Novel Nature Inspired Techniques in Medical Information Retrieval

Miroslav Bursa\textsuperscript{1}, Lenka Lhotská\textsuperscript{1}, Vaclav Chudacek\textsuperscript{1}, Michal Huptych\textsuperscript{1}, Jiri Spilka\textsuperscript{1}, Petr Janku\textsuperscript{2}, and Martin Huser\textsuperscript{2}

\textsuperscript{1} Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University in Prague, Czech Republic
\textsuperscript{2} Obstetrics and Gynaecology Clinic, University Hospital in Brno, Czech Republic

Abstract. 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 $\sim$50,000–120,000 records $\times$ 20 attributes in DB tables, originating from the hospital information system recording over 10 years. This paper concerns only textual attributes with free text input, that means 613,000 text fields in 16 attributes. Each attribute item contains $\sim$800–1,500 characters (diagnoses, medications, etc.). The output of this task is a set of ordered/nominal attributes suitable for rule discovery mining.

Information mining from textual data becomes a very challenging task when the structure of the text record is very loose without any rules. The task becomes even harder when natural language is used and no apriori knowledge is available. The medical environment itself is also 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, medical standards, etc.). Moreover, the typical patient record is filled with typographical errors, duplicates and many (nonstandard) abbreviations.

Nature inspired methods have their origin in real nature processes and play an important role in the domain of artificial intelligence. They offer fast and robust solutions to many problems, although they belong to the branch of approximative methods. The high number of individuals and the decentralized approach to task coordination in the studied species revealed high degree of parallelism, self-organization and fault tolerance.

First, classical approaches such as basic statistic approaches, word (and word sequence) frequency analysis, etc., have been used to simplify the textual data and provide an overview of the data. Finally, an ant-inspired self-organizing approach has been used to automatically provide a simplified dominant structure, presenting structure of the records in the human readable form that can be further utilized in the mining process as it describes the vast majority of the records.

Note that this project is an ongoing process (and research) and new data are irregularly received from the medical facility, justifying the need for robust and fool-proof algorithms.
Keywords: Swarm Intelligence, Ant Colony, Text Mining, Data Mining, Medical Record Processing, Hospital Information System.

1 Introduction

In many industrial, business, healthcare and scientific areas we witness the boom of computers, computational appliances, personalized electronics, high-speed networks, increasing storage capacity and data warehouses. Therefore a huge amount of various data is transferred and stored, often mixed from different sources, containing different data types, unusual coding schemes, and seldom come without any errors (or noise) and omissions.

Even with rapidly increasing computational power of modern computers, the analysis of huge databases becomes very expensive, making the development of novel techniques reasonable. Especially in text processing, the impact of automated methods is crucial. In contrary to classical methods, nature-inspired methods offer many techniques, that can increase speed and robustness of classical methods.

Nature inspired metaheuristics play an important role in the domain of artificial intelligence, offering fast and robust solutions in many fields (graph algorithms, feature selection, optimization, clustering, etc). Stochastic nature inspired metaheuristics have interesting properties that make them suitable to be used in data mining, data clustering and other application areas.

Plenty of nature inspired methods are studied and developed in present. One category is represented by methods, that are inspired by the behavior of ant colonies. These methods have been applied to many problems (often NP-hard). Review can be seen in [4] and [1]. We concentrate on the state-of-the-art nature methods inspired by the social behavior of insect communities, by the swarm intelligence, brain processes and other real nature processes.

Ant colonies inspired many researches to develop a new branch of stochastic algorithms: ant colony inspired algorithms. Based on the ant metaphor, algorithms for both static and dynamic combinatorial optimization, continuous optimization and clustering have been proposed. They show many properties similar to the natural ant colonies, however, their advantage lies in incorporating the mechanisms, that allowed the whole colonies to effectively survive during the evolutionary process.

Cemetery formation and brood sorting are two prominent examples of insects’ collective behavior. However, other types of ant behavior have been observed, for example predator-prey interaction, prey hunting, etc. The most important are mentioned below.

The accuracy for relation extraction in journal text is typically about 60% [5]. 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 [7].