Mahalanobis Distance Based Kennard-Stone 1.0 (MDKS)

NANOBRIDGES
-ACollaborative Project

UNIWERSYTET GDAŃSKI

Marie Curie Actions
EUROPEAN UNION

THE AUTHORS ARE GRATEFUL FOR THE FINANCIAL SUPPORT FROM THE EUROPEAN COMMISSION THROUGH THE MARIE CURIE **IRSES** PROGRAM, NANOBRIDGES PROJECT (FP7-PEOPLE-2011-IRSES, GRANT AGREEMENT NUMBER 295128).

Last updated: 22.05.2015

Mahalanobis Distance based Kennard-Stone 1.0

Background:

The optimal division of dataset into training and independent test subset is an important and critical step in the QSAR modeling analysis. The test-set moleculeswill be predicted well when they are structurally verysimilar to the training-set molecules. The reason is that themodel wouldcapture all features common to the training-setmolecules.There are different techniques available fordivision of the data-set into training and test sets such asstatistical molecular design, self-organising map, clustering, Kennard–Stone (KS) selection, sphere exclusion, etc. [1]. Such rational way of division are superior to the simple random splitting and activity sorting method.
The MDKS algorithm which considers variabilities in both X and Y dimensions, was first proposed by Galvao et al.[2].In statistics, it has been reported that Mahalanobis distance (MD) gives better distance analysis compared with Euclidean approach especially in their applications for detecting detecting outliers [3].

About the algorithm: The algorithm involved in the calculation of MD is as follows:

$$C_x = \frac{1}{(n-1)}(X_c)^T(X_c) \qquad \text{(i)}$$

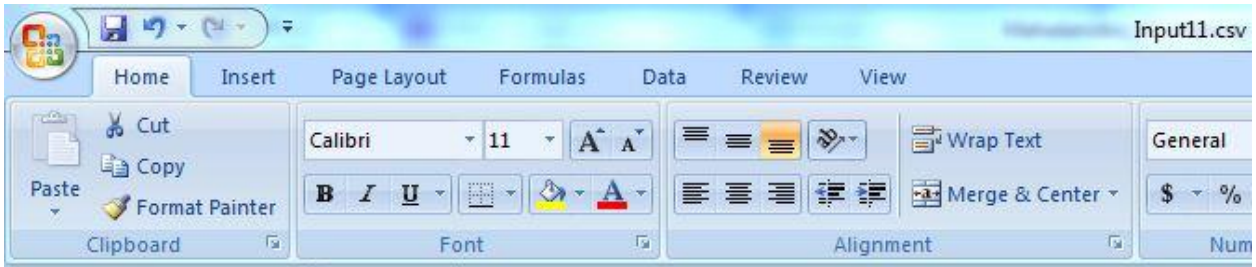
$$MD_i = \sqrt{(x_i - x)C_x{}^{-1}(x_i - \bar{x})^T} \qquad \text{(ii)}$$

Where, C_x is the variance-covariance matrix; X is the data matrix containing n objects. X_c is the column-centered data matrix ($X - \overline{X}$)
The eigen library was used for the calculation of matrix determinant and inverse calculation [4].

The KS algorithm was originally applied to generate a training set when no standard experimental design can be implemented. In this technique, all the objects are considered as candidates for the training set. The steps involved in the selection of training set based on MD-based KS algorithm are as follows [5-6]:

1. The first two compounds of training set are selected by choosing two compounds that are quite farthest apart in terms of MD.
2. Find the compound which has the maximum dissimilarity (maximum minimum distance) from each of the previously selected chemicals and place this chemical in the training set.
3. Repeat step 2 until the desired number of chemicals have been added to the training set.
4. The remaining chemicals were placed in the test set.

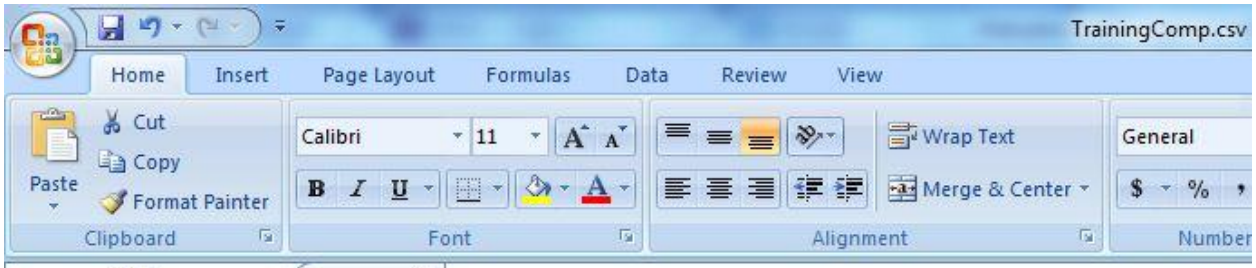
Input file format: This program takes a input file in CSV format.
Enter the non-standardized/non-scaled descriptors matrix.



	A	B	C	D	E	F	G	H	I	J
1	CompNo	S_sCH3	S_ssCH2	S_dsCH	S_aaCH	S_dssC	S_aasC	S_aaaC	S_ssssC	pkI(mM)
2	1	0.99786	-0.41755	1.05029	0.39528	0.68812	0.50956	-0.14907	0.17181	3.173925
3	4	-0.50166	-0.41755	1.00002	0.35905	0.68812	0.47428	-0.14907	0.17181	3.271646
4	5	2.11327	-0.41755	0.91396	-0.05687	0.10788	0.45357	-0.14907	0.17181	2.387216
5	6	-0.50166	-0.41755	0.87386	0.082	0.68812	0.31016	-0.14907	0.17181	2.156145
6	7	0.86771	-0.41755	0.93262	-0.34087	0.68812	0.48256	-0.14907	0.17181	2.910095
7	8	-0.50166	-0.41755	1.0653	0.35235	0.68812	0.51415	-0.14907	0.17181	3.133713
8	9	-0.50166	-0.41755	0.99624	0.22631	0.68812	0.3961	-0.14907	0.17181	2.126679
9	10	-0.50166	0.1371	1.08978	0.30244	0.68812	0.47101	-0.14907	0.17181	3.207608
10	12	-0.50166	-0.00186	0.88027	0.17143	0.68812	0.35283	-0.14907	0.17181	2.486782
11	13	-0.50166	1.13381	0.9997	0.37329	0.68812	0.44626	-0.14907	0.17181	3.332547
12	14	-0.50166	1.38075	1.28003	0.89713	0.68812	0.50564	-0.14907	0.17181	3.251812

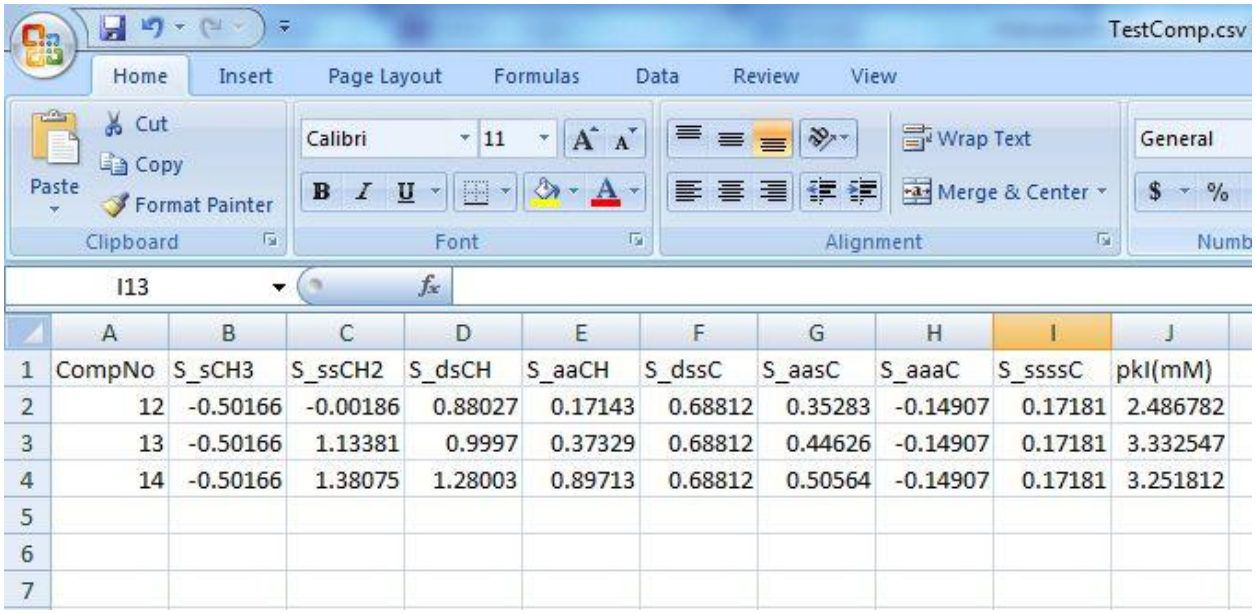
Output file format:

Training set compound



	A	B	C	D	E	F	G	H	I	J
1	CompNo	S_sCH3	S_ssCH2	S_dsCH	S_aaCH	S_dssC	S_aasC	S_aaaC	S_ssssC	pkI(mM)
2	1	0.99786	-0.41755	1.05029	0.39528	0.68812	0.50956	-0.14907	0.17181	3.173925
3	4	-0.50166	-0.41755	1.00002	0.35905	0.68812	0.47428	-0.14907	0.17181	3.271646
4	5	2.11327	-0.41755	0.91396	-0.05687	0.10788	0.45357	-0.14907	0.17181	2.387216
5	6	-0.50166	-0.41755	0.87386	0.082	0.68812	0.31016	-0.14907	0.17181	2.156145
6	7	0.86771	-0.41755	0.93262	-0.34087	0.68812	0.48256	-0.14907	0.17181	2.910095
7	8	-0.50166	-0.41755	1.0653	0.35235	0.68812	0.51415	-0.14907	0.17181	3.133713
8	9	-0.50166	-0.41755	0.99624	0.22631	0.68812	0.3961	-0.14907	0.17181	2.126679
9	10	-0.50166	0.1371	1.08978	0.30244	0.68812	0.47101	-0.14907	0.17181	3.207608
10										
11										

Test set compound



TestComp.csv

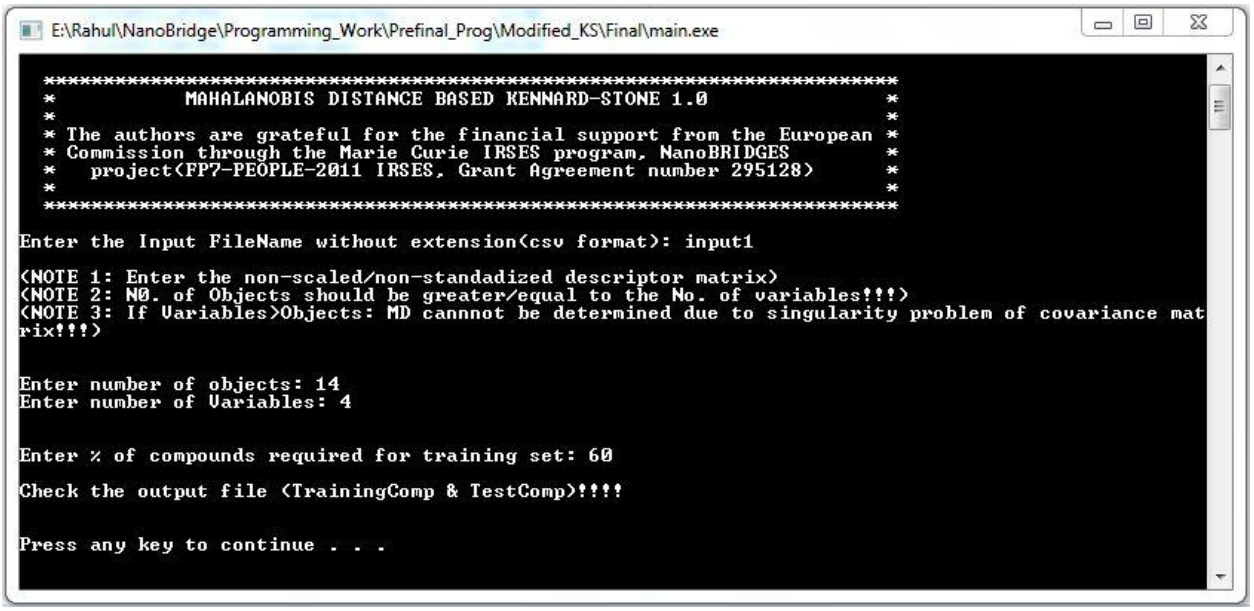
Home | Insert | Page Layout | Formulas | Data | Review | View

Cut, Copy, Format Painter, Paste, Clipboard, Calibri, 11, Font, Wrap Text, Merge & Center, Alignment, General, Numb

I13

	A	B	C	D	E	F	G	H	I	J
1	CompNo	S_sCH3	S_ssCH2	S_dsCH	S_aaCH	S_dssC	S_aasC	S_aaaC	S_ssssC	pkI(mM)
2	12	-0.50166	-0.00186	0.88027	0.17143	0.68812	0.35283	-0.14907	0.17181	2.486782
3	13	-0.50166	1.13381	0.9997	0.37329	0.68812	0.44626	-0.14907	0.17181	3.332547
4	14	-0.50166	1.38075	1.28003	0.89713	0.68812	0.50564	-0.14907	0.17181	3.251812
5										
6										
7										

How to run a program:



E:\Rahul\NanoBridge\Programming_Work\Prefinal_Prog\Modified_KS\Final\main.exe

```
*************************************************************************
*           MAHALANOBIS DISTANCE BASED KENNARD-STONE 1.0                *
*                                                                       *
* The authors are grateful for the financial support from the European *
* Commission through the Marie Curie IRSES program, NanoBRIDGES        *
*   project(FP7-PEOPLE-2011 IRSES, Grant Agreement number 295128)      *
*                                                                       *
*************************************************************************

Enter the Input FileName without extension(csv format): input1

(NOTE 1: Enter the non-scaled/non-standadized descriptor matrix)
(NOTE 2: N0. of Objects should be greater/equal to the No. of variables!!!)
(NOTE 3: If Variables>Objects: MD cannnot be determined due to singularity problem of covariance mat
rix!!!)


Enter number of objects: 14
Enter number of Variables: 4


Enter % of compounds required for training set: 60

Check the output file (TrainingComp & TestComp)!!!!


Press any key to continue . . .
```

Disclaimer

For academic purpose only.

The program MD-based Kennard-Stone 1.0 has been developed in C++ language and is validated on the known data sets. This program is compatible with both 32- and 64-bit Windows operating system. Please report for discrepancy of result for any other dataset. Contact us at any of the following addresses:

Dr. Tomasz Puzyn,
NanaBRIDGESProject
Coordinator,
Faculty of Chemistry,
University of Gdansk,
Gdansk,
Poland 80-952
Email Id: puzi@qsar.eu.org

Dr. Kunal Roy,
Drug Theoretics and Cheminformatics Lab.,
Dept. of Pharmaceutical Technology,
Jadavpur University,
Kolkata, West Bengal,
INDIA-700032
Email Id: kunalroy_in@yahoo.com

Rahul BalasahebAher,
PhdResearch Scholar,
Drug Theoretics and Cheminformatics Lab.,
Dept. of Pharmaceutical Technology,
Jadavpur University,
Kolkata, West Bengal,
INDIA-700032
E-mail Id: rahulba26@gmail.com(*for any queries regarding the program)

References:

1. K. Roy, On some aspects of validation of predictive QSAR models, *Expert Opin. Drug Discov*. 2 (2007), pp. 1567–1577.
2. Galvao, R.K.H., Araujo, M.C.U., José, G.E., Pontes, M.J.C., Silva, E.C. &Saldanha, T.C.B. 2005. A method for calibration and validation subset partitioning. *Talanta* 67, 736-740.
3. Maesschalck, R.D., Jouan-Rimbaud, D., &Massart, D.L. 2000. The Mahalanobis Distance.*Chemometrics and Intelligent Laboratory Systems*50(1), 1-18.
4. http://eigen.tuxfamily.org/index.php?title=Main_Page
5. Kennard, R. W.; Stone, L. A.*Computer Aided Design of Experiments Technometrics* 1969, 11, 137– 148.
6. Snarey, M.; Terrett, N. K.; Willet, P.; Wilton, D. J.Comparison of Algorithms for Dissimilarity-Based Compound Selection. *J. Mol. Graphics Modell.* 1997, 15, 372– 385