J Mol Model (2013) 19:2679–2685
DOI 10.1007/s00894-013-1806-z

SHORT COMMENT

Representing descriptors derived from multiple conformations as uncertain features for machine learning

Ulf Norinder · Henrik Boström

Received: 29 November 2012 / Accepted: 11 February 2013 / Published online: 12 March 2013

Abstract Uncertainty was introduced into the chemical descriptors of 11 datasets by conformational analysis in order to incorporate three-dimensional information and to investigate the resulting predictive performance of a state-of-the-art machine learning method, random forests, for binary classification tasks. A number of strategies for handling uncertainty in random forests were evaluated. The study showed that when incorporating three-dimensional information as uncertainty into chemical descriptors, the use of uniform probability distributions over the range of possible values, in conjunction with fractional distribution of compounds clearly outperforms the use of normal distributions as well as sampling from both normal and uniform distributions. The main conclusion of this study is that, even when distributions of uncertain values are provided, the random forest method can generate models that are almost as accurate from the expected values of these distributions alone. Hence, there seems to be little advantage to using the more elaborate methods of incorporating uncertainty in chemical descriptors when using random forests rather than replacing the distributions with single-point values. The results also show that random forest models with similar performances can also be generated using three-dimensional descriptor information derived from single (lowest-energy or Corina-derived) conformations.

Keywords Machine learning · Random forests · Conformational analysis · Uncertainty · Binary classification

Electronic supplementary material The online version of this article (doi:10.1007/s00894-013-1806-z) contains supplementary material, which is available to authorized users.

U. Norinder
Department of Pharmacy, Uppsala University,
751 23 Uppsala, Sweden

U. Norinder
AstraZeneca R&D, Södertälje, Sweden

H. Boström
Department of Computer and Systems Sciences,
Stockholm University, Forum 100,
164 40 Kista, Sweden
e-mail: henrik.bostrom@dsv.su.se

Present Address:
U. Norinder (✉)
Department of Computational Chemistry,
H. Lundbeck A/S, Ottiliavej 9,
2500 Valby, Denmark
e-mail: ulfn@lundbeck.com

Introduction

The development of pharmaceutical drugs is a costly and time-consuming procedure [1]. Therefore, it has become popular to “front-load” the drug development process with information; in other words, to streamline this process early on, particularly during the lead identification and lead optimization phases, through the appropriate use of data [2]. One aspect of front-loading involves the prediction of various biopharmaceutical properties (e.g., solubility and ADMET properties). For predictive models to be useful, a number of criteria need to be fulfilled, such as robustness, good predictive performance, an appropriate applicability domain, and, in many cases, transparency and interpretability [3].

Most often, a predictive in silico model is formulated by employing a statistical or machine learning algorithm to find a mapping from a particular compound, represented by a set of molecular descriptors, to the output (e.g., a specific biological activity) using available data. Traditionally, each molecular descriptor takes an exact numerical or nominal value. In many cases (e.g., when representing the number of atoms or bonds in a compound) this makes perfect sense, as there is no uncertainty associated with the value. However, other types of descriptors (e.g., log$P$ and other charge-based variables) are not exact: each descriptor value comes with an associated uncertainty, which often is expressed in terms of a standard deviation. Information of this kind is not, however, typically considered in current in silico predictive

Table 1 Data set characteristics

Dataset	End point	No. of compounds	Reference
ace	Angiotensin converting enzyme	114	[20]
ache	Acetylcholinesterase	111	[20]
ames	Ames test	6,512	[21]
bzr	Benzodiazepine receptor	163	[20]
cox2	Cyclooxygenase-2	322	[20]
dhfr	Dihydrofolate reductase	397	[20]
gpb	Glycogen phosphorylase B	66	[20]
hERG	hERG ion channel inhibition	4,667	[22]
solubility	Solubility in buffer at pH7.4	7,493	[23]
therm	Thermolysin	76	[20]
thr	Thrombin	88	[20]

models, usually only the expected (or most likely) value is used. Uncertainty also arises when properties of a molecule are derived from multiple conformations of that molecule; such properties include dipole moments and other electronic properties, as well as (charged) surface areas of various kinds. Thus, a single compound with multiple conformations can have a set of values for a particular descriptor, as the value depends on the specific conformation considered. Since different compounds often have different numbers of conformations, there is no straightforward way of transforming these sets of values into fixed-length feature vectors without losing information or introducing restrictions that come from ordering the conformations. A commonly adopted approach is therefore to represent these multiple values by a single value, which potentially results in a significant loss of information, although attempts have previously been made to sample multiple values [4, 5].

Recently, there has been increasing interest in machine learning methods that are able to learn from uncertain data (such as that mentioned above). Various standard learning algorithms have been adapted to deal with uncertain input features, including support-vector machines [6], decision trees [7, 8], random forests [9, 10], artificial neural networks [11], Bayesian classifiers [12], and rule-based approaches [13–15].

The aim of the study described in the present paper was to investigate whether anything could be gained from representing feature values derived from multiple conformations as uncertain features, in order to potentially take advantage of recent developments in machine learning relating to the handling of such uncertainties.

Nine publicly available and two proprietary (from AstraZeneca) datasets (on solubility and hERG) were used in the study (see Table 1 for a list of names and endpoints, as well as references and supporting information for SMILES and the activity classes of the public datasets).

The three-dimensional structures were generated using Corina (Corina version 3.50, Molecular Networks GmbH, http://www.molecular-networks.com), and subsequent conformational analyses were performed using low-mode sampling with the default settings within Macromodel (Macromodel version 9.5, Schrödinger, LLC, http://www.schrodinger.com). The ten lowest-energy conformations with sufficient dissimilarity were kept for later in order to generate descriptors for each compound under investigation. Dissimilarity was determined by calculating the root mean squared deviation (RMSD) between pairs for all the corresponding atoms within the investigated compound of interest. If the RMSD value did not exceed 1.0 Å, the conformer was considered to be identical to already selected conformers.

Dragon was used to calculate three-dimensional descriptors (Dragon version 1.4.2, Talete s.r.l., http://www.talete.mi.it). The 682 computed molecular descriptors were physicochemical in nature. For a list of the calculated Dragon descriptor sub-blocks, see Table 2.

The datasets in this work concern two-class (binary) classification tasks with, in most cases, balanced classes (i.e., the two classes contain approximately the same number of compounds).

The ten lowest-energy conformations retained from the conformational analysis for each compound were used, and de-

Table 2 Dragon descriptor sub-blocks

Dragon descriptor sub-block	Sub-block number	Number of descriptors
Geometrical descriptors	12	74
RDF descriptors	13	150
3D-MoRSE descriptors	14	160
WHIM descriptors	15	99
GETAWAY descriptors	16	197
*A*log*P*, *M*log*P*	20	2

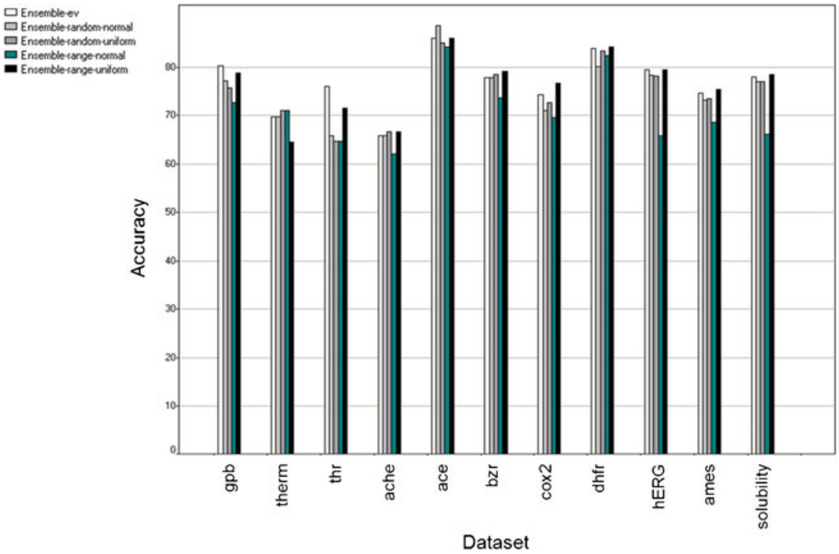

Fig. 1 Average accuracies of the five methods when they were applied to the eleven datasets (datasets increase in size from left to right)

scriptor uncertainty was introduced by identifying the smallest and largest values of each descriptor that occurred in all of the conformations retained from the conformational analysis. The expected value for the descriptor was then taken as the mean of the largest and smallest values found for that descriptor.

The results obtained when using uncertain features (see below) were compared to those achieved using a single value (the expected value). In addition to this, the latter results were compared to the results obtained when implementing two alternative ways of selecting a single value from the set generated from multiple conformations: using the 3D geometry initially generated by Corina for each compound, or using the lowest-energy conformation from the conformational analysis of each compound.

We recently provided a detailed description of the methods employed when introducing uncertainty into ensembles of decision trees [16], so only a brief description of them will be given below. For more details, see [16].

Uncertainty was introduced into the datasets in two ways: by considering a *uniform* probability distribution across the range of possible values (i.e., each value in the interval between the smallest and largest observed values was assumed to be equally probable), or by considering a *normal* distribution, where the interval between the smallest and largest values was assumed to correspond to the 95 % confidence interval. The learning algorithm considered in this study included ensembles of decision trees [17] in the form of so-called random forests [18]. The number of decision trees in each ensemble was 25. Two main

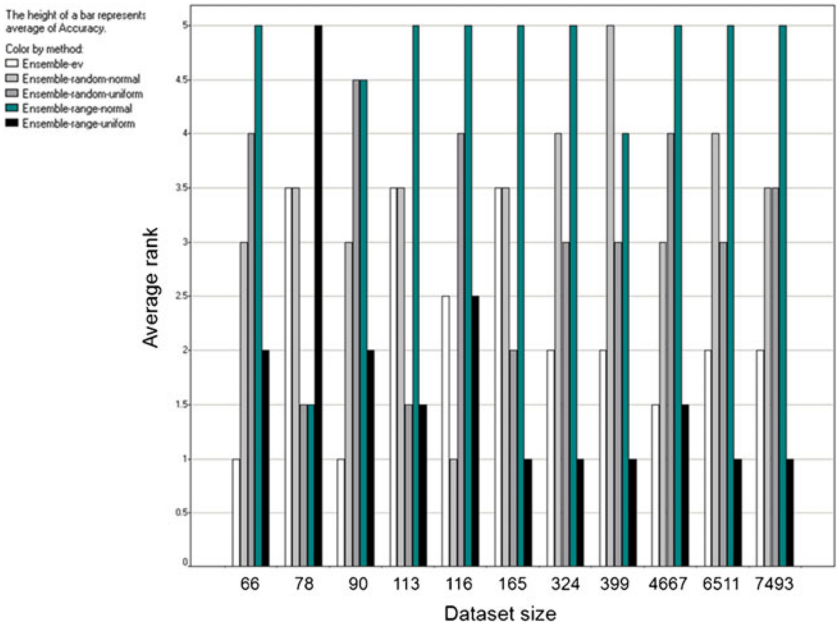

Fig. 2 Average accuracy ranks for the five methods vs. dataset size

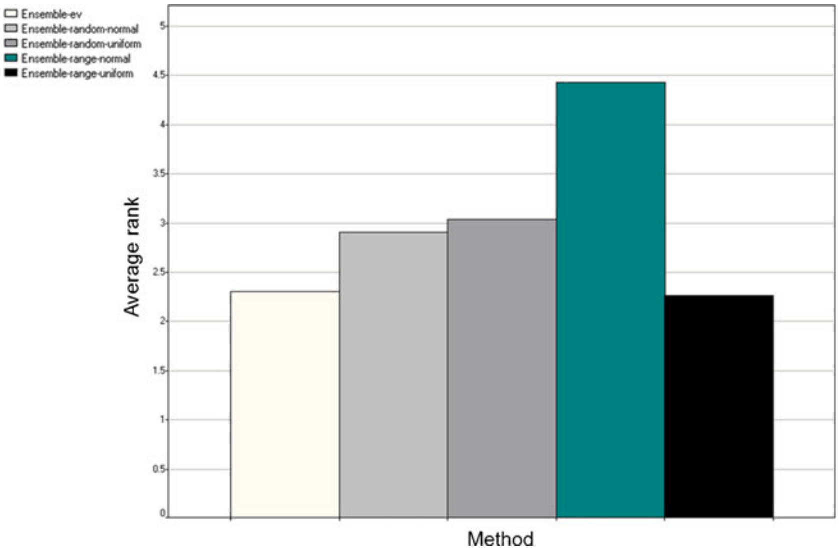

Fig. 3 Average accuracy ranks for the five methods when averaged over all datasets

approaches to handling uncertain numeric features during model construction were considered: *random* (i.e., values were sampled randomly from the distributions prior to tree generation) or *range* (i.e., fractions of compounds may be distributed over multiple nodes during tree growth). Given that uncertainty was introduced in two ways and that there were two approaches to handling uncertain numeric features, there were four possible approaches, which were correspondingly named the "Ensemble-random-normal," "Ensemble-random-uniform," "Ensemble-range-normal," and "Ensemble-range-uniform" methods. In addition, a strategy in which uncertainty was ignored by simply considering the expected value of the distribution, termed "Ensemble-ev", was also implemented as the baseline method. The predictive performance of each model developed was assessed based on the average accuracy resulting from tenfold cross-validation, and the models were then ranked according to average accuracy (the best model was ranked 1, the next best model was ranked 2, and so on).

Results and discussion

The investigation showed that differences in performance (measured in terms of the accuracy resulting from a tenfold cross-validation) among the five methods were rather small for the datasets investigated here, as can be seen in Fig. 1.

A tendency for the Ensemble-range-normal method to be slightly less well performing than the other methods utilized in this study was noted and is depicted in Fig. 1. Also, the methods Ensemble-range-uniform and Ensemble-ev appear to slightly outperform the other methods for larger datasets (Fig. 2).

This superiority of Ensemble-range-uniform and Ensemble-ev is also seen in the ranks of the various methods (Fig. 2): these are the two best-performing methods. Indeed, the overall ranks across all eleven investigated datasets (Fig. 3) show that the most accurate method is Ensemble-range-uniform, followed by Ensemble-ev.

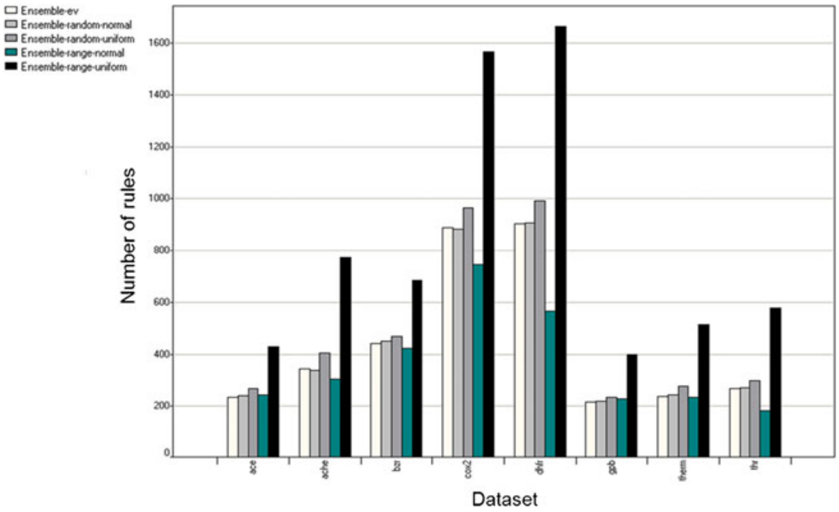

Fig. 4 Number of rules for the models derived by the five methods based on the public datasets

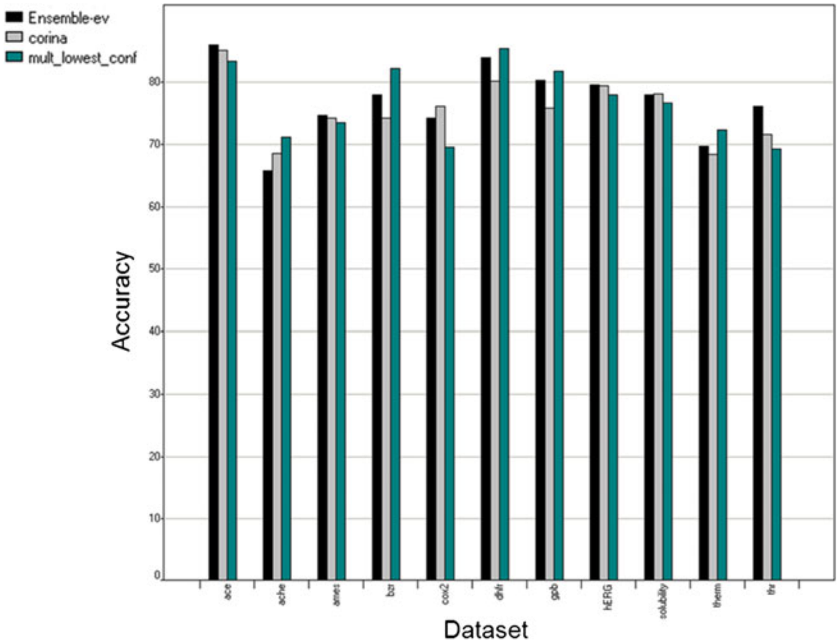

Fig. 5 Average accuracies for models generated from single-valued descriptors by three different methods

Thus, it more advantageous to assume a uniform uncertainty distribution and use the slightly more complex algorithm in which objects (i.e., compounds) are partially distributed according to range rather than to simply use the midpoint (expected) values of the investigated descriptors. Interestingly, the closely related Ensemble-range-normal method is the worst performer for a clear majority of the datasets, as shown in both Figs. 2 and 3. When the average ranks were compared by performing a Friedman test followed by a post hoc Nemenyi test [19], the differences between the two best-performing methods (Ensemble-range-uniform and Ensemble-ev) and the poorest method (Ensemble-range-normal) were found to be significant at the 0.01 level. No other differences were significant at levels below 0.05 according to this test. This difference in predictive performance between the approaches employing normal and uniform distributions may occur because the latter increases the variance (diversity) among the base classifiers compared to the former, which generally has a positive effect on accuracy as long as this diversity does not mean that the accuracies of the individual ensemble members are too low.

It should also be noted that the most accurate method, the Ensemble-range-uniform approach, produces models

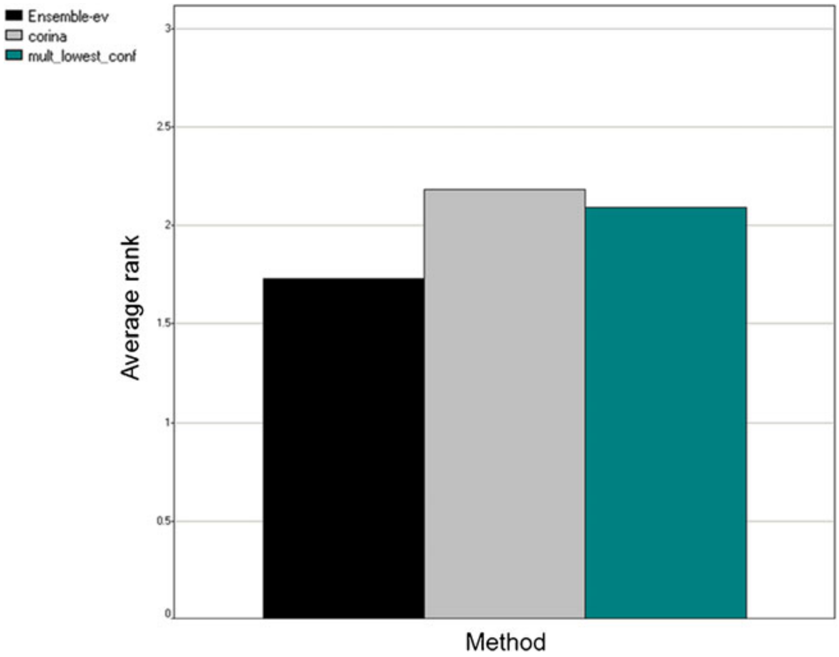

Fig. 6 Average accuracy ranks for three different methods that yield models using descriptors with single values

containing many more rules (leaf nodes), on average approximately twice as many, than the models derived by any of the other four methods, as shown in Fig. 4.

This may be explained by the fact that the weights of the compounds that are distributed over multiple nodes will be more evenly distributed when a uniform rather than normal distribution of the uncertain features is assumed. With more evenly distributed partitioning, tree growth can be expected to continue for longer, resulting in more leaf nodes.

In order to investigate alternative ways of representing the set of descriptor values from multiple conformations, the method using the midpoint values in the intervals (Ensemble-ev) was compared to a method utilizing descriptors calculated from the initial Corina-generated 3-D geometry (termed “Corina”) and a method that used only the lowest-energy conformation (termed “mult_lowest_conf”). The results obtained when using only one descriptor value show relatively small differences in performance among the datasets, as shown in Fig. 5.

The rankings show slightly better performance for the most elaborate approach (Ensemble-ev), which uses a set of conformations to compute the midpoint value for each descriptor (Fig. 6). Again, the overall difference between the methods is not particularly large with respect to accuracy. Indeed, according to the Friedman test [19], none of the differences are significant at the 0.05 level.

This is similar to the observation made by Muehlbachar et al., who stated that using multiple conformations and Boltzmann-weighted descriptors did not improve the statistical quality when deriving QSPR models for logP_{ow}, and that the additional effort involved in generating multiple conformations rather than models based on a single low-energy conformation was not justified [24]. Hechinger et al. also studied the influences of the computational method used for conformer generation and the level of semi-empirical or ab initio quantum mechanical calculation performed on descriptor generation [25]. They, on the other hand, concluded that the use of a single conformer for descriptor generation may mean that the information associated with the descriptor in question is not fully exploited.

In this work, we used uncertainty, which we defined as the difference between the largest and smallest values of each descriptor. A more physics-oriented approach would have been to use the Boltzmann distribution of conformations. However, the purpose of this investigation was to study the possible effect of introducing uncertainty into descriptors derived from 3D conformations on model quality. Using our approach (i.e., using the difference between the largest and smallest values of a descriptor), we maximized the impact of the uncertainty and thus the potential variation in model quality. Using a Bolzmann-weighted distribution of descriptor values would, at best, give the same range of uncertainty for a descriptor; in many cases it would give much smaller variations. This, in turn, would result in models derived from uncertain descriptors that are much more similar to the models derived from descriptor midpoint values (Ensemble-ev) than to the models obtained using our approach to uncertainty, and would therefore limit our ability to study the possible influence of using uncertain descriptors on the statistical quality of models.

The apparent insensitivity of the descriptors to uncertainty prompts the following questions: how different are the computed Dragon 3D descriptors, and how large is the variation in the most important descriptors in comparison with the variation of the other descriptors? To answer these questions, we investigated the two largest public datasets (cox2 and dhfr) by computing the normalized variation (i.e., the range in relation to the average value) for the 25 most important descriptors in the cox2 and dhfr models compared to the rest of the descriptors. The average variations for the most important descriptors and the other descriptors were 0.666 and 0.405, respectively, for the cox2 dataset, and 1.525 and 0.798, respectively, for the dhfr dataset. This shows that, for the latter dataset, the most important descriptors can vary considerably more than the less important ones without compromising model quality (Fig. 1). Models with high predictive performance were derived for both datasets.

Conclusions

The investigations presented in this work indicate that, when performing in silico modeling of binary classification tasks, it is possible to successfully incorporate three-dimensional information into molecular descriptors by utilizing uncertain descriptors. This can be achieved by applying the random forests technique using uniform distributions in conjunction with partially distributed objects (i.e., compounds) according to descriptor range or by simply using the midpoint (expected value) for each descriptor. To further speed up the analysis, if so desired, this work indicates that models of almost the same quality as those obtained using the more elaborate uncertainty scheme can be obtained using a single conformation derived from Corina.

Acknowledgment This work was supported by the Swedish Foundation for Strategic Research through the project *High-Performance Data Mining for Drug Effect Detection* (ref. no. IIS11-0053) at Stockholm University, Sweden.

SMILES and activity classes of the public datasets are provided as supplementary material.

References

1. van de Waterbeemd H, Gifford E (2003) ADMET in silico modeling: towards prediction paradise? Nat Rev Drug Discov 2:192–204
2. Howe TJ, Mahieu G, Marichal P, Tabruyn T, Vugts P (2007) Data reduction and representation in drug discovery. Drug Discov Today 12:45–53

3. Johansson U, Sönströd C, Norinder U, Boström H (2011) The trade-off between accuracy and interpretability for predictive in silico modeling. Fut Med Chem 3:647–663
4. Pissurlenkar RRS, Khedkar VM, Iyer RP, Coutinho EC (2011) Ensemble QSAR: A QSAR method based on conformational ensembles and metric descriptors. J Comp Chem 32:2204–2218
5. Jain AN, Koile K, Chapman D (1994) Compass: Predicting biological activities from molecular surface properties. Performance comparisons on a steroid benchmark. J Med Chem 37:2315–2327
6. Bi J, Zhang T (2005) Support vector classification with input data uncertainty. In: Saul LK, Weiss Y, Bottou L (eds) Advances in Neural Information Processing Systems (NIPS'04), Vancouver, Canada, December 13–18, 2004. MIT Press, Cambridge, pp 161–168
7. Tsang S, Kao B, Yip KY, Ho W-S, Lee SD (2009) Decision trees for uncertain data. In: Golab L, Johnson T, Shkapenyuk V (eds) Proceedings of the 2009 IEEE International Conference on Data Engineering, Shanghai, China, March 29 2009–April 2 2009. IEEE Computer Society, Washington, DC, pp 441–444
8. Qin B, Xia Y, Li F (2009) DTU: a decision tree for uncertain data. In: Theeramunkong T, Kijsirikul B, Cercone N, Ho TB (eds) Proceedings of the 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD), Bangkok, Thailand, April 27–30, 2009. Springer, Heidelberg, pp 4–15
9. Boström H, Norinder U (2009) Utilizing information on uncertainty. In: Johansson R, van Laere J, Mellin J (eds) Proceedings of the 3rd Skövde Workshop on Information Fusion Topics (SWIFT 2009), Skövde, Sweden, October 12−13, 2009. University of Skövde, Skövde, pp 59–62
10. Dudas C, Boström H (2009) Using uncertain chemical and thermal data to predict product quality in a casting process. In: Pei J, Getoor L, de Keijzer A (eds) Proceedings of the First ACM SIGKDD Workshop on Knowledge Discovery from Uncertain Data, Paris, France, June 28, 2009. ACM, New York, pp 57–61
11. Ge J, Xia Y, Tu Y (2010) A discretization algorithm for uncertain data. In: Bringas PG, Hameurlain A, Quirchmayr G (eds) Proceedings of the 21st International Conference on Database and Expert Systems Applications (DEXA): Part II, Bilbao, Spain, August 30–September 3, 2010. Springer, Heidelberg, pp 485–499
12. Qin B, Xia Y, Li F (2010) A Bayesian classifier for uncertain data. In: Shin SY, Ossowski S, Schumacher M (eds) Proceedings of the 2010 ACM Symposium on Applied Computing, Sierre, Switzerland, March 22–26, 2010. ACM, New York, pp 1010–1014
13. Qin B, Xia Y, Prabhakar S (2009) A rule-based classification algorithm for uncertain data. In: Golab L, Johnson T, Shkapenyuk V (eds) Proceedings of the 2009 IEEE International Conference on Data Engineering, Shanghai, China, March 29 2009–April 2, 2009. IEEE Computer Society, Washington, DC, pp 1633–1640
14. Gao C, Wang J (2010) Direct mining of discriminative patterns for classifying uncertain data. In: Rao B, Krishnapuram B, Tomkins A, Yang Q (eds) Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), Washington, DC, USA, July 25–28, 2010. ACM, New York, pp 861–870
15. Qin X, Zhang Y, Li X, Wang Y (2010) Associative classifier for uncertain data. In: Chen L, Tang C, Yang J, Gao Y (eds) Proceedings of the 11th International Conference on Web-Age Information Management (WAIM), Jiuzhaigou, China, July 15–17, 2010. Springer, Heidelberg, pp 692–703
16. Norinder U, Boström H (2012) Introducing uncertainty in predictive modeling—friend or foe? J Chem Inf Model 52:2815–2822. doi:10.1021/ci3003446
17. Quinlan JR (1993) C4.5: programs for machine learning. Morgan Kauffman, San Francisco
18. Breiman L (2001) Random forests. Machine Learning 45:5–32
19. Demsar J (2006) Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res 7:1–30
20. Bruce CL, Jl M, Pickett SD, Hirst JD (2007) Contemporary QSAR classifiers compared. J Chem Inf Model 47:219–227
21. Hansen K, Mika S, Schroeter T, Sutter A, ter Laak A, Steger-Hartmann T, Heinrich N, Müller KR (2009) Benchmark data set for in silico prediction of Ames mutagenicity. J Chem Inf Model 49:2077–2081
22. Gavaghan CL, Hasselgren Arnby C, Blomberg N, Strandlund G, Boyer S (2007) Development, interpretation and temporal evaluation of a global QSAR of hERG electrophysiology screening data. J Comput Aided Mol Des 21:189–206
23. Wood DJ, Buttar D, Cumming JG, Davis AM, Norinder U, Rodgers SL (2011) Automated QSAR with a hierarchy of global and local models. Mol Inf 30:960–972
24. Muehlbacher M, El Kerdawy A, Kramer C, Hudson B, Clark T (2011) Conformation-dependent QSPR models: logPOW. J Chem Inf Model 51:2408–2416
25. Hechinger M, Leonhard K, Marquardt W (2012) What is wrong with quantitative structure–property relations models based on three-dimensional descriptors? J Chem Inf Model 52:1984–1993