A Machine Learning Approach to Extract Drug – Drug Interactions in an Unbalanced Dataset

Jacinto Mata, Ramón Santano, Daniel Blanco,
Marcos Lucero, Manuel J. Maña

Escuela Técnica Superior de Ingeniería. Universidad de Huelva
Ctra. Huelva - Palos de la Frontera s/n. 21819 La Rábida (Huelva)
{jacinto.mata, manuel.mana}@dti.uhu.es,
{ramon.santano, daniel.blanco, marcos.lucero}@alu.uhu.es

Abstract. Drug-Drug Interaction (DDI) extraction from the pharmacological literature is an emergent challenge in the text mining area. In this paper we describe a DDI extraction system based on a machine learning approach. We propose distinct solutions to deal with the high dimensionality of the problem and the unbalanced representation of classes in the dataset. On the test dataset, our best run reaches an F-measure of 0.4702.

Keywords: Drug-drug interaction, machine learning, unbalanced classification, feature selection.

1 Introduction

One of the most relevant problems in patient safety is the adverse reaction caused by drugs interactions. In [3], it is claimed that 1.5 million adverse drug events and tens of thousands of hospital admissions take place each year. A Drug-Drug Interaction (DDI) occurs when the effect of a particular drug is altered when it is taken with another drug. The most updated source to know DDI is the pharmacological specialized literature. However, the automatic extraction of DDI information from this huge document repository is not a trivial problem. In this scenario, text mining techniques are very suitable to deal with this kind of problems.

Different approaches are used in DDI extraction. In [9], the authors propose a hybrid method based on linguistic and pattern rules to detect DDI in the literature. Linguistic rules grasp syntactic structures or semantic meanings that could discover relations from unstructured texts. Pattern-based rules encode the various forms of expressing a given relationship. As far as we know, there are not many works applying machine learning approaches to this task due to the inexistence of available corpora. In [10] a SVM classifier was used to extract DDI into the DrugDDI corpus. However, in the similar problem of protein-protein interaction (PPI) has been widely used obtaining promising effectiveness, as in [7]. The main advantages of this

Proceedings of the 1st Challenge task on Drug-Drug Interaction Extraction (DDIExtraction 2011), pages 59–65,
Huelva, Spain, September 2011.

approach are that they can be easily extended to new set of data and the development effort is considerably lower than manual encoding of rules and patterns.

In this paper we present a machine learning approach to extract DDI using the DrugDDI corpus [10]. Natural Language Processing (NLP) techniques are used to analyze documents and extracting features which represent them. The unbalanced proportion between positive and negative classes in the corpus suggest us the application of sampling techniques. We have experimented with several machine learning algorithms (SVM, Naïve Bayes, Decision Trees, Adaboost) in combination with feature selection techniques in order to reduce the dimensionality of the problem.

The paper is organized as follows. The system architecture is presented in section 2. In Section 3 we describe the set of features that represents each pair of drugs which appears in the documents. Also we present the feature selection methods used to reduce the initial set of attributes. Next, Section 4 describes the techniques that we have used to deal with this unbalanced classification problem. In Section 5 we evaluate the results obtained with the training corpus. The results on the test corpus are presented in Section 6. Finally, the conclusions are in Section 7.

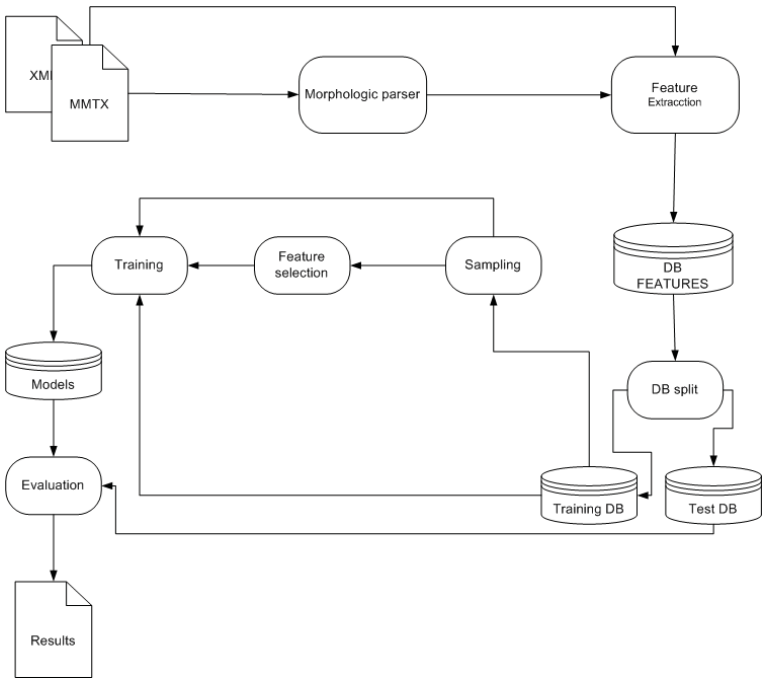

Fig. 1. System Architecture Diagram.

2 System Architecture

Two different document formats has been provided by the organizers, the Unified format and the MMTx format. We have used this last one to develop and testing our system.

The words around the drugs in a sentence have been selected as attributes of the database because they could provide clues about the existence of interaction between two drugs. We have experimented using the words as they appear in the documents and, in other cases, with the lemmas provided by the Stanford University morphologic parser[1].

For each drug pair in a sentence a set of features was extracted. The main features were focused on keywords, distances between drugs and drug semantic types. In the next section, a more detailed description of each attribute is done.

In order to carry out the experimentation, the DB of Features was split in two datasets for training and testing. We have used 2/3 of the original DB for training the classifier. The remaining 1/3 was used to test the system during the development phase.

Before training the classifier we have experimented with two preprocessing techniques. Because this problem is an unbalanced classification task we have carried out sampling techniques. Also, to reduce the dimensionality of the dataset a feature selection technique was performed. To obtain the model, we have experimented with several machine learning algorithms (SVM, Naïve Bayes, Decision Trees, Adaboost).

With each obtained model an evaluation was completed using the test dataset. The results obtained in this evaluation are shown in Section 5.

3 Feature Extraction and Selection

The most important part in this kind of classifying problem is to choose the set of features that represents as well as possible each pair of drugs. It means that we need to find those features that provide important information for differentiating pairs of drugs with interaction of pairs without interactions.

In this section we describe the features we have chosen to build the dataset.

3.1 Features

Firstly, we have extracted the drug ID, which indicates the sentence and the phrase of the dataset to which the drug belongs to.

Secondly, a feature subset composed by keywords was chosen. Each attribute is represented by a binary value that means the presence or absence of this keyword. Three windows of tokens have been considered to locate the keywords: between the first and the second drug, before the first drug and after the second drug. In the last two cases, only three tokens were taken into account.

[1] http://nlp.stanford.edu/index.shtml

In this work, a keyword is a word that could provide relevant information about whether a pair of drugs interacts or not. In order to build the list of keywords we extracted all the words between each pair of drugs, before the first drug or after the second drug, according the case. This set of words was filtered by a short list of stop-words. The POS tag of each word has been taken into account to make the selection. In this sense, we thought that verbs have an important semantic content, so we decided to include all of them into the final list. With respect to the nouns, we did a manual selection choosing those nouns that could be related semantically with drug interactions. Finally, in the case of prepositions, adverbs and conjunctions, we selected those that could be related with negation or frequency.

We have experimented using the keywords as they appear in the documents and, in other cases, with the lemmas provided by the Stanford University morphologic analyzer. In this case, the number of keywords was reduced because distinct verb tenses or plurals of a word were reduced to their lemmas, obtaining a total of 459 attributes.

Next, we added to the feature set the distance, in number of words and phrases, between the drugs. Also we included two features that represent the semantic type of each drug (represented by integer numbers).

Finally, the feature set is completed with the class, a binary value, where 1 means drug interaction and 0 if the pair does not interact.

As we can see in Table 1, we have extracted a total of 600 features from the original dataset to build the develop dataset.

Table 1. Feature set without lemmatization of the keywords.

Feature	Type	Number of features
Drugs ID	Integer	2
Keywords before first drug	Binary	153
Keywords between drugs	Binary	243
Keywords after second drug	Binary	197
Number of words between drugs	Integer	1
Number of phrases between drugs	Integer	1
Drug semantic types	Integer	2
Class	Binary	1
Total		600

3.2 Feature selection

Due to the high dimensionality of the training dataset, we have experimented with chi-squared feature selection method [8]. This method returns a ranking of the features in decreasing order by the value of the chi-squared statistic with respect to the class. We selected the attributes which the statistic had a value greater than 0. The resulting dataset, in the case of keywords without lemmatization, had 496 attributes.

4 Unbalanced Classification

As shown in Table 2, there are 23827 drug pairs in the develop dataset and only 2409 are real drug interactions. Therefore, the positive class is nearly the 10% (9.89%) of the total number of instances. It is a classification task with unbalanced classes. To deal with this problem we have used the SMOTE algorithm [2] in order to balance the classes.

Several classification algorithms have been selected in order to obtain the best effectiveness results with respect to the F-measure of the positive class. We have used the Weka [4] implementation of the following algorithms: RandomForest [1], Naïve Bayes [5], SMO [6] and MultiBoosting [11].

In some cases, to build the classification model, we have applied a cost sensitive matrix in order to penalize false positives.

5 Experimentation on Training Corpus

The develop corpus contains a collection of pharmacological texts labeled with drug interactions. This collection consists of 4267 sentences extracted from a total of 435 documents, which describe the interactions between drugs (Drug Drug Interactions or DDI). From these documents we have extracted 23827 drug pairs as possible cases of interaction. In total, there are 2409 instances corresponding to drug interactions and 21418 instances where there is no interaction between drugs.

Table 2 summarizes the training corpus statistics.

Table 2. Training corpus statistics.

Total different documents (files)	435
Number of documents containing, at least, one drug	412
Number of documents containing, at least, one drug pair	399
Total number of sentences	4267
Total number of drugs	11260
Total number of drug pairs	23827
Number of drug interactions	2409
Total entities that participate in a pair	10374
Average drug per document (documents and sentences with pairs)	25.88
Average drug per sentence (sentences with pairs)	4.67

In the experiment phase, we divided the dataset into two new datasets for training and testing, respectively. The training dataset consists of 2/3 of the total instances (15885). The test dataset consists of the remaining instances (7942).

The distribution of the instances for training and test datasets was done at random, keeping the percentage of instances with drug interaction and no interaction (10% and 90%, respectively).

Table 3 shows the effectiveness results for precision, recall and F-measure on the positive class of the 10 best evaluations. Each row of the table indicates a different

combination of classification algorithm, cost sensitive training, feature selection, sampling and keyword lemmatization.

As can be seen, the best results are obtained with the RandomForest algorithm. Moreover, the cost sensitive training, feature selection, sampling and lemmatization of the keywords contribute to achieve the best F-measures.

Table 3. Evaluation on training corpus. The second column is the classification algorithm. For RandomForest algorithm, the *I* parameter means the number of trees used to train the model. The *CST* column indicates whether the model has been built using a cost sensitive training. Different cost sensitive matrixes have been used in the experimentation phase. The *FS* column shows when feature selection has been carried out. The *Sampling* column has the same meaning with the application of SMOTE algorithm. Finally, *KW Lem.* column shows a lemmatization process has been performed.

RUN	Classification algorithm	CST	FS	Sampling	KW Lem.	Precision	Recall	F-Measure
1	RandomForest (*I* = 50)	X	X	X		0.573	0.617	0.595
2	RandomForest (*I* = 50)	X	X	X	X	0.578	0.610	0.594
3	RandomForest (*I* = 10)	X	X	X	X	0.500	0.654	0.567
4	RandomForest (*I* = 10)	X		X	X	0.492	0.644	0.558
5	RandomForest (*I* = 10)	X	X	X		0.565	0.548	0.556
6	RandomForest (*I* = 10)	X	X	X		0.469	0.677	0.554
7	RandomForest (*I* = 50)	X			X	0.645	0.472	0.545
8	MultiBoosting		X	X		0.674	0.443	0.535
9	RandomForest (*I* = 10)	X		X		0.544	0.520	0.532
10	RandomForest (*I* = 10)	X				0.587	0.471	0.523

6 Results on Test Corpus

In order to send runs with different characteristics, we didn't send the five runs with higher value of F-measure. According to Table 3, runs 1, 2, 4, 7 and 8 were submitted. We chose this strategy because we did not know the characteristics of the test corpus.

In Table 4, we present the results obtained for the five submitted runs. The approaches that obtain the best results on the training dataset coincide with the obtained on the test dataset. Although there are not significant differences between precisions on training and test datasets, a greater decrement in the recall measure do that the F-measure falls a 10% approximately. We think that this decrement in the effectiveness measures is due to a possible overfitting of the classification models.

7 Conclusions

In this paper we have presented a DDI extraction system based on a machine learning approach. We have proposed distinct solutions to deal with the high dimensionality of the problem and the unbalanced representation of classes in the dataset. The results obtained on both datasets are promising and we think that this could be a good starting point for future improvements.

Table 4. Evaluation on test corpus. The second column is the classification algorithm. For RandomForest algorithm, the *I* parameter means the number of trees used to train the model. The *CST* column indicates whether the model has been built using a cost sensitive training. Different cost sensitive matrixes have been used in the experimentation phase. The *FS* column shows when feature selection has been carried out. The *Sampling* column has the same meaning with the application of SMOTE algorithm. Finally, *KW Lem.* column shows a lemmatization process has been performed.

RUN	Classification algorithm	CST	FS	Sampling	KW Lem.	Precision	Recall	F-Measure
1	RandomForest (*I* = 50)	X	X	X		0.5000	0.4437	0.4702
2	RandomForest (*I* = 50)	X	X	X	X	0.4662	0.4291	0.4669
3	RandomForest (*I* = 10)	X		X	X	0.4004	0.4874	0.4397
4	RandomForest (*I* = 50)	X			X	0.6087	0.3152	0.4154
5	MultiBoosting		X	X		0.6433	0.2556	0.3659

References

1. Breiman, L. Random Forests. Machine Learning, 2001. Vol. 45(1):5-32.
2. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P. Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 2002. Vol. 16:321-357.
3. Classen, D.C., Phansalkar, S., Bates, D.W. Critical drug-drug interactions for use in electronic health records systems with computerized physician order entry: review of leading approaches. J. Patient Safety 2011 ,Jun;7(2):61-5.
4. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten. I.H. The WEKA Data Mining Software: An Update; SIGKDD Explorations 2009, Vol. 11, Issue 1.
5. John, G.H., Langley, P. Estimating Continuous Distributions in Bayesian Classifiers. In: Eleventh Conference on Uncertainty in Artificial Intelligence, San Mateo, 338-345, 1995.
6. Keerthi, S.S., Shevade, S.K., Bhattacharyya, C., Murthy, K.R.K. Improvements to Platt's SMO Algorithm for SVM Classifier Design. Neural Computation, 2001. 13(3):637-649.
7. Krallinger, M., Leitner F., Valencia, A. The BioCreative II.5 challenge overview. Proceedings of the BioCreative II. 5 Workshop 2009 on Digital Annotations 2009, 19.
8. Liu, H., Setiono, R., Chi2. Feature selection and discretization of numeric attributes, Proc. IEEE 7th International Conference on Tools with Artificial Intelligence, 338-391, 1995.
9. Segura-Bedmar, I., Martínez, P., de Pablo-Sánchez, C. A linguistic rule-based approach to extract drug-drug interactions from pharmacological documents, March, 2011, BMC BioInformatics, Vol. 12 (Suppl 2):S1.
10. Segura-Bedmar, I., Martinez, P., de Pablo-Sanchez, C. Using a shallow linguistic kernel for drug-drug interaction extraction, Journal of Biomedical Informatics, In Press, Corrected Proof, Available online 24 April 2011, DOI: 10.1016/j.jbi.2011.04.005.
11. Webb, G.I. MultiBoosting: A Technique for Combining Boosting and Wagging. Machine Learning 2000. Vol.40(No.2).