JRS 2012 Data Mining Challenge

Overview

JRS 2012 Data Mining Competition: Topical Classification of Biomedical Research Papers, is a special event of Joint Rough Sets Symposium (JRS 2012, http://sist.swjtu.edu.cn/JRS2012/) that will take place in Chengdu, China, August 17-20, 2012. The task is related to the problem of predicting topical classification of scientific publications in a field of biomedicine. Money prizes worth 1,500 USD will be awarded to the most successful teams. The contest is funded by the organizers of the JRS 2012 conference, Southwest Jiaotong University, with support from University of Warsaw, SYNAT project and TunedIT.

Introduction: Development of freely available biomedical databases allows users to search for documents containing highly specialized biomedical knowledge. Rapidly increasing size of scientific article meta-data and text repositories, such as MEDLINE [1] or PubMed Central (PMC) [2], emphasizes the growing need for accurate and scalable methods for automatic tagging and classification of textual data. For example, medical doctors often search through biomedical documents for information regarding diagnostics, drugs dosage and effect or possible complications resulting from specific treatments. In the queries, they use highly sophisticated terminology, that can be properly interpreted only with a use of a domain ontology, such as Medical Subject Headings (MeSH) [3]. In order to facilitate the searching process, documents in a database should be indexed with concepts from the ontology. Additionally, the search results could be grouped into clusters of documents, that correspond to meaningful topics matching different information needs. Such clusters should not necessarily be disjoint since one document may contain information related to several topics. In this data mining competition, we would like to raise both of the above mentioned problems, i.e. we are interested in identification of efficient algorithms for topical classification of biomedical research papers based on information about concepts from the MeSH ontology, that were automatically assigned by our tagging algorithm. In our opinion, this challenge may be appealing to all members of the Rough Set Community, as well as other data mining practitioners, due to its strong relations to well-founded subjects, such as generalized decision rules induction [4], feature extraction [5], soft and rough computing [6], semantic text mining [7], and scalable classification methods [8]. In order to ensure scientific value of this challenge, each of participating teams will be required to prepare a short report describing their approach. Those reports can be used for further validation of the results. Apart from prizes for top three teams, authors of selected solutions will be invited to prepare a paper for presentation at JRS 2012 special session devoted to the competition. Chosen papers will be published in the conference proceedings.

Contest Participation Rules:

- The competition is open for all interested researchers, specialists and students. Only members of the Contest Organizing Committee cannot participate.
- Participants may submit solutions as teams made up of one or more persons. Each team needs to designate a leader responsible for communication with the Organizers. One person may be incorporated in maximally 2 teams.
- The total number of submission for any single team is limited to 200 solutions.
- Each team is obliged to provide a short report describing their final solution. Reports must contain information such as the name of a team, names of all team members, the last preliminary evaluation score and a brief overview of the used approach. Their length should not exceed 1000 words and they should be sent in the pdf format to JRS12Contest@mimuw.edu.pl by April 2, 2012. Only submissions made by teams that provided the reports will qualify for the final evaluation.

JRS 2012 conference special session: There will be a special session at the JRS 2012 conference devoted to the competition. We will invite authors of selected reports to extend them for publication in the proceedings (after reviews by Organizing Committee members) and presentation at the conference. The invited teams will be chosen based on their rank and innovativeness of approach.

Awards: Top ranked solutions (based on the final evaluation scores) will be awarded with prizes:

- First Prize: 1,000 USD + free JRS 2012 conference registration,
- Second Prize: 500 USD + free JRS 2012 conference registration,
- Third Prize: free JRS 2012 conference registration.

Additionally, at the conference, authors of all papers accepted for presentation at the special session will receive a diploma and a competition T-shirt.

Schedule:

- Jan. 2, 2012: start of the challenge, data sets become available,

- Mar. 30, 2012: deadline for submitting the predictions,
- Apr. 2, 2012: deadline for sending the reports, end of the challenge,
- Apr. 6, 2012: on-line publication of final results, sending invitations for submitting short papers for the special session,
- May 10, 2012: deadline for submissions of camera-ready papers selected for presentation at the JRS special session.

Contest Organizing Committee:

- Andrzej Janusz (Chairman), University of Warsaw
- Hung Son Nguyen, University of Warsaw
- Dominik Ślęzak, University of Warsaw & Infobright Inc.
- Sebastian Stawicki, University of Warsaw
- Adam Krasuski, Main School of Fire Service & University of Warsaw

References:

[1] National Library of Medicine: PubMed: The Bibliographic Database. In McEntyre J., Ostell J.(Eds.): The NCBI Handbook. Available online, http://www.ncbi.nlm.nih.gov/books/NBK21094/

[2] National Library of Medicine: PubMed Central (PMC): An Archive for Literature from Life Sciences Journals. In McEntyre J., Ostell J. (Eds.): The NCBI Handbook. Available online, http://www.ncbi.nlm.nih.gov/books/NBK21087/

[3] National Library of Medicine: Introduction to MeSH - 2012. Available online (2012), http://www.nlm.nih.gov/mesh/introduction.html

[4] Greco S., Pawlak Z., Słowiński R.: Generalized Decision Algorithms, Rough Inference Rules, and Flow Graphs. J. J. Alpigini, J. F. Peters, A. Skowron and N. Zhong (Eds.): Rough Sets and Current Trends in Computing 2002, LNCS 2475, Springer-Verlag, London, UK (2002)

[5] Guyon I. et al.: Feature Extraction: Foundations and Applications. Studies in Fuzziness and Soft Computing. Springer (August 2006)

[6] Hassanien A. E., Suraj Z., Ślęzak D., Lingras P. (Eds.): Rough Computing: Theories, Technologies and Applications. Idea Group Inc (2007)

[7] Stavrianou A., Andritsos P., Nicoloyannis N.: Overview and semantic issues of text mining. SIGMOD Rec. 36, 3, pp. 23-34, (September 2007)

[8] Nguyen H. S.: Scalable Classification Method Based on Rough Sets. In Alpigini J. J., Peters J. F., Skowronek J., Zhong N. (Eds.): Rough Sets and Current Trends in Computing 2002, LNCS 2475, pp. 433-440. Springer-Verlag, London, UK (2002)

Task

Our team has invested a significant amount of time and effort to gather a corpus of documents containing 20,000 journal articles from the PubMed Central open-access subset. Each of those documents was labeled by biomedical experts from PubMed with several MeSH subheadings that can be viewed as different contexts or topics discussed in the text. With a use of our automatic tagging algorithm, which we will describe in details after completion of the contest, we associated all the documents with the most related MeSH terms (headings). The competition data consists of information about strengths of those bonds, expressed as numerical values. Intuitively, they can be interpreted as values of a rough membership function that measures a degree in which a term is present in a given text. The task for the participants is to devise algorithms capable of accurately predicting MeSH subheadings (topics) assigned by the experts, based on the association strengths of the automatically generated tags. Each document can be labeled with several subheadings and this number is not fixed. In order to ensure that participants who are not familiar with biomedicine, and with the MeSH ontology in particular, have equal chances as domain experts, the names of concepts and topical classifications are removed from data. Those names and relations between data columns, as well as a dictionary translating decision class identifiers into MeSH subheadings, can be provided on request after completion of the challenge.

Data format: The data set is provided in a tabular form as two tab-separated values files, namely *trainingData.csv* (the training set) and *testData.csv* (the test set). They can be downloaded only after a successful registration to the competition. Each row of those data files represents a single document and, in the consecutive columns, it contains integers ranging from 0 to 1000, expressing association strengths to corresponding MeSH terms. Additionally, there is a *trainingLables.txt* file, whose consecutive rows correspond to entries in the training set (*trainingData.csv*). Each row of that file is a list of topic identifiers (integers ranging from 1 to 83), separated by commas, which can be regarded as a generalized classification of a journal article. This information is not available for the test set and has to be predicted by participants.
It is worth noting that, due to nature of the considered problem, the data sets are highly dimensional - the number of columns roughly corresponds to the MeSH ontology size. The data sets are also sparse, since usually only a small fraction of the MeSH terms is assigned to a particular document by our tagging algorithm. Finally, a large number of data columns have little (or even none) non-zero values (corresponding concepts are rarely assigned to documents). It is up to participants to decide which of them are still useful for the task.

Format of submissions: The predictions should be submitted in a single text file containing exactly the same number of lines as the test data set. In the consecutive lines, this file should contain identifiers of the predicted topics (integers ranging from 1 to 83) separated by commas and without any spaces. The file *majorityClasses.txt* is an example of a well-formatted submission. It assigns each document to five most frequently occurring classes.
Apart from submitting the solution file, each participating team is required to send a short report (should not exceed 1000 words) describing their final solution. A report should contain the name of a team, names of all team members, the last preliminary evaluation score and a brief overview of the used approach, such as data preprocessing steps, utilized algorithms, parameter tuning techniques, and so on. The reports (pdf file format is preferable) should be send to:
JRS12Contest@mimuw.edu.pl due to 02 April, 2012.

Downloads: You must be logged in and registered to this challenge in order to access the files.

- *Training data* - an information system containing 10,000 objects and 25,640 attributes
- *Test data* - an information system containing 10,000 objects and 25,640 attributes
- *Training decision labels* - generalized topical classification of training objects
- *Example solution* - an exemplary well-formatted submission file

Evaluation of results: The submitted solutions will be evaluated on-line and the preliminary results will be published on the leaderboard. The preliminary score will be computed on a random subset of the test set, fixed for all participants. It will correspond to approximately 10% of the test data size. The final evaluation will be performed after completion of the competition using the remaining part of the test data. Those results will also be published on-line. It is important to note that only teams which send a short report describing their approach before the end of the contest will qualify for the final evaluation. The winning teams will be officially announced during a special session devoted to the competition at the JRS 2012 (http://sist.swjtu.edu.cn/JRS2012/) conference.
Quality of the submissions will be evaluated using a standard measure from Information Retrieval that naturally fits to the considered problem, namely average F-score of the predictions. Let us use the following notation:

$$\begin{aligned} N &- \text{the number of test documents,} \\ \mathrm{TrueTopics}_i &- \text{a set of true topics for a document } i \text{ (given by experts),} \\ \mathrm{PredTopics}_i &- \text{a set of predicted topics for a document } i. \end{aligned}$$

We can define Precision and Recall of a prediction for a single document:

$$\mathrm{Precision}_i = \frac{|\mathrm{TrueTopics}_i \cap \mathrm{PredTopics}_i|}{|\mathrm{PredTopics}_i|}$$

$$\mathrm{Recall}_i = \frac{|\mathrm{TrueTopics}_i \cap \mathrm{PredTopics}_i|}{|\mathrm{TrueTopics}_i|}$$

Average F-score over the test documents will be defined as:

$$\mathrm{Fscore}_i = 2 \cdot \frac{\mathrm{Precision}_i \cdot \mathrm{Recall}_i}{\mathrm{Precision}_i + \mathrm{Recall}_i}$$

$$\mathrm{AvgFscore} = \frac{\sum_{i=1}^{N} \mathrm{Fscore}_i}{N}$$

Intuitively, average F-score combines precision and recall of predictions over the set of all test documents. In case of any draws in the results, the final ranking of participants will be decided based on time of the submissions.