2.3.1 Classifier

Descrição do parâmetro: Classifier é um parâmetro categórico que determina qual equação deve ser usada para a classificação dos documentos de patentes.

Se classifier igual a nb, então temos que $p(c \mid d)$ é calculado de acordo com a expressão:

$$p(c \mid d) = p(c \mid ipc) * p(d \mid c) \quad (3)$$

No qual de forma bem simplificada $p(c \mid d)$ representa a probabilidade de uma área científica c dado um documento (a posteriori), $p(c \mid ipc)$ é a probabilidade de área científica c dado código IPC (a priori) e $p(d \mid c)$ é a probabilidade de um documento dado uma área científica (verossimilhança).

Se classifier igual a priori, então temos que $p(c \mid d)$ é calculado de acordo com a expressão:

$$p(c \mid d) = p(c \mid ipc) \quad (4)$$

Nesse caso, a equação 1 foi simplificada e o produto entre a priori e verossimilhança para cálculo da a posteriori foi reduzido ao cálculo da a priori.

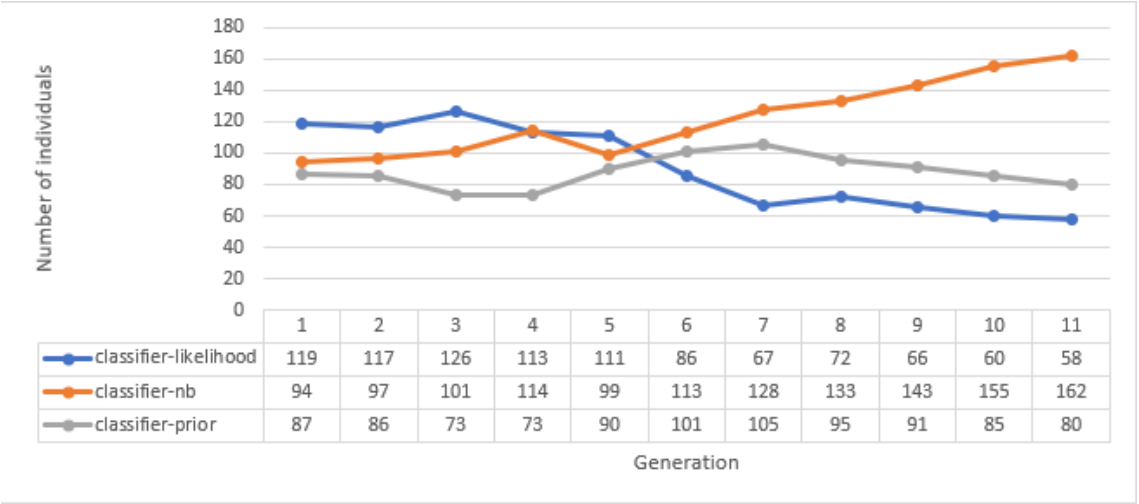
Se classifier igual a likelihood, então temos que $p(c \mid d)$ é calculado de acordo com a expressão:

$$p(c \mid d) = p(d \mid c) \quad (5)$$

Nesse caso, a equação 1 foi simplificada e o produto entre a priori e verossimilhança para cálculo da a posteriori foi reduzido ao cálculo da verossimilhança.

Justificativa para o parâmetro: modelos ensemble em aprendizado de máquina na medida que harmonizam os resultados de uma classificação, tendem a apresentarem resultados superiores aos resultados obtidos pelos classificadores isoladamente Zhang & Ma (2012). O parâmetro classifier visa constatar se a multiplicação entre a priori e a verossimilhança é superior a inferência quando considerada somente o valor da a priori ou somente o valor da verossimilhança como resultado da classificação.

A figura abaixo representa a evolução da presença do parâmetro classifier ao longo das gerações:



	1	2	3	4	5	6	7	8	9	10	11
classifier-likelihood	119	117	126	113	111	86	67	72	66	60	58
classifier-nb	94	97	101	114	99	113	128	133	143	155	162
classifier-prior	87	86	73	73	90	101	105	95	91	85	80

Figura 3: Evolução da presença do parâmetro classifier ao longo das gerações.

Observe que, se comparado aos classificadores a priori e likelihood isoladamente, o classificador naive bayes se tornou predominante na população de classificadores ao longo das gerações, o que nos revela a tendência de que os indivíduos de que usam Naive Bayes são melhores ajustados para a tarefa de classificação. O valor de classifier para o melhor indivíduo encontrado na população foi nb, considerado como melhor valor para o parâmetro.

2.3.2 IPC Digits

O código IPC (https://www.wipo.int/classifications/ipc/en/) é um identificador numérico para uma área tecnológica. Para este estudo, consideramos dividir os códigos IPCs nos primeiros 1 dígito (classificação mais generalizada), 2 dígitos e 4 dígitos (classificação mais especializada).

O dado a priori $p(c \mid ipc)$ é definido como a probabilidade de uma patente pertencer a uma área científica dado uma subdivisão do código ipc. O parâmetro IPC Digits determina em quantos dígitos é necessário subdividir o código IPC para o cálculo da tabela a priori.

Justificativa para o parâmetro: na medida que o código ipc é subdividido, a classificação tecnológica de uma patente é mais especializada. O objetivo do parâmetro IPC Digits é encontrar em qual especialização do código IPC o classificador a priori é mais discriminante para classificar a área científica de uma patente.

A figura abaixo representa a evolução da presença do parâmetro IPC Digits ao longo das gerações:

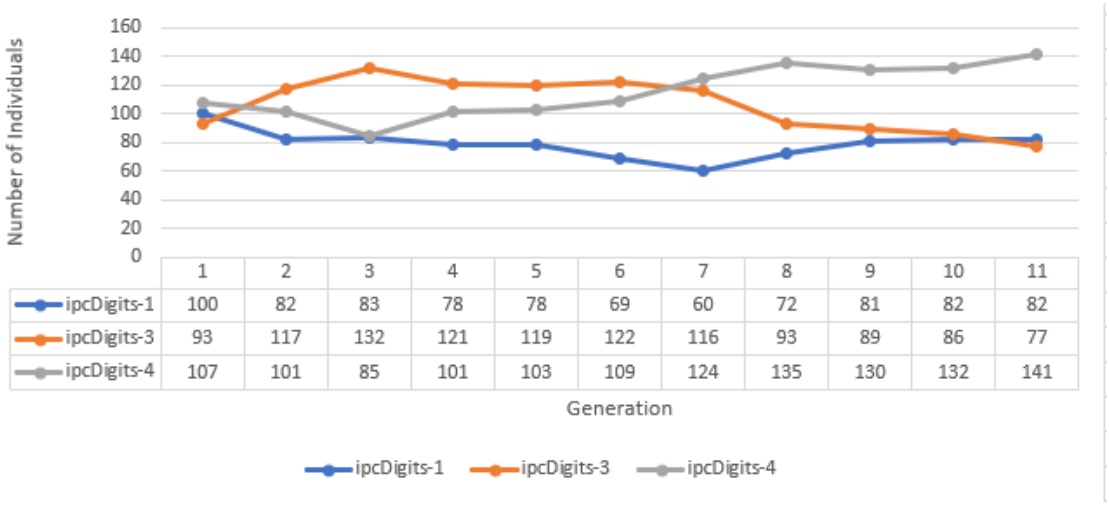


Figura 4: Evolução da presença do parâmetro IPC Digits ao longo das gerações.

Observe que a subdivisão do código IPC nos 4 dígitos tornou-se predominante na população ao longo das gerações. O valor de IPC Digits para o melhor indivíduo encontrado na população foi de 4, considerado como melhor valor para o parâmetro.

2.3.3 Likelihood Amplification

O parâmetro likelihood amplification (la) é uma constante que eleva o valor da verossimilhança na inferência por método bayesiano tal que:

$$p(c \mid d) \;=\; p(d|c)^{la} * \; p(\, c \mid ipc\,) \qquad (6)$$

No qual, de forma bem simplificada, $p(c \mid d)$ representa a probabilidade de uma área científica c dado um documento (a posteriori), $p(c \mid ipc)$ é a probabilidade de área científica c dado código IPC (a priori) e $p(d \mid c)$ é a probabilidade de um documento dado uma área científica (verossimilhança) e la é o parâmetro likelihood amplification.

Justificativa para o parâmetro: A probabilidade de um documento 'd' pertencer a uma área científica 'c' é calculada de acordo com inferência bayesiana apresentada na equação 3 do parâmetro classifier. Mas suponha que o valor de $p(c \mid ipc)$ é muito alto para uma área científica c. Para que o classificador bayesiano seja capaz de inferir corretamente uma instância não pertencente a classe científica c, é necessário que o valor da verossimilhança

$p(d \mid c)$ seja suficientemente grande para inverter a suposição a priori. O parâmetro 'la' é um expoente que ajusta o peso da verossimilhança na classificação.

A figura abaixo representa a evolução da presença do parâmetro likelihood amplification ao longo das gerações:

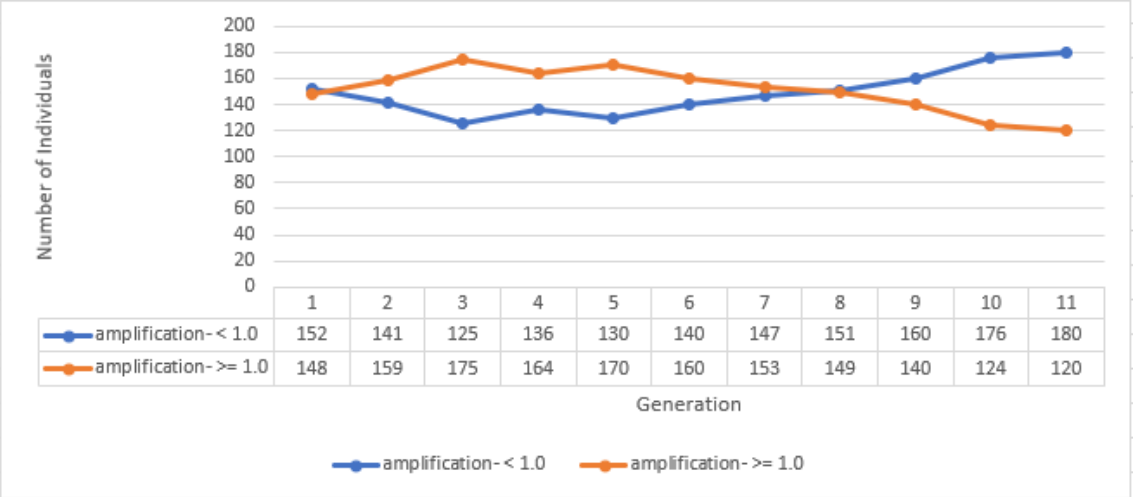


Figura 5: Evolução da presença do parâmetro Likelihood Amplification ao longo das gerações.

Observe que para o gene likelihood amplification, encontramos que os melhores valores encontrados tendem para números inferiores a 1. O valor de la para o melhor indivíduo encontrado na população foi de 0.4 considerado como valor para o constante.

2.3.4 Corte dos tokens

O parâmetro corte dos tokens é uma constante que define o limite inferior que um token deve atingir na expressão abaixo para ser considerado durante a seleção das melhores features do classificador verossimilhança:

$$\text{getWeight (token } p \text{, c)} \geq \left(\sum_{i=0}^{n} \text{getWeight (token } p, c_i\right)) / \text{ct} \quad (7)$$

No qual a função getWeight (token p, c) retorna o peso de um token p em uma área da ciência c, e ct representa a constante de corte dos tokens. Se o peso de um token em uma área científica c for maior que a somatória dos pesos desse mesmo token nas outras áreas científicas dividido por uma constante ct, o token torna-se uma feature do modelo. Caso contrário, ele é descartado.

O algoritmo 1 mostra a implementação em pseudocódigo da função de seleção de features chamada *selectFectures* que recebe como entrada o abstract de uma patente (Patent

Abstract), uma lista de áreas da ciência e uma matriz que diz qual o peso de um token dado uma área da ciência (n-grams) de 3 colunas (Área da ciência, token e peso do token na área da ciência).

Algoritmo 1: Seleção das melhores features de um abstract de patente

function *selectFeatures* (patent-abstract, Glänzel classification, n-grams) **returns** features

tokens ← extrai tokens de patent-abstract

for each science field (fi) **in** Glänzel classification

for each token (p_n)**in** tokens

if (getWeight (token p_n, science field) $\geq$

$$(\sum_{i=0}^{n} \text{getWeight (token } p_n \text{, science field i))/ct)}$$

then features ← token p_n

return features

Justificativa para o parâmetro: a seleção das melhores features para tarefas de classificação é amplamente utilizada em aprendizado de máquina (Manning, Raghavan & Schütze, 2008). O objetivo é remover tokens pouco discriminantes, como stopwords, da equação de classificação. Quanto menor for o valor de corte dos tokens, maior o número de palavras consideradas para a computação da verossimilhança. Por outro lado, quanto maior for o corte dos tokens, apenas jargões ou tokens que representem expressões técnicas serão consideradas. O objetivo é encontrar um valor de corte que não seja tão restritivo a ponto de excluir todos os tokens da equação, mas que também não seja tão abrangente a ponto de permitir a computação de todos os tokens observados.

Não houve convergência de um único valor para a constante de corte dos tokens na população. O valor de corte dos tokens para o melhor indivíduo encontrado na população foi de 3.0 considerado como valor para o constante.

2.3.5 Alisamento das Classes

Retome a equação (3) $p(c \mid d) = p(c \mid ipc) * p(d \mid c)$. Se $p(c \mid ipc) = 0$ ou $p(d \mid c) = 0$, o valor $p(c \mid d)$ será 0 independentemente do valor de $p(d \mid c)$ou $p(c \mid ipc)$.

Alisamento das classes é um parâmetro booleano que indica se inferências com frequência 0 devem receber uma constante maior que 0 para evitar que seja impossível inferir elementos dessa classe no cálculo de probabilidade a posteriori.

Justificativa para o parâmetro: Para atributos que não ocorram dada uma classe científica, a equação a posteri atribui o valor de zero. Na literatura são recomendados diversos métodos de smoothing entre eles o de Laplace Peng, F., Schuurmans, D., & Wang, S. (2004) que foi utilizado neste trabalho como hiperparâmetro. Um processo complementar consiste no incremento da amostra com a finalidade de evitar ou diminuir o número de classes com frequência zero. Este procedimento segue os resultados mostrados em Dominique & Van Pottelsberghe de la Potterie(2002).

A figura abaixo representa a evolução da presença do parâmetro alisamento das classes ao longo das gerações:

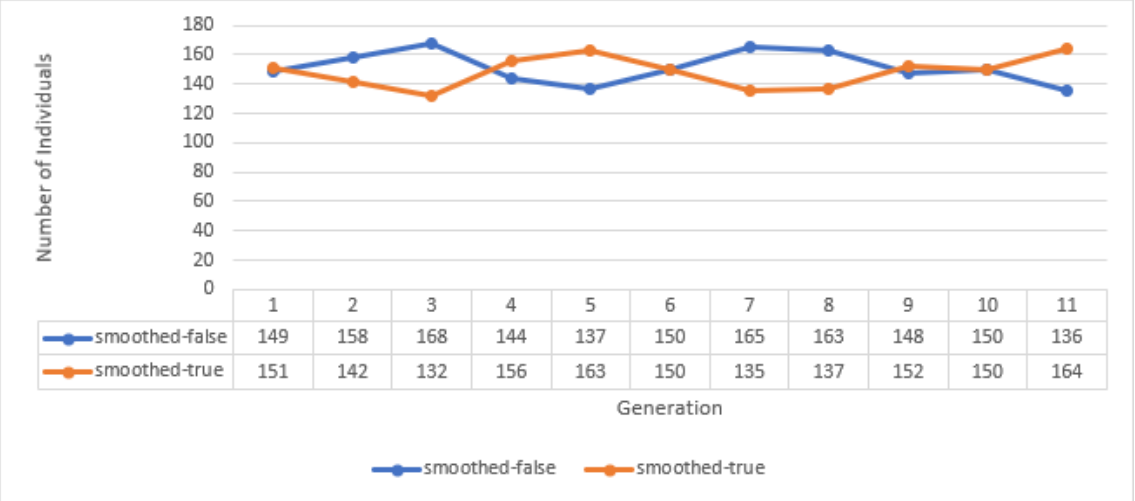


Figura 6: Evolução da presença do parâmetro Alisamento das Classes ao longo das gerações.

A presença dos genes de alisamento das classes flutuou ao longo das gerações. Isso ocorre, pois o alisamento das classes de frequência zero não exerceu influência no valor da acurácia dos indivíduos. Por conta da metodologia de anotação automática de patentes que permite um grande volume da amostra de treino, as chances de haver classes não observadas na amostra de treino é minimizada. O indivíduo de melhor fitness variou para o valor de alisamento das classes.