International Journal of Engineering &Technology, 7 (3.12) (2018) 1096 -1097

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Context Free Grammar Identification from Positive Samples

[1]K.Senthil Kumar, [2]D.Malathi

[1, 2]Department of Computer Science and Engineering, SRMIST, Chennai, India
**Corresponding Author E-mail: [1]Senthilkumar.k@ktr.srmuniv.ac.in,*
[2]Malathi.d@ktr.srmuniv.ac.in

Abstract

In grammatical inference one aims to find underlying grammar or automaton which explains the target language in some way. Context free grammar which represents type 2 grammar in Chomsky hierarchy has many applications in Formal Language Theory, pattern recognition, Speech recognition, Machine learning , Compiler design and Genetic engineering etc. Identification of unknown Context Free grammar of the target language from positive examples is an extensive area in Grammatical Inference/ Grammar induction. In this paper we propose a novel method which finds the equivalent Chomsky Normal form.

Keywords: *CFG, Identification in the Limit, Chomsky Normal form, Parse tree, Grammatical Inference, Pumping Lemma*

1. Introduction

Grammar learning is an active research area in Machine learning.. In grammar learning the main aim is to determine the target language from a given set of examples (which may be or may not be in the target language). With the seminal paper by Gold [4] has made a revolution in the grammar learning theory. In the Chomsky [1] model one has various language models and their respective grammars. In regular language we have either to identify the under lying finite state machine or equivalent regular grammar. In Minimal DFA identification notable contributions are from [4],[5],[6],[7],[8].In this paper we propose context free grammar learning algorithm which uses Chomsky Normal form.
We define our problem as follows
Given: A set of positive examples of an unknown target language
That is S^{+}={s1,s2,…sn} of an unknown target language
Goal: To identify an equivalent CFG G=(NT,,Ter,P,S) in Chomsky Normal Form.
We organize our paper as follows section 2 deals some preliminary notations and definitions, section 3 deals related work, section 4 deals our proposed algorithm and finally conclusion.

2. Preliminary Notations

We use the following notations throughout this paper.
One can find [10] for more details.
Given input alphabet $\sum$ a language is defined as L⊆$\sum$*.A grammar is represented by G=(NT,Ter,P,S) where NT represents a finite nonempty set of variables, Ter is a finite set of terminals(Note: NT∩Ter=Ø), P is a collection of production rules and S is a special start symbol. If the production rules are of form NT→(NTUTer)*.
Then it is known as context free grammar. A Chomsky normal form is a form of Context free grammar where the productions are of the form L→MN or L→c where L,M,N∈NT and c∈Ter.

2.1 Parse tree

A parse tree is a tree with root represents start symbol and interior nodes represent non terminals and leaves represent terminals. For every derivation one can construct a parse tree.

3. Related Work

In the grammar learning one can find 3 important class of learning models, namely

1. Gold's Identification in the limit model:

According to Gold [4] if the machine is given a complete set of examples with every string of the language is appearing at least once then at one point of time the machine will correctly identify the target language. If the machine is given with only positive examples then identification is impossible for any super finite class of languages.

2. Minimally Adequate Teacher (MAT)

This is the model proposed by Angluin[5].In this model each time a query is asked to the Oracle(teacher) by the inference machine(learner).The learner will always get some answer whether the string is in the language or not. With this model Angluin[5]. was able to prove regular languages can be identified with finite set of queries in polynomial time.

3. Probably Approximately Correct(PAC) Model

This is the model proposed by Valiant [6] which uses some probabilistic model of sample data.
In this the learner will have an hypothesis for each input. Samples are drawn from unknown distribution.

Then given ∈, δ>0 Then probability of the difference between actual hypothesis and the predicted which is less than ∈ will be 1-δ.
It is a well known fact in the learning theory is identifying Context Free Grammars from polynomial time with positive examples is NPC. C. la Higuera[2] has presented a beautiful survey on grammatical inference. Tai-Hung Chen,Chun-Han Tseng,Chia-Ping Chen [11] used Minimum Description Length concept to learn Context Free Grammars automatically. They define a cost function which uses two components namely No of bits required to represent the proposed grammar and the second one which represents the no of bits required for parse tree of the corpus using the above grammar John C. Kieffer and En-hui Yang [12] propose a context free grammar which generates a single string, and then they produce a simpler grammar to generate the same string. Ney [13] proposes a context free grammar for speech recognition using Dynamic programming approach.
As in [14] proposes a context free grammar inference algorithm which uses only positive data. This algorithm combines information theory with some heuristic method namely substitutability and frequency technique. Sakakibara[15] presents a context free grammar learning algorithm from unlabelled parse tree. In this paper he uses two types of queries namely structural equivalence queries and structural membership queries. This algorithm is based on Minimally Adequate Teacher Model proposed by Angluin[5].In this paper we use the concept of pumping lemma to find variables of the unknown target grammar.

4. Proposed Algorithm

In this paper we use the concept of pumping lemma to find variables of the unknown target grammar. Pumping lemma says there is a variable in the parse tree (at a deeper level) will repeat again.
Also a well known result in formal language theory which says that every string of length n will have a derivation in 2*n-1 steps if the grammar is in Chomsky Normal Form.

4.1 Proposition

If a string of length n is given then there exists a path in the tree whose height is atmost n.

Proof:

If a string is length of one we have the rule like S→a hence we have a tree with height 1
We assume a string of length n will have a tree of height n.
By mathematical induction if we have a string of length n+1 then at level n there should be a variable which will get us another derivation. Hence the height of the tree is n+1.
We present our algorithm as

Algorithm to find Chomsky Normal form of Unknown Target Language from given positive samples

Input: Given a set of Positive samples of Unknown target language
Output: The corresponding equivalent context free grammar in Chomsky Normal Form

Algorithm_CNF(Char [][n],int n)
{
for each terminal do $X_i \rightarrow a_i$ where $X_i \in NT, a_i \in Ter$
for root do
$S \rightarrow X_i W$ where Xi always produce a terminal.
For each interion unknown node apply pumping lemma to find unknown node.
Rearrage variables;
}

Example:

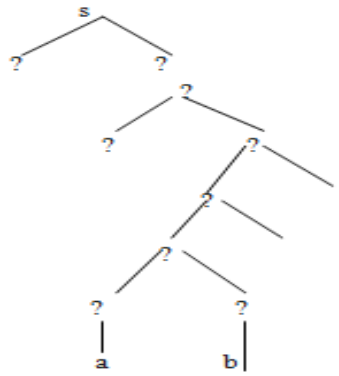

Fig. 1

Suppose we are given set of strings are {aaabbb,aaaabbbb,..}(this set is finite)
In this problem we have two terminals namely a and b
So initial grammar is{ X→a,Y→b}
Obviously S→XW because when we process the string obviously first symbol is a hence
We have { S→XW,X→a,Y→b} and proceeding further we will get all missing nodes in fig 1.
The tentative parse tree is shown above.
Using Pumping one can find that one of the interior node will repeat once again.Also if length of a sub tree is k then k-1th level there should be a variable corresponding to that terminal. Using this idea one can find Chomsky Normal form which will generate equivalent context free grammar of an unknown language.

5. Conclusion

In this paper we propose a novel method of generating unknown target language by a Chomsky normal form.

References

[1] N. Chomsky, "On certain formal properties of grammars," Information and Control, 2(2) (1959), pp.137–167. doi:10.1016/S0019-9958(59)90362-6.
[2] C. la Higuera, "A bibliographical study of grammatical inference", Pattern 270 Recognition, 38(9) (2005), pp.1332–1348.
[3] Y. Sakakibara, "Recent advances of grammatical inference,"Theoretical Computer Science ,185 (1997), pp.15–45.
[4] E. M. Gold,"Language identification in the limit," Information and Control ,10 (1967) ,pp.447–474.
[5] D. Angluin, "Queries and concept learning," Machine Learning 2 ,(1988) ,pp.319– 342
[6] L. Valiant, "A theory of the learnable, "Communications of the ACM, 27 (1984) ,pp.1134–1142.
[7] D. Angluin, "Inductive inference of formal languages from positive data," Information and Control 45(2) ,(1980) ,pp.117–135.
[8] D. Angluin, "Inference of reversible languages," Journal of the ACM 29(3) ,(1982),pp. 741–765.
[9] E. M. Gold, "Complexity of automaton identification from given data," Information and Control 37(3), (1978) ,pp.302–320.
[10] J. E. Hopcroft, J. D. Ullman, R. Motwani, "Introduction to Automata Theory, Languages, and Computation," 3rd Edition, Pearson, 2006.
[11] Tai-Hung Chen,Chun-Han Tseng,Chia-Ping Chen,"Automatic learning of Context free grammars,"ROCLING 2006
[12] John C. Kieffer and En-hui Yang, "Design of context-free grammars for lossless data compression," Proceedings of the 1998 IEEE Information Theory Workshop, pp. 84- 85.
[13] H. Ney, "Dynamic Programming Speech Recognition Using a Context-Free Grammar," Proceedings of ICASSP'87, pp. 69-72.
[14] Alexander Clark,"Learning deterministic context free grammars: The Omphalos competition,"Mach Learn (2007), pp.66:93–110 DOI 10.1007/s10994-006-9592-9.
[15] Yasubumi Sakakibara,"Learning of Context free grammars from Structural data in polynomial time," Theoretical Computer Science ,76 (1990),pp. 223-242