Concept-Based Semi-Automatic Classification of Drugs

Harsha Gurulingappa,*,†,‡ Corinna Kolářik,† Martin Hofmann-Apitius,†,‡ and Juliane Fluck†

Fraunhofer Institute for Algorithms and Scientific Computing, Schloss Birlinghoven, 53754, Sankt Augustin, Germany and Bonn-Aachen International Center for Information Technology B-IT, Dahlmannstrasse 2, 53113, Bonn, Germany

Received March 3, 2009

The anatomical therapeutic chemical (ATC) classification system maintained by the World Health Organization provides a global standard for the classification of medical substances and serves as a source for drug repurposing research. Nevertheless, it lacks several drugs that are major players in the global drug market. In order to establish classifications for yet unclassified drugs, this paper presents a newly developed approach based on a combination of information extraction (IE) and machine learning (ML) techniques. Most of the information about drugs is published in the scientific articles. Therefore, an IE-based framework is employed to extract terms from free text that express drug's chemical, pharmacological, therapeutic, and systemic effects. The extracted terms are used as features within a ML framework to predict putative ATC class labels for unclassified drugs. The system was tested on a portion of ATC containing drugs with an indication on the cardiovascular system. The class prediction turned out to be successful with the best predictive accuracy of 89.47% validated by a 100-fold bootstrapping of the training set and an accuracy of 77.12% on an independent test set. The presented concept-based classification system outperformed state-of-the-art classification methods based on chemical structure properties.

1. INTRODUCTION

The characteristics of a drug are captured by several factors, like physicochemical properties, biological properties, pharmacology, therapeutic indications, and side effects. During the entire life cycle of a drug from its design to synthesis, information about its properties and behavior are recorded in scientific publications and study reports that appear in the form of text. These textual sources comprise the entire wealth of factual statements, assumptions, hypotheses, and conclusions.[1] Efforts have been made to generate and maintain structured information sources such as databases, thesauri, ontologies, and hierarchies. They provide access to the breadth of information that has been accumulated. Such information sources have become a driver of productivity and economic growth leading to a new focus on decision making processes, especially during the design of new drugs.[2]

The pharmacopeias of different countries maintain their own classification system for drugs based on efficacy studies, mechanism of action, clinical outcomes, and market strategies. Two examples are the classification scheme maintained by the United States Pharmacopeia (USP)[3] and the Japanese Pharmacopeia, the Therapeutic Category of Drugs (TCD).[4] Currently, the most commonly used classification system for drugs is the ATC[5] classification system. This scheme hierarchically classifies drugs providing four different levels of granularity based on the organ system, therapeutic, pharmacological, and chemical properties of drugs. ATC considers new drugs entries only upon requests made by manufacturers, regulatory agencies, and researchers. Therefore, the system does not include substances for which no requests have been made or for withdrawn drugs. Within the drug research community, ATC has been used in a systems biology-based framework for finding new targets or medical indications for existing drugs thereby supporting drug repurposing research.[6−8]

Dunkel et al.[9] made a first set toward prediction of ATC classes for yet unclassified drugs or chemical compounds. They developed a web server called SuperPred[10] to predict ATC classes for chemical compounds. It uses a combination of physicochemical and substructure properties for class prediction. The main principle behind the operation of SuperPred is that compounds with similar structure exhibit similar biological activity. SuperPred relies on a basic data set of 2 500 compounds from the SuperDrug database that already have an ATC class annotation and whose structures are represented by structural fingerprints. Additional 3 800 compounds that are structurally similar with a Tanimoto coefficient >0.85 were added to the initial data set to form a larger database of 6 300 compounds. When the user submits a query compound, in the form of a SMILES or MOL file, SuperPred converts it into a structural fingerprint and performs a similarity search against the database of preannotated compounds. The results are given in terms of decreasing Tanimoto coefficient values, providing the most structurally similar database compounds as well as their ATC classes.

On the contrary, several other approaches are available for the determination of general pharmacological classes for drugs. Segura-Bedmar et al.[11] presented a new approach for drug name recognition in biomedical texts and drug classification. Their system combines information obtained from the MetaMap program (MMTx)[12] and nomenclature rules

* Corresponding author. E-mail: hgurulingappa@scai.fhg.de.
† Fraunhofer Institute for Algorithms and Scientific Computing.
‡ Bonn-Aachen International Center for Information Technology B-IT.

10.1021/ci9000844 CCC: $40.75
Published on Web 08/10/2009

ATC Classification System
C Cardiovascular System
C01 Cardiac Therapy
C02 Antihypertensives
C02A Antiadrenergic Agents, Centrally acting
C02B Antiadrenergic Agents, Ganglion Blocking
C03C Antiadrenergic Agents, Peripherally Acting
C04D Arteriolar Smooth Muscle, Acting Agents

TCD Classification System
21 Cardiovascular Agents
211 Cardiotonics
212 Antiarrhythemic Agents
213 Diuretics
214 Antihypertensives
2141 Ganglion Blockers
2142 Hydralazines
2143 Rauwolfias
2144 Angiotensin Converting Enzyme Inhibitors

Figure 1. Comparison of a subset of class labels between the ATC and TCD drug classification systems with indications on the cardiovascular system.

recommended by International Nonproprietary Names (INNs) Program to identify and classify pharmaceutical substances into general drug classes such as "antihypertensive", "antibiotic", "antiviral", etc.

Another approach that can be used to determine class labels for drugs is either hierarchy or ontology alignment[13,14] based on a precondition that drugs are already contained in the hierarchy. Hierarchy alignment is primarily based on the semantic similarity of the class labels. Since different classification schemes characterize drugs by different characteristics, hierarchy alignment is hard to achieve when it is used for aligning drug classes. Figure 1 demonstrates the semantic dissimilarities between categorical labels assigned by the ATC and TCD drug classification systems. ATC divides the class antihypertensives into antiadrenergic agents, centrally acting, antiadrenergic agents, ganglion blocking, for instance.

In contrast to ATC, the TCD divides antihypertensives into ganglion blockers, hydralazines, rauwolifias, and angiotensin converting enzyme inhibitors. Therefore, in order to determine the appropriate ATC class labels for drugs present in TCD, hierarchy alignment will not provide a direct solution, since the semantics of the classes are clearly dissimilar. Considering the limitations associated with the conventional approaches, this paper presents a new framework to predict ATC classes for unclassified drugs by utilizing information about drug's properties mentioned in scientific texts. Text is an interesting resource to find descriptions, and using these descriptions would provide a basis for a new approach, as developed here. The architecture of the framework, the underlying methods, and the results are discussed in the following sections.

2. SYSTEM ARCHITECTURE

The general idea behind this framework is to use textual descriptions about drug's biological and chemical properties as features for predicting their ATC class labels. Figure 2 illustrates the framework as it is currently implemented. The textual basis is titles and abstracts from Medline because of its wide coverage of the biomedical domain. A pattern-matching based procedure is used to extract property terms from the text for the drugs. Subsequently, the extracted terms are mapped to UMLS concepts with the help of the MMTx program. These concepts are used to generate a feature vector for every drug. The drugs and their feature vectors are subjected to the feature selection and ATC class prediction. Four well-established classifiers nearest neighbor, naïve Bayes, decision tree, and support vector machine (SVM)[15,16] are used for class prediction. Their performance is compared to select the one that produces the best result during a validation experiment performed on the training set. Finally, the best performing classifier is used to predict class labels of drugs in the test set, and the results are evaluated.

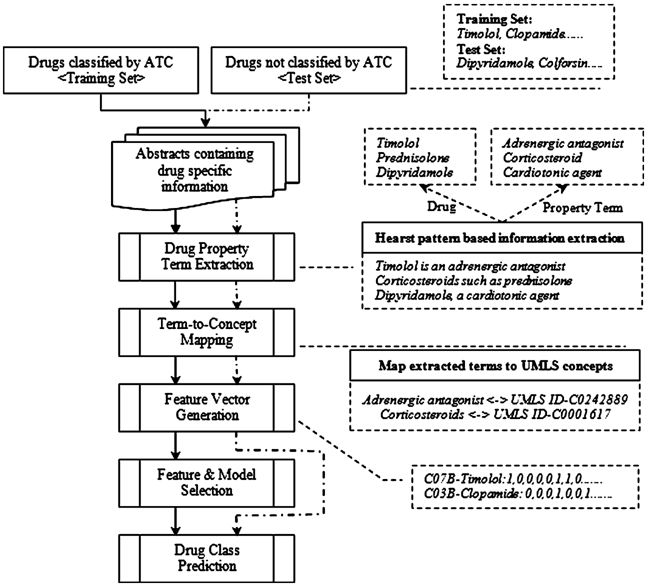
Training Set:
Timolol, Clopamide.....
Test Set:
Dipyridamole, Colforsin....
Drugs classified by ATC
<Training Set>
Drugs not classified by ATC
<Test Set>
Abstracts containing drug specific information
Timolol
Prednisolone
Dipyridamole
Adrenergic antagonist
Corticosteroid
Cardiotonic agent
Drug
Property Term
Drug Property Term Extraction
Hearst pattern based information extraction
Timolol is an adrenergic antagonist
Corticosteroids such as prednisolone
Dipyridamole, a cardiotonic agent
Term-to-Concept Mapping
Map extracted terms to UMLS concepts
Adrenergic antagonist <-> UMLS ID-C0242889
Corticosteroids <-> UMLS ID-C0001617
Feature Vector Generation
C07B-Timolol:1,0,0,0,0,1,1,0......
C03B-Clopamide: 0,0,0,1,0,0,1......
Feature & Model Selection
Drug Class Prediction

Figure 2. Illustration of the system architecture for ATC class prediction. Labels C07B and C03B are the ATC classes annotated on the training data.

3. METHODS

3.1. Data Set Preparation. The performance of the system was analyzed on a portion of ATC-containing drugs with indications on diseases of the cardiovascular system. The superclass cardiovascular agents contains 390 drugs that constitute 13% of the entire drugs classified by ATC with four levels of granularity. An important observation is that out of 390 drugs, only 10 appear in two different subclasses each. In the point of semantics, the class labels at level two are very general, like C02 antihypertensives or C03 diuretics. On the other hand, class labels at level three and four define more specific pharmacological or chemical properties of drugs. Class labels at level four have an insufficient number of drug instances per class in order to be used for supervised learning. Based on the semantics and population of classes at different levels, class labels at level three show an optimal information coverage as well as sufficient per class population to support a supervised learning process. Constraints were defined that classes and substances need to fulfill to be included, i.e.:

(a) A class must contain at least four drugs.

(b) Compounds and chemical substances (e.g., C02DB02 hydralazine) that are not drugs were not considered.

Therefore, noninformative and vacant classes as well as some compounds were excluded. This resulted in 21 level three ATC classes to be used for prediction. All drugs assigned to these classes form a training set of 390 drugs.

Table 1. Class Labels Used for Training and Prediction along with Their ATC Codes and Number of Drugs Contained within Each Class

ATC code	class label	no. of drugs
C03D	potassium sparing agents	6
C07D	calcium channel blockers with cardiac effect	8
C01A	cardiac glycosides	9
C03C	high ceiling diuretics	9
C02A	antiadrenergic agents, centrally acting	10
C03A	low ceiling diuretics, thiazide	10
C09C	angeotensin 2 antagonists	10
C05B	antivaricose agents	11
C02D	arteriolar smooth muscle, agents acting on	12
C02C	antiadrenergic agents, peripherally acting	13
C03B	low ceiling diuretics, nonthiazide	13
C07C	calcium channel blockers with vascular effect	17
C09A	ace inhibitors, plain	18
C01E	cardiac preparations	20
C01D	vasodilators used in cardiac diseases	23
C10A	lipid modifying agents	24
C04A	peripheral vasodilators	26
C01B	antihypertensives; class 1 and 3	30
C07A	beta blocking agents	32
C01C	cardiac stimulants	37
C05A	agents for treatment of hemorrhoids and anal fissures	52
Total		390

Table 1 shows all the class labels used for training and prediction along with the number of drugs contained in each class.

If a drug appears in two different classes, it was used twice in the training set with different class labels. An independent test set consists of 114 drugs with an indication on diseases of the cardiovascular system from the USP and TCD drug classification systems. For all drugs present in the test set, corresponding SMILES representations were extracted from the PubChem[17] database in order to be used as an input for SuperPred. The training and test sets together make up the working set of 504 drugs in total.

3.2. Drug Property Term Extraction. The aim of this process is to automatically identify and extract terms from text that describe pharmacological, systemic, therapeutic, and chemical properties of drugs that are present in the working set. Kolářik et al.[1] has shown that a lexico-syntactic pattern based IE system can be used to extract drug property information from texts. In the framework presented here, a similar kind of approach is followed.

The names and synonyms that include systemic names, trivial names, brand names, and company codes of all the drugs present in the working set provided by KEGG, DrugBank, and PubChem were used to generate a drug dictionary. This dictionary was used to retrieve drug specific corpora from Medline (title and abstracts) with the named entity recognition system ProMiner.[18] Retrieving drug specific corpora helps to reduce the amount of text that has to be subjected to IE by automatically removing those articles that do not contain any drug relevant information. Phrases that follow lexico-syntactic structures, called as Hearst patterns (HP), were extracted from the text. In general, phrases following this pattern, relate noun phrases, one more specific (hyponym), and the other more general (hypernym) by a taxonomic relationship. Some examples of HPs are as follows:

NP_1 is (a|an) NP_0

NP_1 is one of (the|a|an) NP_0

NP_0 such as NP_1, NP_2, (and|or) NP_3

NP_0 stands for noun phrase that represents a drug property term, and NP_1,..., NP_n stand for drugs that are described by NP_0. Examples for Hearst phrases are “timolol is a beta-adrenoceptor blocker” and “a vasodilator like propatyl nitrate”. Extracted phrases semantically not containing drug specific information were filtered out after their extraction. Subsequently, the phrases were automatically fragmented and assigned to their meaningful parts. For example, fragmenting the phrase “adinazolam is a benzodiazepine derivative” would result in adinazolam − a drug name and benzodiazepine derivative − a drug property term. The used Hearst phrase chunker was previously evaluated by Kolářik et al.[1] on an annotated abstract corpus for ibuprofen from Medline and resulted in a precision of 0.97 and a recall of 0.73. Precision measures the percentage of extracted property terms that are semantically valid, whereas recall measures the percentage of semantically valid property terms within the annotated text that have been extracted by the system. As a result of IE, every drug present in the working set was associated with a set of terms describing their properties.

3.3. Term-to-Concept Mapping. After the extraction of property terms from Medline articles, they were mapped to standard biomedical concepts in UMLS metathesaurus. MMTx (version 2.4.C)[19] was used to accomplish this task. This procedure has following advantages:

(a) Synonyms are mapped to one base concept (e.g., 5-HT antagonists and antiserotonergic agents are mapped to C0037753: serotonin antagonists, whereas C0037753 indicates a UMLS identifier).

(b) Term variants are mapped to a single concept (e.g., beta blocker and beta-blocking agent are mapped to C0001745: adrenergic beta antagonists).

The synonym and term variant mapping basically unifies the data and helps to overcome redundancy contained in the data set. Even though UMLS contains over two million concepts, not all identified property terms are present. As a result of this, several terms were only partially mapped or could not be mapped to the UMLS concepts. For example, the term potassium transporting ATPase inhibitor was split and mapped separately into UMLS concepts C0064194: potassium ATPase and C0597607: transport inhibitor. In order to overcome the problems of partial and unmapped terms, a non-UMLS concept list was generated and maintained manually, containing concepts that are not present in UMLS. Figure 3 demonstrates the process of term-to-concept mapping. For the terms that were mapped completely, their UMLS identifiers were directly used as features. Whereas, for terms that were unmapped or partially mapped by MMTx, they were checked for their existence in the non-UMLS concept list. If a concept was already present, then its non-UMLS identifier was used. Otherwise it was considered as a new entry within the concept list and assigned with a new identifier. Every entry within this concept list has a unique identifier, which is different from normal UMLS identifiers. Moreover, every entry in the concept list is associated with its synonyms, abbreviations, and spelling variants to avoid multiple appearances of a single concept. Table 2 shows an example of non-UMLS concepts present in the concept list. Finally, the complete term-to-concept mapping procedure

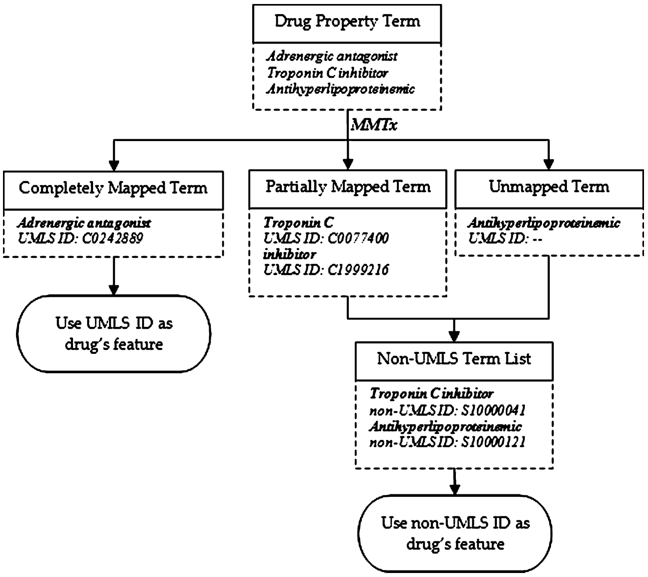

Figure 3. Illustration of the term-to-concept mapping procedure. The term adrenergic antagonist is mapped completely to UMLS concept by MMTx, and its UMLS identifier serves as drug's feature. The terms troponin C inhibitor and antihyperlipoproteinemic are either partially mapped or unmapped to UMLS concepts, and their non-UMLS identifiers are used as drug's features.

Table 2. Examples of Non-UMLS Concepts Present in the Concept List along with Their Non-UMLS Identifiers and Associated Term Variants and Synonyms

non-UMLS ID	non-UMLS concept	synonyms and variants
S10000089	adenylate cyclase inhibitor	adenylate cyclase antagonist adenylate cyclase blocker adenylate cyclase inhibiting agent adenylate cyclase blocking agent, etc.
S10000028	beta 2 adrenergic antagonist	beta 2 adrenergic receptor antagonist beta 2 blocker beta 2 adrenergic blocker beta 2 blocking agent, etc.

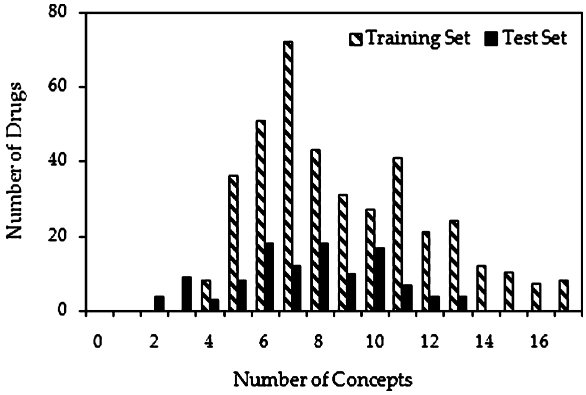

Figure 4. Distribution of number of concepts assigned to drugs in the working set.

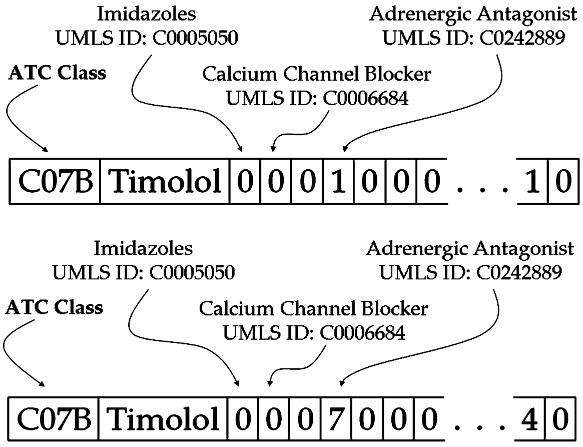

Figure 5. Example of binary feature vector (top) and weighted feature vector (bottom) for a drug in the training set labeled with its corresponding ATC class.

resulted in 368 unique concepts if pooled from all 390 drugs. A quantitative analysis showed that out of 368 concepts, two-thirds (i.e., 248 concepts) were completely mapped UMLS concepts, whereas the remaining one-third (i.e., 120 concepts) were non-UMLS concepts due to either partial or no mapping.

3.4. Feature Vector Generation. The concepts obtained during the previous procedure were used to generate a feature vector for every drug. Every concept represents one feature within that vector. Figure 4 shows the distribution of the number of concepts assigned to drugs in the working set. For the purpose of experimentation, two kinds of feature vectors were generated for every drug, i.e., a binary feature vector and a weighted feature vector of 368 numerical positions. Figure 5 shows an example of both a binary and weighted feature vector for the drug timolol in the training set.

In the binary feature vector, the value of a position is set to 1 if the corresponding concept was obtained for this drug or 0 if the concept was not present. In a weighted feature vector, however, the feature value equals the number of occurrences of the concept in combination with the drug in Hearst phrases. It is set to 0 if the concept was not obtained for this drug. Feature vectors of drug instances in the training set were labeled with the ATC class to which the corresponding drug belongs. The final task was to predict ATC class labels for drug instances in the test set. A list of 368 concepts and a sample data set containing drugs and their extracted properties represented by concept identifiers are available at http://scai.fhg.de/ATC-class-prediction.html.

3.5. Feature and Model Selection. To identify the concepts with the most discriminative power for a class prediction task, the methodology of the feature selection was applied. It is a technique commonly used in machine learning to select a subset of features for building robust learning models.[20] A lot of research work has been devoted to the feature selection, particularly in the text categorization area. Many feature selection algorithms include feature ranking as a primary mechanism to choose an appropriate feature set. They are relatively simple and scalable and have recorded empirical success.[21] Such techniques are also called information−theoretic ranking functions. A well-established chi-square (χ^2) criterion[22,23,27] was used to rank the concepts.

Four numeric values A, B, C, and D were calculated for each concept (f_p), with respect to every individual ATC class (c_i) that was used for training:

A = Number of occurrences of f_p in c_i.

B = Number of occurrences of f_p not in c_i.

C = Number of concepts in c_i that are not f_p.

D = Number of concepts that does not belong to c_i and are not f_p.

Table 3. Results of Feature Ranking with (A) Binary Feature Vectors and (B) Weighted Feature Vectors. Their Corresponding Concept Identifiers and Scores Are Shown

concept	ID	χ^2 score
Binary Feature Vectors		
dihyroxyphenylalanine	C0012315	378.38
steroid	C0338671	345.75
cardenolite	C0007143	345.75
AT1 receptor blocker	C1449680	328.56
vasoconstrictor	C0042397	321.29
Weighted Feature Vectors		
Na ATPase inhibitor	S10000001	425.61
adrenergic agonist	C0001648	402.45
AT1 receptor blocker	C1449680	390.00
Na channel antagonist	C0872271	381.41
dihyroxyphenylalanine	C0012315	366.07

Then, the χ^2 score between c_i and f_p is defined as

$$\chi^2(c_i,f_p) = \frac{N(AD - CB)^2}{(A + C)(B + D)(A + B)(C + D)} \tag{1}$$

where N is the total number of drugs in the training set. Finally, in order to measure the goodness of a concept in a global feature space, the class specific scores of a concept were combined as follows:

$$\chi^2_{\text{avg}}(f_p) = \left(\sum_{i=1}^{21} P(c_i)\right)\chi^2(c_i,f_p) \tag{2}$$

where $P(c_i) = N(c_i)/N$ with $N(c_i)$ = number of drugs belonging to the class c_i.

The χ^2 scores were calculated separately for all concepts occurring in the binary and weighted vectors. Table 3 illustrates an example of concepts ranked by the χ^2 criterion along with their scores. The scored concepts were sorted in a descending order that represents the concepts with highest discriminative power down to the lowest discriminative ones.

A crucial task after ranking the concepts is to select a subset of them that helps to build a good classifier. To accomplish this, a forward selection criterion was implemented. The subset selection was coupled with the model (classifier) selection where four classifiers naïve Bayes, nearest neighbor, decision tree, and SVM were tested. The aim of this procedure was to determine a subset of concepts and a robust classifier that can be employed to predict class labels for drugs present in the test set. For the purpose of supervised learning and class prediction, a popular machine learning toolbox Weka-3.5.3[24] was used. The provided algorithms naive Bayes, IBk, J48, and SMO were applied for naïve Bayes, k-nearest neighbor, decision tree, and SVM, respectively, for classification.

The experiment started by using drugs with binary feature vectors in both the training set and the χ^2 ranking of the concepts. The bit positions in the feature vectors that correspond to top 10% ranked concepts were withheld, and the remaining bit positions were set to zero. This means that only the top ranked 10% concepts were used to create an initial data set. For this data set, 100-fold bootstrapping generated 100 pairs of training and validation sets. Each fold of bootstrapping uses the principle of sampling with replacement on the original data set to partition it into a training and a validation set. The number of drawings equals to the number of instances in the data set where the drawn instances build the training set and the remaining instances form a validation set. Bootstrapping splits the data set into two-thirds (training set) vs one-third (test set/validation set), which is optimal for testing the classifiers. The four classifiers were run on the training sets and tested on the corresponding validation sets to obtain a comparable classification accuracy score. The accuracy was measured as percentage of correctly classified instances in the validation set. An average accuracy for 100 bootstrap experiments was noted as the classification accuracy for that data set. Next, another 10% of the concepts were added to the initial data set, and a similar validation by 100-fold bootstrapping was performed to check the classification accuracy. This process was repeated until all the ranked concepts were used for classification. Finally, the subset of the concepts and the classifier that contributes to the highest classification accuracy (global maximum) was monitored. Initially, the experiments were carried out with the default parameters provided for all four classifiers. Subsequently, the performance was monitored by changing the following key parameters for the individual classifiers:

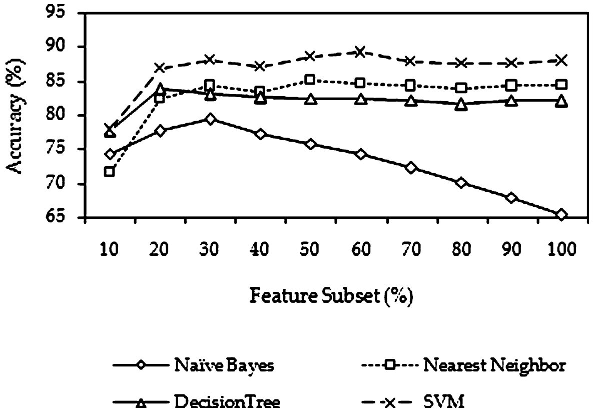

Figure 6. Performance of different classifiers validated by 100-fold bootstrapping using binary feature vectors. The parameters are $k = 1$ for k-nearest neighbor, kernel estimator 'off' for naïve Bayes, pruning 'off' for decision tree and RBF kernel for SVM.

k-nearest neighbor classifier: different number of neighbors (i.e., $k = 1, 2, 3,..., 10$).

Naïve Bayes: kernel estimator (on and off).

Decision Tree: Pruning (on and off).[25]

SVM: Kernel functions (polynomial and RBF kernel).[26]

Figure 6 shows the performance results of the classifiers using binary feature vectors and forward selection of the features. Only the parameters that produced the best overall classification for every individual classifier are reported. Later, a similar kind of experiment was carried out using weighted feature vectors and χ^2 ranking of the concepts present in these vectors. Figure 7 shows the performance results of different classifiers with weighted feature vectors and forward selection of features.

Observations from Figure 6 show that the SVM with RBF kernel outperformed the other three classifiers by a large margin when binary feature vectors were used. SVM generated the global maximum with a classification accuracy of 89.13 ± 2.32% with χ^2 ranked top 60% binary concepts. On the contrary, Figure 7 shows that naïve Bayes with kernel estimator outperformed the other classifiers when weighted feature vectors were used. It generated the global maximum with a classification accuracy of 89.47 ± 2.13% with χ^2

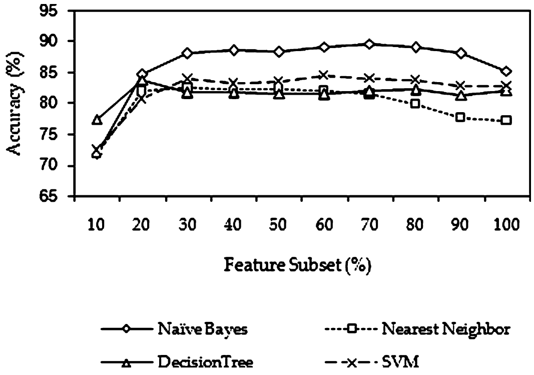

Figure 7. Performance of different classifiers validated by 100 fold bootstrapping using weighted feature vectors. The parameters are $k = 1$ for k-nearest neighbor, kernel estimator 'on' for naïve Bayes, pruning 'off' for decision tree and polynomial kernel for SVM.

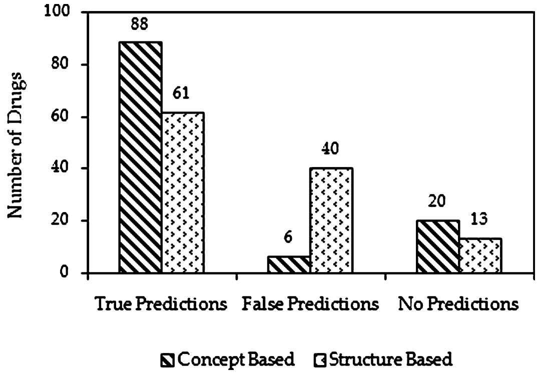

Figure 8. Comparison of concept-based with structure-based ATC class prediction for drugs in the test set.

ranked top 70% weighted concepts. Based on the outcome of classifier comparison, naïve Bayes was selected as a candidate classifier coupled with weighted feature vectors for the classification of drugs in the test set.

4. RESULTS

4.1. Drug Class Prediction. The final task of this framework was to predict ATC classes for drugs present in the test set and evaluate the performance of the classification. Therefore, the following two candidates were applied for this task:

Concept-based supervised classification using naïve Bayes classifier with kernel estimator parameter, weighted feature vectors, and χ^2 ranked top 70% concepts.

Structure-based classification using SuperPred.

SuperPred outputs a ranked list of ATC classes for an input structure. Because only the cardiovascular agents were used for testing, a top ranked class with an indication on the cardiovascular system was used for the evaluation. The proposed solutions for drug class prediction were performed independently, and their results were compared. For evaluation of the classification results, drug instances in the test set were manually annotated with ATC classes by human experts. Figure 8 shows the results of class prediction for drugs in the test set. True predictions indicate drugs that were correctly classified, false predictions indicates drugs that were misclassified, and no predictions indicates that drugs had too sparse features in order to be subjected to concept-based classification or the drug's SMILES representation was inappropriate to be processed by SuperPred.

Table 4. Examples of Misclassified Drugs in the Test Set along with Their KEGG Identifiers and the Number of Features They Contain

drug name	KEGG ID	no. of concepts
indenolol hydrochloride	D01958	11
colforsin daropate	D01697	8
rizatriptan benzoate	D00675	4
dipyridamole	D00302	8
aminophylline hydrate	D05429	6
anhydrous caffeine	D00528	9

The evaluation of the classification results shows that the concept-based approach was able to outperform the structure-based approach by classifying the drugs with an overall classification accuracy of 77.12%. The structure-based approach showed an overall classification accuracy of 53.51%. Thus, in this scenario, the concept-based approach performed significantly better than the structure-based approach. However, a noticeable drawback associated with this approach was that for 20 drugs in the test set, class prediction was not possible since they had sparse features. Sparse features means the number of valid concepts obtained for those drugs, as a result of information extraction, was insufficient for their class prediction. However, for the remaining drugs with sufficient number of concept, the class prediction turned out to be successful. Table 4 shows few examples of drug instances in the test set that were misclassified. In general, these drugs have sufficient number of concepts to be subjected to the ATC class prediction. One possible reason for their misclassification could be that they lack sufficient number of concepts that are relevant with respect to their class.

5. CONCLUSION AND FUTURE WORK

This paper has presented a new framework for concept-based drug class prediction. The experimental results demonstrated that the method is effective in classifying drugs by harvesting drug specific information from the scientific texts. The concept-based approach also showed its superiority in comparison to a structure-based approach in the given scenario.

In an application point of view, there are several scenarios within this framework that can be highlighted. Considering system biology research, where ATC has been used within a network based framework for determination of secondary drug effects,[6−8] the prediction of ATC classes for new drugs can enrich the network topology and help the researchers to arrive at new decisions concerning drug activities. Furthermore, the process of drug property term extraction resulted in various new drug annotation terms that are not mentioned in well curated databases like DrugBank.[1] These newly discovered property terms encompass information about the spectrum of drug targets and multiple therapeutic interventions as well as biochemical effects caused by drugs. Therefore, this new information can be used within a broader area to study poly pharmacology, off target effects, and multitherapeutics concerning drugs of interest. During the term-to-concept mapping procedure, it was observed that only two-thirds of the property terms were mapped to UMLS concepts. This gives an alerting signal for improvement and enhancement of the scope of knowledge covered within the biomedical thesauri and other structured information sources.

There are several strategies that can be implemented to improve the performance of this framework. First, in order to overcome the limitations of feature sparseness, as mentioned in Section 4.1, an application with more sophisticated techniques for extracting the drug properties by a deeper analysis of the structure of the sentence will extract more information about the drugs. Enhancement in the breadth of knowledge sources used for information extraction is another possible way for improvement. Currently, only Medline abstracts and titles have been used. But other information sources, like full text articles and patents, could include more descriptive information, such as extensive analysis of drug interactions, mode of action, toxicity, in vitro/in vivo biological test results, or safety details. Curated databases like DrugBank contain comprehensive information about drugs. Therefore, the utilization of additional information from such sources can help to enrich the feature space of drugs. Currently, an intensive work has been going on to extend the developed framework to include the entire ATC classes and also to solve the multiclass problem associated with few drugs.

To conclude, this framework opens a new possibility to categorize the existing drugs. Predicting ATC classes for yet unclassified drugs can support the standardization of information about drugs. Finally, it is a method that could be used to support strategies aimed at identifying potential secondary applications for existing drugs.

ACKNOWLEDGMENT

We thank all our co-workers, especially Dr. Christoph M. Friedrich and Mr. Roman Klinger for helping in various aspects of machine learning. The work of Harsha Gurulingappa and Corinna Kolářik is funded by the Bonn-Aachen International Center for Information Technologies (http://www.b-it-center.de). We also thank the anonymous reviewers for their comments and recommendations.

REFERENCES AND NOTES

(1) Kolářik, C.; Hofmann-Apitius, M.; Zimmermann, M.; Fluck, J. Identification of new drug classification terms in textual resources. *Bioinformatics* **2007**, *23*, 264–272.

(2) Krallinger, M.; Erhardt, R. A.; Valencia, A. Text-mining approaches in molecular biology and biomedicine. *Drug Discovery Today* **2005**, *10*, 439–445.

(3) USP drug classification system. http://www.genome.jp/kegg-bin/get_htext?br08302.keg; accessed May 29, 2008.

(4) TCD drug classification system. http://www.genome.jp/kegg-bin/get_htext?br08301.keg; accessed May 29, 2008.

(5) ATC drug classification system. http://www.genome.jp/kegg-bin/get_htext?br08303.keg; accessed May 29, 2008.

(6) Nacher, J. C.; Schwartz, J. M. A global view of drug-therapy interactions. *BMC Pharmacol.* **2008**, *8*, 5–13.

(7) Spiro, Z.; Kovacs, I.; Csermely, P. Drug-therapy networks and the prediction of novel drug targets. *J. Biol.* **2008**, *7*, 20–27.

(8) Campillos, M.; Kuhn, M.; Gavin, A. C.; Jensen, L. J.; Bork, P. Drug target identification using side-effect similarity. *Science* **2008**, *5886*, 263–266.

(9) Dunkel, M.; Günther, S.; Ahmed, J.; Wittig, B.; Preissner, R. SuperPred: drug classification and target prediction. *Nucleic Acids Res.* **2008**, *36*, 55–59.

(10) SuperPred: Drug classification and target prediction. http://bioinformatics.charite.de/superpred/; accessed Jul 14, 2008.

(11) Segura-Bedmar, I.; Martínez, P.; Segura-Bedmar, M. Drug name recognition and classification in biomedical texts, A case study outlining approaches underpinning automated systems. *Drug Discovery Today* **2008**, *13*, 816–823.

(12) Aronson, A. R. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc. AMIA Symp.* **2001**, 17–21.

(13) Ryutaro, I.; Hideaki, T.; Shinichi, H. Rule Induction for Concept Hierarchy Alignment. *International Joint Conferences on Artificial Intelligence* **2001**, 26–29.

(14) Lanzenberger, M.; Sampson, J. AlViz - A Tool for Visual Ontology Alignment. In. *IEEE Info. Vis.* **2006**, 430–440.

(15) Huang, J.; Lu, J.; Ling, C. X. Comparing Naive Bayes, Decision Trees, and SVM with AUC and Accuracy. *IEEE Int. Conf. Data Mining* **2003**, 553–556.

(16) A practical guide to support vector classification. http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf (accessed Apr 11, 2008).

(17) PubChem: A free database of chemical structures of small organic molecules and information on their biological activities. http://pubchem.ncbi.nlm.nih/gov (accessed Jul 26, 2008).

(18) Hanisch, D.; Fundel, K.; Mevissen, H. T.; Zimmer, R.; Fluck, J. ProMiner: rule based protein and gene entity recognition. *BMC Bioinformatics* **2005**, *6*, 14–22.

(19) MetaMap Transfer (MMTx) program. http://mmtx.nlm.nih.gov (accessed Jun 14, 2008).

(20) Sebastiani, F. Machine learning in automated text categorization. *ACM Comput. Surv.* **2002**, *34*, 1–47.

(21) Guyon, I.; Elisseeff, A. An introduction to variable and feature selection. *J. Machine Learning Res.* **2003**, *3*, 1157–1182.

(22) Yang, Y.; Liu, X. A re-examination of text categorization methods. *ACM Special Interest Group on Information Retrieval* **1999**, 42–49.

(23) Al-Mubaid, H. Context-Based Technique for Biomedical Term Classification. *IEEE Congress on Evolutionary Computation* **2006**, 5726–5733.

(24) Data Mining with Open Source Machine Learning in Java. http://www.cs.waikato.ac.nz/ml/weka (accessed Aug 11, 2008).

(25) Safavian, S. R.; Landgrebe, D. A Survey of Decision Tree Classifier Methodology. *IEEE Int. Conf. on Systems, Man, and Cybernetics* **1991**, *21*, 660–674.

(26) Chen, G. Y.; Bhattacharya, P. Function Dot Product Kernels for Support Vector Machine. *International Conference on Pattern Recognition* **2006**, *2*, 614–617.

(27) Zheng, Z.; Wu, X.; Srihari, R. Feature selection for text categorization on imbalanced data. *ACM Special Interest Group on Knowledge Discovery and Data Mining* **2004**, *6*, 80–89.

(28) Martin, Y.; Kofron, J.; Traphagen, L. Do structurally similar molecules have similar biological activity. *J. Med. Chem.* **2002**, *45*, 4350–4358.

CI9000844