JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

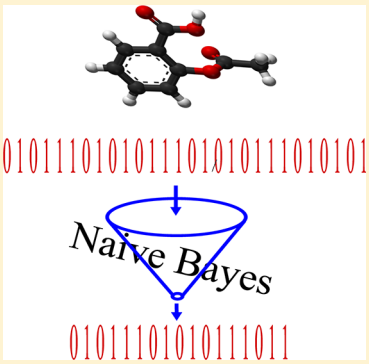
pubs.acs.org/jcim

Note on Naive Bayes Based on Binary Descriptors in Cheminformatics

Joe A. Townsend, Robert C. Glen, and Hamse Y. Mussa*

Unilever Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, U.K.

ABSTRACT: A plethora of articles on naive Bayes classifiers, where the chemical compounds to be classified are represented by binary-valued (absent or present type) descriptors, have appeared in the cheminformatics literature over the past decade. The principal goal of this paper is to describe how a naive Bayes classifier based on binary descriptors (NBCBBD) can be employed as a feature selector in an efficient manner suitable for cheminformatics. In the process, we point out a fact well documented in other disciplines that NBCBBD is a linear classifier and is therefore intrinsically suboptimal for classifying compounds that are nonlinearly separable in their binary descriptor space. We investigate the performance of the proposed algorithm on classifying a subset of the MDDR data set, a standard molecular benchmark data set, into active and inactive compounds.



INTRODUCTION

For illustrative purposes, attention is confined to two-class pattern recognition problems which are ubiquitous in cheminformatics. Generally, in these scenarios, one desires to make predictions on whether a given compound belongs to one of two predefined categories (or classes), based on a set of molecular properties or descriptors. (The generalization to multiclass classification problems will be described briefly.)

A common basic assumption is that a sample data set $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ is sampled from some relevant and underlying population to which both the known and unknown compounds belong. Here N refers to the size of the given sample data set; $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{il}, ..., x_{iL})$ is an L-dimensional descriptor vector representing compound i, where the elements x_{il} are discrete binary values—i.e., the x_{il} attribute is either present ($x_{il} = 1$) or absent ($x_{il} = 0$) in compound i; $y_i \in \{0,1\}$ is the class label of compound i, with $y_i = 1$ if $\mathbf{x}_i$ comes from class one denoted here by ω_1; otherwise, the compound comes from class two denoted here by ω_2.

One way of classifying a compound $\mathbf{x}_i$ is to model the probability $p(\omega_m|\mathbf{x}_i)$ that it belongs to class ω_m (with $m = 1, 2$). Using the given sample, there are various approaches to estimate $p(\omega_m|\mathbf{x}_i)$. In the following discussion, we follow closely the Bayes description of Duda and Hart.[1] For the sake of clarity in the discussion the index i in $\mathbf{x}_i$ is omitted when there is no risk of confusion. In the rest of the paper, the terms "descriptors", "features", and "attributes" are used interchangeably and $m = 1$ or 2 unless stated otherwise.

In the Bayes approach, $p(\omega_m|\mathbf{x})$ can be estimated as

$$p(\omega_m|\mathbf{x}) = \frac{p(\omega_m)p(\mathbf{x}|\omega_m)}{p(\mathbf{x})} \tag{1}$$

where $p(\mathbf{x}) = \sum_{m=1}^{2} p(\omega_m)p(\mathbf{x}|\omega_m)$; $p(\omega_m)$ denotes the a priori class probability; in our context, $p(\mathbf{x}|\omega_m)$ is a class conditional probability.

In the naive Bayes case, it is assumed that conditional on the class ω_m, the elements x_l of the pattern vector $\mathbf{x} = (x_1, x_2, ..., x_l, ..., x_L)$ are independent. This assumption greatly simplifies the estimation of $p(\omega_m|\mathbf{x})$ in eq 1 to

$$p(\omega_m|\mathbf{x}) = \frac{p(\omega_m)\prod_{l=1}^{L} p(x_l|\omega_m)}{p(\mathbf{x})} \tag{2}$$

with

$$p(\mathbf{x}|\omega_m) = \prod_{l=1}^{L} p(x_l|\omega_m) \tag{3}$$

where L denotes the number of distinct descriptors assumed to capture the relevant characteristics and properties of the molecule. Since, in the pattern recognition problem we are concerned with, x_l is discrete and binary-valued, i.e., $x_l \in \{0,1\}$, $p(x_l|\omega_m)$ can be written in the form

$$p(x_l|\omega_m) = \begin{cases} \mu_{l\omega_m}, & \text{if } x_l = 1 \\ 1 - \mu_{l\omega_m}, & \text{if } x_l = 0 \end{cases}$$

Inserting this into eq 3 gives[1]

$$p(\mathbf{x}|\omega_m) = \prod_{l=1}^{L} \mu_{l\omega_m}^{x_l} (1 - \mu_{l\omega_m})^{1-x_l} \tag{4}$$

where $\mu_{l\omega_m} = (1 + J_{x_l|\omega_m})/(2 + N_{\omega_m})$ is a Laplacian corrected[2] first-order class conditional statistical moment; $J_{x_l|\omega_m}$ denotes the total number of times that the lth element of the descriptor vectors $\mathbf{x}$

Received: July 3, 2011

   dx.doi.org/10.1021/ci200303m | J. Chem. Inf. Model. XXXX, XXX, XXX–XXX

that belong to class ω_m takes a value 1, and N_{ω_m} is the total number of descriptor vectors $\mathbf{x}$ in class ω_m.

Inserting eq 4 into eq 2, the resulting expression can be rearranged as

$$d_{\omega_m}(\mathbf{x}) = p(\mathbf{x})p(\omega_m|\mathbf{x}) = p(\omega_m)\prod_{l=1}^{L}\mu_{l\omega_m}^{x_l}(1-\mu_{l\omega_m})^{1-x_l} \quad (5)$$

which can be viewed as a discriminant function.[1,3]

Making use of the fact that discrimination functions are not unique and that monotonic mappings do not affect the rank order among values to be compared,[1,3] one can take the logarithm of eq 5 and by doing so modify it to

$$f_{\omega_m}(\mathbf{x}) = A_{\omega_m} + \mathbf{a}_{\omega_m}{}^{\mathrm{T}}\mathbf{x} \quad (6)$$

where $f_{\omega_m}(\mathbf{x})$ refers to $\ln d_{\omega_m}(\mathbf{x})$ and A_{ω_m} and $\mathbf{a}_{\omega_m}$ are defined as

$$\mathbf{a}_{l\omega_m} = \ln\frac{\mu_{l\omega_m}}{1-\mu_{l\omega_m}} \quad (7)$$

$$A_{\omega_m} = \ln p(\omega_m) + \sum_{l=1}^{L}\ln(1-\mu_{l\omega_m})$$

respectively, where L and the subscripts of $\mathbf{a}_{l\omega_m}$ and A_{ω_m} are as described before.

Obviously, eq 6 is linear with respect to the descriptor vector $\mathbf{x}$; hence, a naive Bayes classifier based on binary descriptors (NBCBBD) is suboptimal (with possible large bias) for classifying nonlinearly separable chemical data sets. The suboptimality of NBCBBD for nonlinear classification problems is a widely known and well-documented concept in the fields of machine learning and statistical pattern recognition. The interested reader is referred to refs 1 and 3. In the following section, we describe how eqs 6 and 7 can be used for selecting relevant cheminformatics descriptors both in two-class and multiclass cheminformatics classification problems.

In passing we note that, for a binary classification, the problem that we have so far confined attention to, an alternative discrimination function, is

$$f(\mathbf{x}) = f_{\omega_1}(\mathbf{x}) - f_{\omega_2}(\mathbf{x}) = (A_{\omega_1} - A_{\omega_2}) + (\mathbf{a}_{\omega_1}{}^{\mathrm{T}} - \mathbf{a}_{\omega_2}{}^{\mathrm{T}})\mathbf{x} \quad (8)$$

where $A_{\omega_1} - A_{\omega_2} = \ln[p(\omega_1)/p(\omega_2)] + \sum_{l=1}^{L}\ln[(1-\mu_{l\omega_1})/(1-\mu_{l\omega_2})]$; $\mathbf{a}_{l\omega_1} - \mathbf{a}_{l\omega_2} = \ln[(\mu_{l\omega_1}(1-\mu_{l\omega_2}))/(\mu_{l\omega_2}(1-\mu_{l\omega_1}))]$, which is commonly referred to as the Odds-Ratio and employed as a feature scoring measure, particularly in Information Retrieval.[4] The Odds-Ratio approach is basically "relevance weighting of search terms (features)".[4] Unfortunately, the Odds-Ratio method sweepingly favors the descriptors associated with large $|\mathbf{a}_{l\omega_1} - \mathbf{a}_{l\omega_2}|$ values and can only be generalized to multiclass problems providing that the multiclass problem is turned into a pairwise separation problem first. These issues may render Odds-Ratio based feature selection schemes unsuitable in the cheminformatics context.

To the best of our knowledge, the class-conditional descriptor independence assumption was relaxed—in a non-naive Bayes setting—in a few (cheminformatics) classification studies.[5−8] For more details on Bayes methods and their application to cheminformatics, the reader is referred to these recent reviews.[9,10]

PROPOSED FEATURE SELECTION SCHEME

The nub of the paper is eq 7. A closer look at this equation reveals that $\mathbf{a}_{l\omega_m} \to 0$ as $\mu_{l\omega_m} \to 0.5$. On the other hand, the values of $\mathbf{a}_{l\omega_m}$ increase sharply as $\mu_{l\omega_m} \to 1$ or 0; where $m = 1, 2$. This means that a descriptor $\mathbf{x}_l$ whose $\mu_{l\omega_1}$ (or $\mu_{l\omega_2}$) is equal (or close) to 0.5 can be omitted from the pattern vector, because this descriptor is less discriminating for both classes. Descriptors associated with small values of $|\mathbf{a}_{l\omega_1} - \mathbf{a}_{l\omega_2}|$, which basically means $\mu_{l\omega_1} \cong \mu_{l\omega_2}$, contain little or no discriminating power and can also be dropped.

Thus in this work $\mathbf{x}_l$ is dropped if $0.5 + \alpha \geq \mu_{l\omega_m} \geq 0.5 - \alpha$, with $m = 1, 2$. In other words, the feature $\mathbf{x}_l$ is discarded whenever the $|\mu_{l\omega_m} - 0.5|$ value is in the range $[0,\alpha]$. A feature is also considered nondiscriminating and dropped if $|\mu_{l\omega_1} - \mu_{l\omega_2}| \leq \beta$. In practice, β and α can be viewed as adjustable parameters. In passing, we note that the "optimal" values of β and α can be estimated, through cross-validation, using the training set and the linear equation (eq 6). However, in this paper, we elected to estimate the values of β and α in an ad hoc manner as we considered this to be more insightful. Note the α and β parameters can only take non-negative values.

The arguments above can be easily extended to the multiclass case. In this scenario, whenever $\alpha \geq |\mu_{l\omega_m} - 0.5|$, $(m = 1, 2, ...)$, $\mathbf{x}_l$ is dropped. A feature is also removed from the descriptor vector if the values of all (or the majority) of the probabilities for the $\mu_{l\omega_m}$ for $\omega_1, \omega_2, ...$ classes are equal (or similar in the sense described in the two-class scenario).

In order to test the selection performance of our feature selecting method in the cheminformatics context, binary probabilistic classifiers based on the Parzen−Rosenblatt Window[11] and NB approaches were used. The NB classifier is as described above; see eq 2. The Parzen−Rosenblatt Window based probabilistic classifier can be expressed as[5]

$$p(\omega_m|\mathbf{x}) = c_m \sum_{\mathbf{x}_i \in \omega_m}^{N_{\omega_m}} K(\mathbf{x}_i, \mathbf{x}; \lambda) \quad (9)$$

with

$$c_m = \frac{p(\omega_m)}{N_{\omega_m}p(\mathbf{x})}$$

where $p(\omega_m)$, N_{ω_m}, and $p(\mathbf{x})$ are as defined before. $K(\mathbf{x}_i, \mathbf{x}; \lambda)$ is a kernel function; $\mathbf{x}$ denotes the compound to be classified; $\mathbf{x}_i$ are the training data points; λ is a tunable parameter.[5]

To test the generalization ability of the binary probabilistic classifiers, one can use a variety of measures. In this work classifier performances were measured using the so-called area under the curve (AUC) "metric"; for completeness, we also present the corresponding receiver operating characteristics (ROCs).[12]

RESULTS AND DISCUSSIONS

For illustrative purposes, the proposed algorithm was tested on a bioactivity data set that was used by Lowe et al.[5]—a data set taken from the MDL Drug Data Report (MDDR) database. This data set has been used in a number of classification studies.[5,13−16] It consisted of 8293 compounds and 11 activity classes, which is a subset of the 102 533 structures and 11 activity classes reported in previous classification studies.[13−15] The 8293 compounds, and 11 activity classes and their (MDDR) codes are as summarized

dx.doi.org/10.1021/ci200303m | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 1. 8293 Structures/Compounds to Classify According to Their Biological Activity

activity class	MDDR activity code	no. of active compounds
5HT3 antagonist	06233	752
5HT1A agonists	06235	827
5HT reuptake inhibitors	06245	359
D2 antagonist	07701	395
renin inhibitors	31420	1130
angiotensin II AT1 antagonist	31432	943
thrombin inhibitors	37110	803
substance P antagonist	42731	1246
HIV protease inhibitors	71523	750
cyclooxygenase inhibitors	78331	636
protein kinase C inhibitors	78374	452

Table 2. Average Number of Features in Each Activity Class That the Proposed Feature Selector Selected and Their Corresponding Standard Deviations As Described in the Main Text

activity class code	original number of features	selected number of features	standard deviation
06233	166	100	0.328
06235	166	93	0.000
06245	166	98	0.227
07701	166	100	0.150
31420	166	105	0.000
31432	166	105	0.365
37110	166	102	0.276
42731	166	99	0.296
71523	166	94	0.000
78331	166	94	0.188
78374	166	88	0.557

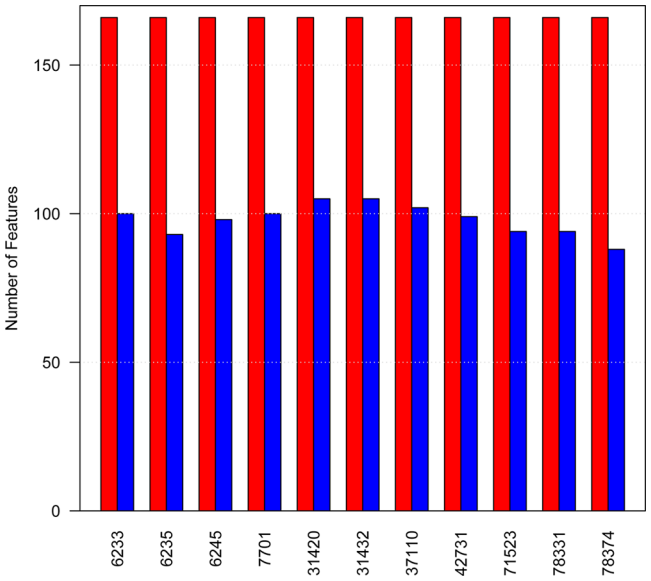


Figure 1. Plot showing the number of features selected and the original number of features across the different activity classes. Here blue (red) denotes feature selection was (was not) performed.

in Table 1 (as reported in ref 5). Column 1 and 3 of the table show the activity class and how many compounds (out of the 8293) are

Table 3. Performance of the Proposed Feature Selector on Classifying the Structures in the Given 11 Activity Classes[a]

activity class code	AUC(FSF)	AUC(SSF)
06233	0.924	0.919
06235	0.890	0.894
06245	0.890	0.892
07701	0.886	0.884
31420	0.981	0.981
31432	0.986	0.983
37110	0.909	0.897
42731	0.851	0.870
71523	0.905	0.885
78331	0.956	0.952
78374	0.853	0.852

[a]Columns 1, 2, and 3 refer to activity code, AUC values on the different activity classes yielded by NB classifiers using an FSF (full set of features) and SSF(selected set of features), respectively.

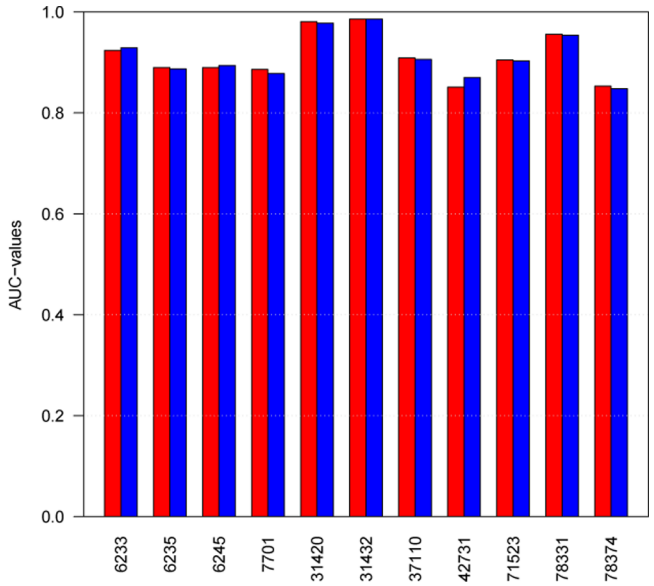


Figure 2. Plots of the AUC values for the NB classifiers across the different activity classes. Blue (red) denotes feature selection was (was not) performed.

active against the given target. This means that the classification test problem that we employed was, by design, an induced two-category classification problem.

Initially, L elements of the pattern vectors were generated using MACCS fingerprints.[17] All structures/compounds were represented by 166 MACCS structure keys, the descriptors, computed by MOE[18]—i.e., L was 166.

In each activity class, a leave-one-out scheme was employed, i.e., each time 8292 data points were used for both feature selection and constructing the corresponding classifier, and the remaining one data point was used for testing. In other words, for a given activity class, 8293 different sets of features were selected and in each case the 8292 training data vectors consisting of the corresponding selected feature set were used to construct classifiers based on eqs 2 and 9.

In the classifiers based on the Parzen−Rosenblatt Window approach, the employed kernel was the Aitchison−Aitken (AA)

dx.doi.org/10.1021/ci200303m | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

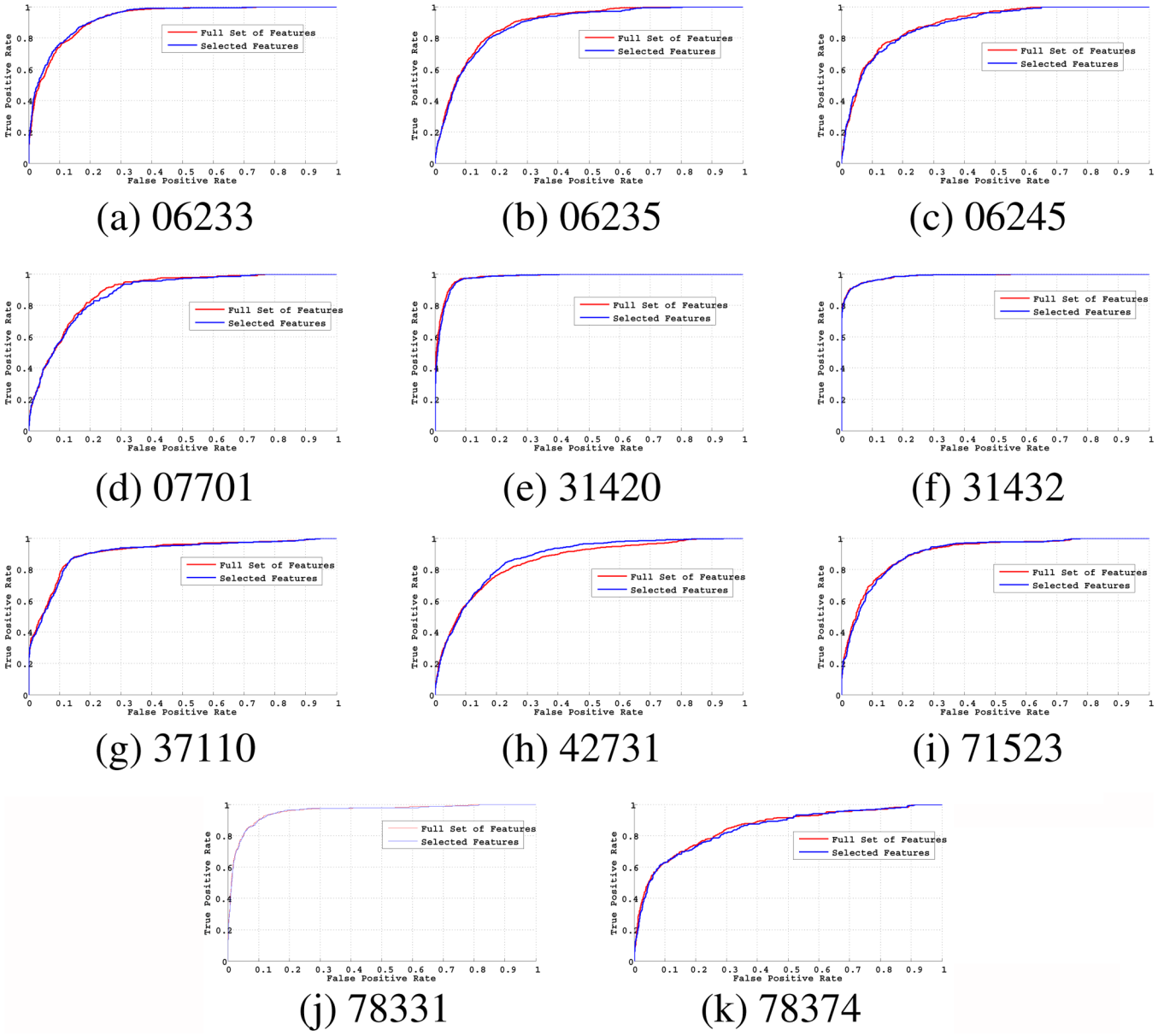


Figure 3. Plots showing the ROC curves of classification performances yielded by the generated NB classifiers on all activity classes.

kernel function[19] which has recently been shown to be a positive definite kernel by one of us (H.Y.M.).[20] The AA kernel is a kernel function defined over the discrete descriptor space and can be given as

$$K(\mathbf{x}_i, \mathbf{x}; \lambda) = \lambda^{q-r(\mathbf{x}_i-\mathbf{x})}(1-\lambda)^{r(\mathbf{x}_i-\mathbf{x})} \quad (10)$$

where $\mathbf{x}$ and $\mathbf{x}_i$ are as defined before and λ is a tunable smoothing variable $(0.5 < \lambda \leq 1)$. In the case of $\lambda = 1$, there is no smoothing; $r(\mathbf{x}_i - \mathbf{x}) = (\mathbf{x}_i - \mathbf{x})^{\mathrm{T}}(\mathbf{x}_i - \mathbf{x})$ is the number of disagreements in corresponding components of $\mathbf{x}_i$ and $\mathbf{x}$; q is the number of features. Note that when no feature selection is made, q is L (= 166).

Results. In each activity class, $\mu_{l\omega_1}$ and $\mu_{l\omega_2}$ were computed for each feature x_l using the training data set. Then, as described in the previous section, x_l was dropped whenever the value of $|\mu_{l\omega_1} - \mu_{l\omega_2}|$ was in the range $[0,\beta]$. A feature was also discarded if $|\mu_{l\omega_1}|$, $|\mu_{l\omega_2}|$, or both were in the range $[0,\alpha]$.

For a given activity class, $|\mu_{l\omega_1} - \mu_{l\omega_2}|$ was computed for each feature. These values were then ranked in descending order. As the largest possible $|\mu_{l\omega_1} - \mu_{l\omega_2}|$ value is one, in this work any feature whose corresponding $|\mu_{l\omega_1} - \mu_{l\omega_2}|$ value ≤ 0.04 was considered less discriminating and dropped. This means that in all 11 activity classes β was set to 0.04. The retained set of features were further screened by appropriately choosing the value of α for the given activity class. In this work the "optimal" α value was chosen by plotting $\ln[\mu_{l\omega_{1/2}}/(1-\mu_{l\omega_{1/2}})]$ (here: $\omega_{1/2}$ denotes ω_1 and ω_2) against $\mu_{l\omega_{1/2}}$ and then visually inspecting how the $\ln[\mu_{l\omega_{1/2}}/(1-\mu_{l\omega_{1/2}})]$ values spread out about zero. In all 11 activity classes an α parameter value of 0.02 was considered appropriate, i.e., if $0.5 + 0.02 \geq \mu_{l\omega_{1/2}} \geq 0.5 - 0.02$, the features corresponding to these values were discarded.

As described in the previous section, the feature selection process discussed above was repeated 8293 times for a given

dx.doi.org/10.1021/ci200303m | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 4. Performance of the Proposed Feature Selector on Classifying the Structures in the Given 11 Activity Classes[a]

activity class code	AUC(FSF)	AUC(SSF)
06235	0.985	0.983
06233	0.997	0.997
06245	0.990	0.990
07701	0.960	0.980
31420	0.998	0.998
31432	0.999	0.999
37110	0.999	0.999
42731	0.999	0.999
71523	0.999	0.999
78331	0.999	0.999
78374	0.999	0.998

[a]Columns 1, 2, and 3 refer to activity code, AUC values on the different activity classes yielded by NB classifiers using a FSF (full set of features) and SSF(selected set of features), respectively.

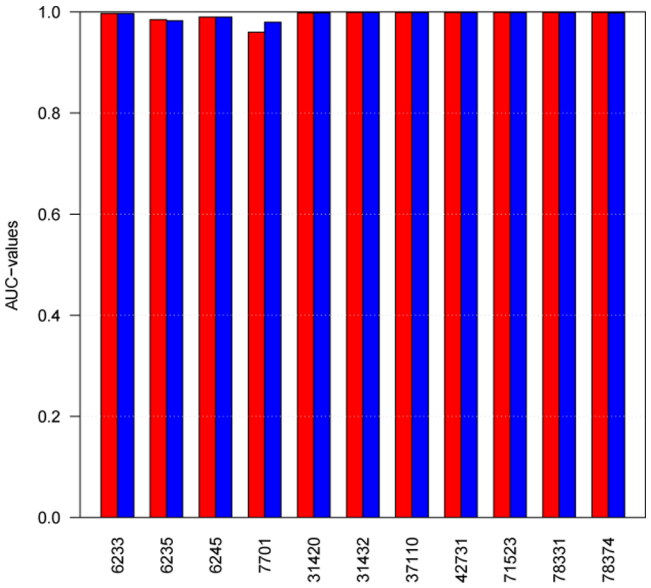


Figure 4. Plots of the AUC values for the Parzen–Rosenblatt Window classifiers across the different activity classes. Blue (red) denotes feature selection was (was not) performed.

activity class, which yielded 8293 sets of selected features for this particular activity class. The arithmetic mean (rounded to the nearest integer) of the sizes of these sets of features selected and the corresponding standard deviation (for each activity class) are shown in columns 3 and 4 of Table 2. Column 2 of this table shows the original number of features (166 of them) assumed to adequately represent the compounds. Clearly the proposed algorithm yielded significant reduction in the number of features across the different activity classes, and this is also illustrated in Figure 1. In the best case, the number of features were reduced to about half of the original value while in the worst case, the number was to reduced to about 37% of the original value. In the figure, the blue and red bars denote the selected and original number of features, respectively; in the rest of the paper, a similar color code scheme was followed in all the figures.

In each activity class, the 8292 data points employed to select the discriminating features were also used to generate classifiers for a given activity class. This means a leave-one-out scheme was employed to generate classifiers, i.e., 8293 classifiers were constructed for a given activity class. In this work, $\mathbf{x}$ is assigned to class ω_1 if $p(\omega_1|\mathbf{x}) \geq p(\omega_2|\mathbf{x})$; otherwise, the pattern vector (the compound) is assigned to class ω_2.

The generalization ability of the generated NB classifiers (based on selected features) on the 11 activity classes (one per activity class) are shown in Column 3 of Table 3. For comparison, NB classifiers were also constructed using all the 166 features. The classification performances of these NB classifiers on the 11 data sets are shown in Column 2. These classification performances were illustrated in Figure 2, where the blue and red bars indicate that the selected and original number of features were employed to generate the classifiers, respectively. The panels in Figure 3 show the receiver operating characteristic (ROC) curves associated with the AUC values shown in Figure 2 and Table 3. Each panel demonstrates a pair of curves; the blue curve denotes the ROC curve showing prediction results yielded by a NB classifier built on pattern vectors consisting of selected features; the red curve shows the ROC curve of the classification results predicted by a NB classifier generated on pattern vectors consisting of the original set of features, and the panel label refers to the corresponding activity class code.

The two sets of AUC values in the table were clearly paired. Furthermore, one could not assume a Gaussian distribution for these sets of AUC values. Thus a nonparametric test statistic was used to investigate whether or nor the two sets of results were statistically significantly different. Choosing 0.05 level of significance, a two-tailed Wilcoxon signed rank test conducted on the two sets of AUC values gave a p-value of 0.3074. We, therefore, concluded the two sets of results were not statically significantly different at the assumed level of significance.

For comparison we repeated the process described above: this time round the generated classifiers were based on the Parzen–Rosenblatt Window approach given in eq 9. In this case, λ, the adjustable parameter in our kernel function, was set to 0.9 across the 11 activity classes. Table 4 and Figures 4 and 5 clearly show patterns similar to those observed in the classification performances yielded by NB classifiers. A two-tailed Wilcoxon signed rank test (using a significance level of 0.05) gave a p-value of 1.0, which indicated that there was no statistically significant difference between the two sets of AUC-values in the Parzen–Rosenblatt Window approach case.

In passing we note that across the different activity classes, the Parzen–Rosenblatt Window based classifiers systematically outperformed the NB classifiers. In fact, a two-tailed Wilcoxon signed rank test (using 0.05 level of significance) gave a p-value of 0.0009766

In summary, with the simple, but insightful approach through which the α and β values were chosen, the proposed feature selector reduced substantially the number of discriminating features in all activity classes. In some activity classes, the number of features were reduced to about half of their original value. Although reducing the number features did not improve the classification performances of the classifiers, in this work at least, it did not degrade the generalization ability of the classifiers, either.

dx.doi.org/10.1021/ci200303m | J. Chem. Inf. Model. XXXX, XXX, XXX–XXX

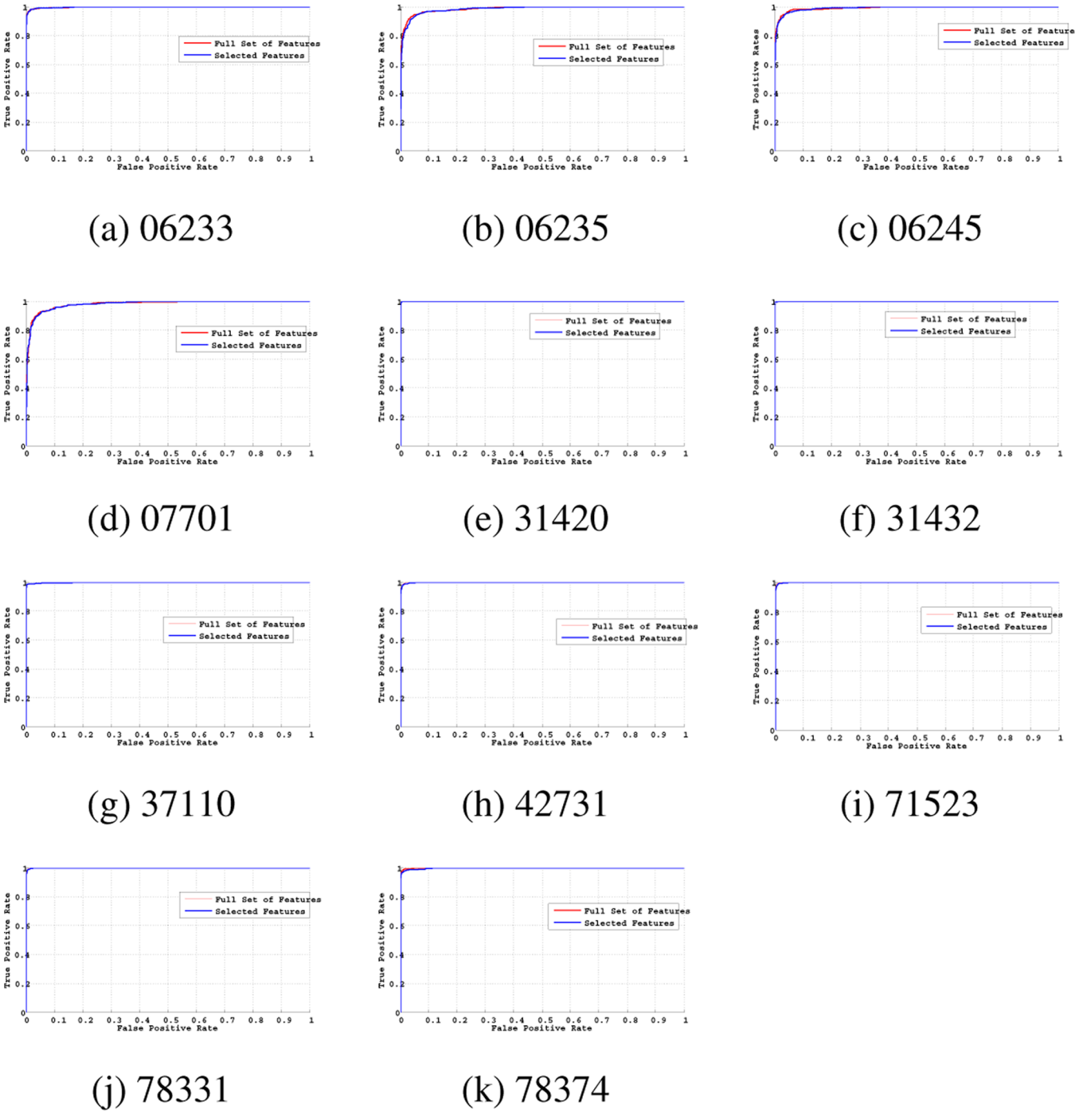


Figure 5. Plots showing the ROC curves of classification performances yielded by the generated Parzen−Rosenblatt Window classifiers on all activity classes.

CONCLUSION

We have described how NBCBBD can be employed for feature selection. It was also demonstrated that NBCBBD is by nature a linear classifier which means that it is suboptimal in classifying compounds that are only nonlinearly separable in their binary descriptor space. We used 11 data sets from the MDDR database to test the performances of the proposed feature selector. The results indicate that the new feature selector is certainly promising.

AUTHOR INFORMATION

Corresponding Author

*E-mail: hym21@cam.ac.uk.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors acknowledge the useful discussions on this work with Dr J. B. O. Mitchell and Dr. B. Brooks. We would also like to thank Mr. R. L. Marchese Robinson for providing the datasets employed in this work. The authors would like to thank Unilever (H.Y.M. and R.C.G.) and Microsoft Research Division (J.A.T.) for financial support.

REFERENCES

(1) Duda, R. O.; Hart, P. E. *Pattern Classification and Scene Analysis*, 1st ed.; John Wiley & Sons, Ltd: New York, NY, 1973.

(2) Good, I. J. *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*, 1st ed.; MIT Press: Cambridge, MA, 1965.
(3) Bishop, C. *Pattern Recognition and Machine Learning*, 1st ed.; Springer-Verlag: New York, 2006.
(4) Robertson, S. E.; Jones, K. S. J. Relevance weighting of search terms. *J. Am. Soc. Inf. Sci.* **1976**, *27*, 129−146.
(5) Lowe, R.; Mussa, H. Y.; Mitchell, J. D.; Glen, R. C. Classifying Molecules Using a Sparse Probabilistic Kernel Binary Classifier. *J. Chem. Inf. Model.* **2011**, *51*, 1539−1544.
(6) Angelopoulos, N; Hadjiprocopis, A; Walkinshaw, M. D. Bayesian Model Averaging for Ligand Discovery. *J. Chem. Inf. Model.* **2009**, *49*, 1547−1557.
(7) Abdo, A.; Chen, B.; Mueller, C.; Salim, N.; Willett, P. Ligand-Based Virtual Screening Using Bayesian Networks. *J. Chem. Inf. Model.* **2010**, *50*, 1012−1020.
(8) Abdo, A.; Salim, N. New Fragment Weighting Scheme for the Bayesian Inference Network in Ligand-Based Virtual Screening. *J. Chem. Inf. Model.* **2011**, *51*, 25−32.
(9) Klon, A. E. Bayesian Modeling in Virtual High Throughput Screening. *Comb. Chem. High Throughput Screen.* **2009**, *12*, 469−483.
(10) Bender, A. Bayesian Methods in Virtual Screening and Chemical Biology. *Methods Mol. Biol.* **2011**, *672*, 175−196.
(11) Parzen, E. On estimation of a probability density function and mode. *Ann. Math. Stat.* **1962**, *33*, 1065−1076.
(12) Fawcett, T. An introduction to ROC analysis. *Patt. Rec. Lett.* **2006**, *27*, 861−874.
(13) Wilton, D. J.; Harrison, R. F.; Willett, P.; Delaney, J.; Lawson, K.; Mullier, G. Virtual screening using binary kernel discrimination: Analysis of pesticide data. *J. Chem. Inf. Model.* **2006**, *46*, 471−477.
(14) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures. *Org. Biomol. Chem.* **2004**, *2*, 3256−3266.
(15) Nigsch, F.; Mitchell, J. B. O. How to winnow actives from inactives: Introducing Molecular Orthogonal Sparse Bigrams (MOSBs) and multi-class winnow. *J. Chem. Inf. Model.* **2008**, *48*, 306−318.
(16) Cannon, E. O.; Amin, A.; Bender, A.; Sternberg, M. J. E.; Muggleton, S. H.; Glen, R. C.; Mitchell, J. B. O. Support vector inductive logic programming outperforms the naive Bayes classiïňĄer and inductive logic programming for the classiïňĄcation of bioactive chemical compounds. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 269−280.
(17) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273−1280.
(18) *MOE (The Molecular Operating Environment)*, version 2009.10; Chemical Computing Group Inc.: Montreal, Canada H3A 2RT, 2009.
(19) Aitchison, J.; Aitken, C. G. G. Multivariate binary discrimination by the kernel method. *Biometrika* **1976**, *63*, 413−420.
(20) Mussa, H. Y. The Aitchison - Aitken discrete kernel function is after all a positive definite kernel. Statistics and Probability Letters (in revision).