Review of doctoral thesis "Machine Learning of Analysis by Reduction"

Author: RNDr. Petr Hoffmann

Reviewer: RNDr. Daniel Průša, Ph.D.
Czech Technical University in Prague, Faculty of Electrical Engineering
Karlovo náměstí 13, 121 35 Praha 2

The thesis deals with the machine learning of analysis by reduction based on training samples. The considered model performing analysis is the restarting automaton. This topic is very well motivated, namely by the challenge to automatically analyze sentences of a natural language, having regard to linguistic databases which provide many samples of reductions in the form of dependency trees.
The author identified three main research goals:

- To study complexity of the problem from the theoretical point of view.
- To propose a learning algorithm for a suitable variant of the restarting automaton.
- To propose a testing methodology and apply it to the learned automata.

The achieved results are described in a text consisting of 140 pages. The introduction presents basic notions, gives a very good overview of known results on the topic of grammatical inference and defines studied problems. The author successfully managed to define the problem of learning from samples of analysis by reduction formally. The stated three goals are fulfilled in chapters five, six and seven. Theoretical and application results can be distinguished.

The theoretical ones are a very strong part of the work. They include a number of nontrivial theorems that significantly disseminate knowledge on restarting automata, especially with regard to the possibility of learning. The author shows that he is familiar with a large number of proof techniques as well as able to come with his own original approaches. The proofs are in most cases comprehensible, with a satisfactory level of detail. I only find it more difficult to understand two very long proofs (Theorems 5.2.8 and 6.4.7) occupying several pages. A revision followed by a simplification would be very helpful here, possibly in a combination with improving the structure.

The author also proves his analytical skills in the application part. He takes into account the obtained theoretical limits and proposes a practical method for learning analysis by reduction. Moreover, a suitable methodology is proposed for testing the trained automata. The presented empirical evaluation shows that the method has a good potential. It is my opinion here that scope of the proposal, implementation, testing and the achieved correct rate of the method meet the thesis demands. On the other hand, it should be noted that relatively simple languages were used and samples of reductions were generated with the help of a restarting automaton. This means that a natural source of training data, represented e.g. by the already mentioned dependency trees, has not been used. It is thus not possible to answer the question whether the method would perform well if an automaton is learned from a large amount of samples. It would certainly be interesting if the author could apply the method at least to a small subset of natural sentences.

The next evaluation focuses on clarity and structure of the text. Overall, the thesis is written very carefully. I have not encountered nearly any grammatical errors neither typos, individual parts are comprehensive and synoptical. Nevertheless, from a higher point of view I would recommend to shorten explanation of known notions and to improve thematic grouping of results into chapters.

A large space in the introduction is devoted to facts known from elementary courses of theoretical computer science (notions like finite-state automaton, regular languages, graph). Similarly, principles of basic techniques like reduction in polynomial time are being clarified when used in proofs. This kind of information is unnecessary. The thesis is not intended for a reader who is not aware of such facts. It can not serve as a textbook, which complements this knowledge. It only makes sense to

specify notions and notations that could be ambiguous, because of there are more conventions used in the literature. A different case are more specialized notions like the grammatical inference or restarting automaton. Their introduction is desirable.

The tiles of the sixth and seventh chapter give the impression that they are devoted primarily to the algorithm proposal and its verification. In reality, however, both chapters contain many theoretical results (a taxonomy of k-reversible restarting automata, linear grammars and languages). Related theorems have impact on the practical realization of the learning process, on the other hand, they do not participate directly in the algorithm. It would be better to move them into a standalone chapter.

The last remark addresses undifferentiated significance of proved results. All of them are presented as theorems. It would be preferable to distinguish auxiliary statements and present them as lemmas. This applies for example to many claims of properties of words and subwords in the fifth chapter.

In conclusion, I can say that the author achieved good results and clearly fulfilled the defined objectives. He managed to publish the results and presented them at international conferences. He undoubtedly proved his ability of independent creative work. In my opinion, the thesis satisfies all the conditions for gaining the Ph.D. degree, therefore I **recommend accepting** it.

Suplementary comments:

- Theorem 5.2.2 – we can at least say that the considered problem belongs in the polynomial hierarchy to the class $\sum_2^P$.
- Definition 6.2.1 contains one of few inaccuracies or typos – the meaning of word v_0 is unclear. How is it related to the word v? If v_0 has been introduced in the previous text, it should be clearly stated where it was.
- Section 7.3.1 – the comparison of experimental results to the method LARS on fixed languages says how many reductions were needed to learn correct automata, however, it is not explained if this amount is always sufficient or whether just suitable training sets of the mentioned sizes were found.
- Section 7.3.2, Table 7.1 – it would be good to add one more row (for reversibility 4) in the case of 80 samples, since the best result is in the last row (reversibility 3). As it is, it is not clear whether the trend of improving (value *prod*) continues or not.

Questions:

1. Is not it a limitation from a practical point of view that a *nondeterministic* reducing automaton has been chosen as the studied model? How much is it time-consuming to simulate the learned automata and how does time complexity of the simulation grow in the number of metainstructions?
2. In addition to natural language, do you see any other practical domains which we can consider as suitable for the proposed learning method? Such domains should have natural sources of training data (samples of analysis by reduction), similarly like linguistic databases.

Prague, June 10, 2013

Daniel Průša

RNDr. Daniel Průša, Ph.D.

Review of doctoral thesis "Machine Learning of Analysis by Reduction"

Author: RNDr. Petr Hoffmann

Reviewer: RNDr. Daniel Průša, Ph.D.
Czech Technical University in Prague, Faculty of Electrical Engineering
Karlovo náměstí 13, 121 35 Praha 2

The thesis deals with the machine learning of analysis by reduction based on training samples. The considered model performing analysis is the restarting automaton. This topic is very well motivated, namely by the challenge to automatically analyze sentences of a natural language, having regard to linguistic databases which provide many samples of reductions in the form of dependency trees.
The author identified three main research goals:

- To study complexity of the problem from the theoretical point of view.
- To propose a learning algorithm for a suitable variant of the restarting automaton.
- To propose a testing methodology and apply it to the learned automata.

The achieved results are described in a text consisting of 140 pages. The introduction presents basic notions, gives a very good overview of known results on the topic of grammatical inference and defines studied problems. The author successfully managed to define the problem of learning from samples of analysis by reduction formally. The stated three goals are fulfilled in chapters five, six and seven. Theoretical and application results can be distinguished.

The theoretical ones are a very strong part of the work. They include a number of nontrivial theorems that significantly disseminate knowledge on restarting automata, especially with regard to the possibility of learning. The author shows that he is familiar with a large number of proof techniques as well as able to come with his own original approaches. The proofs are in most cases comprehensible, with a satisfactory level of detail. I only find it more difficult to understand two very long proofs (Theorems 5.2.8 and 6.4.7) occupying several pages. A revision followed by a simplification would be very helpful here, possibly in a combination with improving the structure.

The author also proves his analytical skills in the application part. He takes into account the obtained theoretical limits and proposes a practical method for learning analysis by reduction. Moreover, a suitable methodology is proposed for testing the trained automata. The presented empirical evaluation shows that the method has a good potential. It is my opinion here that scope of the proposal, implementation, testing and the achieved correct rate of the method meet the thesis demands. On the other hand, it should be noted that relatively simple languages were used and samples of reductions were generated with the help of a restarting automaton. This means that a natural source of training data, represented e.g. by the already mentioned dependency trees, has not been used. It is thus not possible to answer the question whether the method would perform well if an automaton is learned from a large amount of samples. It would certainly be interesting if the author could apply the method at least to a small subset of natural sentences.

The next evaluation focuses on clarity and structure of the text. Overall, the thesis is written very carefully. I have not encountered nearly any grammatical errors neither typos, individual parts are comprehensive and synoptical. Nevertheless, from a higher point of view I would recommend to shorten explanation of known notions and to improve thematic grouping of results into chapters.

A large space in the introduction is devoted to facts known from elementary courses of theoretical computer science (notions like finite-state automaton, regular languages, graph). Similarly, principles of basic techniques like reduction in polynomial time are being clarified when used in proofs. This kind of information is unnecessary. The thesis is not intended for a reader who is not aware of such facts. It can not serve as a textbook, which complements this knowledge. It only makes sense to

specify notions and notations that could be ambiguous, because of there are more conventions used in the literature. A different case are more specialized notions like the grammatical inference or restarting automaton. Their introduction is desirable.

The tiles of the sixth and seventh chapter give the impression that they are devoted primarily to the algorithm proposal and its verification. In reality, however, both chapters contain many theoretical results (a taxonomy of k-reversible restarting automata, linear grammars and languages). Related theorems have impact on the practical realization of the learning process, on the other hand, they do not participate directly in the algorithm. It would be better to move them into a standalone chapter.

The last remark addresses undifferentiated significance of proved results. All of them are presented as theorems. It would be preferable to distinguish auxiliary statements and present them as lemmas. This applies for example to many claims of properties of words and subwords in the fifth chapter.

In conclusion, I can say that the author achieved good results and clearly fulfilled the defined objectives. He managed to publish the results and presented them at international conferences. He undoubtedly proved his ability of independent creative work. In my opinion, the thesis satisfies all the conditions for gaining the Ph.D. degree, therefore I **recommend accepting** it.

Suplementary comments:

- Theorem 5.2.2 – we can at least say that the considered problem belongs in the polynomial hierarchy to the class $\sum_2^P$.
- Definition 6.2.1 contains one of few inaccuracies or typos – the meaning of word v_0 is unclear. How is it related to the word v? If v_0 has been introduced in the previous text, it should be clearly stated where it was.
- Section 7.3.1 – the comparison of experimental results to the method LARS on fixed languages says how many reductions were needed to learn correct automata, however, it is not explained if this amount is always sufficient or whether just suitable training sets of the mentioned sizes were found.
- Section 7.3.2, Table 7.1 – it would be good to add one more row (for reversibility 4) in the case of 80 samples, since the best result is in the last row (reversibility 3). As it is, it is not clear whether the trend of improving (value *prod*) continues or not.

Questions:

1. Is not it a limitation from a practical point of view that a *nondeterministic* reducing automaton has been chosen as the studied model? How much is it time-consuming to simulate the learned automata and how does time complexity of the simulation grow in the number of metainstructions?
2. In addition to natural language, do you see any other practical domains which we can consider as suitable for the proposed learning method? Such domains should have natural sources of training data (samples of analysis by reduction), similarly like linguistic databases.

Prague, June 10, 2013

Daniel Průša

RNDr. Daniel Průša, Ph.D.