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 guaranteed.
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 compounds 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.
6. Calculate distance between all remaining compounds and all constructed sphere centers.

7. Select compounds with smallest (or largest) distance and go to step 2.

\(^a\)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:
  - $c = 0.3$
  - $TSET = 268$, $CVSET = 33$, $PSET = 32$

- Details of the Study:
  - Original descriptor pool = 248
  - Reduced descriptor pool = 51
  - Type I, Type II & Type III models generated.
## Type I Models

<table>
<thead>
<tr>
<th>SE</th>
<th>AB</th>
</tr>
</thead>
<tbody>
<tr>
<td>V7CH 19</td>
<td>V7CH 19</td>
</tr>
<tr>
<td>N7CH 20</td>
<td>N7CH 20</td>
</tr>
<tr>
<td>MOLC 8</td>
<td>NAB 15</td>
</tr>
<tr>
<td>NAB 15</td>
<td>MDE 14</td>
</tr>
<tr>
<td>WTPT 3</td>
<td>MDE 23</td>
</tr>
<tr>
<td>SHDW 5</td>
<td>MDE 44</td>
</tr>
<tr>
<td>lumo</td>
<td>PND 1</td>
</tr>
<tr>
<td>NITR 3</td>
<td>PND 3</td>
</tr>
<tr>
<td>CHAA 2</td>
<td>NITR 4</td>
</tr>
<tr>
<td>CHAA 3</td>
<td>WPSA 3</td>
</tr>
</tbody>
</table>

$R^2$ | .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

*AB: Activity Binning, SE: Sphere Exclusion*
Type I Outliers

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

- 5 Descriptor Model (5-3-1)

<table>
<thead>
<tr>
<th>Method</th>
<th>TSET</th>
<th>CVSET</th>
<th>PSET</th>
</tr>
</thead>
<tbody>
<tr>
<td>AB</td>
<td>0.71</td>
<td>0.86</td>
<td>0.74</td>
</tr>
<tr>
<td>SE</td>
<td>0.68</td>
<td>0.71</td>
<td>0.80</td>
</tr>
</tbody>
</table>

- 7 Descriptor Model (7-3-1)

<table>
<thead>
<tr>
<th>Method</th>
<th>TSET</th>
<th>CVSET</th>
<th>PSET</th>
</tr>
</thead>
<tbody>
<tr>
<td>AB</td>
<td>0.64</td>
<td>0.70</td>
<td>0.81</td>
</tr>
<tr>
<td>SE</td>
<td>0.60</td>
<td>0.76</td>
<td>0.78</td>
</tr>
</tbody>
</table>

\(^a\)AB: Activity Binning, SE: Sphere Exclusion
Type II

9 Descriptor Model (9-6-1)

<table>
<thead>
<tr>
<th>Method</th>
<th>TSET</th>
<th>CVSET</th>
<th>PSET</th>
</tr>
</thead>
<tbody>
<tr>
<td>AB</td>
<td>0.54</td>
<td>0.76</td>
<td>0.71</td>
</tr>
<tr>
<td>SE</td>
<td>0.65</td>
<td>0.73</td>
<td>0.79</td>
</tr>
</tbody>
</table>

*AB: Activity Binning, SE: Sphere Exclusion*
# Type III

<table>
<thead>
<tr>
<th></th>
<th>TSET</th>
<th>CVSET</th>
<th>PSET</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AB</td>
<td>SE</td>
<td>AB</td>
</tr>
<tr>
<td>5-3-1</td>
<td>0.61</td>
<td>0.62</td>
<td>0.61</td>
</tr>
<tr>
<td>7-3-1</td>
<td>0.58</td>
<td>0.55</td>
<td>0.57</td>
</tr>
<tr>
<td>8-3-1</td>
<td>0.56</td>
<td>0.53</td>
<td>0.61</td>
</tr>
<tr>
<td>9-3-1</td>
<td>0.59</td>
<td>0.56</td>
<td>0.62</td>
</tr>
</tbody>
</table>

\(^a\)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.