

*Trabalho de Pesquisa Operacional*

***Otimização de Árvores Sintáticas  
aplicada à Detecção de Células  
Cancerígenas***

*Juliana Oliveira Ferreira  
Bacharelado em Ciência da Computação  
UNIFAL-MG*

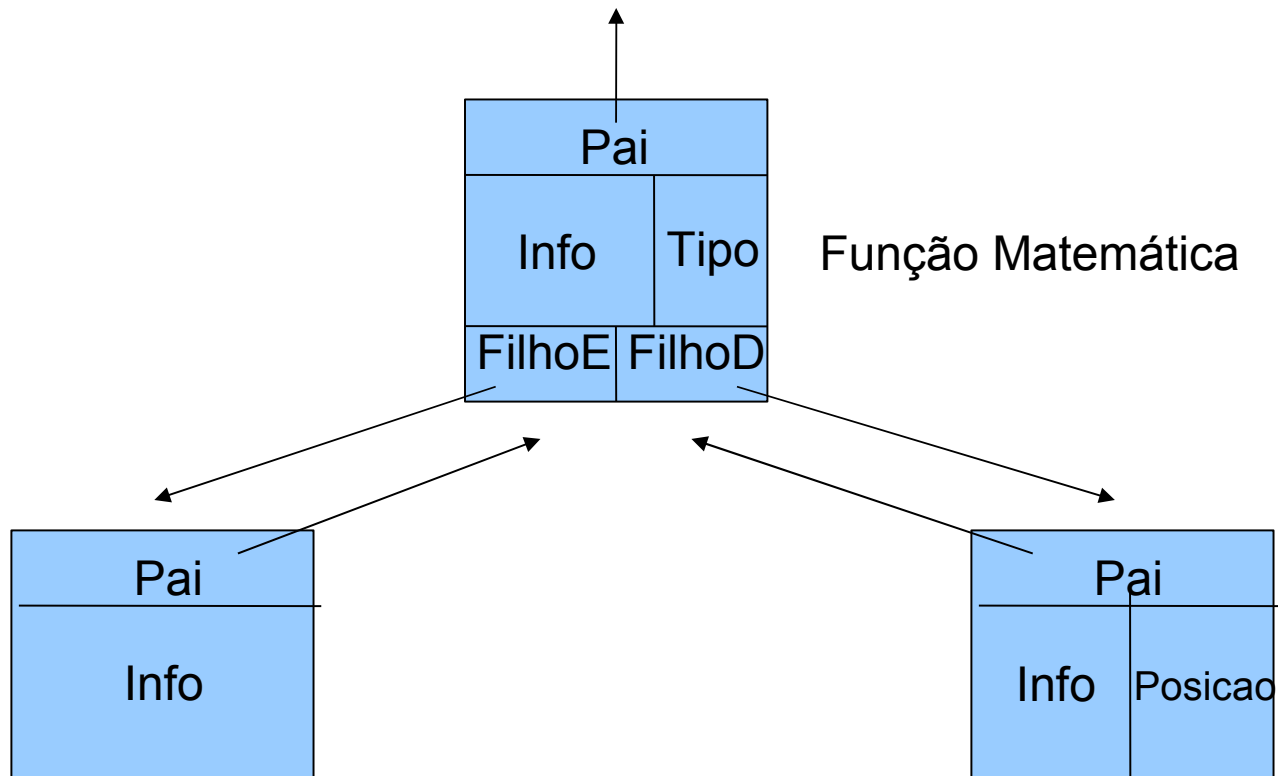
# Objetivo

- *Gerar uma técnica eficiente para o reconhecimento de células com câncer maligno, baseando-se em características básicas da célula.*
- *Poder identificar quais características das células são realmente necessárias para distinguir as com câncer maligno, das com câncer benigno.*

# *Representação*

- *Baseada em Árvores Sintáticas*
- *A árvore armazena expressões que contém :*
  - ✓ *Funções Matemáticas*
  - ✓ *Constantes*
  - ✓ *Variáveis (características do objeto a ser identificado)*
- *Toda árvore terá como raiz a função sigmóide logística.*

# *Representação*



# *Gramática*

*Esta pesquisa utiliza Gramática Livre de Contexto (Gramática da Árvore Sintática) para gerar a Linguagem reconhecida pelo sistema.*

$$\mathbf{G = \{V,T,R,P\}}$$

$$\mathbf{V = \{<ROOT>, <SIGM>, <EXP>, <BINARY>, <UNARY>, <LEAF>, <SUM>, <SUB>, <PROD>, <CONST>, <VAR>\}}$$

$$\mathbf{T = R \cup \{x_1, x_2, \dots, x_n\} \cup \{(,)\}}$$

$$\mathbf{P = <ROOT>}$$

# Gramática

*O conjunto de regras  $R$  é formado pelas fórmulas a seguir:*

$\langle \text{ROOT} \rangle \rightarrow \langle \text{SIGM} \rangle$

$\langle \text{SIGM} \rangle \rightarrow 1 / (1 + e^{(-1 * \langle \text{EXP} \rangle)})$

$\langle \text{EXP} \rangle \rightarrow (\langle \text{UNARY} \rangle) \mid (\langle \text{BINARY} \rangle) \mid (\langle \text{LEAF} \rangle)$

$\langle \text{BINARY} \rangle \rightarrow \langle \text{SUM} \rangle \mid \langle \text{SUB} \rangle \mid \langle \text{PROD} \rangle$

$\langle \text{UNARY} \rangle \rightarrow \langle \text{SIGM} \rangle$

$\langle \text{LEAF} \rangle \rightarrow \langle \text{CONST} \rangle \mid \langle \text{VAR} \rangle$

$\langle \text{SUM} \rangle \rightarrow \langle \text{EXP} \rangle + \langle \text{EXP} \rangle$

$\langle \text{SUB} \rangle \rightarrow \langle \text{EXP} \rangle - \langle \text{EXP} \rangle$

$\langle \text{PROD} \rangle \rightarrow \langle \text{EXP} \rangle * \langle \text{EXP} \rangle$

$\langle \text{CONST} \rangle \rightarrow y \mid y \in \mathbb{R}$

$\langle \text{VAR} \rangle \rightarrow s_1 \mid s_2 \mid \dots \mid s_n$

# *Problemas com a Gramática*

*Na primeira Gramática utilizada, as funções matemáticas eram:*

- ✓ Adição*
- ✓ Subtração*
- ✓ Multiplicação*
- ✓ Divisão*
- ✓ Potenciação*
- ✓ Raiz*

*Tratar Exceções.*

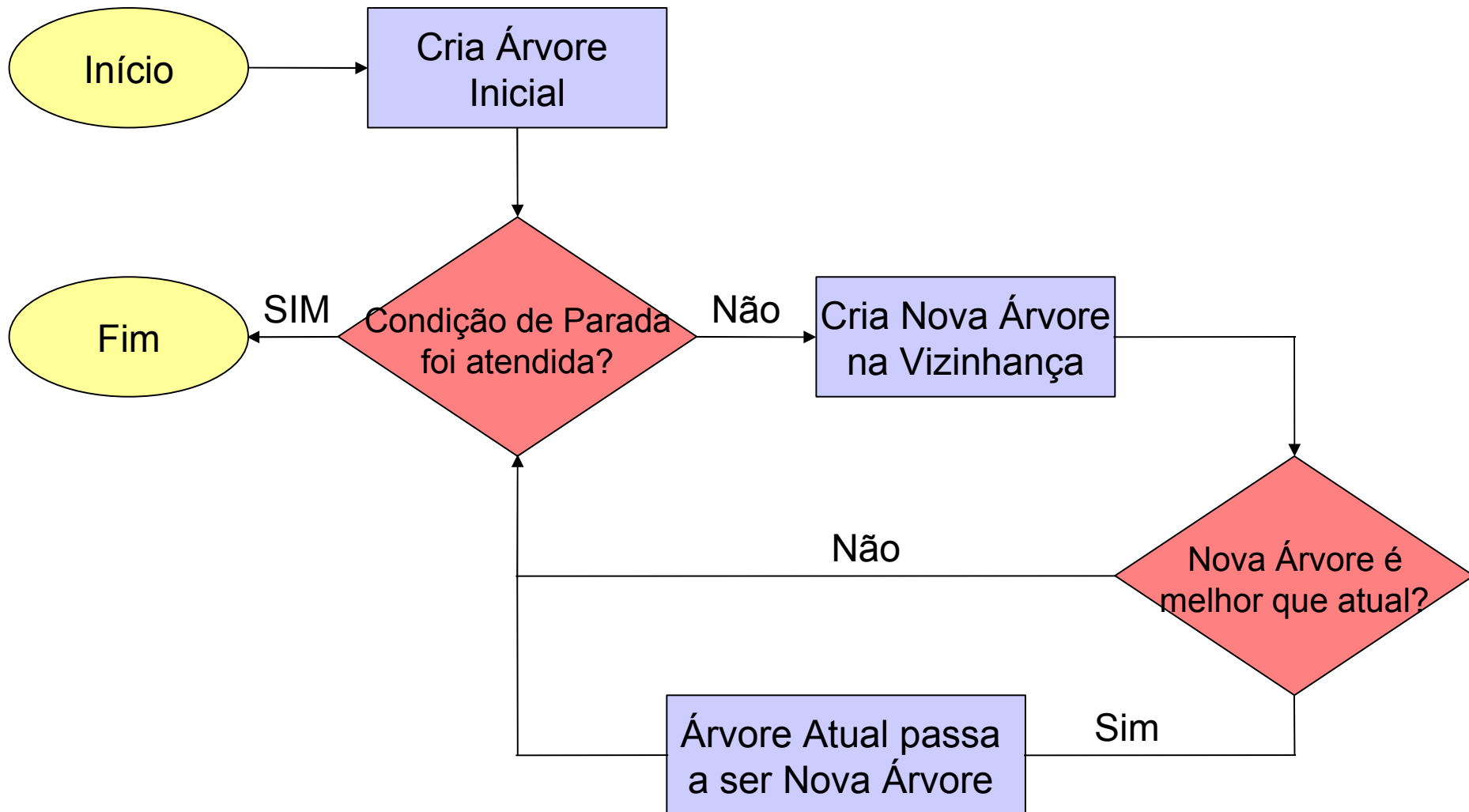
# *Base de Dados*

- *Este trabalho utiliza a base de testes com dados reais, conhecida como Proben1, publicada no Relatório Técnico de Prechelt (1994), focando na base de células cancerígenas.*
- *A base de testes descreve informações retiradas de células de câncer de mama, obtidas através de imagens digitalizadas. Para cada uma das 699 amostras da base, são informadas 9 características das células, como por exemplo, espessura, uniformidade, formato, etc.*
- *Para cada célula, é também indicado se a mesma possuía câncer maligno ou benigno.*

# *Base de Dados*

- *A Base de Dados câncer é composta por 3 arquivos que possuem os mesmos dados, porém em ordem diferentes, para avaliar a real capacidade de classificação do Sistema.*
- *Cada Arquivo é dividido em três partes:*
  - ✓ *Conjunto de Treinamento*
  - ✓ *Conjunto de Validação*
  - ✓ *Conjunto de Teste*

# *Algoritmo de Treinamento*



# *Solução Inicial(Inicialização da Árvore)*

- Nenhuma heurística foi utilizada para gerar a árvore inicial
- O método aqui utilizado é baseado no Algoritmo *Ramped Half-and-half* (RHH), explicado em Eiben (2003) na forma Full method.
- A árvore é gerada aleatoriamente com altura 3.

# *Condição de Parada*

- É analisada pelo conjunto de Validação
- Pode ser verificada de duas maneiras:
  - ✓ Quantidade de acerto
  - ✓ Valor do SSE

# *Quantidade de Acerto*

*Para cada célula faça*

*se ((expressão.pocessa(padrao i) > 0,5) e (padrao i == 1))*

*ou ((expressão.pocessa(padrao i) < 0,5) e (padrao i == 0))entao*

*quantidade de acertos = quantidade de acerto + 1;*

*end*

*end*

# *Valor do SSE*

- Soma dos Erros Quadráticos (*Sum Square Error – SSE*).
- Largamente utilizada na arquitetura de diferentes Redes Neurais Artificiais.
- Mais detalhes podem ser obtidos em Haykin (1998).

$$SSE = \sum_{i=1}^n (\text{esperado}(i) - \text{obtido}(i))^2$$

# *Otimização(Condição de Troca)*

- Assim como nas Redes Neurais Artificiais, este classificador está sujeito ao erro, que deve ser minimizado. No caso desta pesquisa, a minimização é focada na localização de fórmulas mais adequadas para a classificação.
- Para analisar essas fórmulas é usado o conjunto de treinamento

# *Otimização(Condição de Troca)*

Existem três formas de analisar se uma árvore é melhor que a outro:

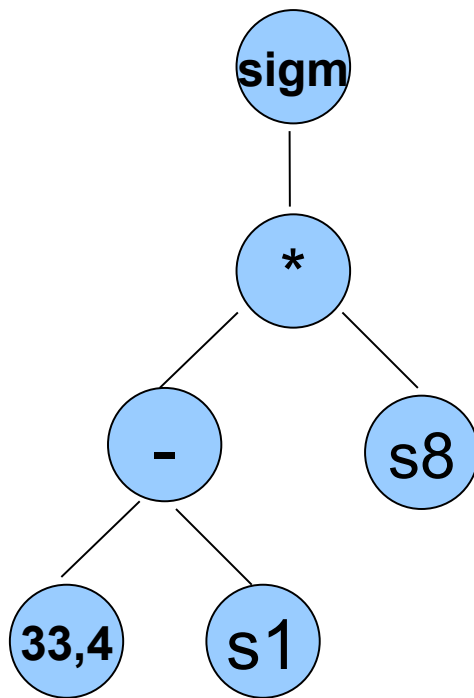
- ✓ Quantidade de acerto
- ✓ Valor do SSE
- ✓ Quantidade de acerto e Valor do SSE

# *Navegação no Espaço de Busca*

- *Durante o processo de busca, este trabalho causa perturbações na melhor árvore conhecida (árvore atual), com o objetivo de criar uma árvore em sua vizinhança.*

# *Mutação 1*

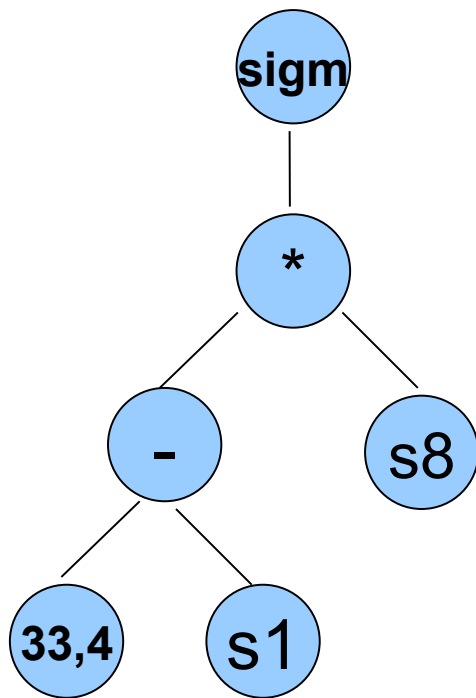
Arvore Atual:



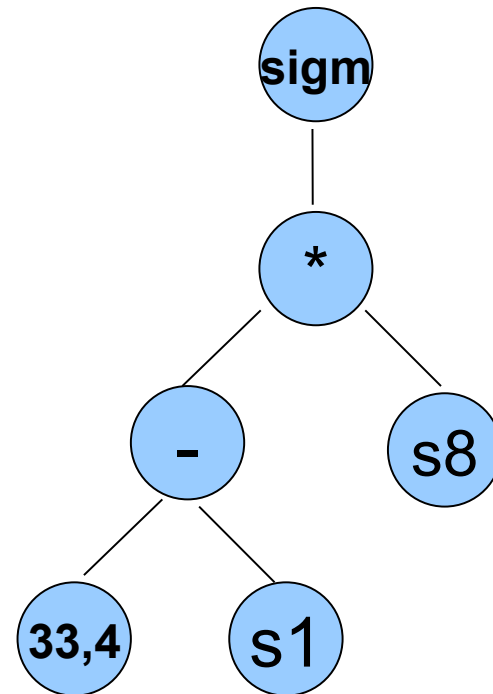
Arvore Nova:

# *Mutação 1*

Arvore Atual:

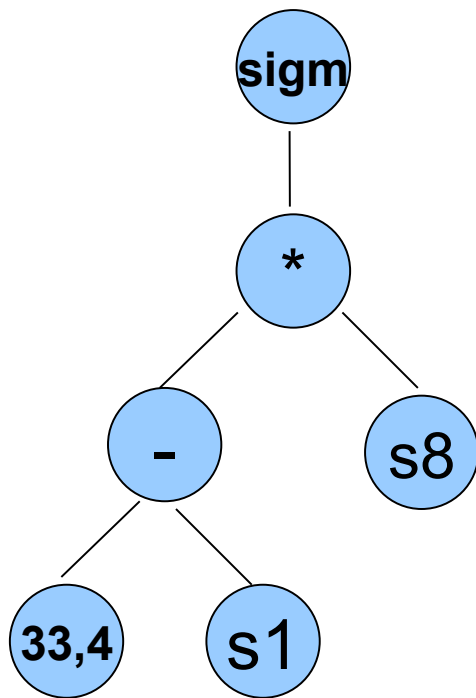


Arvore Nova:

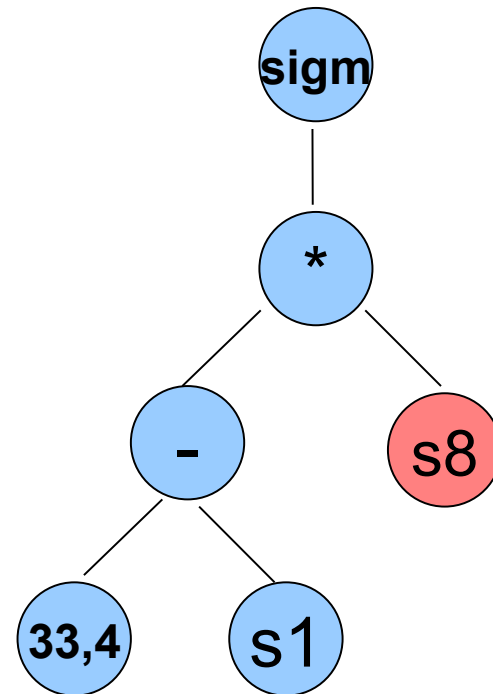


# Mutação 1

Arvore Atual:

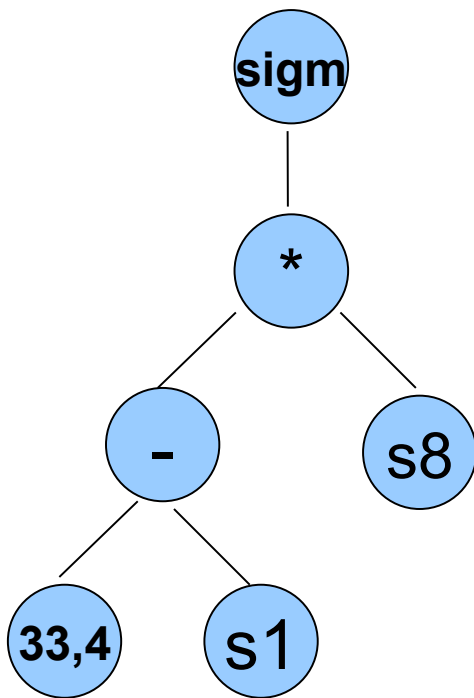


Arvore Nova:

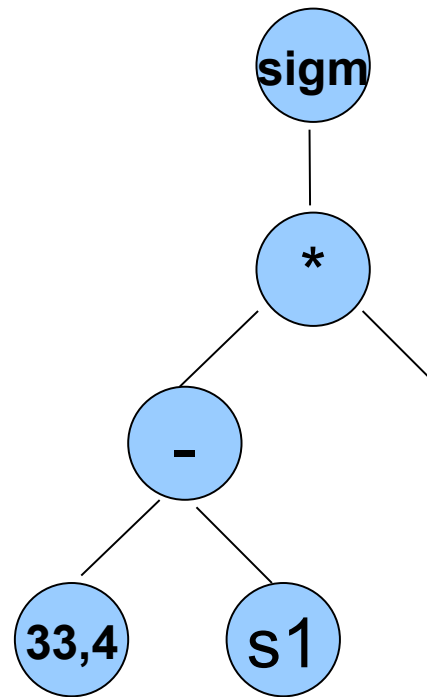


# Mutação 1

Arvore Atual:

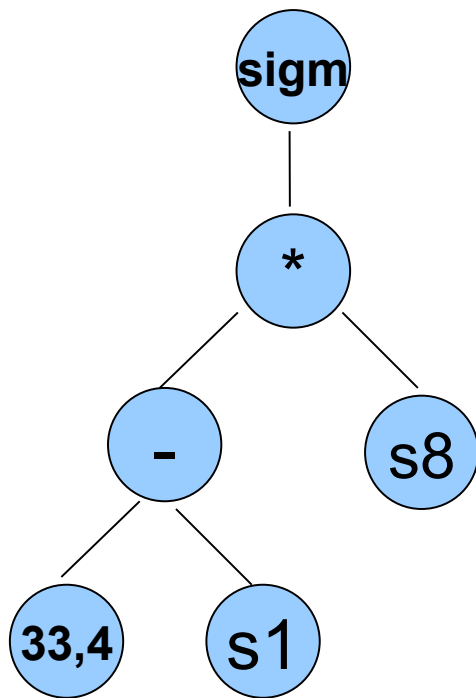


Arvore Nova:

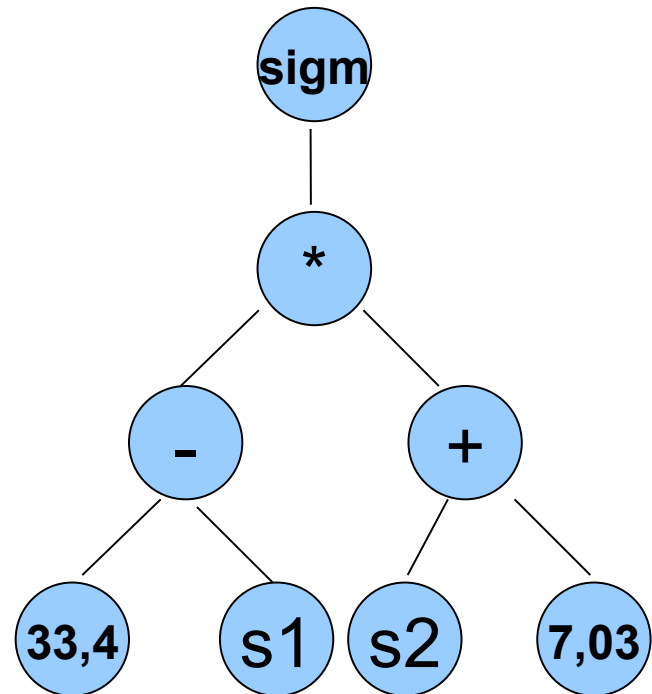


# Mutação 1

Arvore Atual:

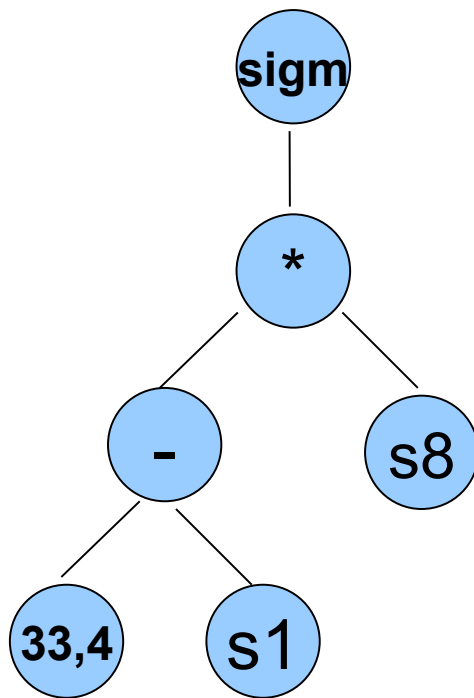


Arvore Nova:



# *Mutação 2*

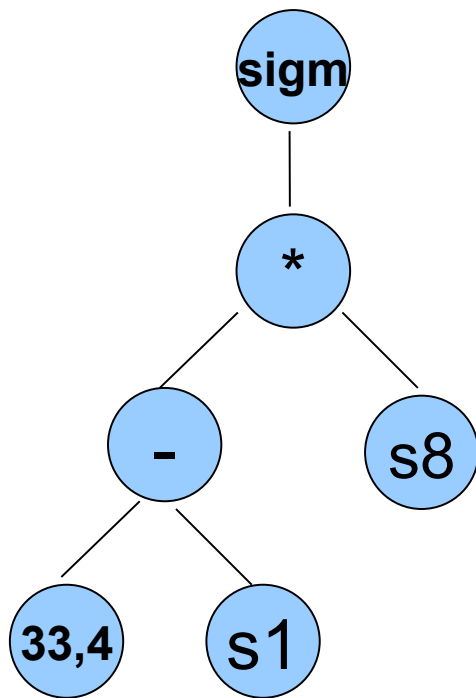
Arvore Atual:



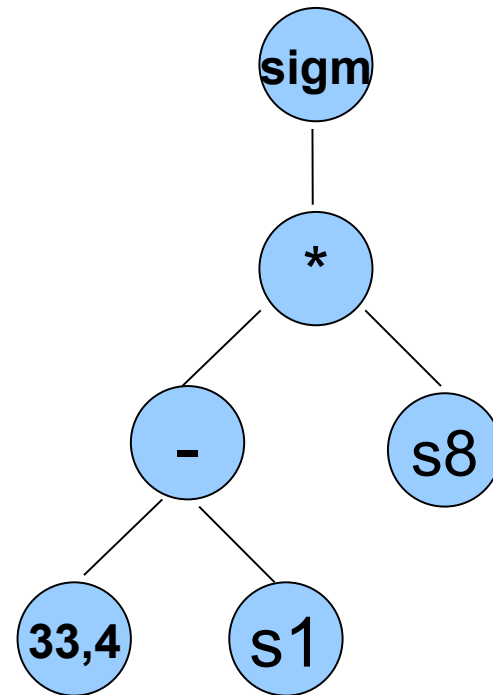
Arvore Nova:

# Mutação 2

Arvore Atual:

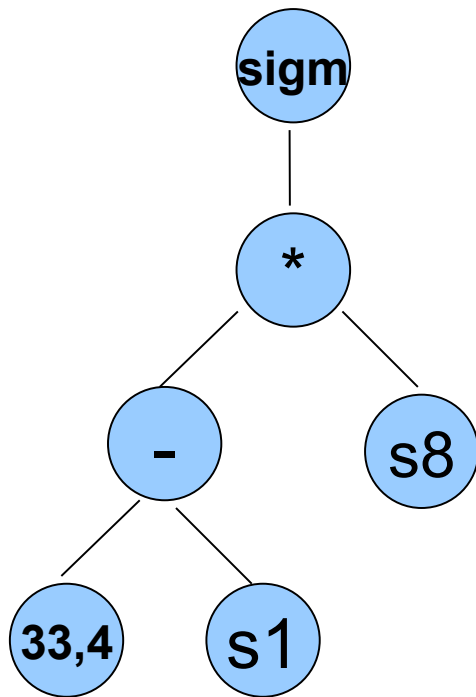


Arvore Nova:

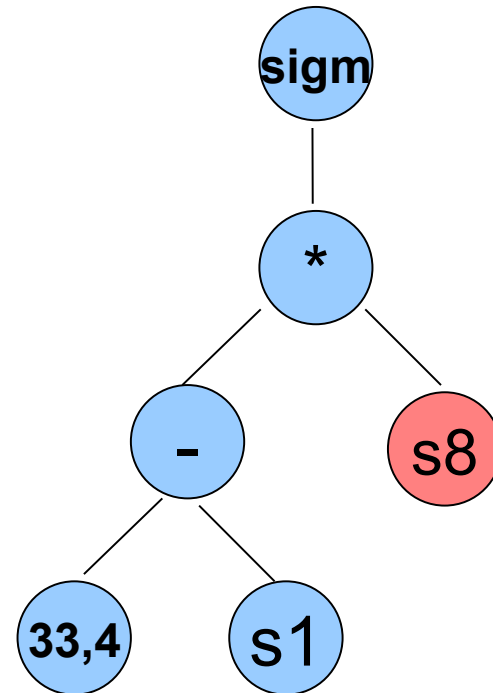


# Mutação 2

Arvore Atual:

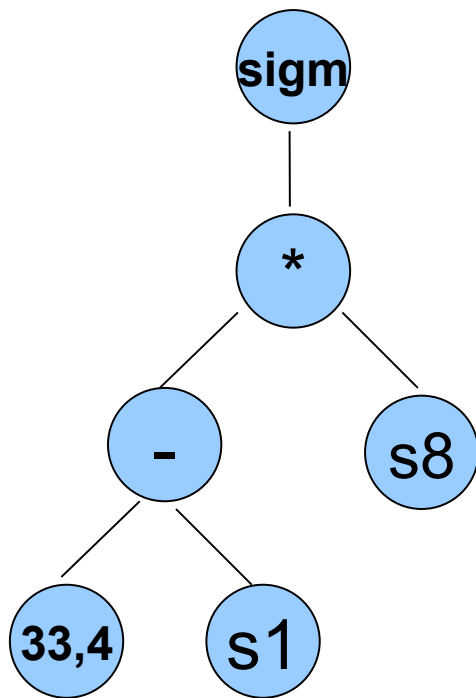


Arvore Nova:

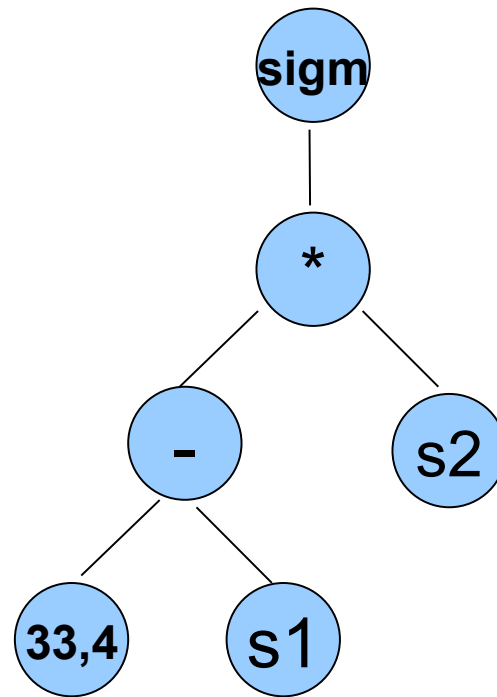


# Mutação 2

Arvore Atual:



Arvore Nova:



# *Algoritmo Subida na Encosta*

*As árvores aceitas como atual sempre melhoram na medida em que o código é executado, sendo assim esta implementação utiliza um Algoritmo de Subida na Encosta.*

# *Algoritmo Subida na Encosta*

*Problema:*

*Ficar preso em Mínimos e Máximos locais.*

*Solução???*

# *Resultados*

## ***Maiores Percentuais de Acerto***

| <b>Arquivo</b> | <b>Percentual de Acerto</b> | <b>Características utilizadas</b> |
|----------------|-----------------------------|-----------------------------------|
| cancer1.dt     | 99,42 %                     | 7                                 |
| cancer2.dt     | 96,55 %                     | 8                                 |
| cancer3.dt     | 97,12 %                     | 8                                 |

# *Resultados*

## *Cancer1*

### *SSE e Quantidade de Acerto*

| Porcentagem de acerto | Variáveis utilizadas |
|-----------------------|----------------------|
| 98,85%                | 9                    |
| 99,42%                | 7                    |
| 98,85%                | 8                    |

# *Resultados*

## *Cancer1*

### *SSE*

| Porcentagem de acerto | Variáveis utilizadas |
|-----------------------|----------------------|
| 96,55%                | 8                    |
| 97,13%                | 9                    |
| 93,10%                | 8                    |

# *Resultados*

## *Cancer1*

### *Quantidade de Acerto*

| Porcentagem de acerto | Variáveis utilizadas |
|-----------------------|----------------------|
| 96,55%                | 9                    |
| 98,28%                | 8                    |
| 97,70%                | 9                    |

# *Bibliografia*

**Eiben, E., Smith, J. E.,** *Introduction to Evolutionary Computing.* Natural Computing Series. MIT Press. Springer. Berlin. 2003.

**Prechelt, L. ,** *Proben1: A set of Neural Network Benchmark Problems and Benchmarking Rules.* Technical Report 21/94, Fakultät für Informatik, Universität Karlsruhe, 76128 Karlsruhe, Germany, September, 1994.

**Oliveira, H. C. B. de;** Otimização de Árvores Sintáticas Aplicada ao Problema de Classificação de Padrões; *XL Simpósio Brasileiro de Pesquisa Operacional.* 2008.

**Prechelt, L. ,** *Proben1: A set of Neural Network Benchmark Problems and Benchmarking Rules.* Technical Report 21/94, Fakultät für Informatik, Universität Karlsruhe, 76128 Karlsruhe, Germany, September, 1994.