Exposé:

Semi-Supervised Learning: Can Text Mining Help Identify Genes Associated with Obesity?

Martin Schmidt, 26. 9. 2005
Supervisors: Prof. Ulf Leser, Prof. Gudrun Brockmann, Dr. Armin Schmitt, Jörg Hakenberg
Time period: October 2005 - March 2006

1. Background

The Obesity Gene Map [1] (OGM) project is an online collection of data about markers, genes and mutations linked with the disease obesity. The collaborators of this project collect the data by reviewing the relevant literature. The project members search for relevant publications in the biomedical database PubMed [2], in obesity and genetic journals, in the collaborators' personal collection of reprints and in papers made available to them by the scientific community. Over 600 genes, markers and chromosomal regions have been assembled so far [3].

As the current status paper of the OGM notes, "the interest (among biomedical researchers) for the compendium of putative human obesity genes continues to be very strong" [3]. The online version of the OGM allows users to query as well as browse through the assembled information [4]. It links the entities contained in the database with the publications in which it is proven or at least suggested to be related to obesity. The text line or lines that serve as evidence for this relation are identified in the entry of each entity. Someone researching a certain gene related to obesity can therefore use the OGM as a shotcut to publications he might be interested in, or gather evidence for an association between a gene and obesity very quickly.

The OGM is collected and maintained by hand. Since it contains data reviewed by experts and linked with evidence supporting its correctness, it is very reliable. This makes the OGM very interesting for text mining projects in the biomedical area: Due to its reliability, the OGM can be used as a gold standard for text mining algorithms.

Also, text mining algorithms could find an interesting field of application in the OGM project. The update of the Obesity Gene Map that was released in March 2005 only contained publications that were "published up to October 2004" [3]. This shows the immense effort it takes to review all the relevant literature by hand. With this project, we want to explore if a share of the work going into finding the relevant publications can be done automatically.

2. Project Statement

This project focuses on extracting associations between genes and the disease obesity from text. We will examine how many of the genes that are included in the OGM can be found by the chosen text mining procedures. The success of this effort will be determined by the percentage of genes in the OGM that the algorithm finds.

We will also determine whether this approach is able to identify genes or relevant publications that are not included in the OGM for various reasons[1]. Even finding one such source that stands up to verification by an expert might be a success. An example publication is [5], which is not included in the OGM 2004 update since it was published after October 2004 (see above).

[1] Publication appeared after last update, publication was rejected, publication has been overlooked so far.

3. Supervised and semi-supervised Learning

“Supervised Machine Learning is a technique to learn (a separation function) from examples of its inputs and outputs” [6]. These examples are called training data. Depending on the precision required, the effort to supply this data can be costly and has to be done by hand. Most text mining applications use this technique to single out useful information

Semi-supervised machine learning is a machine learning technique related to supervised learning. It requires a limited amount of training data to learn a rough first separation function. It then uses instances of data it has classified as new instances of training data to improve that separation function. Algorithms employing semi-supervised learning techniques have been shown to outperform supervised algorithms on specific problems [7]. More information about both supervised and semi-supervised machine learning can be found in [8].

4. Procedure

The project will be undertaken in a cooperation effort with Prof. Brockmann and Dr. Schmitt of the Institute for Animal Sciences, who will provide the needed biological expertise.

The search algorithm we will implement first collects all entries in the PubMed database that show up as results of a keyword search for the word "obesity" or its simplest derivatives. The content of all hits will be analyzed for co-occurrences with the word "obesity". Those groups of terms, henceforth called “contexts”, that show up as significantly overrepresented in this analysis will be included in a feature space.

We take a number of gene names that are associated with obesity and bear unambiguous names from the OGM. An example woud be the “uncoupling protein 1” gene. A negative example is the “hedgehog” gene. Two groups of texts are collected through searches with these gene names: The first are positive texts: They mention one of the picked gene names in a context already included in the feature space, and they are aready included in the OGM entry of that gene. The second group consists of negative texts. These mention one of the genes in the list, but not in context with obesity. The two groups are used as training data for a classification algorithm employing semi-supervised machine learning to make use of its aforementioned advantage.

The found articles are now searched for unlabeled text blocks. These text blocks contain a gene name, but it is not clear whether these gene names are associated with obesity. The classification algorithm trained in step 2 classifies these unlabeled examples into positive and negative contexts. All previously unknown gene names in the paragraphs named positive are added to the list of gene names. All previously unknown contexts that are encountered are added to the feature space in a generalized form. The method for doing this can be found in [9]. The search for labeled and unlabeled contexts is repeated iteratively until no new gene names or contexts are added.

We will write the software required by the project in the Java programming language. We will also make use of a number of existing libraries: The Xerces XML Parser library (more information under [10]) and the text mining libraries of the WBI group.

Literature

[1] http://obesitygene.pbrc.edu/

[2] http://www.pubmed.org/

[3] L. Perusse, T. Rankinen, A. Zuberi, Y. C. Chagnon, S. J. Weisnagel, B. Argyropoulos , B. Walts, E. E. Snyder, C. Bouchard: The human obesity gene map: the 2004 update. Obes Res. 2005 Mar;13 3): 381-490.

[4] E. E. Snyder, B. Walts, L. Perusse, Y. C. Chagnon, S. J. Weisnagel, T. Rankinen, C. Bouchard: The human obesity gene map: the 2003 update. Obes Res. 2004 Mar;12(3):369-439.

[5] Sarruf et al.: Cyclin D3 Promotes Adipogenesis through Activation of Peroxisome Proliferator-Activated Receptor {gamma}. Mol. Cell. Biol..2005; 25: 9985-9995.

[6] S. J. Russell, P. Norvig: Artificial Intelligence A Modern Approach, Second Edition. 2003 Pearson Education Inc., Upper Saddle River, New Jersey.

[7] U. Brefeld, C. Büscher, T. Scheffer: Multi-View Discriminative Sequential Learning. Proc.ECML, 2005

[8] B. Novak: Use of Unlabelled Data in Supervised Machine Learning. Proc. SIKDD, pp. 1-4, Ljubljana, Slovenia, October 2004.

[9] J.-H. Kim, M. Hilario: Learning Information Extraction Rules for Protein Annotation from Unannotated Corpora. CICLing 2005.

[10] http://xml.apache.org/xerces-j/