Novel Nature Inspired Techniques in Medical Data Mining

Miroslav Bursa and Lenka Lhotska
{bursam|lhotska}@fel.cvut.cz
CTU in Prague, Czech republic

Introduction

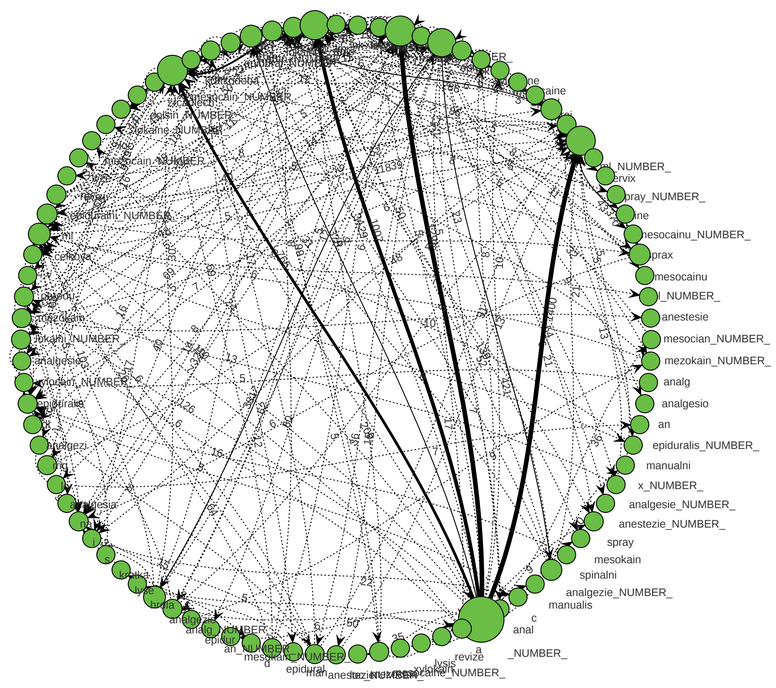
Information mining from textual data becomes a very challenging task when the structure of the text record is loose without any rules. The task becomes even harder when natural language is used and no apriori knowledge is available. The medical environment is very specific: the natural language used in textual description varies with the personality creating the record, however it is restricted by terminology (i.e. medical terms, etc.). Moreover, the typical patient record is filled with typographical errors, duplicates and many (nonstandard) abbreviations. The accuracy for relation extraction in journal text is typically about 60 % [3]. A perfect accuracy in text mining is nearly impossible due to errors and duplications in the source text. Even when linguists are hired to label text for an automated extractor, the inter-linguist disparity is about 30 %. The best results are obtained via an automated processing supervised by a human [5]

Dataset

In this work we have studied, evaluated and proposed different swarm intelligence techniques for mining information from loosely structured medical textual records with no apriori knowledge. We describe the process of mining a large dataset of ∼50,000–120,000 records × 20 attributes in DB tables. Each attribute item contains ∼800–1,500 characters (diagnoses, medications, etc.). The output of this task is a set of ordered/nominal attributes suitable for rule discovery mining.

Attribute transition graph: Overview

The overview of one small (in field length) attribute is visualized in the figure. Only a subsample (about 5 %) of the dataset could be displayed in this paper, as the whole set would render into a uncomprehensible black stain. The vertices (literals) are represented as a green circle, the size reflects the literal frequency. Edges represent transition states between literals; edge stroke shows the transition rate (probability) of the edge.



The vertices are (usually) organized depending on the position in the text (distance from the starting point) as the have the highest potence. Number literal (a wildcard) had the highest potence, as many quantitative measures are contained in the data (age, medication amount, etc.). Therefore it has been fixed to the following literal, spreading into the graph via multiple nodes. This allowed to organize the chart visualization in more logical manner. Time needed to organize such graph was about 5–10 minutes.

Swarm Intelligence and Semiautomatic Processing

The Ant Colony Optimization (ACO) [2] is a metaheuristic approach, inspired by the ability of ants to discover shortest path between nest and food source. The process is guided by deposition of a chemical substance (pheromone). As the ants move, they deposit the pheromone on the ground (amount of the pheromone deposited is proportional to the quality of the food source discovered). Such pheromone is sensed by other ants and the amount of pheromone changes the decision behavior of the ant individual. The ant will more likely follow a path with more pheromone.

Automated layout of transition graph is very comfortable, however the contents of the attribute is so complicated, that a human intervention is inevitable.

A semi-automated (corrected by a human expert) organized transition graph showing the most important relations in one textual attribute. An aid of a human expert has been used in semi-automated approach (see the corresponding figure) where the automated layout has been corrected by the expert. The correction time has been about 20–30 seconds only.

Conclusion and Discussion

The main advantage of the nature inspired concepts lies in automatic finding relevant literals and group of literals that can be adopted by the human analysts and furthermore improved and stated more precisely. The use of induced probabilistic models in such methods increased the speed of loosely structured textual attributes analysis and allowed the human analysts to develop lexical analysis grammar more efficiently in comparison to classical methods. The speedup (from about 5–10 minutes to approx 20–30 seconds) allowed to perform more iterations, increasing the yield of information from data that would be further processed in rule discovery process. However, the expert intervention in minor correction is still inevitable.

References

[1] M. Bursa, L. Lhotska, and M. Macas. Hybridized swarm metaheuristics for evolutionary random forest generation. *Proc. of the 7th Intl Conf. on Hybrid Intell. Systems 2007 (IEEE CSP)*, pages 150–155, 2007.

[2] M. Dorigo and T. Stutzle. *Ant Colony Optimization*. MIT Press, Cambridge, MA, 2004.

[3] D. Freitag and A. K. McCallum. Information extraction with HMMs and shrinkage. *Proceedings of the AAAI Workshop on Machine Learining for Information Extraction*, 1999.

[4] P.-P. Grasse. La reconstruction du nid et les coordinations inter-individuelles chez bellicositermes natalensis et cubitermes sp. la thÃl'orie de la stigmergie: Essai d'interprÃl'tation des termites constructeurs. *Insectes Sociaux*, 6:41–81, 1959.

[5] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of the ICML*, pages 282–289, 2001.

Funding

This research project has been supported by the research program no. MSM 6840770012 "Transdisciplinary Research in the Area of Biomedical Engineering II" of the CTU in Prague, sponsored by the Ministry of Education, Youth and Sports of the Czech Republic. This work has been developed in the BioDat research group `http://bio.felk.cvut.cz`.