Trimming the complexity of Ranking by Pairwise Comparison

Samuel Hiard (ULg)

February 11, 2013

In computer science research, and more specifically in bioinformatics, the size of databases never stops to increase. This can be an issue when trying to answer questions that imply algorithms in nonlinear polynomial time with regards to the number of objects in the database, the number of attributes or the number of associated labels per objects. This is the case of the Ranking by Pairwise Comparison (RPC) algorithm. This algorithm builds a model which is able to predict the label preference for a given object, but the computation needs to be performed in an order of $N*(N-1)/2$ in terms of the number N of labels. Indeed, a pairwise comparator model is needed for each possible pair of labels. Our hypothesis is that a significant part of the set of comparators often contains redundancy and/or noise, so that trimming the set could be beneficiary. We implemented several methods, starting from the simplest one, which merely chooses a set of T comparators ($T < N*(N-1)/2$) at random, to a more complex approach based on partially randomized greedy search.

This thesis will provide a detailed overview of the context we are working in, provide the reader with required background, describe existing preference learning algorithms including RPC, investigate on possible trimming methods and their accuracy, then will conclude on the relevance and robustness of the trimming approximation.

After implementing and executing the procedure, we could see that using between $N/2$ and $2N$ comparators was sufficient to keep up with the original RPC algorithm, as long as a smart trimming method is used, and sometimes even outperforms it on noisy datasets. Also, comparing the use of base models in regression mode vs. classification mode showed that models built in regression mode may be more robust when using the original RPC. We thus empirically show that, in the particular case of RPC, reducing the complexity of the method gives similar or better results, which means that problems that could not be addressed by this algorithm, or at least not in an acceptable period of time, now can be. We also found that the regression mode yields RPC to be often more robust regarding its base learner parameters, meaning that the quest of optimality, which can also be time-consuming, is less difficult.

Yet research on this topic is not over, and we could think of different means to further improve the RPC algorithm or investigate other innovative approaches, which will be discussed in the future work section. Also, the trimming method is not limited to RPC and could be applied to other algorithms which aggregate information provided

by a set of models, e.g. the whole multitude of ensemble models used in machine learning.