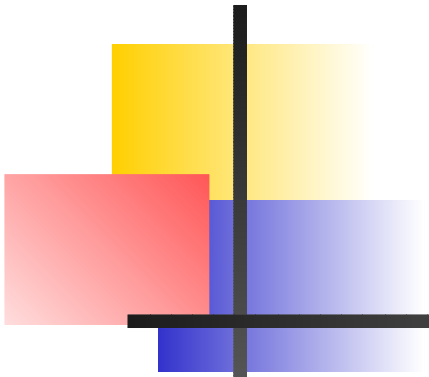

Sphere Exclusion Method for Set Generation

Rajarshi Guha

Penn State University

Dividing a Data Set

- Both the training & prediction set should be representative of the whole data set.
- Ideally, prediction set should also mirror the training set.

Methods for Set Selection

- Random.
- Activity Sampling.
- Clustering Methods.
  - KSOM.
  - K - means algorithm.
  - Kennard Stone.
  - Maximum Dissimilarity.

Problems with Clustering

- Different clusters have different density of points.
- Closeness of TSET & PSET is not gauranteed.

Sphere Exclusion

- Use *probe spheres* to set a similarity limit.
- Radii of the spheres is given by,

$$R = c\left(\frac{V}{N}\right)^{1/K}$$

- Depends on a user defined constant, c, called the *Disimilarity Level*.

Algorithm

1. Select compound with highest activity and add to TSET.
2. Construct sphere centered at this point, radius R.
3. All compunds within the sphere go into the TSET.
4. Exclude the points selected in 3 from the dataset.
5. If there are no more compounds, exit.

Algorithm

6. Calculate distance between all remaining compunds and all constructed sphere centers.
7. Select compunds with smallest (or largest) distance and go to step 2.

a

[a] A. Golbraikh et al, *J. Comp. Aid. Mol. Des.*

What Does It Give Us?

- Generates a TSET & PSET.
- It is difficult to exactly get a TSET of specified size.
- As a result we need to vary c by trial and error.
- Once you have the TSET, randomly select a CVSET (if required) from it.

Results

- Data Set
  - pcDHFR dataset.
  - 333 molecules.
- Holistic Descriptors
  - BCUT & Galvez topological indices (from Dragon).
  - 63 descriptors after reduction in Dragon.

Results

- Set Generation:
  - c = .3
  - TSET = 268, CVSET = 33, PSET = 32
- Details of the Study:
  - Original descriptor pool = 248
  - Reduced descriptor pool = 51
  - Type I, Type II & Type III models generated.

Type I Models

	SE	AB
	V7CH 19	V7CH 19
	N7CH 20	N7CH 20
	MOLC 8	NAB 15
	NAB 15	MDE 14
	WTPT 3	MDE 23
	SHDW 5	MDE 44
	lumo	PND 1
	NITR 3	PND 3
	CHAA 2	NITR 4
	CHAA 3	WPSA 3
R^2	.5358	.5125

V7CH: 7th order valence chain

N7CH: number of 7th order chains

MOLC: molecular connectivity

NAB: number of aromatic bonds

WTPT: sum of heteroatom ID's

SHDW: std shaadow area on XZ plane

CHAA: HBMIX descriptors

NITR: weighted at. surface area of N's

MDE: molecular distance edge desc.

PND: superpendentic index

WPSA: CPSA descriptor

[a]AB: Activity Binning, SE: Sphere Exclusion

Type I Outliers

- Sphere Exclusion: 2 outliers
- Acitivity Binning: 1 outlier
- Outliers are different for the two methods.

Type II

- 5 Descriptor Model (5-3-1)

Method	TSET	CVSET	PSET
AB	0.71	0.86	0.74
SE	0.68	0.71	0.80

- 7 Descriptor Model (7-3-1)

Method	TSET	CVSET	PSET
AB	0.64	0.70	0.81
SE	0.60	0.76	0.78

a

[a]AB: Activity Binning, SE: Sphere Exclusion

Type II

- 9 Descriptor Model (9-6-1)

Method	TSET	CVSET	PSET
AB	0.54	0.76	0.71
SE	0.65	0.73	0.79

a

[a]AB: Activity Binning, SE: Sphere Exclusion

Type III

	TSET		CVSET		PSET	
	AB	SE	AB	SE	AB	SE
5-3-1	0.61	0.62	0.61	0.68	0.70	0.82
7-3-1	0.58	0.55	0.57	0.58	0.73	0.98
8-3-1	0.56	0.53	0.61	0.52	0.77	0.93
9-3-1	0.59	0.56	0.62	0.54	0.73	0.86

a

[a]AB: Activity Binning, SE: Sphere Exclusion

Conclusions

- It seems to work well for Type I models, but this could be due to chance
- The method does not seem to be consistent
- There is no marked improvement in use sphere exclusion