist/Magpie.jar
>>
> read simple-model.in
data = new data.materials.CompositionDataset
data import magpie/datasets/small_set.txt
        Imported 612 entries
data attributes properties directory magpie/Lookup Data/
data attributes properties add set general
        Added 22 new properties.
data attributes generate
        Generated 145 attributes.
data target delta_e
        Set target property to delta_e
model = new models.regression.WekaRegression trees.RandomForest
model train $data
model crossvalidate $data 10
print model validation stats

Variable: model - Command: validation stats
Number Tested: 612
Pearson's Correlation (R): 0.9735
Spearman's Correlation (Rho): 0.9652
Kendall's Correlation (Tau): 0.8422
MAE: 1.8602e-01
RMSE: 2.6699e-01
MRE: 0.5296
ROC AUC: 0.9440
```

通过read 01-simple-model.in
程序会一步一步执行输入文件中的代码行

每一行的代码都会进行显示，执行完毕后也会输出一行执行的结果

Print语句之后，程序输出了最后的结果：
Number Tested: 612
Pearson's Correlation (R): 0.9735
Spearman's Correlation (Rho): 0.9652
Kendall's Correlation (Tau): 0.8422
MAE: 1.8602e-01
RMSE: 2.6699e-01
MRE: 0.5296
ROC AUC: 0.9440

1-3：预测与实际结果的相关系数
MAE：mean-absolute error
RMSE：root mean squared error
MRE：mean relative error
AUC：area under curve

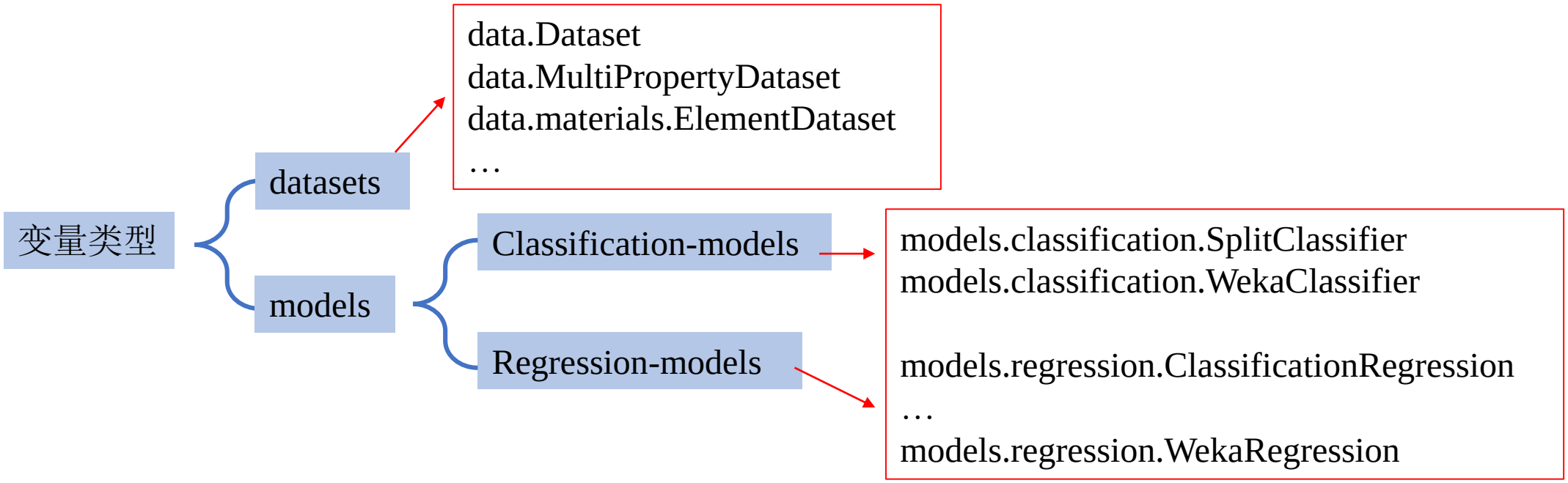
5. Magpie 命令语句形式

- 程序内部命令
  - read 文件名(.in 文件)
  - list (输出目前的变量和内容摘要)

```
> list
      Name                  Type    About
      data    CompositionDataset    Number of entries:  612 - Number of features: 145
     model       WekaRegression    Trained: true - Validated: true
```

  - ……
- 变量类型以及其对应的命令
  - 创建变量
    - data = new data.Dataset
    - 形式：<variable name> = new <class name> <options...>
  - 对变量进行操作
    - data import data.csv
    - 形式：<variable name> <command...>

6. Magpie中的主要内容： 变量类型



详细的变量介绍 请参考软件包中的 \code\magpie\doc文件夹中的 variables.html

如何使用 variables.html 这个命令说明文档？

网页文档中首先展示了所有可用的文档类型

Datasets

Each of these dataset objects can be used to represent different kinds of data, both in terms of how Magpi and what kind of attributes it can generate.

data.Dataset: Usage: *No options to set*
data.MultiPropertyDataset: Usage: *No options*
data.materials.CompositionDataset: Usage: *No options*
data.materials.ElementDataset: Usage: *No options*
data.materials.PrototypeDataset: Usage: <Structure description filename>

Models

Magpie is equipped with the ability to generate many different kinds of models. This includes models for clas or predicting the value of some property.

Classification Models

Classifiers are used decide which group an entry belongs out of a finite list of options.

models.classification.CumulantExpansionClassifier: Usage: <filename> <order>
models.classification.SplitClassifier: Usage: *No options*
models.classification.WekaClassifier: Usage: <Weka classifier> [<classifier options...>]

可以点击每一个变量类型，查看它可以执行的操作

如何使用 variables.html 这个命令说明文档？

创建变量

data = new data.Dataset

形式：<variable name> = new <class name> <options...>

对变量进行操作

data import data.csv

形式：<variable name> <command...>

指令的分析

Example1. data变量的 import 指令

命令的一般形式为： **[output =] <variable name> <command…>**

本例中不需要output

import <filename> [<options...>] – Import data by reading a file

filename: Name of file to import data from

options...: Any options used when parsing this dataset (specific to type of Dataset)

说明文档

data.csv

应用方式： data import data.csv

变量名

命令

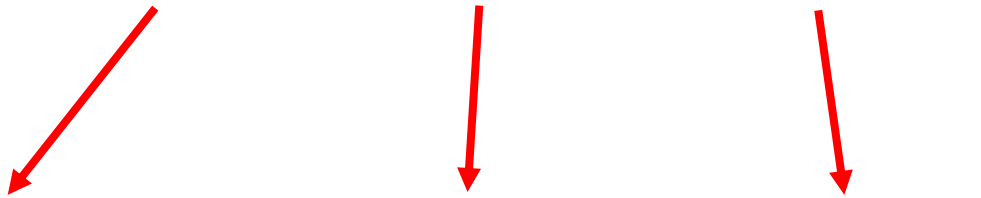
如何使用 variables.html 这个命令说明文档？

Example2. 如果想对data变量进行划分

1. 首先查看data变量的命令，找到可以实现这个功能的命令：

<output> = subset <number|fraction> – Generate a random subset from this dataset
number|fraction: Either the fraction or number of entries to select
output: New dataset containing random selection from this dataset

2. 命令的一般形式为： **[output =] <variable name> <command…>**



Subset 后的 Number|fraction 代表着可以接受两种输入方式，如果是小于1的数，就认为是一个fraction。

应用方式： datasubset = data subset 0.5

A harder case

背景介绍

数据集： glass.data

位置：code/datasets文件夹中

训练算法：wekaclassification 的随机子空间

要求：

将所有的混合类（AC）修改成（CR）

只用全部元素为金属元素的数据进行训练

对Zr-Ni-Cu相图进行预测

保存预测结果

glass.data数据形式：

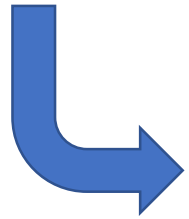
Comp：元素组成配比

Gfa：（三种可能的标志）

AM：amorphous

AC：AM and CR

CR：crystalline



```
comp gfa{AM,AC,CR}
Ag20Al25La55 AM
Ag15Al10Mg75 AM
Ag25Al10Mg65 AM
Ag25Al20Mg55 AM
Ag35Al10Mg55 AM
Ag35Al20Mg45 AM
Ag45Al20Mg35 AM
Ag10Ce6Cu84 AM
Ag10Ce10Cu80 AM
Ag15Ce6Cu79 AM
Ag20Ce6Cu74 AM
Ag20Ce10Cu70 AM
Ag25Ce6Cu69 AM
Ag30Ce6Cu64 AM
Ag30Ce10Cu60 AM
```

A harder case

输入文件: 02-harder-model.in

```
data = new data.materials.CompositionDataset
data import datasets/glass.data
data target gfa
data modify ClassEliminationModifier AC CR

data attributes properties directory magpie/Lookup Data/
data attributes properties add set general
data attributes generate

data filter include AllMetalsFilter
model = new models.classification.WekaClassifier &
        meta.RandomSubSpace -P 0.5 -S 1 -I 20 &
        -W weka.classifiers.trees.RandomForest -- &
        -I 10 -K 0 -S 1 -num-slots 1

model train $data

save model gfa

search = data clone -empty

gen = new data.utilities.generators.PhaseDiagramCompositionEntry
          3 3 -alloy 2 Zr Ni Cu

search target gfa

eval = new utility.tools.BatchModelEvaluator $model 10000
eval add $search $gen

save search Zr-Ni-Cu stats
```

创建一个data数据类型，用import方法导入glass.data 进行训练，将训练的目标定位 gfa （glass forming ability）

用modifier 方法，将原来的AC类全部修改为CR类，将3个类的分类变为2类分类问题

attribute generate 方法产生特征属性

data filter 方法对data变量进行筛查，剔除含有非金属元素的化合物成分，使得data只包含金属化合物

model设置为 WekaClassifier 后面的输入是这个学习算法的参数，它的具体算法为RandomSubspace

model train $data 来对模型进行训练，并将模型保存

建立一个search数据集，它复制了data这个数据集的模式(-empty)，但没有复制数据集的实际数据。将search的target也定位gfa

建立一个gen数据集，产生Zr，Ni，Cu三种元素以不同比例组成的，每隔2%含量变化的化合物集合

创建一个eval类，将model最为参数传入eval
在eval中添加search 和gen这两个数据，同时对gen进行预测
保存search 为Zr-Ni-Cu，文件格式为csv

A harder case

！说明：在使用中，variables.html 文档中只是简单得接受了各个变量可以使用的命令，但是没有具体介绍具体的一些参数。

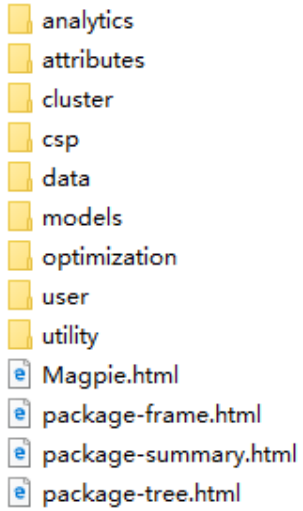
例如，我们要实现 filter 方法：data filter include AllMetalFilter

例如：

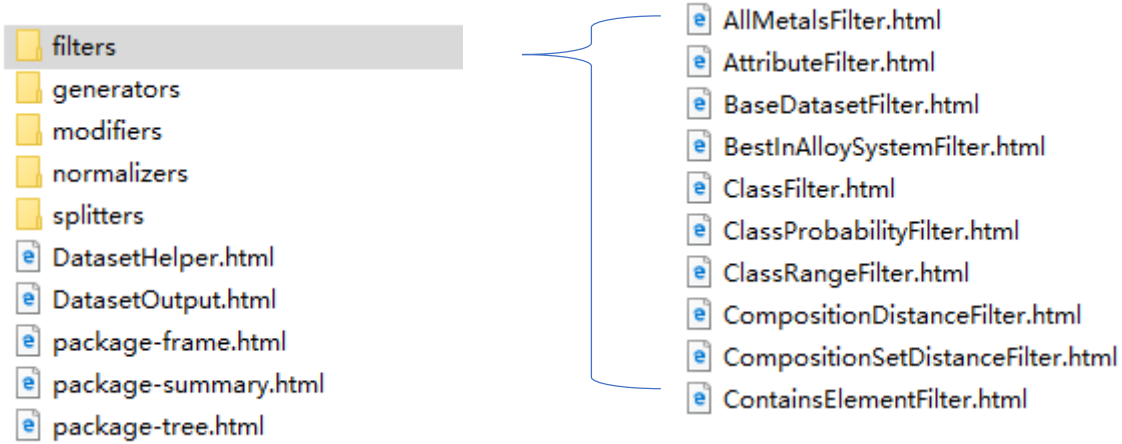
filter <include|exclude> <method> [<options...>] – Run dataset through a filter
include/exclude: Whether to include/exclude only entries that pass the filter
method: Filtering method. Name of a BaseDatasetFilter ("?" to print available methods)
options...: Options for the filter

说明中没有说明可以共选择的method，此时需要我们自己进入程序文件夹中搜索可以使用的方法。方法如下

进入 code\magpie\javadoc\magpie 文件夹



可以在data -> utilities 中找到filter一个文件夹
其中包含了各种功能的filter，以及他们的方法接受



其他方法也可以通过这个方式找到它的子功能以及使用的方法

A harder case: visualization using Excel

Output file: Zr-Ni-Cu.csv

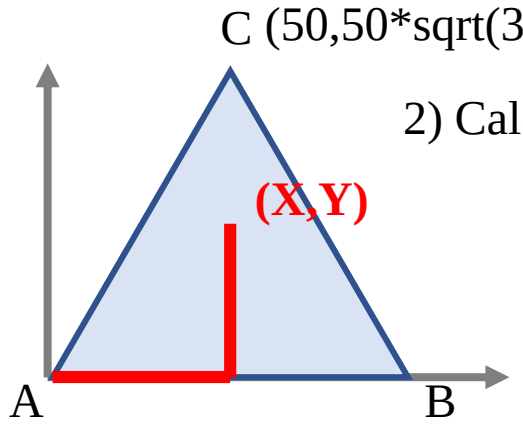
Result output:

- name
- measured=Nan (no measurement)
- predicted: predicted class
- probability: probability of the prediction

	A	B	C	D
1	name	measured	predicted	probability
2	Zr0.02Cu0.96Ni0.02	NaN	1.00E+00	9.05E-01
3	Zr0.02Cu0.94Ni0.04	NaN	1.00E+00	8.90E-01
4	Zr0.04Cu0.94Ni0.02	NaN	1.00E+00	9.20E-01
5	Zr0.02Cu0.92Ni0.06	NaN	1.00E+00	8.55E-01
6	Zr0.04Cu0.92Ni0.04	NaN	1.00E+00	8.85E-01
7	Zr0.06Cu0.92Ni0.02	NaN	1.00E+00	8.80E-01

First, convert name into composition

Zr=MID(A2,FIND("Z",A2)+2,FIND("C",A2)-FIND("Z",A2)-2)
Cu=MID(A2,FIND("C",A2)+2,FIND("N",A2)-FIND("C",A2)-2)
Ni=RIGHT(A2,LEN(A2)-FIND("N",A2)-1)



2) Calculate the coordinate of a point

X=100* (B%+0.5(1-A%-B%))
Y=100* sqrt(3)*0.5(1-A%-B%)

3) Get a list of coordinate X,Y and the target

	A	B	C
1	X	Y	probability
2	57	5.19	1
3	58	10.38	1
4	56	6.92	1
5	57	12.11	1
6	56	10.38	1
7	54	6.92	1
8	56	13.84	1
9	55	12.11	1
10	54	10.38	1
11	53	8.65	1
12	55	15.57	1
13	54	13.84	1
14	53	12.11	1
15	52	10.38	1

4) Sort according to probability (copy to a different sheet)

5）Select a color scale:

Prob>0.9

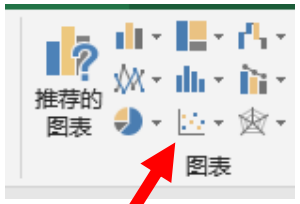
Prob>0.7

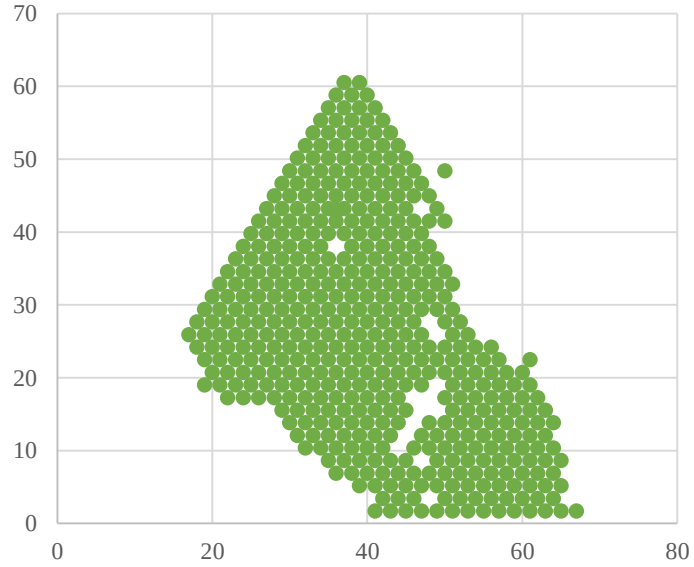
Prob>0.5

460	0.705	49	43.25
461	0.705	45	50.17
462	0.705	43	53.63
463	0.705	42	55.36
464	0.705	41	57.09
465	0.705	50	48.44
466	0.7	20	17.3
467	0.7	18	20.76
468	0.7	17	22.49
469	0.7	34	6.92
470	0.7	37	5.19
471	0.7	48	3.46

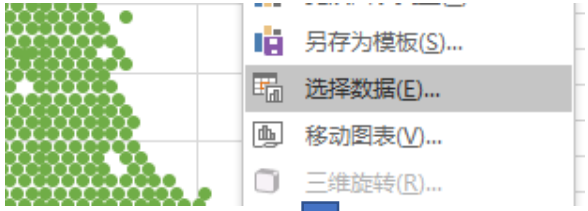
Plot each section with different probability with different color on the same graph

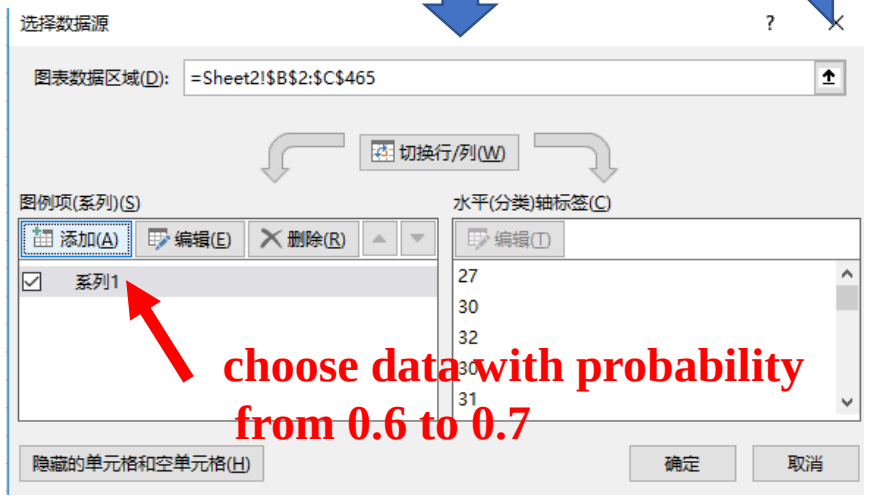
6) Choose the first group of data and plot scatter graph





7) Add data for the next section





8) Complete the graph