Derivation of Relative Drug Likelihood

The Relative Drug Likelihood (RDL) approach is based on an assertion that a desirable value of a property is one that increases the probability of identifying a drug, not simply a value that is similar to known drugs. Bayesian probability theory allows us to quantitatively estimate this; Bayes' theorem states:

$$P(D|X) = \frac{P(X|D)P(D)}{P(X)}.$$

Here, P(D|X) is the probability of a compound being a drug given the value of a property X, in Bayesian terms this is known as the posterior. P(X|D) is the probability of the property X given that a compound is a drug, known as the likelihood. P(X) is the probability distribution for the property X for all compounds, whether drugs or not, and is known as the evidence. Finally, P(D) is the probability of a compound being a successful drug, given no further information, which is the 'prior probability' of a compound being a drug (a very small number, based on historical evidence!).

In traditional Bayesian inference, we would then use Bayes' theorem to compute the posterior probability of a compound not being a drug given the value of a property X:

$$P(D'|X) = \frac{P(X|D')P(D')}{P(X)}.$$

We could conceivably compare the posterior probability $P(D|X)$ against $P(D'|X)$ to determine whether or not to classify the compound as a drug; since the evidence P(X) is a constant, we need only compare the numerators $P(X|D)P(D)$ against $P(X|D')P(D')$. In the case of multiple properties $X_1, \dots, X_n$, assuming conditional independence of the properties (as in naive Bayes classification) would allow us to classify a compound as a drug or non-drug by comparing

$$P(D)\prod_{i=1}^{n} P(X_i|D)$$

against

$$P(D')\prod_{i=1}^{n} P(X_i|D').$$

However, since P(D) << P(D'), clearly we cannot use the above decision rule to simply classify a compound as a drug or non-drug based on its property values alone. Instead, given the simpler objective of determining if a compound is *more likely* to be a drug based on its property values, we take the ratio between the posterior probability of a compound being a drug and not being a drug to arrive at the equation:

$$\frac{P(D|X)}{P(D'|X)} = \frac{P(X|D)}{P(X|D')}\frac{P(D)}{P(D')}.$$

A desirable value for a property is one for which this ratio is relatively high, i.e. the probability of a compound being a drug is increased relative to the probability of it not being a drug. P(D) and P(D') are constants and hence this ratio is directly proportional to the ratio of the likelihoods of property X for drug and non-drugs. Therefore, our measure of desirability is the *relative likelihood*

$$d(x) = \frac{P(X=x|D)}{P(X=x|D')}.$$

We fit the desirability function for a property by binning the property values using the method described in Bickerton *et al.* [1]. As this method is sensitive to statistical anomalies caused by small numbers of compounds in a bin, bins containing less than 5 compounds are combined with neighbouring bins to ensure that each bin has more than this minimum number. The resulting desirability functions are estimated by fitting a curve to the desirability value for each bin using locally weighted scatterplot smoothing (LOESS) [2].

The individual desirability functions above can be combined into a metric, the Relative Drug Likelihood (RDL), by taking the geometric mean of the relative likelihoods of the individual properties:

$$\mathrm{RDL} = \exp\left(\frac{1}{n}\sum_{i=1}^{n} \ln\left(d_i(x_i)\right)\right).$$

References

1 Bickerton, G.R., Paolini, G.V., Besnard, J., Muresan, S., and Hopkins, A.L. (2012) Quantifying the chemical beauty of drugs. *Nature Chemistry*, 4, 90-98.

2 Cleveland, W.S. and Devlin, S.J. (1988) Locally-Weighted Regression: An Approach to Regression Analysis by Local Fitting. *Journal of the American Statistical Association*, 83, 596-610.