JOURNAL OF CHEMICAL INFORMATION AND MODELING

Erratum

pubs.acs.org/jcim

Erratum for "In Silico Target Predictions: Defining a Benchmarking Data Set and Comparison of Performance of the Multiclass Naïve Bayes and Parzen-Rosenblatt Window"

Alexios Koutsoukas, Robert Lowe, Yasaman KalantarMotamedi, Hamse Y. Mussa, Werner Klaffke, John B. O. Mitchell, Robert C. Glen,* and Andreas Bender*

J. Chem. Inf. Model. **2013**, *53* (8), 1957−1966. DOI:10.1021/ci300435j

Supporting Information

Incorrect handling during the data extraction step resulted in unsuitable ligand−protein pairs (inactive data points) reaching the final active data set that was subsequently employed in the study. To rectify the observed problem, the results reported in the study were re-evaluated by repeating the experimental steps and are presented and discussed below to verify that the conclusions previously made were not materially altered. Re-extraction of the data revealed that the correct number of ligand−protein annotations should have been 132 281 instead of the 155 208 that were previously reported. The inclusion of incorrect data points was expected to decrease the performance of the models and thus a re-evaluation of the experimental process was performed.

1. DISTRIBUTION OF ACTIVITY CLASSES IN THE RE-EXTRACTED DATA SET

Re-extracting the data set and subsequently retaining only the activity classes with at least 20 data points reduced the number of kinases in the final data set. The total number of activity classes counted in the re-extracted data set was 678, instead of the 895 that were previously reported. Kinases were the protein family that was mostly affected, where the initial 307 activity classes were reduced to 111, distribution of protein classes in the re-extracted data set is shown in Figure 1e (where e stands for erratum). The proportion of kinases among the activity classes was reduced from an initial 34% to 16% in the final data set.

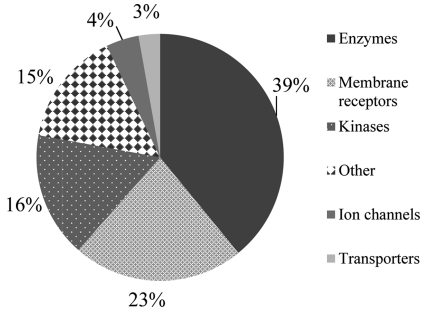

Figure 1e. Protein target class distribution in the training data set employed for the re-extracted data set. It can be seen that the majority of activity classes are enzymes, membrane receptors, and kinases. Kinases were reduced from an initial 34% to 16% in the final data set.

2. RE-EVALUATION OF RETRIEVAL RATE OF TRUE POSITIVES IN TOP-*K* POSITIONS

In this section both algorithms were re-evaluated using 5-fold cross-validation (5FCV) on the re-extracted data sets and the results obtained were compared to the previously reported ones to investigate the effects of inclusion of incorrect data points. Recall of true positives among the top-*k* positions by both algorithms was improved as shown in Figure 2e and Table 1e. Performance, as measured by recall of true positives (%) in the top-3 positions, was observed to increase by 7.2% (from the previously reported 78% to 85.2%) and by 6.5% (from the previous 74% to 80.5%) for the PRW and NB respectively, averaged over 5FCV. A comparison of the newly obtained results with the previously reported ones is shown in Table 1e.

3. RE-EVALUATION OF CLASS SIZE AND STRUCTURAL DIVERSITY ON THE PERFORMANCE OF THE MODELS

The effect of class size and intraclass structural diversity on the performance of the models was re-evaluated and is presented in

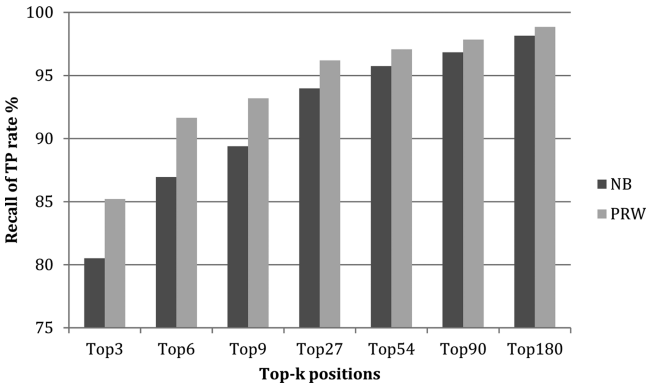

Figure 2e. Recall of true positive rate (%) achieved by Naïve Bayes (NB) and Parzen-Rosenblatt window (PRW) among the top 3, 6, 9, 27, 54, 90, and 180 positions, respectively, of the data set. Both algorithms retrieved more than 80% of the correct labels among the top-3 positions, where NB achieved performance of 80.5% and the PRW retrieved 85%. Furthermore, among the top-9 positions the PRW's performance exceeded 90%, achieving a recall of 93.2%, while NB achieved 89.4%.

Received: May 13, 2014
Published: June 18, 2014

Table 1e. Recall of True Positives among the Top-*k* Positions Achieved by the Parzen-Rosenblatt Window (PRW) and Naïve Bayes (NB) Classifiers[a]

Recall of TP in the Top-*k* Positions					
reported in the original paper			re-evaluated results		
top-*k* positions	PRW	NB	top-*k* positions	PRW	NB
3	78%	74%	**3**	85.2%	80.5%
6	86%	80.5%	**6**	91.6%	86.9%
9	87.6%	83.1%	**9**	93.2%	89.4%
27	90.5%	87.2%	**27**	96.2%	94%
54	91.8%	88.3%	**54**	97.1%	95.7%
90	92.2%	89.4%	**90**	97.8%	96.8%
180	93.4%	91.1%	**180**	98.9%	98.2%

[a]Recall reported previously in this study is shown on the left, while the re-evaluated recall is shown on the right. Removal of incorrect data points resulted in the overall improvement of the observed recall of true positive instances in the top-3 positions averaged over 5FCV by 7% (from 78% to 85.2%) and 6% (74% to 80.5%) for PRW and NB, respectively. The improvement in the top-9 positions was measured to be 5.6% (from 87.6% to 93.2%) and 6.3% (from 89.4% to 93.2%) for the PRW and NB, respectively.

Figure 3e. Here, similar behavior was observed to that previously reported in Figure 4 of the original study. Activity classes with average intraclass Tanimoto similarity (Tc) higher or equal of 0.4 (classes with low structural diversity) perform well overall and are retrieved among the top 100 positions (shown on the axis on the left). Similar performance was observed among classes with a large number of (≥200) data points. In contrast, classes with a small number of data points and low intraclass structural similarity (middle bottom section) remained more difficult to predict and presented larger variation in performance (distribution on the vertical *z*-axis).

A re-examination of the protein classes that were difficult to predict is presented in Figure 4e. Here it can be observed that kinase classes, despite being reduced in number in the final data set, remained among the most challenging classes to classify.

4. EXTERNAL DATA SET

A re-evaluation of the performance of the two machine learning algorithms on the external data set assembled from WOMBAT is presented in Figure 5e. Here similar results were obtained to those previously reported in Figure 7 of the original study. Similarly to the previous results, the NB demonstrated better performance than the PRW for those compounds with Tanimoto similarity less than or equal to 0.4, while the PRW performed better for compounds with similarity of 0.4 or higher to the active training data set in ECFP_4 space.

5. MULTICLASS LAPLACIAN MODIFIED NAÏVE BAYES CLASSIFIER

(a) In eq 4, it should read

$$p(\omega_\alpha|\boldsymbol{x}) = \frac{p(\omega_\alpha)}{p(\boldsymbol{x})}\left[\prod_{i=1}^{d}\frac{p(\omega_\alpha|f_i)}{p(\omega_\alpha)}p(f_i)\right]$$

(b) In eq 6 (and in subsequent equations), $(f_i|\omega_\alpha)$ should read $(\omega_\alpha|f_i)$.

(c) In eq 7, $p'^{(f_i|\omega_\alpha)}$ should read $p'(\omega_\alpha|f_i)$.

(d) In eq 9, a term is missing—i.e., this equation should read

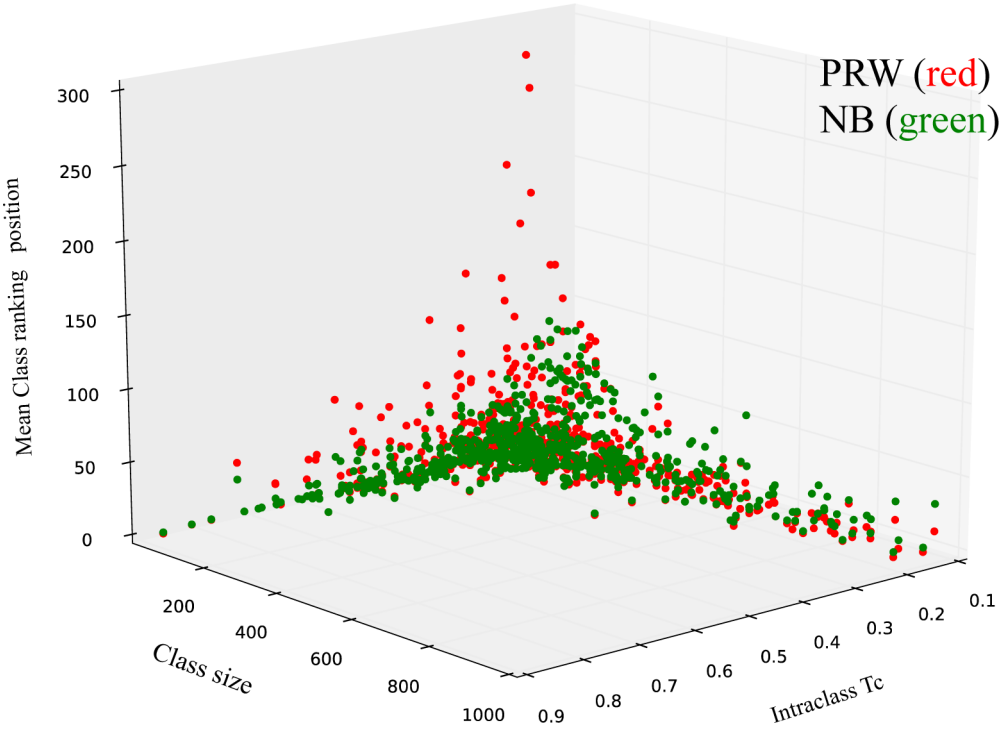

Figure 3e. Effect of class size and intraclass Tanimoto similarity (in ECFP_4 space) on the performance of the Parzen-Rosenblatt window (PRW) and Naïve Bayes (NB) classifiers, respectively. Classes with intraclass Tc ≥ 0.4 perform well overall and are retrieved in the top 100 positions or better (shown on the axis on the left). Similar performance was observed among classes with a large number of ≥200 data points. In contrast, classes with a small number of data points and low intraclass Tc (middle bottom section) are more difficult to predict and present larger variation in performance (distribution on the vertical *z*-axis).

dx.doi.org/10.1021/ci500255v | J. Chem. Inf. Model. 2014, 54, 2180−2182

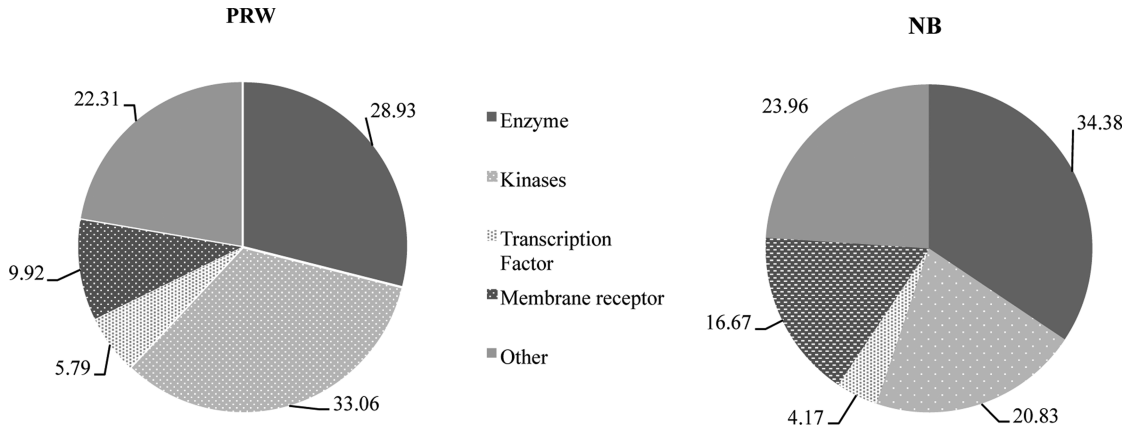

Figure 4e. Distribution of protein targets found among classes with "poor performance" (defined as classes with a mean ranking below the top 27 positions). Kinases and enzymes were found to be among the most challenging classes to classify, together constituting 62% of classes for PRW and 55% for NB.

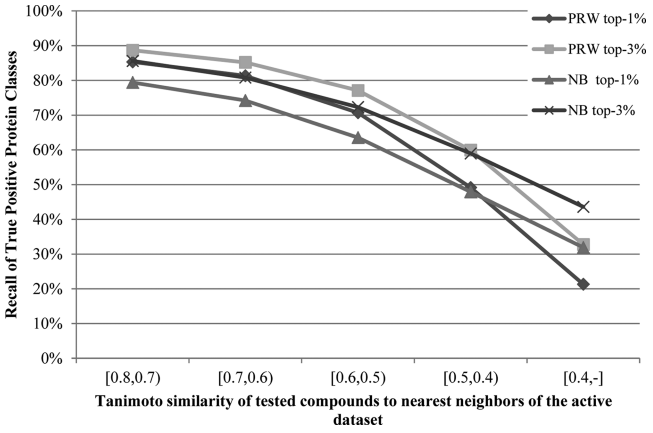

Figure 5e. Re-evaluated recall for the external data set among the top-9 and top-27 positions achieved by NB and PRW versus the similarity to the nearest neighbor of the training set measured by Tanimoto coefficient in ECFP_4 space.

$$S_{\omega\alpha}(\boldsymbol{x}) = \Sigma_i f_i \log\left[\frac{N^+_{i\omega_\alpha} + 1}{N_i^{+} * p(\omega_\alpha) + 1}\right] + \log\frac{\Pi_{i=1}^{d} p(f_i)}{p(\boldsymbol{x})}$$

In all the above, ω_α, $\boldsymbol{x}$, and f_i are as described in the original publication. d is the dimensionality of the feature vectors.

■ ASSOCIATED CONTENT

Supporting Information

This material is available free of charge via the Internet at http://pubs.acs.org.

dx.doi.org/10.1021/ci500255v | J. Chem. Inf. Model. 2014, 54, 2180−2182