POLITECHNIKA WROCŁAWSKA

WYDZIAŁ ELEKTRONIKI

KIERUNEK: Informatyka

SPECJALNOŚĆ: Inżynieria Systemów Informatycznych

PRACA DYPLOMOWA

MAGISTERSKA

Nowe mechanizmy uczenia w modelu sGCS

New learning algorithms in sGCS model

autor : Michał Stempkowski

Opiekun pracy:

dr hab. inż. Olgierd Unold,

prof. nadzw. PWr

OCENA PRACY:

WROCŁAW 2016

[Wstęp 5](#_Toc443726186)

[Cel pracy 7](#_Toc443726187)

[Pojęcia podstawowe 8](#_Toc443726188)

[Algorytm CYK 16](#_Toc443726189)

[GCS 19](#_Toc443726190)

[Główna pętla algorytmu 19](#_Toc443726191)

[Indukcja gramatyki 20](#_Toc443726192)

[Parsowanie zdań 21](#_Toc443726193)

[Ścisk 22](#_Toc443726194)

[Operatory pokrycia 22](#_Toc443726195)

[Operator pokrycia terminalnego 23](#_Toc443726196)

[Operator pokrycia uniwersalnego 23](#_Toc443726197)

[Operator pokrycia startowego 24](#_Toc443726198)

[Operator pokrycia agresywnego 24](#_Toc443726199)

[Operator pokrycia pełnego 24](#_Toc443726200)

[Algorytm genetyczny 25](#_Toc443726201)

[Metody selekcji 25](#_Toc443726202)

[Inwersja 26](#_Toc443726203)

[Mutacja 26](#_Toc443726204)

[Krzyżowanie 27](#_Toc443726205)

[Korekcja gramatyki 27](#_Toc443726206)

[Usuwanie produkcji redundantnych 27](#_Toc443726207)

[Usuwanie produkcji nieproduktywnych 28](#_Toc443726208)

[Usuwanie produkcji nieosiągalnych 28](#_Toc443726209)

[sGCS 30](#_Toc443726210)

[Algorytm CYK+ 30](#_Toc443726211)

[Traceback 32](#_Toc443726212)

[Normalizacja 32](#_Toc443726213)

[Przystosowanie w sGCS 33](#_Toc443726214)

[Estymacja prawdopodobieństw 33](#_Toc443726215)

[neg-sGCS 33](#_Toc443726216)

[Implementacja 35](#_Toc443726217)

[Użytkowanie aplikacji 36](#_Toc443726218)

[Wykorzystanie gui\_manager.py 37](#_Toc443726219)

[Wykorzystanie console\_fetcher.py 48](#_Toc443726220)

[Artefakty 49](#_Toc443726221)

[Biblioteka 54](#_Toc443726222)

[Warstwowa budowa biblioteki 56](#_Toc443726223)

[Moduły pomocnicze 57](#_Toc443726224)

[GUI 59](#_Toc443726225)

[Warstwa wykonawców 92](#_Toc443726226)

[Właściwy algorytm 97](#_Toc443726227)

[Właściwy algorytm – część indukcyjna 121](#_Toc443726228)

[Warstwa rdzenia 138](#_Toc443726229)

[Warstwa danych 142](#_Toc443726230)

[Parametry algorytmu 145](#_Toc443726231)

[Wyniki badań 147](#_Toc443726232)

[Standardowa losowa populacja 147](#_Toc443726233)

[GCS 147](#_Toc443726234)

[sGCS 153](#_Toc443726235)

[neg-sGCS 158](#_Toc443726236)

[Reguły charakterystyczne 164](#_Toc443726237)

[GCS 165](#_Toc443726238)

[sGCS 168](#_Toc443726239)

[neg-sGCS 172](#_Toc443726240)

[Podsumowanie 176](#_Toc443726241)

[Bibliografia 178](#_Toc443726242)

Wstęp

Od zarania ludzkości gromadzenie i przetwarzanie informacji jest ważnym zadaniem. Z kolejnymi wiekami człowiek gromadził tych informacji coraz więcej i wykorzystuje coraz to bardziej wymyślne i automatyczne metody analizy tych informacji. Wraz z rozwojem komputerów stały się one istotną częścią mechanizmu poznawczego człowieka.

Jednym z ważniejszych problemów, jakie napotykamy przy próbie wykorzystania komputerów do wspomagania zdobywania wiedzy, jest ich sposób rozumowania i pozyskiwania wiedzy. Człowiek analizując problemy myśli w sposób wysoce abstrakcyjny, zauważa zależności, formułuje reguły pomiędzy obiektami, tworzy sieć połączeń „jeżeli A, to B”. Komputery posiadają coraz większą moc obliczeniową, jednakże nadal są dosyć ograniczone w kwestii samodzielnego zdobywania wiedzy.

Uczenie maszynowe (ang. Machine Learning) to jedna z gałęzi informatyki, która skupia się na rozszerzeniu umiejętności komputerów w rozpoznawaniu wzorców oraz samodzielnym uczeniu się. Zahacza one o inne dziedziny nauki, takie jak psychologia, czy statystyka. Obejmuje ono problematykę pisania programów, które potrafią pozyskiwać nową wiedzę i poprawiać istniejącą na podstawie zyskiwanych informacji.

Nie da się zanegować faktu, że uczenie maszyn jest zagadnieniem ważnym i ma wiele zastosowań. Umiejętność napisania aplikacji zdolnych do adaptowania się, znajdowania wzorców i symulujących proces ludzkiego wnioskowania jest ważna, zaś aplikacje takie znajdują wiele zastosowań we współczesnym świecie:

* Przetwarzanie języka naturalnego;
* Rozpoznawanie mowy – tworzenie aplikacji komputerowych wykorzystujących mowę ludzką jako interfejs sterowania;
* Systemy ekspertowe – systemy wspomagające proces podejmowania decyzji, wspomagające proces analizy złożonych problemów ekonomicznych, medycznych i nie tylko;
* Data mining – analiza i odnajdowanie zależności w dużych strukturach danych;
* Bioinformatyka – rozpoznawanie powtarzających się struktur łańcuchów RNA, czy DNA, rozpoznawanie oraz analiza zachowania chorób;
* Rozpoznawanie obrazów – wyszukiwanie wzorców z obrazów. Zamiana analogowej reprezentacji na cyfrową, odnajdywanie abstrakcyjnych kształtów;
* Robotyka – tworzenie automatów zdolnych do zdobywania wiedzy na temat środowiska, w którym zostały umieszczone, zdolność do autonomicznego funkcjonowania i wykonywania czynności;
* Wnioskowanie logiczne.

W poniższej pracy skupimy się na jednym z mechanizmów uczenia maszynowego, a konkretnie na systemie GCS (Grammar Classifier System) . Jest to system uczenia nadzorowanego, skupiający się na problemie klasyfikacji. Został wprowadzony w publikacji „Ewolucyjne wnioskowanie gramatyczne” przez Unolda [1]. Wynikiem działania algorytmu jest zbiór reguł gramatyki, czyli zestaw reguł opisujący dane zjawisko.

Cel pracy

Celem poniższej pracy jest zbadanie konkretnych aspektów funkcjonowania algorytmów GCS oraz sGCS [2], ich wpływu na cały proces uczenia oraz zaproponowanie nowych mechanizmów systemu sGCS. Skupi się ona przede wszystkim na zbadaniu wpływu reguł początkowych na proces uczenia. Zbada również wpływ zdań negatywnych na proces uczenia sGCS i zaproponuje modyfikacje niezbędne do uruchomienia algorytmu sGCS w trybie uczenia z wykorzystaniem zdań negatywnych. Badania te powinny dać pole do rozwoju nowych mechanizmów uczenia systemu sGCS.

Praca miałą zajmować się jeszcze zbadaniem dwóch zagadnień – zbadać możliwości wykorzystania w algorytmie więcej niż jednej populacji oraz genetycznego algorytmu wyspowego [3]. Oprócz tego na wstępnym etapie prac podjęto próbę zrównoleglenia algorytmu na karcie graficznej przy wykorzystaniu technologii CUDA [4]. Niestety oba te pomysły musiały być zarzucone, gdyż zabrakło czasu na ich dalszy rozwój.

Rozdział „Pojęcia podstawowe” dostarczy niezbędną teoretyczną widzę na temat gramatyk oraz funkcjonowania samego algorytmu.

Sam algorytm ze wszystkimi wariantami zostanie omówiony w rozdziałach „Algorytm CYK”, „GCS” oraz „sGCS”.

Rozdział „neg-sGCS” jest poświęcony modyfikacjom, jakim należy poddać algorytm sGCS, aby uczył się też z wykorzystaniem zdań negatywnych.

Kolejny rozdział („Implementacja”) omawia szczegółowo budowę oraz użytkowanie dostarczonej wraz z pracą biblioteki sgcs.

W rozdziale „Wyniki badań” znajdujmą się wyniki badań wpływu zdań negatywnych oraz regułcharakterystycznych na proces uczenia.

Ostatni rozdział stanowi podsumowanie pracy.

Pojęcia podstawowe

Poniżej podano definicje pojęć, które występują w tej pracy. Żródła te zebrano z wielu źródeł oraz poprzednich prac [1] [2] [5] [6] [7] [8].

Definicja (Relacja)

Relacją na zbiorze nazywamy dowolny podzbiór iloczynu kartezjańskiego . Jeżeli jest relacją na zbiorze Z, to zależność możemy zapisać w postaci .

Przykład

Przykładem relacji jest relacja mniejszości na zbiorze liczb naturalnych.

Definicja (Zwrotność)

Relację na zbiorze nazywamy zwrotnią, jeżeli .

Przykład

Relacja jest zwrotna, gdyż .

Definicja (Symetryczność)

Relację na zbiorze nazywamy przechodnią, jeżeli .

Definicja (Przechodniość)

Relację na zbiorze nazywamy przechodnią, jeżeli .

Przykład

Relacja jest przechodnia, gdyż .

Definicja (Słaba antysymetryczność)

Relację na zbiorze nazywamy słabo antysymetryczną, jeżeli .

Przykład

Relacja jest słabo antysymetryczna, gdyż .

Definicja (Częściowy porządek)

Częściowym porządkiem w zbiorze nazywamy taką relację (gdzie ), która jest zwrotnia, przechodnia i słabo antysymetryczna.

Przykład

Relacja jest częściowym porządkiem w zbiorze .

Definicja (Porządek liniowy)

Porządek jest liniowy w zbiorze , jeżeli .

Przykład

Porządek jest liniowy, każde 2 elementy są porównywalne.

Definicja (Zbiór częściowo uporządkowany)

Zbiór częściowo uporządkowany to zbiór z relacją częściowego porządku . Zbiór częściowo uporządkowany oznaczamy .

Definicja (Ograniczenie górne)

Niech będzie częściowym porządkiem i niech . Element nazywamy ograniczeniem górnym zbioru , jeżeli . Dualnie definiujemy ograniczenie dolne.

Przykład

Dla częściowego porządku 0 jest ograniczeniem dolnym a 1 ograniczeniem górnym.

Definicja (Kres górny)

Kresem górnym zbioru X nazywamy najmniejszy element zbioru . Kres górny oznaczamy . Dualnie definiujemy kres górny (i oznaczamy go ).

Przykład

Dla częściowego porządku 0 jest kresem dolnym i nie istnieje kres górny (zawsze jest jakaś inna liczba bliższa jedynki i ją nie będącą).

Definicja (Zbiór skierowany)

Niech będzie zbiorem częściowo uporządkowanym. Wtedy zbiór jest zbiorem skierowanym, jeżeli.

Definicja (Porządek zupełny)

Zbiór jest porządkiem zupełnym, jeżeli P ma element najmniejszy oraz każdy skierowany podzbiór X zbioru P ma kres górny.

Definicja (Porządek regularny)

Zbiór jest porządkiem regularnym, jeżeli w każdym niepustym zbiorze .

Definicja (Dobry porządek)

Porządek nazywamy dobrym, jeżeli jest liniowy i regularny.

Definicja (Alfabet)

Alfabetem nazywamy skończony, niepusty zbiór.

Definicja (Symbol)

Elementy należące do alfabetu nazywamy symbolami.

Definicja (Słowo, słowo nad alfabetem)

Słowem (wyrazem) nad alfabetem nazywamy każdy skończony ciąg symboli taki, że .

Definicja (Słowo puste)

Symbolem oznaczamy słowo puste, czyli słowo niezawierające żadnego symbolu.

Definicja (Długość słowa)

Długość słowa oznaczamy, jako . Nazywamy w ten sposób liczbę symboli składających się na dane słowo.

Definicja (Język, język nad alfabetem)

Językiem nad alfabetem nazywamy dowolny zbiór słów nad alfabetem . Będziemy oznaczać go, jako .

Definicja (Język nad dowolnym alfabetem)

Zbiór pusty jest językiem nad dowolnym alfabetem.

Definicja (Gramatyka formalna)

Gramatyką formalną nazywamy czwórkę , w której:

– skończony zbiór symboli pomocniczych (zwanych nieterminalnymi).

– skończony zbiór symboli końcowych (zwanych terminalnymi).

– relacja skończona zwana zbiorem produkcji (wzór).

– wyróżniony symbol pomocniczy, zwany symbolem początkowym gramatyki.

Definicja (Bezpośrednia wyprowadzalność)

Mając gramatykę , jeżeli , wówczas pisząc stwierdzamy, że można wyprowadzić z .

Definicja (Wyprowadzalność)

Wyprowadzalność możemy zdefiniować, jako relację bezpośredniej wyprowadzalności z dodatkową właściwością przechodności. Wyprowadzalność z oznaczamy, jako . Czyli , gdy istnieje skończony ciąg wyprowadzeń pośrednich taki, że .

Definicja (Osiągalność)

Mając gramatykę , nazywamy osiągalnym, jeżeli .

Definicja (Produktywność)

Mając gramatykę , nazywamy produktywnym, jeżeli .

Definicja (Język generowany przez gramatykę)

Jezyk L(G) generowany przez gramatykę G jest ciągiem symboli takim, że .

Definicja (Równoważność gramatyk)

Dwie gramatyki są równoważne, jeżeli generują taki sam język, czyli .

Nakładając dodatkowe ograniczenia na gramatykę możemy zdefiniować 4 typy gramatyki. Typy te nazywamy hierarchią Chomsky'ego.

Definicja (Gramatyka kombinatoryczna)

nazywamy gramatyką kombinatoryczną. Nie posiada ona żadnych ograniczeń. Języki generowane przez gramatykę nazywamy rekurencyjnie przeliczalnymi.

Definicja (Gramatyka kontekstowa)

Gramatyka to taka gramatykę, że jej reguły spełniają następującą zależność: , gdzie Nazywamy ją gramatyką typu 1 (kontekstową). Język L(G1) nazywamy językiem kontekstowym.

Definicja (Gramatyka bezkontekstowa)

Gramatyką to taka gramatyka, że jej reguły spełniają zależność: , gdzie A => Tau.

Definicja (Gramatyka lewostronnie liniowa)

Gramatykę nazywamy lewostronnie liniową, jeżeli każda jej produkcja jest postaci: , gdzie .

Definicja (Gramatyka prawostronnie liniowa)

Gramatykę nazywamy prawostronnie liniową, jeżeli każda jej produkcja jest postaci: , gdzie .

Definicja (Gramatyka regularna)

Gramatyki prawostronnie liniowe i lewostronnie liniowe nazywa się gramatykami regularnymi.

Definicja(Postać Chomsky’ego, PNC, CNF)

Gramatyka bezkontekstowa jest w postaci Chomsky'ego, jeżeli wszystkie jej reguły są zapisane w jednej z trzech postaci: , gdzie .

Twierdzenie (Równoważność CNF)

Dla każdej gramatyki bezkontekstowej istnieje równoważna gramatyka bezkontekstowa w postaci Chomsky'ego

Odniesień do dowodu należy szukać w pracy Skórzewskiego [7]

Definicja (klasyfikator GCS)

Klasyfikator GCS jest zbiorem postaci , gdzie:

– Lewa strona produkcji, .

– Prawa strona produkcji, .

– ilość użyć zdania w rozkładzie zdań poprawnych.

– ilość użyć zdania w parsowaniu zdań niepoprawnych.

– ilość punktów zdobyta przez algorytm w wyniku parsowania zdań poprawnych (profit).

– ilość punktów zdobyta przez algorytm w wyniku parsowania zdań niepoprawnych (debt).

– przystosowanie klasyfikatora.

Symbolem będziemy oznaczać zbiór wszystkich klasyfikatorów gramatyki G.

Definicja (Gramatyka GCS)

Gramatykę generowaną podczas działania algorytmu GCS będziemy nazywać gramatykę bezkontekstową złożoną z reguł postaci CNF. W odniesieniu do tej gramatyki będzie się stosować alternatywną reprezentację następującej postaci: , gdzie:

– zbiór symboli nieterminalnych;

–zbiór symboli terminalnych.;

– produkcje terminalne, ;

– produkcje nieterminalne, ;

– symbol startowy ;

– symbol uniwersalny .

Definicja(Drzewo wyprowadzenia/derywacji/rozkładu)

Niech będzie gramatyką bezkontekstową. Drzewo jest drzewem wyprowadzenia (lub derywacji, albo rozkładu) dla gramatyki , jeżeli:

* ;
* ;
* Jeżeli wierzchołek wewnętrzny jest etykietowany symbolem , to ;
* Jeżeli synowie wierzchołka drzewa etykietowanego symbolem mają kolejno etykiety , to jest produkcją ze zbioru P.

Definicja (Stochastyczna gramatyka bezkontekstowa)

Stochastyczna gramatyka bezkontekstowa sGCS to taki zbiór, że , gdzie:

* (N, T, P, S) jest gramatyką bezkontekstową;

jest funkcją prawdopodobieństwa przydzielającą każdej regule taką wartość prawdopodobieństwa, że suma prawdopodobieństw dla reguł o lewej stronie A wynosi 1.

Definicja (klasyfikator sGCS)

Klasyfikator sGCS jest zbiorem postaci , gdzie:

– Lewa strona produkcji, .

– Prawa strona produkcji, .

– ilość użyć zdania w rozkładzie zdań poprawnych.

– ilość zastosowań produkcji w najbardziej prawdopodobnych drzewach rozbiorów zdań.

– ilość punktów zdobyta przez algorytm w wyniku parsowania zdań poprawnych (profit).

– prawdopodobieństwo produkcji.

– funkcja przystosowania (ang. fitness).

Definicja (Znormalizowane prawdopodobieństwo)

Prawdopodobieństwo znormalizowane to funkcja o sygnaturze: . Przyporządkowuje ona każdej regule prawdopodobieństwo , zachowując przy tym zależność .

Definicja (Gramatyka sGCS)

Gramatykę sGCS przedstawiamy w następującej postaci: . Jedyny nowy element w stosunku do gramatyki GCS to , które jest znormalizowaną funkcją prawdopodobieństwa.

Definicja (Zdanie)

Mając alfabety A, B (takie, że ) oraz gramatykę G nad alfabetem A, przez zdanie będziemy rozumieć dwójkę postaci , gdzie:

* ;

Zatem zdanie jest krotką zawierającą słowo nad pewnym alfabetem oraz informację, czy należy ono do danej gramatyki.

Definicja (Przystosowanie)

Po wykonaniu parsowania na całym zbiorze uczącym każdy klasyfikator otrzymuje miarę swojej użyteczności, zwaną przystosowaniem. Wyliczamy je przy pomocy tak zwanej funkcji przystosowania, zgodnie z poniższymi wzorami:

– miara użyteczności klasyfikatora niebiorącego udziału w parsowaniu;

– klasyczna funkcja przystosowania;

– funkcja płodności;

– ilość wykorzystań klasyfikatora w parsowaniu zdań pozytywnych;

– ilość wykorzystań klasyfikatora w parsowaniu zdań negatywnych;

– suma punktów zdobytych za parsowanie zdań poprawnych;

– suma punktów zdobytych za parsowanie zdań niepoprawnych;

– waga rozbioru zdania poprawnego;

– waga rozbioru zdania niepoprawnego;

– waga funkcji klasycznej;

– waga funkcji płodności;

– maksymalna ilość punktów zdobytych przez różnicę ;

– minimalna ilość punktów zdobytych przez różnicę .

Definicja (Pozytywna reprezentacja gramatyki)

Pozytywna reprezentacja gramatyki to pewien podzbiór zdań należących do gramatyki G.

Definicja (Pełna reprezentacja gramatyki)

Pozytywna reprezentacja gramatyki to pewien podzbiór wszystkich możliwych zdań (zarówno należących, jak i nie do gramatyki G).

Algorytm CYK

Podczas indukcji gramatyki ważne jest ustalenie, czy dane zdanie należy do gramatyki (*ang. membership query*). Najbardziej popularnymi algorytmami parsowania gramatyki są algorytm CYK oraz Earleya. Algorytm CYK [1](nazwa pochodzi od nazwisk twórców, Cocke’a, Youngera oraz Kasamiego) to dynamiczny algorytm mający za zadanie sprawdzenie, czy dane zdanie należy do gramatyki. Jego złożoność to , gdzie – długość sprawdzanego zdania, a – rozmiar gramatyki. Warunkiem niezbędnym do zastosowania tego algorytmu jest posiadanie gramatyki przedstawionej za pomocą zbioru reguł w postaci normalnej Chomsky’ego. Poniżej przedstawiono pseudokod algorytmu:

Rysunek 1: Algorytm CYK

procedure CYK

begin

n := len(x)

for I := 1 to n do

;

for j := 2 to n do

for i := 1 to n do

begin

;

for k := 1 to j – 1 k do

;

end

end

Działanie algorytmu polega na sprawdzaniu, czy , gdzie jest podłańcuchem zdania , rozpoczynającym się na pozycji , mającym długość . Jeżeli istnieje takie wyprowadzenie, wówczas algorytm zapamiętuje ten fakt. Stosując indukcję po wyprowadzamy coraz dłuższe łańcuchy, aż po osiągnięciu możemy rozstrzygnąć, czy Ponieważ , zatem .

Dane funkcjonowania algorytmu CYK są zapisywane w tablicy trójkątnej o wymiarach . Zawartość komórki [j, i] tej tabeli odpowiada zbiorowi .

Rysunek 2: Przykładowa gramatyka

* S -> AB
* A -> a
* B -> b
* A -> AB
* B -> AA
* A -> BB

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **a** | **b** | **a** | **b** | **a** |
| A | B | A | B | A |
| S, A | - | S, A | - |  |
| B | - | B |  |  |
| B, A | A |  |  |  |
| B, S, A |  |  |  |  |

Rysunek 3: Przykładowa tabela CYK

Powyżej przedstawiono przykładowe wypełnienie tabeli CYK dla zdania oraz gramatyki G. Tabelę tą rozszerzono dodatkowo o zerowy wiersz zawierający zdanie w celu lepszej reprezentacji idei algorytmu.

Pierwszy wiersz tabeli uzupełnia się poprzez wpisanie produkcji wyprowadzających pojedyncze symbole nieterminalne. Następnie należy przejść do uzupełniania wierszu drugiego. Uzupełnia się go takimi produkcjami, że wyprowadzają one symbole nieterminalne z wyższych wierszy. Przykładowowo chcąc uzupełnić komórkę [3,3] musimy najpierw uzupełnić wiersze 1 oraz 2. Następnie znajdujemy taki zestaw produkcji , że . Zapisujemy wszystkie odnalezione w ten sposób wartości w komórce [3, 3]. Ponieważ komórka nie zawiera żadnych symboli, więc w efekcie nie znajdujemy żadnej produkcji. Następnie szukamy w ten sam sposób kolejnych wartości , tym razem z wykorzystaniem komórek , . Nie istnieje żadna produkcja postaci , więc kombinacja tych dwóch symboli nic nam nie daje. Następnie próbujemy dopasowania . Przyglądając się zestawowi reguł przedstawionemu na [..] widzimy, że pod wartość możemy podstawić symbol . Dodajemy go więc do zbioru symboli w komórce . Nie istnieje również żadna produkcja postaci . Skończyły nam się kombinacje symboli w komórkach , nie mamy również kolejnych kombinacji komórek do wzięcia. Przechodzimy do uzupełniania kolejnej komórki – czyli .

Postępując w ten sposób uzupełniamy wszystkie komórki tabeli CYK. Dysponując pełnym zestawem symboli w jesteśmy w stanie ustalić czy dane zdanie zostało sklasyfikowane jako należące do gramatyki, czy też nie. Zdanie należy do gramatyki, jeżeli w tej komórce znajduje się symbol startowy gramatyki. Wówczas zdanie jest wyprowadzalne z symbolu startowego, co możemy pokazać odtwarzając całe wyprowadzenie na podstawie tabeli CYK. Jak widzimy zdanie należy do gramatyki .

Tablica CYK zawiera wszystkie możliwe rozkłady zdania.

Wariantem, jaki możemy wprowadzić do tego algorytmu, jest wprowadzenie powielania symboli nieterminalnych w tabeli . Zakładając, że mamy zdanie postaci oraz gramatykę, na którą składają się wyłącznie reguły i otrzymujemy tabelę danych, którą ilustruje Rysunek 4.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **a** | **a** | **a** | **a** | **a** |
| S | S | S | S | S |
| S | S | S | S |  |
| S, S | S, S | S, S |  |  |
| S, S, S, S, S | S, S, S, S, S |  |  |  |
| S, S, S, S, S, S, S, S, S, S, S, S, S, S |  |  |  |  |

Rysunek 4: Przykładowa tabela CYK z powtarzającymi się produkcjami

Widać zatem, że CYK może w przypadku takiej reprezentacji osiągnąć złożoność . Wynika to z faktu, że liczba wszystkich możliwych wyprowadzeń ciągu o długości jest funkcją wykładniczą . Reprezentacja taka bywa jednak przydatna gdy chcemy później odtworzyć drzewo wywodu, w szczególności jeżeli zapamiętujemy dodatkowo informacje jednoznacznie określające skąd wziął się dany symbol. W aplikacji opisanej w tej pracy magisterskiej zdecydowano się na taką właśnie reprezentację.

GCS

Algorytm GCS jest wieloetapowym algorytmem mającym na celu wygenerowanie gramatyki bezkontekstowej służącej do rozwiązania problemu klasyfikacji. Został on zaproponowany w publikacji Unolda [1]. Na gramatykę składają się reguły terminalne oraz nieterminalne zapisane w postaci normalnej Chomsky’ego (CNF) [7]. Generowanie odbywa się z wykorzystaniem zarówno pozytywnych, jak i negatywnych zdań w procesie uczącym (zgodnie z twierdzeniem Golda niemożliwa jest indukcja jedynie na podstawie zdań poprawnych).

Główna pętla algorytmu

Rysunek 5: Algorytm GCS

procedure GCS

begin

wczytaj zestaw zdań uczących

wczytaj parametry algorytmu

inicjuj gramatykę

while (not kryterium stopu) do

begin

indukuj gramatykę

oblicz dla

if then uruchom algorytm genetyczny na gramatyce .

end

end

Na samym początku algorytm wczytuje zestaw danych uczących. Zbiór ten zawiera zarówno zdania pozytywne, jak i negatywne. Następnie wczytywane są również parametry modelu, czyli partametry konfiguracyjne algorytmu. Algorytm GCS ma ich całkiem sporo i zostaną one szczegółowo omówione później.

Kolejnym krokiem jest inicjacja gramatyki . Zadaniem inicjacji jest wygenerowanie populacji początkowej zawierającej reguł nieterminalnych. Są one generowane w sposób losowy z wykorzystaniem symboli nieterminalnych ze zbioru . Część lub całość reguł początkowych może być dostarczonych przez użytkownika. Następnie zaczyna się główna pętla algorytmu, która będzie wykonywana do czasu spełnienia warunków stopu. Warunkiem stopu może być osiągnięcie pewnego pułapu funkcji przystosowania (zazwyczaj oczekuje się wartości 100%), lub osiągnięcie pewnej ilości kroków. Implementacja towarzysząca tej pracy umożliwia dodatkowo zakończenie procesu po upływie określonego czasu . W każdym kroku przeprowadzamy indukcję gramatyki, ocenę jej przystosowania oraz (o ile pozwala na to wartość ) uruchomienie algorytmu genetycznego (obecna implementacja umożliwia wykorzystanie innego algorytmu ewolucyjnego).

Indukcja gramatyki

Rysunek 6:Procedura IndukujGramatykę

procedure Indukuj gramatykę (G)

begin

inicjuj parametry produkcji

if then korekcja gramatyki

for each do

begin

parsuj zdanie gramatyką

aktualizuj parametry produkcji

end

for each do oblicz

end

Indukcję gramatyki przeprowadza się poprzez parsowanie zdanie po zdaniu wszystkich zdań ze zbioru uczącego az wykorzystaniem mechanizmów tworzenia nowych reguł. Dane dotyczące klasyfikatorów zbieramy w . Po zakończeniu parsowania wszystkich zdań obliczamy przystosowanie wszystkich klasyfikatorów.

Parsowanie zdań

Rysunek 7: Procedura parsowania zdania

procedure Parsuj zdanie (r, G)

begin

, jest -tym słowem zdania o długości

inicjuj tablicę reprezentującą otoczenie

for to do

begin

inicjuj zbiory

, jest komunikatem zewnętrznym reprezentującym aktualny stan otoczenia

if then

begin

zastosuj operator pokrycia terminalnego dla

if then zastosuj operator pokrycia startowego dla

end

, czyli uaktywnij akcje w otoczeniu

end

for j to do

for to do

begin

inicjuj zbiory

, podciąg zdania o długości , rozpoczynający się od -tej

pozycji

,

czyli utwórz komunikat reprezentujący aktualny stan otoczenia

if then

begin

if then zastosuj operator pokrycia

agresywnego dla D

if then zastosuj operator pokrycia pełnego dla

end

, czyli uaktywnij akcję w otoczeniu

end

end

Na samym początku zadaniem algorytmu ma być uzupełnienie pierwszego wiersza tabeli . Od zwykłego wypełniania pierwszego wiersza przy pomocy algorytmu różni się to tylko tym, że w przypadku natrafienia na nieznany symbol uruchamiamy operator pokrycia terminalnego. Dodatkowo jeżeli zdanie składa się z pojedynczego słowa i należy przy tym do gramatyki wówczas możemy uruchomić dodatkowo operator pokrycia startowego. Zadaniem tych operatorów jest dodanie nowych reguł, z symbolem terminalnym po prawej stronie, do zbioru reguł terminalnych oraz rozszerzenie zbioru przez uwzględnienie tych dodatkowych produkcji.

Proces uzupełniania wierszy kolejnych również jest tak naprawdę algorytmem CYK rozszerzonym o operatory pokrycia. Są to operatory pokrycia agresywnego oraz pełnego. Są one wykorzystywane do stworzenia nowych reguł nieterminalnych o prawej stronie . Reguły te w klasycznym wydaniu algorytmu GCS są dodawane z tzw. ściskiem, załączona implementacja umożliwia zmianę tego zachowania.

Ścisk

Rysunek 8: Algorytm ścisku

procedure Dodaj ze ściskiem (nowy, P)

begin

for to do

begin

, losowa podpopulacja z o wielkości

end

end

Operatory pokrycia startowego, pełnego oraz startowego, podobnie jak algorytm ewolucyjny dodają reguły do populacji z tak zwanym ściskiem. Jego zadaniem jest dodanie nowej reguły do populacji w taki sposób, żeby jej rozmiar nie uległ zwiększeniu, poprzez zastąpienie nową regułą podobnej, acz mało skutecznej metody już istniejącej w populacji. Parametrami są osobnik, nazwany nowy oraz populacja reguł . Na początku procedura ścisku wybiera z podpopulację reguł o rozmiarze . Bierze z tej populacji najsłabszego osobnika i dodaje go do zbioru zbioru . Procedurę tą kontynuujemy, dopóki nie dysponujemy zbiorem K o rozmiarze . Ponieważ dodawaliśmy tam tylko najsłabszych osobmików, zatem jest to zbiór reguł o potencjalnie niskich współczynnikach fitness, które nie wydają się być potrzebne w naszej populacji.Następnie w tej populacji jest szukany najbardziej podobny do nowego osobnik . Miarą podobieństwa jest tutaj ilość takich samych symboli występoujących w obu regułach na tych samych pozycjach (na przykład podobieństwo reguł oraz wynosi 1, gdyż w przypadku obu reguł ich prawe dziecko to . Reguła ta zostaje zastąpiona przez nowego osobnika.

Operatory pokrycia

Zadaniem operatorów pokrycia jest dodawanie nowych reguł do populacji. Zaimplementowana biblioteka pozwala zmienić domyślną metodę dodawania reguł przez operatory, aczkolwiek domyślnym zachowaniem jest dodawanie przez operatory pokrycia terminalnego oraz uniwersalnego reguł jako nowe pozycje, zaś startowego, agresywnego oraz pełnego ze ściskiem. Dodatkowo umożliwia zastosowanie prawdopodobieństwa uruchomienia również względem operatorów wymagających domyślnie zmiennych dyskretnych (wykorzystanie klasyczne również jest możliwe, 100% jest odpowiednikiem true, zaś 0% - false. Czyni to jednak cały model spójniejszym oraz daje nowe możliwości (np. symbol uniwersalny wyprowadzający tylko część symboli terminalnych. Możliwe jest też rozszerzenie o elityzm.

Operator pokrycia terminalnego

Rysunek 9: Operator pokrycia terminalnego

procedure Operator pokrycia terminalnego (a)

begin

utwórz

if then zastosuj operator pokrycia uniwersalnego dla

end

Operator pokrycia terminalnego jest uruchamiany, gdy w populacji brakuje reguły wyprowadzającej dany symbol terminalny. Zadaniem tego operatora jest wówczas dodanie takiej reguły do populacji (co powoduje zwiększenie rozmiaru populacji), a następnie uruchomienie operatora pokrycia uniwersalnego, jeżeli pozwala na to wartość .

Operator pokrycia uniwersalnego

Rysunek 10: Operator pokrycia uniwersalnego

procedure Operator pokrycia uniwersalnego (a)

begin

utwórz

end

Operator pokrycia uniwersalnego ma za zadanie wyprowadzić produkcje postaci . Ponieważ operator pokrycia uniwersalnego jest uruchamiany razem z operatorem pokrycia terminalnego, uruchamianego przy każdym nowym słowie, więc , czyli .

Operator pokrycia startowego

Rysunek 11: Operator pokrycia startowego

procedure Operator pokrycia startowego (a)

begin

utwórz

dodaj ze ściskiem do

end

Uruchomienie operatora pokrycia startowego jest uzależnione od zmiennej . Dodatkowo jest on stosowany wyłącznie w przypadku pozytywnych zdań długości 1. Zadaniem tego operatora jest dodanie do populacji produkcji wyprowadzających dane symbole terminalne bezpośrednio z symbolu startowego. Są one dodawane do populacji ze ściskiem.

Operator pokrycia agresywnego

Rysunek 12: Operator pokrycia agresywnego

procedure Operator pokrycia agresywnego (D)

begin

utwórz

dodaj ze ściskiem do

end

Uruchomienie operatora pokrycia agresywnego jest wykonywane z prawdopodobieństwem . Operator ten dodajeze ściskiem nową regułę do populacji reguł nieterminalnych.

Operator pokrycia pełnego

Rysunek 13: Operator pokrycia pełnego

procedure Operator pokrycia pełnego (D)

begin

utwórz

dodaj ze ściskiem do

end

Operator pokrycia pełnego tworzy produkcję, której lewa strona jest symbolem startowym, zaś prawa należy do do zbioru . Jest on uruchamiany tylko w ostatniej komórce tabeli , wyłącznie w przypadku rozbioru zdania pozytywnego. Nowa reguła jest następnie dodawana ze ściskiem do pozostałych reguł.

Algorytm genetyczny

Rysunek 14: Algorytm genetyczny

procedure Algorytm genetyczny (G)

begin

if liczba losowa z then zastosuj krzyżowanie na i

zastosuj mutację na

zastosuj mutację na

if liczba losowa z then zastosuj inwersję na

if liczba losowa z then zastosuj inwersję na

dodaj ze ściskiem do

dodaj ze ściskiem do

end

Algorytm genetyczny wraz z jego działaniem został szczegółowo opisany w [9]. Na początku ma miejsce selekcja dwóch osobników z populacji reguł nieterminalnych. Rodzaj selekcji jest zależny od zmiennych decyzyjnych oraz . Dostępne rodzaje selekcji to selekcja metdoą losową, ruletki lub turniejową. Następnie na kopiach wybranych produkcji są wykonywane operacje genetyczne (krzyżowanie, mutacja oraz inwersja). Wynik działania zostaje dodany do populacji ze ściskiem, chroniąc równocześnie przed usunięciem najlepszych produkcji przy pomocy mechanizmu elityzmu.

Metody selekcji

Istnieje wiele metod selekcji, tutaj jednak zostaną opisane trzy najpopularniejsze.

Selekcja ruletkowa

W selekcji ruletkowej szansa na wybranie klasyfikatora to wartość jego przystosowania w stosunku do sumy wszystkich klasyfikatorów. Bierzemy wartości dla wszystkich klasyfikatorów i budujemy z nich ruletkę; tj. każdemu klasyfikatorowi przypisujemy wycinek koła proporcjonalny do prawdopodobieństwa jego wyboru. Selekcja jest dokonywana po losowym obrocie koła ruletki.

Selekcja turniejowa

W selekcji turniejowej losujemy populację o rozmiarze . Z tak wylosowanej podpopulacji wybieramy klasyfikator o najwyższej wartości funkcji przystosowania .

Selekcja losowa

Zwraca losowy klasyfikator zgodnie z rozkładem normalnym, nie uwzględniając przy tym funkcji przystosowania.

**Inwersja**

Rysunek 15: Operator inwersji

procedure Inwersja(p)

begin

utwórz

end

Operator inwersji jest uruchamiany z prawdopodobieństwem dla każdej produkcji. Powoduje on odwrócenie kolejności symboli po prawej stronie produkcji.

Mutacja

Rysunek 16: Operator mutacji

procedure Mutacja(p)

begin

utwórz

if then

if then

if then

end

Mutacja również jest uruchamiana dla każdej produkcji. Z prawdopodobieństwem zmianie może ulec lewa strona produkcji, pierwszy symbol prawej strony produkcji oraz drugi symbol prawej strony produkcji. Test prawdopodobieństwa powtarzamy dla każdego z elementów, również symbol X jest losowy i przy każdym z tych testów może przyjąć inną wartość.

Krzyżowanie

Rysunek 17: Operator krzyżowania

procedure Krzyżowanie(, )

begin

utwórz

utwórz

if then

begin

end

else

begin

end

end

Krzyżowanie dwóch produkcji powoduje zamianę pierwszego lub drugiego symbolu prawej strony dwóch reguł , . Następnie reguły te zamieniają się lewymi stronami. Krzyżowanie jest wykonywane z prawdopodobieństwem .

Korekcja gramatyki

Rysunek 18: Algorytm korekcji gramatyki

procedure Korekcja gramatyki(G)

begin

usuń produkcje redundantne z

usuń produkcje nieproduktywne z

usuń produkcje nieosiągalne z

end

Algorytm korekcji gramatyki ma za zadanie usunąć z niej zbędne produkcje, które powstają w trakcie indukcji oraz ewolucji gramatyki. Wykonuje to zadanie stosując usuwanie produkcji redundantnych, nieproduktywnych oraz nieosiągalnych.

Usuwanie produkcji redundantnych

Usuwanie produkcji polega na przejrzeniu populacji i usunięciu powtarzających się reguł. Ze względu na konstrukcję populacji w dostarczonej implementacji duplikaty reguł nie są w ogóle dodawane do populacji, więc krok ten nie jest potrzebny.

Usuwanie produkcji nieproduktywnych

Rysunek 19: Algorytm usuwania produkcji nieproduktywnych

procedure Usuń produkcje nieproduktywne(G)

begin

repeat

for each do

if then

until się nie zmienia

end

Kolejnym zadaniem jest usunięcie produkcji nieproduktywnych, czyli takich reguł z których nie da się wyprowadzić ciągu złożonego wyłącznie z symboli nieterminalnych.

Na samym początku w zbiorze zapisujemy takie reguły, które bezpośrednio wyprowadzają symbole terminalne. Następnie przeglądamy populacje reguł, dodając do zbioru takie produkcje, które bezpośrednio wyprowadzqają jedynie symbole znajdujące się już w . Postępując w ten sposób cały czas powiększamy zbiór . Korzystając z indukcyjnej właściwości wyprowadzalności w zbiorze zawsze znajdują się wyłącznie symbole wyprowadzające co najmniej jeden ciąg złożony wyłącznie z symboli nieterminalnych. Kiedy nie ulegnie on już zmianie oznaczać to będzie, że w mamy zbiór wszystkich symboli wyprowadzających co najmniej jeden ciąg złożony wyłącznie z symboli terminalnych. Wystarczy teraz pozostawić takie reguły, których prawe strony składają się jedynie z symboli w .

Usuwanie produkcji nieosiągalnych

Rysunek 20: Algorytm usuwania produkcji nieosiągalnych

procedure Usuń produkcje nieosiągalne(G)

begin

repeat

for each do

begin

end

until i się nie zmienia

end

Produkcje nieosiągalne to takie produkcje, których nie da się osiągnąć zaczynając wyprowadzenie z .

Tworzymy dwa zbiory – , czyli zbiór symboli osiągalnych z oraz - zbiór produkcji osiągalnych. W szczególności jest symbolem osiągalnym z , toteż od początku znajduje się w . Następnie przeglądamy reguły, dodając do produkcje o lewej stronie z , dodając symbole z ich prawej strony do . Tym sposobem coraz bardziej powiększamy zbiór symboli osiągalnych z , powiększamy równocześnie zbiór produkcji osiągalnych. Kiedy żaden z tych zbiorów nie ulega już zmianie oznacza to, że w jest zbiór wszystkich symboli osiągalnych z , więc w celu usunięcia produkcji nieosiągalnych wystarczy jedynie przypisać go jako nowy zbiór .

sGCS

Kolejnym algorytmem wykorzystywanym w tej pracy magisterskiej jest algorytm sGCS (*ang. Stochastic Grammar-based Classifier System*). System ten stanowi rozszerzenie algorytmu GCS o obsługę omówionych wcześniej stochastycznych gramatyk bezkontekstowych. Został on wprowadzony i szczegółowo przedstawiony w pracy magisterskiej Kępy [10], a następnie zmodyfikowany i ulepszony w pracy Pasieki [2]. Poniższy rozdział skupi się na omówieniu różnic pomiędzy dwoma algorytmami. Pojawią się nowe elementy, niewystępujące w algorytmie GCS, niezbędne do indukowania struktury oraz parametrów gramatyki stochastycznej.

Algorytm CYK+

Pseudokod algorytmu CYK+ wydaje się znacząco różnić od pseudokodu algorytmu CYK. Jeżeli jednak zastanowić się nad samą funkcjonalnością to wprowadza on w zasadzie dwie rzeczy – po pierwsze pamięta w danych komórkach prawdopodobieństwa danych poddrzew rozbioru zdania, po drugie zapamiętuje informacje pomocnicze w odtworzeniu drzew lub drzewa wywodu. Obie wersje pseudokodu możemy obejrzeć w pracy Kępy [10], tutaj jednak skupimy się wyłącznie na wyjaśnieniu wprowadzonych funkcjonaności.

Rysunek 21: Przykładowa gramatyka stochastyczna

* S -> AB (1)
* A -> a (0,4)
* B -> b (0,5)
* A -> AB (0,3)
* B -> AA (0,5)
* A -> BB (0,3)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **a** | **b** | **a** | **b** | **a** |
| A(0,4) | B(0,5) | A(0,4) | B(0,5) | A(0,4) |
| S(0,2), A(0,06) | - | S(0,2), A(0,06) | - |  |
| B(0,012) | - | B(0,012) |  |  |
| B(0,0018), A(0,0048) | A(0,0018) |  |  |  |
| B(0,00096), S(0,00072), A(0,00072) |  |  |  |  |

Rysunek 22: Przykładowa tabela CYK+

Patrząc na Rysunek 21, przedstawiający reguły gramatyki stochastycznej oraz na Rysunek 22, przedstawiający przykładowy rozkład zdania ababa przy pomocy tych reguł widać, że komórki tabeli CYK+ uzupełnia się dokładnie w ten sam sposób jak w przypadku CYK. Dodatkowo w pierwszym wierszu wpisujemy prawdopodobieństwo dla reguły terminalnej wyprowadzającej dany symbol terminalny. W przypadku kolejnych wierszy sytuacja trochę się komplikuje – często symbol w danej komórce da się wyprowadzić na więcej niż jeden sposób. Istnieje wiele sposobów obliczania prawdopodobieństwa w takich systuacjach (wprowadzają je między innymi Pasieka [2]), poniżej przedstawiono dwa najpopularniejsze:

Podejście Viterbi

To podejście dąży do maksymalizacji prawdopodobieństwa wyprowadzenia danego ciągu. Jeżeli w komórce tabeli CYK+ znajduje się już dany symbol, wówczas zapamiętujemy większe z dwóch prawdopodobieństw (patrz wzór poniżej).

Rysunek 23: Prawdopodobieństwo przy podejściu Viterbi

Podejście Baum-Welch

W tym podejściu sumujemy prawdopodobieństwa wszystkich poddrzew wyprowadzenia danego ciągu (wzór poniżej).

Rysunek 24: Prawdopodobieństwo przy podejściu Baum-Welch

Poza samym prawdopodobieństwem musimy również przechowywać w jakiś sposób dane umożliwiające łątwe odtworzenie drzewa derywacji przez algorytm Traceback. Najłatwiej tego dokonać dodając do każdego symbolu w tabeli CYK system koordynatów umożliwiający jednoznaczne określenie kto jest jego rodzicem. W obecnej implementacji wykonuje się przy pomocy listy krotek , które są następującej postaci:

Rysunek 25: Przechowywanie śladu drzewa derywacji

– wiersz obecnie rozpatrywanego symbolu.

– kolumna obecnie rozpatrywanego symbolu.

– z której to z kolei kombinacji komórek uzyskano symbol. Wartość pozwala łatwo odtworzyć informację wg poniższego wzoru:

– wartość umożliwiająca jednoznaczne określenie, który symbol z komórki o współrzędnych jest rodzicem.

– wartość umożliwiająca jednoznaczne określenie, który symbol z komórki o współrzędnych jest rodzicem.

– reguła wyprowadzająca z obecnego symbolu oraz .

– prawdopodobieństwo policzone przy pomocy jednego z podejść.

Traceback

Rysunek 26: Algorytm Traceback

procedure Traceback(CYK)

begin

function aux(węzeł)

begin

aktualizuj fitness

if then

begin

end

end

aux()

end

Algorytm umożliwiający odtwarzanie drzew derywacji. Jest on niezbędny do prawidłowego aktualizowania funkcji przystosowania algorytmu . Zazwyczaj porusza się on po najbardziej prawdopodobnym drzewie rozkładu, chociaż istnieją też inne implementacje [2].

Normalizacja

Dodając nowe i usuwając stare reguły należy pamiętać o tym, że zgodnie z definicją […] prawdopodobieństwo musi być znormalizowane. Obecna implementacja dokonuje tego w locie, obliczając prawdopodobieństwo zgodnie z poniższym wzorem:

Rysunek 27: Wzór na normalizację prawdopodobieństwa

Wartości te są pamiętane, dzięki czemu nie jest konieczne liczenie ich za każdym razem. Zatem normalizacja odbywa się poprzez podzielenie prawdopodobieństwa interesującej nas reguły przez prawdopodobieństwo wszystkich reguł o takiej samej lewej stronie.

W przypadku dodawania nowych reguł ich prawdopodobieństwo jest losowane. Później wartość ta staje się zależna od wartości funkcji przystosowania.

Przystosowanie w sGCS

W przypadku funkcji przystosowania za Pasieką […] stosujemy funkcję przystosowania opartą na ilości wystąpień zdania w drzewach wyprowadzeń zdań pozytywnych.

Rysunek 28: Przystosowanie w sGCS

gdzie – ilość wystąpień danej reguły w rozbiorze zdań pozytywnych.

Estymacja prawdopodobieństw

Kolejnym ważnym elementem jest moduł estymacji. Zadaniem tego modułu jest obliczenie wartości prawdopodobieństw reguł. Jest ono przeliczane po każdym zakończonym etapie indukcyjnym.

Proces estymacji wykorzystuje obliczone wcześniej wartości parametru , aktualizowany przez każde uruchomienie algorytmu Traceback. Poniżej przedstawiono sposób obliczania prawdopodobieństwa na etapie estymacji:

Rysunek 29: Estymacja prawdopodobieństw w sGCS

neg-sGCS

Algorytm sGCS nie wykorzystuje w procesie uczenia zdań negatywnych. W dotychczasowych badaniach [10] [2] ten obiecujący algorytm często osiągał stosunkowo słabe rezultaty. Jest to spowodowane niepełnością wiedzy w przypadku podawaniu algorytmowi jedynie pozytywnych przykładów – o ile jest zawsze powinno być możliwe wygenerowanie dla pozytywnej reprezentacji języka takiej grtamatyki stochastycznej , że , to dotychczasowe badania [10] pokazują, że znacznie bardziej prawdopodobne jest wygenerowanie nadgramatyki, czyli takiej gramatyki , że . Prowadzi to do nieprawidłowego rozpoznawania zdań negatywnych jako pozytywne, czego chcielibyśmy uniknąć. Praca ta podejmuje próbę realizacji wariantu sGCS, który byłby w stanie uczyć się na zdaniach negatywnych.

Algorytm neg-sGCS nie różni się praktycznie niczym od algorytmu sGCS. Jedyne zmiany to:

* Wykorzystanie zdań negatywnych w procesie uczenia (tj. wysyłanie ich w ogóle do parsera CYK+);
* Stosowanie funkcji fitness GCS – stosujemy tutaj system nagrody i kary (profit oraz debt) za parsowanie zdań pozytywnych oraz negatywnych. Stosujemy tutaj wagi nagrody za parsowanie zdań pozytywnych oraz kary za parsowanie zdań negatywnych jak w przypadku GCS opisanego w publikacji Unolda [1], czyli odpowiednio 1 i 2. Początkowo istniała obawa, że zastosowanie takich wag spowoduje zmianę zachowania algorytmu w drugą stronę – tj. generowanie gramatyk u których rozpoznawanie zdań pozytywnych stoi na niskim poziomie. Jak pokazały jednak badania zjawisko to nie wystąpiło;

1. Implementacja

Aplikacja została zaimplementowana jako biblioteka języka Python w wersji 3.4. Może ona zatem być wykorzystywana w dowolnym środowisku posiadającym dostęp do interpretera języka Python w wersji 3.4 lub nowszej oraz niezbędne biblioteki.

Biblioteki niezbędne do wykorzystania aplikacji zostały wymienione w pliku „requirements.txt” (znajdującym się w głównym katalogu projektu). Są to:

* PyHamcrest – wymagany do uruchomienia testów. Jest to biblioteka dostarczająca zestaw rozszerzeń do standardowego modułu testów jednostkowych unittest, które czynią testy wysoce ekspresywnymi;
* Matplotlib – biblioteka umożoliwająca tworzenie wykresów dla języka Python, jest wykorzystywana do rysowania diagramów procesu uczenia gramatyki;
* Mock – wykorzystywany do tworzenia atrap obiektów (ang. Mock Object) na potrzeby testowania aplikacji. W bibliotece wykorzystany w testach jednostkowych, modułowych oraz komponentowych;
* Psutil – wykorzystywany do pobierania informacji o obecnym obciążeniu procesora i zajętej pamięci niezależnie od systemu operacyjnego;

Wszystkie te biblioteki możemy bardzo łatwo zainstalować przy pomocy narzędzia do instalowania modułów pip.

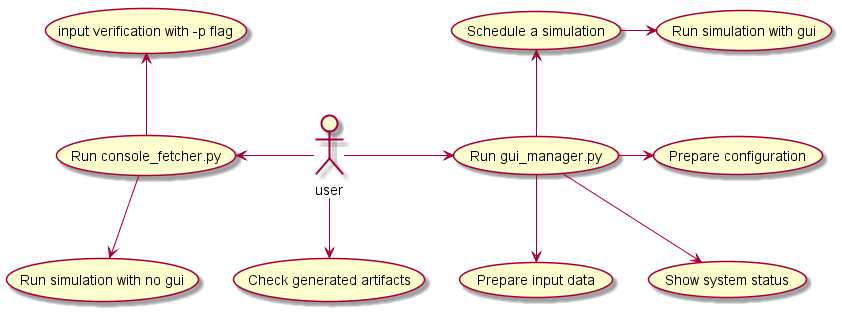
pip jest wygodnym menedżerem pakietów języka Python. Stanowi ważne narzędzie dla programistów rozwijających aplikacje w tym języku. Umożliwia łatwe instalowanie oraz usuwanie bibliotek języka python. zazwyczaj wystarczy polecenie , aby pip automatycznie pobrał pakiet wraz ze wszystkimi niezbędnymi zależnościami, zainstalował je i uczynił widocznymi z poziomu dowolnej aplikacji w języku Python uruchamianej na naszym komputerze. Oprócz ręcznego instalowania bibliotek jedna po drugiej istnieje możliwość zebrania ich nazw w pliku tekstowym (nazwijmy go na przykład requirements.txt), umieszczając każdą nazwę w osobnej linii, a następnie uruchomić tylko polecenie . Biblioteka wykorzystuje to drugie podejście – wywołanie na najwyższym poziomie projektu spowoduje zainstalowanie wszystkich zależności niezbędnych do uruchomienia algorytmu w trybie konsolowym.

W celu uruchomienia aplikacji w trybie okienkowym należy dodatkowo posiadać zainstalowany pakiet PyQt oraz wykorzystywane przez niego biblioteki Qt w odpowiedniej wersji. Ponieważ proces ten różni się znacznie na różnych systemach, a jest on niezbędny tylko do uruchomienia przykładowej otoczki graficznej dostarczanej biblioteki, więc nie zostanie on tutaj omówiony. Informacji na temat instalacji biblioteki PyQt oraz Qt na posiadanym systemie należy szukać na stronie twórców PyQt, Qt oraz na forach poświęconych konkretnym systemom.

Podczas pisania aplikacji wykorzystywano metodologię wytwarzania oprogramowania TDD. Dzięki temu pokrycie aplikacji stoi na całkiem wysokim poziomie 88%. Poziom ten byłby wyższy gdyby nie ograniczenia czasowe uniemożliwijące prawidłowe pokrycie testami jednej z najwyższych warstwy aplikacji – warstwy logiki biznesowej. Poza testami jednostkowymi w aplikacji znajdują się również testy modułowe oraz komponentowe (za komponent uznajemy tutaj całą włąściwą część algorytmu, pozbawioną dodatków typu części odpowiedzialne za rysopwanie diagramów, obsługę GUI itp.). W przypadku testów modułowych oraz komponentowych również wykorszystano biblioteki PyHamcrest unttest oraz Mock.

1. Użytkowanie aplikacji

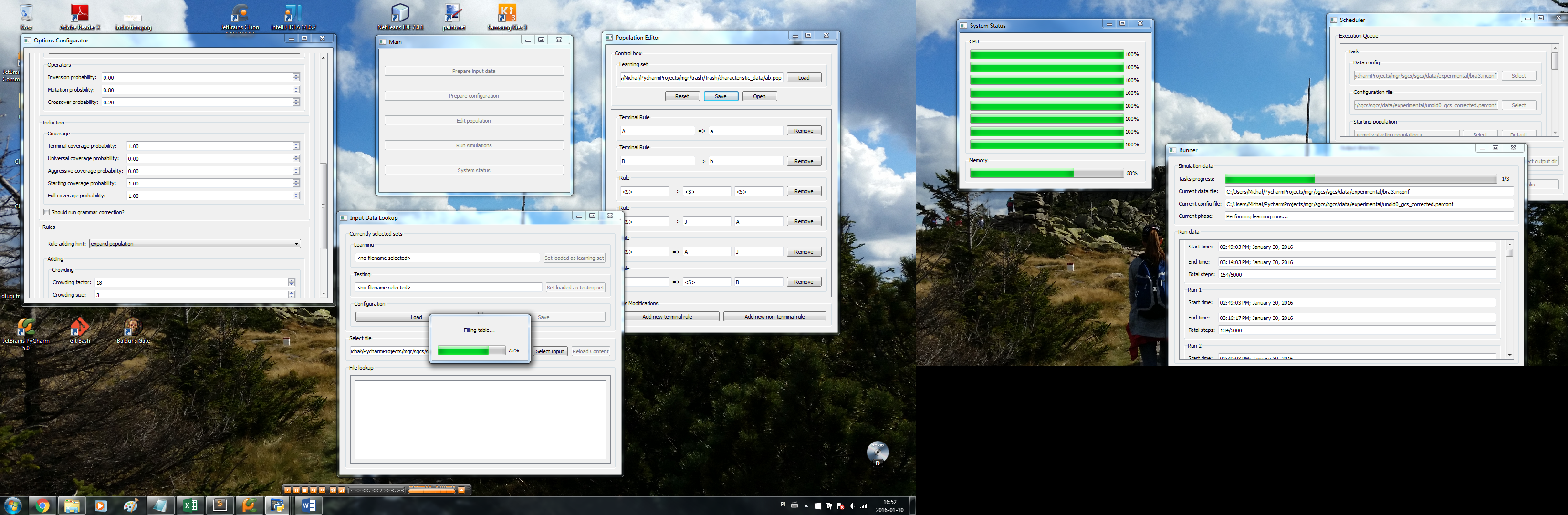
Dodatkowo do biblioteki zostały dołączone skrypty umożliwiające uruchomienie klasycznego modułu GCS/sGCS bez konieczności pisania własnego kodu. Aplikację możemy uruchomić na jeden z dwóch sposobów – z wykorzystaniem skryptu gui\_manager.py albo console\_fetcher.py (przypadki użycia na Rysunek 30).



Rysunek 30: Diagram przypadków użycia aplikacji

1. Wykorzystanie gui\_manager.py

GCS oraz sGCS to algorytmy o wysokim poziomie złożoności. Posiadają kilkadziesiąt parametrów, mogą przyjmować różne dane wejściowe, zaś procesy generowania gramatyki mogą być żmudne i czasochłonne. Dlatego też w trakcie projektowania aplikacji zdecydowano się na zastosowanie wielookienkowego GUI, mając na celu ułatwienie wykorzystania wielu monitorów w pracy oraz łatwe przygotowywanie zbiorów danych do kolejnych testów, podczas gdy poprzednie są jeszcze w trakcie wykonywania. Przykładowe zdjęcie funkcjonowania aplikacji przedstawiono poniżej.



Rysunek 31: Aplikacja w trakcie użytkowania

Poniżej znajduje się omówienie kolejnych okien dostępnych w stworzonej aplikacji oraz prawidłowy sposób ich wykorzystania.

Menu główne

Po uruchomieniu skryptu pojawia się okno menu głównego. Przedstawia je Rysunek 32. Zawiera ono listę przycisków, które można uruchomić. Są to:

Prepare input data – podgląd plików ze zdaniami gramatyki napisanymi w formacie adabingo oraz przygotowywanie zestawów zbiorów uczących – testowych;

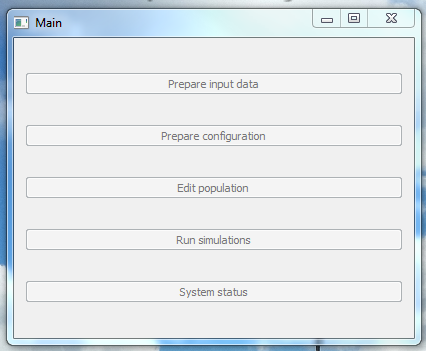
Prepare configuration – umożliwia ustawienie pożądanych parametrów algorytmu i zapisanie tych ustawień w postaci pliku konfiguracyjnego. Dzięki temu jest możliwe późniejsze uruchamianie testów dla różnych kombinacji danych wejściowych i parametrów algorytmu bez konieczności ponownego ich wprowadzania;

Edit population – naciśnięcie tego przycisku spowoduje wyświetlenie edytora populacji. Z jego pomocą możliwy jest wygodny podgląd populacji wygenerowanej przez algorytm, czy też na przykład przygotowanie ręcznie całości lub fragmentu populacji początkowej;

Run simulations – Przycisk ten prowadzi do menu Schedulera, umożliwiającego przygotowanie zestawu testów do uruchomienia;

System status – wynikiem naciśnięcia tego przycisku jest otwarcie okna statusu systemu – okno to pokazuje obciążenie każdego z rdzeni procesora oraz ilość zajętej pamięci RAM. Narzędzie to umożliwia dokonanie oceny czy nasz algorytm w optymalny sposób wykorzystuje dostępne zasoby obliczeniowe.

Każde z wymienionych okien jest singletonem – jeżeli jedno gdzieś istnieje, to nie można w ramach tej samej aplikacji otworzyć drugiego takiego samego (co równocześnie nie przeszkadza w otwarciu okna innego typu).

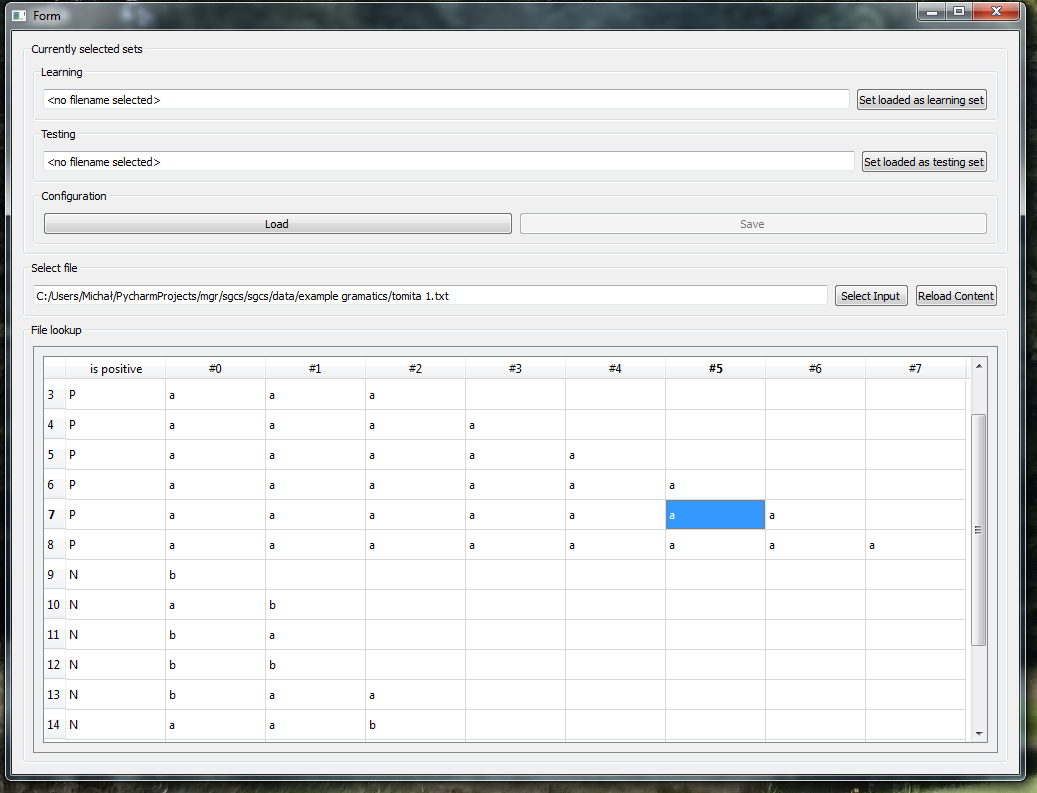


Rysunek 32: Okno menu głównego

Prepare input data

Okno to (Rysunek 33) umożliwia wczytanie pliku wejściowego z gramatyką. Można tego dokonać przy pomocy przycisku Select Input. Wówczas program załaduje wskazany plik do pamięci, podda go dokładnie takiemu samemu procesowi parsowania, jaki będzie miał miejsce przy uruchomieniu algorytmu, po czym wyświetli wszystkie występujące w nim zdania w formie czytelnej tabeli. Jeżeli plik jest duży, pojawi się okienko dialogowe informujące nas o postępie operacji oraz szacowanym postępem wypełniania tabeli. Aplikacja zakłada, że dane wejściowe są zawsze podane w formacie adabingo, aczkolwiek nie wykorzystuje wszystkich informacji oferowanych przez ten format, w wyniku czego zawartość pierwszego wiersza oraz długość zdania w kolejnych nie są weryfikowane ani wykorzystywane. Niemniej nie są to elementy opcjonalne i muszą w podanym pliku wystąpić; w przeciwnym wypadku mozliwe jest uzyskanie błędnie sparsowanego pliku. Można spowodować odświeżenie zawartości tabeli (na przykład w celu upewnienia się, że po dokonaniu modyfikacji zewnętrznym programem dane są nadal prawidłowo parsowane przez aplikację) naciskając przycisk Reload Content.

Po załadowaniu pliku możemy oznaczyć go jako zbiór uczący lub testowy (przyciski Set loaded as learning set oraz Set loaded as testing set). Po oznaczeniu pliku jego ścieżka bezwzględna pojawia się odpowiednio w polu Learning albo Testing. Kiedy wybierzemy oba zbiory, odblokowaniu ulega przycisk Save – możemy taką konfigurację zapisać jako plik konfiguracji \*.inconf. Plik ten będzie zawierał ścieżki bezwzględne do obu podanych zbiorów (Learning oraz Testing).



Rysunek 33: Okno przygotowania danych wejściowych

Prepare configuration

Okno to (Rysunek 34) umożliwia przygotowanie zestawu parametrów dla algorytmu. W pierwszej kolejności warto wybrać najpierw rodzaj planowanego algorytmu. Pole Template jest odpowiedzialne za określenie czy konfiguracja jest planowana dla sGCS, czy GCS. Ze względu na fakt posiadania innych parametrów drzewo konfiguracji tych algorytmów nieco się różni i zmiana wartości tego pola może spowodować zresetowanie wartości pozostałych pól. Poniżej zostały opisane poszczególne pola.

Reset – przycisk powodujący zresetowanie pól do wartości domyślnych.

Save – przycisk umożliwiający zapisanie obecnego stanu konfiguracji.

Open – przycisk umożliwiający wczytanie wcześniej zapisanego pliku ustawień.

Template – rodzaj stosowanego algorytmu. Akceptowane wartości: sGCS i GCS (neg-sGCS to sGCS z zastosowaniem odpowiedniej funkcji przystosowania).

Should evolution be run? – odznaczenie tego pola spowoduje wyłączenie algorytmu genetycznego – gramatyki będą tworzone w oparciu wyłącznie na samym mechanizmie indukcji.

Max algorithm runs – maksymalna ilość uruchomień cyklu uczącego, jaką ma wykonać aplikacja.

Max algorithm execution time – ograniczenie umożliwiające zakończenie pojedynczego uruchomienia uczącego, jeżeli ten przekroczył podany czas. Jeżeli krokowi uczenia zabraknie czasu, w ten sposób, to jest traktowany jakby przekroczył ograniczenie maksymalnej ilości kroków indukcyjnych.

Max evolution steps – maksymalna ilość kroków ewolucyjnych, jaką może wykonać aplikacja w pojedynczym uruchomieniu uczącym.

Satisfying fitness – przy jakiej wartości parametru fitness proces uczenia uzna gramatykę za wyuczoną (np. jeżeli ustawimy tę wartość na 100%, to uruchomienie uczące zostanie uznane za sukces tylko w przypadku poprawnego sklasyfikowania wszystkich zdań).

Statistics configuration – Tutaj można wybrać rodzaj zastosowanej funkcji fitness. Obecnie w systemie są dostępne dwie funkcje fitness – Classical (przedstawiona w Unold1) oraz Pasieka (przedstawiona w Pasieka). Wybór ten decyduje również o tym, czy będziemy dokonywali parsowania zdań negatywnych w procesie uczenia – funkcja Pasieka wykorzystuje jedynie informacje zdobyte w wyniku parsowania zdań pozytywnych, zatem w przypadku jej wybrania zdania negatywne będą ignorowanie, klasyczna wykorzystuje je w procesie uczenia, więc będą one wykorzystane. W trybie GCS mamy możliwość wykorzystania wyłącznie funkcji klasycznej, w trybie sGCS obu.

Selectors – W bloku opisującym algorytm genetyczny istnieje możliwość wybrania do dziesięciu różnych selektorów. Ilość selektorów wiąże się z ilością reguł branych pod uwagę w procesie działania algorytmu genetycznego oraz wiąże się z arnością operatorów genetycznych (co zostanie dokładniej opisane w sekcji omawiającej moduł sgcs.evolution.evolution\_operators). Z selektorów mamy do wyboru selekcję turniejową, losową oraz metodą ruletki.

Operators – Tutaj można nadać prawdopodobieństwo uruchomienia każdego z operatorów genetycznych. Dostępne operatory to operator krzyżowania, mutacji oraz inwersji

Induction – Ta część okna konfiguracyjnego daje możliwość ustalenia prawdopodobieństw operatorów pokrycia. Zaproponowano ujednolicony model parametrów operatorów pokrycia – wszystkie są wyrażone za pomocą prawdopodobieństwa. Więcej szczegółów na ten temat w omówieniu modułu sgcs.induction.coverage\_operators.

Should run grammar corrections? – zaznaczenie tego pola spowoduje usunięcie produkcji nieosiągalnych oraz nieproduktywnych na końcu każdego etapu indukcji gramatyki (tj. tuż przed uruchomieniem algorytmu genetycznego).

Rules adding hint – W tym miejscu możemy sprecyzować ograniczenia na rozmiar zbioru – jeżeli to zrobimy, to wówczas ilość reguł nieterminalnych od chwili uruchomienia algorytmu nie ulegnie zmianie i będzie utrzymywana na stałym poziomie. Dostępne opcje to None (wówczas ta decyzja jest zależna od wartości domyślnej operatorów pokrycia), expand population (ścisk jest wówczas stosowany tylko podczas działania algorytmu ewolucyjnego, operatory pokrycia rozszerzają populację), control population size (elitism DISABLED) – operatory pokrycia dodają nowe reguły ze ściskiem, control population size (elitism ENABLED) dodatkowo uwzględnia mechanizm elityzmu. Słowo hint w przypadku tego pola jest ważne – mianowicie reguła ta zostanie nadpisana, jeżeli w jej wyniku miałoby nastąpić zwiększenie rozmiaru populacji powyżej wartości maksymalnej (patrz Max non terminal rules population size poniżej).

Crowding – w bloku tym mamy dwie kontrolki – Crowding factor oraz Crowding size. Pierwszy z nich decyduje o rozmiarze populacji ścisku (współczynnik cf), drugi o rozmiarze podpopulacji, z której wybieramy każdego członka populacji ścisku (współczynnik cs).

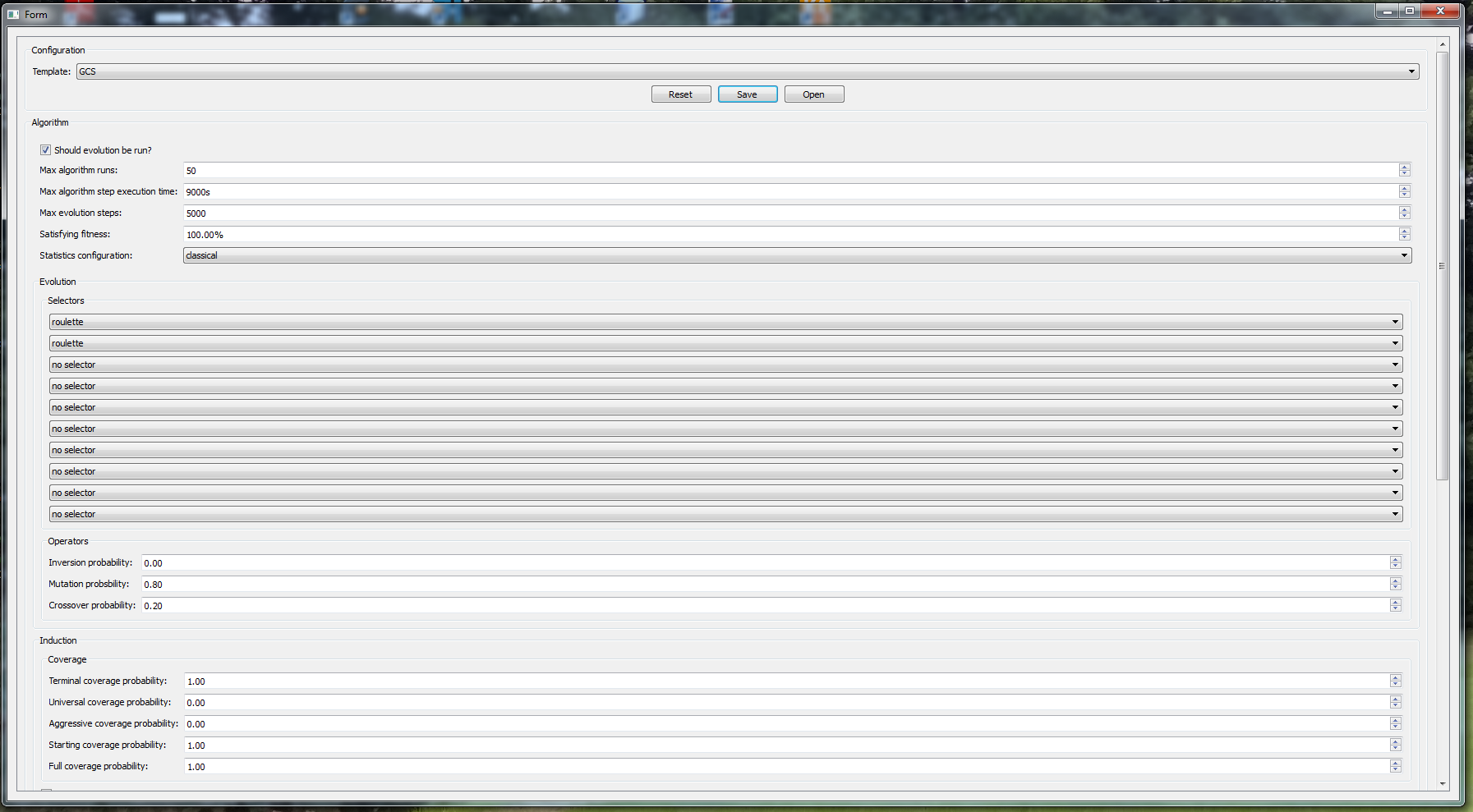
Elitism – na tym etapie możliwe jest podjęcie decyzji na temat elityzmu stosowanego w aplikacji. Możemy go włączyć lub wyłączyć (pole Is elitism used?), można tu też ustalić rozmiar elity.

Max non terminal rules population size: Maksymalny, nieprzekraczalny rozmiar nieterminalnej części populacji. Wartość ta nie może być w algorytmie nigdy przekroczona i jeżeli by to miało nastąpić, to taki precedens ma priorytet nad ustawieniami ścisku i nawet jeżeli jest on domyślnie wyłączony, zostanie zastosowany.

Starting population size – rozmiar populacji reguł przed uruchomieniem algorytmu. Na początku działania algorytmu populacja będzie się składać z reguł dostarczonych w pliku populacji początkowej, powiększonej ewentualnie o losowe reguły w ilości brakującej do tej wartości.

Max non terminal symbols – ilość symboli nieterminalnych, która może wystąpić w algorytmie. Jest to wartość stosunkowo ważna – od rozmiaru tego parametru zależy w końcu przedział symboli, na którym dokonujemy wszelkich operacji losowych. Im mniejszy, tym częściej przy konieczności losowania symbolu będziemy trafiać na te posiadające już jakieś reguły. Z kolei większe wartości pozwolą nam na generowanie bardziej złożonych gramatyk. Ponieważ symbole specjalne (tj. symbol startowy oraz uniwersalny) posiadają specjalne oznaczenia, zatem parametr ten nie posiada górnego ograniczenia.

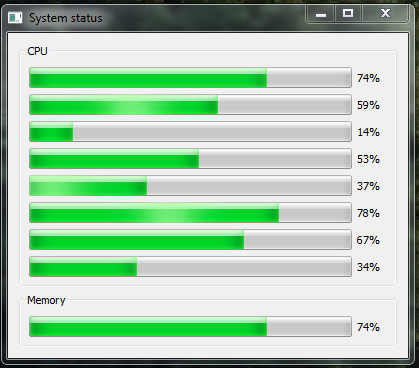
Pliki konfiguracyjne posiadają rozszerzenie \*.parconf.



Rysunek 34: Edycja konfiguracji

System status

Okno to (Rysunek 35) pełni funkcję czysto diagnostyczną, jego zadaniem jest pokazywanie obecnego obciążenia wszystkich rdzeni procesora oraz zużycie pamięci. Informacja ta jest przydatna, gdyż każde uruchomienie cyklu uczącego jest niezależne od pozostałych i możliwe jest puszczenie ich w niezależnych procesach w celu maksymalnego wykorzystania maszyny, na której algorytm został uruchomiony. Dzięki takiemu podglądowi jesteśmy w stanie ocenić czy algorytm w optymalny sposób wykorzystuje zapewnione mu zasoby. Uruchomienie niektórych zbiorów danych może spowodować wielogodzinne stuprocentowe obciążenie procesora, z czym należy się liczyć podczas uruchamiania algorytmu. Jeżeli nasz sprzęt nie jest przygotowany na takie obciążenia (np. nie dysponuje wystarczająco sprawnym układem chłodzącym) uruchamianie algorytmu nie jest wskazane, jako że bez przedsięwzięcia dodatkowych środków ostrożności może ono doprowadzić do wyłączenia komputera, a nawet do trwałego uszkodzenia sprzętu. Biblioteka zawiera jednowątkową wersję algorytmu, aczkolwiek do interfejsu graficznego jest obecnie podpięty jedynie wariant wielowątkowy.



Rysunek 35: Okno System Status

Edit population

Z pomocą przedstawionego poniżej okna możliwe jest zarządzanie populacjami. Można przy jego pomocy otworzyć istniejący plik populacji \*.pop (uzyskany na przykład w wyniku działania algorytmu), czy też utworzyć taką populację od podstaw (na przykład w celu wykorzystania jej w algorytmie jako populacji początkowej). Należy tutaj pamiętać o tym, że wewnątrz aplikacji celem przyspieszenia obliczeń wszystkie symbole gramatyki posiadają wewnętrzne identyfikatory będące liczbami całkowitymi i na potrzeby komunikacji z użytkownikiem są tłumaczone przez warstwę pośrednią (jest to jedna z kluczowych funkcjonalności modułu sgcs.datalayer). W celu prawidłowej obsługi reguł przez poniższe narzędzie niezbędne jest zachowanie kontekstu reguł – umożliwia to plik zbioru uczącego. Jeżeli w jakikolwiek sposób zmodyfikujemy nasz zbiór uczący, dodając, usuwając, modyfikując lub zmieniając kolejność zdań, obiekty populacji mogą przestać być prawidłowo interpretowane. Dlatego też, jeżeli planujemy dokonać jakichkolwiek zmian w zestawie zdań uczących, to po ich dokonaniu należy ponownie wygenerować powiązane pliki \*.pop, lub też przynajmniej upewnić się, że ich zawartość po wczytaniu jest zgodna z naszymi oczekiwaniami.

Ustalenie kontekstu

W celu ustalenia kontekstu populacji należy nacisnąć przycisk Load i wybrać plik, co powinno poskutkować odblokowaniem przycisków Save i Open przy właściwie sformatowanym pliku danych wejściowych, lub komunikat o błędzie w przypadku wystąpienia błędu w trakcie parsowania. Bez ustalenia kontekstu populacji nie będzie możliwe jej wczytanie lub zapisanie, gdyż pliki populacji wykorzystują wewnętrzną reprezentację reguł i podczas odczytu oraz zapisu musi nastąpić tłumaczenie danych na przyjazne człowiekowi i odwrotnie.

Modyfikacje populacji

Populację można zmodyfikować przy pomocy przycisków Remove, Add new terminal rule oraz Add new non-terminal rule. W przypadku reguł nieterminalnych w każdym okienku należy podać identyfikator symbolu. Dla reguł terminalnych po lewej stronie produkcji podajemy identyfikator symbolu, po prawej zaś symbol terminalny. Identyfikator symbolu rządzi się następującymi prawami:

<identyfikator\_symbolu> := <identyfikator\_symbolu\_nieterminalnego> | <identyfikator\_symbolu\_specjalnego>

<identyfikator\_symbolu\_specjalnego> := ‘<S>' | ‘<U>’

<identyfikator\_symbolu\_nieterminalnego := <ciag\_wielkich\_liter>

<ciag\_wielkich\_liter> := <wielka\_litera> | <wielka\_litera> <ciag\_wielkich\_liter>

<wielka\_litera> := ‘A’...’Z’

<symbol\_terminalny> := dowolny ciąg znaków niebędących znakami białymi

Przykłady identyfikatorów symboli nieterminalnych:

A, H, AABC

Należy zauważyć, iż w populacji może istnieć więcej symboli, niż jest znaków alfabetu. W pierwszej kolejności algorytm wykorzystuje do reprezentacji pojedynczą wielką literę alfabetu (A, B i tak dalej), w dalszej kolejności następuje dołożenie kolejnej litery, lub zwiększenie litery na kolejnej pozycji (tj. mamy A, B, …, Y, Z, AA, AB, …, AY, AZ, BA, BB, …). Ludzka reprezentacja zezwala na dowolne symbole z powyższego zbioru, aczkolwiek należy pamiętać, że algorytm wykorzystuje podczas swojego działania jedynie k pierwszych symboli (gdzie k to ilość symboli terminalnych zdefiniowana w danych konfiguracyjnych \*.parconf ). Przy pomocy <S> oznaczamy startowy symbol gramatyki, przy pomocy <U> symbol uniwersalny.

Zapisywanie/otwieranie populacji

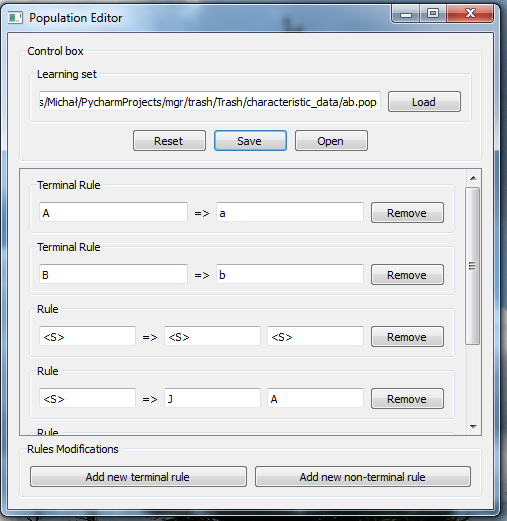
Przy pomocy przycisków Save / Open jest możliwe zapisanie lub wczytanie istniejącej populacji.

Ponowne generowanie plików populacji

Jeżeli dokonujemy zmian w zbiorze uczącym, możemy wygenerować populację ponownie, podejmując następujące kroki:

1. Wczytujemy stary zbiór uczący.
2. Otwieramy powiązany z nim plik populacji.
3. Wczytujemy nowy zbiór uczący.
4. Zapisujemy otwartą populację.

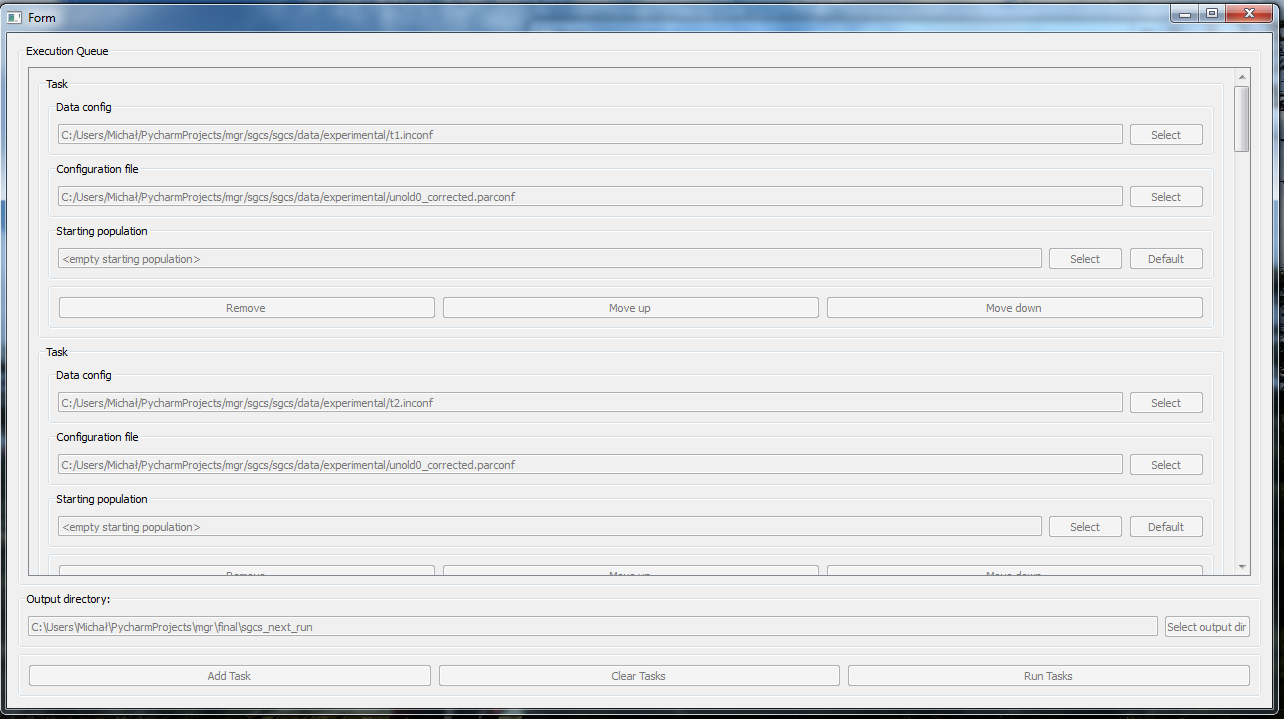
Tym sposobem jest możliwe dokonanie konwersji reguł populacji pomiędzy różnymi zbiorami uczącymi.



Rysunek 36: Edytor populacji

Scheduler

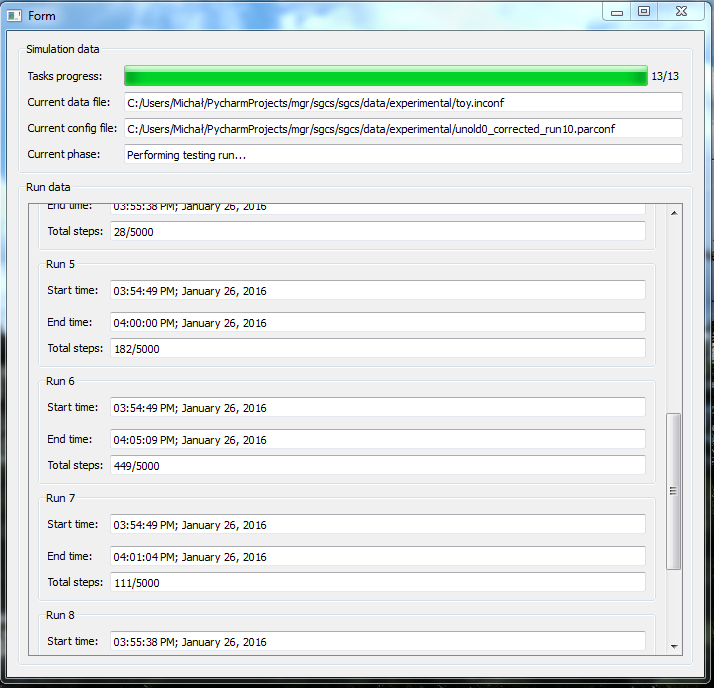
W menu tym (Rysunek 37) można przygotować zestaw badań do przeprowadzenia. Naciskając przycisk Add Task, dodajemy kolejne zadanie obliczeniowe. Przez zadanie rozumiemy pełne odpalenie algorytmu (cykl uczący i testowy) dla zadanego pliku danych wejściowych (skonstruowanego wcześniej przy pomocy Prepare input data), konfiguracji (przygotowanego przy pomocy Prepare configuration) i startowej populacji. Jeżeli w trakcie działania programu trzeba będzie coś zmienić, możliwe jest przerwanie wykonania, usunięcie obliczonych, poprawienie pożądanych rzeczy i wreszcie ponowne wystartowanie procesu obliczeń. Ścieżkę do folderu z danymi wynikowymi ustalamy poprzez naciśnięcie przycisku Select output dir. Folder ten zostanie w całości wyczyszczony na wczesnym etapie działania algorytmu, więc należy wybrać go z rozwagą w celu uniknięcia przypadkowego skasowania ważnych danych! Można również zmieniać kolejność wykonania zadań przy pomocy przycisków Move up, Move down.



Rysunek 37: Menu Scheduler

Runner

Zadaniem tego okna (Rysunek 38) jest poinformowanie użytkownika o obecnym stanie obliczeń algorytmu. W Simulation data można znaleźć szczegółowe informacje na temat obecnie realizowanego zadania, takie jak nr obecnie wykonywanego zadania, ilość wszystkich zadań, ścieżkę do pliku z danymi wejściowymi, konfiguracją oraz obecnie wykonywany etap prac (patrz omówienie modułu sgcs.gui.runner). W obszarze Run data znajduje się osobna rubryka dla każdego cyklu uczącego, z datą początkową, końcową oraz ilością wykonanych kroków ewolucyjnych. Kiedy wszystkie zadania zostaną wykonane, na ekranie pojawi się informujący nas o tym odpowiedni komunikat. Wszystkie artefakty generowane przez algorytm są zbierane na końcu każdego zadania, a następnie zapisywane do sprecyzowanego w menu Schedulera katalogu wyjściowego. Każde zadanie posiada osobny katalog numerowany w kolejności zgodnej z kolejnością zadań w Schedulerze.



Rysunek 38: Okno przebiegu algorytmu

1. Wykorzystanie console\_fetcher.py

Drugi skrypt uruchamiający aplikację został napisany podczas uruchamiania testów w chmurze. Jego powstanie było związane z jednej strony z trudnościami w postawieniu pełnego środowiska graficznego w chmurze oraz z drugiej strony w celu stworzenia prostego konsolowego interfejsu, który można by wykorzystywać w skryptach konsolowych. Jego składnia jest następująca:

console\_fetcher [-p] –o output\_dir [-s] [input\_file config\_file [starting\_population]]

Parametr –p oznacza, że nie algorytm nie uruchomi się, a jedynie wyśle na standardowe wyjście informację o przygotowanej konfiguracji (dzięki czemu możliwe jest wcześniejsze upewnienie się, że skrypt wywołano z właściwymi parametrami). Parametr –o jest obowiązkowy – po nim podajemy ścieżkę do folderu, w którym będą zapisywane artefakty. Następnie należy podać dowolną ilość naprzemiennie występujących plików z danymi wejściowymi i plików konfiguracji. Jeżeli chcielibyśmy dodatkowo podać konfigurację początkową, należy wówczas podać również flagę –s, która sprawi, że algorytm nie będzie konstruował zadań na podstawie dwójki, lecz trójki parametrów. Te trzy parametry to dane wejściowe, konfiguracja i startowa populacja. Jako że całość może wydawać się dosyć skomplikowana, poniżej zostało przedstawione kilka scenariuszy uruchomienia skryptu.

console-fetcher –o ~/data t1.inconf sgcs.parconf t2.inconf gcs.parconf

Powyższe polecenie uruchomi algorytm o konfiguracji sgcs.parconf (czyli najprawdopodobniej sgcs) dla zbioru tomita 1, a następnie gcs dla tomita 2. Artefakty z działania obu algorytmów zostaną zapisane w folderze ~/data.

console-fetcher –p –o ~/data t1.inconf sgcs.parconf t2.inconf gcs.parconf

Flaga –p spowoduje, że program wyświetli tylko informację o tym co by zrobił w przypadku uruchomienia tej samej komendy bez flagi –p.

console-fetcher –s –o ~/data t1.inconf sgcs.parconf t1.pop t2.inconf gcs.parconf t2.pop

Wymienione wcześniej algorytmy rozpoczną działanie odpowiednio z populacją początkową t1.pop oraz t2.pop .

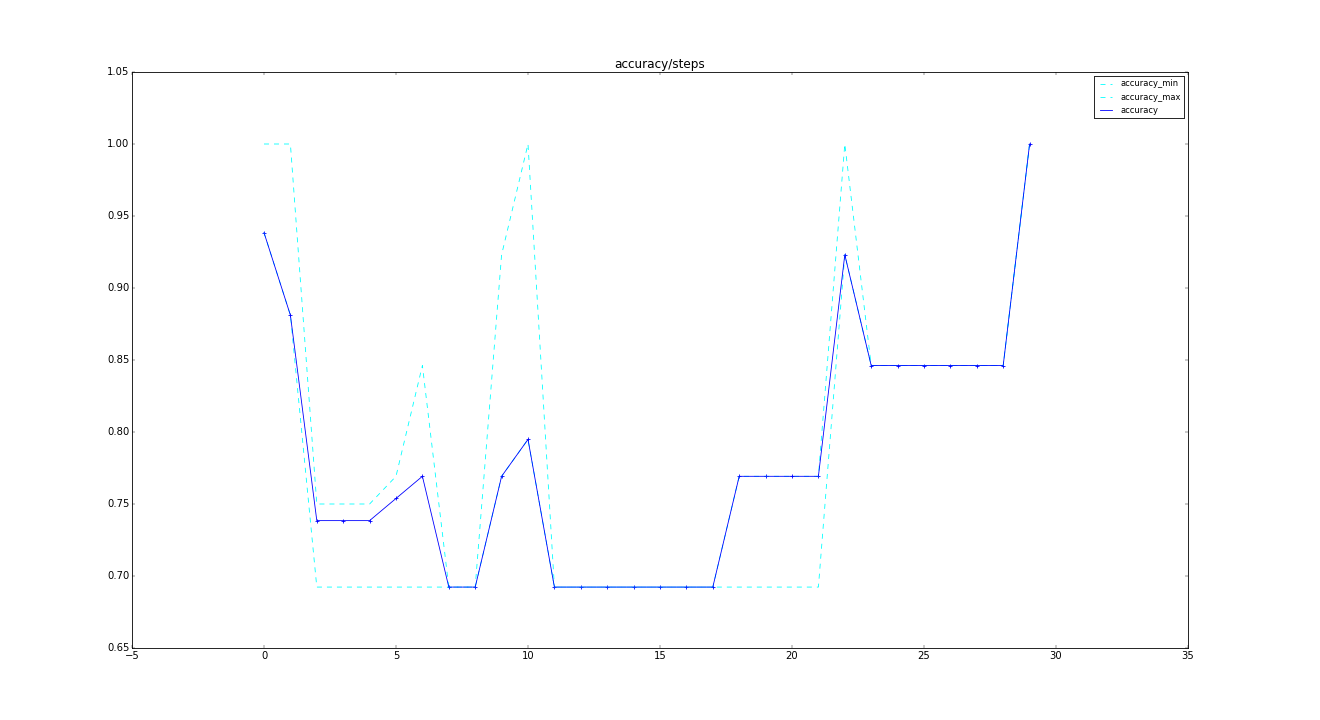
console-fetcher –s –o ~/data t1.inconf sgcs.parconf t1.pop t2.inconf gcs.parconf

Skrypt otrzymał złe argumenty (pomimo flagi –s drugie z planowanych uruchomień nie posiada jawnie podanego pliku populacji. Zostanie wyświetlone ostrzeżenie, po czym skrypt zakończy działanie.

1. Artefakty

Wynikiem działania algorytmu jest wygenerowanie katalogu dla każdego zdefiniowanego zadania. Poniżej została omówiona typowa zawartość takiego katalogu:

pliki \*.png (accurracy.png, fallout,png, falsediscovery.png, falseomisiion.png, missrate.png, negpredictive.png, precision.png, sensivity.png, specifity.png, fitness.png) – są to diagramy średnich dla wszystkich uruchomień cyklu uczącego miar macierzy pomyłek (ang. confussion matrix) od kroku ewolucyjnego (Rysunek 39). Można na nich znaleźć również wartość minimalną oraz maksymalną (przerywana linia).



Rysunek 39: Przykładowy diagram generowany przez system

final\_population.pop – W tych plikach znajduje się mało czytelny dla ludzkiego oka zapis stanu populacji wykorzystanej w cyklu testowym (Rysunek 40). Pliki są zapisywane w formacie json. Stosuje się w nich następującą reprezentację wewnętrzną:

* Całość jest listą;
* Pierwszy element listy to typ populacji reguł (stochastyczna lub nie);
* Pozostałe elementy to wszystkie reguły populacji o nieokreślonej kolejności, na każdą z nich składają się:
  + Nieznormalizowane prawdopodobieństwo (tylko w przypadku populacji stochastycznych);
  + Rodzic, lewe i prawe dziecko – rodzic zawsze jest symbolem nieterminalnym, lewe dziecko może być symbolem terminalnym lub też nie, prawe dziecko jest symbolem nieterminalnym lub wartością null.

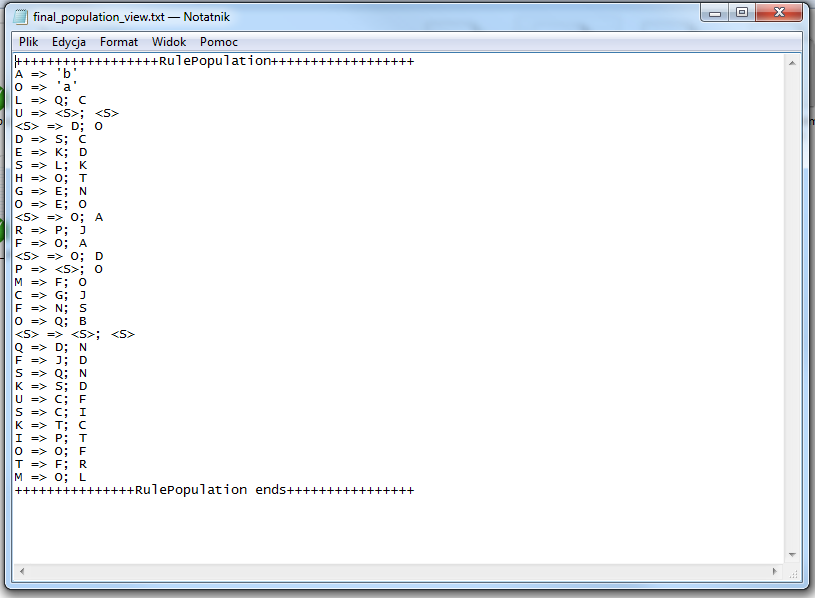
Symbol x zostanie zapisany w postaci pojedynczej liczby całkowitej y wg następujących zasad:

* y = -101 – occurence(x) dla symboli terminalnych, gdzie occurence(x) to wartość równa numerowi pierwszego wystąpienia symbolu terminalnego w zbiorze uczącym (np. jeżeli pierwsze zdanie w zbiorze uczącym to „a b a a c”, wówczas symbole a, b, c, otrzymają identyfikatory -101, -102, -103;
* -100 <= y <= 100, jeżeli x to symbol specjalny. W obecnej implementacji biblioteki istnieją tylko dwa symbole specjalne – startowy (y=1) oraz uniwersalny (y=2), reszta może zostać rozbudowana w przyszłości wraz z algorytmem;
* y = 101 + ord(x) jeżeli x to symbol nieterminalny. Przez ord(x) rozumiemy takie odwzorowanie identyfikatora symbolu x na zbiór liczb całkowitych, że A=1, B=2, itd.;



Rysunek 40:Przykładowy plik wynikowej populacji

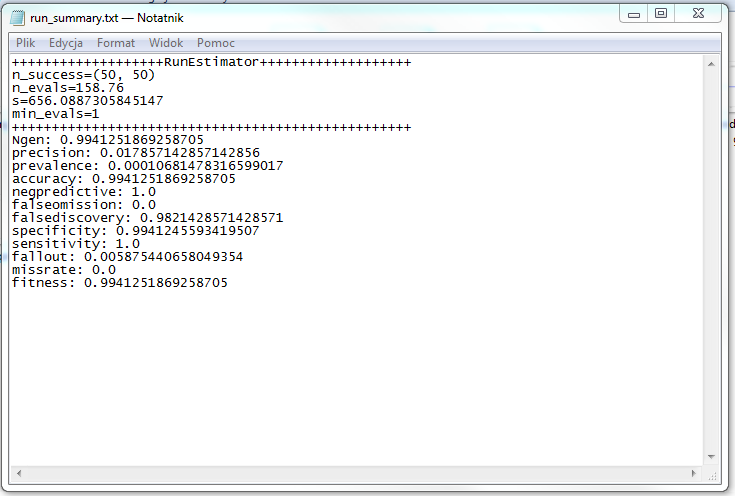
final\_population\_view.txt – Czytelny zapis populacji wykorzystanej w cyklu testowym (Rysunek 41). Dzięki temu plikowi jest możliwy szybki podgląd populacji wynikowej bez konieczności uruchamiania edytora populacji.



Rysunek 41: Przykładowy plik podglądu końcowej populacji

grammar\_estimator.grest – plik .json będący zrzutem obiektu oceny jakości procesu uczenia. Dane te to nic innego jak tekstowa reprezentacja diagramów macierzy omyłek.

run\_summary.txt – podsumowanie przebiegu algorytmu (Rysunek 42). Zawiera ilość uruchomień uczących zakończonych sukcesem (n\_success), średnią ilość kroków ewolucyjnych potrzebnych do znalezienia rozwiązania (n\_evals), odchylenie standardowe tejże (s) oraz minimum (min\_evals). Oprócz tego znajduje się tutaj macierz omyłek dla etapu testowego.



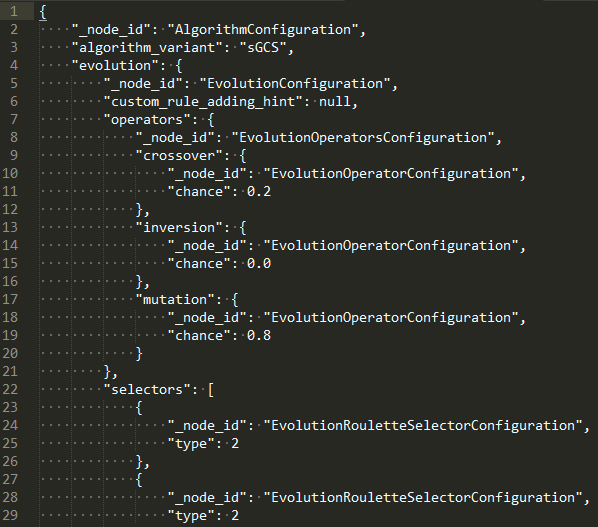
Rysunek 42: Przykładowy plik podsumowania przebiegu algorytmu

\*.inconf – plik danych wejściowych (Rysunek 43). Zawiera pełną ścieżkę do zbioru testowego oraz uczącego zapisane w formacie json.



Rysunek 43: Przykładowy plik danych wejściowych

\*.parconf – plik konfiguracji (Rysunek 44). Zawiera wartości wszystkich parametrów algorytmu zapisane w formacie json.



Rysunek 44: Przykładowy plik konfiguracji

1. Biblioteka

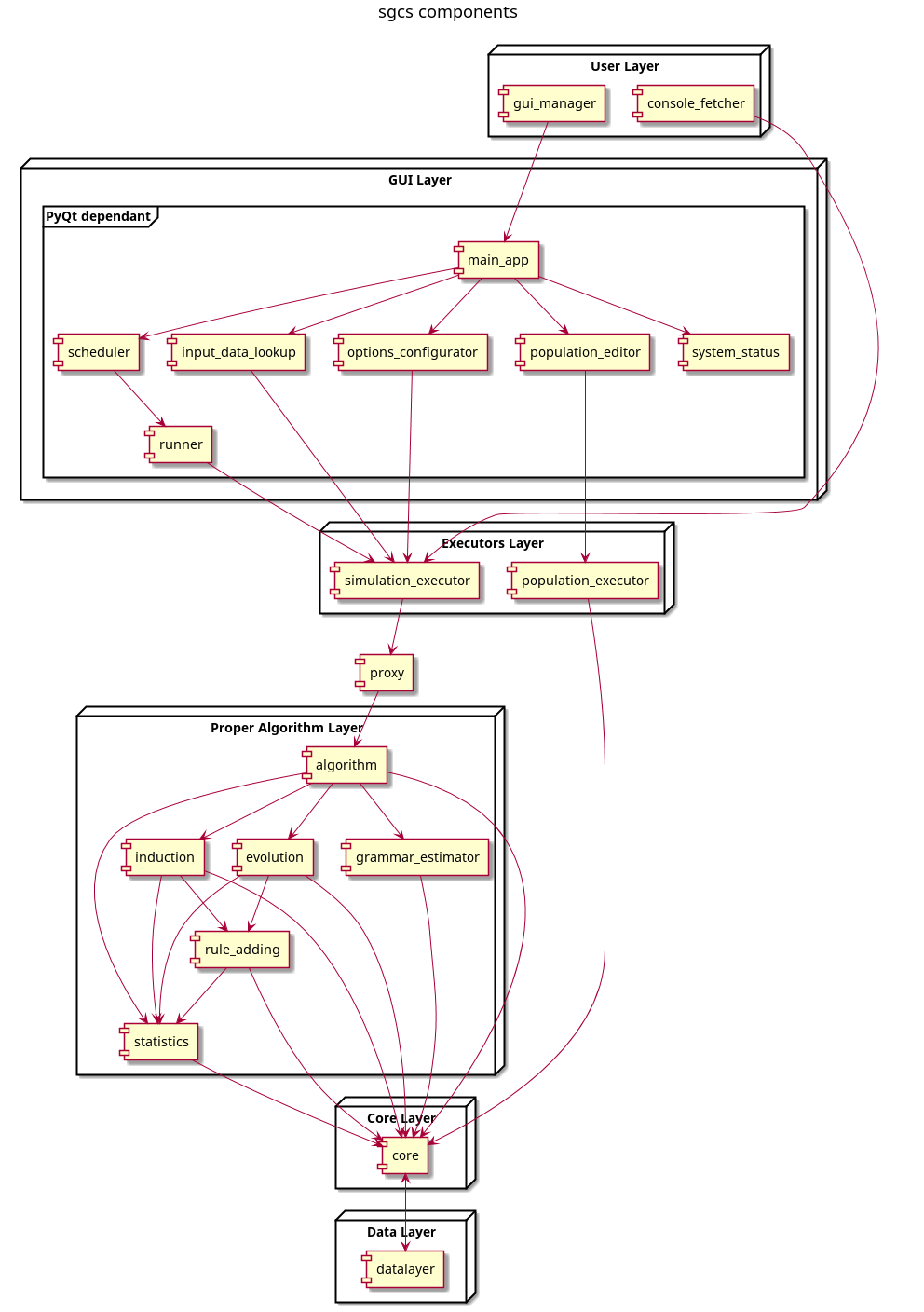
Na najwyższym poziomie projektu znajduje się plik „requirements.txt”. Zawiera on listę wszystkich bibliotek niezbędnych do dziania biblioteki, z wyjątkiem PyQt4, na którym oparto gui i który należy zainstalować osobno, w celu korzystania z elementów interfejsu graficznego.

Instalacji zależności wymienionych w pliku requirements.txt można dokonać następującym poleceniem:

sudo pip install –r requirements.txt

Dodatkowo zależnie od system jedna z wymienionych w tym pliku bibliotek (matplotlib) może wymagać dodatkowych bibliotek, które to zależności należy rozwiązać zgodnie z zaleceniami twórców biblioteki (na przykład na ubuntu polecenie sudo apt-get install python3-matplotlib rozwiązuje wszystkie problemy z zależnościami).

Warstwowa budowa biblioteki



Rysunek 45: Diagram modułowy biblioteki sgcs

W całej budowie biblioteki (Rysunek 45) można wyróżnić sześć warstw logicznych:

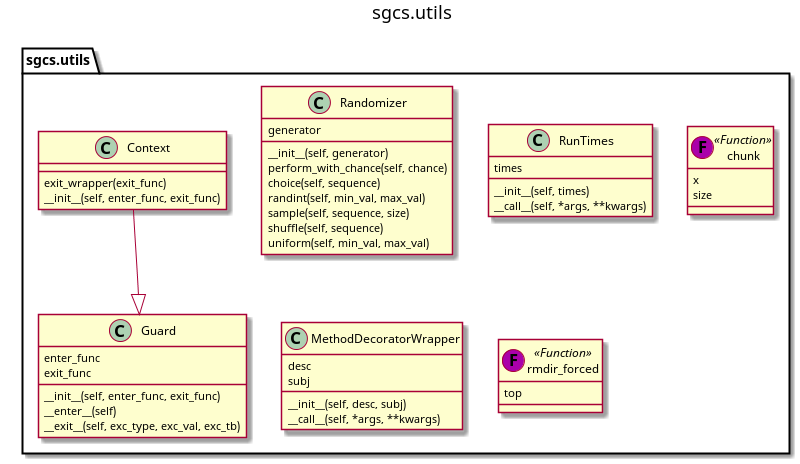
* Warstwa Użytkownika – pierwsza warstwa, na jaką natrafia użytkownik podczas uruchamiania aplikacji. Jest to skrypt inicjujący GUI lub aplikację konsolową;
* Warstwa GUI – warstwa z którą użytkownik aplikacji również ma bezpośrednią styczność (przynajmniej w przypadku aplikacji z interfejsem graficznym, w przypadku aplikacji konsolowej interfejsem użytkownika są parametry wywołania skryptu);
* Warstwa Wykonawców – jest to warstwa realizująca sporą część logiki biznesowej, czyli zbieranie artefaktów, edycję danych wejściowych algorytmu itp.;
* Warstwa Właściwego Algorytmu – warstwa zawierająca „czysty” algorytm GCS/sGCS/neg-sGCS, nieskażona logiką wyższych, ani niższych warstw;
* Warstwa Rdzenia – warstwa odpowiadająca za przechowywanie w pamięci najbardziej podstawowych elementów środowiska algorytmu, ściśle powiązana z warstwą niższą;
* Warstwa Danych – warstwa odpowiadająca za serializację oraz deserializację danych Warstwy Rdzenia.

Dodatkowo mamy jeszcze moduł sgcs.gui.proxy rozszerzający możliwości warstwy algorytmu w celu łatwiejszego wykorzystania go przez GUI. Moduł sgcs.core bywa też używany w niektórych klasach GUI, co trochę przełamuje zaprezentowany diagram zależności, ale chroni przed wprowadzaniem niepotrzebnych warstw abstrakcji. Oprócz tego w bibliotece występują moduły pomocnicze wykorzystywane praktycznie w każdym innym module – z tego powodu nie zostały naniesione na diagram.

Moduły pomocnicze

Biblioteka posiada zbiór kilku klas oraz funkcji wykorzystywanych praktycznie w całej aplikacji. Zostały one zebrane w jeden wygodny w wykorzystaniu moduł – sgcs.utils. Praktycznie każdy moduł wykorzystuje sporą część zawartych w nim funkcjonalności.

sgcs.utils (Rysunek 46)



Rysunek 46: Moduł sgcs.utils

Randomizer

Klasa zapewniająca warstwę abstrakcji pomiędzy całą biblioteką a standardową klasą biblioteczną random.Random. Dzięki temu jest możliwe wykorzystanie innej klasy zapewniającej dostęp do lepszej maszyny losowej (na przykład klasę wykorzystującą specyficzną architekturę maszyny, na której uruchamiamy naszą aplikację) albo łatwe uruchomienie całej aplikacji z dokładnie tym samym seedem, co może być niezwykle przydatne w przypadku rozwijania oraz debugowania algorytmu w znacznym stopniu opartego w końcu na losowości.

generator – pole przechowujący prawdziwy obiekt generatora (np. instancję klasy random.Random)

\_\_init\_\_(self, generator)\_\_ - konstruktor pobierający instancję obiektu prawdziwego generatora, dla którego klasa ma być fasadą.

perform\_with\_chance(self, chance) – metoda zwraca True, jeżeli liczba wylosowana z przedziału [0,1] <= chance, False w pp.

choice(self, sequence) – metoda zwracająca losowy obiekt z sekwencji sequence (wszystkie mają równą szansę na bycie wylosowanym).

randint(self, min\_val, max\_val) – zwraca losową liczbę całkowitą z przedziału [min\_val, max\_val].

sample(self, sequence, size) – zwraca size obiektów z sekwencji sequence bez powtórzeń.

shuffle(self, sentence) – zwraca kopię sekwencji sequence, nadając jej losowy porządek.

uniform(self, min\_val, max\_val) – zwraca losową rzeczywistą liczbę zmiennoprzecinkową z przedziału [min\_val, max\_val].

RunTimes

Klasa udostępniająca cukier syntaktyczny na wywołania typu:

run\_one\_more\_time = RunTimes(1)

while remainder > self.ALPHABET\_SIZE or run\_one\_more\_time():

chunk(x, size)

Funkcja dzieląca x na listy długości size, a następnie zwracająca taką listę list. Jeżeli x % size != 0 wówczas ostatni chunk zostanie dopełniony elementami z początku x.

Guard

Cukier syntaktyczny umożliwiający łatwe tworzenie zasobów wykorzystywanych w popularnej w języku Python bloku kontekstu with/as.

enter\_func – delegat wywołany przy wejściu w blok with/as.

exit\_func – delegat wywoływany przy opuszczeniu bloku with/as.

Context

Klasa dziedzicząca po klasie Guard, zapewniającą nieco cukru syntaktycznego na funkcję wyjścia (o ile nie ma potrzeby korzystania z zaawansowanych właściwości bloku with/as, i potrzebny jest prosty strażnik, który po prostu wywoła funkcję na początku i końcu bloku, wówczas powinno korzystać się z tej klasy zamiast klasy Guard.

MethodDecoratorWrapper

Wrapper przydatny przy tworzeniu dekoratorów metod potrzebujących dostępu do instancji obiektu, którego metodę dekorują.

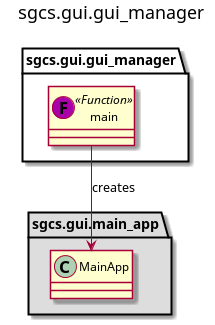
rmdir\_forced(top)

Funkcja usuwająca rekurencyjnie cały folder wraz z jego zawartością.

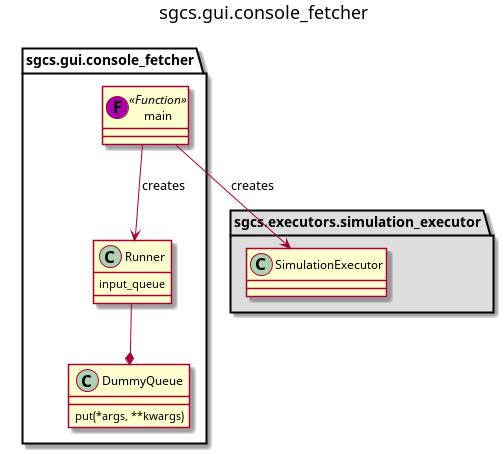
GUI

Ponad całą biblioteką znajduje się dwa proste interfejsy umożliwiające użytkownikowi przetestowanie możliwości biblioteki zaraz po dostarczeniu i wykorzystanie w pełni jej możliwości. Jest to wspominany już w tej pracy konsolowy skrypt console\_fetcher.py oraz utworzone w PyQt GUI uruchamiane przy pomocy skryptu gui\_manager.py. Poniżej znajdują się diagramy zależności obu skryptów.

sgcs.gui.gui\_manager (Rysunek 47)

Skrypt gui\_manager.py powoduje inicjalizację środowiska PyQt, a następnie uruchomienie widgetu MainApp, odpowiadającego za obsługę opisanego wcześniej menu głównego aplikacji. Następuje oddelegowanie kontroli do MainApp, powraca ona do gui\_manager.py dopiero podczas czyszczenia środowiska PyQt podczas finalizowania działania aplikacji.

Rysunek 47: Moduł sgcs.gui.gui\_manager

sgcs.gui.console\_fetcher.py (Rysunek 48)

Rysunek 48: Moduł sgcs.gui.console\_fetcher

W przypadku skryptu console\_fetcher.py postawiono na minimum funkcjonalności, jego uruchomienie powoduje natychmiastowe uruchomienie algorytmu. Z tego też powodu jest potrzeba istnienia uproszczonych klas Runner oraz DummyQueue.

Runner

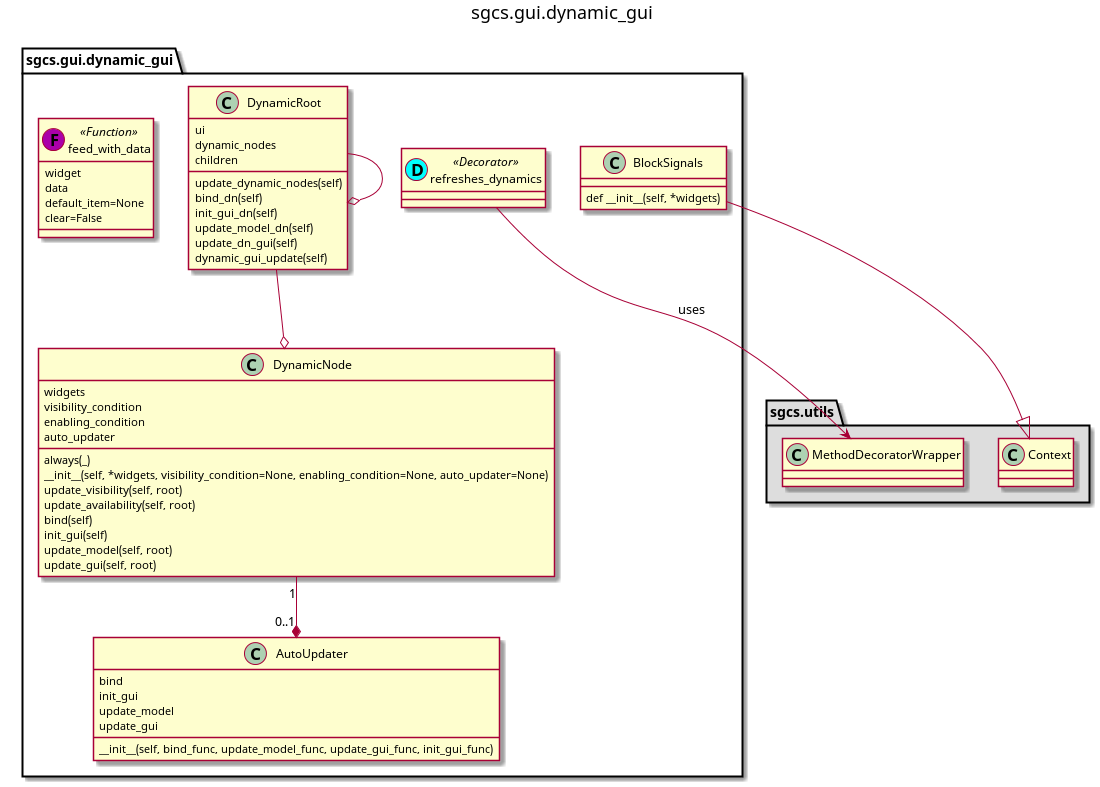
Klasa ta służy do przechowywania oraz ustalenia kolejki zadań, które mają być wykonane przez naszą bibliotekę. W przypadku console\_fetchera nie mamy do czynienia z interaktywną konfiguracją listy zadań (która to jest generowana automatycznie na podstawie podanej listy argumentów wywołania skryptu), toteż klasa ta jest wydmuszką i odpowiedzialność na wygenerowaniu listy zadań spoczywa na funkcji main. Oprócz tego zawiera ona pole input\_queue, również wydmuszki, która w przypadku interaktywnego gui jest wykorzystywana przez podprocesy algorytmu do zwracania szczegółowych danych diagnostycznych, niewykorzystywanych w przypadku uruchomienia aplikacji w trybie chmury.

DummyQueue

Klasa ta jest wydmuszką klasy multiprocessing.Queue, umożliwiającej asynchroniczną wymianę informacji pomiędzy procesami aplikacji. W tej aplikacji można rozróżnić dwa systemy przesyłania danych z procesów – jako wartość zwracana poleceń imap oraz właśnie przy pomocy tej kolejki. Ponieważ jednak przy pomocy kolejki wysyłamy jedynie informacje diagnostyczne, które mają większą rację bytu przy zastosowaniu gui (nieobecnego w przypadku uruchomienia aplikacji w trybie chmury), więc DummyQueue implementuje wyłącznie pustą metodę put (wszystkie informacje diagnostyczne są ignorowane).

sgcs.gui.dynamic\_gui (Rysunek 49)

Moduł sgcs.gui.dynamic\_gui pełni w zasadzie tą samą rolę dla wszystkich powiązanych z gui modułów bilbioteki co utils dla wszystkich modułów biblioteki – jego zadaniem jest zgromadzenie całego wspólnego kodu dla wszystkich modułów GUI. Większość klas modułu skupia się według idei dynamicznego gui – umożliwiają one tworzenie prostego wiązania danych, co umożliwia programowanie z wykorzystaniem wzorca projektowego Model-View-Viewmodel, domyślnie niewspieranego przez bibliotekę PyQt (istnieje wsparcie dla tegoż wzorca w przypadku kilku bardziej skomplikowanych widgetów, jednakże oddzielenie logiki od implementacji zgodnie z zalecaniami MVVM nie jest możliwe w przypadku prostszych widgetów bez zastosowania dodatkowego kodu.

feed\_with\_data(widget, data, default\_item=None, clear=False)

Rysunek 49: Moduł sgcs.gui.dynamic\_gui

Funkcja ta spowoduje wypełnienie widgetu dysponującego metodami clear(self), addItems(self, items) oraz setCurrentIndex(self, index) danymi data. Można wykorzystać tą metodę na przykład w celu utworzenia comboboxa na podstawie listy napisów. Pole default\_item umozliwia sprecyzowanie domyślnie zaznaczonej opcji. Podanie opcjonalnego argumentu clear z wartością False spowoduje wyczyszczenie dotychczasowej zawartości widgetu. Funkcja ta zakłada przechowywanie przez widget danych w porządku i może nie działać prawidłowo, jeżeli widget nie zapewnia kolejności pozycji.

refreshes\_dynamics

Dekorator, którym można udekorować metodę klasy dziedziczącej po klasie DynamicRoot. Funkcja w ten sposób udekorowana będzie zawsze pod koniec swojego działania wołać metodę DynamicRoot.update\_dynamic\_nodes(self) w celu odświeżenia dynamicznych elementów gui.

BlockSignals

Obiekt kontekstu (można go wykorzystać w bloku with/as), który powoduje zablokowanie wysyłania sygnałów PyQt przez podane widgety (przydatne przy modyfikowaniu wartości, które mogłyby zostać potraktowane jako akcja użytkownika w przypadku gdy nie jest to pożądane zachowanie. BlockSignals jest bezpieczne ze względu na wielokrotne wywołanie (tj. jeżeli wewnątrz bloku A BlockSignals znajdzie się blok BlockSignals B, oba blokujące ten sam widget x, wówczas x zostanie prawidłowo odblokowany pod koniec najbardziej zewnętrznego bloku A, a nie po opuszczeniu bloku B).

\_\_init\_\_(self, \*widgets) – Konstruktor, wszystkie widgets zostaną zablokowane.

DynamicRoot

Klasa realizująca bindowanie widoków View do obiektów ViewModel.

ui – pole przechowujące obiekt View.

dynamic\_nodes – pole przechowujące listę tzw. dynamicznych węzłów, umożliwiających sprecyzowanie logiki interfejsu bez konieczności modyfikowania jego samego.

children – jest możliwe podczepianie jednego obiektu typu DynamicRoot do drugiego jako dziecko w celu utworzenia struktury drzewiastej. Dzięki temu żądając odświeżenia gui lub modelu korzenia takiej struktury w efekcie nastąpi odświeżenie jej w całości. Ten mechanizm działa tylko w przypadku struktur drzewiastych i nie jest zabezpieczony przed wystąpieniem cyklu!

update\_dynamic\_nodes(self) – powoduje odświeżenie widoczności oraz dostępności gui na podstawie wszystkich dynamicznych węzłów.

bind\_dn(self) – powoduje podpięcie logiki wystawionej przez metody bind(self) wszystkich dynamicznych węzłów.

init\_gui\_dn(self) – powoduje inicjację GUI poprzez wywołanie wszystkich metod init\_gui(self) dynamicznych węzłów.

update\_model\_dn(self) – powoduje odświeżenie modelu na podstawie obecnego stanu gui (czyli dokładnie wywołanie metod update\_model() wszystkich dynamicznych węzłów).

update\_dn\_gui(self) – powoduje odświeżenie gui na podstawie obecnego stanu modelu, blokując przy tym wszystkie sygnały PyQt na czas procesu.

dynamic\_gui\_update(self) – powoduje odświeżenie widoczności, dostępności oraz stanu GUI na podstawie obecnego modelu.

DynamicNode

Klasa umożliwiająca zdefiniowanie części lub całości logiki obsługi widgetu. Może ona przejąć kontrolę przede wszystkim nad widocznością widgetu, jego dostępnością oraz powiązać go z konkretnym viewmodelem. Jeden obiekt DynamicNode reprezentuje jakiś fragment logiki, w efekcie są one powiązane z widgetami relacją wiele do wielu – jeden dynamiczny węzeł może opisywać zachowanie wielu widgetów, ale równocześnie kilka innych węzłów może skupiać się na pozostałych aspektach jednego z tych widgetów itd.

widgets – pole precyzujące widgety, których widoczność i dostępność będzie kontrolowana przez ten obiekt.

visibility\_condition – delegat określający warunki widoczności widgetów.

enabling\_condition – delegat określający warunki dostępności widgetów.

auto\_updater – instancja obiektu AutoUpdater, zajmującego się powiązaniem modelu z widokiem.

always(\_) - pomocniczy delegat zwracający zawsze True.

\_\_init\_\_(self, \*widgets, visibility\_condition=None, enabling\_condition=None, auto\_updater=None) – konstruktor. Jeżeli nie poda się któregoś z parametrów opcjonalnych, wówczas ten aspekt zostanie zignorowany przez ten konkretny węzeł (na przykład niesprecyzowanie delegata visibility\_condition sprawi, że węzeł ten nie będzie zarządzać widocznością widgetów).

AutoUpdater

Klasa grupująca funkcje bind, init\_gui, update\_model oraz update\_gui w jednym miejscu.

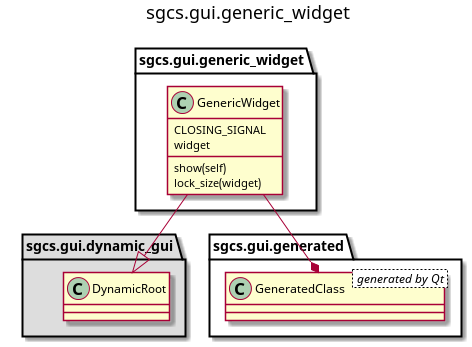
bind – delegat zawierający logikę, która powinna zostać wykonana na późnym etapie tworzenia gui (np. rejestracja na sygnały).

init\_gui – delegat zawierający logikę inicjującą GUI przy tworzeniu interfejsu po raz pierwszy.

update\_model – delegat odświeżający model na podstawie obecnego stanu GUI.

update\_gui – delegat odświeżający GUI na podstawie obecnego stanu modelu.

sgcs.gui.generic\_widget (Rysunek 50)



Rysunek 50: Moduł sgcs.gui.generic\_widget

Klasa GenericWidget stanowi rozszerzenie klasy DynamicRoot o kilka przydatnych mechanizmów. Po pierwsze stanowi pomost pomiędzy klasami wygenerowanymi przy pomocy QtDesignera a zdefiniowanymi przez użytkownika widokami. Po drugie nadpisuje obsługę zamknięcia okna w taki sposób, żeby został wysłany sygnał informujący o tym wydarzeniu (brakujące zachowanie w obecnej wersji PyQt). Po trzecie umożliwia łatwą blokadę rozmiarów okna (kolejna czynność, której nie da się wykonać w PyQt za pomocą jednej metody).

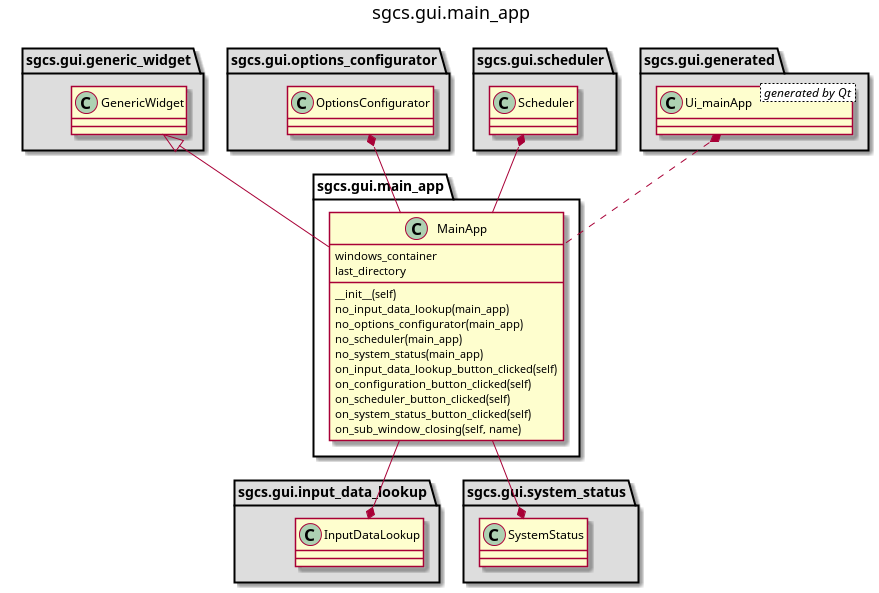
CLOSING\_SIGNAL – identyfikator sygnału wysyłanego podczas zamykania okna. Jedyny argument sygnału to nazwa klasy widgeta, który jest właśnie zamykany.

widget – obiekt QtGui.QtWidget.

show(self) – metoda umożliwiająca pokazanie widgetu

lock\_size(widget) – statyczna metoda umożliwiająca zablokowanie możliwości zmiany rozmiarów okna.

sgcs\_main\_app (Rysunek 51)



Rysunek 51: Moduł sgcs.gui.main \_app

Klasa odpowiadająca za wyświetlenie i obsługę menu głównego. Wszystkie otwierane przez nią okna są singletonami – w danej chwili może być otwarte wyłącznie jedno okno danego typu, nie możemy na przykład uruchomić dwóch okien statusu systemu równocześnie. Jest to związane z wykorzystaniem specyficznych zasobów systemowych przez większość okien, co sprawia, że utworzenie ich kilku instancji mogłoby prowadzić do dziwnych problemów (i tylko w kilku przypadkach miałoby jakiś sens).

windows\_container – fabryka okien zajmująca się kontrolą ich czasu życia

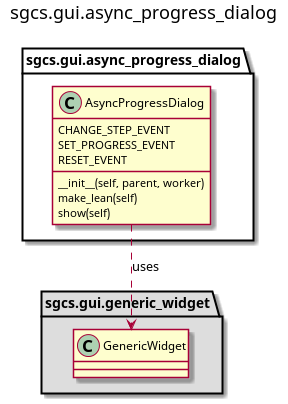
last directory – zmienna przechowująca ostatni otwarty w jakiejkolwiek części aplikacji folder. Ułatwia to powrót do ostatnio używanych danych (nie musimy za każdym razem szukać tego samego folderu).

no\_input\_data\_lookup/no\_options\_configurator/no\_scheduler/no\_system\_status – funkcje sprawdzające czy dane okno zostało już stworzone.

on\_input\_data\_lookup\_button\_clicked/on\_configuration\_button\_clicked/on\_scheduler\_button\_clicked/on\_system\_status\_button\_clicked – funkcje obsługujące otwieranie okien poszczególnych aplikacji.

on\_sub\_window\_closing – funkcja obsługująca proces zamknięcia okna aplikacji, nasłuchuje na GenericWidget.CLOSING\_SIGNAL.

sgcs.gui.async\_progress\_dialog (Rysunek 52)



Rysunek 52: Moduł sgcs.gui.async\_progress\_dialog

Klasa umożliwiająca wyświetlenie maksymalnie odchudzonego dialogu z komunikatem oraz paskiem postępu. Po wyświetleniu go można sterować jego zachowaniem przy pomocy odpowiednich sygnałów.

CHANGE\_STEP\_EVENT – identyfikator sygnału wysyłanego w celu zmiany tekstu obecnego komunikatu.

SET\_PROGRESS\_EVENT – identyfikator sygnału wysyłanego w celu zmiany wartości obecnego postępu.

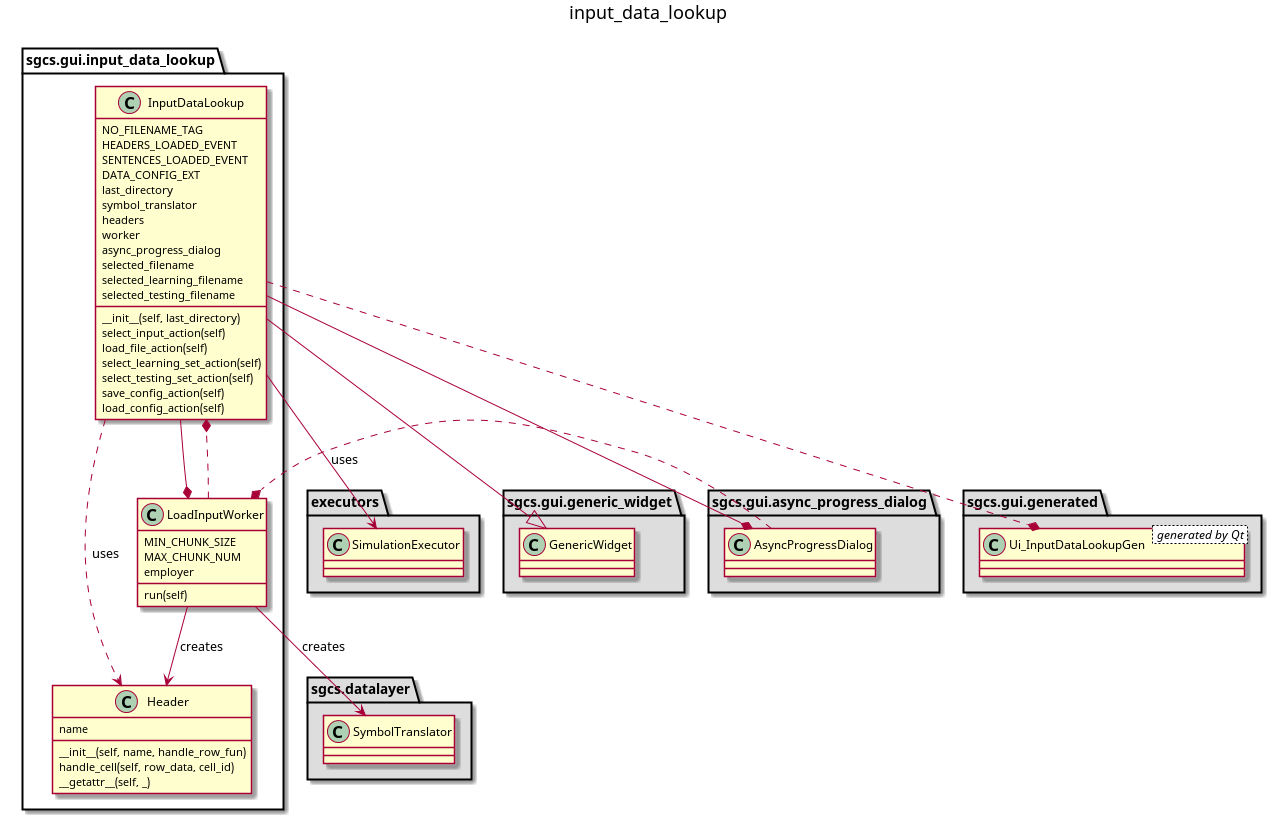
RESET\_EVENT – identyfikator sygnału powodującego schowanie/ponowne wyświetlenie okna dialogowego.

\_\_init\_\_(self, parent, worker) – jako parent należy podać okno – rodzica komunikatu. Worker jest obiektem, który będzie emitować sygnały (na przykład osobny wątek).

make\_lean(self) - „odchudza” komunikat.

show(self) – powoduje wyświetlenie komunikatu.

sgcs.gui.input\_data\_lookup (Rysunek 53)



Rysunek 53: Moduł sgcs.gui.inout\_data\_lookup

Moduł ten zawiera klasy niezbędne do obsługi podglądu plików z danymi wejściowymi do gramatyk oraz tworzenie plików konfiguracyjnych danych wejściowych \*inconf.

InputDataLookup

Główna klasa odpowiedzialna za obsługę menu „Prepare input data”.

NO\_FILENAME\_TAG – napis widoczny do czasu wyboru pliku.

HEADERS\_LOADED\_EVENT – identyfikator sygnału wysyłanego, gdy już znamy wszystkie nagłówki kolumn tabeli.

SENTENCES\_LOADED\_EVENT – identyfikator sygnału wysyłanego, gdy tabela została załadowana i jest już gotowa do użytku.

DATA\_CONFIG\_EXT – stała przechowująca rozszerzenie plików konfiguracyjnych danych wejściowych („.inconf”).

last\_directory – ostatni otwarty folder.

symbol\_translator – obiekt wykorzystywany do wczytania danych z dysku (obiekt typu SymbolTranslator).

headers – nagłówki tabeli.

worker – uchwyt wątku, któremu zleca się wczytanie pliku.

async\_progress\_dialog – uchwyt okna z postępem wczytywania wyświetlanego w trakcie ładowania danych wejściowych.

selected\_filename – wybrany plik danych wejściowych.

selected\_learning\_filename – ścieżka do wybranego zestawu uczącego.

selected\_testing\_filename – ścieżka do wybranego zestawu testowego.

\*\_action(self) – funkcje obsługujące naciśnięcie odpowiedniego przycisku.

LoadInputWorker

Wątek zajmujący się wczytaniem danych do tabeli. Wbrew pozorom nie jest to zadanie aż tak trywialne, jak mogłoby się wydawać – mimo wykonywania najbardziej czasochłonnej operacji, czyli ładowania danych z dysku do pamięci w osobnym wątku to w przypadku dużych danych wejściowych samo dodawanie danych do widgetu tabeli może spowodować zawieszenie się głównego okna. Dlatego też worker ten dokonuje podziału danych wejściowych na chunki odpowiedniego rozmiaru, po czym wczytuje je jeden po drugim. Chunk ma rozmiar co najmniej równy MIN\_CHUNK\_SIZE i równocześnie taki, żeby ilość chunków nie przekroczyła MAX\_CHUNK\_NUM. Worker na czas swojej pracy powoduje również wyświetlenie komunikatu ze statusem prac.

MIN\_CHUNK\_SIZE – minimalny rozmiar chunka.

MAX\_CHUNK\_NUM – maksymalna ilość chunków, jakie mogą powstać.

employer – uchwyt zleceniodawcy zadania, czyli InputDataLookup.

run(self) – główna metoda workera, w celu wykonania asynchronicznego nie należy wołać jej bezpośrednio, lecz zamiast tego QtCore.Qthread.start.

Header

Klasa zajmująca się obsługą prawidłowego wyświetlania zawartości danej kolumny tabeli.

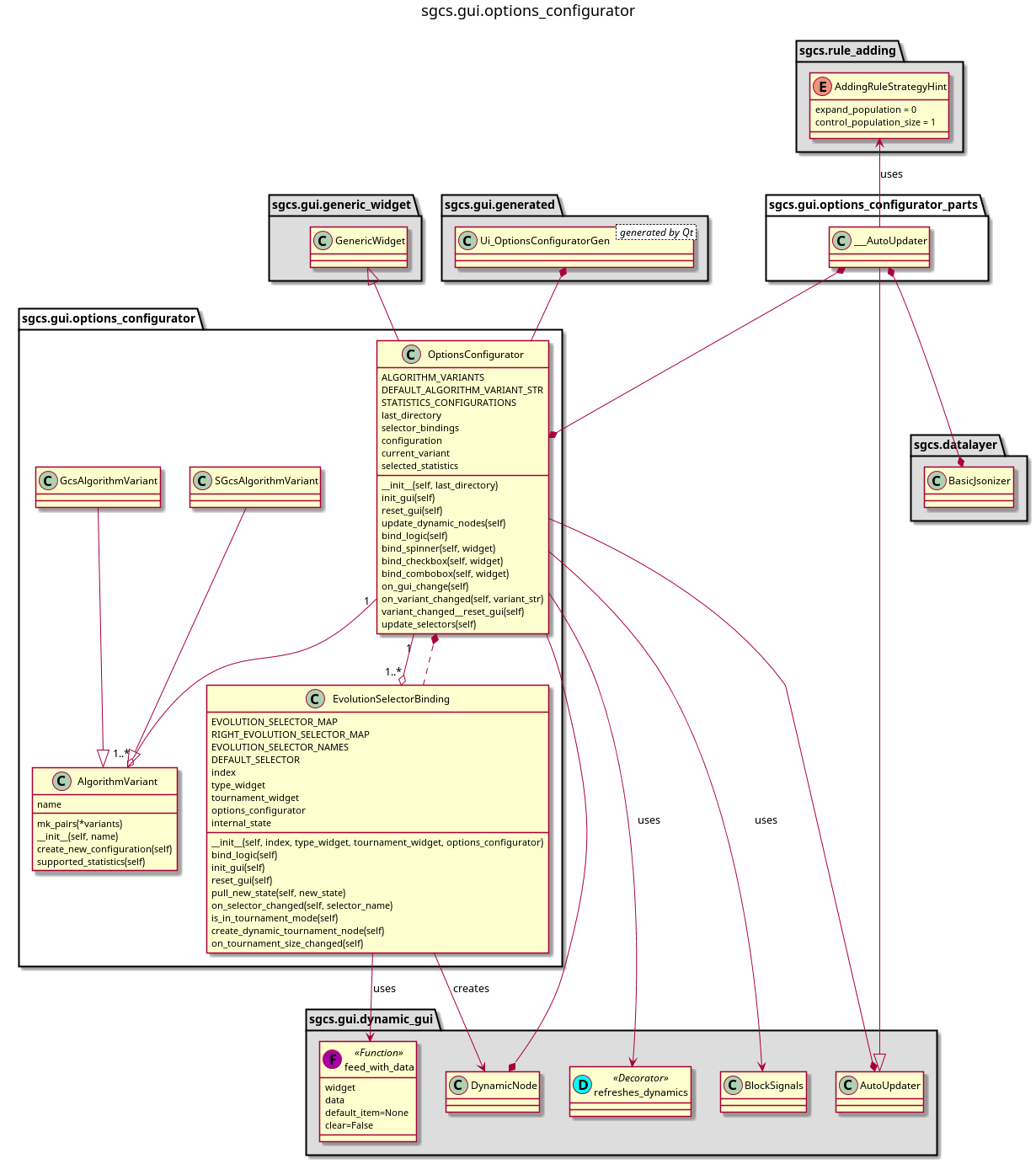
name – nagłówek danej kolumny.

\_\_init\_\_(self, name, handle\_row\_fun) - konstruktor, należy przekazać mu nagłówek kolumny name oraz delegata zdolnego do odpowiedniego wyświetlenia zawartości danej kolumny.

handle\_cell(self, row\_data, cell\_id) – funkcja zajmująca się prawidłowym wypełnieniem danej komórki tabeli.

sgcs.options\_configurator i sgcs.options\_configurator\_parts (Rysunek 54)

Rysunek 54: Moduły sgcs.options\_configurator i sgcs.options\_configurator\_parts

Widok „Prepare configuration” jest najbardziej skomplikowaną częścią GUI, co widać, chociażby po jego diagramie klas. Diagram ten byłby jeszcze bardziej rozbudowany, gdyby nie zwijać wszystkich składowych obiektów AutoUpdater z przestrzeni nazw sgcs.gui.options\_configurator\_parts w jedną klasę \_\_\_AutoUpdater. Jest to spowodowane niezwykłym rozbudowaniem tego menu (ma za zadanie obsłużyć dobór około 30 parametrów algorytmu), dynamiczną budową (nie wszystkie pola są zawsze dostępne, istnieją zależności pomiędzy niektórymi parametrami) oraz dosyć wczesnym napisaniem go, kiedy jeszcze struktura GUI nie była aż tak uporządkowana.

AlgorithmVariant

Abstrakcyjna klasa stosowana do rozróżnienia pomiędzy dostępnymi algorytmami na poziomie GUI. Obecnie posiada dwie konkretne implementacje: *GcsAlgorithmVariant* oraz *SgcsAlgorithmVariant*.

name – wyświetlana nazwa algorytmu

mk\_pairs(\*variants) – statyczna metoda tworząca mapę variant.name: variant.

create\_new\_configuration – abstrakcyjna metoda, której wywołanie powinno spowodować utworzenie domyślnego obiektu konfiguracji dla danego wariantu algorytmu.

supported\_statistics – lista nazw funkcji fitness, które są wspierane przez dany algorytm. Od doboru funkcji fitness (wykorzystującej lub nie zdania negatywne) zależy również czy proces uczenia będzie przebiegał również na zdaniach negatywnych.

\_\_\_AutUpdater

Grupa klas dziedziczących po klasie AutoUpdater umiejscowionych w przestrzeni nazw sgcs.gui.options\_configurator\_parts, jak sama nazwa przestrzeni wskazuje, są klasami przechowującymi fragmenty logiki menu „Prerare Configuration”. W przestrzeni tej możemy odnaleźć następujące klasy:

* ClassicalStatisticsAutoUpdater – klasa obsługująca część menu odpowiedzialną za klasyczną funkcję płodności;
* CrowdingAutoUpdater – klasa obsługująca część menu odpowiedzialną za funkcjonowanie ścisku;
* ElitismAutoUpdater – klasa obsługująca część menu odpowiedzialną za funkcjonowanie elityzmu;
* EvolutionAutoUpdater – klasa obsługująca część menu odpowiedzialną za parametry operatorów ewolucyjnych;
* InductionAutoUpdater – klasa obsługująca część menu odpowiedzialną za parametry operatorów pokrycia oraz korekcję gramatyki;
* MainBoxAutoUpdater – klasa obsługująca część menu odpowiedzialną za wczytywanie i zapisywanie konfiguracji oraz wybór wariantu algorytmu;
* RootAutoUpdater – klasa obsługująca część menu odpowiedzialną za ogólne parametry algorytmu (ilość uruchomień cyklu uczacego, satysfakcjonujący poziom przystosowania, rodzaj funkcji przystosowania);
* RulesAutoUpdater – klasa obsługująca część menu odpowiedzialną za ogólny stan populacji reguł (maksymalny rozmiar populacji, ilość symboli nieterminalnych itd.).

Każda z tych klas pilnuje, aby parametry, za które jest odpowiedzialna, były zgodne z tym, co jest widoczne w danym momencie w menu. Dbają również o widoczność i dostępność części menu (na przykład chowając część dotyczącą parametryzacji klasycznej funkcji fitness, jeżeli zdecydujemy się na inną niż klasyczna funkcję fitness).

EvolutionSelectorBinding

Jest to klasa zapewniająca obsługę części menu związanej z doborem selektorów algorytmu genetycznego. Jest to jeden z protoplastów architektury DynamicRoot, który nie został ze względu na swoją skomplikowaną logikę oraz brak czasu w pełni przetłumaczony na nowy mechanizm. Zadaniem tej klasy jest obsługa wyboru selektorów algorytmu genetycznego oraz udostępnienie pól dodatkowych parametrów, jeżeli dany selektor takie parametry posiada.

EVOLUTION\_SELECTOR\_MAP – mapa odwzorowująca nazwę na węzeł konfiguracyjny selektora.

RIGHT\_EVOLUTION\_SELECTOR\_MAP – mapa odwzorowująca węzeł konfiguracyjny węzeł selektora na nazwę.

EVOLUTION\_SELECTOR\_NAMES – lista kluczy EVOLUTION\_SELECTOR\_MAP.

DEFAULT\_SELECTOR – nazwa selektora domyślnego.

index – nr selektora, za który odpowiada instancja.

type\_widget – uchwyt do pola combobox odpowiedzialnego za wybór typu selekcji.

tournament\_widget – uchwyt do pola spinner odpowiedzialnego za określenie rozmiaru turnieju.

options\_configurator – uchwyt do głównej klasy okna konfiguracji.

internal\_state – przechowywana węzeł konfiguracyjny selektora.

bind\_logic(self) – metoda odpowiedzialna za rejestrację funkcji obsługujących wydarzenia oraz późną inicjalizację niektórych wartości pól.

init\_gui(self) – metoda odpowiedzialna za inicjalizację menu (dostępne pola oraz zakres wartości).

reset\_gui(self) – funkcja odświeżania GUI.

pull\_new\_state(self, new\_state) – metoda obsługująca zmianę selektora z równoczesnym odświeżeniem menu (tj. zmianę stanu selektora na przykład poprzez wczytanie istniejącej konfiguracji).

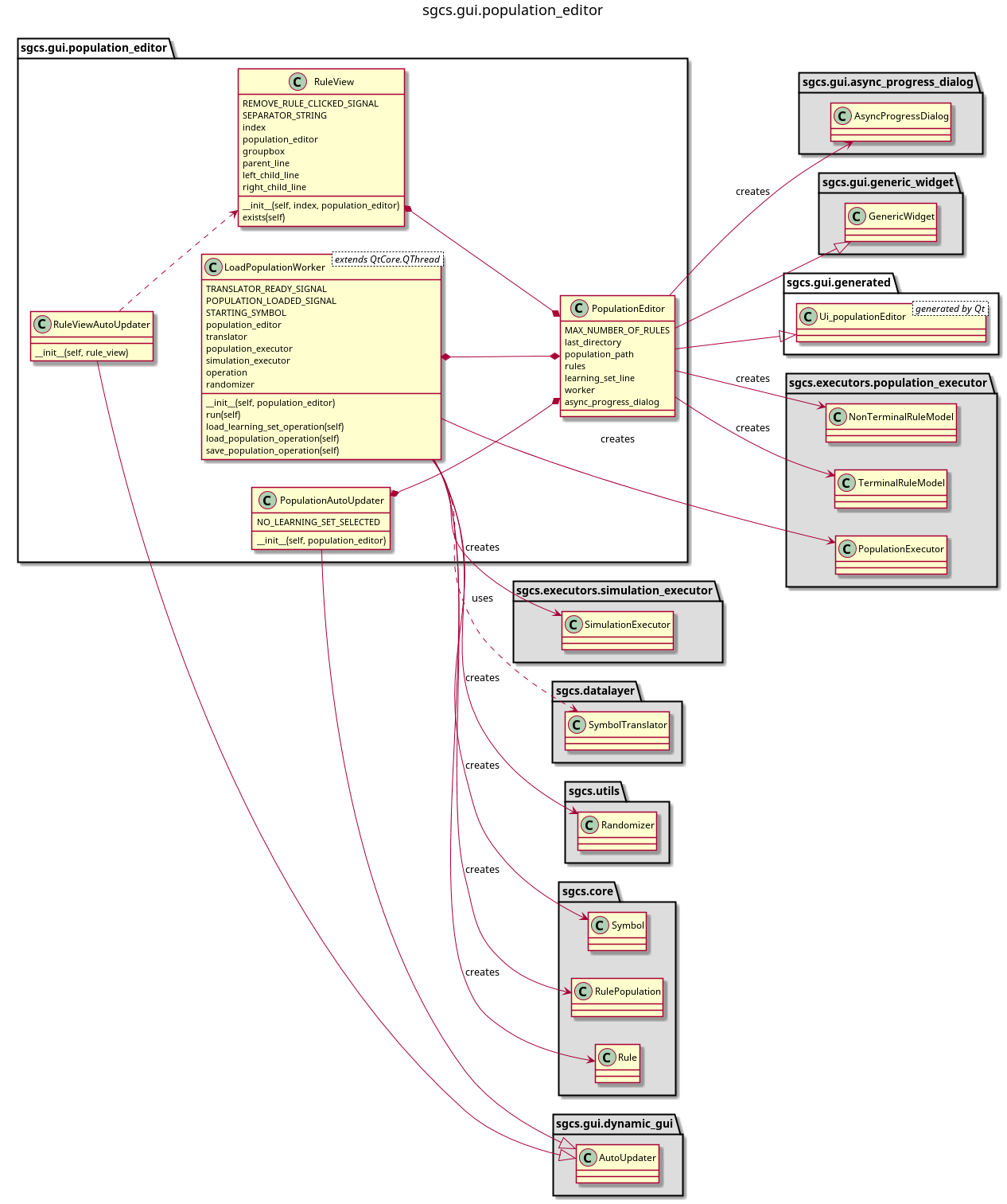
on\_selector\_change(self) – metoda obsługująca zmianę selektora przez bezpośrednią interakcję użytkownika z kontrolką (nie poprzez wczytanie istniejącej konfiguracji).

is\_in\_tournament\_mode(self) – metoda zwracająca True, jeżeli obecną selekcją jest selekcja turniejowa, False w przeciwnym przypadku.

create\_dynamic\_tournament\_node(self) – metoda tworząca dynamiczny węzeł odpowiedzialny za widoczność kontrolki rozmiaru turnieju.

on\_tournament\_size\_changed(self) – metoda obsługująca zmianę rozmiaru turnieju.

sgcs.gui.population\_editor



Okno „Edit population jest tworzone i sterowane przy pomocy klas z modułu sgcs.gui.population\_editor.

RuleViewAutoUpdater

Klasa ta zapewnia synchronizację danych oraz GUI pojedynczej reguły.

PopulationAutoUpdater

Klasa ta zapewnia synchronizację ścieżki do zbioru uczącego z odpowiednim widgetem.

NO\_LEARNING\_SET\_SELECTED – napis wykorzystywany w kontrolce w przypadku braku ścieżki.

RuleView

Klasa ta zarządza widokiem oraz edycją reguły. Informuje również klasę RuleEditor o potrzebie usunięcia reguły.

REMOVE\_RULE\_CLICKED\_SIGNAL – identyfikator sygnału usunięcia reguły.

SEPARATOR\_STRING – napis rozdzielający lewą oraz prawą stronę produkcji.

index – identyfikator reguły.

population\_editor – uchwyt obiektu PopulationEditor.

groupbox – groupbox, w którym klasa umieszcza wszystkie swoje elementy GUI (musi być widoczna na zewnątrz w celu podłączenia do GUI PopulationEditor).

parent\_line, left\_child\_line, right\_child\_line – uchwyty linii tekstu, które składają się na regułę.

exists(self) – Zwraca True, jeżeli istnieje reguła odpowiadająca temu obiektu (tj. w tabieli reguł na pozycji index coś jest).

LoadPopulationWorker

Zadaniem tej klasy jest odciążenie głównego wątku z kilku czasochłonnych procesów. Każdemu uruchomieniu takiego procesu będzie towarzyszyć wyświetlenie odpowiedniego okna, apelującego do użytkownika o cierpliwość. Obecnie Worker może w danej chwili realizować jedno z trzech zadań:

* Wczytywanie zbioru uczącego;
* Wczytywanie pliku populacji;
* Zapisywanie populacji do pliku.

TRANSLATOR\_READY\_SIGNAL – identyfikator sygnału wysyłanego zaraz po wczytaniu pliku ze zbiorem uczącym.

POPULATION\_LOADED\_SIGNAL – identyfikator sygnału wysyłanego po wczytaniu pliku populacji.

STARTING\_SYMBOL – stała, symbol startowy gramatyki.

population\_editor – uchwyt do instancji PopulationEditor.

translator – zmienna przechowująca obiekt SymbolTranslator. Bez tego obiektu nie jest możliwe tłumaczenie z wewnętrznej reprezentacji symboli na przyjazną człowiekowi oraz odwrotnie. Zmienna zostaje zainicjowana po wczytaniu zbioru uczącego.

population\_executor – zmienna przechowująca instancję obiektu PopulationExecutor, zajmującą się logiką biznesową obsługi populacji.

simulation\_executor – zmienna przechowująca instancję obiektu SimulationExecutor, zajmującego się logiką biznesową uruchamiania algorytmu oraz zbierania artefaktów.

operation – obecnie realizowane przez workera zadanie.

randomizer – instancja randomizera niezbędna dla prawidłowego utworzenia niektórych części systemu (na przykład populacji reguł).

run(self) – główna metoda wątku. W celu asynchronicznego wywołania należy wykorzystać metodę QtCore.Qthread.start(). Przed wywołaniem którejkolwiek z tych dwóch metod należy ustawić odpowiednią wartość pola operation (na przykład load\_learning\_set\_operation w celu załadowania zbioru uczącego).

load\_learning\_set\_operation(self) – operacja wczytująca zbiór uczący. Wczytanie zbioru uczącego powoduje zainicjowanie zmiennej translator odpowiednim instancją klasy SymbolTranslator, co umożliwia (i odblokowuje na poziomie GUI) wczytywanie oraz zapisywanie populacji. Należy pamiętać, że GUI chroni jedynie przed wczytaniem lub zapisaniem populacji w przypadku braku istnienia instancji SymbolTranslator. Nie chroni jednak przed wczytywaniem lub zapisywaniem populacji z innym zestawem uczącym niż ten, przy pomocy którego wygenerowano populację reguł. Nie może przed tym chronić, gdyż jest to często zachowanie pożądane (tworzymy zestaw reguł wspólnych dla kilku gramatyk, po czym zapisujemy je dla każdej z nich). Brak ostrożności może jednak spowodować powstanie źle wygenerowanych zbiorów.

load\_population\_operation – operacja umożliwiająca wczytanie istniejącego pliku populacji. Przed wczytaniem należy upewnić się, że wybrano pożądany zbiór uczący.

save\_population\_operation – operacja umożliwiająca zapisanie obecnej populacji do pliku. Przed zapisem należy upewnić się, że wybrano pożądany zbiór uczący.

PopulationEditor

Klasa ta niejako łączy wszystkie powyższe, zapewniając komunikację pomiędzy nimi, dodatkowo zajmuje się aktualizacją modelu populacji (dodawanie i usuwanie reguł w odpowiedzi na wysyłane przez siebie oraz workera sygnały).

MAX\_NUMBER\_OF\_RULES – stała definiująca jak wiele reguł może znaleźć się w populacji.

last\_directory – pole przechowujące ścieżkę ostatnio otwieranego przez nas folderu, dzięki czemu szukanie plików rozpoczynamy od miejsca, gdzie je ostatnio skończyliśmy.

population\_path – ścieżka do pliku ze zbiorem uczącym.

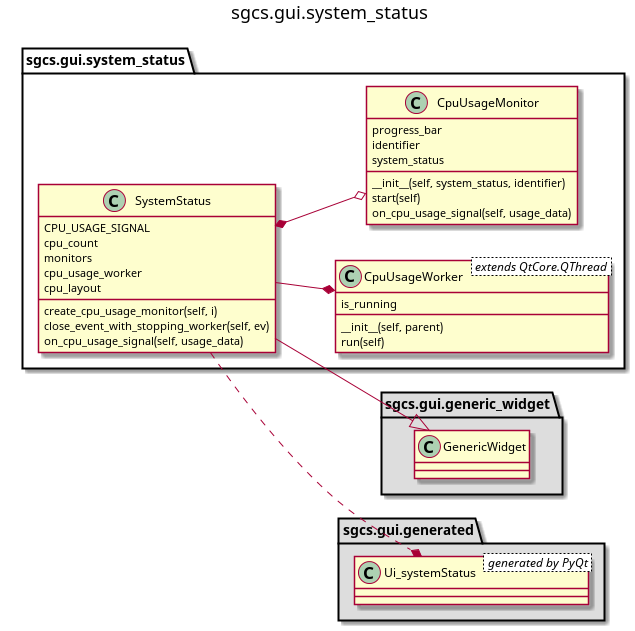
rules – lista obiektów RuleModel, będących reprezentacją reguł na potrzeby tego okna.

learning\_set\_line – linia odpowiadająca za wyświetlanie ścieżki do zbioru uczącego.

worker – instancja klasy LoadPopulationWorker.

async\_progress\_dialog – instancja widgetu asynchronicznego okna postępu (komunikat wyświetlany w trakcie trwania długich operacji).

sgcs.gui.system\_status



Zadaniem tego modułu jest monitorowanie stanu systemu (dokładnie pamięci oraz procesora) i informowanie o jego użyciu przez użytkownika. Nie jest to moduł bezpośrednio związany z samym algorytmem sGCS, jednakże jest niezwykle przydatny. Dzięki niemu można w łatwy sposób śledzić wykorzystanie wszystkich rdzeni oraz ilość zarezerwowanej pamięci. Domyślnie algorytm stara się optymalnie wykorzystywać przydzielone mu zasoby, toteż takie narzędzie jest przydatne w ustalaniu czy system rzeczywiście optymalnie wykorzystuje przydzielone mu zasoby i czy nie da się przyspieszyć całego procesu poprzez modernizację maszyny uruchamiającej aplikację.

CpuUsageWorker

Klasa ta w regularnych odstępach (1s) odpytuje system przy pomocy biblioteki psutil o jego obciążenie procesorów i ilość wolnej pamięci. Następnie wysyła tę informację przy pomocy sygnału SystemStatus.CPU\_USAGE\_SIGNAL.

is\_running – zmienna boolowska umożliwiają prawidłowe zakończenie działania wątku (po zmianie jej wartości na False nie dalej jak sekundę później nastąpi zakończenie działania wątku).

\_\_init\_\_(self, parent) – konstruktor. W tym przypadku pole parent to widget, do którego CpuUsageWorker ma wysyłać sygnały informujące o użyciu procesora (domyślnie widget powiązany z instancją SystemStatus).

run(self) – główna funkcja wątku. Zamiast niej powinno wywoływać się metodę QtCore.Qthread.start w celu asynchronicznego wykonania.

CpuUsageMonitor

Klasa zajmująca się nasłuchiwaniem na sygnały wykorzystania procesora. Jej zadaniem jest obsługa tylko jednego rdzenia.

progress\_bar – widget pasek postępu powiązany z obserwowanym rdzeniem procesora. Przyjmuje wartości od 0% do 100% zużycia.

identifier – Liczba naturalna jednoznacznie identyfikująca obserwowany rdzeń.

start(self) – Wywołanie tej metody powoduje rejestrację klasy na sygnały użycia procesora.

on\_cpu\_usage\_signal(self, usage\_data) – funkcja zajmująca się obsługą przybyłego sygnału zużycia procesora (aktualizuje progress\_bar).

SystemStatus

Klasa tworząca i koordynująca działanie pojedynczej instancji CpuUsageWorker oraz co najmniej jednej instancji CpuUsageMonitor. Oprócz tego zajmuje się obserwowaniem i informowaniem o zużyciu pamięci.

CPU\_USAGE\_SIGNAL – identyfikator sygnału zużycia procesora.

cpu\_counts – zmienna przechowująca całkowitą ilość rdzeni w systemie.

monitors – instancje klasy CpuUsageMonitor (tyle ile wynosi cpu\_counts).

cpu\_usage\_worker – instancja obiektu CpuUsageWorker.

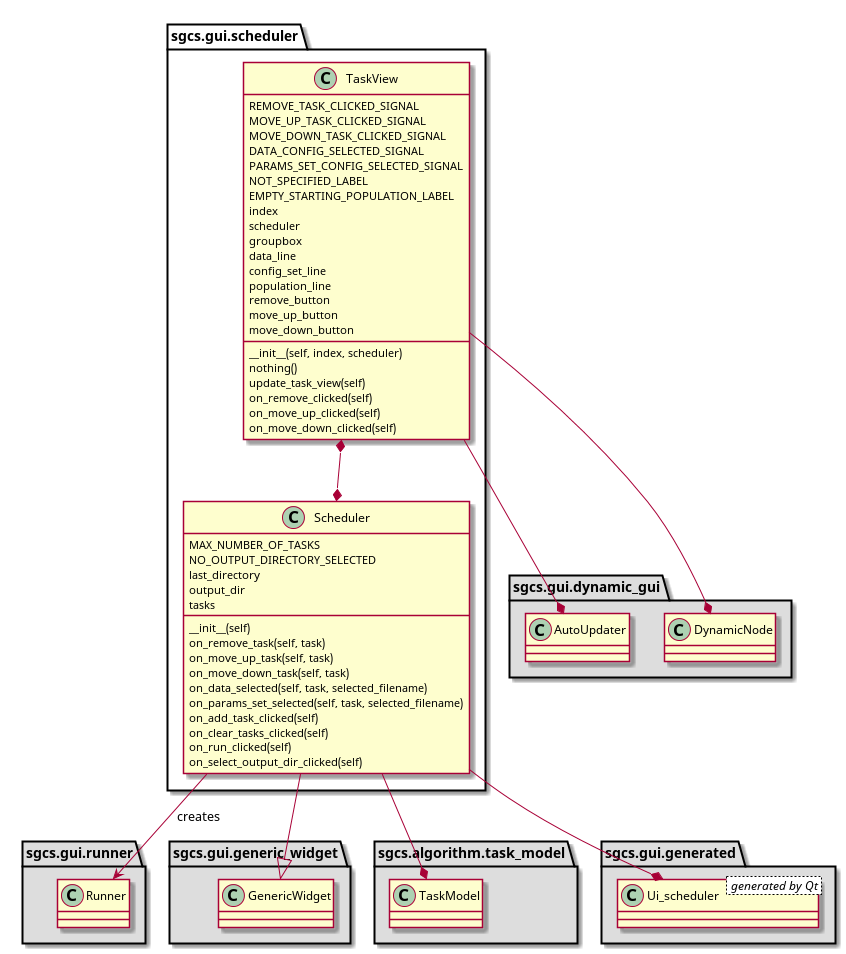
cpu\_layout – część interfejsu, do którego dodają się paski postępu monitorów.

create\_cpu\_usage\_monitor(self, i) – metoda nadzorująca utworzenie oraz rejestrację pojedynczego monitora. Parametr i to numer rdzenia, którego obserwacją ma zająć się monitor.

close\_event\_with\_stopping\_worker(self, ev) – metoda upewniająca się, że w przypadku zamknięcia okna CpuUsageWorker również prawidłowo zakońćzy działanie (wywoływane automatycznie tuż przed oryginalną obsługą wydarzenia zamknięcia).

on\_cpu\_usage\_signal(self, usage\_data) – metoda rejestrowana na sygnały zużycia procesora. Wyciąga z nich informację o zużyciu pamięci i dokonuje odpowiedniej aktualizacji GUI.

sgcs.gui.scheduler



Moduł ten odpowiada za obsługę menu „Scheduler”. Znajduje się tutaj zatem logika umożliwiająca przygotowanie zestawu zadań (tj. odpaleń konkretnego wariantu algorytmu z konkretnymi parametrami dla konkretnych danych wejściowych), zakolejkowanie ich i ustalenie ścieżki folderu na artefakty z przebiegów algorytmu. Odpowiedzialność za uruchomienie i monitoring przebiegu spoczywa na module sgcs.gui.runner, do którego ten moduł wysyła jedynie listę zaplanowanych zadań oraz ścieżkę folderu z artefaktami. Po uruchomieniu algorytmu modyfikacja tych danych nie będzie już możliwa.

TaskView

ViewModel obiektu zadania. Wyświetla jego dane (ścieżkę konfiguracji danych wejściowych \*inconf, konfiguracji parametrów \*.parconf oraz populacji początkowej \*.pop) i zajmuje się ich obsługą.

REMOVE\_TASK\_CLICKED – identyfikator sygnału informujący, że powiązany z ViewModelem model zadania powinien zostać usunięty.

MOVE\_UP\_TASK\_CLICKED\_SIGNAL – identyfikator sygnału informujący, że powiązany z ViewModelem model zadania powinien być przesunięty w górę kolejki.

MOVE\_DOWN\_TASK\_CLICKED\_SIGNAL - identyfikator sygnału informujący, że powiązany z ViewModelem model zadania powinien być przesunięty w dół kolejki.

DATA\_CONFIG\_SELECTED\_SIGNAL - identyfikator sygnału informujący o podaniu nowej ścieżki do danych wejściowych \*.inconf.

PARAMS\_SET\_CONFIG\_SELECTED\_SIGNAL - identyfikator sygnału informujący o podaniu nowej ścieżki do parametrów algorytmu \*.parconf.

NOT\_SPECIFIED\_LABEL – napis wyświetlany w przypadku niesprecyzowanych danych wejściowych lub parametrów algorytmu.

EMPTY\_STARTING\_POPULATION\_LABEL – napis wyświetlany, gdy nie podano ścieżki do populacji początkowej (przyjmowana jest wówczas domyślnie populacja pusta, rozszerzona ewentualnie o losowe reguły początkowe, oczywiście o ile tak stwierdzą parametry algorytmu).

index – indeks powiązanego modelu,

scheduler – uchwyt do obiektu klasy Scheduler.

groupbox – GUI generowane przez viewmodel, wystawione na potrzeby dołączenia do GUI Schedulera.

data\_line – element GUI wyświetlający obecną ścieżkę do pliku \*.inconf.

config\_set\_line – element GUI wyświetlający obecną ścieżkę do pliku \*.parconf.

population\_line – element GUI wyświetlający obecną ścieżkę do pliku \*.pop.

remove\_button – przycisk umożliwiający usunięcie powiązanego modelu.

move\_up\_button – przycisk umożliwiający przesunięcie zadania w górę kolejki.

move\_down\_button – przycisk umożliwiający przesunięcie zadania w dół kolejki.

nothing() - pusty delegat, przydatny w miejscach gdzie prawdziwy delegat nie jest potrzebny.

update\_task\_view(self) – metoda zajmująca się aktualizacją GUI na podstawie modelu.

on\_remove\_clicked(self) – metoda obsługująca naciśnięcie przycisku usunięcia zadania.

on\_move\_up\_clicked(self) – metoda obsługująca naciśnięcie przycisku przesunięcia zadania w górę kolejki.

on\_move\_down\_clicked(self) – metoda obsługująca naciśnięcie przycisku przesunięcia zadania w dół kolejki.

Scheduler

Klasa zarządzająca modelem kolejki zadań. Nasłuchuje na sygnały wysyłane przez obiekty TaskView i dokonuje odpowiednich modyfikacji kolejki.

MAX\_NUMBER\_OF\_TASKS – maksymalna ilość zadań w kolejce. GUI jest dynamicznie generowane i jest w stanie obsłużyć maksymalnie taką liczbę zaplanowanych zadań.

NO\_OUTPUT\_DIRECTORY\_SELECTED – napis wyświetlany zamiast ścieżki folderu na artefakty, jeżeli ten nie został jeszcze sprecyzowany.

last\_directory – zmienna pamiętająca ostatnio otwarty folder.

output\_dir – zmienna pamiętająca ścieżkę folderu na artefakty.

tasks – model kolejki zadań.

on\_remove\_task(self, task) – metoda zarejestrowana na sygnał REMOVE\_TASK\_CLICKED\_SIGNAL, usuwająca zadanie task z kolejki.

on\_move\_up\_task(self, task) – metoda zarejestrowana na sygnał MOVE\_UP\_TASK\_CLICKED\_SIGNAL, przenosząca zadanie task w górę kolejki.

on\_move\_down\_task(self, task) – metoda zarejestrowana na sygnał MOVE\_DOWN\_TASK\_CLICKED\_SIGNAL, przenosząca zadanie task w dół kolejki.

on\_data\_selected(self, task, selected\_filename) – metoda zarejestrowana na sygnał DATA\_CONFIG\_SELECTED\_SIGNAL, aktualizująca ścieżkę \*.inconf danego zadania task.

on\_params\_set\_selected(self, task, selected\_filename) – metoda zarejestrowana na sygnał PARAMS\_SET\_CONFIG\_SELECTED\_SIGNAL, aktualizująca ścieżkę \*.parconf danego zadania task.

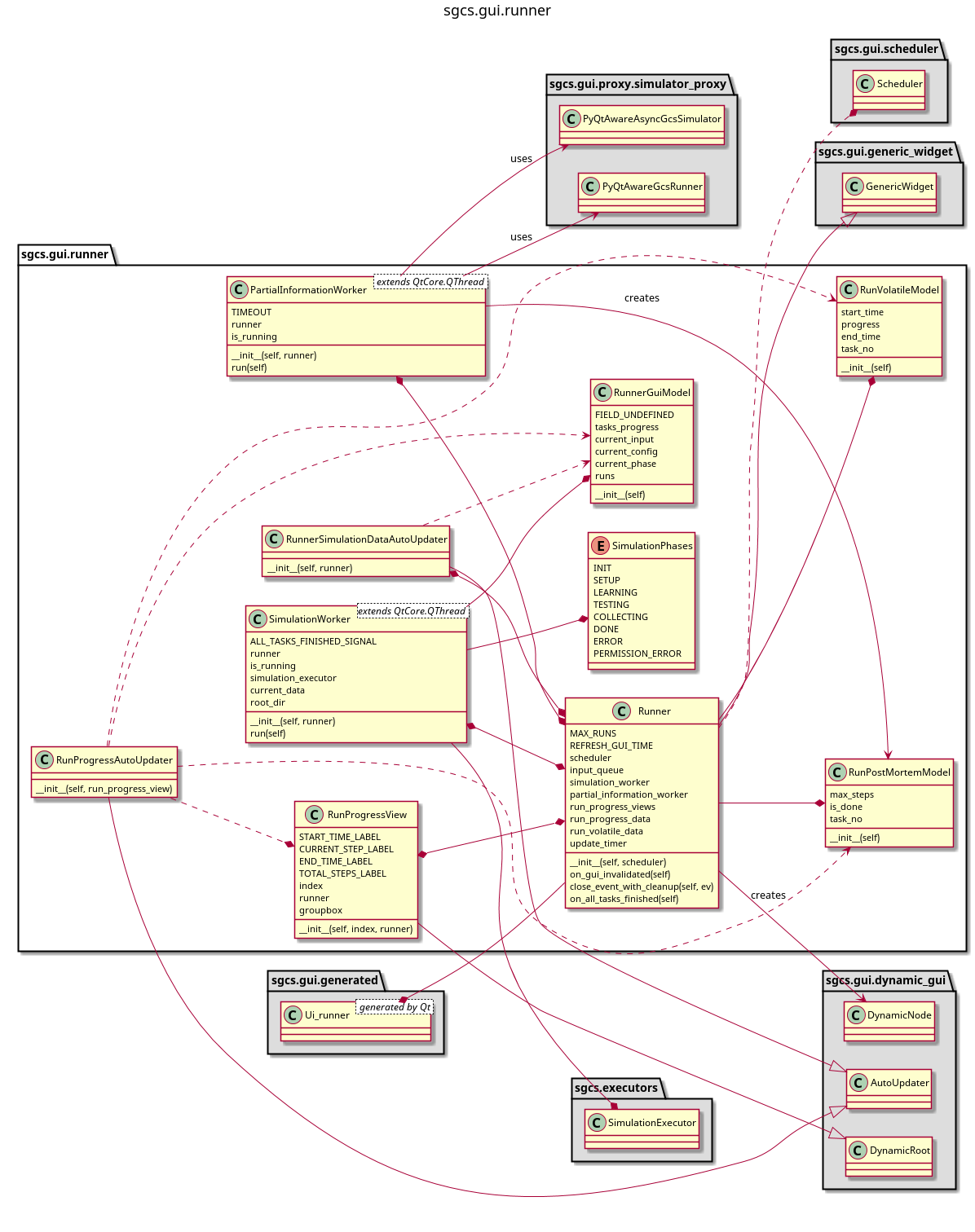
on\_add\_task\_clicked(self) – metoda zarejestrowana na naciśnięcie przycisku dodania nowego zadania. Powoduje ona dodanie nowego pustego (tj. z niesprecyzowaną ścieżką konfiguracji wejściowej i parametrów oraz pustą populacją) na końcu kolejki.

on\_clear\_tasks\_clicked(self) – metoda zarejestrowana na naciśnięcie przycisku wyczyszczenia kolejki zadań. Usuwa wszystkie zakolejkowane zadania.

on\_run\_clicked(self) – przekierowuje obecną kolejkę zadań i sterowanie do modułu sgcs.gui.runner. Oczywiście pod warunkiem poprawności danych (ścieżka artefaktów musi być podana, każde zadanie musi posiadać sprecyzowane ścieżki do pliku \*.inconf i \*.parconf, plik \*.pop jest opcjonalny).

on\_select\_output\_dir\_clicked(self) – medosa obsługująca wybór ścieżki folderu na artefakty.

sgcs.gui.runner

Moduł ten odpowiada za uruchomienie algorytmu i zapewnienie możliwości obserwacji jego działania przy pomocy GUI.

RunPostMortemModel

Model przechowujący niezmienne lub rzadko zmieniające się (na przykład pod koniec przebiegu uczącego). Nie zawiera danych przesyłanych przy pomocy kolejki Runner.input\_queue. W kodzie modułu ma czasami miejsce zamienne stosowanie terminologii „task” w odniesieniu do zadania (czyli całego uruchomienia algorytmu dla danej populacji startowej, parametrów i danych wejściowych) oraz w odniesieniu do przebiegu uczącego (pojedyncze uruchomienie etapu uczącego dla algorytmu). Należy więc zachować szczególną ostrożność w analizowaniu kodu modułu pod tym względem.

max\_steps – maksymalna ilość kroków ewolucyjnych, jaką może podjąć algorytm.

is\_done – czy dany przebieg uczący został zakończony.

task\_no – numer przebiegu uczącego.

RunVolatileModel

Model przeznaczony do przechowywania danych ulotnych. Są to dane ulegające dosyć częstym zmianom, przesyłane z algorytmu przy pomocy kolejki Runner.input\_queue. Mogą czasami nie być aktualne z obecnym stanem algorytmu, w szczególności jeżeli jest ich generowanych w danym momencie dużo, aczkolwiek są to nadal dane diagnostyczne i nie muszą być w 100% aktualne, tak długo jak pozostają w miarę aktualne.

start\_time – data rozpoczęcia przebiegu uczącego.

progress – numer obecnego kroku ewolucyjnego.

end\_time – data zakończenia przebiegu ewolucyjnego.

task\_no – numer przebiegu uczącego.

RunProgressAutoUpdater

Klasa odpowiedzialna za aktualizowanie obecnego stanu przebiegu uczącego.

RunProgressView

Klasa odpowiedzialna za prezentację obecnego stanu przebiegu uczącego. Zajmuje się początkowym ustawieniem stylu całego bloku oraz zapewnieniem widoczności odpowiednich jego elementów, podczas gdy samo zadanie odświeżania wartości zostało oddelegowane do instancji RunProgressAutoUpdater. Wyświetla GUI tylko dla przebiegów rzeczywiście istniejących (to jest RunProgressView o warotści pola index = 60 nic nie wyświetli jeżeli jest tylko 50 zaplanowanych przebiegów cyklu uczącego).

START\_TIME\_LABEL – stała, część dynamicznego GUI.

CURRENT\_STEP\_LABEL – stała, część dynamicznego GUI.

END\_TIME\_LABEL – stała, część dynamicznego GUI.

TOTAL\_STEPS\_LABEL – stała, część dynamicznego GUI.

index – numer przebiegu uczącego.

runner – uchwyt instancji klasy Runner.

groupbox – GUI generowane przez tą klasę, wystawione na potrzeby podpięcia do reszty GUI klasy Runner.

SimulationPhases

Klasa przechowująca fazy, które mają miejsce podczas działania algorytmu, wraz z ich opisami. W trakcie działania algorytmu wyróżniamy następujące fazy:

* INIT – faza początkowa, podczas której ma miejsce głównie przygotowanie interfejsu.
* SETUP – faza podczas której jest brane kolejne zadanie z kolejki, ma miejsce update GUI i przygotowanie środowiska do uruchomienia algorytmu.
* LEARNING – faza podczas której są wykonywane asynchronicznie przebiegu uczące w liczbie sprecyzowanej w odpowiednim pliku konfiguracyjnym \*.parconf.
* TESTING – faza testowa następująca po fazie uczącej, podczas której ma miejsce przebieg testowy algorytmu.
* COLLECTING – faza zbierania artefaktów. Po zakończeniu działania algorytmu wszystkie dane zgromadzone z jego przebiegu, wynikowa populacja, itd. są zapisywane w wyznaczonym wcześniej folderze. Po tym kroku następuje przejście do kroku ENVIRONMENT jeżeli kolejka zadań jest niepusta, przejście do kroku DONE w pp.
* DONE – przejście do tej fazy następuje po opróżnieniu kolejki zadań. Można bezpiecznie zamknąć okno Runner.
* ERROR – pojawia się w momencie, gdy w trakcie działania wystąpił krytyczny błąd i nie jest możliwe prawidłowe kontynuowanie działania algorytmu.
* PERMISSION\_ERROR – błąd pojawiający się, gdy podczas działania algorytm nie ma prawa dostępu do jakiegoś miejsca. Najczęściej wiąże się to z brakiem praw do zapisu w folderze z artefaktami, gdyż na przykład jakiś plik jest otwarty przez zewnętrzną aplikację. Algorytm będzie ponawiać próbę dostępu co sekundę, pozostając w tym stanie do czasu rozwiązania przez użytkownika problemu z dostępem do zasobu, po czym wznowi działanie od miejsca, gdzie je ostatnio skończył.

RunnerSimulationDataAutoUpdater

Klasa odpowiedzialna za aktualizację danych dotyczących całej obsługi kolejki zadań. Jej zadaniem jest wyświetlanie postępu w realizacji zadań z kolejki oraz ścieżki do obecnie wykorzystywanych plików wejściowych, co pozwala łatwo zidentyfikować wykonywane właśnie zadanie.

RunnerGuiModel

Model reprezentujący obecny stan kolejki zadań.

FIELD\_UNDEFINED – napis wyświetlany w przypadku, gdy dana wartość jest niezdefiniowana (na przykład czas startu przebiegu uczącego, który jeszcze się nie zaczął).

tasks\_progress – numer obecnie wykonywanego zadania (lub FIELD\_UNDEFINED jeżeli wykonywanie zadań jeszcze się nie rozpoczęło).

current\_input – ścieżka do pliku \*.inconfig obecnie wykonywanego zadania (lub FIELD\_UNDEFINED jeżeli wykonywanie zadań jeszcze się nie rozpoczęło).

current\_config – ścieżka do pliku \*.parconfig obecnie wykonywanego zadania (lub FIELD\_UNDEFINED jeżeli wykonywanie zadań jeszcze się nie rozpoczęło).

current\_phase – faza SimulationPhase, w której obecnie Runner się znajduje.

runs – kopia kolejki zadań wykonana na podstawie tej w klasie Scheduler.

SimulationWorker

Klasa odpowiedzialna za obsługę kolejki zadań i uruchamianie algorytmu dla kolejnych zadań.

ALL\_TASKS\_FINISHED\_SIGNAL – identyfikator sygnału informującego, że wszystkie zadania z kolejki zostały wykonane.

runner – uchwyt do instancji Runner.

is\_running – zmienna boolowska wykorzystywana do delikatnego wyłączenia workera. Niestety zatrzymanie SimulationWorkera nie jest łatwym zadaniem z całymi pulami procesów zarządzanymi przez niego i może nie nastąpić natychmiast, toteż po zatrzymaniu algorytmu w połowie wykonania zaleca się ponowne uruchomienie aplikacji w celu uzyskania pewności, że wątek ten został ostatecznie zniszczony. Jest to o tyle ważne, że właśnie ten wątek jest odpowiedzialny za uruchomienie całego algorytmu, nieprawidłowo zaś zatrzymany algorytm może pozostawić osierocone procesy nadal zajmujące sporą część zasobów komputera. Ponowne uruchomienie aplikacji po nagłym przerwaniu algorytmu daje 100% pewności, że wszystkie zasoby zostały zwolnione.

simulation\_executor – instancja obiektu SimulationExecutor zajmującego się logiką biznesową przebiegu symulacji oraz zbieraniem artefaktów.

current\_data – instancja RunnerGuiModel.

root\_dir – ścieżka do folderu na artefakty.

run(self) – główna metoda wątku. Zdecydowanie nie zaleca się jej ręcznego wywoływania, jako że jest to metoda zajmująca niezwykle dużo czasu (wykonuje po kolei wszystkie zadania). Zamiast tego dobrze jest wystartować ją w osobnym wątku poprzez wywołanie metody QtCore.Qthread.start(self).

PartialInformationWorker

Klasa zajmująca się odbieraniem wiadomości diagnostycznych przesyłanych z wnętrza algorytmu przy pomocy Runner.input\_queue. Są to czasy rozpoczęcia, zakończenia przebiegu uczącego oraz obecny krok ewolucyjny.

TIMEOUT – maksymalny czas (w sekundach) oczekiwania na wiadomość w Runner.input\_queue. W przypadku przekroczenia tego czasu worker sprawdzi swój warunek zakończenia pracy i jeżeli będzie miał ją kontynuować, to przejdzie do dalszego czekania na wiadomość. Przerwy w oczekiwaniu na wiadomość są konieczne w celu umożliwienia workerowi bezpieczne zakończenie pracy w przypadku problemu z działaniem algorytmu. Wybrano 10 sekund uniknięcia obciążania procesora częstym łapaniem wyjątków i sprawdzaniem warunków działania. Oznacza to też, że w przypadku przedwczesnego zakończenia działania algorytmu wątek zakończy swoje działanie najpóźniej w 10 sekund (czas obsługi pojedynczej wiadomości jest w tym przypadku pomijalnie mały).

runner – uchwyt instancji klasy Runner.

is\_running – zmienna boolowska, której przestawienie na False spowoduje zakończenie działania wątku w co najwyżej 10 sekund (patrz opis pola TIMEOUT powyżej).

run(self) – główna metoda wątku. W celu asynchronicznego wywołania należy stosować QtCore.Qthread.start(self).

Runner

Klasa odpowiedzialna za obsługę GUI podczas działania algorytmu. W zasadzie oddelegowuje ona całą logikę do swoich klas składowych, sama przechowując kilka współdzielonych przez nie obiektów Zapewnia też prawidłowe zwalnianie zasobów w przypadku przedwczesnego zakończenia działania algorytmu, wyświetlenie odpowiedniego powiadomienia po planowanym zakończeniu działania algorytmu oraz okresowe odświeżanie GUI.

MAX\_RUNS – maksymalna ilość przebiegów procesu uczenia. Jeżeli algorytm uruchomi większą ilość cykli uczących, wówczas nie będą one prawidłowo wyświetlane. Wartość tą można zwiększyć, aczkolwiek PyQt zawsze zarezerwuje zasoby niezbędne do stworzenia maksymalnego GUI, co może spowodować wolniejsze działanie programu.

REFRESH\_GUI\_TIME – wartość tego pola definiuje co ile milisekund następi odświeżenie GUI.

scheduler – uchwyt schedulera, w którym zaprojektowano to uruchomienie.

input\_queue – kolejka wiadomości typu multiprocessing.Queue. Wykorzystywana do przesyłania wiadomości diagnostycznych z wnętrza działającego algorytmu.

simulation\_worker – instancja klasy SimulationWorker.

partial\_information\_worker – instancja klasy PartialInformationWorker.

run\_progress\_views – lista (o długości MAX\_RUNS) obiektów typu RunProgressView.

run\_progress\_data – lista obiektów RunVolatileModel.

update\_timer – instancja obiektu QtCore.QTimer, wywołujący odświeżenie GUI co REFRESH\_GUI\_TIME milisekund.

on\_gui\_invalidated(self) – metoda wymuszająca odświeżenie całego GUI.

close\_event\_with\_cleanup(self, ev) – rozszerzenie standardowej obsługi zamknięcia okna, wymuszające zatrzymanie wszystkich podległych wątków w pierwszej kolejności.

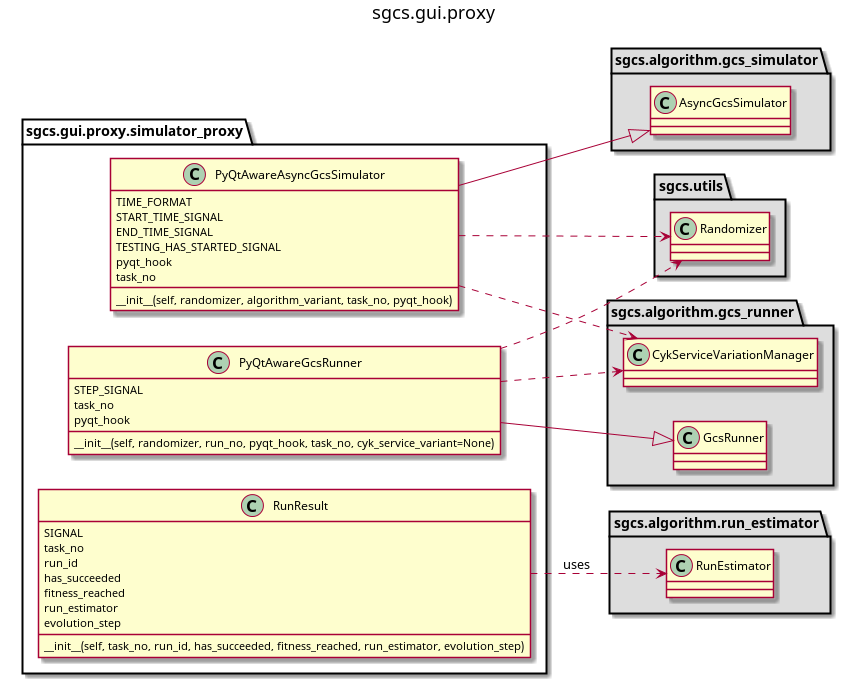
on\_all\_tasks\_finished(self) – metoda wywoływana w chwili planowanego ukończenia wszystkich zadań – powoduje wyświetlenie stosownego komunikatu.

sgcs.gui.proxy

W bibliotece podjęto starania mające na celu uniezależnienie biblioteki od wykorzystywanego GUI. Niestety GUI potrzebuje różnego rodzaju informacji na różnych etapach działania algorytmu i pełne rozdzielenie ich od siebie okazało się nietrywialnym zadaniem. Ostatecznie zdecydowano się na następujące rozwiązanie – cały algorytm został napisany w tak, że będzie on praktycznie nieświadomy istnienia GUI, natomiast moduł sgcs.gui.proxy stanowić będzie rozszerzenie jego najwyższego poziomu abstrakcji w celu udostępnienia niezbędnych GUI danych. Umożliwia to uruchamianie algorytmu w całkowitej separacji od GUI.

Zatem moduł ten zawiera prosty protokół asynchronicznej komunikacji pomiędzy algorytmem a GUI oraz klasy świadome tego nowego mechanizmu.

Wbrew mylnemu nazewnictwu stosowanemu wewnątrz modułu nie zawiera on żadnych zależności względem biblioteki PyQt. Stanowi rozszerzenie interfejsu algorytmu, które może być wykorzystane przez dowolne GUI, nawet przez aplikację konsolową.

RunResult

Klasa odpowiedzialna za wysyłanie pełnego zestawu danych zakończonego przebiegu uczącego algorytmu.

SIGNAL – identyfikator sygnału wysyłanego po zakończeniu przebiegu uczącego algorytmu.

task\_no – numer aktualnie wykonywanego zadania.

run\_id – numer przebiegu uczącego algorytmu.

has\_succeded – zmienna boolowska informująca czy przebieg zakończył się sukcesem.

fitness\_reached – osiągnięty poziom fitness f z przedziału [0;1]

run\_estimator – obiekt run\_estimator zawierający statystyki danego przebiegu.

evolution\_step – ilość wykonanych kroków ewolucyjnych.

PyQtAwareGcsRunner

Klasa rozszerzająca klasę GcsRunner o wysyłanie sygnałów kroku ewolucyjnego po każdym zakończonym kroku ewolucyjnym.

STEP\_SIGNAL – identyfikator sygnału zakończonego kroku ewolucyjnego.

task\_no – numer wykonywanego zadania.

pyqt\_hook – uchwyt do kolejki Runner.input\_queue z pomocą której następuje wysyłanie wiadomości.

PyQtAwareAsyncGcsSimulator

Klasa rozszerzająca klasę AsyncGcsSimulator o wysyłanie sygnałów rozpoczęcia, zakończenia oraz przejścia do cyklu testowego algorytmu.

TIME\_FORMAT – stała, format zapisu wysyłanego czasu.

START\_TIME\_SIGNAL – identyfikator sygnału czasu rozpoczęcia działania danego przebiegu algorytmu (zarówno uczącego, jak i testowego).

END\_TIME\_SIGNAL – identyfikator sygnału czasu zakończenia działania danego przebiegu algorytmu (zarówno uczącego, jak i testowego).

TESTING\_HAS\_STARTED – identyfikator sygnału rozpoczęcia cyklu testowego algorytmu.

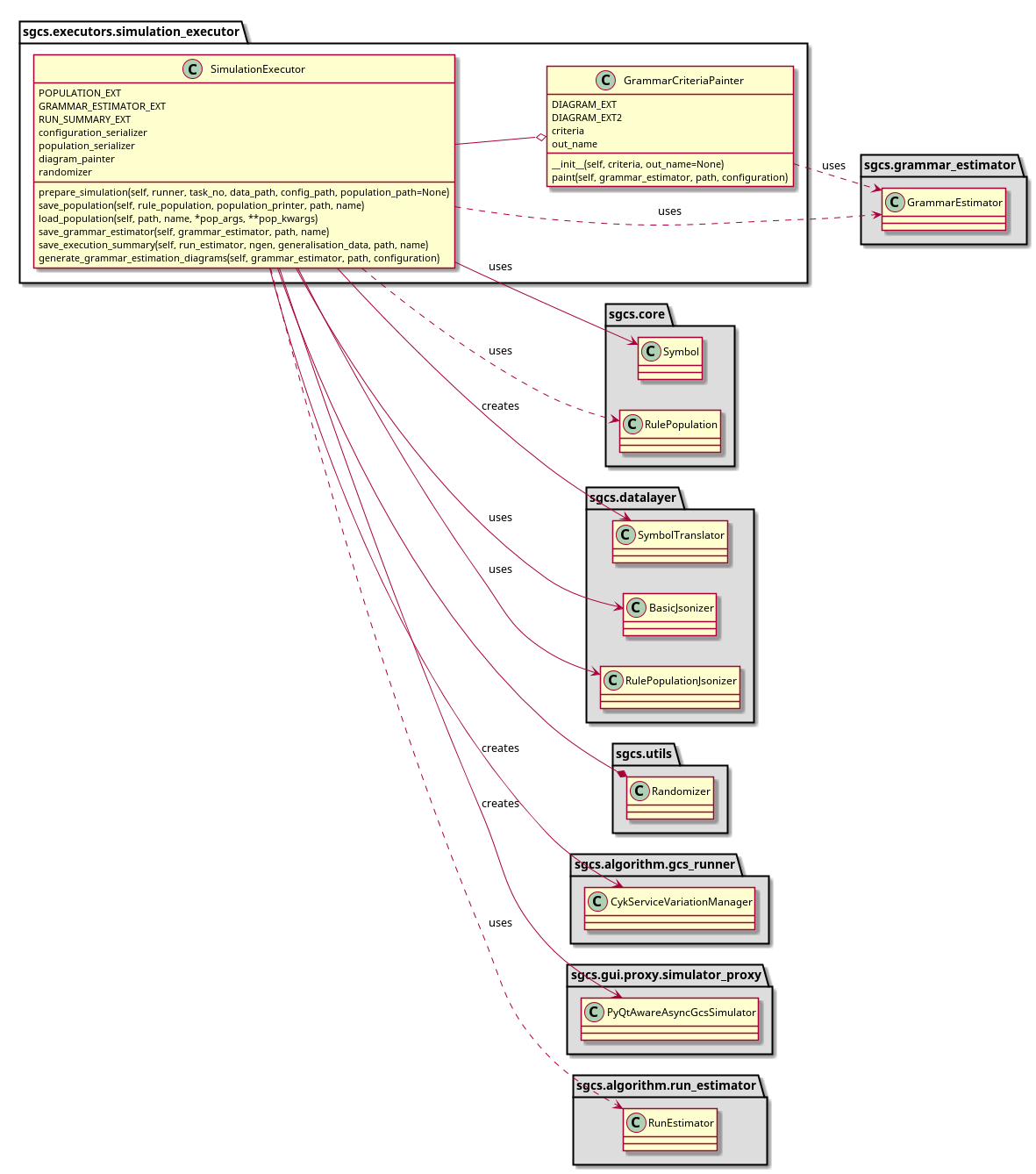
task\_no – numer wykonywanego zadania.

pyqt\_hook – uchwyt do kolejki Runner.input\_queue z pomocą której następuje wysyłanie wiadomości.

Warstwa wykonawców

Cała logika biznesowa biblioteki znajduje się w przestrzeni nazw sgcs.executors. Warstwa ta posiada stosunkowo niewiele funkcjonalności – główne zadania egzekutorów to tłumaczenie danych z reprezentacji wewnętrznej na łatwą do odczytu dla człowieka, generowanie diagramów oraz zbieranie artefaktów z działania algorytmu.

sgcs.executors.simulation\_executor



Moduł ten skupia się na obsłużeniu logiki występującej podczas działania algorytmu. Zajmuje się prawidłowym uruchomieniem algorytmu oraz zapisywaniem artefaktów, w tym rysowaniem diagramów.

SimulationExecutor

Klasa ta łączy większość logiki systemu, zapewniając możliwość załadowania niezbędnych danych z pomocą modułu sgcs.datalayer, zainicjować klasy z modułu sgcs.algorithm odpowiednimi danymi, a po wykonaniu wygenerować artefakty i zapisać je ponownie z pomocą modułu sgcs.datalayer do folderu artefaktów.

POPULATION\_EXT – stała, rozszerzenie plików populacji wykorzystywane w całym systemie („.pop”).

GRAMMAR\_ESTIMATOR\_EXT – stała, rozszerzenie plików oceny gramatyki wykorzystywane w całym systemie („.grest”).

RUN\_SUMMARY\_EXT – stała, rozszerzenie plików podsumowania symulacji wykorzystywane w całym systemie („.txt”).

configuration\_serializer – instancja BasicJsonizer, która umożliwia tłumaczenie węzłów konfiguracji na format \*.json i odwrotnie.

population\_serializer – instancja RulePopulationJsonizer, która umożliwia tłumaczenie obiektów typu RulePopulation na format \*.json i odwrotnie.

diagram\_painter – lista instancji klasy GrammarCriteriaPainter dla wszystkich wspieranych rodzajów diagramów.

randomizer – instancja obiektu Randomizer.

prepare\_simulation(self, runner, task\_no, data\_path, config\_path, population\_path=None) – metoda konstruująca z podanych danych delegata (który otrzymawszy obiekt konfiguracji uruchomi cały algorytm, a po jego zakończeniu zwróci krotkę zawierającą wyniki wykonania algorytmu), obiekt konfiguracji oraz delegata umożliwiającego wypisanie w czytelny dla człowieka sposób reguł populacji.

save\_population(self, rule\_population, population\_printer, path, name) – metoda umożliwiająca zapisanie populacji reguł rule\_population w folderze path w dwóch plikach – szczegółowego i ciężkiego w odczycie przez człowieka pliku name.pop oraz łatwego w odczycie pliku name\_view.txt.

load\_population(self, path, name, \*pop\_args, \*\*pop\_kwargs) – metoda umożliwiająca odczytanie populacji z pliku path/name.pop. Ponieważ populacja reguł wymaga dodatkowych parametrów przy tworzeniu, parametry te muszą być również dostarczone (pop\_args, pop\_kwargs).

save\_grammar\_estimator(self, grammar\_estimator, path, name) – metoda powoduje zapisanie obiektu typu GrammarEstimator w pliku path/name.grest

save\_execution\_summary(self, run\_estimator, ngen, generalization\_data, path, name) – metoda generuje plik podsumowania przebiegu algorytmu, a następnie zapisuje go w path/name.txt

generate\_grammar\_estimation\_diagrams(self, grammar\_estimator, path, configuration) – metoda powoduje wygenerowanie wszystkich wspieranych (lista diagram\_painter) rodzajów diagramów i zapisanie ich w folderze path.

GrammarCriteriaPainter

Klasa zdolna do wygenerowania pojedynczego diagramu. W obecnej postaci klasa brakuje trochę do bycia skończonej – wiele cech diagramów, które mogłyby być parametryzowalne, nie są, oprócz tego brak wsparcia dla generowania diagramów w postaci grafiki wektorowej, co byłoby bardzo wygodne w szczególności właśnie w przypadku diagramów. Większość logiki umożliwiającej to jest już zaimplementowana i niewiele potrzeba, aby rozszerzyć tę klasę o nowe możliwości, tymczasem spełnia ona wszystkie wymagania niezbędne do rysowania wymaganych przez aplikację diagramów.

DIAGRAM\_EXT – domyślne rozszerzenie grafiki rastrowej, do którego są zapisywane diagramy.

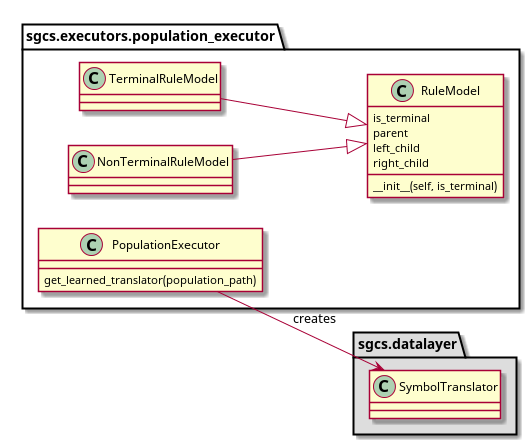
DIAGRAM\_EXT2 – domyślne rozszerzenie grafiki wektorowej, do której mogłyby być zapisywane diagramy.

criteria – nazwa i równocześnie identyfikator kryterium, które ma być przedstawione na diagramie.

out\_name – opcjonalna nazwa, która ma być stosowana na diagramach zamiast criteria (alternatywna reprezentacja).

paint(self, grammar\_estimator, path, configuration) – metoda rysująca, a następnie zapisująca diagram do plików (na razie tylko postać rastrowa, wektorowa może zostać dodana w przyszłości).

sgcs.executors.population\_executor



Moduł skupia się na zapewnieniu obsługi populacji poza działającym algorytmem. Jest to stosunkowo prosty moduł, większość funkcjonalności jest realizowana przez warstwy niższe (sgcs.core oraz sgcs.datalayer).

RuleModel

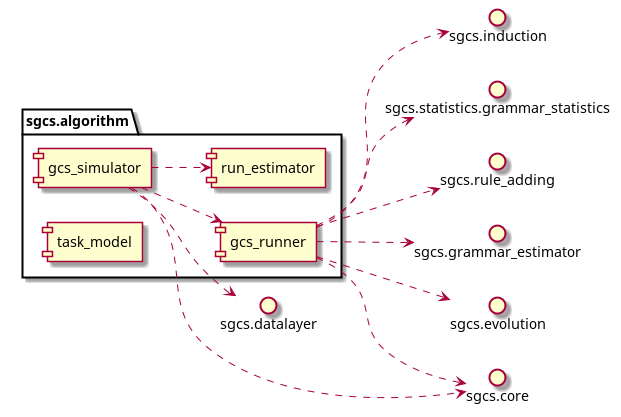
Prosta klasa do przechowywania reguł w czytelnej dla człowieka formie. Posiada dwie klasy dziedziczące po niej (TerminalRuleModel oraz NonTerminalRuleModel).

PopulationExecutor

Klasa umożliwiająca dostęp do obiektów populacji poza działającym algorytmem.

get\_learned\_translator(population\_path) – po podaniu ścieżki do zbioru uczącego population\_path wczytuje go w całości, w efekcie zwracając obiekt SymbolTranslator. Obiekt ten jest zdolny do tłumaczenia symboli w dwie strony – z reprezentacji wewnętrznej do przyjaznej człowiekowi i odwrotnie, jest zatem niezbędny do interaktywnej obsługi populacji poza działającym algorytmem.

Właściwy algorytm



Właściwy algorytm zawiera całą logikę działania algorytmu klasyfikacji gramatycznej, całkowicie odseparowaną od interfejsu graficznego oraz od niskopoziomowych funkcjonalności jak na przykład obsługa systemu plików. Warstwa ta zapewnia spójny interfejs do uruchamiania algorytmu w dowolnym wariancie – domyślnie wspiera GCS, sGCS oraz neg-sGCS wraz z licznymi parametrami i mechanizmami usprawniającymi ich działanie.

sgcs.algorithm.task\_model

Moduł ten zawiera jedynie jedną prostą klasę – TaskModel. Jego pozostałe funkcjonalności zostały przeniesione do innych modułów i powinien zostać usunięty po przeniesieniu klasy TaskModel.

TaskModel

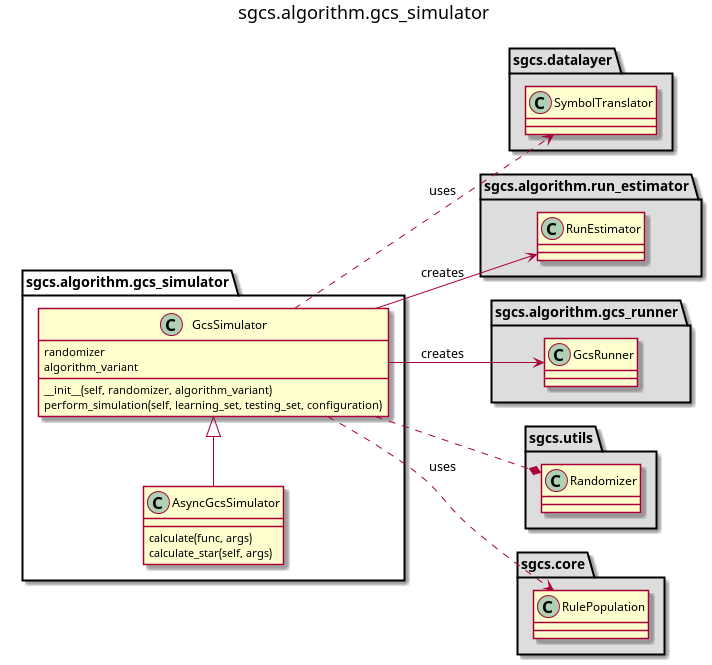
Klasa mająca za zadanie reprezentować pojedyncze zadanie w kolejce zaplanowanych uruchomień algorytmów. Zawiera 3 pola przechowujące ścieżki do związanych z zadaniem zasobów.

data\_configuration – ścieżka do pliku \*.inconf dla zadania.

params\_configuration – ścieżka do pliku \*.parconf dla zadania.

population\_configuration – ścieżka do pliku \*.pop dla zadania.

sgcs.algorithm.gcs\_simulator

Moduł ten zawiera klasy GcsSimulator oraz AsyncGcsSimulator. Są to klasy najwyższego poziomu wykonujące w swoim ciele cały algorytm klasyfikacji gramatyki.

GcsSimulator

Wbrew nazwie klasa ta zajmuje się obsługą dowolnego algorytmu klasyfikacji gramatycznej. Obiektem decydującym o różnym zachowaniu jest algorithm\_variant, aczkolwiek na tym poziomie jest on tylko przekazywany w dół struktury do klasy GcsRunner. GcsSimulator uruchamia configuration.max\_algorithm\_runs przebieg uczący, po czym przechodzi do pojedynczego przebiegu testowego. Całość jest wykonana sekwencyjnie.

randomizer – uchwyt obiektu Randomizer.

algorithm\_variant – obiekt typu CykServiceVariationManager.

perform\_simulation(self, learning\_set, testing\_set, configuration) – funkcja pobiera SymbolTranslator wygenerowany na podstawie zbioru uczącego, SymbolTranslator wygenerowany na podstawie zbioru testowego oraz obiekt konfiguracji. Dokonuje przebiegów uczacych, a następnie testowych, po czym zwraca krotkę zawierającą dane zebrane podczas procesu.

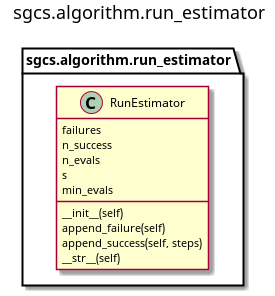
AsyncGcsSimulator

Klasa ta w przeciwieństwie do swojej nadklasy uruchamia cykle uczące równolegle zamiast sekwencyjnie. Na samym początku zostaje utworzona pula procesów o takiej liczności jak ilość rdzeni logicznych dostępnych na danym komputerze (czyli uwzględnia ilość rdzeni fizycznych oraz technologie typu Hyper-Threading). Następnie wszystkie przebiegi cyklu uczącego są wykonywane asynchronicznie przez tą pulę procesów. Główny proces zajmuje się jedynie zbieraniem wyników oraz ewentualnym łączeniem niektórych danych (np. przeliczanie średniej wartości funkcji fitness w danym kroku ewolucyjnym dla wszystkich przebiegów uczących algorytmu). Dzięki temu wykorzystanie asynchronicznego zamiast sekwencyjnego symulatora pozostaje transparentne i daje zauważalne przyspieszenie w przypadku parsowania gramatyk wymagających dużej ilości kroków ewolucyjnych do wyuczenia się gramatyki.

calculate(func, args) – funkcja pomocnicza wywołująca pierwszy parametr z listą parametrów przekazaną w postaci drugiego argumentu.

calculate\_star – funkcja przygotowana do wykorzystania z multiprocessing.imap, w wygodny sposób przekazuje parametry do funkcji, zapewnia inną wartość seeda dla każdego z procesów oraz rozszerza zwracaną wartość funkcji o dodatkowe informacje wymuszone przez asynchroniczność wykonania obliczeń.

sgcs.algorithm.run\_estimator



Klasa ta odpowiada za gromadzenie statystyk wykonania algorytmu. Po każdym ukończeniu przebiegu uczącego jest wołana metoda append\_failure lub append\_success. W efekcie po zakończeniu uczenia instancja tej klasy jest w stanie podać wartości n\_success, n\_evals, s oraz min\_evals umożliwiające ocenę przebiegu procesu uczenia.

failures – ilość przebiegów uczących zakończonych porażką.

n\_success – parametr nSuccess (ilość przebiegów uczących zakończonych znalezieniem satysfakcjonującej gramatyki).

n\_evals – parametr nEvals (średnia ilość kroków potrzebnych do znalezienia satysfakcjonującej gramatyki).

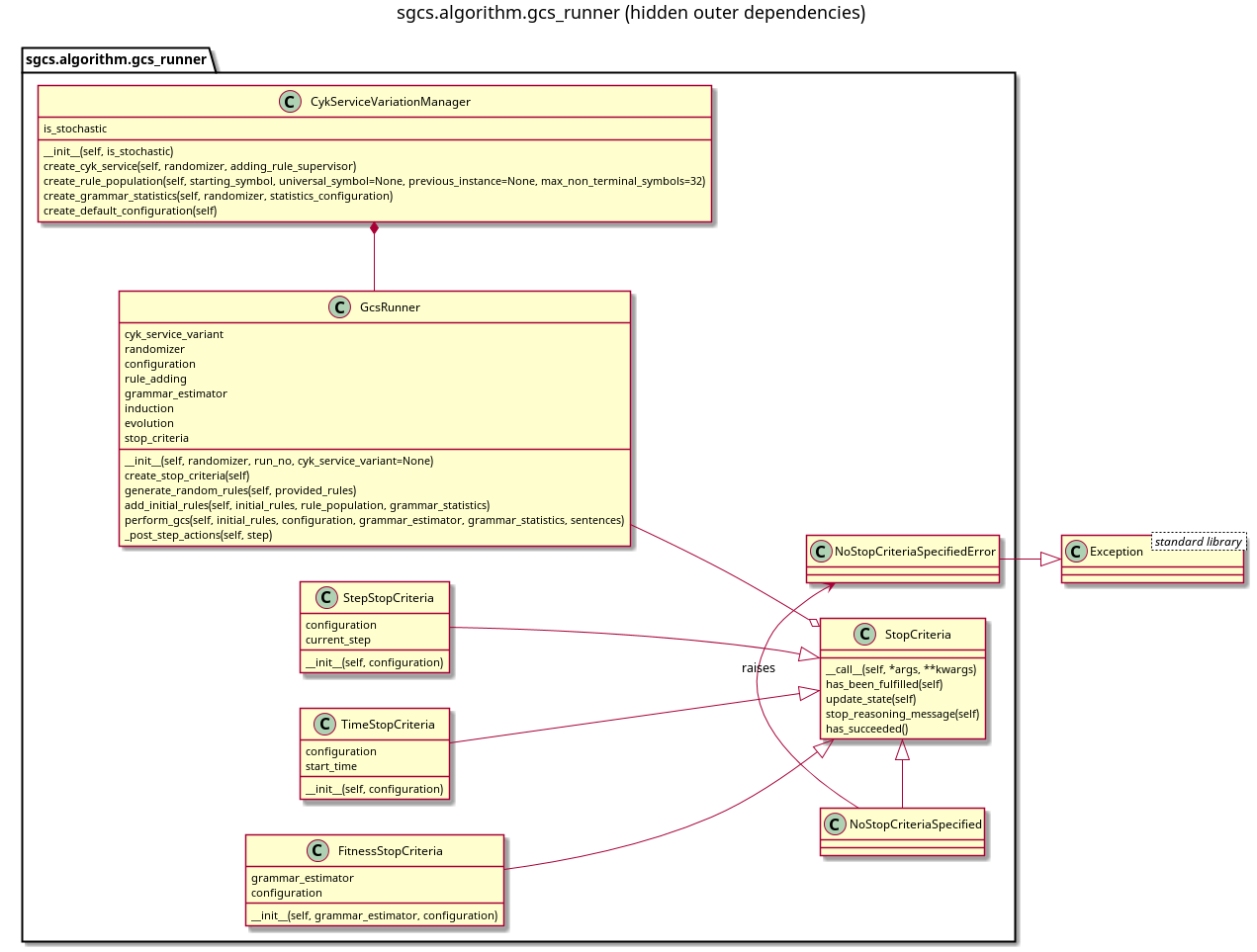
s – parametr s (odchylenie standardowe wartości nEvals).

min\_evals – parametr minEvals (minimalna ilość kroków potrzebnych do znalezienia satysfakcjonującej gramatyki).

append\_failure(self) – zapisuje przebieg jako zakończony porażką.

append\_success(self, steps) – zapisuje przebieg jako zakończony sukcesem (wraz z ilością kroków).

sgcs.algorithm.gcs\_runner

Moduł sgcs.algorithm.gcs\_runner gromadzi całą logikę pojedynczego przebiegu algorytmu w jedną metodę. Poszczególne klasy tego modułu posiadają wiele zależności od innych modułów i przedstawienie ich w czytelny sposób na diagramie jest praktycznie niemożliwe, dlatego też zależności międzymodułowe nie zostały przedstawione. Jest on zależny od wielu klas należących do następujących modułów:

* sgcs.utils;
* sgcs.rule\_adding;
* sgcs.core;
* sgcs.statistics,grammar\_statistics;
* sgcs.grammar\_estimator;
* sgcs.evolution;
* sgcs.induction;

CykServiceVariationManager

Klasa ta zarządza różnicami w zachowaniu pomiędzy algorytmami stochastycznymi (sGCS, neg-sGCS) a klasycznymi (GCS). Zamysłem jest tutaj wystawienie wspólnego API dla obu grup algorytmów, ukrywając w ciałach metod wszelkie różnice pomiędzy nimi. Dzięki temu klient tej klasy (GcsRunner) może pozostać w całkowitej nieświadomości czy wykorzystywany przez niego algorytm jest stochastyczny, czy też nie.

is\_stochastic – przechowuje informację czy obecnie wykorzystywany algorytm jest stochastyczny, czy też nie.

create\_cyk\_service(self, randomizer, adding\_rule\_supervisor) – metoda tworzy instancję typu CykService w przypadku algorytmu klasycznego lub StochasticCykService w przypadku algorytmu stochastycznego.

create\_rule\_population(self, starting\_symbol, universal\_symbol=None, previous\_instance=None, max\_non\_terminal\_symbols=32) – metoda tworząca instancję RulePopulation lub StochasticRulePopulation.

create\_rule\_statistics(self, randomizer, statistics\_configuration) – metoda tworząca odpowiednią instancję liczącą statystyki gramatyki (klasyczną lub pasieki). Należy zwrócić uwagę, że decyzja jest w tym przypadku zależna od tego czy dana konfiguracja akceptuje uczenie na zdaniach negatywnych (Pasieka nie, klasyczna tak).

create\_default\_configuration(self) – metoda tworząca domyślny obiekt konfiguracji dla danego algorytmu.

StopCriteria

Abstrakcyjna klasa realizująca pojedynczy warunek stopu dla głównej pętli algorytmu.

Przykład użycia:

criterias = [StopCriteria1(), StopCriteria2()]

while not any(cr() for cr in criterias):

…

message = „Success!” if any(cr.has\_succeded() for cr in criterias) else „Failure!”

print(message)

has\_been\_fullfilled(self) – metoda sprawdzająca spełnienie warunku.

update\_state(self) – metoda aktualizująca wewnętrzny stan instancji kryterium. Kiedy wywołujemy metodę \_\_call\_\_ kryterium, wówczas najpierw zostanie wywołana metoda update\_state, a następnie dopiero zwrócony rezultat has\_been\_fullfilled.

stop\_reasoning\_message(self) – metoda wypisująca przyczynę zatrzymania (przydatne przy kilku warunkach stopu).

has\_succeded() – metoda informująca czy nastąpienie tego warunku stopu powoduje sukces, czy też nie. Dzięki temu możemy mieszać różne kryteria razem i na podstawie ich stanu po opuszczeniu pętli ustalić, czy nastąpił sukces, czy też porażka. Domyślnie zwraca wartość False.

NoStopCriteriaSpecified

Klasa dziedzicząca po StopCriteria. Przy próbie wywołania metody has\_been\_fullfilled podnosi wyjątek NoStopCriteriaSpecifiedError. Jest to przydatne gdy dysponuje się klasą nieposiadającą domyślnych kryteriów stopu i chce się mieć pewność, że użytkownik zapewni własne.

FitnessStopCriteria

Warunek stopu spełniony, kiedy przejście gramatyki zakończy się uzyskaniem fitness na poziomie configuration.satisfying\_fitness. Jego metoda has\_succeded() zwraca wartość True.

grammar\_estimator – uchwyt do instancji GrammarEstimator, która pozwala sprawdzić obecną wartość fitness gramatyki.

configuration – uchwyt do węzła konfiguracyjnego pozwalający odczytać wartość configuration.satisfying\_fitness.

StepStopCriteria

Warunek stopu spełniony, kiedy current\_step przekroczy configuration.max\_algorithm\_steps. current\_step, jest preinkrementowany przy każdym wywołaniu metody update\_state (czyli również przy \_\_call\_\_).

configuration – uchwyt do węzła konfiguracyjnego pozwalający odczytać wartość configuration.max\_algorithm\_steps.

current\_step – obecny krok (zaczyna liczenie od 0).

TimeStopCriteria

Warunek stopu spełniony, kiedy obecny czas jest większy od start\_time + configuration.max\_execution\_time.

configuration – uchwyt do węzła konfiguracyjnego pozwalający na odczytanie wartości configuration.max\_execution\_time

start\_time – czas w chwili utworzenia tego obiektu.

GcsRunner

Klasa ta realizuje główną pętlę algorytmu w sposób niezależny od tego, czy jest stochastyczny i czy wykorzystuje zdania negatywne w procesie uczenia. Wykorzystuje wszystkie warunki stopu (FitnessStopCriteria, TimeStopCriteria oraz StepStopCriteria).

cyk\_service\_variant – klasa zajmująca się obsługą różnic pomiędzy różnymi wariantami agorytmu.

randomizer – uchwyt instancji Randomizer.

configuration – uchwyt węzła konfiguracyjnego.

rule\_adding – instancja klasy RuleAddingSupervisor.

grammar\_estimator – obiekt klasy GrammarEstimator.

induction – instancja klasy CykService.

evolution – instancja klasy EvolutionService.

stop\_criteria – lista obiektów StopCriteria wykorzystywanych jako warunek stopu głównej pętli algorytmu.

create\_stop\_criteria(self) – Funkcja inicjująca listę stop\_criteria nowymi instancjami klas FitnessStopCriteria, StepStopCriteria oraz TimeStopCriteria.

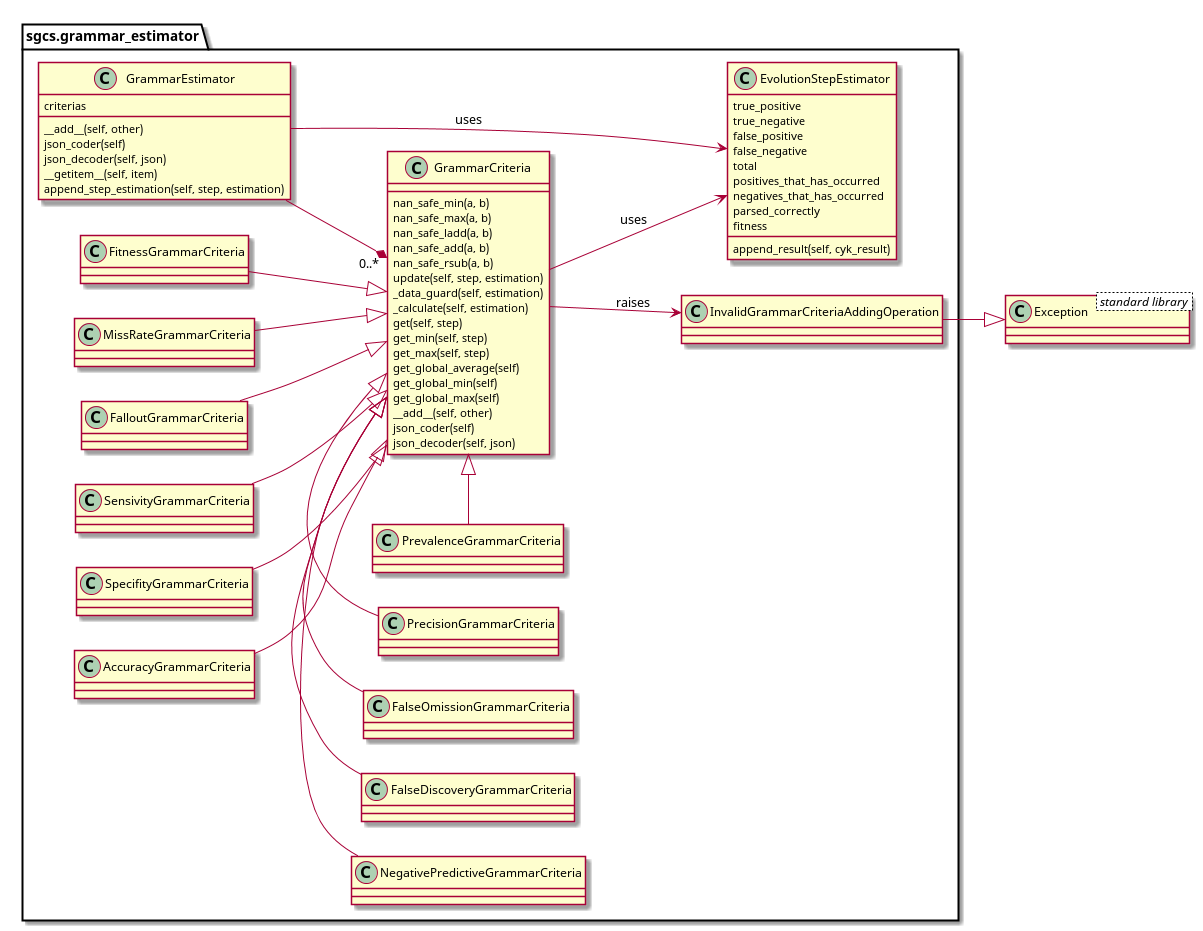
generate\_random\_rules(self, provided\_rules) – metoda zwiększająca zbiór reguł początkowych o configuration.rule.random\_starting\_population\_size – len(provided\_rules).

add\_initial\_rules(self, initial\_rules, rule\_population, grammar\_statistics) – metoda dodająca listę reguł startowych initial\_rules do populacji rule\_population z uwzględnieniem aktualizacji grammar\_statistics.

perform\_gcs(self, initial\_rules, configuration, grammar\_estimator, grammar\_statistics, sentences) – główna metoda algorytmu. Inicjuje zestaw reguł początkowych i wykonuje kolejne kroki ewolucyjne aż do momentu nastąpienia któregoś z warunków stopu.

\_post\_step\_actions(self, step) – metoda zawierająca zestaw działań wykonywanych po zakończeniu pojedynczego kroku ewolucyjnego. Istniejąca implementacja jest pusta, daje to jednak możliwości rozwoju klasom dziedziczącym (jak na przykład AsyncGcsRunner).

sgcs.grammar\_estimator

Moduł ten wykorzystuje się do gromadzenia informacji o kondycji algorytmu. Na podstawie ilości należących sparsowanych, nienależących sparsowanych, nienależących niesparsowanych oraz należących niesparsowanych zdań są obliczane wartości macierzy omyłek dla każdego kroku ewolucyjnego (oraz samego testu generalizacji).

EvolutionStepEstimator

Klasa przechowująca ilość zdań pozytywnych sparsowanych (True Positive), ilość zdań negatywnych niesparsowanych (True Negative), ilość zdań pozytywnych niesparsowanych (False Negative) oraz ilość zdań negatywnych sparsowanych (False Positive). Zawiera jeszcze kilka pomocniczych miar i sama zajmuje się kontrolą tych wszystkich wartości – wystarczy wywołać jej metodę append\_result wraz z obiektem CykResult zwróconym przez instancję CykService.

true\_positive – pole służące do odczytu ilości sparsowanych zdań prawdziwych.

true\_negative – pole służące do odczytu ilości niesparsowanych zdań negatywnych.

false\_positive – pole służące do odczytu ilości sparsowanych zdań negatywnych.

false\_negative – pole służące do odczytu ilości niesparsowanych zdań pozytywnych.

total – ilość wszystkich sparsowanych i niesparsowanych zdań

positives\_that\_has\_occurred – zmienna opisująca ilość pozytywnych zdań, które wystąpiły.

negatives\_that\_has\_occurred – zmienna opisująca ilość negatywnych zdań, które wystąpiły.

parsed\_correctly – ilość prawidłowo sparsowanych zdań.

fitness – osiągnięty w kroku poziom fitness.

GrammarCriteria

Klasa odpowiedzialna za gromadzenie analizę wielu kroków ewolucyjnych algorytmów. Przyjmuje kolejne obiekty EvolutionStepEstimator ze wszystkich uruchomień i na ich podstawie generuje daną statystykę na przestrzeni wielu kroków, w każdym z nich obliczając ją na podstawie wszystkich dostępnych przebiegów algorytmu. Wiele innych klas dziedziczy w tym module po tej klasie, nie ma jednak sensu poświęcać im osobnego miejsca na omówienie, jako że implementują one po prostu metody \_calculate oraz \_data\_guard zgodnie ze wzorem na statystykę, którą reprezentują (co jest zawarte w ich nazwie).

nan\_safe\_min(a, b) – funkcja min odporna na wartość 'nan' (zwraca drugą z wartości w przypadku gdy pierwsza nie jest liczbą).

nan\_safe\_max(a, b) – funkcja max odporna na wartość 'nan' (zwraca drugą z wartości w przypadku gdy pierwsza nie jest liczbą).

nan\_safe\_ladd(a, b) – operacja dodawania lewostronnie odporna na wartość 'nan' (zwraca b w przypadku gdy a nie jest liczbą).

nan\_safe\_add(a, b) – operacja dodawania odporna na wartość 'nan' (zwraca drugą z wartości w przypadku gdy pierwsza z nich nie jest liczbą).

nan\_safe\_rsub(a, b) – prawostronnie odporna na wartość 'nan' operacja odejmowania (zwraca a w przypadku gdy b nie jest liczbą).

update(self, step\_estimation) – aktualizuje przebieg statystyki na podstawie kolejnego obiektu step\_estimation.

\_data\_guard(self, estimation) – abstrakcyjna metoda, której konkretyzacje powinny zwracać True, jeżeli dla tego estimation dana statystyka istnieje (na przykład statystyka Precision nie jest zdefiniowana jeżeli żadne zdanie nie zostało sparsowane). Zazwyczaj wystarczy, aby funkcja ta pilnowała czy mianownik wzoru danej statystyki jest niezerowy.

\_calculate(self, estimation) – abstrakcyjna metoda, której konkretyzacje zajmują się liczeniem danej statystyki (dla poprawnych danych, funkcja ta wywołuje się tylko wtedy, kiedy metoda \_data\_guard zwróciła True).

get(self, step) – pobranie średniej wartości statystyki w danym kroku.

get\_min(self, step) – pobranie wartości minimalnej statystyki w danym kroku.

get\_max(self, step) – pobranie wartości maksymalnej statystyki w danym kroku.

get\_global\_average(self) – pobranie globalnej średniej statystyki na podstawie wszystkich kroków.

get\_global\_min(self) – pobranie globalnej wartości minimalnej statystyki na podstawie wszystkich kroków.

get\_global\_max(self) – pobranie globalnej wartości maksymalnej statystyki na podstawie wszystkich kroków.

\_\_add\_\_(self, other) – umożliwia dodawanie do siebie obiektów GrammarCriteria. W wyniku dodania dwóch obiektów tworzony jest trzeci, który posiada połączoną wiedzę dwóch pozostałych na temat gramatyki. Funkcja przydatna w przypadku asynchronicznego wykonania przebiegów uczących, kiedy to trzeba zgromadzić wiedzę zdobytą przez osobne procesy.

json\_coder(self) – metoda umożliwiająca serializację do napisu w formacie json.

json\_decoder(self, json) – metoda umożliwiająca wczytanie wewnętrznego stanu obiektu z obiektu json. Metoda nie sprawdza poprawności obiektu json.

GrammarEstimator

Klasa gromadząca wszystkie statystyki gramatyki. Przechowuje je w postaci mapy nazwa – obiekt GrammarCriteria w polu criterias. Udostępnia dostęp do tych statystyk przez nazwę, interfejs do dodawania kolejnych obiektów EvolutionStepEstimation oraz operację dodawania obiektów GrammarEstimator.

criterias – pole mapujące nazwę statystyki do obiektu GrammarCriteria. Domyślnie mapa ta jest zainicjowana statystykami z macierzy pomyłek, można ją dowolnie zmodyfikować.

\_\_add\_\_(self, other) – funkcja umożliwiająca dodawanie obiektów GrammarEstimator, szczególnie przydatna przy asynchronicznym wykonywaniu przebiegów uczących.

json\_coder(self) – metoda umożliwiająca serializację do napisu w formacie json.

json\_decoder(self, json) - metoda umożliwiająca wczytanie wewnętrznego stanu obiektu z obiektu json. Metoda nie sprawdza poprawności obiektu json.

\_\_get\_item\_\_(self, item) – metoda wbudowana wprowadzająca poniższy cukier syntaktyczny:

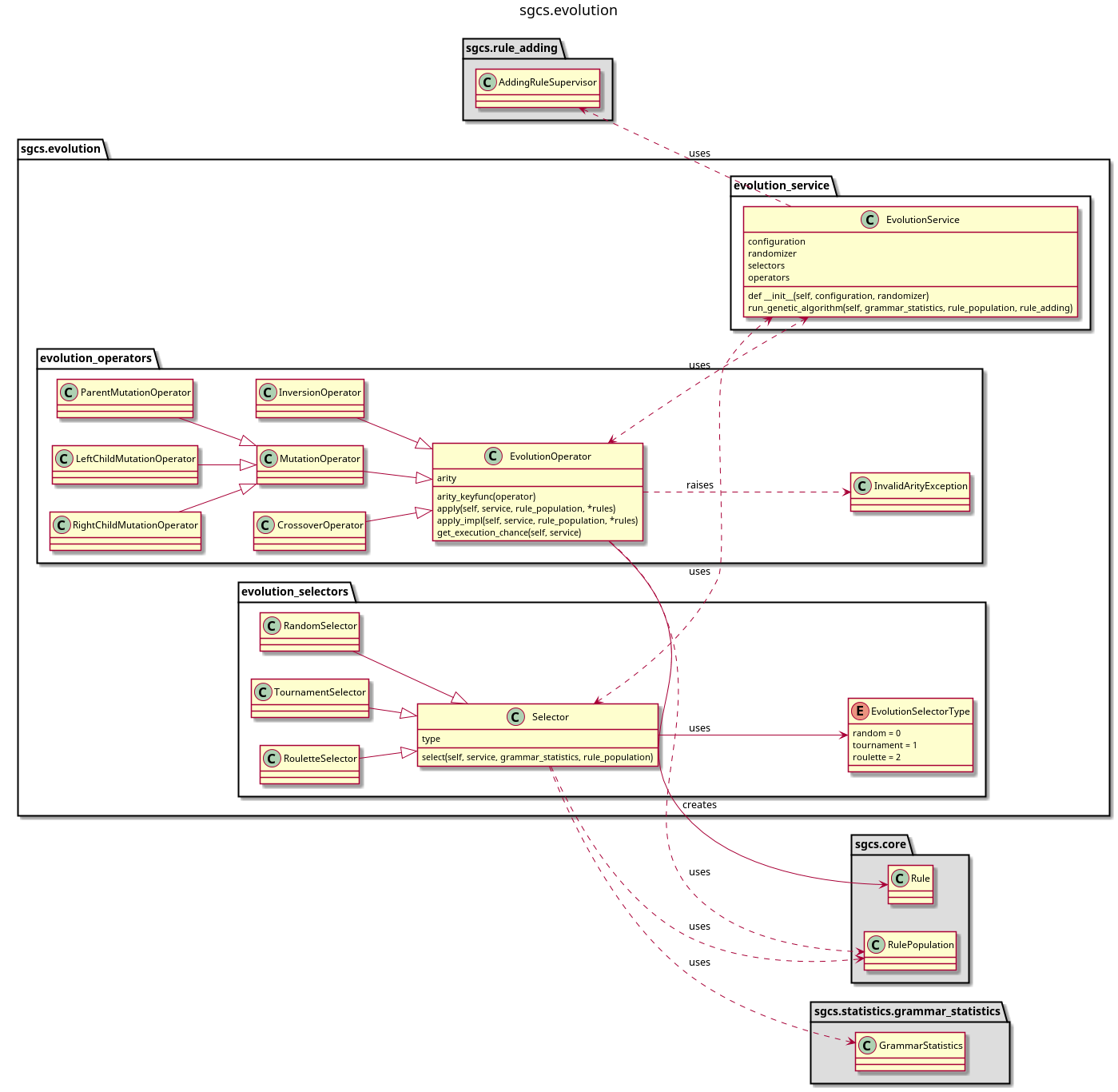
fitness\_criteria = grammar\_criteria['fitness']

Który tak naprawdę oznacza:

fitness\_criteria = grammar\_criterias.criterias['fitness']

append\_step\_estimation(self, step\_estimation) – metoda dodająca obiekt EvolutionStepEstimator step\_estimation do wszystkich statystyk w criterias (poprzez zawołanie na nich metody update).

sgcs.evolution



Moduł ewolucji zawiera całą logikę algorytmu genetycznego wykorzystywanego w algorytmie. Algorytm genetyczny można by w łatwy sposób podmienić na jakiś inny algorytm ewolucyjny, obecnie jednak moduł posiada jedynie implementację algorytmu genetycznego. Moduł ten ma na celu wygenerować nowe reguły na podstawie istniejącej populacji i jej stanu (przechowywanego w grammar\_statistics), a następnie przekazać wygenerowane reguły do modułu sgcs.rule\_adding.

EvolutionSelector

Klasa abstrakcyjna reprezentująca pojedynczy selektor.

type(self) – metoda zwracająca typ selektora (niezbędne do zgrania z węzłami konfiguracyjnymi algorytmu).

select(self, service, grammar\_statistics, rule\_population) – abstrakcyjna metoda dokonująca selekcji jednej reguły z rule\_population, przy pomocy wiedzy otrzymanej z grammar\_statistics oraz service (EvolutionService).

RandomSelector

Klasa dziedzicząca po EvolutionSelector, realizująca logikę selekcji losowej.

TournamentSelector

Klasa dziedzicząca po EvolutionSelector, realizująca logikę selekcji turniejowej.

RouletteSelector

Klasa dziedzicząca po EvolutionSelector, realizująca logikę selekcji ruletkowej.

EvolutionOperator

Abstrakcyjna klasa realizująca logikę operatora ewolucyjnego. Zdecydowano się tutaj na wysoce abstrakcyjną definicję operatora jako funkcję przyjmującą n reguł na wejściu i zwracającą k nowych reguł na wyjściu, co umożliwia dodawanie praktycznie dowolnych operatorów genetycznych oraz czyni model algorytmu genetycznego wysoce jednolitym. Szczegóły na temat wykorzystania operatorów o nietypowej n-arności i k-krotności wartości zwracanej zostaną podane przy okazji omówienia klasy EvolutionService.

arity – arność, czyli ilość argumentów przyjmowanych na wejściu przez algorytm.

arity\_keyfunc – metoda zwracająca wartość pola arity, wygodna w użyciu w funkcjach sortujących.

apply(self, service, rule\_population, \*rules) – metoda wywołująca operator genetyczny na podanych regułach. Jeżeli len(rules) != self.arity wówczas zostaje zgłoszony wyjątek InvalidArityException informujący o nieprawidłowej arności wywołania operatora genetycznego (operator musi być zawsze wywoływany dokładnie z tyloma argumentami, ile wynosi jego arność). Następnie metoda sprawdza, czy operator powinien zostać zaaplikowany (każdy operator genetyczny jest aplikowany z pewnym prawdopodobieństwem). Jeżeli tak zwraca wynik działania metody apply\_impl, w przeciwnym przypadku zwraca niezmodyfikowane reguły.

apply\_impl(self, service, rule\_population, \*rules) – abstrakcyjna metoda skupiająca się na wykonaniu samej operacji genetycznej.

get\_execution\_chance(self, service) – metoda zwracająca wartość prawdopodobieństwa zaaplikowania danego operatora (zazwyczaj wartość ta jest odczytywana z odpowiedniego pola service.configuration.operators).

CrossoverOperator

Binarny operator genetyczny krzyżownia.

InversionOperator

Unarny operator genetyczny inwersji.

MutationOperator

Abstrakcyjny unarny operator genetyczny mutacji. Jest ukonkretniony przez trzy klasy, które zajmują się osobno mutacją poszczególnych wartości reguły (tj. każdy z nich traktujemy jako niezależny operator, wykonując niezależne losowanie czy dana mutacja ma być zaaplikowana itd.).

ParentMutationOperator

Unarny operator genetyczny mutacji ojca reguły.

LeftChildMutationOperator

Unarny operator genetyczny mutacji lewego dziecka reguły.

RightChildMutationOperator

Unarny operator genetyczny mutacji prawego dziecka reguły.

EvolutionService

Klasa odpowiedzialna za obsługę całego algorytmu genetycznego. Działanie algorytmu genetycznego odbywa się w następujących krokach

1. Wybierz x reguł z powtórzeniami z populacji, gdzie x to ilość podanych w konfiguracji selektorów przez użytkownika (mogą być to selektory różnego rodzaju) i zapisz je jako selected\_rules.
2. Wybierz elitę populacji elite (jeżeli elityzm jest włączony).
3. Zgrupuj operatory względem arności.
4. Dla każdej dostępnej arności n:
   1. Dla każdego dostępnego operatora op o arności n i krotności wartości zwracanej k:
      1. Podziel selected\_rules na krotki długości n (powtarzając początek selected\_rules w ostatniej krotce jeżeli len(selected\_rules) % n != 0.
         1. Zaaplikuj op do wszystkich krotek, listy k wyników zsumuj i zapisz jako nowe selected\_rules.
5. Dodaj selected\_rules z elityzmem do populacji przy pomocy instancji klasy AddingRuleSupervisor z modułu sgcs.rule\_adding.

Grupowanie operatorów jest stabilne względem listy operators (tak więc kolejność dwóch operatorów o tej samej arności na liście nie ulegnie zmianie).

configuration - uchwyt do węzła konfiguracyjnego.

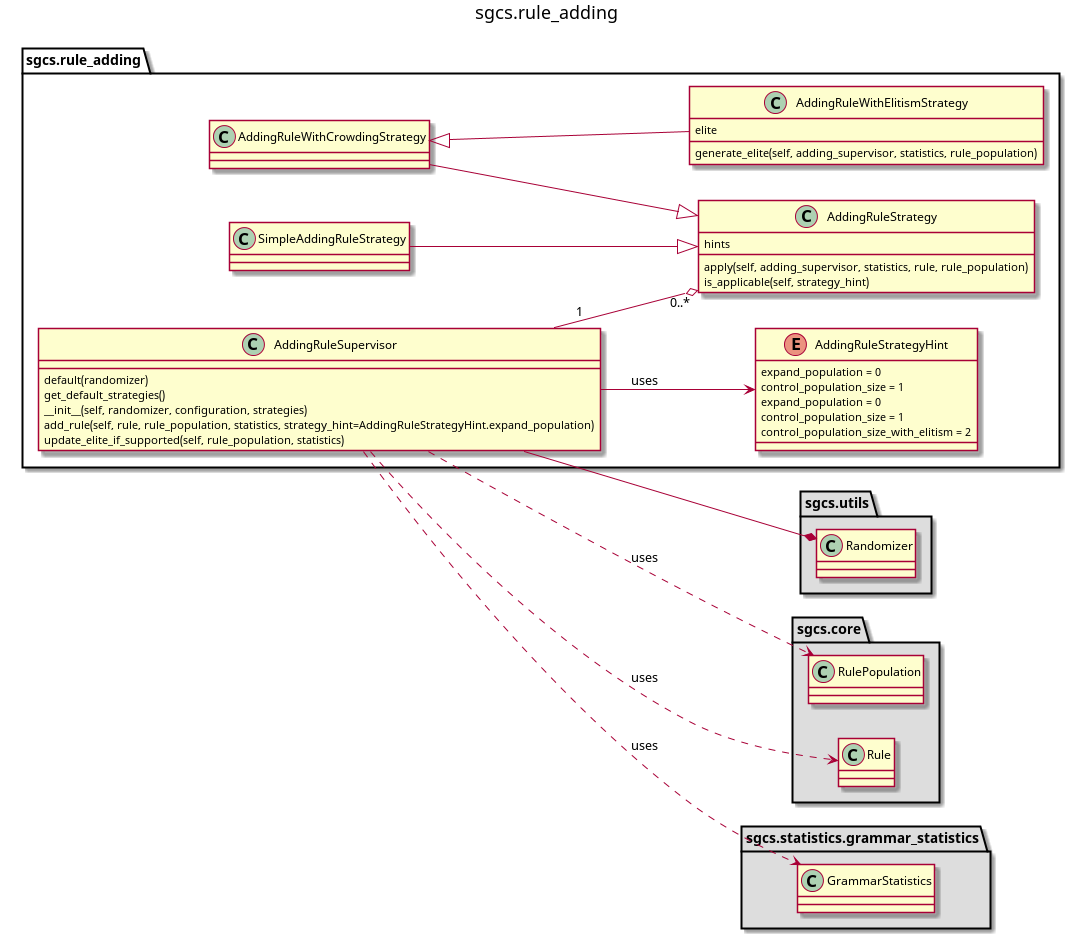
randomizer – uchwyt do instancji klasy Randomizer.

selectors – lista wspieranych selektorów (o rzeczywiście wykorzystanych decyduje konfiguracja).

operators – lista operatorów genetycznych.

run\_genetic\_algorithm(self, grammar\_statistics, rule\_population, rule\_adding, configuration) – Metoda uruchamiająca algorytm genetyczny.

sgcs.rule\_adding

Zadaniem tego modułu jest nadzorowanie dodawania nowych reguł do populacji. Praktycznie każde dodawanie nowej reguły odbywa się za pośrednictwem tego modułu.

AddingRuleSupervisor

Klasa odpowiedzialna za zarządzanie dodawaniem reguł.  
default(randomizer) – funkcja tworząca instancję o domyślnej konfiguracji.

get\_default\_strategies() - funkcja zwracająca listę zawierającą domyślny zestaw strategii.

add\_rule(self, rule, rule\_population, statistics, strategy\_hint = AddingRuleStrategyHint.expand\_population) – metoda powodująca dodanie reguły do populacji z uwzględnieniem strategy\_hint. Należy tutaj zwrócić uwagę na słowo „hint” - AddingRuleSupervisor może nadpisać to zachowanie w jednym z przypadków:

* Jeżeli strategy\_hint == AddingRuleStrategyHint.expand\_population, a osiągnęliśmy już maksymalny rozmiar populacji, wówczas będzie miało dodanie reguły z AddingRuleStrategyHind.control\_population\_size;
* Jeżeli strategy\_hint wymuszałby dodanie reguły do pustej populacji bez dokonywania zmiany jej rozmiaru, wówczas supervisor doda populację, zwiekszając tym samym jej rozmiar (do 1).

Pomijając te wyjątkowe sytuacje AddingRuleSupervisor znajdzie obiekt odpowiedniej strategii i z jej pomocą doda regułę do populacji.

update\_elite\_if\_supported(self, rule\_population, statistics) – funkcja aktualizująca stan elity przechowywanej przez funkcję. Metoda ta musi być wywoływana osobno, gdyż w przypadku dodawania kilku nowych reguł elita powinna być sprecyzowana przed dodaniem jakiejkolwiek z nich, nie powinna być generowana przy każdym dodaniu reguły.

AddingRuleStrategyHint

Klasa ta pełni funkcję wyliczenia z następującymi możliwymi wartościami:

* expand\_population;
* control\_population\_size;
* control\_population\_size\_with\_elitism.

AddingRuleStrategy

Abstrakcyjna klasa definiująca podstawową logikę funkcjonowania strategii dodawania reguł.

hints – lista AddingRuleStrategyHint, w przypadku których strategia jest aplikowalna.

apply(self, adding\_supervisor, statistics, rule, rule\_population) – metoda abstrakcyjna powodująca zaaplikowanie danej strategii.

is\_applicable(self, strategy\_hint) – metoda zwracająca True, jeżeli daną strategię da się zastosować z danym strategy\_hint.

SimpleAddingRuleStrategy

Najprostsza strategia dodawania reguł – poprzez rozszerzenie zbioru. Reguła zostaje dodana, obiekt statistics poinformowany o zmianach.

AddingRuleWithCrowdingStrategy

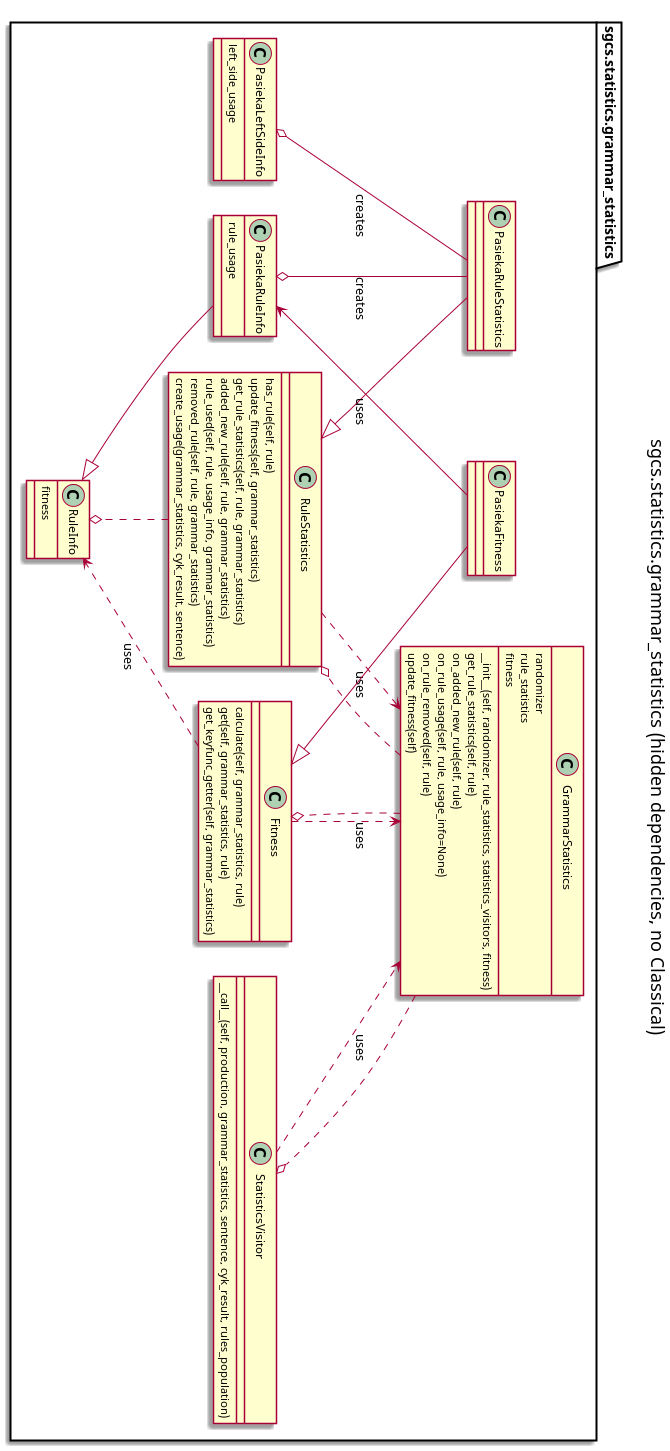
Strategia wykorzystująca mechanizm ścisku do dodawania reguł.

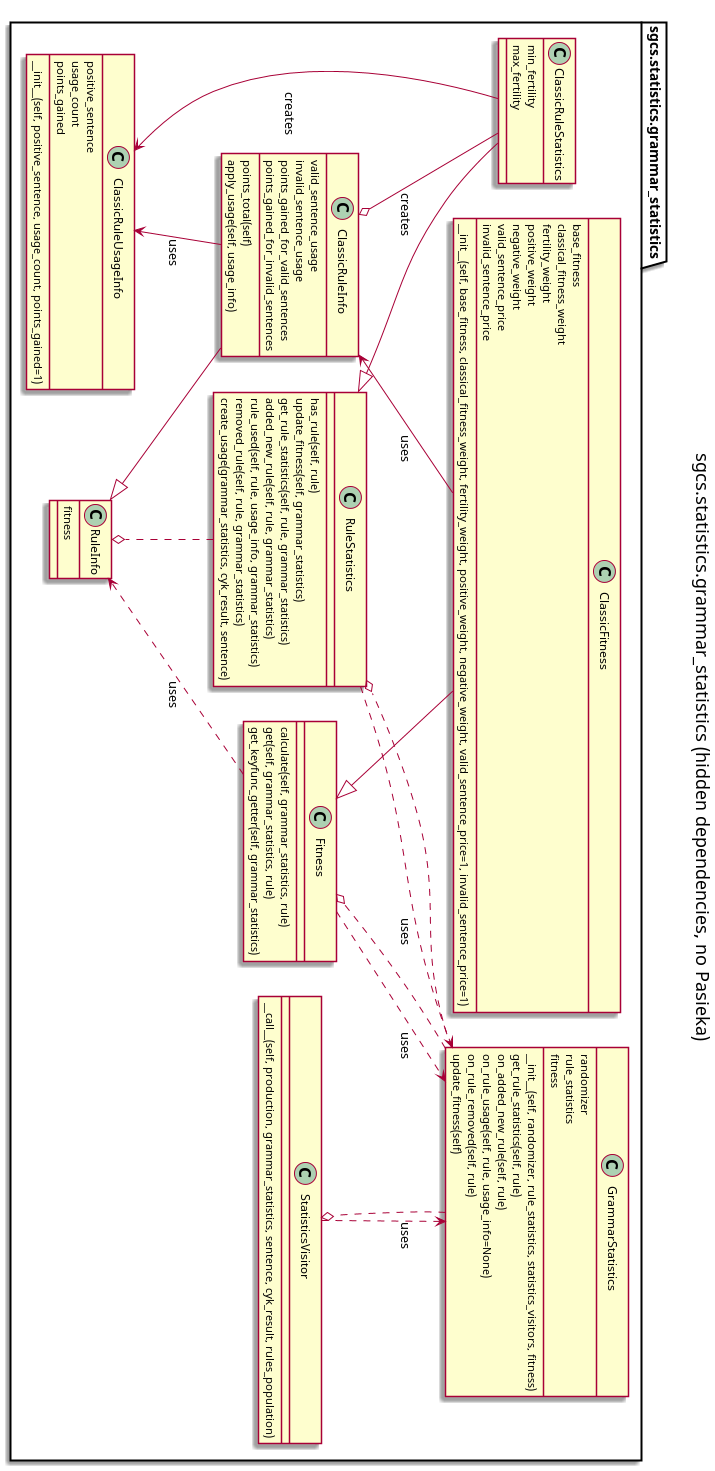
AddingRuleWithElitismStrategy

Strategia wykorzystująca mechanizm ścisku oraz elityzm do dodawania reguł.

generate\_elite(self, adding\_supervisor, statistics, rule\_population) – funkcja generująca elitę. Nie trzeba jej ręcznie wywoływać – zostanie automatycznie wywołana przez metodę AddingRuleSupervisor.update\_elite\_if\_supported.

sgcs.statistics.grammar\_statistics





Powyższe dwa diagramy dopiero po połączeniu dają całkowity obraz modułu sgcs.statistics.grammar\_statistics. Niestety nawet bez dołączania zewnętrznych zależności modułu jest to obraz wysoce nieczytelny. Dlatego też zdecydowano się na przedstawienie modułu w postaci dwóch diagramów – na jednym ukryto klasy związane z funkcją przystosowania Pasieka, na drugim ukryto klasy związane z klasyczną funkcją przystosowania. Obie grupy są od siebie wzajemnie niezależne, natomiast część wspólna obu diagramów została zdublowana.

Moduł ten zajmuje się obliczaniem wartości funkcji przystosowania.

GrammarStatistics

Główna klasa zajmująca się obsługą parametru fitness. Większość jej funkcji oddelegowuje swoją logikę do obiektu rule\_statistics, co czyni ją praktycznie fasadą tejże klasy.

randomizer – uchwyt obiektu typu Randomizer.

rule\_statistics – instancja obiektu dziedziczącego po abstrakcyjnym RuleStatistics. Klasa ta zajmuje się obsługą większości logiki GrammarStatistics, pozostawiając mu jedynie wspólną dla wszystkich funkcji fitness logikę.

fitness – instancja obiektu dziedziczącego po klasie Fitness.

statistics\_visitors – lista instancji obiektu StatisticsVisitor. Jest on na tyle generyczny, że bez względu na ilość i rodzaj obsługiwanych funkcji fitness nie ma potrzeby umieszczania na niej więcej niż jednego obiektu tego typu – inne obiekty mogłyby być przydatne przy zastosowaniu naprawdę niestandardowych funkcji przystosowania (patrz opis klasy StatisticsVisitor).

default(randomizer, configuration) – funkcja tworząca instancję GrammarStatistics przygotowaną pod względem liczenia funkcji klasycznej.

sgcs\_variant(randomizer, configuration) – funkcja tworząca instancję GrammarStatistics przygotowaną pod względem liczenia funkcji Pasieka.

get\_rule\_statistics(self, rule) – metoda zwracająca statystyki reguły rule (postać zależna od implementacji statystyk).

on\_added\_new\_rule(self, rule) – metoda obsługująca utworzenie nowej reguły.

on\_rule\_usage(self, rule\_usage, usage\_info=None) – metoda wywoływana, gdy reguła zostanie użyta. Można dołączyć opcjonalny parametr UsageInfo (umożliwiający zbieranie dodatkowych informacji przez bardziej skomplikowane funkcje fitness jak na przykład funkcja klasyczna).

on\_rule\_removed(self, rule) – metoda obsługująca usunięcie reguły.

update\_fitness(self) – przeliczenie funkcji fitness na podstawie obecnie zebranych statystyk.

RuleStatistics

Klasa abstrakcyjna, której konkretyzacje realizują większość logiki GrammarStatistics. Posiada ona dwie dodatkowe metody:

has\_rule(self, rule) – zwraca True, jeżeli reguła rule już istnieje.

create\_usage(grammar\_statistics, cyk\_result, sentence) – tworzy obiekt UsageInfo, jeżeli jest on wspierany przez funkcję fitness. Obiekty te nie mają określonego z góry interfejsu, jako że są to obiekty służące właśnie do przekazywania danych typowych wyłącznie dla poszczególnych funkcji fitness i to one są ich producentem i konsumentem. Jeżeli nie ma potrzeby wykorzystania takiego obiektu, zwrócona zostanie wartość None.

RuleInfo

Klasa przechowująca informacje niezbędne do wyliczenia wartości fitness (oraz obecną wartość tegoż parametru).

StatisticsVisitor

Obiekt ten powstał w celu bycia wykorzystanym przez algorytm traceback. W chwili kiedy zostanie wywołany jak funkcja, w odpowiedni sposób zapisze informację o wykorzystaniu reguły. Ukrycie logiki liczenia funkcji fitness czyni go wysoce uniwersalnym, dzięki czemu wydaje się, że raczej nie będzie konieczności dodawania kolejnych wizytatorów (przy założeniu, że funkcje przystosowania nie zaczną wymagać wysoce specyficznych danych).

Fitness

Abstrakcyjna klasa reprezentująca sam wzór na funkcję przystosowania.

calculate(self, grammar\_statistics, rule) – abstrakcyjna metoda licząca wartość funkcji przystosowania.

get(self, grammar\_statistics, rule) – metoda zwracająca wartość funkcji przystosowania. Jeżeli wartość ta nie została jeszcze nigdy policzona, wówczas wywołanie tej metody spowoduje te obliczenia.

get\_keyfunc\_getter(self, grammar\_statistics) – metoda zwracająca delegata funkcji get, który jest wygodny w zastosowaniu w wyrażeniach funkcyjnych (jak na przykład polecenie map() ).

Classic\_\_\_

Funkcja klasyczna wykorzystuje szereg różnorodnych klas w celu wyliczenia wartości fitness. Są to:

* ClassicRuleStatistics;
* ClassicRuleInfo;
* ClassicRuleUsageInfo;
* ClassicFitness.

Klasy te nie wymagają większego komentarza – są ukonkretnionymi klasami omówionych klas abstrakcyjnych, mających za zadanie obsłużyć liczenie klasycznej funkcji fitness. Warto tutaj jedynie zauważyć fakt istnienia klasy ClassicRuleUsageInfo – w klasie tej są zbierane informacje jak punktów zdobytych przez regułę, czy zdanie było pozytywne oraz pozostałe informacje wymagane przez klasyczną funkcję przystosowania.

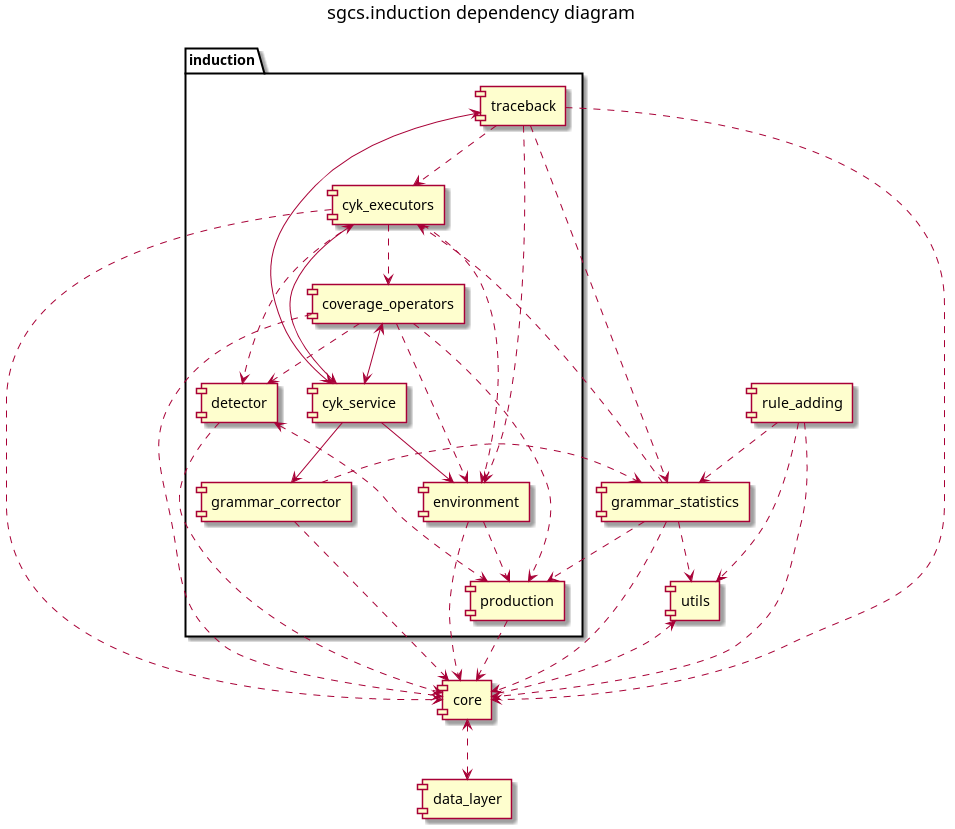
Pasieka\_\_\_

Funkcja Pasieka również posiada swój zestaw klas niezbędnych do wyliczenia wartości przystosowania. Są to:

* PasiekaRuleStatistics;
* PasiekaFitness;
* PasiekaRuleInfo;
* PasiekaLeftSideInfo.

PasiekaLeftSideInfo gromadzi informacje na temat wykorzystania reguł o Symbolu lewej strony produkcji, trzymanie tej informacji jest niezbędne i nie ma sensu jej duplikować w każdej regule posiadającej ten symbol w lewej części produkcji. Pokazuje to równocześnie uniwersalność modułu, który może ukryć w ten sposób pod omówionymi wcześniej klasami spore fragmenty nietrywialnej mechaniki, nie wymagając przy tym samym modyfikacji interfejsu.

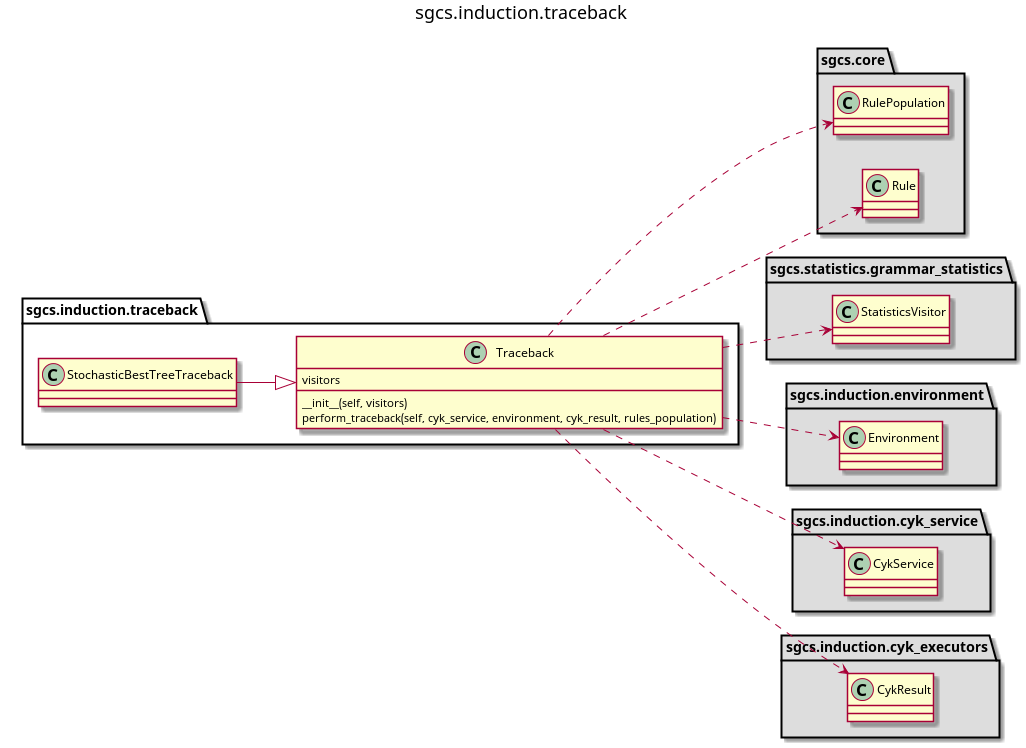
Właściwy algorytm – część indukcyjna

Powyżej znajduje się diagram zależności modułów indukcyjnej części algorytmu genetycznego. Jest to oczywiście najważniejsza i najbardziej złożona część algorytmu. Wszystkie moduły związane z indukcją gramatyki umieszczono w przestrzeni nazw sgcs.induction.

sgcs.induction.cyk\_service

Moduł ten zawiera dwie klasy – CykService oraz dziedziczącą po niej StochasticCykService. Klasy te gromadzą obiekty z całego modułu indukcji i składają je w jedną, dostępną dla wszystkich składowych strukturę, w zależności od wykorzystanej klasy struktura ta jest nastawiona na indukcję gramatyk stochastycznych lub klasycznych. Znajduje się tu też główna pętla indukcji gramatyki.

sgcs.induction.traceback



Moduł ten zawiera klasy implementujące rożne warianty algorytmu traceback. Wykorzystują one wzorzec wizytatora – traceback zapewnia jedynie spójny interfejs do poruszania się po drzewie rozbioru zdania, zaś sam wizytator decyduje w jaki sposób wykorzysta zebrane podczas wędrówki dane. Pozwala to na separację systemu nagradzania odwiedzonych węzłów od samej metody poruszania się po drzewie rozbioru.

Traceback

Standardowa implementacja algorytmu traceback wykorzystywana przez algorytm GCS. Algorytm ten gorliwie przechodzi po wszystkich możliwych drzewach rozkładu zdania.

visitors – lista wizytatorów. Wizytatorzy muszą mieć następującą sygnaturę:

1. Aktualnie odwiedzana produkcja.
2. Obiekt GrammarStatistics.
3. Obiekt Sentence (reprezentacja całego zdania).
4. Obiekt CykResult.
5. Obiekt RulePopulation.

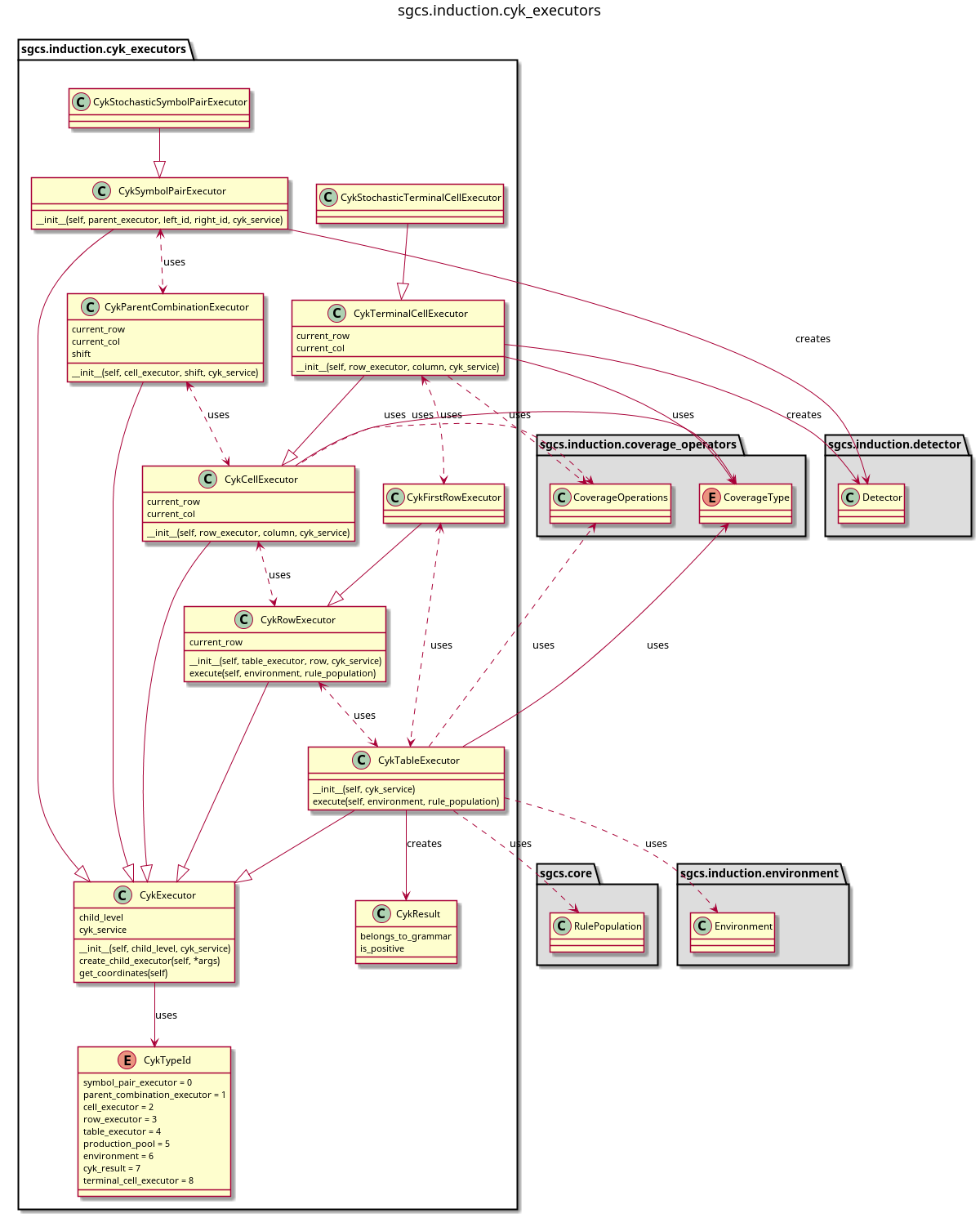
Kolejność tych parametrów ma znaczenie.

perform\_traceback(self, cyk\_service, environment, cyk\_result, rules\_population) – metoda powodująca wywołanie algorytmu traceback. Obiekt environment stanowi reprezentację tablicy CYK, cyk\_result zawiera między innymi werdykt przydzielony mu przez algorytm oraz informację o tym, czy zdanie w rzeczywistości należy do gramatyki, czy też nie.

StochasticBestTreeTraceback

Implementacja algorytmu traceback wykorzystywana przez algorytmy stochastyczne. Algorytm ten przechodzi jedynie po najbardziej prawdopodobnym drzewie rozkładu zdania. Dziedziczy po klasie Traceback i nie rozszerza istniejącego interfejsu.

sgcs.induction.cyk\_executors

Moduł sgcs.induction.cyk\_executors jest rozbitą na czynniki pierwsze pętlą algorytmu CYK. Kolejni wykonawcy skupiają się na coraz głębszych elementach algorytmu CYK, wszyscy są tworzeni przez odpowiednio przygotowaną wcześniej fabrykę. Takie rozdrobnienie tego kluczowego miejsca algorytmu umożliwia łatwe modyfikowanie funkcjonowania procesu indukcji.

CykTypeId

Wyliczenie przedstawiające identyfikatory typów klas, z jakimi mamy do czynienia w module. Jest ono wykorzystywane przez fabrykę obiektów, w efekcie czego żadna z klas powiązanych z poniższymi wartościami nie jest tworzona w inny sposób niż przy pomocy fabryki. Domyślnie moduł ten promuje następujący podział klas (który może zostać rozszerzony, jeżeli zmienimy zachowanie fabryki:

1. table\_executor – klasa wykorzystywana na najwyższym poziomie, odpowiada za wypełnianie całej tablicy CYK.
2. row\_executor – klasa odpowiadająca za uzupełnienie całego wiersza tablicy CYK.
3. cell\_executor – klasa odpowiedzialna za uzupełnienie pojedynczej komórki tabeli CYK.
4. parent\_combination\_executor – klasa odpowiedzialna za wpisanie do komórki tabeli CYK reguł powstałych z reguł z pojedynczej kombinacji komórek – rodziców.
5. symbol\_pair\_executor – najniższy poziom uzupełniania tabeli CYK, generująca reguły powstałe z pojedynczej pary symboli.
6. environment – klasa zarządzająca zawartością tabeli CYK.
7. production\_pool – klasa zarządzająca zawartością pojedynczej komórki tabeli CYK.
8. cyk\_result – klasa niosąca informacje o rezultacie parsowania (oraz o prawdziwej przynależności zdania).

Klasy oznaczone identyfikatorami typów 1-5 są klasami wykonawców (uporządkowanych malejąco względem ich odpowiedzialności). Klasa bezpośrednio kontrolowana przez wykonawcę nadrzędnego będzie od tej pory nazywana „podwykonawcą”.

CykResult

Prosta klasa przechowująca wynik parsowania zdania przy pomocy algorytmu CYK oraz prawdziwą przynależność zdania.

belongs\_to\_grammar – zmienna boolowska ustalająca czy zdanie rzeczywiście należy do gramatyki.

is\_positive – zmienna boolowska stwierdzająca czy CYK zdołał sparsować zdanie.

CykExecutor

Abstrakcyjna klasa dostarczająca trochę podstawowej logiki klasom wykonawców. Zawiera uchwyt do instancji CykService, jest w stanie stworzyć podwykonawcę oraz podać swoje koordynaty w tabeli CYK.

child\_level – identyfikator podwykonawcy.

cyk\_service – uchwyt instancji CykService.

create\_child\_executor(self, \*args) – metoda zwracająca obiekt podwykonawcy. Musimy wiedzieć dokładnie jakie parametry są wymagane przez konstruktor.

get\_coordinates(self) – metoda zwracająca koordynaty wykonawcy w tabeli CYK.

CykTableExecutor

Klasa najwyższego poziomu algorytmu CYK. Jest w stanie ustalić, czy zdane zostało sparsowane przez tabelę CYK, może uruchomić operatory pokrycia zarejstrowane na CoverageType.no\_starting\_symbol oraz zlecić podwykonawcom wypełnianie wierszy tabeli CYK. Jej podwykonawcy to CykFirstRowExecutor (dla pierwszego wiersza) oraz CykRowExecutor (dla kolejnych).

\_\_init\_\_(self, cyk\_service) – konstruktor.

execute(self, environment, rule\_population) – metoda wywołująca algorytm CYK.

CykRowExecutor

Podwykonawca klasy CykTableExecutor. Przeprowadza algorytm CYK na poziomie pojedynczego wiersza, przeznaczony do wszystkich wierszy powyżej pierwszego). Jego podwykonawcą jest CykCellExecutor.

current\_row – zwraca numer obecnego wiersza.

\_\_init\_\_(self, table\_executor, row, cyk\_service) – kontruktor. Przekazywanie siebie podwykonawcy często ma miejsce w klasach wykonawców.

CykFirstRowExecutor

Klasa uzupełniająca pierwszy wiersz tabeli CYK. Dziedziczy po CykRowExecutor. Jej podwykonawcą jest CykTerminalStateExecutor lub CykStochasticTerminalCellExecutor (w przypadku algorytmu stochastycznego).

CykCellExecutor

Klasa uzupełniająca pojedynczą komórkę (powyżej pierwszego wiersza). Może wywołać operatory pokrycia zarejestrowane na CoverageType.no\_effector\_type oraz zlecić podwykonawcom wpisywanie przodukcji powstałych z konkretnych par komórek – rodziców. Jej podwykonawcą jest CykParentCombinationExecutor.

current\_col – zwraca numer obecnej kolumny.

CykCellTerminalExecutor

Klasa uzupełniająca pojedynczą komórkę pierwszego wiersza tabeli CYK. W przypadku nieznalezienia żadnych produkcji może wywołać operatory pokrycia zarejestrowane na typ pokrycia CoverageType.unknown\_terminal\_symbol. Nie posiada podwykonawcy.

CykStochasticTerminalCellExecutor

Klasa dziedzicząca po CykCellTerminalExecutor, uzupełniająca dodatkowo prawdopodobieństwa produkcji.

CykParentCombinationExecutor

Klasa tworząca wszystkie produkcje powstałe z jednej pary komórek – rodziców. Jej podwykonawcą jest CykSymbolPairExecutor lub CykStochasticSymbolPairExecutor w przypadku gramatyk stochastycznych.

shift – koordynat przesunięcia w tabeli CYK – określa jednoznacznie koordynaty komórek rodziców.

CykSymbolPairExecutor

Klasa tworząca produkcje powstałe z pojedynczej kombinacji symboli. Nie ma podwykonawcy.

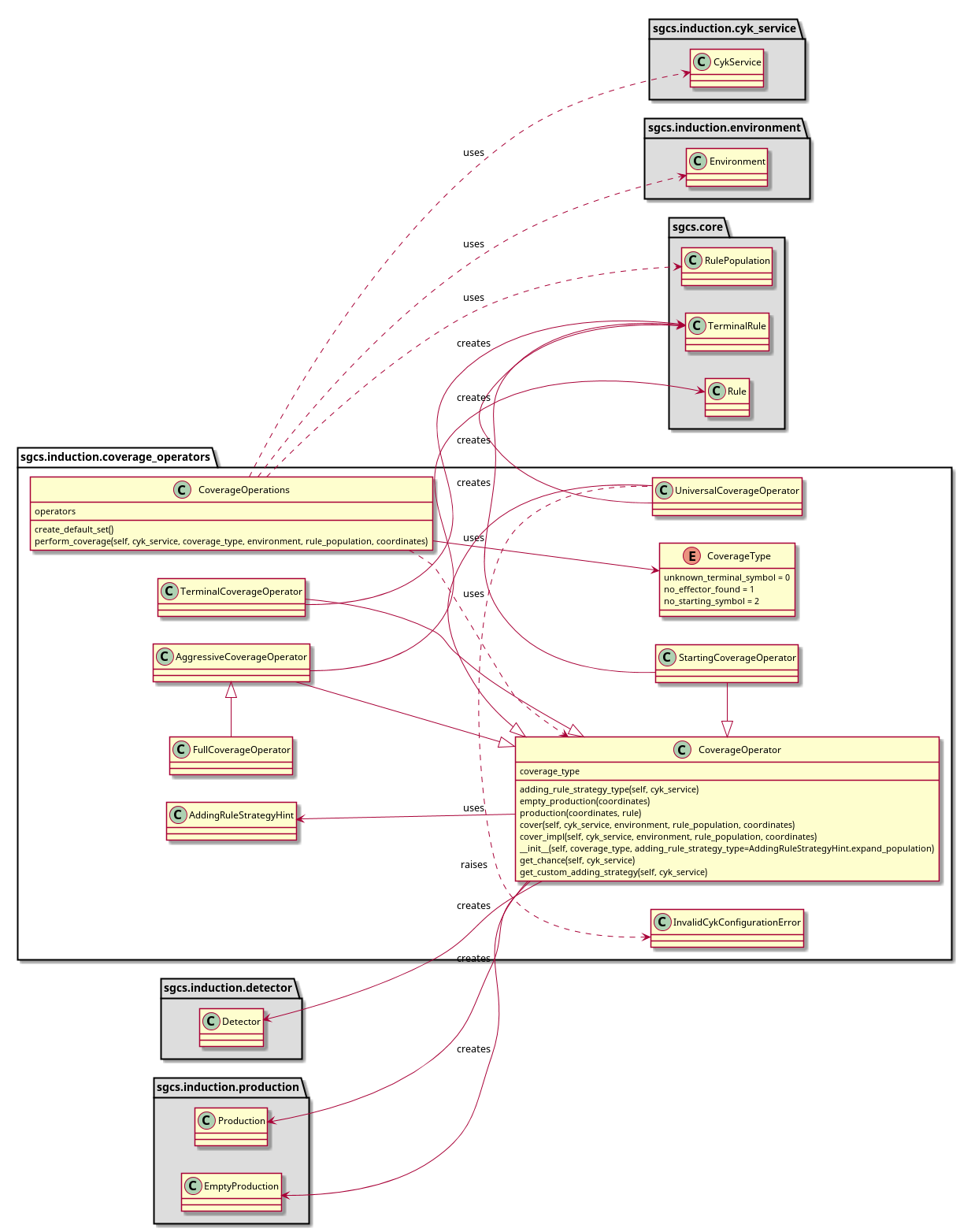
left\_id – koordynat lewego symbolu – jednoznacznie określa symbol z lewej komórki odpowiedzialny za tworzone produkcje.

right\_id – koordynat prawego symbolu – jednoznacznie określa symbol z prawej komórki odpowiedzialny za tworzone produkcje.

CykStochasticSymbolPairExecutor

Klasa rozszerzająca CykSymbolPairExecutor o aktualizację prawdopodobieństw produkcji.

sgcs.induction.coverage\_operators

Moduł sgcs.induction.coverage\_operators dostarcza implementację standardowych operatorów pokrycia. Zostały one oddzielone od modułu sgcs.induction.cyk\_executors przy pomocy identyfikatorów typu pokrycia CoverageType. Kiedy w głównej pętli algorytmu następuje potrzeba uruchomienia operatorów pokrycia, wówczas jest wywoływana metoda CoverageOperations.perform\_coverage z odpowiednim identyfikatorem typu pokrycia, na który zareagują wyłącznie operatory pokrycia zarejestrowane na ten typ. Na przykład przy natrafieniu na nieznany symbol nieterminalny jest wykorzystywany identyfikator CoverageType.unknown\_terminal\_symbol, na który reagują odpowiednio operatory pokrycia terminalnego oraz uniwersalnego. Umożliwia to łatwe rozszerzanie algorytmu o nowe operatory pokrycia bez potrzeby wprowadzania modyfikacji do głównej pętli algorytmu, zmiany dotyczące operatorów pokrycia dokonuje się wyłącznie w module operatorów pokrycia. Po stworzeniu nowych reguł są one dodawane do populacji reguł z pomocą klasy AddingRuleSupervisor. Operatory pokrycia posiadają domyślne strategie dodawania reguł (na przykład operator pokrycia agresywnego zawsze dodaje reguły ze ściskiem), aczkolwiek można nadpisać to zachowanie przy pomocy odpowiedniego ustawienia węzła konfiguracyjnego tego operatora.

W celu ujednolicenia modelu wszystkie operatory posługują się prawdopodobieństwem – oznacza to, że jest możliwe uruchomienie na przykład operatora pokrycia uniwersalnego z pewnym prawdopodobieństwem (zamiast dyskretnie zawsze lub nigdy).

CoverageType

Typ wyliczeniowy przechowujący dostępne rodzaje operatorów pokrycia – są to obecnie unknown\_terminal\_symbol, no\_effector\_found oraz no\_starting\_symbol. CoverageType najłatwiej zrozumieć jako powód, dla którego może być konieczne uruchomienie operatora pokrycia – uruchamiamy je gdy natrafimy na nieznany symbol gramatyki, nie udało się wpisać nic do jednej z komórek tabeli CYK lub komórka nie zawiera symbolu startowego gramatyki.

CoverageOperations

Klasa agregująca operatory pokrycia. Jest zdolna do wygenerowania ich podstawowego zestawu oraz uruchomienia operatorów pokrycia danego typu, kiedy zaistnieje taka potrzeba.

operators – lista operatorów pokrycia

create\_default\_set() - funkcja zwracająca domyślną listę (czyli zawierającą wszystkie dostarczone wraz z tym modułem) operatorów pokrycia.

perform\_coverage(self, cyk\_service, coverage\_type, environment, rule\_population, coordinates) – funkcja ta znajduje wszystkie operatory zarejestrowane na coverage\_type i uruchamia dany operator (który ostatecznie zostanie uruchomiony lub też nie, zależnie od prawdopodobieństwa oraz innych dodatkowych warunków wymaganych przez operator). Jeżeli udało się wygenerować nową produkcję, to jest ona dodawana do tabeli CYK, zaś nowa reguła z nią związana dodana z pomocą AddingRuleSupervisor do populacji reguł. Metoda próbuje zachłannie wszystkich operatorów pokrycia.

CoverageOperator

Klasa abstrakcyjna dostarczająca części wspólnej logiki operatorom pokrycia.

coverage\_type – powiązany z operatorem identyfikator typu pokrycia.

\_\_init\_\_(self, coverage\_type, adding\_rule\_strategy\_type = AddingRuleStrategyHint.expand\_population) – konstruktor, pobierający typ pokrycia operatora oraz domyślną strategię dodawania nowych reguł.

adding\_rule\_strategy\_type(self, cyk\_service) – zwraca domyślną dla operatora strategia dodawania nowych reguł, chyba że została ona nadpisana w konfiguracji własną strategią.

empty\_production(coordinates) – Funkcja pomocnicza zwracająca obiekt pustej produkcji dla koordynatów coordinates.

production(coordinates, rule) – Funkcja pomocnicza zwracająca obiekt produkcji powiązany z regułą rule i koordynatami coordinates.

cover(self, cyk\_service, environment, rule\_population, coordinates) – metoda z prawdopodobieństwem wywołania operatora pokrycia zwraca rezultat metody cover\_impl (której zachowanie jest zależne od implementacji konkretnego operatora pokrycia, ale spodziewa się od niej zwrócenia nowej produkcji, lub też nie).

cover\_impl(self, cyk\_service, environment, rule\_population, coordinates) – abstrakcyjna metoda, która jeżeli warunki uruchomienia operatora są spełnione (na przykład operator pokrycia startowego próbuje pokryć zdanie należące do gramatyki) wykona odpowiednią operację pokrycia.

get\_chance(self, cyk\_service) – abstrakcyjna metoda zwracająca prawdopodobieństwo uruchomienia operatora pokrycia (zazwyczaj jest to wartość zapisana w którymś z węzłów konfiguracji).

get\_custom\_adding\_strategy(self, cyk\_service) – abstrakcyjna metoda zwracająca podaną własną strategię dodawania reguł (zazwyczaj jest to wartość zapisana w którymś z węzłów konfiguracji).

TerminalCoverageOperator

Klasa reprezentująca operator pokrycia terminalnego. Typ pokrycia tego operatora to CoverageType.unknown\_terminal\_symbol. Indywidualna strategia dodawania reguł jest zlokalizowana w cyk\_service.configuration.coverage.operators.terminal.adding\_hint, zaś prawdopodobieństwo w cyk\_service.configuration.coverage.operators.terminal.chance. Domyślna strategia dodawania nowych reguł to AddingRuleStrategyHint.expand\_population.

UniversalCoverageOperator

Klasa reprezentująca operator pokrycia uniwersalnego. Typ pokrycia tego operatora to CoverageType.unknown\_terminal\_symbol. Indywidualna strategia dodawania reguł jest zlokalizowana w cyk\_service.configuration.coverage.operators.universal.adding\_hint, zaś prawdopodobieństwo w cyk\_service.configuration.coverage.operators.universal.chance. Domyślna strategia dodawania nowych reguł to AddingRuleStrategyHint.expand\_population. Korzystanie z tego operatora wymaga sprecyzowania symbolu uniwersalnego w populacji reguł – jeżeli symbol ten nie jest zdefiniowany, przy próbie użycia operatora zostanie zgłoszony wyjątek InvalidCykConfigurationError (gdyż oznacza to sprzeczność w samej konfiguracji).

StartingCoverageOperator

Klasa reprezentująca operator pokrycia startowego. Typ pokrycia tego operatora to CoverageType.no\_starting\_symbol. Indywidualna strategia dodawania reguł jest zlokalizowana w cyk\_service.configuration.coverage.operators.starting.adding\_hint, zaś prawdopodobieństwo w cyk\_service.configuration.coverage.operators.starting.chance. Domyślna strategia dodawania nowych reguł to AddingRuleStrategyHint.control\_population\_size. Operator ten posiada dodatkowy warunek uruchomienia – można go wykorzystać wyłącznie w zdaniu pozytywnym.

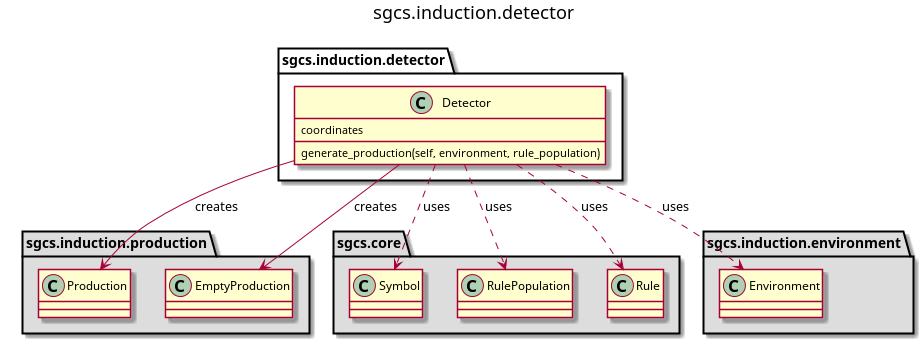
AggressiveCoverageOperator

Klasa reprezentująca operator pokrycia agresywnego. Typ pokrycia tego operatora to CoverageType.no\_effector\_found. Indywidualna strategia dodawania reguł jest zlokalizowana w cyk\_service.configuration.coverage.operators.aggressive.adding\_hint, zaś prawdopodobieństwo w cyk\_service.configuration.coverage.operators.aggressive.chance. Domyślna strategia dodawania nowych reguł to AddingRuleStrategyHint.control\_population\_size.

FullCoverageOperator

Klasa reprezentująca operator pokrycia pełnego. Dziedziczy po AggressiveCoverageOperator. Typ pokrycia tego operatora to CoverageType.no\_starting\_symbol. Indywidualna strategia dodawania reguł jest zlokalizowana w cyk\_service.configuration.coverage.operators.full.adding\_hint, zaś prawdopodobieństwo w cyk\_service.configuration.coverage.operators.full.chance. Domyślna strategia dodawania nowych reguł to AddingRuleStrategyHint.control\_population\_size.

sgcs.induction.detector

Moduł zawierający klasę Detector.

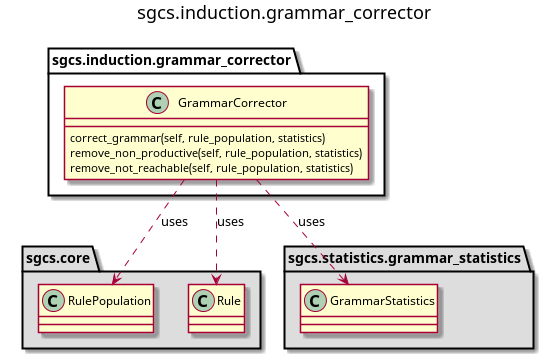
Detector

Wygodny wrapper na koordynaty produkcji. Zawiera również logikę generowania produkcji na podstawie symboli jednoznacznie wskazywanych przez krotkę coordinates w tabeli CYK.

coordinates – jednoznaczne określenie położenia symboli wskazywanych przez detektor. Dysponując tablicą CYK można powiedzieć, których dwóch symboli dotyczy.

generate\_productions(self, environment, rule\_population) – metoda znajdująca wszystkie możliwe produkcje dla danego detektora.

sgcs.induction.grammar\_corrector

Moduł zawierający klasy odpowiedzialne za korekcję gramatyki. Obecnie wspiera usuwanie reguł nieproduktywnych oraz nieosiągalnych; usuwanie reguł redundantnych nie jest potrzebne, gdyż budowa populacji reguł chroni przed wystąpieniem powtarzających się reguł.

GrammarCorrector

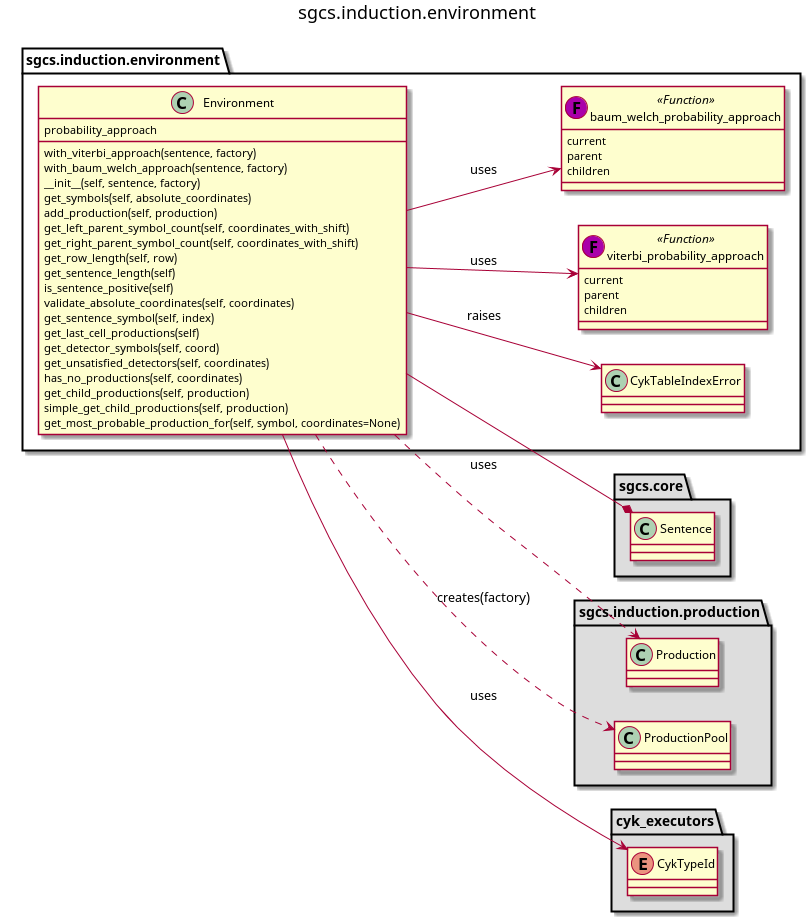
Klasa odpowiedzialna za korekcję gramatyki.

correct\_grammar(self, rule\_population, statistics) – metoda wywołująca remove\_non\_productive, a następnie remove\_not\_reachable.

remove\_non\_productive(self, rule\_population, statistics) – metoda usuwająca reguły, z których nie da się wyprowadzić reguł terminalnych.

remove\_not\_reachable(self, rule\_population, statistics) – metoda usuwająca reguły, których nie da się wyprowadzić z symbolu startowego gramatyki.

sgcs.induction.environment

Moduł ten zawiera klasy ściśle powiązane z obsługą tabeli CYK.

Environment

Klasa reprezntująca tabelę CYK, zdolna do obsługi zarówno danych stochastycznych, jak i klasycznych.

probability\_approach – delegat sposobu liczenia prawdopodobieństw kolejnych komórek. Dostępni obecnie delegaci to viterbi\_probablity\_approach (liczenie prawdopodobieństw metodą Viterbiego) oraz baum\_welch\_probability\_approach (liczenie prawdopodobieństw metodą Bauma-Welcha). Jeżeli klasa nie jest wykorzystywana przez algorytm stochastyczny, wówczas pole to przybiera wartość None.

with\_viterbi\_approach(sentence, factory) – funkcja tworząca instancję Environment z podejściem Viterbiego.

with\_baum\_welch\_approach(sentence, factory) – funkcja tworząca instancję Environment z podejściem Bauma-Welcha.

\_\_init\_\_(self, sentence, factory) – konstruktor pobierający zdanie, na podstawie którego zostanie wygenerowana tabela CYK oraz fabrykę factory, z której będzie możliwe uzyskanie instancji obiektów ProductionPool (reprezentujących pojedyncze komórki tabeli).

get\_symbols(self, absolute\_coordinates) – metoda zwracająca wszystkie efektory z komórki położonej na pozycji absolute\_coordinates.

add\_production(self, production) – metoda dodająca nową produkcję (wraz z ewentualnym przeliczeniem prawdopodobieństw).

get\_left\_parent\_symbol\_count(self, coordinates\_with\_shift) – metoda zwracająca ilość efektorów w komórce lewego rodzica.

get\_right\_parent\_symbol\_count(self, coordinates\_with\_shift) – metoda zwracająca ilość efektorów w komórce prawego rodzica.

get\_row\_length(self, row) – metoda zwracająca rzeczywistą długość wiersza tabeli CYK (tabela CYK jest tabelą trójkątną).

get\_sentence\_length(self) – metoda zwracająca długość zdania parsowanego w tabeli.

is\_sentence\_positive(self) – metoda zwracająca True, jeżeli zdanie należy do gramatyki.

get\_sentence\_symbol(self, i) – metoda zwracająca i-ty Symbol zdania.

get\_last\_cell\_productions(self) – metoda zwracająca produkcje ostatniej komórki tabeli CYK.

get\_detector\_symbols(self, coord) – metoda zwracająca symbole rodziców w postaci krotki kolekcji.

get\_unsatisfied\_detectors(self, coordinates) – metoda zwracająca detektory, dla których nie dopasowano żadnych reguł (przydatne przy generowaniu nowych).

has\_no\_productions(self, coordinates) – metoda zwraca True, jeżeli w danej komórce tabeli CYK nie zapisano żadnych produkcji.

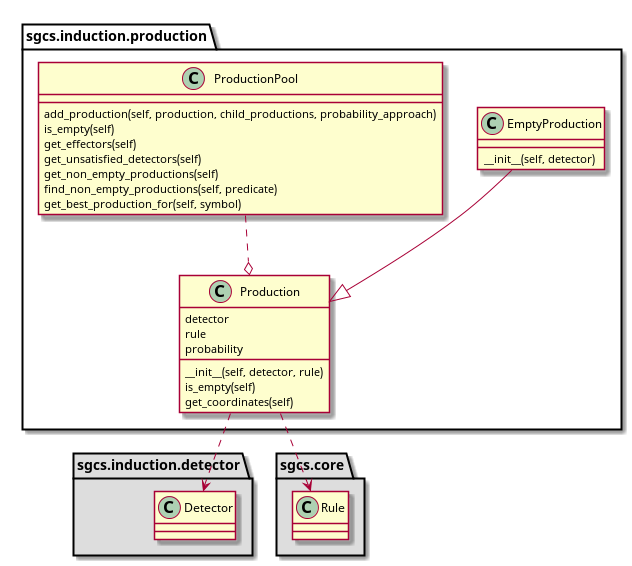
simple\_get\_child\_productions(self, production) – metoda zwracająca produkcje produkowane przez produkcję production (wykorzystywane między innymi przy przemierzaniu drzewa algorytmem Traceback). Jeżeli mamy do czynienia z algorytmem stochastycznym, prawdopodobieństwa są również zwracane.

get\_most\_probable\_production\_for(self, symbol, coordinates=None) – Zwraca z komórki CYK o koordynatach coordinates produkcję o rodzicu symbol taką, że jej prawdopodobieństwo jest największe. Jeżeli pominie się pole coordinates, wówczas metoda przyjmie koordynaty ostatniej komórki tabeli CYK.

CykTableIndexError

Wyjątek zgłaszany w przypadku wykroczenia indeksem poza granice tabeli CYK.

sgcs.induction.production

Moduł służący do zarządzania produkcjami w systemie. Przez produkcję należy tutaj rozumieć obiekt wiążący detektor z regułą populacji.

Production

Produkcja jest obiektem relacji klasy Detector oraz Rule. Oznacza to, że daje on wiedzę, skąd dana reguła została uzyskana oraz w jakim miejscu tabeli CYK się znajduje. Są to dane niezmiernie przydatne, ułatwiają między innymi wykorzystanie algorytmu traceback, który musi tylko poruszać się po zapamiętanej wcześniej sieci współrzędnych.

detector – pole przechowujące obiekt klasy Detector.

rule – pole przechowujące obiekt klasy Rule.

probability – obiekt przechowujący wartość prawdopodobieństwa danej produkcji (jest to prawdopodobieństwo całego drzewa generowanego przez daną produkcję, nie zaś prawdopodobieństwo pojedynczej produkcji). To pole jest opcjonalne – jeżeli klasa jest wykorzystywana poza algorytmem stochastycznym, wówczas jego wartość jest ignorowana.

is\_empty(self) – metoda zwracająca True, jeżeli produkcja jest „pusta”. Pusta, czyli nie generuje żadnej reguły (rule = None).

get\_coordinates(self) – zwraca współrzędne przechowywane przez detector.

EmptyProduction

Produkcja nieposiadającą reguły. Klasa ta w zasadzie nie rozszerza klasy Production, a stanowi jedynie ekspresywny alias na Production(detector, None).

ProductionPool

Pula produkcji, wykorzystywana między innymi do reprezentacji zawartości pojedynczej komórki w tabeli CYK.

add\_production(self, production, child\_productions, probability\_approach) – metoda dodająca produkcję do puli. Wykonywane są przy okazji różne czynności pomocnicze jak zapisywanie efektorów, przeliczanie prawdopodobieństw itp.

is\_empty(self) – zwraca True, jeżeli pula nie zawiera niepustych produkcji, False w przeciwnym przypadku.

get\_effectors(self) – zwraca efektory zarejestrowane w puli.

get\_unsatisfied\_detectors(self) – zwraca detektory pustych populacji (przydatne przy generowaniu nowych reguł).

get\_non\_empty\_productions(self) – zwraca wszystkie niepuste produkcje.

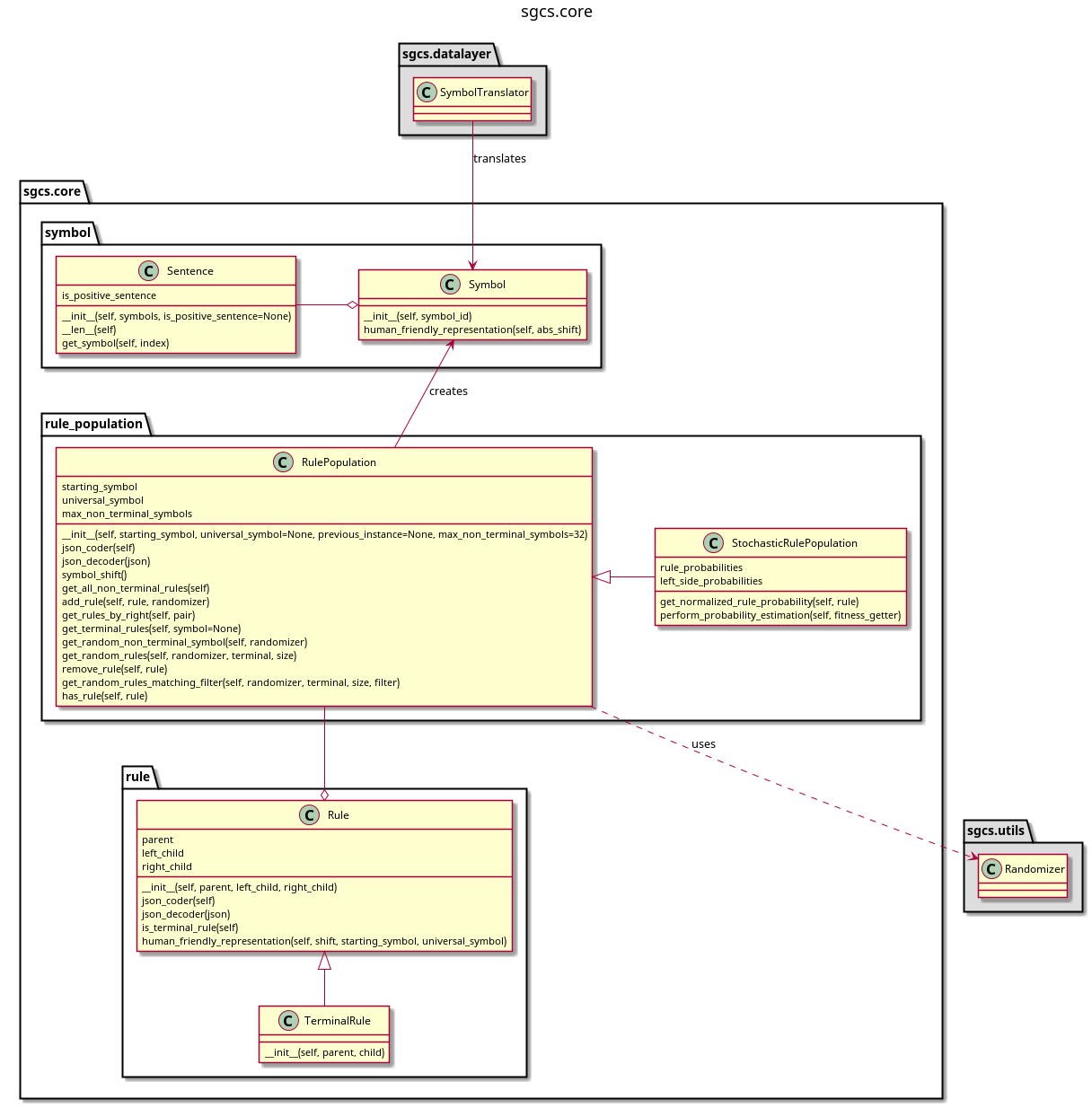
find\_non\_empty\_productions(self, predicate) – zwraca wszystkie niepuste produkcje spełniające predykat predicate.

get\_best\_production\_for(self, symbol) – zwraca najbardziej prawdopodobną produkcję dla efektora symbol.

Warstwa rdzenia

Warstwa ta jest sercem algorytmu. Klasy zawarte w jej jedynym module sgcs.core są wykorzystywane w wielu miejscach aplikacji, zazwyczaj reprezentują najbardziej podstawowe pojęcia związane z algorytmem wnioskowania gramatycznego. Są to klasy reprezentujące podstawowe dane pamięci algorytmu.

sgcs.core

RulePopulation

Klasa odpowiedzialna za zarządzanie populacją reguł.

starting\_symbol – reprezentacja symbolu startowego.

universal\_symbol – reprezentacja symbolu uniwersalnego.

max\_non\_terminal\_symbols – stała, określa maksymalną ilość symboli nieterminalnych.

json\_coder(self) – metoda dokonująca tłumaczenia populacji reguł na format .json.

json\_decoder(self, json) – metoda wczytująca stan wewnętrzny obiektu z obiektu w formacie .json.

symbol\_shift – wartość bezwzględna, od której zaczyna się numerowanie indeksów (każdy index taki, że |index| < symbol\_shift jest traktowany jako index zarezerwowany na symbol specjalny).

get\_all\_non\_terminal\_rules(self) – metoda zwracająca wszystkie reguły nieterminalne.

add\_rule(self, rule, randomizer) – dodaje nową regułę do populacji, dokonując losowania prawdopodobieństwa, jeżeli jest to wymagane.

get\_rules\_by\_right(self, pair) – zwraca wszystkie reguły, których prawa strona produkcji = pair.

get\_terminal\_rules(self, symbol=None) – zwraca wszystkie terminale reguły lub terminalne reguły produkujące symbol (jeżeli podany).

get\_random\_non\_terminal\_symbol(self, randomizer) – tworzy losowy symbol.

get\_random\_rules(self, randomizer, terminal, size) – zwraca maksymalnie size losowych reguł.

remove\_rule(self, rule) – usuwa wybraną regułę z populacji.

get\_random\_rules\_matching\_filter(self, randomizer, terminal, size, filter) – metoda zwracająca maksymalnie size losowych reguł, które spełniają warunek filter.

has\_rule(self, rule) – metoda zwraca True, jeżeli reguła rule znajduje się w populacji.

StochasticRulePopulation

Klasa rozszerzająca klasę RulePopulation o obsługę reguł stochastycznych.

rule\_probabilities – mapa reguła – prawdopodobieństwo.

left\_side\_probabilities – mapa ojciec reguły – prawdopodobieństwo symbolu.

get\_normalized\_rule\_probability(self, rule) – metoda zwracająca znormalizowaną wartość prawdopodobieństwa reguły.

perform\_probability\_estimation(self, fitness\_getter) – metoda powodująca wyliczenie nowych prawdopodobieństw dla reguł.

Rule

Klasa stanowiąca reprezentację reguły.

parent – rodzic Rule.

left\_child – lewe dziecko Rule.

right\_child – prawe dziecko Rule.

STARTING\_SYMBOL\_REPR – stała, reprezentacja symbolu startowego.

UNIVERSAL\_SYMBOL\_REPR – stała, reprezentacja symbolu uniwersalnego.

json\_coder(self) – metoda dokonująca tłumaczenia reguły na format .json.

json\_decoder(self, json) – metoda wczytująca stan wewnętrzny obiektu z obiektu w formacie .json.

is\_terminal\_rule(self) – zwraca True, jeżeli reguła jest terminalna.

human\_friendly\_representation(self, shift, starting\_symbol, universal\_symbol) – funkcja tłumacząca obiekt Rule na obiekt łatwy w odczytaniu przez człowieka.

TerminalRule

Cukier syntaktyczny na wyrażenie Rule(A, B, None).

Sentence

Obiekt stanowiący reprezentację zdania.

is\_positive\_sentence – czy zdanie jest tak naprawdę zdaniem pozytywnym.

\_\_init\_\_(self, symbols, is\_positive\_sentence=None) – konstruktor. Pobiera zestaw symboli, z jakich zostanie skonstruowane oraz informację czy podane zdanie w rzeczywistości należy do gramatyki, czy też nie.

\_\_len\_\_(self) – zwraca długość zdania.

get\_symbol(self, index) – Zwraca symbol na pozycji index.

Symbol

Klasa stanowiąca reprezentację pojedynczego symbolu gramatyki. Jeżeli ma się do czynienia z instancją Symbol reprezentującą symbol terminalny, to klasa ta nie posiada informacji wystarczającej do odtworzenia skrywanej wartości – potrzeba do tego dodatkowej wiedzy w postaci obiektu SymbolTranslator z modułu sgcs.datalayer.

human\_friendly\_representation(self, abs\_shift) – Wyświetla ludzką reprezentację symbolu (jak zostało to szczegółowo omówione przy oknie „Population Editor”). abs\_shift jest wartością przesunięcia symboli gramatyki, czyli powinna być tu użyta wartość RulePopulation.rule\_shift. Metody tej powinniśmy używać tylko wyłącznie w przypadku gdy dany Symbol jest symbolem nieterminalnym – nie należy stosować tej metody w przypadku symboli terminalnych oraz specjalnych. Najlepiej zostawić wówczas proces tłumaczenia samej klasie SymbolTranslator, która potrafi prawidłowo wykorzystywać tę metodę.

from\_human\_friendly\_representation(human\_repr, abs\_shift) – Metoda umożliwiająca odtworzenie reprezentacji wewnętrznej symbolu nieterminalnego z ludzkiej reprezentacji. Powstała ona z myślą o byciu wywoływaną przez klasę SymbolTranslator i wszelkie jej wywołania w innym kontekście powinny być wykonywane ze szczególną ostrożnością.

Warstwa danych

Warstwa ta skupia się na prawidłowej serializacji i deserializacji danych zawartych w module sgcs.core. Zawiera tylko jeden moduł – sgcs.datalayer.

sgcs.datalayer

Moduł odpowiedzialny za serializację, deserializację oraz zapisywanie i wczytywanie danych z lokalnego systemu plików.

SymbolTranslator

Klasa zapewniająca dwustronne tłumaczenie symboli z wewnętrznej na ludzką reprezentację i odwrotnie. Jest to klasa o stosunkowo wysokim poziomie abstrakcji, nie wymagając od użytkownika wiedzy na temat sposobu przechowywania danych.

create(path) – funkcja tworząca domyślnie skonfigurowaną instancję klasy SymbolTranslator z gorliwym ładowaniem pliku path w chwili pierwszej potrzeby tłumaczenia.

\_\_init\_\_(self, tokenizer, negative\_allowed=True) – konstruktor pobierający instancję tokenizera (obiekt zwracający dane wejściowe już w poddanej tokenizacji formie) oraz informację czy zdania negatywne również powinny być parsowane (w przypadku negative=False zdania negatywne zwracane przez tokanizera będą ignorowane).

get\_sentences(self) – zwraca generator zwracający kolejne zdania gramatyki. Zdania są pobierane leniwie z tokenizera (od klas położonych niżej zależy czy są wczytywane gorliwie, czy też leniwie).

symbol\_to\_word(self, symbol) – klasa tłumacząca symbol. Generuje reprezentację dla symboli nieterminalnych (z wykorzystaniem metod klasy sgcs.core.Symbol), zwraca reprezentację symboli specjalnych oraz odczytuje na podstawie zdobytej podczas parsowania wiedzy symbole terminalne. Aby metoda ta prawidłowo zadziałała, symbol terminalny musiał wystąpić już w jakimś zdaniu, aczkolwiek ponieważ klasa ta jest jedynym źródłem zdań, a zatem również symboli terminalnych, to każdy symbol musiał już wystąpić w jakimś zdaniu podczas działania algorytmu. Niekoniecznie jest to prawdą w przypadku edycji populacji poza działającym algorytmem i to jest powodem, dla którego należy przy jej dokonywaniu zachować szczególną ostrożność.

word\_to\_symbol(self, word) – metoda tłumacząca słowo terminalne na instancję klasy Symbol. Dotyczą się jej te same uwagi i obostrzenia co metody symbol\_to\_word.

rule\_population\_to\_string(self, rule\_population) – funkcja zwracająca napis stanowiący wygodną w czytaniu reprezentację populacji reguł. Klasa ta wykorzystuje metodę symbol\_to\_word, co może być potencjalnym źródłem problemów napotykanych przy wykorzystaniu tej metody.

EagerTokenizer

Klasa, której zadaniem jest podział linii zwróconych przez obiekt data\_fetcher na łatwe w parsowaniu tokeny. Nazwa może być myląca, EagerTokenizer dokonuje leniwej tokenizacji.

\_\_init\_\_(self, data\_fetcher) – konstruktor przyjmujący obiekt data\_fetcher odpowiedzialny za dostarczanie zdań w postaci linii tekstu.

get\_chunk\_generator(self) – metoda tworząca generator zwracający kolejne linie w postaci listy tokenów

EagerFileFetcher

Klasa odpowiedzialna za gorliwe wczytywanie pliku z danymi wejściowymi. Zamiast tej klasy można zdecydować się na LazyFileFetcher (wczytujący plik leniwie).

\_\_init\_\_(self, path) – konstruktor zapamiętujący ścieżkę path z danymi wejściowymi.

get\_chunk\_generator(self) – zwraca generator zwracający kolejne zdania (w sposób leniwy lub chowając pod sobą listę).

BasicJsonizer

Klasa tłumacząca obiekty dziedziczące po klasie SimpleJsonNode na format .json i odwrotnie. Tłumaczona klasa oraz jej składowe musi być typu podstawowego albo równocześnie dziedziczyć po klasie SimpleJsonNode i być na liście expected\_classes. Klasę tę wykorzystuje się głównie do tłumaczenia węzłów konfiguracji.

node\_id – stała będąca kluczem, pod którym zapisuje się w json nazwy klas, dzięki czemu jest możliwe ich późniejsze odtworzenie.

\_\_init\_\_(self, expected\_classes) – konstruktor pobierający listę obiektów klas, które mają być obsługiwane przez instancję (wszystkie powinny dziedziczyć po klasie SimpleJsonNode).

to\_json(self, obj) – metoda tłumacząca obj na napis w formacie json.

from\_json(self, json) – metoda tłumacząca napis w formacie json na obiekt.

RulePopulationJsonizer

Klasa odpowiedzialna za serializację obiektów typu RulePopulation.

make\_binding\_map(binding\_list) – funkcja przyjmująca listę klas, z których następnie tworzy mapę nazwa klasy – klasa.

\_\_init\_\_(self, bindings) – konstruktor przyjmujący mapę nazwa-klasa obsługiwanych klas (na przykład RulePopulation oraz klasy dziedziczące).

to\_json(self, population) – klasa zwracająca napis w formacie json będący tłumaczeniem obiektu population.

from\_json(self, json, randomizer, \*args, \*\*kwargs) – metoda zwracająca obiekt obiekt populacji tworzony na podstawie napisu w formacie json. Ponieważ populacje potrzebują zazwyczaj dodatkowych parametrów, są one również przekazywane.

Parametry algorytmu

Algorytm przetestowano dla trzech wariantów algorytmu klasyfikacji gramatycznej – GCS, sGCS oraz neg-sGCS. Poniżej przedstawiono omówienie parametrów, z jakimi uruchamiano algorytm.

– czy należy uruchomić algorytm genetyczny (tak/nie)? Podano: tak.

– lista typów stosowanej selekcji (gdzie typy to losowa, ruletka, turniej). Podano: {ruletka, ruletka}

– zmienna zezwalająca na uruchomienie korekcji gramatyki (tak, nie). Podano: nie.

– prawdopodobieństwo zastosowania operatora pokrycia startowego ([0,1]). Podano: 1.

– prawdopodobieństwo zastosowania operatora pokrycia pełnego ([0,1]). Podano: 1.

– prawdopodobieństwo zastosowania operatora pokrycia uniwersalnego ([0,1]). Podano: 0.

– maksymalna ilość kroków ewolucyjnych ([1, …]). Podano: 5 000.

– liczba iteracji ([1, …]). Podano: 50.

– rozmiar populacji (([1, …]). Podano: 40.

– liczba startowych produkcji nieterminalnych ([0, …]). Podano: 30.

– liczba symboli nieterminalnych ([1, …]). Podano: 19.

– prawdopodobieństwo krzyżowania dla algorytmu genetycznego ([0,1]). Podano: 0.2.

– prawdopodobieństwo mutacji dla lagorytmu genetycznego ([0,1]). Podano: 0.8.

– prawdopodobieństwo inwersji ([0,1]). Podano 0.

– prawdopodobieństwo zastosowania operatora pokrycia agresywnego ([0,1]). Podano: 0.

– współczynnik ścisku ([0, ..]). Podano: 18.

– podpopulacja ścisku ([1, …]). Podano: 3.

–kwota bazowa. Podano 0.5.

– współczynnik zmniejszania kwoty bazowej. Podano:0.5.

– rozmiar elity. ([0, …]). Podano: 0.

– waga rozbioru zdania poprawnego. Podano: 1.

– waga rozbioru zdania niepoprawnego. Podano: 2.

– waga funkcji klasycznej przystosowania. Podano: 1.

– waga funkcji płodności. Podano: 0.

– miara użyteczności klasyfikmatora niebiorącego udziału w parsowaniu. Podano: 0.5.

- zmienna informująca czy powinniśmy uczyć się też na podstawie zdań negatywnych. Ustawiona na false w przypadku sGCS, true w pozostałych algorytmach.

Algorytm sGCS oraz neg-sGCS korzystają dodatkowo z następujących parametrów:

– rodzaj algorytmu stosowanego do obliczania prawdopodobieństwa derywacji (Viterbi lub Baum-Welch). Podano: Viterbi.

Wyniki badań

Algorytm był uruchamiany na laptopie wyposażonym w 8 rdzeni logicznych oraz na zestawie serwerów postawionych w chmurze Amazon Web Services, o łącznej mocy obliczeniowej 100 rdzeni logicznych. Ze względu na wysoką przenaszalność biblioteki uruchomienie aplikacji w chmurze okazało się być stosunkowo prostym zadaniem, po uprzednim zrezygnowaniu z zapewnienia funkcjonującego gui PyQt, którego instalacja i konfiguracja okazały się niełatwym zadaniem na serwerach firmy Amazon. Algorytmy poddano trzem rodzajom testów. Po pierwsze powtórzono badania GCS oraz sGCS, przedstawione w pracy Kępy [10], co umożliwiło pokazanie poprawności implementacji i dało bazę do której można było porównać wyniki kolejnych badań. Następnie przetestowano działanie algorytmu neg-sGCS. Wreszczcie przetestowano wpływ reguł charakterystycznych na proces uczenia algorytmu. Dodatkowo rozpoczęto badania wpływu powtarzających się zdań w zbiorze uczącym na proces uczenia, aczkolwiek zabrakło czasu na przeprowadzenie sensownej ilości eksperymentów, toteż badania te nie zostaną szerzej omówione.

Standardowa losowa populacja

Na początku przeprowadzono badania z zastosowaniem populacji złożonej wyłącznie z losowych reguł. Nie określono z góry żadnego zestawu reguł charakterystycznych i nie dodano ich do programu. Poniżej przedstawiono wyniki tych badań – GCS, sGCS oraz neg-sGCS

GCS

Algorytm GCS uruchomiono ze standardowymi parametrami, wymienionymi w poprzednim rozdziale. Stosowano klasyczną funkcję przystosowania.

Tomita 1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 137,98 | 100,00% | 100,00% | 100,00% |
| 2 | 7,60 | 100,00% | 100,00% | 100,00% |
| 3 | 1,64 | 100,00% | 100,00% | 100,00% |
| 4 | 1,64 | 100,00% | 100,00% | 100,00% |
| 5 | 1,52 | 100,00% | 100,00% | 100,00% |
| Average: | 30,08 | 100,00% | 100,00% | 100,00% |

Tomita 2

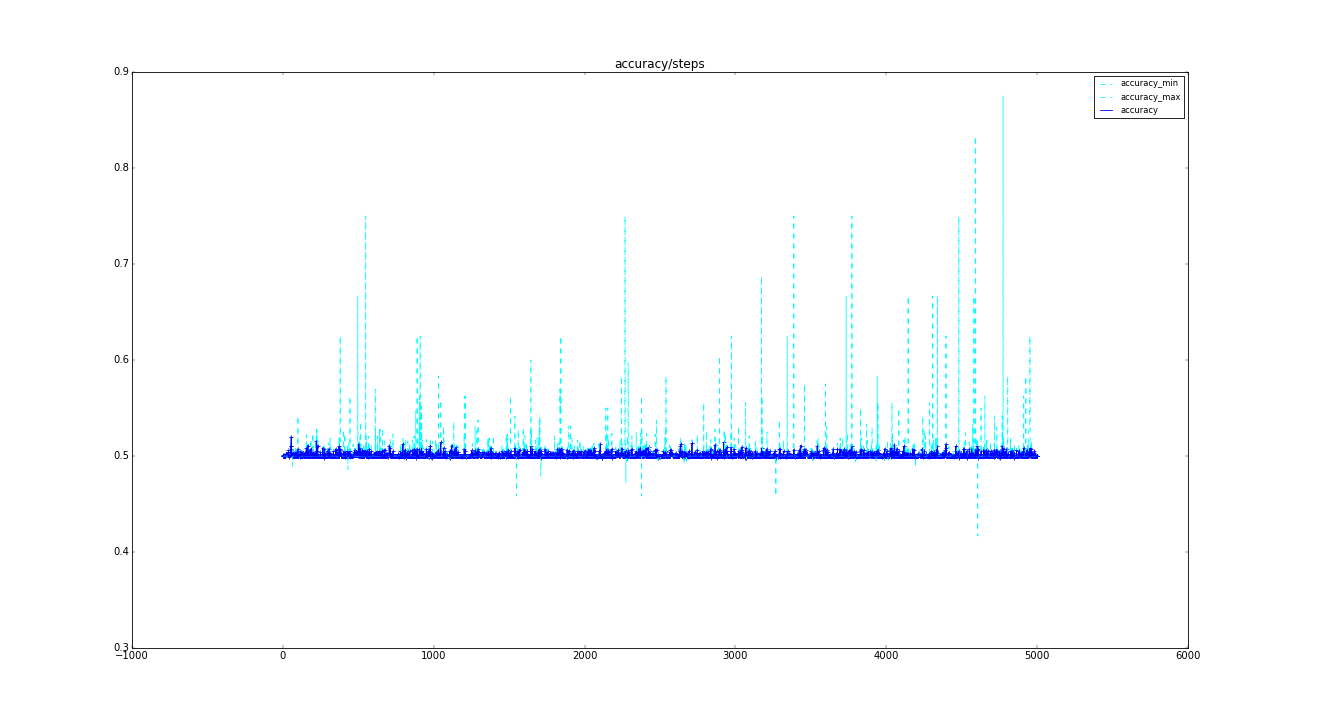
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 16,08 | 100,00% | 100,00% | 100,00% |
| 2 | 98,30 | 100,00% | 100,00% | 100,00% |
| 3 | 38,86 | 100,00% | 100,00% | 100,00% |
| 4 | 5,96 | 100,00% | 100,00% | 100,00% |
| 5 | 36,23 | 100,00% | 100,00% | 100,00% |
| Average: | 39,09 | 100,00% | 100,00% | 100,00% |

W przypadku gramatyk Tomita 1 oraz Tomita 2 nie ma wielkiego zaskoczenia. Każdy wyuczony zestaw klasyfikatorów uzyskał 100% przystosowania w teście generalizacji.

Tomita 3

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 937,43 | 63,32% | 93,91% | 89,50% |
| 2 | nan | 100,00% | 0,00% | 14,42% |
| 3 | 1637,43 | 83,28% | 77,62% | 78,44% |
| 4 | 2271,70 | 100,00% | 100,00% | 100,00% |
| 5 | 3353,38 | 62,80% | 93,20% | 88,82% |
| Average: | 2049,99 | 81,88% | 72,95% | 74,24% |

Gramatyka Tomita 3 ze względu na posiadanie błędnych przykładów w zbiorze uczącym [1] okazała się trudną gramatyką do wyuczenia. W przypadku drugiego uruchomienia żaden z 50 cykli uczących nie zdołał znaleźć rozwiązania. Jak widać na rysunku [..] algorytm w drugim przebiegu przez większość czasu osiągał przystosowanie rzędu 50%, czasami tylko osiągając krótkotrwałą poprawę. W pozostałych przypadkach otrzymano przystosowanie ok. 74% , co daje trochę słabszy rezultat niż w przypadku dotychczasowych prac [2] [10].



Tomita 4

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 1347,46 | 87,58% | 100,00% | 95,59% |
| 2 | 2086,47 | 5,34% | 83,84% | 56,00% |
| 3 | 1830,18 | 99,99% | 98,07% | 98,75% |
| 4 | 2240,00 | 91,26% | 67,54% | 75,95% |
| 5 | 1847,62 | 0,12% | 99,99% | 64,56% |
| Average: | 1870,35 | 56,86% | 89,89% | 78,17% |

Tomita 5

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 710,63 | 90,26% | 53,04% | 59,24% |
| 2 | 958,00 | 66,71% | 73,34% | 72,24% |
| 3 | 941,26 | 90,18% | 63,09% | 67,61% |
| 4 | 970,24 | 81,79% | 69,96% | 71,93% |
| 5 | 824,94 | 72,04% | 86,28% | 83,91% |
| Average: | 881,01 | 80,20% | 69,14% | 70,99% |

Tomita 6

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 2132,31 | 100,00% | 100,00% | 100,00% |
| 2 | 1960,82 | 86,25% | 32,69% | 50,54% |
| 3 | 1741,39 | 97,11% | 1,71% | 33,51% |
| 4 | 1924,84 | 42,82% | 90,19% | 74,40% |
| 5 | 1861,03 | 100,00% | 100,00% | 100,00% |
| Average: | 1924,08 | 85,24% | 64,92% | 71,69% |

Tomita 7

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 1848,32 | 60,91% | 80,10% | 79,36% |
| 2 | 1792,28 | 55,63% | 84,97% | 83,84% |
| 3 | 2124,00 | 46,52% | 97,89% | 95,92% |
| 4 | 2109,53 | 78,81% | 100,00% | 99,19% |
| 5 | 1506,62 | 48,79% | 93,75% | 92,03% |
| Average: | 1876,15 | 58,13% | 91,34% | 90,07% |

Pomiary dla Tomita 4-7 również wydają się być nieznacznie niższe niż te osiągnięte w publikacji Unolda [1]. Nie odbiegają one jednak zbytnio od wyników przedstawionych w pracy Kępy [10]. Może to być spowodowane potencjalnym błędem implementacji lub rozbieżnością któregoś z parametru. Ponieważ jednak wyniki pozostają z grubsza takie same (jedyna większa różnica to trochę gorsze rezultaty Tomita 4, które uzyskało przystosowanie rzędu 100% w pracy Kępy [10]), a sam algorytm w znacznej mierze opiera się na losowości i może generować różne wyniki, więc zdecydowano się na kontynuowanie badań i porównianie wyników pozostałych badań z otrzymanymi powyżej.

ab

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 1118,50 | 100,00% | 100,00% | 100,00% |
| 2 | 1776,17 | 100,00% | 100,00% | 100,00% |
| 3 | 2034,00 | 100,00% | 100,00% | 100,00% |
| 4 | 295,67 | 100,00% | 100,00% | 100,00% |
| 5 | 676,33 | 100,00% | 100,00% | 100,00% |
| Average: | 1180,13 | 100,00% | 100,00% | 100,00% |

anbn

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 798,80 | 100,00% | 100,00% | 100,00% |
| 2 | 1181,26 | 100,00% | 100,00% | 100,00% |
| 3 | 1036,11 | 100,00% | 99,98% | 99,98% |
| 4 | 832,00 | 100,00% | 99,97% | 99,97% |
| 5 | 1108,95 | 100,00% | 100,00% | 100,00% |
| Average: | 991,42 | 100,00% | 99,99% | 99,99% |

bra1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 267,86 | 0,00% | 99,04% | 98,09% |
| 2 | 479,07 | 0,00% | 99,04% | 98,09% |
| 3 | 813,74 | 0,00% | 99,04% | 98,09% |
| 4 | 813,74 | 0,00% | 99,04% | 98,09% |
| 5 | 276,65 | 0,00% | 99,04% | 98,09% |
| Average: | 530,21 | 0,00% | 99,04% | 98,09% |

bra3

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 213,88 | 0,68% | 99,62% | 65,99% |
| 2 | 495,58 | 0,53% | 99,55% | 65,90% |
| 3 | 539,28 | 12,32% | 98,97% | 69,52% |
| 4 | 248,03 | 0,60% | 99,64% | 65,96% |
| 5 | 252,22 | 0,84% | 98,88% | 65,56% |
| Average: | 349,80 | 2,99% | 99,33% | 66,59% |

pal2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 348,67 | 100,00% | 92,46% | 92,49% |
| 2 | 2180,25 | 100,00% | 100,00% | 100,00% |
| 3 | 3506,80 | 100,00% | 100,00% | 100,00% |
| 4 | nan | 33,86% | 96,18% | 95,94% |
| 5 | 550,00 | 100,00% | 98,88% | 98,88% |
| Average: | 1646,43 | 86,77% | 97,50% | 97,46% |

toy

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 845,33 | 0,00% | 99,59% | 99,43% |
| 2 | 1500,00 | 0,08% | 99,76% | 99,76% |
| 3 | 2087,83 | 0,00% | 99,83% | 99,67% |
| 4 | 1138,33 | 0,00% | 99,77% | 99,61% |
| 5 | 594,40 | 0,08% | 99,76% | 99,76% |
| Average: | 1233,18 | 0,03% | 99,74% | 99,65% |

W przypadku gramatyk bezkontekstowych również uzyskaliśmy wysokie wyniki. ab, anbn, pal2 osiągnęły przystosowanie bliskie 100%. Wysoki wynik cechuje również gramatyki toy i bra1, aczkolwiek widać, że w przypadku żadnej z tych gramatyk algorytm nie poradził sobie dobrze ze zdaniami pozytywnymi. O ile nie dziwi to zbytnio w przypadku to, to jest trochę niepokojąca w przypadku bra3. Z drugiej strony jest to gramatyka o dużej ilości symboli terminalnych,m wymagająca całkiem pokaźnego zestawu reguł do bycia przedstawioną. Poniżej znajduje się przykładowy zestaw reguł opisujący gramatykę bra3 (za Unold [1]) oraz przykładowy zestaw wygenerowany przez jeden z przebiegów:

A

O => 'b' C => 'e'

R => 'c' N => 'a'

B => 'f' R => 'd'

T => C; S M => N; <S>

<S> => R; B <S> => R; M

<S> => <S>; C <S> => M; R

<S> => O; C <S> => N; R

<S> => <S>; <S> <S> => O; <S>

D => R; <S> B => C; D

<S> => D; B <S> => N; D

E => D; F I => <S>; E

G => <S>; E C => I; O

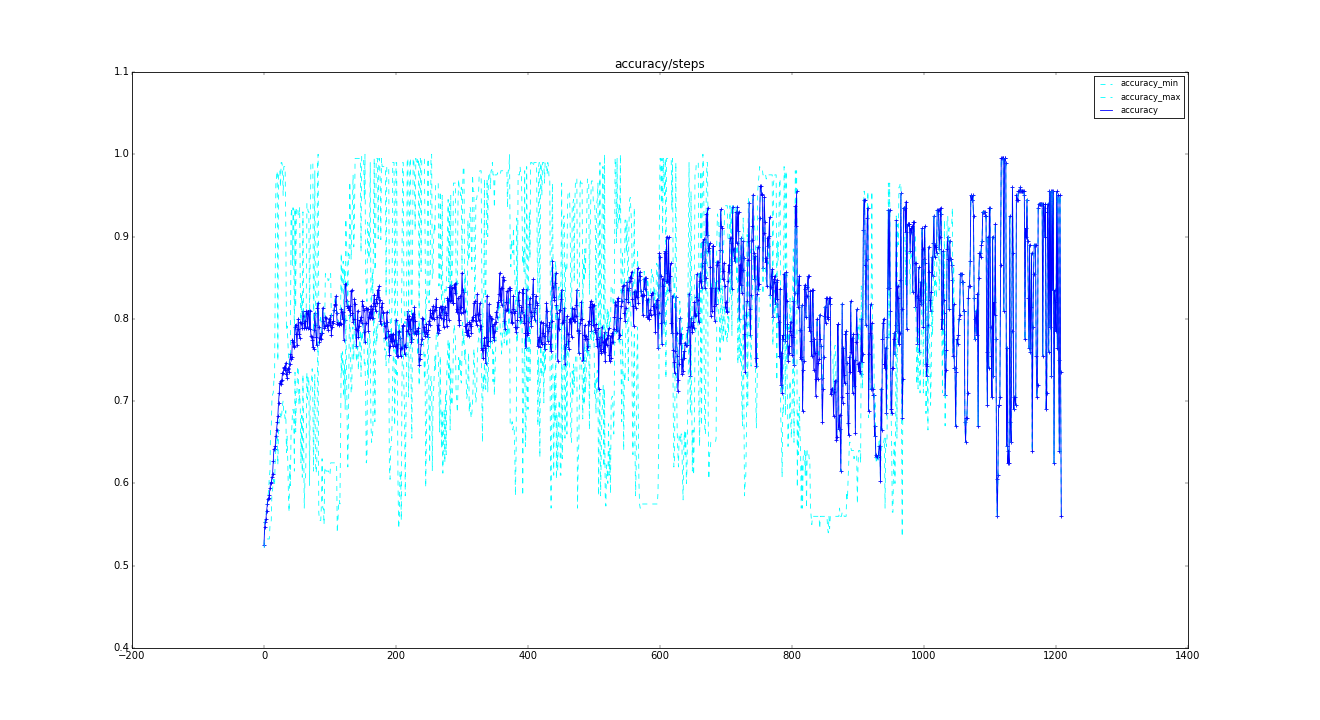
B => O; I M => C; I

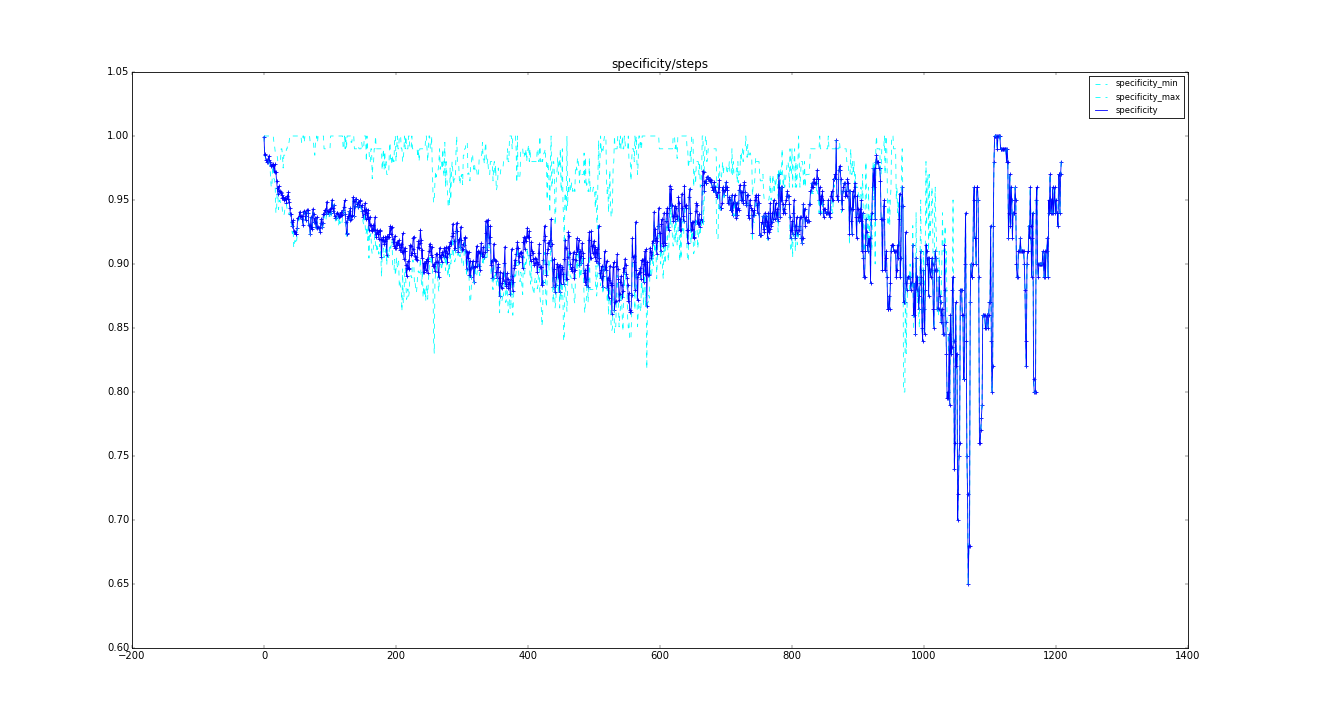
Q => O; B G => O; E

<S> => F; M <S> => D; <S>

U => M; T U => Q; U

Rysunek [..] przedstawia przystosowanie dla bra3 w kolejnych krokach algorytmu, rysunek […] pokazuje miejsce w którym tkwi problem uczenia się gramatyki bra3 – algorytm osiąga już na samym początku bardzo wysokie wyuczenie zdań negatywnych, ale nie klasyfikuje prawidłowo wszystkich zdań pozytywnych.   
Z czasem radzi sobie z coraz większą ilością zdań pozytywnych, kosztem niestety rozpoznawania zdań negatywnych. To oraz uzyskany specyficzny zestaw reguł pokazują, że zazwyczaj zostaje wybrana populacja reguł wysoce wyspecjalizowanych, co prowadzi do uzyskania niskiego wyniku testu generalizacji dla zdań pozytywnych.





sGCS

Algorytm sGCS uruchomiono ze standardowym zestawem parametrów, czyli w szczególności z zastosowanie algorytmu Viterbiego do obliczania prawdopodobieństwa drzew derywacji. Zastosowano funkcję fitness przedstawioną w pracy Pasieki [2], czyli parametr przystosowania był oparty jedynie na ilości zastosowań w parsowaniu zdań (pozytywnych, jako że oczywiście tylko takie wzięły udział w procesie uczenia).

Tomita 1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 1,00 | 100,00% | 100,00% | 100,00% |
| 2 | 1,00 | 100,00% | 100,00% | 100,00% |
| 3 | 1,00 | 100,00% | 100,00% | 100,00% |
| 4 | 1,00 | 100,00% | 100,00% | 100,00% |
| 5 | 1,00 | 100,00% | 100,00% | 100,00% |
| Average: | 1,00 | 100,00% | 100,00% | 100,00% |

Tomita 2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 51,98 | 100,00% | 93,96% | 93,96% |
| 2 | 45,32 | 100,00% | 99,78% | 99,78% |
| 3 | 84,38 | 100,00% | 98,30% | 98,30% |
| 4 | 71,24 | 100,00% | 99,99% | 99,99% |
| 5 | 89,80 | 100,00% | 99,78% | 99,78% |
| Average: | 68,54 | 100,00% | 98,36% | 98,36% |

Tomita 3

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 1,00 | 100,00% | 0,00% | 14,42% |
| 2 | 1,00 | 100,00% | 0,00% | 14,42% |
| 3 | 1,00 | 100,00% | 0,00% | 14,42% |
| 4 | 1,00 | 100,00% | 0,00% | 14,42% |
| 5 | 1,00 | 100,00% | 0,00% | 14,42% |
| Average: | 1,00 | 100,00% | 0,00% | 14,42% |

Tomita 4

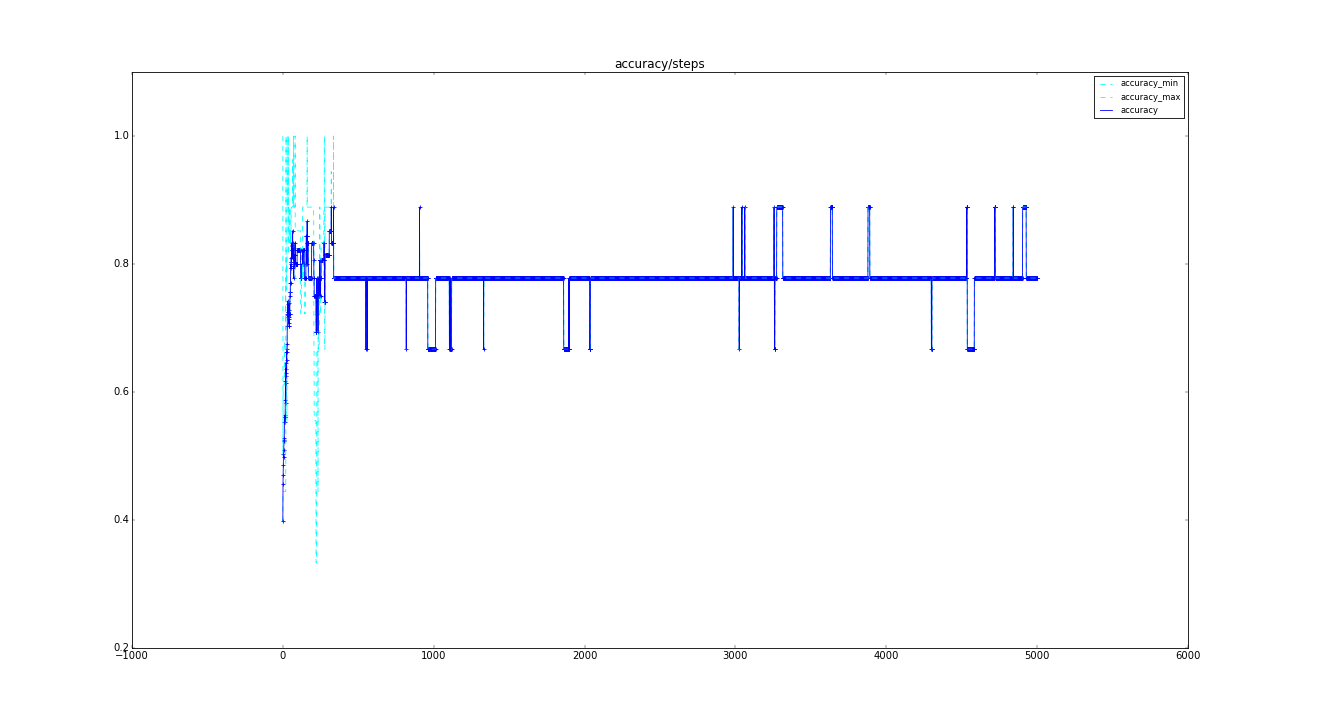
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 2,18 | 100,00% | 0,00% | 35,47% |
| 2 | 2,48 | 100,00% | 0,00% | 35,47% |
| 3 | 2,02 | 100,00% | 0,00% | 35,47% |
| 4 | 1,60 | 100,00% | 0,00% | 35,47% |
| 5 | 1,60 | 100,00% | 0,00% | 35,47% |
| Average: | 1,98 | 100,00% | 0,00% | 35,47% |

W przypadku gramatyk Tomita 1-4 nie ma żadnych zaskoczeń. Wyniki są praktycznie identyczne z pracą Kępy [10]

Tomita 5

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 267,96 | 100,00% | 80,00% | 83,33% |
| 2 | 75,86 | 100,00% | 80,00% | 83,33% |
| 3 | 99,10 | 100,00% | 80,00% | 83,33% |
| 4 | 51,43 | 100,00% | 80,00% | 83,33% |
| 5 | 85,30 | 100,00% | 80,00% | 83,33% |
| Average: | 115,93 | 100,00% | 80,00% | 83,33% |

Gramatyka Tomita 5 uzyskała interesujący rezultat, generując za każdym razem praktycznie identyczną gramatykę o przystosowaniu 83,33%. W pracy Kępy [10] widać o wiele większy rozrzut tych wartości, osiągają one również średnio gorsze rezultaty (NGen na poziomie 67,13%, NegGen oscyluje pomiędzy 3,36% a 88,71%, PosGen również nigdy nie osiąga 100%). Jak widać to na diagramie […], w zestawieniu z plikiem run\_summary.txt, w prawie każdym przypadku bardzo szybko znajduje rozwiązanie, tylko raz nie kończy poniżej 500 kroków i zaczyna błądzić w celu znalezienia rozwiązania



Tomita 6

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 143,74 | 100,00% | 0,00% | 33,34% |
| 2 | 179,30 | 100,00% | 0,00% | 33,34% |
| 3 | 195,49 | 100,00% | 0,00% | 33,34% |
| 4 | 229,34 | 100,00% | 0,00% | 33,34% |
| 5 | 314,53 | 100,00% | 0,00% | 33,34% |
| Average: | 212,48 | 100,00% | 0,00% | 33,34% |

Tomita 6 osiąga wyniki zbieżne z pracą Kępy [10].

Tomita 7

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 1,50 | 100,00% | 0,00% | 3,84% |
| 2 | 1,38 | 100,00% | 0,00% | 3,84% |
| 3 | 1,46 | 100,00% | 0,00% | 3,84% |
| 4 | 1,36 | 100,00% | 0,00% | 3,84% |
| 5 | 1,44 | 100,00% | 0,00% | 3,84% |
| Average: | 1,43 | 100,00% | 0,00% | 3,84% |

Da się zauważyć również odstępstwo od pracy Kępy [10] w przypadku gramatyki Tomita 7. Jest ono o tyle dziwne, że oba algorytmy uzyskawszy dokładnie takie same wyniki PosGen oraz NegGen powinny mieć również identyczne wyniki NGen. Tymczasem przy każdym przebiegu zaprezentowanym w pracy Kępy znajdujemy wynik 0,02%. Sprawdzono dokładnie zbiór uczący i wynik ów nie jest możliwy do uzyskania – zbiór testowy Tomita 7 to 65 535 zdań, z czego 2515 zdań pozytywnych. Uzyskanie wyuczenia zdań pozytywnych na poziomie 100% w teście generalizacji oznacza więc skuteczność na poziomie 3,84%, a nie 0,02% jak przedstawiono w pracy magisterskiejKępy. Zakładając zatem, że w pracy tej popełniono błąd, lub wykorzystywano zbiór Tomita 7 o innym zestawie zdań testowych, to prawdopodobnie uzyskano identyczny wynik (również 100% skuteczności w rozpoznaniu zdań pozytywnych i 0% w rozpoznaniu zdań negatywnych).

ab

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 65,70 | 100,00% | 0,00% | 7,18% |
| 2 | 33,00 | 100,00% | 71,83% | 73,85% |
| 3 | 17,67 | 100,00% | 72,03% | 74,04% |
| 4 | 17,90 | 100,00% | 0,03% | 7,21% |
| 5 | 36,60 | 100,00% | 0,05% | 7,23% |
| Average: | 34,17 | 100,00% | 28,79% | 33,90% |

W przypadku ab uzyskano wyniki o wiele lepsze niż w [Kępa] (średnio 8,87%), aczkolwiek gorsze niż w [Pasieka] (92%). Widzimy jednak po pojedynczych uruchomieniach, że algorytm zachowuje się w sposób daleki od stabilnego w przypadku indukcji tej gramatyki (bardzo udane przebiegi 2 i 3, fatalny rezultat w 1). Podobnie zachowuje się on w pracy [Kępa].

anbn

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 22,92 | 100,00% | 75,01% | 75,01% |
| 2 | 60,40 | 100,00% | 75,01% | 75,01% |
| 3 | 27,76 | 100,00% | 99,71% | 99,71% |
| 4 | 28,78 | 100,00% | 87,67% | 97,67% |
| 5 | 28,52 | 100,00% | 75,01% | 75,01% |
| Average: | 33,68 | 100,00% | 82,48% | 84,48% |

Dobre rezultaty otrzymujemy również w przypadku anbn. Jak widać to w tabeli […] zawsze mamy 100% skuteczności w rozpoznaniu zdań pozytywnych, co do negatywnych to oscyluje ona w granicach 75-100%. Jest to zatem wynik zbieżny z [Kępa] (około 99%) oraz [Pasieka] (95%).

bra1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 14,84 | 100,00% | 100,00% | 100,00% |
| 2 | 14,21 | 100,00% | 50,48% | 50,96% |
| 3 | 31,65 | 100,00% | 67,31% | 67,62% |
| 4 | 33,40 | 100,00% | 50,48% | 50,96% |
| 5 | 17,98 | 100,00% | 50,48% | 50,96% |
| Average: | 22,42 | 100,00% | 63,75% | 64,10% |

Duża rozbieżność skuteczności rozpoznania zdań negatywnych bra1, zaobserwowana w pracy Kępy [10], wystąpiła również przy okazji tych badań. Uzyskano wyniki podobne do przedstawionych przez poprzednie prace.

bra3

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 71,54 | 0,00% | 99,66% | 65,79% |
| 2 | 25,85 | 7,15% | 96,00% | 65,82% |
| 3 | 152,72 | 9,15% | 90,88% | 63,10% |
| 4 | 137,26 | 5,77% | 85,52% | 58,41% |
| 5 | 78,39 | 0,49% | 96,47% | 63,84% |
| Average: | 93,15 | 4,51% | 93,71% | 63,39% |

W przypadku bra3 da się zauważyć interesujący rezultat – algorytm uzyskuje jedynie kilka procent skuteczności w rozpoznawaniu zdań pozytywnych, a w zamian za to bardzo dużą skuteczność w rozpoznawaniu zdań negatywnych. Najwyraźniej miało tu miejsce zjawisko przeuczenia, co potwierdzają uzyskane reguły:

R => 'd' (0.45)

R => 'a' (0,46)

U => 'f' (0.82)

B => 'e' (0.5)

M => 'b' (1.0)

J => 'c' (1.0)

B => R; R (0.5)

<S> => M; B (0.23)

<S> => J; <S> (0.01)

<S> => <S>; B (0.03)

<S> => <S>; U (0.23)

<S> => M; <S> (0.03)

<S> => <S>; <S>(0.23)

<S> => J; U (0,24)

U => R; U (0.18)

C => M; B (1)

R => R; <S> (0.09)

pal2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 40,20 | 100,00% | 66,93% | 67,06% |
| 2 | 15,40 | 100,00% | 52,02% | 52,20% |
| 3 | 11,20 | 100,00% | 50,20% | 50,39% |
| 4 | 151,10 | 100,00% | 66,93% | 67,06% |
| 5 | 15,11 | 100,00% | 50,20% | 50,39% |
| Average: | 46,60 | 100,00% | 57,26% | 57,42% |

toy

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 449,60 | 26,67% | 93,58% | 93,48% |
| 2 | 227,10 | 100,00% | 0,09% | 0,25% |
| 3 | 286,44 | 0,00% | 97,19% | 97,04% |
| 4 | 269,90 | 0,95% | 84,64% | 84,50% |
| 5 | 198,90 | 0,00% | 99,87% | 99,71% |
| Average: | 286,39 | 25,52% | 75,07% | 75,00% |

W przypadku gramatyk pal2 oraz toy uzyskano podobne rezultaty jak w dotychczasowych pracach. Statystykę toy zaniża tutaj wyjątkowo niefortunny przebieg nr 2 (gdzie jak widać uzyskano jedynie 0,09% skuteczności w rozpoznaniu zdań negatywnych). Również wartości PosGen wskazują na przeuczenie się algorytmu.

Przedstawione do tej pory wyniki odbiegają nieznacznie od osiągniętych przez pozostałe prace. Różnice te pozostają jednak w akceptowalnej granicy, toteż będzie możliwe przetestowanie nowego wariantu algorytmu oraz wpływu reguł charakterystycznych na proces indukcji.

neg-sGCS

Jest to nowy mechanizm sGCS zaproponowany przez pracę. Opiera się on na algorytmie sGCS, któremu umożliwiono uczenie się również na przykładach negatywnych, wykorzystując część zagadnień z GCS jako składowe algorytmu będącego również w stanie wykorzystać przykłady negatywne, jak chociażbyklasyczną funkcję przystosowania. Właśnie ją zastosowano do określania przystosowania poszczególnych reguł.

Tomita 1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 1,06 | 100,00% | 100,00% | 100,00% |
| 2 | 1,06 | 100,00% | 100,00% | 100,00% |
| 3 | 1,02 | 100,00% | 100,00% | 100,00% |
| 4 | 1,08 | 100,00% | 50,01% | 50,02% |
| 5 | 1,08 | 100,00% | 50,01% | 50,02% |
| Average: | 1,06 | 100,00% | 80,00% | 80,01% |

Tomita 2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 108,45 | 100,00% | 100,00% | 100,00% |
| 2 | 143,50 | 100,00% | 100,00% | 100,00% |
| 3 | 124,70 | 100,00% | 100,00% | 100,00% |
| 4 | 95,33 | 100,00% | 100,00% | 100,00% |
| 5 | 95,33 | 100,00% | 100,00% | 100,00% |
| Average: | 113,46 | 100,00% | 100,00% | 100,00% |

Nowa wariant nie wpłynął na proces uczenia gramatyki Tomita2. Jak w przypadku GCS i sGCS uzyskano 100% przystosowania.

Tomita 3

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | nan | 100,00% | 0,00% | 14,42% |
| 2 | 3998,67 | 100,00% | 82,95% | 85,41% |
| 3 | 665,50 | 82,07% | 45,63% | 50,88% |
| 4 | 3210,00 | 100,00% | 99,51% | 99,58% |
| 5 | 3210,00 | 100,00% | 99,51% | 99,58% |
| Average: | 2771,04 | 96,41% | 65,52% | 69,97% |

Podczas indukcji gramatyki Tomita 3 udało się uzyskać całkiem wysokie rezultaty. Przebiegi 2, 4, 5 uzyskały rezultaty bliskie 100% przystosowania. Są one niestety zaniżone przez wyniki przebiegów 3 i 1. Jest to spowodowane niepowodzeniem w pełnym nauczeniu się na zbiorze uczącym w którymkolwiek z 50 uruchomień uczących przebiegu 1 oraz wygenerowaniem bardzo wyspecjalizowanego zestawu reguł w przebiegu 3. Nadal jest to wynik znacznie lepszy niż przy zastosowaniu zwykłego sGCS.

Tomita 4

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 2430,33 | 85,86% | 100,00% | 81,48% |
| 2 | 1826,10 | 0,58% | 100,00% | 70,37% |
| 3 | 2189,60 | 86,47% | 90,12% | 97,02% |
| 4 | 2625,80 | 91,26% | 100,00% | 72,04% |
| 5 | 2625,80 | 91,26% | 100,00% | 66,87% |
| Average: | 2339,53 | 71,09% | 98,02% | 77,56% |

W przypadku Tomita 4 osiągnął on rezultaty zbliżone do tych osiąganych przez wariant GCS. Po wynikach widać, że algorytm preferował prawidłowe rozpoznanie zdań negatywnych, w przypadku prawie każdego przebiegu uzyskano 100% przystosowania do zdań negatywnych przy równoczesnym zachowaniu wysokiej skuteczności identyfikowania zdań pozytywnych.

Tomita 5

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 650,07 | 45,75% | 88,63% | 81,48% |
| 2 | 816,70 | 78,58% | 68,73% | 70,37% |
| 3 | 886,63 | 82,15% | 100,00% | 97,02% |
| 4 | 880,07 | 93,66% | 67,72% | 72,04% |
| 5 | 1199,29 | 100,00% | 60,25% | 66,87% |
| Average: | 886,55 | 80,03% | 77,07% | 77,56% |

W przypadku Tomita 5 rezultaty wszystkich trzech algorytmów są zbliżone do siebie, neg-sGCS plasuje się pośrodku ze swoją skutecznością.

Tomita 6

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 1355,76 | 100,00% | 100,00% | 100,00% |
| 2 | 1302,38 | 41,19% | 95,03% | 77,09% |
| 3 | 1822,06 | 66,83% | 100,00% | 88,94% |
| 4 | 2121,09 | 96,04% | 100,00% | 98,68% |
| 5 | 1971,26 | 18,66% | 91,12% | 66,97% |
| Average: | 1714,51 | 64,54% | 97,23% | 86,34% |

Dla Tomita 6 algorytm neg-sGCS uzyskał najlepszy rezultat ze wszystkich trzech gramatyk, pozostawiając wariant sGCS daleko w tyle. Nie odbiega on znacząco w wynikach od GCS (widzimy nawet podobny rozrzut w przypadku obu algorytmów, oba uzyskują 100% przystosowania w pierwszym przebiegu, oba posiadają również dosyć kiepskie rezultaty, aczkolwiek GCS częściej uzyskuje dobre wyniki dla zdań pozytywnych, zaś neg-sGCS dla zdań negatywnych).

Tomita 7

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 2833,33 | 1,75% | 100,00% | 96,23% |
| 2 | 2353,50 | 48,79% | 93,75% | 92,03% |
| 3 | 2480,67 | 48,79% | 93,76% | 92,03% |
| 4 | 2235,25 | 72,09% | 81,17% | 80,83% |
| 5 | 2410,50 | 46,56% | 95,23% | 93,36% |
| Average: | 2462,65 | 43,60% | 92,78% | 90,90% |

W indukcji Tomita 7 neg-sGCS zdecydowanie króluje nad wariantem sGCS, uzyskując około 90% skuteczności w miejsce 4% sGCS. Wyniki są porównywalne z rezultatami osiągniętymi przez GCS.

ab

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 1849,17 | 100,00% | 100,00% | 100,00% |
| 2 | 1641,86 | 100,00% | 100,00% | 100,00% |
| 3 | 1534,17 | 100,00% | 100,00% | 100,00% |
| 4 | 234,00 | 100,00% | 100,00% | 100,00% |
| 5 | 903,33 | 100,00% | 91,58% | 92,19% |
| Average: | 1232,51 | 100,00% | 98,32% | 98,44% |

Dla ab neg-sGCS generuje wyniki niemal tak dobre jak GCS. Zostały one zaburzone jedynie przez przebieg 5, którego skuteczność (92,19%) również trudno nazwać daleką od satysfakcjonującej.

anbn

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 1445,39 | 100,00% | 100,00% | 100,00% |
| 2 | 1204,33 | 100,00% | 100,00% | 100,00% |
| 3 | 1251,89 | 100,00% | 99,82% | 99,82% |
| 4 | 1119,24 | 100,00% | 99,84% | 99,84% |
| 5 | 1370,39 | 100,00% | 100,00% | 100,00% |
| Average: | 1278,25 | 100,00% | 99,93% | 99,93% |

Dla anbn wszystkie algorytmy osiągają bardzo dobry poziom generalizacji bliski 100%.

bra1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 422,80 | 0,00% | 99,04% | 98,09% |
| 2 | 488,24 | 0,00% | 99,04% | 98,09% |
| 3 | 484,35 | 0,00% | 96,12% | 95,21% |
| 4 | 199,93 | 0,00% | 99,04% | 98,09% |
| 5 | 261,57 | 0,00% | 99,04% | 98,09% |
| Average: | 371,38 | 0,00% | 98,46% | 97,51% |

bra1 stoi na znacznie wyższym poziomie niż miało to miejsce w przypadku sGCS w poprzednich pracach [2] [10]. Mimo to podobnie jak przy zastosowaniu GCS mamy tutaj do czynienia z praktycznie zerowym rozpoznaniem zdań pozytywnym, co oznacza przeuczenie algorytmu na etapie uczenia. Poniżej przedstawiono jeden z zestawów reguł uzyskiwany w wyniku działania algorytmu.

I => 'a' (0.5)

Q => 'b' (1)

<S> => <S>; <S> (0.5)

<S> => I; Q (0.5)

H => I; Q (1)

I => I; H (0.5)

bra3

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 585,13 | 22,70% | 98,85% | 72,97% |
| 2 | 570,40 | 3,61% | 99,06% | 66,62% |
| 3 | 501,77 | 2,02% | 98,96% | 66,01% |
| 4 | 219,90 | 4,91% | 98,52% | 66,70% |
| 5 | 238,50 | 0,62% | 99,83% | 66,11% |
| Average: | 423,14 | 6,77% | 99,04% | 67,68% |

Dla neg-sGCS z bra3 uzyskano wynik zbieżny z wynikami pozostałych algorytmów. Podobnie jak jego poprzednicy neg-sGCS nie jest w stanie w prawidłowy sposób wyuczyć się gramatyki, tworząc wysoce wyspecjalizowane reguły, co kończy się niską skutecznością w rozpoznawaniu zdań pozytywnych. Jest to jednak dosyć powszechna przypadłość, dotykająca również implementacji sGCS przedstawionej w pracy Pasieki [2].

pal2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 2869,17 | 100,00% | 100,00% | 100,00% |
| 2 | 1653,67 | 100,00% | 97,76% | 97,77% |
| 3 | 1767,00 | 100,00% | 98,52% | 98,53% |
| 4 | 1158,00 | 100,00% | 100,00% | 100,00% |
| 5 | 1552,00 | 100,00% | 100,00% | 100,00% |
| Average: | 1799,97 | 100,00% | 99,26% | 99,26% |

pal2 to gramatyka, w której neg-sGCS odniósł zdecydowany sukces. Trzy z pięciu przebiegów kończą się odnalezieniem gramatyki w 100% poprawnej, pozostałe dwa przebiegi generują zestawy reguł zapewniające skuteczność rzędu 98%. Są to wyniki lepsze od obu poprzednich algorytmów.

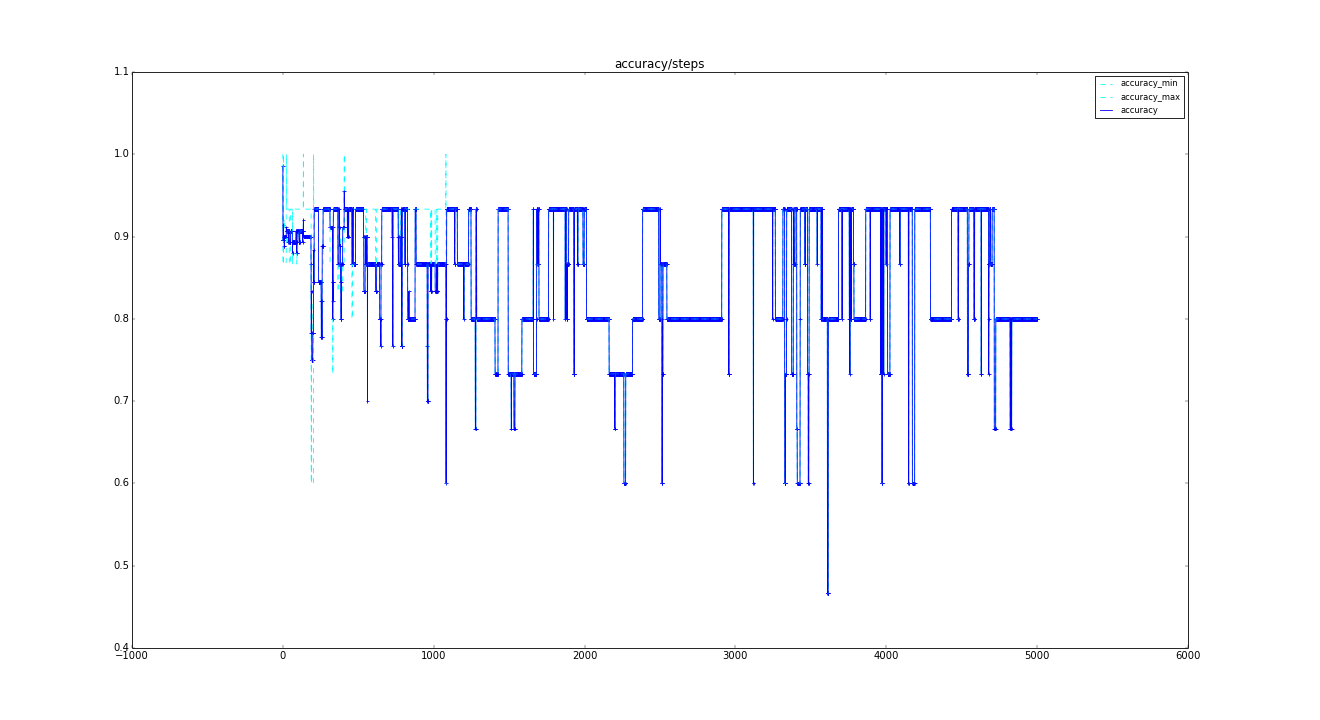
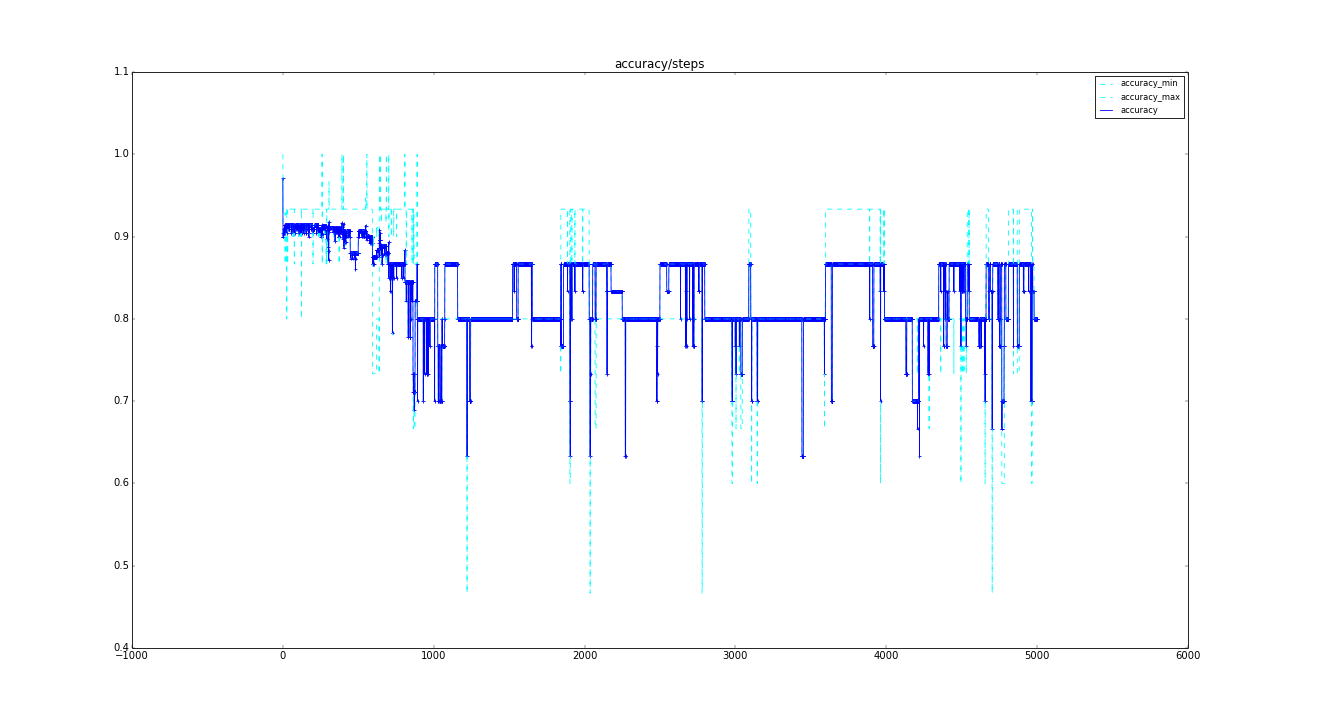
toy

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 1417,75 | 0,00% | 99,09% | 98,93% |
| 2 | 1433,57 | 0,00% | 99,87% | 99,71% |
| 3 | 1810,80 | 0,00% | 99,96% | 99,80% |
| 4 | 551,00 | 0,00% | 99,34% | 99,18% |
| 5 | 865,40 | 0,00% | 99,81% | 99,65% |
| Average: | 1215,70 | 0,00% | 99,61% | 99,45% |

Podobnie jak w przypadku poprzednich algorytmów neg-sGCS ma problem z poprawnym rozpoznaniem zdań pozytywnych. Pod pozornie wysokim wynikiem prawie 100% kryje się wysoce wyspecjalizowany zestaw klasyfikatorów, odpowiadający „Nie” w większości przypadków i wysoki rezultat jest uzależniony od budowy zbioru testowego (którego prewalencja to jedyne 0,16%). Mimo to są to najlepsze wyniki spośród wszystkich algorytmów.

Nowy algorytm stanowi ciekawe rozszerzenie algorytmu sGCS. Mimo, że pozornie zdania pozytywne powinny wystarczać w procesie uczenia to widać, że użycie również zdań negatywnych w znacznym stopniu poprawia wynik. Wyniki generalizacji neg-sGCS plasują się między wynikami sGCS a GCS w przypadku prostszych gramatyk, w przypadku tych bardziej skomplikowanych zdaża się czasami (w przypadku Tomita 6, Tomita 7, bra3 i pal2), że osiągnięty wynik jest lepszy niż poprzednich gramatyk. Niestety wykorzystanie zdań negatywnych znacznie wydłuża proces uczenia, sprawiając, że zazwyczaj zajmuje on co najmniej tyle czasu, ile w przypadku GCS. Nie da się jednak zaprzeczyć, że jeżeli interesuje nas wygenerowanie gramatyki stochastycznej to okazuje się on o wiele skuteczniejszym algorytmem niż sGCS.

Jak widać to po poniższych przebiegach wykresu Accuracy od kroku ewolucyjnego (przykład przedstawia przebieg dla algorytmu Tomita 2, aczkolwiek ten wzorzec występuje również w przypadku innych gramatyk) algorytm neg-sGCS ma większą niż GCS tendencję do „błądzenia” jeżeli nie zdoła znaleźć rozwiązania w pewnej ilości kroków. Należy w szczególności skupić się na części diagramu dotyczącej funkcjonowania obu algorytmów poniżej 1000 kroków – reszta przedstawia pojedyncze uruchomienia i jedyny interesujący fakt jaki można z niej wywnioskować, to że w przypadku neg-sGCS większa ilość uruchomień zdołała „zabłądzić” w okolice wyższych kroków. Tymczasem na przedziale [0;1000] widzimy, że neg-sGCS bardziej dąży średnią wartością do niższych wartości przystosowania niż jego poprzednik. Ta cecha algorytmu powoduje właśnie częstsze niepowodzenia w znalezieniu w stu procentach przystosowanego rozwiązania, z drugiej strony może być porządana w przypadku gramatyk o większej złożoności, gdzie takie szersze poszukiwanie rozwiązania jest porządane.



Reguły charakterystyczne

Tomita 1:

Tomita 2:

Tomita 3:

Tomita 4:

Tomita 5:

Tomita 6:

Tomita 7:

ab:

anbn:

bra1:

bra3:

pal2:

toy:

Powyżej przedstawiono zestaw reguł z jakimi uruchamiano wszystkie algorytmy w testach reguł charakterystycznych. Pozostałe brakujące reguły początkowe zostały wylosowane przed każdym uruchomieniem uczącym. Podczas wyboru zdecydowano się na kładzenie nacisku na rekurencyjne reguły z symbolem startowym– jako trudniejsze do wygenerowania przez algorytm, a także reguły opisujące interesujące nas fragmenty gramatyki. Dla żadnej gramatyki nie zdefiniowano więcej niż połowy potrzebnych reguł, ograniczając się do minimum zdolnego do szybkiego zdobycia niezerowej wartości fitness. Reguły nie były w żaden sposób zabezpieczone przed usunięciem, więc zabezpieczenie populacji przed ich usunięciem poprzez upewnienie się, że już od samego początku będą posiadać niezerową wartość fitness wymusiło zazwyczaj zdefiniowanie całej grupy powiązanych reguł.

Ograniczenia czasowe nie pozwoliły na ukończenie piątej serii eksperymentów, dlatego każdej gramatyce towarzyszą tu tylko 4 pomiary.

GCS

Tomita 1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 1,00 | 100,00% | 100,00% | 100,00% |
| 2 | 1,00 | 100,00% | 100,00% | 100,00% |
| 3 | 1,00 | 100,00% | 100,00% | 100,00% |
| 4 | 1,00 | 100,00% | 100,00% | 100,00% |
| Average: | 1,00 | 100,00% | 100,00% | 100,00% |

Tomita 2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 450,60 | 100,00% | 100,00% | 100,00% |
| 2 | 492,72 | 100,00% | 100,00% | 100,00% |
| 3 | 572,02 | 100,00% | 100,00% | 100,00% |
| 4 | 521,56 | 100,00% | 100,00% | 100,00% |
| Average: | 509,23 | 100,00% | 100,00% | 100,00% |

Gramatyki Tomita 1-2 są bardzo prostymi gramatykami, obecnie osiągającymi już100% przystosowania, więc dodanie reguł nie wpłynęło na ich proces uczenia (poza znacznym wydłużeniem uczenia Tomita 2 spowodowanym zmniejszeniem przestrzeni akceptowalnych rozwiązań).

Tomita 3

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 1818,75 | 76,10% | 79,58% | 79,07% |
| 2 | 1395,14 | 100,00% | 100,00% | 100,00% |
| 3 | 1933,11 | 100,00% | 100,00% | 100,00% |
| 4 | 1641,25 | 100,00% | 100,00% | 100,00% |
| Average: | 1697,06 | 94,03% | 94,90% | 94,77% |

Względem poprzednich przeprowadzonych testów widać znaczną poprawę Tomita 3. Dodanie części reguł odpowiedzialnych za generowanie parzystej liczby a po lewej stronie spowodowało również zwiększenie ilości uruchomień cykli uczących zakończonych sukcesem.

Tomita 4

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 588,87 | 91,26% | 100,00% | 96,90% |
| 2 | 813,31 | 91,26% | 100,00% | 96,90% |
| 3 | 588,87 | 91,26% | 100,00% | 96,90% |
| 4 | 1178,58 | 91,26% | 100,00% | 96,90% |
| Average: | 792,41 | 91,26% | 100,00% | 96,90% |

W przypadku Tomity 4 dodano reguły wykorzystujące w znacznym stopniu symbol startowy gramatyki. Spowodowało to skrócenie czasu poszukiwania rozwiązania, ustabilizowanie skuteczności wyników na wysokim poziomie (które miały do tej pory dosyć spory rozrzut 50-100%).

Tomita 5

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 1893,63 | 100,00% | 100,00% | 100,00% |
| 2 | 1429,97 | 100,00% | 100,00% | 100,00% |
| 3 | 1372,72 | 51,45% | 62,84% | 60,94% |
| 4 | 1745,68 | 85,85% | 65,22% | 68,66% |
| Average: | 1610,50 | 84,33% | 82,02% | 82,40% |

Dodanie reguł generujących ciąg a lub b o długości 2 spowodowały wygenerowanie stuprocentowo przystosowanej gramatyki w dwóch z czterech przypadków. Zmniejszyło jednak również przestrzeń rozwiązań, co w efekcie doprowadziło do wydłużenia procesu uczenia.

Tomita 6

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 579,74 | 100,00% | 100,00% | 100,00% |
| 2 | 588,15 | 66,83% | 63,74% | 64,77% |
| 3 | 955,44 | 100,00% | 100,00% | 100,00% |
| 4 | 992,82 | 66,83% | 100,00% | 88,94% |
| Average: | 779,04 | 83,42% | 90,94% | 88,43% |

W przypadku Tomity 6 zdecydowano się na dodanie reguł o wysokiej płodności. Poskutkowało to niewielkim wzrostem skuteczności oraz skróceniem czasu uczenia o połowę.

Tomita 7

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 532,03 | 100,00% | 100,00% | 100,00% |
| 2 | 632,07 | 100,00% | 100,00% | 100,00% |
| 3 | 392,13 | 100,00% | 100,00% | 100,00% |
| 4 | 607,62 | 100,00% | 100,00% | 100,00% |
| Average: | 540,96 | 100,00% | 100,00% | 100,00% |

Tomitę 7 również uruchomiono z regułami o wysokiej płodności. Dodanie ich skróciło trzykrotnie czas uczenia i spowodowało generowanie populacji klasyfikatorów zapewniających 100% przystosowania.

ab

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 1635,50 | 100,00% | 100,00% | 100,00% |
| 2 | 2266,80 | 100,00% | 100,00% | 100,00% |
| 3 | 1637,44 | 100,00% | 100,00% | 100,00% |
| 4 | 2114,80 | 100,00% | 100,00% | 100,00% |
| Average: | 1913,64 | 100,00% | 100,00% | 100,00% |

Dodanie reguł o wysokiej płodności spowodowało podwojenie czasu uczenia algorytmu dla ab, równocześnie zapewniając wygenerowanie gramatyki o 100% przystosowania.

anbn

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 1660,92 | 100,00% | 100,00% | 100,00% |
| 2 | 1607,36 | 100,00% | 99,84% | 99,84% |
| 3 | 1610,96 | 100,00% | 100,00% | 100,00% |
| 4 | 1513,92 | 100,00% | 100,00% | 100,00% |
| Average: | 1598,29 | 100,00% | 99,96% | 99,96% |

Ze względu na prostą budowę zdecydowano się na dodanie jedynie trzech reguł charakterystycznych do anbn. Wybór padł na reguły umożliwiające wyprowadzenie najprostszych zdań. Nie wpłynęło to w żadnym stopniu na proces uczenia, poza niewielkim wydłużeniem potrzebnego czasu ze względu na okrojenie przestrzeni akceptowalnych rozwiązań.

bra1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 2,42 | 0,00% | 99,04% | 98,09% |
| 2 | 2,42 | 0,00% | 99,04% | 98,09% |
| 3 | 2,00 | 0,00% | 99,04% | 98,09% |
| 4 | 2,00 | 0,00% | 99,04% | 98,09% |
| Average: | 2,21 | 0,00% | 99,04% | 98,09% |

bra3

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 363,33 | 1,02% | 99,66% | 66,13% |
| 2 | 337,06 | 2,12% | 99,00% | 66,07% |
| 3 | 249,55 | 1,98% | 99,23% | 66,18% |
| 4 | 238,98 | 0,73% | 99,42% | 65,87% |
| Average: | 297,23 | 1,46% | 99,33% | 66,06% |

Dodanie reguł wysoce płodnych reguł generujących ciągi liter ‘a’ nie wpłynęły na uczenie bra1 oraz bra3.

pal2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 3848,00 | 100,00% | 99,70% | 99,70% |
| 2 | 1622,00 | 100,00% | 98,94% | 98,94% |
| 3 | nan | 33,86% | 96,18% | 95,94% |
| 4 | 1258,00 | 100,00% | 100,00% | 100,00% |
| Average: | 2242,67 | 83,47% | 98,71% | 98,65% |

Dodanie reguł generujących ciągi a w sposób „palindromiczny” (tj. dodając ‘a’ po lewej stronie dodajemy je też po stronie prawej) spowodowało niewielki wzrost skuteczności algorytmu dla pal2.

toy

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 2258,33 | 0,00% | 99,89% | 99,73% |
| 2 | 1301,00 | 0,00% | 99,91% | 99,75% |
| 3 | 1948,00 | 0,00% | 99,98% | 99,82% |
| 4 | 2564,50 | 0,00% | 99,90% | 99,74% |
| Average: | 2017,96 | 0,00% | 99,92% | 99,76% |

W przypadku toy wykorzystano wysoko umieszczone w drzewach derywacji produkcje, których bezpośredni sąsiaedzi produkują symbole terminalne. Spowodowało to nieznaczną poprawę, nie rozwiązując problemu rozpoznawania zdań pozytywnych.

sGCS

Tomita 1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 1,00 | 100,00% | 100,00% | 100,00% |
| 2 | 1,00 | 100,00% | 100,00% | 100,00% |
| 3 | 1,00 | 100,00% | 100,00% | 100,00% |
| 4 | 1,00 | 100,00% | 100,00% | 100,00% |
| Average: | 1,00 | 100,00% | 100,00% | 100,00% |

Tomita 2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 2003,89 | 100,00% | 100,00% | 100,00% |
| 2 | 1943,88 | 85,71% | 99,95% | 99,95% |
| 3 | 1370,39 | 100,00% | 100,00% | 100,00% |
| 4 | 1940,65 | 100,00% | 100,00% | 100,00% |
| Average: | 1814,70 | 96,43% | 99,99% | 99,99% |

Tomita 3

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 1,00 | 100,00% | 0,00% | 14,42% |
| 2 | 1,00 | 100,00% | 0,00% | 14,42% |
| 3 | 1,00 | 100,00% | 0,00% | 14,42% |
| 4 | 1,00 | 100,00% | 0,00% | 14,42% |
| Average: | 1,00 | 100,00% | 0,00% | 14,42% |

Tomita 4

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 1,00 | 100,00% | 0,00% | 35,47% |
| 2 | 1,00 | 100,00% | 0,00% | 35,47% |
| 3 | 1,00 | 100,00% | 0,00% | 35,47% |
| 4 | 1,00 | 100,00% | 0,00% | 35,47% |
| Average: | 1,00 | 100,00% | 0,00% | 35,47% |

Tomita 5

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 62,16 | 100,00% | 80,00% | 83,33% |
| 2 | 117,16 | 100,00% | 80,00% | 83,33% |
| 3 | 75,70 | 99,99% | 0,02% | 16,68% |
| 4 | 68,56 | 100,00% | 60,00% | 66,67% |
| Average: | 80,90 | 100,00% | 55,01% | 62,50% |

Tomita 6

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 67,72 | 100,00% | 0,01% | 33,34% |
| 2 | 58,60 | 100,00% | 0,01% | 33,34% |
| 3 | 58,72 | 100,00% | 100,00% | 100,00% |
| 4 | 64,68 | 100,00% | 0,01% | 33,34% |
| Average: | 62,43 | 100,00% | 25,01% | 50,01% |

Tomita 7

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 1,00 | 100,00% | 0,00% | 3,84% |
| 2 | 1,00 | 100,00% | 0,00% | 3,84% |
| 3 | 1,00 | 100,00% | 0,00% | 3,84% |
| 4 | 1,00 | 100,00% | 0,00% | 3,84% |
| Average: | 1,00 | 100,00% | 0,00% | 3,84% |

Dodanie reguł charakterystycznych do populacji reguł stochastytcznych nie wpłynęło w żadnym stopniu na proces generowania gramatyk Tomita L1-L7. Reguły te były z reguły dosyć szybko usuwane.

ab

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 17,20 | 100,00% | 2,69% | 9,67% |
| 2 | 14,20 | 100,00% | 100,00% | 100,00% |
| 3 | 15,90 | 100,00% | 100,00% | 100,00% |
| 4 | 16,80 | 100,00% | 100,00% | 100,00% |
| Average: | 16,03 | 100,00% | 75,67% | 77,42% |

W przypadku gramatyki ab dodanie reguł charakterystycznych spowodowało znaczną poprawę rozpoznania zdań negatywnych.

anbn

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 30,90 | 100,00% | 99,99% | 99,99% |
| 2 | 31,30 | 100,00% | 75,01% | 75,01% |
| 3 | 42,80 | 100,00% | 98,79% | 98,79% |
| 4 | 31,46 | 100,00% | 99,71% | 99,71% |
| Average: | 34,12 | 100,00% | 93,38% | 93,38% |

bra1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 20,84 | 100,00% | 84,14% | 84,29% |
| 2 | 23,50 | 100,00% | 96,75% | 96,78% |
| 3 | 20,00 | 100,00% | 75,72% | 75,95% |
| 4 | 18,80 | 100,00% | 100,00% | 100,00% |
| Average: | 20,79 | 100,00% | 89,15% | 89,26% |

bra3

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 80,38 | 0,00% | 99,37% | 65,59% |
| 2 | 53,08 | 2,97% | 69,95% | 47,18% |
| 3 | 112,64 | 0,00% | 99,76% | 65,85% |
| 4 | 36,88 | 9,15% | 72,24% | 53,34% |
| Average: | 70,75 | 3,03% | 85,33% | 57,99% |

W przypadku gramatyk anbn, bra1 oraz bra3 nie nastąpiła zauważalna poprawa. bra1 zdaje się mieć trochę lepsze średnie wyniki, aczkolwiek nie odbiegają one zbytnio od tych osiągniętych przed dodanie reguł.

pal2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 14,70 | 100,00% | 93,18% | 93,21% |
| 2 | 50,20 | 100,00% | 0,23% | 0,61% |
| 3 | 26,60 | 100,00% | 2,41% | 2,78% |
| 4 | 52,60 | 100,00% | 52,02% | 52,20% |
| Average: | 36,03 | 100,00% | 36,96% | 37,20% |

Dosyć interesujące zachowanie, a mianowicie pogorszenie skuteczności algorytmu można zaobserwować w przypadku pal2. Okazało się, że algorytm nadał dosyć duże prawdopodobieństwa regułom charakterystycznym, i ostatecznie generował gramatyki z przeważającą ilością symboli ‘a’, będące zawsze nadzbiorem populacji palindromów nad alfabetem , co z kolei spowodowało niskie rezultaty NegGen, a co za tym idzie niski wynik NGen (ze względu na niską prewalencję 0,39% zbioru testowego). Poniżej porzedstawiono gramatykę wygenerowaną przeaz najgorsze, drugie uruchomienie:

A => 'a' (1)

C => 'b' (1)

<S> => A; A (0.17)

<S> => C; C (0.23)

O => A; <S> (0.62)

<S> => O; O (0.24)

<S> => <S>; C (0.15)

O => C; <S> (0.38)

<S> => O; A (0.14)

<S> => <S>; <S> (0.05)

<S> => C; <S> (0,02)

toy

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 51,67 | 53,33% | 98,14% | 98,07% |
| 2 | nan | 0,00% | 99,93% | 99,77% |
| 3 | 104,67 | 59,05% | 94,11% | 94,06% |
| 4 | 56,00 | 0,00% | 99,60% | 99,44% |
| Average: | 70,78 | 28,10% | 97,95% | 97,84% |

Dodanie reguł zapewniających krótką ścieżkę z symbolu startowego do terminalnych skróciło czas niezbędny do znalezienia rozwiązania oraz poprawiło nieco skuteczność algorytmu dla zdań pozytywnych.

Reguły charakterystyczne w niewielkim stopniu wpłynęły na proces uczenia sGCS. Było to głównie spowodowane dosyć szybkim wyrzucaniem takich reguł. Czasami odnosiły one efekt odwrotny od zamierzonego (na przykład w przypadku pal2 algorytm zbyt na nich polegał i nie dysponując zdaniami negatywnymi, które mogłyby wyprowadzić z błędu zazwyczaj wygenerował nadzbiór szukanej gramatyki). Na działanie tego algorytmu mogłoby mieć wpływ zablokowanie możliwości usuwania reguł charakterystycznych z populacji reguł. Niestety zabrakło czasu na przeprowadzenie badań pod tym kątem.

neg-sGCS

Tomita 1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 1,00 | 100,00% | 100,00% | 100,00% |
| 2 | 1,00 | 100,00% | 100,00% | 100,00% |
| 3 | 1,00 | 100,00% | 100,00% | 100,00% |
| 4 | 1,00 | 100,00% | 100,00% | 100,00% |
| Average: | 1,00 | 100,00% | 100,00% | 100,00% |

Tomita 2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 869,74 | 100,00% | 100,00% | 100,00% |
| 2 | 709,78 | 100,00% | 99,98% | 99,98% |
| 3 | 722,32 | 100,00% | 100,00% | 100,00% |
| 4 | 670,54 | 100,00% | 99,62% | 99,62% |
| Average: | 743,10 | 100,00% | 99,90% | 99,90% |

Tomita 3

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | nan | 100,00% | 0,00% | 14,42% |
| 2 | 1,00 | 100,00% | 100,00% | 100,00% |
| 3 | 6,00 | 100,00% | 100,00% | 100,00% |
| 4 | 1,00 | 100,00% | 100,00% | 100,00% |
| Average: | 2,67 | 100,00% | 75,00% | 78,61% |

Tomita 4

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 384,88 | 91,26% | 100,00% | 96,90% |
| 2 | 520,91 | 91,26% | 100,00% | 96,90% |
| 3 | 303,20 | 91,26% | 100,00% | 96,90% |
| 4 | 480,50 | 91,26% | 100,00% | 96,90% |
| Average: | 422,37 | 91,26% | 100,00% | 96,90% |

Jeżeli było to możliwe to dodanie reguł charakterystycznych dla doświadczeń z gramatykami Tomita 1-4 spowodowało poprawę algorytmu neg-sGCS, dzięki czemu prawie wszędzie udało mu się uzyskać 100% przystosowania. Jedyny słaby rezultat wystąpił przy pierwszym przebiegu dla gramatyki Tomita 3, aczkolwiek jest to spowodowane niezdolnością algorytmu do wyuczenia się gramatyki (nie nastąpiło to w żadnym z 50 cykli uczących).

Tomita 5

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 1285,27 | 66,53% | 100,00% | 94,42% |
| 2 | 1726,89 | 77,30% | 68,83% | 70,25% |
| 3 | 1827,58 | 68,90% | 70,87% | 70,55% |
| 4 | 1515,50 | 40,20% | 99,05% | 89,24% |
| Average: | 1588,81 | 63,23% | 84,69% | 81,12% |

Tomita 6

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 614,77 | 66,83% | 100,00% | 88,94% |
| 2 | 473,82 | 66,83% | 100,00% | 88,94% |
| 3 | 916,06 | 100,00% | 100,00% | 100,00% |
| 4 | 1056,03 | 66,83% | 100,00% | 88,94% |
| Average: | 765,17 | 75,12% | 100,00% | 91,71% |

Jak widać Tomita 5 i 6 nie osiągnęły praktycznie żadnej poprawy.

Tomita 7

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 3,75 | 100,00% | 100,00% | 100,00% |
| 2 | 1049,29 | 100,00% | 100,00% | 100,00% |
| 3 | 920,09 | 100,00% | 100,00% | 100,00% |
| 4 | 211,09 | 100,00% | 100,00% | 100,00% |
| Average: | 546,06 | 100,00% | 100,00% | 100,00% |

Podobnie jak w przypadku GCS zastosowanie reguł charakterystycznych powoduje zmniejszenie wartości nEvals oraz generowanie zestawów reguł zapewniających 100% przystosowania w teście generalizacji.

ab

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 810,57 | 100,00% | 100,00% | 100,00% |
| 2 | 1741,78 | 100,00% | 100,00% | 100,00% |
| 3 | 578,43 | 100,00% | 100,00% | 100,00% |
| 4 | 89,50 | 100,00% | 100,00% | 100,00% |
| Average: | 805,07 | 100,00% | 100,00% | 100,00% |

anbn

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 1536,55 | 100,00% | 100,00% | 100,00% |
| 2 | 1934,55 | 100,00% | 99,49% | 99,49% |
| 3 | 1730,27 | 100,00% | 100,00% | 100,00% |
| 4 | 1580,14 | 100,00% | 100,00% | 100,00% |
| Average: | 1695,38 | 100,00% | 99,87% | 99,87% |

bra1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 2,00 | 0,00% | 99,04% | 98,09% |
| 2 | 2,00 | 0,00% | 99,04% | 98,09% |
| 3 | 2,00 | 0,00% | 99,04% | 98,09% |
| 4 | 2,00 | 0,00% | 99,04% | 98,09% |
| Average: | 2,00 | 0,00% | 99,04% | 98,09% |

bra3

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 281,85 | 0,67% | 99,57% | 65,96% |
| 2 | 202,52 | 0,69% | 99,35% | 65,81% |
| 3 | 267,44 | 0,61% | 99,60% | 65,96% |
| 4 | 224,59 | 3,28% | 99,54% | 66,82% |
| Average: | 244,10 | 1,31% | 99,52% | 66,14% |

pal2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 301,67 | 100,00% | 100,00% | 100,00% |
| 2 | 1134,33 | 100,00% | 99,78% | 99,78% |
| 3 | 462,00 | 100,00% | 100,00% | 100,00% |
| 4 | 223,20 | 100,00% | 100,00% | 100,00% |
| Average: | 530,30 | 100,00% | 99,95% | 99,95% |

Gramatyki ab, anbn, bra1, bra3 oraz pal2 pozostają na względnie tym samym poziomie skuteczności. Zmniejszeniiu za to z pewnością ulega czas niezbędny do wyuczenia tych gramatyk, w szczególności w przypadku bra1 (spadek z nEvals=371 do 2). Powraca też znany niestety z symulacji GCS problem wyuczenia zdań pozytywnych gramatyk nawiasowych (bra1, bra3).

toy

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Run | nEvals | PosGen [%] | NegGen [%] | NGen [%] |
| 1 | 1932,75 | 0,00% | 99,88% | 99,72% |
| 2 | 1047,00 | 0,00% | 99,94% | 99,78% |
| 3 | 1327,00 | 0,00% | 99,94% | 99,78% |
| 4 | 1347,00 | 0,00% | 99,59% | 99,44% |
| Average: | 1413,44 | 0,00% | 99,84% | 99,68% |

W przypadku toy algorytm osiągnął najlepsze dotychczas przystosowanie NGen. Niestety, dzieje się to kosztem przystosowania PosGen, które tak jak w przypadku sGCS spadło do 0%.

Jak widać połączenie dodawania reguł charakterystycznych do początkowej populacji oraz stosowanie algorytmu neg-sGCS stanowi dobre połączenie i daje bardzo dobre rezultaty.

Podsumowanie

Algorytmy generujące klasyfikatory gramatyczne są interesujące i ułatwiają zrozumienie skomplikowanych zależności, jakimi są powiązane występujące w świecie rzeczywistym zjawiska. Kolejne prace rozwijają dalsze mechanizmy oraz badają wpływ poszczególnych elementów na funkcjonowanie całego mechanizmu. Praca ta pokazała potencjał stosowania zdań negatywnych w procesie uczenia oraz zaproponowała sposób obsługi takich zdań przy pomocy prostej implementacji neg-sGCS. Jest to najprawdopodobniej kierunek, w którym powinien pójść dalszy rozwój algorytmu. Niestety neg-sGCS nie radzi sobie idealnie z każdą gramatyką i zdarzają mu się zachowania odwrotne do tych spotykanych w sGCS – a konkretnie czasami generuje on gramatyki dobrze rozpoznające zdania negatywne, mające problemy z rozpoznawaniem zdań pozytywnych. Kluczowy może tu być odpowiedni dobór parametrów – być może należałoby zmniejszyć wagę rozbioru negatywnego zdania w przypadku neg-sGCS.

Inną możliwością jest równoczesne symulowanie kilku populacji i próba wymiany informacji pomiędzy nimi przy pomocy genetycznego algorytmu wyspowego [3]. Można by wówczas uruchomić obok siebie kilka algorytmów, na przykład sGCS i neg-sGCS i pozwalać im na wymianę informacji. Podejście to jednak wymagałoby ustalenia sposobu wymiany reguł, które będąc regułami stworzonymi w całkowicie innej populacji inną ścieżką ewolucji mogłyby się okazać nieprzenaszalne pomiędzy populacjami. Można by tu zastanowić się nad zastosowaniem podejścia Pittsburg lub hybrydowego [2] w algorytmie genetycznym lub jego części migracyjnej w celu przesyłania całych zbiorów reguł zamiast pojedynczych osobników.

Równoczesne generowanie kilku populacji może również sprawiać poważny problem czasowy, jako że na wygenerowanie i przetestowanie pojedynczej gramatyki trzeba było czekać czasami nawet kilkanaście godzin. Taka ewolucja równoległa daje jednak pole opisu do dalszego zrównoleglenia algorytmu, w wyniku czego poprzez dysponowanie odpowiednią architekturą obliczeniową (na przykład własną chmurą) można zredukować wpływ dodawania kolejnych populacji.

Interesujące rezultaty dało również wykorzystanie reguł charakterystycznych w procesie uczenia. Zastosowanie ich w przypadku GCS oraz neg-sGCS zazwyczaj poskutkowało poprawą działania tych algorytmów. Wykorzystanie ich w przypadku sGCS nie dało dużego zysku, co było czasami spowodowane przez przedwczesne usunięcie tych reguł w procesie uczenia, a czasami uniwersalność i płodność tych reguł zwiększyła szansę sGCS na stworzenie nadgramatyki zamiast poszukiwanej (jak miało to miejsce na przykład w przypadku badań nad gramatyką pal2). Tak czy inaczej generalnie dodawanie reguł charakterystycznych do populacji początkowej znacznie poprawia rezultaty działania algorytmów i jest to praktyka, którą warto stosować. Interesującym zagadnieniem mogłoby być zautomatyzowanie wyszukiwania reguł charakterystycznych, dzięki czemu taki zestaw reguł mógłby być w jakiś sposób wstępnie określany przez inny algorytm uruchamiany przez systemem sGCS.

Bibliografia

|  |  |
| --- | --- |
| [1] | O. Unold, Ewolucyjne wnioskowanie gramatyczne, Wrocław: Oficyna wydawnicza Politechniki Wrocławskiej, 2006. |
| [2] | G. Pasieka, Praca magisterska. Nowe mechanizmy adaptacji w modelu sGCS, Wrocław, 2014. |
| [3] | A. Reza and Z. Koorush, "A multilevel evolutionary algorithm for optimizing numerical functions," *International Journal of Industrial Engineering Computations,* vol. 2, 2011. |
| [4] | Y. Tsuruokazy and J. Tsujii, "Iterative CKY parsing for Probabilistic Context-Free Gramamrs," in *Natural Language Processing - IJCNLP 004*, First International Joint Conference, Hainan Island, China, Revised Selected Papers, 2004, pp. 52-60. |
| [5] | L. Pacholski i W. Charatonik, Logika dla informatyków. Materiały do zajęć, ósme poprawione red., Wrocław: Uniwersytet Wrocławski, 2010. |
| [6] | M. Skrzypczyk, Praca magisterska. LibGCS - biblioteka dla systemu xGCS, Wrocław, 2011. |
| [7] | P. Skórzewski, Praca magisterska. Gramatyki i automaty probabistyczne, Poznań, 2010. |
| [8] | M. E. Błędowski, Projekt inżynierski: Implementacja uczącego się systemu klasyfikującego GCS, Wrocław, 2012. |
| [9] | J. Arabas, Wykłady z algorytmów ewolucyjnych, Warszawa: WNT, 2009, pp. 89-107. |
| [10] | M. Kępa, Praca magisterska: Stochastyczny model GCS, Wrocław, 2008. |

[Rysunek 1:Przykładowa tabela CYK 18](#_Toc443730699)

Spis definicji

Alfabet 10

Bezpośrednia wyprowadzalność 11

Częściowy porządek 8

Długość słowa 10

Dobry porządek 10

Drzewo wyprowadzenia/derywacji/rozkładu 13

Gramatyka bezkontekstowa 12

Gramatyka formalna 10

Gramatyka GCS 13

Gramatyka kombinatoryczna 11

Gramatyka kontekstowa 12

Gramatyka lewostronnie liniowa 12

Gramatyka prawostronnie liniowa 12

Gramatyka regularna 12

Gramatyka sGCS 14

Język generowany przez gramatykę 11

Język nad dowolnym alfabetem 10

Język, język nad alfabetem 10

klasyfikator GCS 13

klasyfikator sGCS 14

Kres górny 9

Ograniczenie górne 9

Osiągalność 11

Pełna reprezentacja gramatyki 16

Porządek liniowy 9

Porządek regularny 10

Porządek zupełny 9

Postać Chomsky’ego, PNC, CNF 12

Pozytywna reprezentacja gramatyki 16

Produktywność 11

Przechodniość 8

Przystosowanie 15

Relacja 8

Równoważność CNF 12

Równoważność gramatyk 11

Słaba antysymetryczność 8

Słowo puste 10

Słowo, słowo nad alfabetem 10

Stochastyczna gramatyka bezkontekstowa 14

Symbol 10

Symetryczność 8

Wyprowadzalność 11

Zbiór częściowo uporządkowany 9

Zbiór skierowany 9

Zdanie 14

Znormalizowane prawdopodobieństw 14

Zwrotność 8