The INFUSIS Project – Data and Text Mining for *In Silico* Modeling

Henrik Boström[1,2], Ulf Norinder[3], Ulf Johansson[4], Cecilia Sönströd[4], Tuve Löfström[4], Elzbieta Dura[5], Ola Engkvist[6], Sorel Muresan[6], Niklas Blomberg[6]

[1]Informatics Research Centre, University of Skövde, [2]Dept. of Computer and Systems Sciences, Stockholm University, [3]AstraZeneca R&D Södertälje, [4]School of Business and Informatics, University of Borås, [5]Lexware Labs, [6]AstraZeneca R&D Mölndal

Abstract

The INFUSIS project is a three-year collaboration between industry and academia in order to further the development of new effective methods for generating predictive and interpretable models from machine learning and text mining to solve drug discovery problems.

Introduction

One of the most intensive areas of research within the pharmaceutical industry today is to collect and analyze data on absorption, distribution, metabolism, excretion and toxicity (ADMET) [1]. The overall purpose is to learn how various compounds interact with the human body in order to guide drug development projects in the search for promising compounds. Specifically, compounds unsuitable as drug candidates, e.g., due to toxicity, should be detected as early as possible. Hence, a lot of effort is spent on *front loading* the drug development projects, i.e., a substantial amount of analysis is invested in the very early phases of the projects.

Currently, a commonly adopted approach is to leverage large libraries of chemicals (acquired or synthesized to meet stringent quality criteria)) and use high-throughput screening (HTS) to test for biological activity. Promising compounds found in this way become the focus for continued research, which typically leads to further synthesis and screening. Synthesis and screening processes are, however, often time consuming and costly, making it desirable to estimate the biological activity, as well as ADMET properties, before synthesis. When computer software is used for this initial modeling, the procedure is referred to as in silico modeling [1]. If successful, in silico modeling saves much time and investments by excluding non-promising compounds, thus allowing earlier focus on drug candidates with high potential.

Several aspects of *in silico* modeling are investigated in the INFUSIS project[1] (*INformation FUSion for In Silico modeling in pharmaceutical research*), which is a collaboration of University of Skövde, University of Borås, AstraZeneca AB and Lexware Labs, running from January 2009 to December 2011, with funding from the Swedish Knowledge Foundation and the industrial partners. The research problems

[1] www.his.se/infusis

addressed by the project include handling of uncertainty of measurements or descriptors, improve interpretability of predictive models as well as advancing ensemble techniques to improve predictive performance and robustness. The INFUSIS project also tries to improve predictive modeling by fusing information from various sources such as unstructured texts where corpus technology is used to uncover relevant information. The challenges of the addressed problems and some results that have been achieved so far are presented in the following sections.

Handling uncertainty

This part of the project investigates the question: to what extent can information on descriptor value uncertainty be exploited to improve *in silico* modeling. Standard decision tree and forest learning algorithms have been extended with the ability to build models from uncertain data specified by probability distributions rather than specific values. Empirical investigations on selected *in silico* modeling datasets with uncertain data have been undertaken comparing different strategies for representing uncertainty and strategies for handling such distributions during tree building.

Different approaches to handling uncertain numerical features have been explored when using the random forest algorithm for generating predictive models. The two main approaches are: i) sampling from probability distributions prior to tree generation, which does not require any change to the underlying tree learning algorithm, and ii) adjusting the algorithm to allow for handling probability distributions, similar to how missing values typically are handled, i.e., partitions may include fractions of examples. In [2], an experiment with six datasets concerning the prediction of various chemical properties was presented, where 95% confidence intervals were included for one of the 92 numerical features. In total, five approaches to handling uncertain numeric features were compared: ignoring the uncertainty, sampling from distributions that are assumed to be uniform and normal respectively and adjusting tree learning to handle probability distributions that are assumed to be uniform and normal respectively. The experimental results show that all approaches that utilize information on uncertainty indeed outperform the single approach ignoring this, both with respect to accuracy and area under ROC curve. A decomposition of the squared error of the constituent classification trees shows that the highest variance is obtained by ignoring the information on uncertainty, but that this also results in the highest mean squared error of the constituent trees. In [3], a similar experiment was presented on predicting product quality in a casting process.

Future work includes extending the empirical investigation to a larger number of datasets and also to a larger number of uncertain features. Another direction for future work includes investigating the effectiveness of the two main approaches (sampling vs. distributing fractions of examples) also for uncertain categorical features.

Increasing interpretability

When interpretable models are required, additional demands such as brevity, i.e., important relationships are described with as few rules as possible, can be placed on

models. An important issue is to develop algorithms that are able to optimize such properties. Furthermore, it is desirable that any parameters of such algorithms are easy to use and that they affect the results in a reasonably predictable way, e.g., allowing users to trade various interpretability properties against each other and also against different accuracy measurements. This project studies the use of *in silico* concept description modelling for drug discovery. The focus has so far been on techniques producing decision trees and ordered rule sets, also called decision lists.

The decision list algorithm Chipper [4], specifically aimed at concept description, has so far been evaluated in two different studies on medicinal chemistry data sets. In [5], three different decision list algorithms (JRip, PART and Chipper) were evaluated on a data set concerning the interaction of molecules with a human gene that regulates heart functioning (hERG). The main results were that decision list algorithms can obtain predictive performance not far from the state-of-the-art method random forests, but also that algorithms focusing on accuracy alone may produce complex decision lists that are very hard to interpret. The experiments also showed that by sacrificing accuracy only to a limited degree, comprehensibility (measured as both model size and classification complexity, i.e., the average number of tests needed for a classification) can be improved remarkably.

In [6], the task studied was how to obtain accurate and comprehensible QSAR models. The data sets used were 8 publicly available medicinal chemistry datasets, with six different feature sets containing up to 1024 attributes. Three techniques (J48 decision trees and JRip and Chipper decision lists) were evaluated on predictive performance, measured as accuracy, and comprehensibility, measured as model size. The results on accuracy showed that J48 obtains superior accuracy, followed by Chipper, and then JRip. On comprehensibility, the results were reversed; JRip obtained the smallest models, followed by Chipper, with J48 producing the largest models. Regarding the effect of feature reduction on accuracy, all techniques were seen to benefit from feature reduction, which almost always resulted in increased accuracy. For model size, however, feature reduction was seen not to be universally beneficial; only J48 produced smaller models for the reduced datasets, while both decision list algorithms actually produced larger models on average. The overall conclusion is that, for these datasets, there exists a definite trade-off between accuracy and interpretability.

Future work consists of a more detailed study of the effect of feature reduction on decision lists and further development of the Chipper algorithm.

Another way of obtaining interpretable models is to generate transparent representations of opaque models, an activity named rule extraction. We have previously developed a rule extraction algorithm based on genetic programming, called G-REX. [7].

A recent addition to G-REX [8] is the ability to explicitly focus on extracting rules for a specific set of instances, similar to transductive learning. Another recent study [9] utilizes the inherent inconsistency of genetic

search to form an imaginary ensemble, which is then used as a guide when selecting one specific tree, as the comprehensible model.

Current work includes further development of the G-REX framework, but also hybrid techniques producing comprehensible models. More specifically, we intend to explore the connection between rule extraction and techniques utilizing semi-supervised and transductive learning.

Advancing ensemble techniques

One major open research problem concerns the relationship between ensemble diversity and accuracy, which is not completely understood, especially for classification problems. Furthermore, several different studies show that the correlation between proposed diversity measures and test set accuracy is remarkably low, see e.g., [10,11]. Because of this, there is no widely accepted diversity measure that can be used for ensemble design. Currently, various researchers instead try very different approaches, resulting in a steady stream of highly specialized and quite technical ensemble algorithms.

Naturally, when presenting a novel algorithm, the implicit claim is that the new algorithm, in some aspect, represents the state-of-the-art. Obviously, the most important criterion is predictive performance, typically measured using either accuracy or AUC. A recent study [12], using 32 publicly available data sets from the drug discovery domain, however, showed that several straightforward techniques producing ANN ensembles were more than able to match the performance of the widely acknowledged ensemble techniques GASEN [13] and NegBagg [14]. Especially NegBagg, which is a fairly recent algorithm, was constantly outperformed by most of the standard bagging versions included in the study. Nevertheless, the results for GASEN were even more striking, showing that it was most often detrimental to apply GASEN at all. Or, put in another way, creating an ensemble of all available ANNs was normally a stronger choice than using the subset suggested by GASEN.

This project investigates how diversity measures can be utilized for choosing and combining models to further improve predictive performance. The overall goal is to develop a robust method that effectively incorporates measures of diversity to produce highly accurate ensemble models. Based on the findings in [12], further development of straightforward techniques, which only implicitly target diversity, is prioritized.

Another current study investigates how feature reduction should be applied to ANN ensemble training. Naturally, feature reduction in general has been heavily investigated, but studies targeting feature reduction for classifiers specifically trained to be part of ensembles are quite rare. Our algorithm, aimed at producing “optimal” different feature sets for the base classifiers, is based on genetic search and uses fitness functions combining accuracy and diversity measures.

Finally, it could be noted that more accurate ensembles would probably be beneficial for black-box rule extraction techniques.

Fusing information from multiple sources

Unstructured texts are among the most important additional information sources. Of particular interest are reports with experimental data involving chemical processes important in tracing a certain biochemical activity, e.g., toxicity. Two tasks must be performed in order to obtain relevant data from texts in biochemistry. The first one is a special named entity recognition task: compound names and chemical processes need to be identified in free text. The other one is mapping of the names identified in texts to compounds in some suitable database. In this task, name ambiguity and variability constitute the two chief problems to be addressed [15].

We use text corpus technology tools to uncover relevant information from texts. Culler is an information retrieval system based on natural language processing. It allows versatile and precise data extraction from natural language processed text collections, called corpora. Culler is adapted in the project to allow finding names of chemical compounds. The adapted tool may hence be used to compile new sets of compounds. At the moment there are over 3,000 names of chemical substances available as one concept class in queries in Culler corpora, available at http://bergelmir.iki.his.se/culler/.

One corpus, called Diabetes, is a selection of about 200,000 abstracts on diabetes from PubMed. Patterns of the actual use of chemical nomenclature in research texts have been extracted from this corpus. There are significant differences in how terms are registered in lexicons and how they are actually used [16], making the task of proper identification of chemical compounds in texts a non-trivial task despite availability of large libraries of chemical compounds. Chemlist is the library used in the project [17]. The text sources encompass a broad selection of PubMed abstracts on obesity. The selection counts about 860,000 abstracts and it is currently being turned into a Culler corpus.

Concluding remarks

The INFUSIS project aims to contribute with tools and techniques for data and text analysis to support decision making in the domain of medicinal chemistry. In particular, presented and planned contributions include handling of uncertain data, generating interpretable models, utilizing diversity and feature reduction for ensembles, and using text analysis to compile compound sets related to biochemical activities.

Acknowledgments

This work was supported by the INFUSIS project (www.his.se/infusis) at the University of Skövde, Sweden, in partnership with University of Borås, AstraZeneca, Lexware Labs and the Swedish Knowledge Foundation under grant 2008/0502.

References

[1] H. van de Waterbeemd and E. Gifford, “Admet in silico modelling: towards prediction paradise?” Nat Rev Drug Discov, vol. 2, no. 3, pp 192–204, 2003.

[2] H. Boström and U. Norinder, “Utilizing Information on Uncertainty for In Silico Modeling using Random Forests”, Proc. of the 3rd Skövde Workshop on Information Fusion Topics, pp 59-62, 2009.

[3] C. Dudas and H. Boström, “Using uncertain chemical and thermal data to predict product quality in a casting process”, Proc. of the First ACM SIGKDD Workshop on Knowledge Discovery from Uncertain Data, pp 57–61, 2009.

[4] U. Johansson, C. Sönströd, T. Löfström and H. Boström, “Chipper – A Novel Algorithm for Concept Description”, Proc. of the Scandinavian Conference on Artificial Intelligence, pp 133-140, 2008.

[5] C. Sönströd, U. Johansson, U. Norinder, and H. Boström, “Comprehensible Models for Predicting Molecular Interaction with Heart-Regulating Genes”, Proc. of the International Conference on Machine Learning and Applications, pp 559 – 564, 2008.

[6] C. Sönströd, U. Johansson and U. Norinder, “Generating Comprehensible QSAR models”, Proc. of the 3rd Skövde Workshop on Information Fusion Topics, pp 44-48, 2009.

[7] U. Johansson, R. König and L. Niklasson, “Rule Extraction from Trained Neural Networks using Genetic Programming”, Proc. of the International Conference on Artificial Neural Networks, supplementary proceedings, pp 13-16, 2003.

[8] U. Johansson and L. Niklasson, “Evolving Decision Trees Using Oracle Guides”, Proc. of the IEEE Symposium on Computational Intelligence and Data Mining, pp 238-244, 2009.

[9] U. Johansson, R. König, T. Löfström and L. Niklasson, ”Using Imaginary Ensembles to Select GP Classifiers”, EuroGP, 2010, In Press.

[10] L. I. Kuncheva and C. J. Whitaker, “Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy”, Machine Learning, (51):181-207, 2003.

[11] U. Johansson, T. Löfström and L. Niklasson, “The Importance of Diversity in Neural Network Ensembles - An Empirical Investigation”, Proc. of the International Joint Conference on Neural Networks, pp 661-666, 2007.

[12] U. Johansson, T. Löfström and U. Norinder, “Evaluating Ensembles on QSAR Classification”, Proc. of Skövde Workshop on Information Fusion Topics, pp 59-62, 2009.

[13] Z.-H. Zhou, J.-X. Wu and W. Tang. “Ensembling Neural Networks: Many Could Be Better Than All”, Artificial Intelligence, Vol. 137, No. 1-2:239-263, 2002.

[14] M. M. Islam, X. Yao, S. M. Shahriar Nirjon, M. A. Islam and K. Murase, “Bagging and boosting negatively correlated neural networks”. IEEE transactions on systems, man, and cybernetics, Part B: Cybernetics, 38(3):771-84, 2008.

[15] Y. Tsuruoka, J. McNaught, S. Ananiadou, “Normalizing biomedical terms by minimizing ambiguity and variability“, BMC Bioinformatics, Vol. 9, No. Suppl 3, 2008.

[16] E. Dura, O. Engkvist and S. Muresan, “Names of chemical compounds within drug discovery context“, Proc. of the 3rd Skövde Workshop on Information Fusion Topics, pp 55-58, 2009.

[17] K. M. Hettne, R. H. Stierum, M. J. Schuemie, P. J. M. Hendriksen, B. J. A. Schijvenaars, E. M. van Mulligen, J. Kleinjans, and J. A. Kors, “A dictionary to identify small molecules and drugs in free text“, Bioinformatics, September 16, 2009.