Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Coordenação de Pós-Graduação em Ciência da Computação

Explorando Aprendizagem Ativa para Reduzir o Esforço Manual na Geração de Gabaritos para Resolução de Entidades

Diego Fernandes de Araújo

Dissertação submetida à Coordenação do Curso de Pós-Graduação em Ciência da Computação da Universidade Federal de Campina Grande - Campus I como parte dos requisitos necessários para obtenção do grau de Mestre em Ciência da Computação.

Área de Concentração: Ciência da Computação
Linha de Pesquisa: Sistemas de Informação e Banco de Dados

Prof. Carlos Eduardo Santos Pires
(Orientador)

Campina Grande, Paraíba, Brasil

A663e Araújo, Diego Fernandes de.
Explorando aprendizagem ativa para reduzir o esforço manual na geração de gabaritos para resolução de entidades / Diego Fernandes de Araújo. – Campina Grande, 2019.
76 f. : il. color.

Dissertação (Mestrado em Ciência da Computação) – Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática, 2019.
"Orientação: Prof. Dr. Carlos Eduardo Santos Pires".
Referências.

1. Bases de Dados. 2. Resolução de Entidades. 3. Aprendizagem de Máquina. 4. Deduplicação. 5. Gabarito. 6. Aprendizagem Ativa. 7. Classificação. I. Pires, Carlos Eduardo Santos. II. Título.

CDU 004.65(043)

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECÁRIA SEVERINA SUELI DA SILVA OLIVEIRA CRB -15/225

"EXPLORANDO APRENDIZAGEM ATIVA PARA REDUZIR O ESFORÇO MANUAL NA GERAÇÃO DE GABARITOS PARA RESOLUÇÃO DE ENTIDADES"

DIEGO FERNANDES DE ARAÚJO

DISSERTAÇÃO APROVADA EM 22/05/2019

CARLOS EDUARDO SANTOS PIRES, Dr., UFCG
Orientador

LEANDRO BALBY MARINHO, Dr., UFCG
Examinador

BERNADETTE FARIAS LÓSCIO, Dra., UFPE
Examinadora

CAMPINA GRANDE - PB

Resumo

Diversos métodos de Resolução de Entidades (RE) têm sido desenvolvidos na academia e indústria ao longo dos anos com o intuito de identificar entidades (e.g. registros) duplicadas em bases de dados a fim de tratá-las. Para avaliar a qualidade dos resultados de tais métodos, é necessário compará-los com um gabarito, que consiste em um documento contendo todos os pares de registros duplicados conhecidos em uma base de dados. A geração desses gabaritos para bases de dados reais é feita de forma manual a partir da inspeção de todas as combinações de pares de registros existentes nessas bases. Isso apresenta complexidade quadrática, com relação ao(s) tamanho(s) da(s) base(s) de dados, o que acarreta na necessidade de bastante tempo para realização da tarefa e na possibilidade de introdução de erros. Em virtude disto, alguns trabalhos apresentam abordagens automáticas ou semiautomáticas para geração de gabaritos para a tarefa de RE que, no entanto, ou não são aplicáveis a domínios variados ou ainda requerem um esforço manual considerável. Neste trabalho é proposta GTGenERAL, uma abordagem semiautomática que combina resultados de múltiplos algoritmos de RE juntamente com Aprendizagem Ativa para gerar gabaritos, com redução de esforço manual. Experimentos usando bases de dados reais mostram que a abordagem é capaz de gerar gabaritos próximos àqueles gerados pela abordagem do estado da arte, enquanto reduz substancialmente o esforço manual empreendido no processo.

Palavras-chave: Resolução de Entidades, Deduplicação, Gabarito, Aprendizagem de Máquina, Aprendizagem Ativa, Classificação.

Abstract

Several methods of Entity Resolution (ER) have been developed both at academia and industry over the years, with the aim to identify duplicate entities (e.g. records) in datasets. To evaluate the efficacy of such methods, it is necessary to compare their results with a ground-truth, which consists of a document containing all known duplicate record pairs in a dataset. In general, the generation of ground-truths for real datasets is done manually from the inspection of all combinations of pairs of records in a dataset. However, this is subject to error and presents quadratic complexity, with respect to the size(s) of the dataset(s), requiring a long time to be performed. In this context, some works present (semi)automatic approaches for the generation of ground-truths for the ER task. However, such approaches are either not applicable to several domains or still require a considerable manual effort. In this work, we propose GTGenERAL, a semiautomatic approach which combines results from multiple algorithms of ER together with Active Learning to generate ground-truths employing reduced manual effort. Experiments using real datasets show that, with great manual effort reduction, GTGenERAL is able to generate ground-truths close to those generated by the state-of-the-art approach, while substantially reducing the manual effort undertaken in the process.

Keywords: Record Linkage, Deduplication, Ground-truth, Machine Learning, Active Learning, Classification.

Agradecimentos

Feliz quem reconhece que a caminhada da vida, em seus diversos estágios, não se pode fazer sozinho. Com isso em mente, inicio os agradecimentos pela conclusão dessa etapa tão conturbada da minha vida, que foi o mestrado, lembrando daquele que certamente esteve comigo em todos os momentos: Deus.

Foi Ele a quem tanto recorri, ora pedindo inspiração e capacidade para realizar as tarefas que me eram incubidas, ora pedindo ânimo para seguir a jornada. Principalmente quando, ao sair da UFCG, chegava no apartamento onde residia, sem a possibilidade de contar com a presença da minha família, como era tão costumeiro aos meus colegas. Mas Ele sempre esteve presente, e me fazia lembrar disso ao ler uma folha de caderno colada na parede do meu quarto com os dizeres "Não tenha medo, pois Eu estou aqui.". E isso me reconfortava, por mais simples que pudesse parecer.

Gratidão imensa à minha família: meu pai Sebastião e minha mãe Edleuza, que me incentivaram no caminho da educação, e meu irmão Thiago com sua esposa Lidiane que, juntamente com meus pais, sempre estiveram torcendo por mim. Família é o que temos de mais importante nessa vida. De nada adiantam conquistas profissionais ou títulos acadêmicos, se não houver uma família com que compartilhar os frutos resultantes desses êxitos.

À minha esposa Jussara agradeço pelo incentivo inicial e constante que me deu para que buscasse o mestrado. Foi ela que enxergou capacidade em mim, quando eu mesmo não enxergava. Foi por ela, também, que mantive o foco na conclusão dessa jornada.

Agradecimentos aos familiares e amigos, que estiveram sempre na torcida, em especial a Renan e Thalisson, ambos grandes amigos que a vida me trouxe. Ao primeiro, já mestre, por apontar o caminho a ser percorrido e pelos conselhos inúmeros. Ao segundo, por ter compartilhado o desejo pela pós-graduação e anseios decorrentes dessa escolha, e por ter iniciado a busca por esse caminho juntamente comigo.

Agradeço a Poliano pelo acolhimento em Campina Grande e pelas palavras de motivação, mesmo quando extremamente atarefado e, consequentemente, preocupado com seu mestrado e trabalho.

Agradeço aos professores do programa de pós-graduação da UFCG: Gustavo, Leandro, Nazareno e André Augusto, pelo aprendizado promovido em suas aulas, que foi aplicado

durante a execução da pesquisa realizada e que certamente será consolidado no decorrer da minha vida acadêmica.

À Universidade Estadual da Paraíba, que desde 2007 permeia minha vida enquanto aluno, professor e técnico-administrativo, agradeço o crescimento pessoal e profissional proporcionados, além do incentivo para capacitação no mestrado a partir da liberação das atividades laborais.

Aos integrantes do Laboratório de Qualidade de Dados (LQD), inúmeros agradecimentos:

- A Demetrio, por ter me apontado tantos caminhos acerca da pesquisa, o que me fez muitas vezes chamá-lo de coorientador, mas principalmente pelas tantas conversas e risos que amenizavam a tensão do trabalho;

- A Thiagão, com quem tive o primeiro contato na prévia de sua qualificação, ocasião em que o professor Carlos disse "Diego, esse é você daqui a um ano.". Ótima pessoa em quem se espelhar. O agradecimento pelo apoio técnico com os servidores disponibilizados para realização da pesquisa, mas principalmente pelas conversas e apoio ao ouvir meus desabafos, em especial na noite do churrasco quase frio;

- À Andreza e Veruska, que de tão indissociáveis, são citadas aqui juntas mesmo. Sei que prometi várias páginas de agradecimentos para ambas, mas espero que minhas atitudes em relação a vocês transpareçam o quão grato sou por compartilharmos os anseios do mestrado que cursamos juntos e pelo apoio mútuo que empreendemos. Vocês foram, por tantas vezes, acalanto quando estive em Campina, e minhas mãos e pés, quando lá não pude estar e precisei resolver algo;

- A Lucas que, apesar da passagem rápida pelo laboratório, engrossava o caldo das boas conversas junto das meninas;

- A Tiago Finlandês, patoense forte, pelo apoio no início da minha jornada na UFCG e pelas piadas a nível Tiago Brasileiro;

- A Ígor, pelos debates acerca de aprendizagem de máquina, que tanto me auxiliaram na execução da minha pesquisa;

- A Cleilton que, apesar de não ter concluído a jornada que iniciamos juntos na UFCG, compartilhou comigo tantos anseios em relação a ter que cursar um mestrado longe de casa, tendo que se desdobrar para cuidar da família e dar conta das responsabilidades do trabalho.

Por fim, os agradecimentos àqueles que tiveram participação direta no êxito deste trabalho de mestrado:

- Dimas, que em um curto espaço de tempo trouxe questionamentos e sugestões que contribuíram sobremaneira para a melhoria da pesquisa realizada, além do primoroso auxílio nas revisões realizadas na escrita dos textos resultantes do trabalho;

- Professor Carlos que, com sua orientação e cobranças, ajudou-me a delinear e executar a pesquisa, sempre buscando a primazia no trabalho realizado. Certamente levarei muitas de suas atitudes enquanto profissional em minha carreira futura.

Conteúdo

Lista de Símbolos

AA - Aprendizagem Ativa

AdInTDS - *Adaptive and Interactive Training Data Set Selection for Entity Resolution*

AM - Aprendizagem de Máquina

AS - *Annealing Standard*

GTGenERAL - ***G**round-**T**ruth **Gen**erator for **E**ntity **R**esolution with **A**ctive **L**earning*

RE - Resolução de Entidades

STERSWin - ***S**election of **T**raining Dataset for **E**ntity **R**esolution with **S**liding **Win**dow*

Lista de Figuras

Lista de Tabelas

Lista de Algoritmos

Capítulo 1

Introdução

A possibilidade de os sistemas de informações reunirem informações de diversas fontes de dados, muitas vezes heterogêneas, aliada ao elevado volume de dados por elas armazenados, potencializa a existência de problemas referentes à qualidade dos dados. Dentre os problemas, pode-se destacar a presença de registros duplicados em bases de dados, os quais consistem em múltiplas representações de uma mesma entidade do mundo real [2; 37]. Diversos motivos podem levar à ocorrência desses registros duplicados como erros de digitação, mudança de endereço ou nomes de pessoas que não foram devidamente atualizados, entre outros [11].

Na Tabela 1, é possível observar um exemplo fictício em que a entidade do mundo real *Michael Jackson* possui seus dados duplicados em uma base de dados devido a erros de digitação ocorridos no preenchimento dos atributos *Nome* e *Nascimento* e à falta de alguns dados no campo *Endereço* de um dos registros.

Tabela 1.1: Exemplo de registro duplicado em uma base de dados.

Id	**Nome**	**Nascimento**	**Endereço**
79	Michael Jackson	05/05/1980	R.: Hollywood Boulevard Nº 15
...	...	...	
981	Mychael Jakson	05/05/1908	R.: Hollywood B. Nº 15

Diversos métodos de Resolução de Entidades (RE)[1] têm sido desenvolvidos na academia

[1]Também conhecida como deduplicação de dados, correspondência entre entidades ou resolução de dados.

e indústria ao longo dos anos com o intuito de identificar registros duplicados em bases de dados. Para determinar os pares duplicados, tais métodos baseiam-se na similaridade entre os valores dos atributos, como endereço, nome e data de nascimento, dos pares de registros [17; 19; 31]. O valor de similaridade gerado após a comparação dos valores dos atributos é utilizado para classificar cada par de registros como duplicado ou não-duplicado.

A classificação dos pares de registros, para os diversos métodos de RE, ocorre a partir de dois tipos de abordagens: i) aquelas que confiam no conhecimento do especialista no domínio da(s) base(s) de dados, as quais geralmente combinam uso de funções de distância e regras lógicas para esse fim; ou ii) abordagens que, a partir de dados de exemplo, são treinadas para identificar esses pares de registros, como aquelas que utilizam algoritmos de Aprendizagem de Máquina (AM)[2] [19].

Independentemente da abordagem de classificação empregada, a qualidade dos resultados produzidos pelos métodos de RE é normalmente mensurada com o uso de métricas como *Precision*[3] e *Recall*[4], apresentadas nas Equações 1.1 e 1.2, as quais necessitam ter acesso às ocorrências de casos verdadeiro positivos (VP), falso positivos (FP) e falso negativos (FN) identificados pelo método de RE utilizado. Para tal, é necessário comparar os registros duplicados identificados pelo método de RE com um gabarito, que consiste em um conjunto contendo os pares de registros duplicados conhecidos em uma base de dados [12; 25; 35].

$$precision = \frac{|VP|}{|VP| + |FP|} \tag{1.1}$$

$$recall = \frac{|VP|}{|VP| + |FN|} \tag{1.2}$$

1.1 Motivação e Problematização

Para bases de dados reais, a estratégia *naive* para a geração do gabarito consiste em um ou mais especialistas no domínio da(s) base(s) de dados realizarem a inspeção dos pares formados pelo Produto cartesiano aplicado sobre todos os registros existentes na(s) base(s) de

[2]Em inglês, *Machine Learning*.

[3]*Precision* calcula a proporção de pares duplicados corretos recuperados.

[4]*Recall* afere a proporção de pares duplicados corretamente recuperados.

dados. Nesse caso, a tarefa apresenta complexidade quadrática com relação ao(s) tamanho(s) da(s) base(s) de dados, o que acarreta na necessidade de bastante tempo para avaliação e possibilidade de introdução de erros [11; 30; 39]. Essa dificuldade leva à existência de poucas bases de dados reais com gabaritos disponíveis em repositórios públicos, as quais geralmente são pequenas e restritas a poucos domínios, o que dificulta a realização de experimentos relacionados a RE mais abrangentes [51].

Considerando, por exemplo, uma base de dados real com 100 registros, o número de pares de registros a serem avaliados, após a aplicação do Produto cartesiano sobre todos os registros, é de 4.950. Se a base de dados a ser considerada for apenas dez vezes maior, o número de avaliações passa para 499.500. Esse crescimento quadrático torna inviável a atuação de um especialista humano no domínio da base de dados, responsável pela geração do gabarito para todos os pares de registros.

Dessa forma foi identificada como possibilidade de pesquisa científica a proposição de uma abordagem para geração de gabaritos para avaliar a tarefa de RE. Tal abordagem deve: i) maximizar a correta seleção de pares de registros duplicados e ii) minimizar a aplicação de esforço humano para realização da tarefa.

1.2 Relevância

A existência de poucas bases de dados com gabarito disponíveis em repositórios públicos acarreta na dificuldade de profissionais da indústria e membros da comunidade científica validarem novos métodos relacionados à tarefa de RE.

Nesse sentido, alguns trabalhos científicos têm apresentado abordagens automáticas para geração de gabaritos para a tarefa de RE. Por exemplo, os autores de [25] se baseiam em informações disponíveis nos metadados dos registros avaliados para identificar aqueles que são duplicados. Entretanto, observa-se que a solução apresentada não é generalizável para múltiplos domínios de bases de dados.

Outros trabalhos propõem abordagens semiautomáticas para geração de gabaritos. Por exemplo, os autores de [51] realizam inspeção manual de registros conflitantes entre algoritmos de RE. Entretanto, observa-se que, apesar de haver redução na intervenção humana obtida (quando comparada à estratégia *naive*), ainda há necessidade de uma quantidade con-

siderável de esforço manual para geração de gabaritos para a tarefa de RE.

Desse modo, a necessidade de uma abordagem semiautomática para geração de gabaritos para RE, independente de domínio, que procure reduzir consideravelmente o esforço manual empregado no processo enquanto busque maximizar a quantidade de pares de registro corretos identificados, ressalta a importância deste trabalho.

Assim, com o intutito de treinar algoritmos de AM de maneira acurada para auxiliar na composição de um gabarito, podem ser exploradas estratégias de Aprendizagem Ativa (AA)[5], que buscam gerar conjuntos de treinamento com pouca intervenção manual, sem a necessidade de conjuntos de dados previamente rotulados imposta por estratégias como *Oversampling* e *Undersampling*, as quais manipulam sinteticamente esses dados para compor conjuntos de treinamento [45].

É importante destacar que a tarefa de geração de gabaritos para RE pode se confundir com o desenvolvimento de um classificador ótimo para RE, porém este último em geral necessita realizar sua tarefa de forma eficiente, enquanto a primeira prioriza a eficácia do gabarito a ser gerado, não se preocupando com o tempo necessário para a geração do mesmo.

1.3 Objetivos

O objetivo geral deste trabalho é propor uma abordagem semiautomática para geração de gabaritos que faça uso de algoritmos de AM, a partir da aplicação de estratégias de AA conjuntamente ao uso de resultados de múltiplos algoritmos de RE.

Para que seja possível alcançar o objetivo geral, os seguintes objetivos específicos são necessários:

- estudar formas de combinar resultados de algoritmos de RE;
- identificar estratégias de AA, aplicadas ao problema de RE, que possam ser utilizadas para gerar conjuntos de treinamento capazes de treinar algoritmos de AM;
- compreender e selecionar algoritmos de AM a serem utilizados no processo de classificação dos pares de registros que irão compor gabaritos para RE;

[5]Em inglês, *Active Learning*.

- avaliar a abordagem para geração de gabarito para RE proposta, em termos de esforço manual empreendido para geração do gabarito e em relação à quantidade de entidades duplicadas corretamente identificadas.

1.4 Contribuições

Este trabalho apresenta as seguintes contribuições:

- GTGenERAL - ***G**round-**T**ruth **Gen**erator for **E**ntity **R**esolution with **A**ctive **L**earning*, uma abordagem semiautomática que combina resultados de múltiplos algoritmos de RE e algoritmos de AM para gerar gabaritos para a tarefa de RE, capaz de minimizar o esforço manual empreendido no processo enquanto busca maximizar a seleção de pares de registros que têm a maior probabilidade de serem duplicados; a fim de evitar o emprego de grande esforço manual no processo, a abordagem faz uso de estratégias de AA que buscam selecionar um conjunto de treinamento reduzido, porém representativo, para treinar os algoritmos de AM. GTGenERAL faz uso tanto de estratégias de AA que empregam heurísticas que assumem monotonicidade, as quais partem do pressuposto que um par de registros com alto valor de similaridade é mais provável representar a mesma entidade do mundo real do que um par com baixo valor de similaridade [1; 16], como aquelas que não consideram essa propriedade [3; 13];

- STERSWin, uma estratégia de AA que, por meio do deslizamento de janelas sobre um conjunto de pares obtidos a partir de resultados conflitantes de algoritmos de RE, assumindo monotonicidade, é capaz de compor um conjunto de treinamento balanceado e representativo para treinar classificadores de RE;

- Um estudo experimental empregando bases de dados reais de diversos domínios, para avaliar a abordagem proposta tanto utilizando a estratégia de AA proposta, quanto a estratégia de AA disponível em [13], a qual não assume monotonicidade.

1.5 Organização do Trabalho

Este documento está estruturado da seguinte forma: no Capítulo 2, são abordados os conceitos-chave empregados para compreensão do trabalho. No Capítulo 3, são discutidos trabalhos relacionados. O problema abordado é definido no Capítulo 4, onde também é descrita em detalhes a solução do estado-da-arte. No Capítulo 5 a abordagem GTGenERAL é proposta como alternativa para solucionar o problema abordado neste trabalho. Ainda, é apresentada uma estratégia de AA, que assume monotonicidade, a ser aplicada em GTGenERAL. No Capítulo 6, a abordagem proposta é avaliada usando bases de dados reais. Finalmente, o Capítulo 7 conclui o trabalho e apresenta direcionamentos para trabalhos futuros.

Capítulo 2

Fundamentação Teórica

Neste capítulo, são apresentados os conceitos necessários para compreensão da abordagem proposta neste trabalho. Os conceitos abordados são: i) Tipos de Classificadores para Resolução de Entidades; ii) Aprendizagem de Máquina para Resolução de Entidades; iii) Aprendizagem Ativa para Geração de Conjuntos de Treinamento Representativos; iv) Monotonicidade em Resolução de Entidades; e v) Comitê de Classificadores.

2.1 Tipos de Classificadores para Resolução de Entidades

A tarefa de RE, cujo processo é apresentado resumidamente na Figura 2.1 para uma base de dados, pode ser considerada um problema de classificação [16]. Em um primeiro momento, os dados são pré-processados, tornando-se padronizados. Em seguida, a fim de reduzir a complexidade do processo de RE, tornando-o mais eficiente, os registros podem ser indexados, ou seja, aqueles registros com valores semelhantes para um ou mais atributos são agrupados para serem comparados entre si. Com a indexação, são gerados pares de registros para que, na terceira etapa, possam ser comparados usando-se funções de similaridade, as quais retornam valores que indicam o quão similar é cada par. A cada par de registros, é associado a um valor de similaridade. Esses valores de similaridade, por fim, são utilizados na etapa de classificação para decidir se o par representa uma entidade duplicada ou não [6; 23; 28].

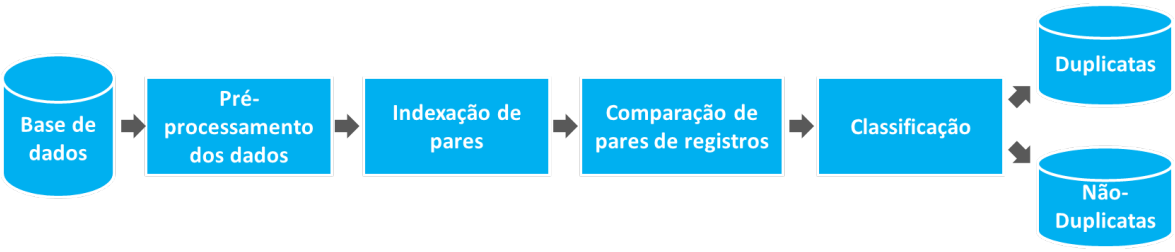


Figura 2.1: Processo de Resolução de Entidades para uma base de dados, adaptado de [11]

Em particular, na etapa de classificação de pares de registros (quanto à sua duplicidade), podem ser aplicadas abordagens: i) não-supervisionadas, as quais fazem uso de valores de similaridade previamente computados, sem que se tenha acesso a qualquer informação sobre as características dos pares de registros que dizem respeito a duplicados ou não-duplicados; e ii) supervisionadas, as quais utilizam dados de treinamento para aprender como classificar os pares de registros. Mais especificamente, dentre essas abordagens, estão aquelas que fazem uso de algoritmos de AM, os quais necessitam de um conjunto de treinamento que contenha pares de registros com seus respectivos valores de similaridade e rótulos que informem se cada par pertence à classe de registros duplicados ou não-duplicados [11; 19].

Dado o fato de as abordagens de RE envolverem sempre a classificação de pares de registros, a partir deste ponto, quando se utilizar o termo "classificador não-supervisionado"estará sendo referenciada qualquer abordagem de RE não-supervisionada. Da mesma maneira, quando for utilizado o termo "classificador supervisionado", dado o escopo deste trabalho, estará sendo referenciada qualquer abordagem supervisionada que faça uso de algoritmos de AM para classificar pares de registros.

2.2 Aprendizagem de Máquina para Resolução de Entidades

A AM estuda maneiras de o computador aprender tarefas e melhorar a execução destas tarefas a partir da experiência [36]. Algoritmos de AM basicamente fazem uso de modelos matemáticos que buscam a função ideal $f : X \rightarrow Y$ que melhor reflete o problema, onde X representa o domínio das variáveis de entrada do modelo e Y o domínio do valor a ser predito, como uma classe, por exemplo. Os algoritmos buscam a partir de um conjunto de

treinamento $T = \{(x_1, y_1), ..., (x_n, y_n)\}$ estimar uma função $g : X \rightarrow Y \mid g \approx f$.

No caso de AM aplicada ao problema de RE, o conjunto de treinamento pode, por exemplo, ser da forma de um vetor $T = \{(x_1^1, x_1^2, ..., x_1^i, y_1), ..., (x_n^1, x_n^2, ..., x_n^i, y_n)\}$. Em T, cada $x_j^i \in [0, 1]$ representa a similaridade entre os valores do i-ésimo atributo do j-ésimo par de registros contido na base de treino. Como mostrado no conjunto T da Figura 2.2, a classe a qual cada par pertence é representada por $y_j \in \{0, 1\}$, com $y_j = 1$ se o par referenciar um registro duplicado, (a_1, b_1), por exemplo; ou $y_j = 0$, caso contrário, como os pares (a_1, b_2) e (a_2, b_1) [15; 19].

T=

	X			Y
ID-Pares	Nome	Nasc	End	Classe
(a1,b1)	0.6	0.8	1.0	1
(a1,b2)	0.0	0.15	0.0	0
(a2,b1)	0.2	0.1	0.5	0
(a2,b2)	0.9	0.85	0.95	1

Figura 2.2: Exemplo de conjunto de treinamento para AM aplicada a RE.

Para que o modelo a ser treinado possa ser acurado, é importante que T possua exemplos que representem bem as características dos mais diversos conjuntos de registros abordados no problema [11]. Como os dados utilizados para treinamento nas abordagens de AM são obtidos tradicionalmente a partir de amostragem aleatória, se faz necessário que a amostra selecionada aleatoriamente seja suficientemente grande, de forma que possa incluir a maior quantidade possível de pares de registros, inclusive aqueles mais raros que possuam padrões de dados bastante distintos dos demais [29; 43].

Quando aplicados à tarefa de RE, os algoritmos de AM, ao selecionar aleatoriamente pares para compor o conjunto de treinamento, tendem a ser penalizados com amostras pouco representativas. Isso se deve ao fato de as bases de dados geralmente sofrerem com o problema de desbalanceamento de classes, em que o subconjunto de pares duplicados é geralmente muito menor que o subconjunto de pares não-duplicados [4; 25]. Frente a isso, podem ser utilizadas, por exemplo, estratégias de AA com o intuito de re-

duzir o tamanho desses conjuntos sem implicar em perda de representatividade dos mesmos.

2.3 Aprendizagem Ativa para Geração de Conjunto Treinamento Representativo

Para amenizar o problema de geração de conjuntos de treinamento a partir de bases de dados desbalanceadas, podem ser utilizados métodos como i) *Oversampling*, que reduz o número de amostras da classe majoritária em um conjunto de dados [8]; ou ii) *Undersampling*, que cria amostras sintéticas da classe minoritária a partir dos dados existentes [26]. Entretanto, tais métodos necessitam que os dados possuam os rótulos referentes às classes consideradas, inviabilizando sua aplicação para geração de gabaritos proposta neste trabalho.

Dessa forma, considerando bases de dados que não possuam qualquer rotulação prévia, podem ser utilizadas estratégias de Aprendizagem Ativa (AA) para geração de conjuntos de treinamento. Nessas estratégias, algoritmos de AM, em conjunto com um especialista humano (oráculo), buscam selecionar dados informativos para treinamento, acarretando na geração de conjuntos rotulados balanceados menores, porém suficientes para treinar algoritmos acurados [16; 43]. No âmbito de AA, são considerados dados informativos aqueles que, sem apresentar redundância entre si, são capazes de representar as características da(s) base(s) de dados [13; 44]

O processo de AA, ilustrado na Figura 2.3, consiste basicamente em treinar o modelo utilizado pelo algoritmo de AM com um pequeno conjunto de treinamento inicial T e, iterativamente, selecionar novos dados da base para que o oráculo rotule-os. A cada iteração, os dados são adicionados a T, que ajudará a melhorar o modelo treinado, tornando-o capaz, no caso do problema de RE, de classificar corretamente um maior número de novos pares de registros. O processo é repetido até que um valor relativo à eficácia do modelo seja alcançado ou um orçamento, que determina o número máximo de rotulações, seja esgotado [11; 43].

A seleção de dados é realizada a partir da aplicação de heurísticas, como AA baseada em amostra por incerteza ou consulta a um comitê de classificadores [16; 44]. O conjunto T, resultante da seleção, pode ser utilizado para treinar algoritmos de AM de forma tão acurada quanto aqueles algoritmos que fazem uso de grandes conjuntos de treinamento gerados por

amostragem aleatória [16].

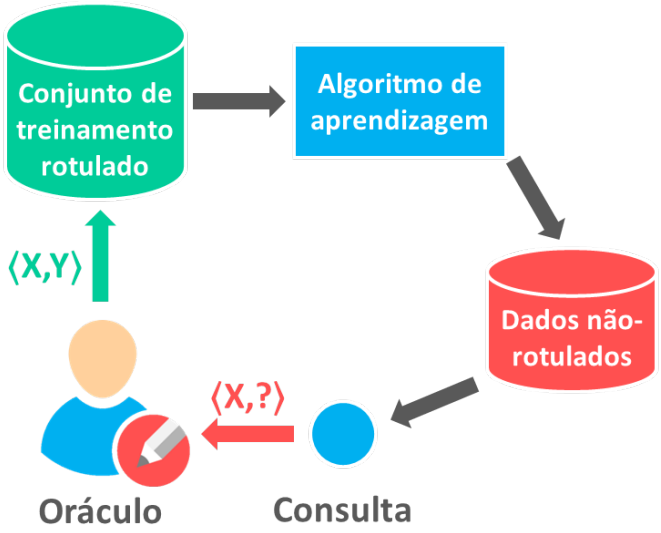


Figura 2.3: Esquema geral de funcionamento de AA, adaptado de [43].

2.4 Monotonicidade em Resolução de Entidades

A monotonicidade é uma propriedade matemática que diz respeito à relação de ordem mantida entre dois conjuntos de maneira que esta relação é estritamente não-decrescente ou estritamente não-crescente [40].

Formalmente, seja S um conjunto ordenado de valores e f uma função definida sobre S, tem-se que:

- f é monótona não-decrescente quando $\forall x, y \in S, (x > y \Rightarrow f(x) \geq f(y))$;
- f é monótona não-crescente quando $\forall x, y \in S,\ (x > y \Rightarrow f(x) \leq f(y))$.

Em RE, embora a monotonicidade não seja uma regra, ela pode ser observada em inúmeras bases de dados [16]. Por exemplo, na relação existente entre os valores de similaridade (textual) entre pares de registros $(0,0;\ 0,1;\ 0,2;\ 0,3;...;0,9;\ 1,0)$ e a disposição de pares de registros duplicados existentes na base de dados. Nesses casos, espera-se que quão maiores os valores de similaridade, maior será a probabilidade de que os pares com esses valores representem registros duplicados.

Essa relação pode ser observada, por exemplo, na Figura 2.4, onde são apresentados dois gráficos que destacam a concentração de pares de registros duplicados para duas das bases de dados utilizadas nos experimentos do Capítulo 6. Ambos os gráficos apresentam a relação entre valores de similaridade para atributos[6] das bases de dados em questão, calculados a partir da função de similaridade *Levenshtein*[7]. Os contornos auzis destacam a região de concentração de duplicatas, próximas a valores altos de similaridade, enquanto as não-duplicatas, representadas por contornos vermelhos, apresentam-se distribuídas próximas a baixos valores de similaridade.

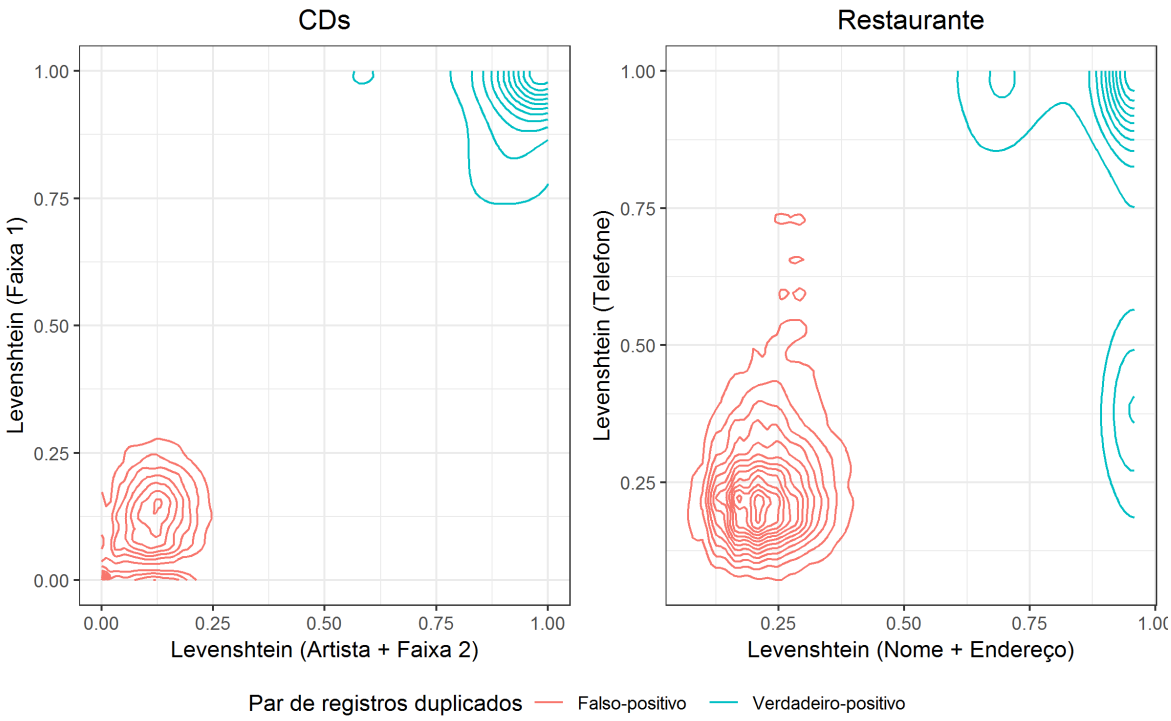


Figura 2.4: Presença de monotonicidade em bases de dados reais.

Apesar de ser uma hipótese que não se pode garantir que ocorra em todas as bases de dados, como alertado em [3] e [13], a hipótese da monotonicidade mantém-se em muitas bases de dados [1; 16]. Este fato estimulou o desenvolvimento de estratégias de AA aplicadas a RE baseadas em monotonocidade [1; 14; 16; 38; 49], assumindo que essa propriedade

[6]Para melhor compreensão do leitor, o nome dos atributos das bases de dados utilizadas, geralmente abreviados e em inglês, são apresentados nesta dissertação em português.

[7]De maneira geral, a função de similaridade *Levenshtein* calcula o custo mínimo necessário para transformar uma *string* A em uma *string* B através de operações que envolvem inserção, remoção ou substituição de caracteres [19; 37].

torna mais viável a geração de um bom conjunto de treinamento para o processo de AM no contexto de RE.

2.5 Comitê de Classificadores

Dentre as alternativas para geração do conjunto inicial de treinamento para estratégias de AA, estão aquelas que se baseiam no conceito de comitê de classificadores, cuja ideia central é que a união de classificadores distintos tende a fornecer um resultado melhor do que aqueles que seriam obtidos por um único classificador [10]. Esse conceito pode ser utilizado i) a partir da aplicação de classificadores supervisionados sobre uma base de dados, de modo que iterativamente seus resultados sejam combinados e seus conflitos rotulados por um especialista, para serem utilizados para retreinar e aprimorar os próprios classificadores (Figura 2.5) como ocorre em [5; 21; 42]; ou ii) a partir da combinação de classificadores não-supervisionados cujo resultado do conflito entre eles é utilizado como entrada para outra tarefa que não envolve o retreinamento desses classificadores (Figura 2.6) [9; 51].

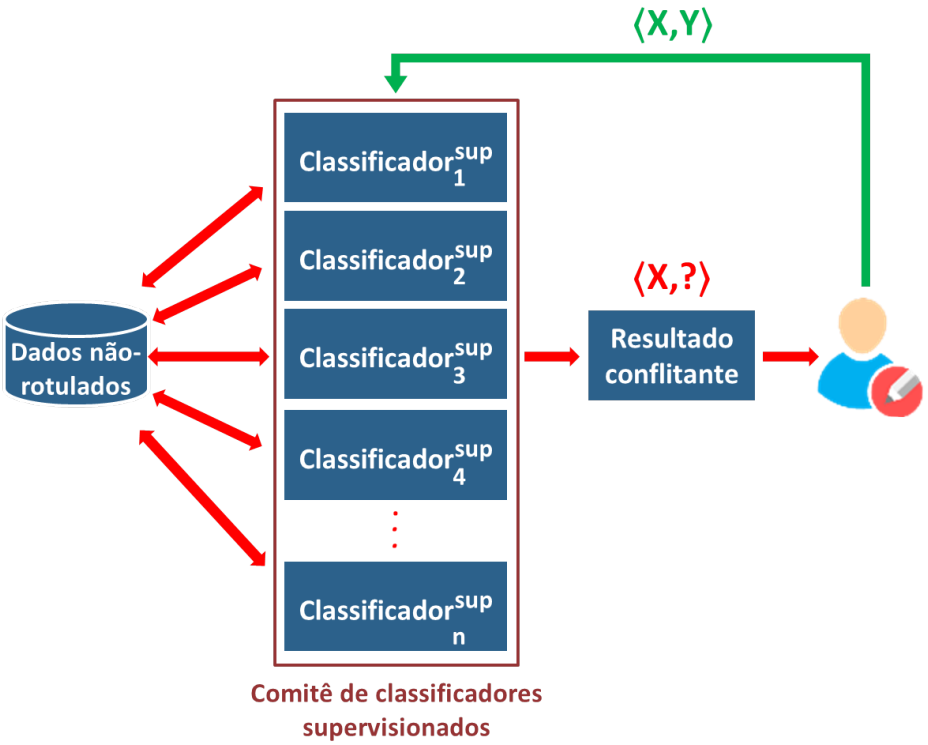


Figura 2.5: Comitê de classificadores supervisionados.

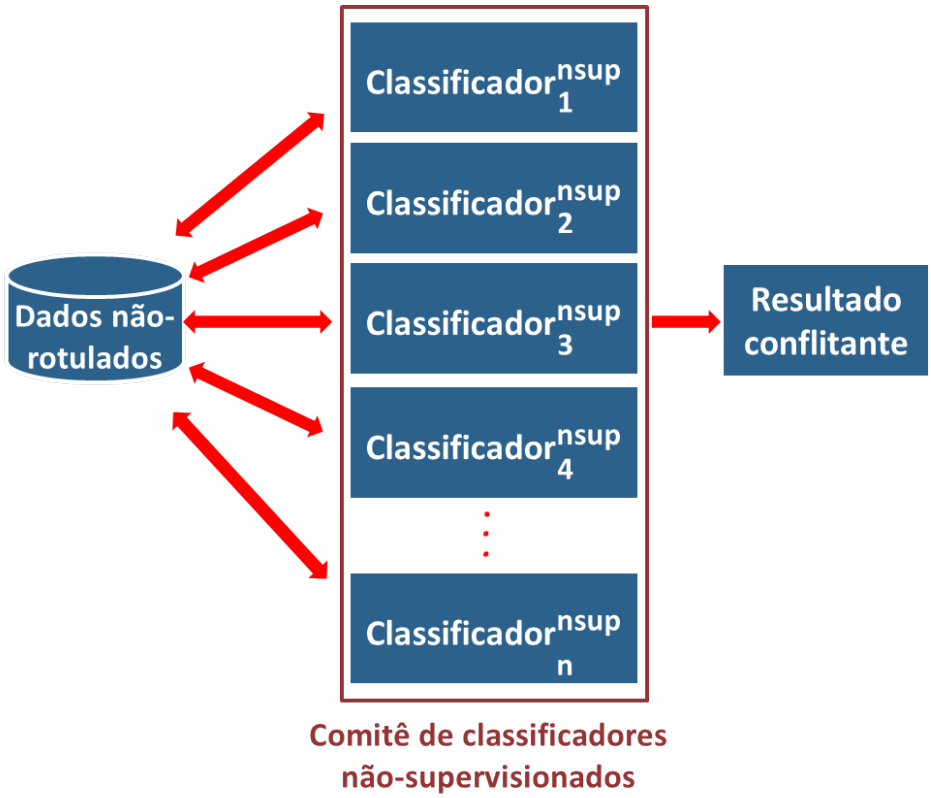


Figura 2.6: Comitê de classificadores não-supervisionados.

2.6 Considerações Finais

Neste capítulo, foi apresentada uma visão geral sobre os principais tópicos necessários à compreenssão deste trabalho. Inicialmente foram apresentadas as etapas do processo de Resolução de Entidades com destaque para os tipos de classificadores que podem ser aplicados nesse processo. Em seguida, foram discutidos os conceitos relativos à classificação supervisionada em RE a partir de algoritmos de Aprendizagem de Máquina. Ainda, foi apresentado o conceito de Aprendizagem Ativa, que pode ser aplicado para gerar conjuntos de treinamento a partir de bases de dados desbalanceadas, que não possuam qualquer rotulação prévia. Posteriormente, foi abordado o conceito de monotonicidade, característica matemática presente em diversas bases de dados que tem sido explorada para desenvolver estratégias de AA específicas para o problema de RE. Por fim, foi apresentado o conceito de Comitê de Classificadores, o qual é utilizado na abordagem para geração de gabaritos para RE proposta neste trabalho. No capítulo seguinte, são apresentados os trabalhos do estado da arte

relacionados ao problema e solução proposta apresentados neste trabalho.

Capítulo 3

Trabalhos Relacionados

Neste capítulo, são discutidos trabalhos que abordam o problema de geração de gabaritos tanto em RE, quanto em outras áreas, com destaque para aqueles que o fazem com auxílio de AM. Além disso, são apresentados trabalhos que discutem a aplicação de AA na tarefa de RE.

Os artigos selecionados foram encontrados a partir da realização de consultas aos engenhos de busca *ACM Digital Library*[8] e *IEEE Xplore Digital Library*[9], consistindo de trabalhos publicados entre 2005 e 2018, a partir da sentença de palavras-chave apresentadas na Tabela 3.1. Além disso, com base nas referências apresentadas nos artigos encontrados, foram selecionados trabalhos adicionais para compor a pesquisa.

3.1 Geração de Gabaritos para a Tarefa de RE

Uma proposta de geração automática de gabaritos para avaliar a deduplicação de registros bibliográficos da plataforma gerenciadora de referências *Mendeley*[10] é apresentada em [25]. O processo baseia-se na análise de metadados de documentos inseridos pelos usuários em suas bases de documentos. Para criar um conjunto de dados representativo, o sistema seleciona amostras de documentos de diferentes usuários, com diferentes níveis de ruído, provenientes de uma base de dados, como arXiv[11], que possuam o mesmo identificador e os

[8] `https://dl.acm.org/`

[9] `https://ieeexplore.ieee.org/Xplore/home.jsp`

[10] `https://www.mendeley.com/`

[11] `https://arxiv.org/`

Tabela 3.1: Palavras-chave utilizadas nas buscas dos trabalhos relacionados.

	Sentença de palavras-chave
1	(*"ground truth"* OR *"ground-truth"* OR *"gold standard"*) AND (*"duplicate detection"* OR *"deduplication"* OR *"record linkage"* OR *"data matching"* OR *"entity matching"* OR *"entity resolution"* OR *"merge-purge"* OR *"record matching"* OR *"information retrieval"*)
2	(*"active learning"*) AND (*"duplicate detection"* OR *"deduplication"* OR *"record linkage"* OR *"data matching"* OR *"entity matching"* OR *"entity resolution"* OR *"merge-purge"* OR *"record matching"* OR *"information retrieval"*)

consideram como duplicados desde que possuam uma porcentagem de similaridade acima de um limiar pré-estabelecido. Apesar de apresentar bons resultados, a solução não é generalizável para outros domínios, como a abordagem GTGenERAL proposta neste trabalho, por depender de elementos específicos de registros bibliográficos tratados pelo sistema legado do *Mendeley*.

Com uma abordagem independente de domínio, os autores de [51] sugerem um processo semiautomático para geração de gabaritos para deduplicação de registros chamado *Annealing Standard* (*AS*). O processo faz uso de resultados de RE gerados por n classificadores e inclui inspeção manual realizada de forma iterativa. Nessa abordagem, a intervenção humana só se faz necessária quando os pares automaticamente classificados como duplicados na iteração corrente discordam do resultado anterior. Os experimentos, realizados apenas sobre uma base de dados privada, apresentam um resultado próximo ao gabarito real em termos de qualidade, com necessidade de menos esforço humano se comparados à geração *naive* de gabaritos para a tarefa de RE.

Na geração do *AS*, por exemplo, se um par de registros é apontado por $n - 1$ classificadores como possível registro duplicado, este par é considerado conflitante devido ao fato de ter sido classificado como registro não-duplicado por um único classificador e, assim, tem

que ser inspecionado manualmente. Na abordagem proposta no presente trabalho (GTGenERAL), a qual se baseia no processo de geração do *AS*, nem todos os conflitos gerados entre os classificadores participantes do processo são inspecionados manualmente, mas sim utilizados como entrada para geração de um conjunto de treinamento para classificadores supervisionados por meio de estratégias de AA.

3.2 Geração de Gabaritos utilizados em outras Áreas da Computação

Em outras áreas da Computação, duas estratégias para geração de gabaritos são amplamente utilizadas, com o intuito principal de reduzir a intervenção humana necessária no processo: *Crowdsourcing* e Aprendizagem de Máquina.

3.2.1 Aplicação de Técnicas de *Crowdsourcing*

A tarefa de *Crowdsourcing* consiste em distribuir através de plataformas como Amazon Mechanical Turk[12] e Figure Eight[13] pequenas tarefas a indivíduos que as realizam por uma recompensa geralmente financeira. Nesse sentido, trabalhos voltados aos mais diversos objetivos como "Transcrição de imagens históricas de documentos manuscritos" [22], "Descrição de imagens disponíveis na *internet*" [33], "Transcrição de diálogos de notícias e *talk shows*" [47] e "Detecção de objetos em imagens" [48] propõem a utilização de tais mecanismos de *Crowdsourcing*, seja para gerar o gabarito a partir da simples rotulação de exemplos disponibilizados aos indivíduos ou para, a partir da rotulação de uma quantidade pré-estabelecida de exemplos, treinar modelos de AM que serão utilizados para classificação de mais exemplos que devem compor o gabarito.

Em todos esses trabalhos, além da necessidade de realização de investimento monetário para utilização da maioria das plataformas, surge a preocupação com a qualidade do resultado gerado devido aos distintos graus de conhecimento dos indivíduos e com a possibi-

[12] `https://www.mturk.com/`

[13] `https://www.figure-eight.com/`

lidade de inserção de dados ruidosos por indivíduos mal-intencionados (*spammers*) [34; 41; 46; 52]. Considerando esses possíveis problemas e, especialmente, por ser projetada para lidar com os mais diversos domínios de bases de dados, que demandam a atuação de especialistas desses domínios, a abordagem GTGenERAL não faz uso de qualquer estratégia de *Crowdsourcing*.

3.2.2 Aplicação de Técnicas de Aprendizagem de Máquina

Outra alternativa que pode ser utilizada com fins de reduzir o esforço manual na geração de gabaritos é a aplicação de algoritmos de AM, que é empregada em GTGenERAL conjuntamente a resultados de classificadores não-supervisionados.

Alinhados a esse pensamento, os autores de [24] fazem uso de um processo semiautomático para gerar gabaritos para avaliar a tarefa de análise de imagens complexas de documentos árabes. O método proposto faz uso de AM por meio da aplicação de redes neurais. Considerando a variedade entre classes de documentos, o processo fornece um modelo dedicado para cada classe de documentos, de modo que seja capaz de aprender as características de cada classe, adaptando-se às mudanças que ocorrem, recebendo como entrada as correções feitas pelos usuários.

Em [20], a estratégia de AM é aplicada com o intuito de automatizar o processo de geração de gabaritos para avaliar a tarefa de marcação de músicas disponíveis em plataformas digitais. O processo baseia-se na fusão de decisões de múltiplos anotadores automáticos a fim de compor o conjunto de treinamento para um algoritmo de regressão, alcançando aprendizagem acurada com boas taxas de predição.

3.2.2.1 Estratégias de Aprendizagem Ativa Aplicadas ao Problema de RE

Diversos trabalhos [1; 3; 4; 13; 14; 16; 38; 42; 49; 50] têm abordado a aplicação de AA para a tarefa de RE. Tais trabalhos, além de preocuparem-se em gerar o conjunto de treinamento com menor esforço, lidam com o problema de desbalanceamento de classes nas bases de dados, como discutido no Capítulo 2. Essas estratégias podem ser divididas entre aquelas que assumem a hipótese de monotonicidade (em que, quanto maior a similaridade

entre um par de registros, mais provável que ele represente uma registro duplicado) como em [1; 14; 16; 38] e aquelas que não confiam nesta propriedade, como AdInTDS*[14], proposta em [13] e utilizada nos experimentos realizados neste trabalho, e aquelas apresentadas em [3; 4; 42; 50].

Na abordagem AdInTDS, vetores de similaridade são agrupados em *clusters* de onde são extraídos exemplos para rotulação. Recursivamente, esses *clusters* são divididos em *clusters* menores e novas rotulações são realizadas, até que um grau de pureza pré-estabelecido seja atingido, em que o *cluster* contenha a maioria de vetores representando pares de registros duplicados ou a maioria de pares não-duplicados. Na abordagem GTGenERAL, para diminuir o esforço manual na geração do conjunto de treinamento dos classificadores supervisionados, a abordagem pode empregar qualquer estratégia de AA (independentemente da hipótese de monotonicidade).

3.3 Comparativo das Abordagens para Geração de Gabarito Apresentadas

A abordagem para geração de gabaritos para RE apresentada em [25], desenvolvida especificamente para ser utilizada com dados provenienetes da plataforma *Mendeley*, não é aplicável a outros domínios. Já a abordagem para geração de gabaritos para RE proposta em [51] é independente de domínio, no entanto demanda grande intervenção manual por contar apenas com a atuação de especialistas no processo. Por sua vez, buscando reduzir a intervenção humana no processo de geração de gabaritos utilizados em outras áreas da Computação, os trabalhos apresentados em [24] e [20] fazem uso de Aprendizagem de Máquina, entretanto apresentam processos específicos para os domínios em que são aplicadas.

Na Tabela 3.2, é apresentada uma visão geral das diferenças entre as abordagens para geração de gabaritos supracitadas em relação à abordagem GTGenERAL proposta neste trabalho.

Para ilustrar essas diferenças foram consideradas cinco características relacionadas ao contexto deste trabalho:

[14] *Adaptive and Interactive Training Data Set Selection for Entity Resolution*.

- **Abordagem semiautomática para geração de gabarito:** informa se a abordagem faz uso ou não de processo semiautomático para gerar gabaritos;
- **Geração de gabarito para a tarefa de RE:** indica se a abordagem é utilizada para gerar gabaritos para avaliar a tarefa de RE;
- **Independente de domínio:** informa se a abordagem pode ser aplicada a mais de um domínio de dados;
- **Utilização de Aprendizagem de Máquina:** aponta se a abordagem faz uso de técnicas de AM no processo de geração de gabaritos;
- **Utilização de Aprendizagem Ativa:** indica se a abordagem, que faz uso de técnicas de AM no processo de geração de gabaritos, aplica estratégias de AA para compor o conjunto de treinamento utilizado.

Tabela 3.2: Tabela comparativa das abordagens para geração de gabarito apresentadas.

Trabalhos	**Abordagem semi-automática para geração de gabarito**	**Geração de gabarito para a tarefa de RE**	**Independente de domínio**	**Utilização de Aprendizagem de Máquina**	**Utilização de Aprendizagem de Ativa**
[25]	N	S	N	N	N
[51]	S	S	S	N	N
[24]	S	N	N	S	N
[20]	N	N	N	S	N
GTGenERAL	**S**	**S**	**S**	**S**	**S**

3.4 Considerações Finais

Neste capítulo, foram apresentados trabalhos científicos relativos à geração de gabaritos específicos para a tarefa de RE, assim como trabalhos relativos à geração de gabaritos em outras áreas da Computação. Além disso foram discutidos trabalhos científicos que envolvem a aplicação de Aprendizagem Ativa na tarefa de RE, dos quais uma parte assume o conceito

de monotonicidade em suas estratégias, enquanto outros não consideram essa característica. No capítulo seguinte, é definido e formalizado o problema abordado neste trabalho.

Capítulo 4

Definição do Problema e Solução do Estado-da-Arte

Neste capítulo é definido e formalizado matematicamente o problema de geração de gabaritos para RE abordado neste trabalho. Além disso, é detalhada a abordagem do estado-da-arte para geração de gabaritos para RE, *Annealing Standard*, a qual é usada como base para a abordagem proposta neste trabalho.

4.1 Formalização

Dado $R = \{r_1, r_2, ..., r_n\}$ um conjunto de registros de uma ou mais bases de dados, as quais possuam erros e variações em seus dados acarretando em duplicidade de registros, tem-se que $(R \times_< R)$ é o conjunto de todos os pares (r_j, r_k) com $j < k$. A fim de identificar todos os pares de registros (r_j, r_k) duplicados em R, as abordagens de RE fazem uso de um classificador c cujo resultado é o conjunto C de todos os pares $(r_j, r_k) \in (R \times_< R)$ que foram considerados duplicados por c.

Dessa forma, seja $\mathcal{C} = \{C_1, C_2, ..., C_n\}$ o conjunto de resultados de n classificadores para um ou mais conjuntos de registros, tem-se que $\mathcal{C}^{sup} = \{C_1^{sup}, C_2^{sup}, ..., C_n^{sup}\}$ é o conjunto equivalente a $\mathcal{C}$ composto apenas por resultados de n classificadores que fazem uso de abordagens de classificação supervisionada (c^{sup}). De forma análoga, $\mathcal{C}^{nsup} = \{C_1^{nsup}, C_2^{nsup}, ..., C_m^{nsup}\}$ é o conjunto equivalente a $\mathcal{C}$ composto apenas por resultados de m classificadores que fazem uso de abordagens de classificação não-supervisionada (c^{nsup}).

Para que classificadores supervisionados ou não-supervisionados sejam avaliados, faz-se necessário que o conjunto de registros sobre o qual são aplicados possuam um gabarito, que é definido formalmente da seguinte maneira:

Definição 1. (*Gabarito para Resolução de Entidades*).

Seja R um conjunto de registros de uma ou mais bases de dados, tem-se que um gabarito GS consiste no conjunto de todos registros duplicados conhecidos em R.

Em processos de geração manual de gabaritos e em estratégias de AA aplicados a RE, pares de registros $(r_i, r_j) \in (R \times_< R)$ são apresentados a humanos especialistas no domínio de R (óraculos) para que, após análise dos valores dos atributos de cada par, seja determinado se representam registros duplicados ou não. Formalmente, um oráculo é definido como segue:

Definição 2. (*Oráculo para rotulação de pares de registros*).

Um oráculo baseado em inspeção manual ζ é uma função $\zeta : (R \times_< R) \rightarrow \{d, n\}$, onde, para cada $(r_i, r_j) \in (R \times_< R)$:

$$\zeta(r_i, r_j) = \begin{cases} d, & \text{se } (r_i, r_j) \text{ indica um par de registros duplicados,} \\ n, & \text{caso contrário.} \end{cases}$$

Associado a ζ há um orçamento $b > 0$ que indica a quantidade máxima de rotulações manuais que podem ser realizadas.

O problema abordado neste trabalho consiste em gerar um gabarito aproximado para avaliar classificadores de RE com um número reduzido de intervenções humanas realizadas durante o processo. O problema pode ser formalizado da seguinte maneira:

Definição 3. (*Geração de Gabarito Aproximado para RE*).

Sejam R um conjunto de registros de uma ou mais bases de dados, $\mathcal{C} = \{C_1, C_2, \cdots, C_n\}$ um conjunto de resultados de classificadores sobre R, $\zeta : (R \times_< R) \rightarrow \{d, n\}$ um oráculo baseado em inspeção manual, $b \in \mathbb{Z}$ um limite de orçamento de inspeções manuais e GS o gabarito de registros duplicados em R. O problema de geração de gabarito aproximado para

RE consiste em:

Gerar: um gabarito aproximado $G' \subset (R \times_< R)$ com base em $\mathcal{C}$

Maximizar: $|G' \cap GS|$

Minimizar: $(|G'| - |GS|)^2$

Sujeito a: $uso(\zeta) \leq b$

4.2 *Annealing Standard*: abordagem do estado-da-arte

O problema apresentado na Seção 4.1 é tratado no trabalho apresentado em [51], cujo gabarito resultante, denominado *Annealing Standard* (*AS*), é composto por $PA \cup PM$, onde $PA = \bigcap_{i=1}^{n} C_i^{nsup}$ é o conjunto de pares de registros rotulados automaticamente por classificadores não-supervisionados e $PM = \bigcup_{i=1}^{n} C_i^{nsup} \backslash PA$ é o conjunto de pares de registros rotulados manualmente por um especialista do domínio de R.

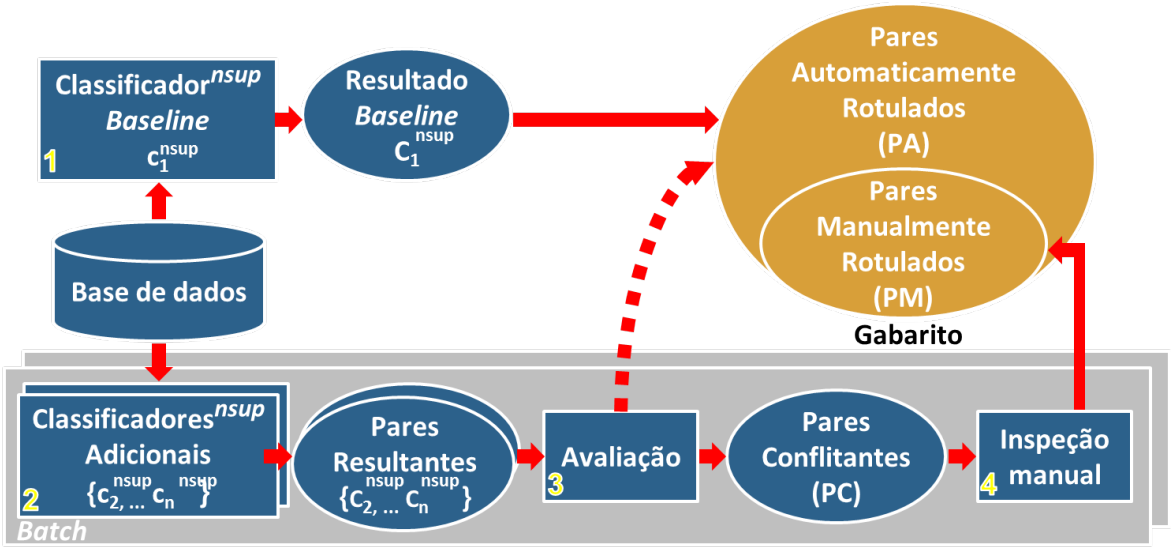


Figura 4.1: Abordagem semiautomática para geração de gabarito para RE, adaptada de [51].

Conforme ilustrado na Figura 4.1, na abordagem de geração do *AS*, um primeiro classificador não-supervisionado (c_1^{nsup}) é aplicado sobre a base de dados (passo 1) e seu resultado (conjunto de possíveis registros duplicados) é estabelecido como *baseline* formando a primeira versão do gabarito com pares classificados unicamente de maneira automática. A cada

iteração seguinte, um novo classificador não-supervisionado ($c_k^{nsup}, k > 1$) é aplicado sobre a base de dados (passo 2) e seus pares resultantes são comparados com o resultado existente no gabarito atual (passo 3).

Os pares de registros classificados pelo algoritmo de RE corrente cuja classe difere daquela apresentada no conjunto de pares previamente classificados automaticamente (PA) devem ser inspecionados manualmente (passo 4) e inseridos no conjunto de pares rotulados manualmente (PM).

Conforme mais iterações forem executadas, com mais classificadores não-supervisionados, a tendência é que o conjunto de pares manualmente rotulados cresça, permanecendo no conjunto de pares automaticamente rotulados apenas os pares classificados igualmente por todos os classificadores não-supervisionados. No Apêndice A pode ser visto um exemplo detalhado da geração do *AS*.

4.3 Considerações Finais

Neste capítulo, o problema abordado neste trabalho foi definido e formalizado matematicamente, além de ter sido detalhada a abordagem do estado da arte para geração de gabaritos para RE, *Annealing Standard*. No capítulo seguinte, são apresentadas em detalhes a abordagem para geração de gabaritos para RE e a estratégia de AA, propostas neste trabalho.

Capítulo 5

Abordagem GTGenERAL

Este capítulo apresenta GTGenERAL, uma abordagem semiautomática que combina resultados de múltiplos classificadores não-supervisionados e AA, para gerar gabaritos com esforço manual reduzido para avaliar a tarefa de RE. Em seguida, é apresentada STERSWin - ***S**election of **T**raining Dataset for **E**ntity **R**esolution with **S**liding **Win**dow*, uma estratégia de AA que, assumindo monotonicidade, busca selecionar um conjunto de treinamento balanceado e representativo.

5.1 Abordagem GTGenERAL

De forma semelhante ao *AS*, o gabarito produzido pela abordagem proposta neste trabalho é composto por um conjunto de pares de registros automaticamente rotulados (PA) e um conjunto de pares manualmente rotulados (PM), ambos gerados a partir de resultados provenientes de um comitê de classificadores não-supervisionados. A formação desses conjuntos se dá conforme os passos descritos a seguir e ilustrados na Figura 5.1, assim como no Algoritmo 1.

Inicialmente, um classificador não-supervisionado (c_1^{nsup}) é aplicado sobre a base de dados (passo 1) de maneira que os pares de registros classificados por ele como possíveis duplicados devem compor a primeira versão do gabarito (*baseline*). A partir das próximas iterações, um novo classificador não-supervisionado ($c_k^{nsup}, k > 1$) classifica os pares de registros da base de dados (passo 2) e, posteriormente, esses pares são comparados com o resultado existente na versão atual do gabarito (passo 3). Assim, o conjunto PA é povo-

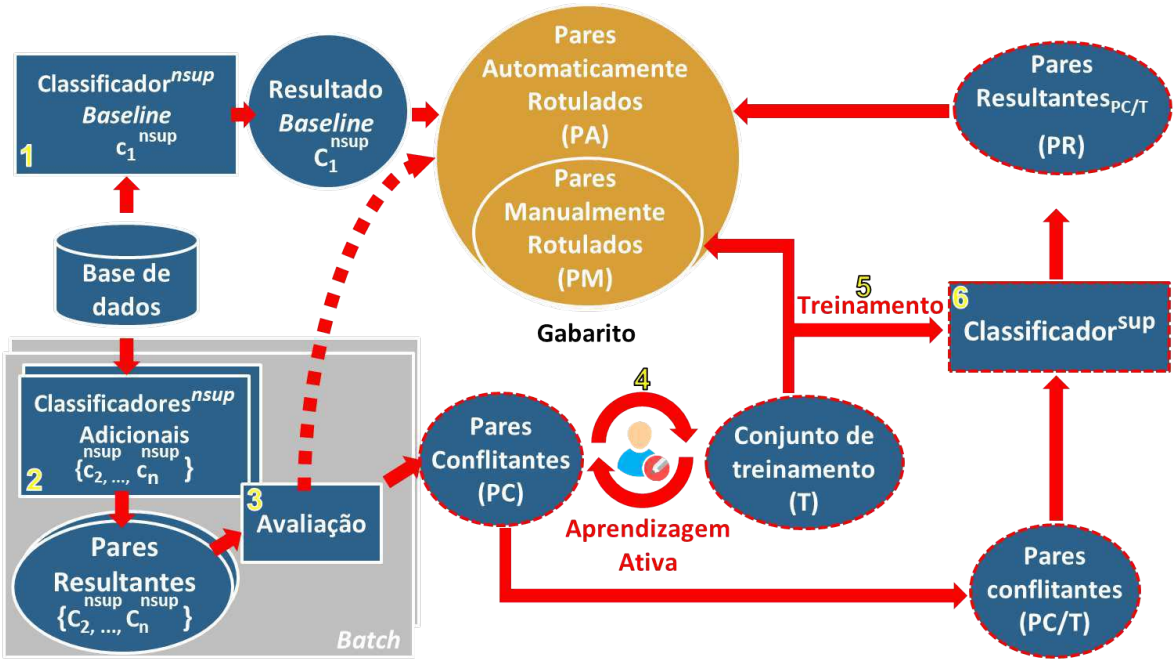


Figura 5.1: GTGenERAL: Abordagem semiautomática para geração de gabarito para RE com o uso de AA.

ado inicialmente pela intersecção dos resultados de n classificadores não-supervisionados, $PA = \bigcap_{i=1}^{n} C_i^{nsup}$.

Em seguida, é gerado um conjunto PC de pares de registros conflitantes entre os n classificadores não-supervisionados participantes do comitê, $PC = \bigcup_{i=1}^{n} C_i^{nsup} \backslash PA$. Por não terem sido classificados igualmente por todos os classificadores, supõe-se que os pares presentes em PC sejam de classificação não-óbvia. Dessa forma, espera-se que esses pares possam auxiliar na geração de um conjunto de treinamento com bom nível de informatividade, a ser utilizado nas etapas seguintes do processo.

Essa acumulação de pares conflitantes no conjunto PC busca, também, sanar um problema apresentado pelo *AS* que torna mais dispendioso o processo em termos de esforço humano: neste, uma inspeção manual tem de ser realizada sempre que um resultado conflitante surge, mesmo que gerado por apenas um classificador. Por se tratar de um processo cumulativo, a ordem com que os classificadores são aplicados sobre o conjunto de registros não influencia no número de conflitos existentes. Dessa forma, não importa se um classificador, bom ou ruim, é utilizado no início ou no fim desse processo de GTGenERAL.

Na Tabela 5.1, por exemplo, considerando um conjunto de registros R =

$\{a, b, c, d, e, f, g, h, i, j, k, l\}$, são apresentados os resultados de seis classificadores não-supervisionados com pares de registros considerados duplicados. A partir do conflito existente entre esses resultados é formado o conjunto PC, o qual encontra-se disposto na última linha da referida tabela.

Classificadores não-supervisionados
$C_1^{nsup} = \{(a, b); (c, d); (g, k); (i, j)\}$
$C_2^{nsup} = \{(a, b); (c, d); (e, f); (g, h); (i, l)\}$
$C_3^{nsup} = \{(a, b); (c, f); (d, h); (e, g); (h, j); (i, l)\}$
$C_4^{nsup} = \{(a, b); (b, i); (d, h); (e, f); (g, h)\}$
$C_5^{nsup} = \{(a, b); (d, h); (f, g); (i, l)\}$
$C_6^{nsup} = \{(a, b); (d, h); (f, i); (g, k); (h, j)\}$
Pares conflitantes
$PC = \{(b, i); (c, d); (c, f); (d, h); (e, f); (e, g); (f, g); (f, i); (g, h); (g, k); (h, j); (i, j); (i, l)\}$

Tabela 5.1: Exemplo de formação do conjunto PC.

Parte dos pares de registros constantes em PC deverão ser classificados posteriormente por classificadores supervisionados. Assim, como discutido no Capítulo 2, cada par de registros é associado a um vetor de similaridade referente à comparação dos valores de atributos dos registros de R, gerados a partir da aplicação de funções de comparação escolhidas pelo especialista no domínio de R.

Com o conjunto PC devidamente composto, é aplicada sobre ele uma estratégia de AA $f^{AA} : PC \rightarrow T$ (passo 4) que seleciona um conjunto reduzido de pares de registros a serem rotulados por um óraculo ζ e comporem o conjunto T, de modo que o número de rotulações não pode ultrapassar o orçamento pré-estabelecido (b), i.e., $T \subset PC$ e $|T| \leq b(\zeta)$. Como os pares de registros que compõem T foram rotulados manualmente, esses pares passam a compor, também, o conjunto PM de pares manualmente rotulados que fazem parte do gabarito, i.e., $PM \coloneqq T$.

No passo 5, os pares de registros de T são utilizados para treinar um classificador supervisionado (c^{sup}). Uma vez treinado, no passo 6, esse classificador será utilizado para classificar os pares do conjunto PC que não foram selecionados para T, i.e., $PC \backslash T$. Os pares duplicados classificados nesse processo automático são armazenados no conjunto PR

de pares resultantes, $c^{sup} : (PC \backslash T) \rightarrow PR$ e, por fim, são adicionados ao conjunto de pares automaticamente rotulados, $PA \coloneqq PA \cup PR$. Ao término do processo, tem-se o gabarito formado por pares de registros rotulados automática e manualmente, $GT \coloneqq PA \cup PM$.

Algoritmo 1 GTGenERAL

Input:

- $\mathcal{C}^{nsup}$: conjunto de resultados de classificadores não-supervisionados
- c^{sup}: classificador supervisionado
- f^{AA}: estratégia de AA
- ζ: oráculo humano
- b: orçamento máximo para rotulações manuais

Output: $Gabarito\ para\ RE\ (GT)$

1: $PA \coloneqq C_1^{nsup}$
2: $PC \coloneqq \emptyset$
3: **for** $i \coloneqq 2$ to $|\mathcal{C}^{nsup}|$ **do**
4: $\quad PA \coloneqq PA \cap C_i^{nsup}$
5: **end for**
6: $PC \coloneqq \bigcup_{i=1}^{n} C_i^{nsup} \backslash PA$
7: $T \coloneqq \cup_{(r_1,r_2)\in f^{AA}(PC)} \langle r_1, r_2, \zeta(r_1, r_2)\rangle$ ▷ T é preenchido até o orçamento b ser alcançado
8: $PM \coloneqq T$
9: $train(c^{sup}, T)$ ▷ c^{sup} é treinado com o conjunto de treinamento T
10: $RP \coloneqq c^{sup}(PC \backslash T)$ ▷ Os pares não rotulados pelo oráculo são classificados por c^{sup}
11: $PA \coloneqq PA \cup RP$
12: $GT \coloneqq PA \cup PM$
13: **return** GT

5.2 STERSWin: Estratégia de AA baseada em Janelas de Similaridades

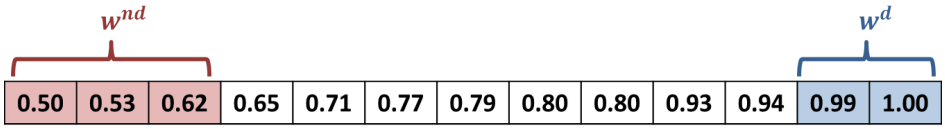
Nesta seção, é proposta uma estratégia de AA baseada em monotonicidade, a ser utilizada na abordagem GTGenERAL. A estratégia, avaliada nos experimentos do Capítulo 6, faz uso de janelas deslizantes aplicadas sobre um conjunto ordenado de pares de registros conflitantes PC, com o intuito de selecionar os melhores exemplos de pares de registros para treinamento.

Uma vez que o problema de desbalanceamento entre classes é intrínseco ao problema de RE, a estratégia de AA proposta faz uso dos conceitos de comitê de classificadores e monotonicidade, conjuntamente. A combinação desses conceitos tem por finalidade favorecer a geração de um conjunto balanceado com elementos informativos, que possa ser utilizado no treinamento de classificadores supervisionados.

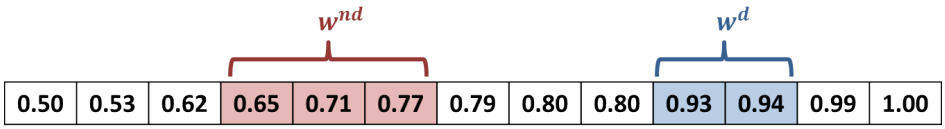
Considerando que cada par $(r_j, r_k) \in PC$ tem associado a si um valor de similaridade $sim \in [0, 1]$, pela monotonicidade, pode-se assumir que os pares com maior valor de similaridade tendem a representar registros duplicados, enquanto aqueles com menor valor de similaridade tendem a representar registros não-duplicados.

Assumindo PC ordenado crescentemente de acordo com os valores de similaridade, tem-se os pares de registros de menor similaridade próximos ao extremo inferior de PC e, de maneira análoga, próximo ao extremo superior, pares com maior similaridade. Dessa maneira, a partir de cada um dos extremos, é possível iterativamente deslizar duas janelas w^d e w^{nd} de tamanho n e m (com $n \geq 1$ e $m \geq 1$), respectivamente, em direção ao centro de PC. Dessa forma, do subconjunto de registros contido entre os limites de w^d é possível ser extraído um par de registros potencialmente duplicado, assim como de w^{nd} um potencialmente não-duplicado, ambos para serem rotulados e adicionados ao conjunto de treinamento T.

Para exemplificar, considere os seguintes valores de similaridade de pares de registros: $\{0.50, 0.53, 0.62, 0.65, 0.71, 0.77, 0.79, 0.8, 0.8, 0.93, 0.94, 0.99, 1.0\}$, contidos no conjunto de pares conflitantes PC apresentado na Seção 5.1. Assumindo w^{nd} e w^d com tamanhos 3 e 2, respectivamente, são produzidos os cenários de deslizamento das janelas em duas iterações consecutivas conforme mostrado na Figura 5.2.



(a) Iteração 1.



(b) Iteração 2.

Figura 5.2: Exemplo de deslizamento das janelas w^{nd} e w^{d} em duas iterações consecutivas.

A seleção realizada em cada janela visa buscar pares informativos, o que não se pode garantir apenas pela monotonicidade, a qual apenas sugere onde estarão concentrados pares possivelmente duplicados, assim como pares possivelmente não-duplicados. Uma maneira de buscar esses pares informativos dentro das janelas é basear-se na divergência entre os classificadores que identificaram cada registro como possível duplicado, como discutido na Seção 2.5.

Um par de registros com valor de similaridade alto, apontado por muitos classificadores como possível duplicado, pode ser considerado fácil de classificar e, assim, ser pouco informativo para treinamento. A partir desse princípio, aqueles pares indicados como duplicados por poucos classificadores, dentre vários existentes, seriam os mais apropriados para rotulação manual. De maneira análoga, um par de registros residente no extremo inferior de PC, mas apontado como possível registro duplicado por uma quantidade significativa de classificadores, em relação aos pares vizinhos, seria o mais indicado a ser investigado e rotulado.

Assim, para considerar a divergência de classificadores atrelada à escolha de um par, se faz necessário que cada $(r_j, r_k) \in PC$ tenha associado a si, também, a informação a respeito do número de classificadores que o apontaram como duplicado ($\#clasf$), de modo que sim armazene alguma estatística (e.g. média, mínimo ou máximo) referente às similaridades calculadas por cada um dos classificadores. Dessa forma, considerando o conjunto PC apresentado como exemplo na Seção 5.1, tem-se as seguintes quantidades de classificadores que apontaram cada par como possível registro duplicado: $\{1, 2, 1, 4, 2, 1, 1, 1, 2, 2, 2, 1, 3\}$.

Na Figura 5.3, é apresentada uma visão geral da estratégia de AA STERSWin[15] que, com um orçamento pré-estabelecido de rotulações, seleciona os exemplos para treinamento a partir do uso de janelas deslizantes aplicadas sobre um conjunto PC ordenado pelo valor de sim.

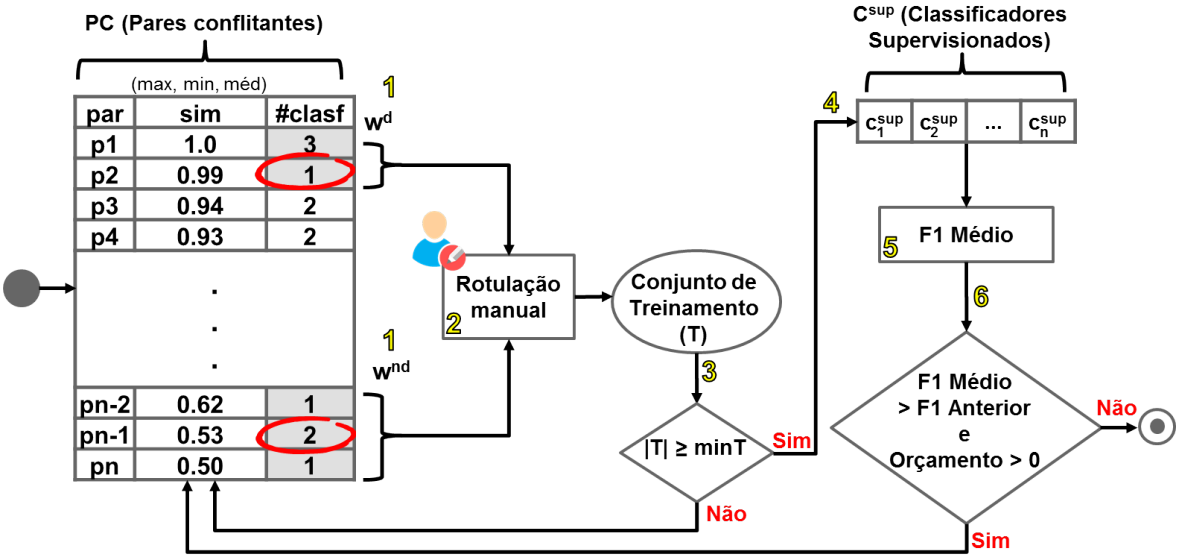


Figura 5.3: Visão geral da estratégia de AA STERSWin.

No passo 1, a partir dos pares agrupados pelas janelas w^d e w^{nd}, a estratégia, baseada em $\#clasf$, seleciona um par de registros de cada janela (a cada deslizamento) para ser rotulado no passo 2 por um humano (oráculo) como um exemplo de par duplicado ou não-duplicado. Os pares rotulados devem compor o conjunto T o qual, a partir de um tamanho pré-definido $minT$ (passo 3), servirá para treinar classificadores supervisionados existentes no conjunto C^{sup} no passo 4. Após o treinamento, cada classificador gera um valor de F1[16] e a média desses valores é calculada no passo 5. Para que a estratégia possa decidir se novos pares devem ser selecionados para compor o conjunto de treinamento, no passo 6, é verificado se há orçamento disponível para rotulação e se o valor médio de F1 calculado é maior que o valor calculado anteriormente (caso a iteração atual não seja a primeira).

Dado que STERSWin baseia-se na monotonicidade relativa aos níveis de similaridade, a presença tanto de valores de similaridade pertencentes a níveis altos, quanto pertencentes a níveis baixos é importante para tentar gerar um conjunto de treinamento balanceado.

[15]No Apêndice B, é apresentado o algoritmo STERSWin.

[16]F1 calcula a média harmônica entre *precision* e *recall*.

Nesse sentido, como o conjunto PC é composto por pares de registros apontados como duplicados por pelo menos um classificador, é natural que não ocorram valores de similaridade pertencentes a níveis mais baixos como $(0.1, 0.2, 0.3)$, por exemplo, o que pode prejudicar a seleção de possíveis exemplos de pares não-duplicados.

Entretanto, como cada par pertencente a PC possui a informação referente a quantos classificadores o apontaram como possível duplicado, pode-se a partir dessa informação definir um limiar k em que os pares apontados por uma quantidade de classificadores acima desse limiar seriam possíveis duplicados e, do contrário, possíveis não-duplicados.

Assim, com o intuito de potencializar a capacidade da estratégia de AA STERSWin em selecionar um conjunto de treinamento balanceado, o conjunto PC deve ser particionado. Desse modo, tem-se $PC^{nd} \subset PC$ (com $\#clasf <= k$), o conjunto com possível maior concentração de registros não-duplicados e $PC^{d} \subset PC$ (com $\#clasf > k$), o conjunto com possível maior concentração de registros duplicados. Dessa maneira, a cada iteração da execução de STERSWin, w^{nd} deve ser deslizada sobre PC^{nd}, enquanto sobre PC^{d} é deslizada w^{d}.

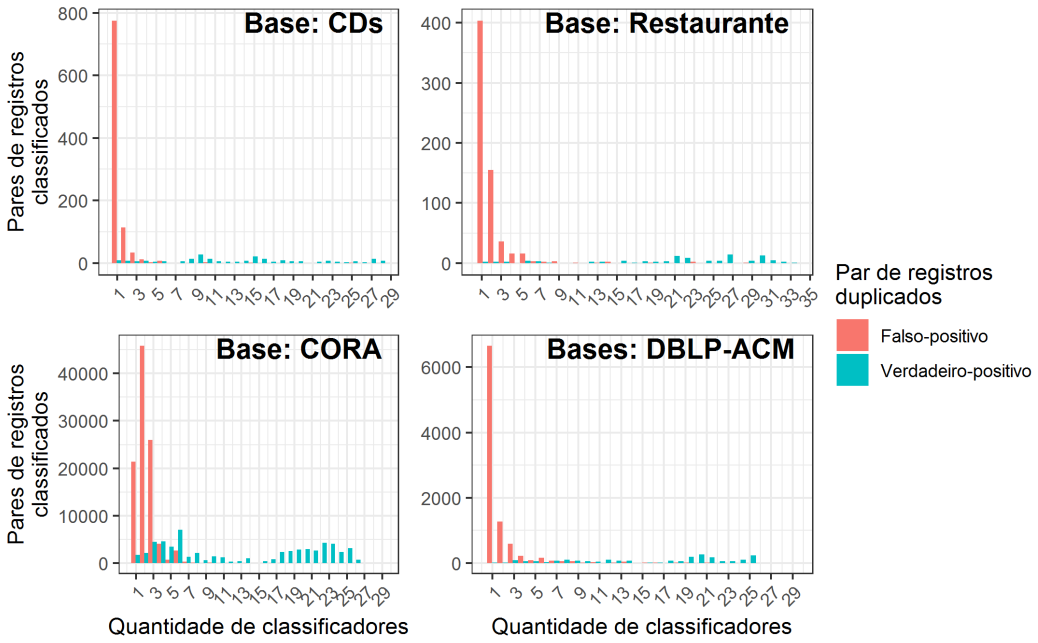


Figura 5.4: Concentração de pares de registros falso-positivos em PC.

Das bases de dados consideradas nos experimentos e apresentadas na Tabela 6.1, verificou-se empiricamente (Figura 5.4) que pares de registros pertencentes a PC aponta-

dos por apenas um ou dois classificadores como possíveis duplicados, em sua maioria são casos falso-positivos. Assim, $k = 2$ apresenta-se como um limiar aceitável para dividir o conjunto PC.

5.3 Considerações Finais

Neste capítulo, foi proposta a abordagem GTGenERAL, que combina os conceitos de comitê de classificadores e Aprendizagem Ativa para gerar gabaritos para RE. Além disso, foi introduzida a estratégia de AA STERSWin baseada em janelas deslizantes, a qual, explorando o conceito de monotonicidade relacionado ao problema de RE, busca selecionar um conjunto de treinamento balanceado e representativo. No capítulo seguinte, é apresentado um amplo estudo experimental realizado para avaliar GTGenERAL e STERSWin.

Capítulo 6

Avaliação Experimental

Neste capítulo, são apresentados os resultados experimentais do estudo. Os experimentos foram projetados com o intuito de mensurar o esforço manual empreendido para geração de gabaritos para a tarefa de RE a partir da abordagem GTGenERAL, bem como sua eficácia com respeito aos pares corretamente adicionados ao gabarito. A abordagem aqui proposta foi avaliada utilizando tanto a estratégia de AA STERSWin que assume monotonicidade, quanto a estratégia de AA AdInTDS proposta em [13]. Nos experimentos, ambas as variantes da abordagem GTGenERAL foram comparadas com a abordagem do estado da arte *Annealing Standard* (*AS*) proposta em [51].

Para avaliar a abordagem GTGenERAL, foram utilizadas bases de dados reais de múltiplos domínios, com gabarito disponível, cujos detalhes são apresentados na Tabela 6.1. Como se pode observar, diferentemente das demais, a base CORA possui um número de pares duplicados bem maior que o número de registros. Isso ocorre porque ela possui grupos de vários registros que representam uma mesma entidade do mundo real. Além disso, pode-se perceber que a porcentagem de registros duplicados presentes nas bases de dados é bastante pequena, o que dificulta a seleção de um conjunto de treinamento para algoritmos de AM.

6.1 Experimentos e Questões de Pesquisa

As avaliações experimentais apresentadas neste capítulo são divididas em dois blocos de avaliação. O primeiro bloco engloba questões de pesquisa relativas à estratégia de AA baseada em janelas deslizantes, STERSWin. O segundo bloco considera questões de pesquisa

Tabela 6.1: Características das bases de dados utilizadas nos experimentos

Base de dados	**Domínio dos dados**	**Quantidade de registros**	**Quantidade de pares de registros duplicados**	**Quantidade total de pares de registros**	**Porcentagem de pares de registros duplicados**
CDs[a]	CDs musicais	9.763	299	47.623.920	0,001%
CORA[b]	Publicações científicas	1.879	64.578	837.865	3,660%
Restaurante[c]	Restaurantes	864	112	372.816	0,030%
DBLP-ACM[d]	Publicações científicas	2.616 / 2.294	2.224	6.001.104	0,074%

[a] Disponível em `https://hpi.de/naumann/projects/data-quality-and-cleansing/annealing-standard.html`

[b] Disponível em `http://www.cs.utexas.edu/users/ml/riddle/data.html`

[c] Disponível em `http://www.cs.utexas.edu/users/ml/riddle/data.html`

[d] Disponível em `https://dbs.uni-leipzig.de/en/research/projects/object_matching/fever/benchmark_datasets_for_entity_resolution`

relativas à abordagem para geração de gabaritos para RE, GTGenERAL.

Para a avaliação da estratégia de AA STERSWin foram consideradas as seguintes questões de pesquisa:

- (**QP1**) Qual o impacto do tamanho das janelas utilizado em STERSWin na qualidade do conjunto de treinamento e na quantidade de inspeções manuais necessárias?

- (**QP2**) O ponto de partida (Extremos ou Região de Incerteza) para deslizamento das janelas influencia na qualidade do conjunto de treinamento gerado e na quantidade de inspeções manuais necessárias?

Para a avaliação da abordagem GTGenERAL, as questões de pesquisa a seguir foram observadas:

- (**QP3**) A quantidade de classificadores não-supervisionados utilizados influencia o resultado do processo de geração do gabarito, em termos de esforço manual empreendido e de eficácia?

- (**QP4**): A qualidade dos classificadores não-supervisionados utilizados influencia o resultado do processo de geração do gabarito, em termos de eficácia e esforço manual empreendido?

- (**QP5**): Qual o efeito da utilização de seleção randômica de pares, ao invés da aplicação de abordagens de Aprendizagem Ativa, para geração do conjunto de treinamento em termos de eficácia?

6.2 Métricas Utilizadas

Dado o objetivo deste trabalho, as métricas utilizadas consistem em mensurar o esforço manual empreendido na geração do gabarito e a correspondente eficácia da abordagem proposta. Como cada base de dados possui seu próprio gabarito, cada inspeção manual é simulada (e computada) a partir de cada consulta realizada ao gabarito, o que equivale a uma participação do oráculo. Em relação à eficácia, como espera-se que um gabarito possua todos os pares duplicados corretamente identificados, é utilizada a métrica F-Measure (F1), que calcula a média harmônica das métricas *Precision* e *Recall*. Os valores computados pela métrica F1 estão compreendidos entre 0,0 e 1,0, de maneira que quão mais próximo de 1,0, melhor o ajuste entre *Precision* e *Recall*.

6.3 Configuração dos Experimentos

A abordagem *baseline* para geração de gabaritos para RE, *Annealing Standard*, foi implementada em Java 8. Já a estratégia de AA STERSWin foi implementada em Python 3.7, utilizando o tamanho mínimo do conjunto de treinamento $(T) = 20$, de modo que fosse possível dividir tal conjunto em blocos menores, a fim de realizar o treinamento inicial dos classificadores supervisionados utilizados no processo. Além disso, foi aplicado o limiar para particionamento de $PC\ (k) = 2$, conforme discutido no Capítulo 5. A estratégia AdInTDS, que teve seus códigos disponibilizados pelos autores, foi utilizada com os melhores parâmetros identificados no estudo realizado em [13] (pureza mínima do *cluster* = 0,95, acurácia do oráculo = 1,0, margem de erro de amostra = 0,1, método de seleção de *cluster* = Far). Para ambas as estratégias de AA, foi estabelecido um orçamento máximo de duzentas rotulações manuais para os experimentos realizados, baseado na observação empírica da quantidade média de rotulações realizadas pela estratégia de AA do estado da arte, AdInTDS. Os classificadores supervisionados utilizados (*Decision Tree*, *Random Forest* e

SVM) [32], comumente estudados e aplicados em trabalhos que envolvem AM e RE [1; 4; 13; 50], foram empregados a partir do pacote Python scikit-learn[17].

Os conjuntos de pares de registros identificados como possíveis duplicados, utilizados como entrada para a abordagem GTGenERAL, são provenientes de classificadores não-supervisionados elaborados pelos autores deste estudo. Para cada base de dados, foram elaborados pouco mais de 30 classificadores a partir do DuDe Toolkit[18], de modo que os classificadores diferem entre si em relação às funções de distância e limiares utilizados. Conforme ilustrado na Figura 6.1, com o intuito de representar classificadores de eficácias distintas utilizados em situações reais, os classificadores foram elaborados de modo que seus resultados possuíssem valores de F1 bem distribuídos. Por exemplo, para base de dados Restaurante observa-se que existem classificadores com valores de F1 muito baixos, próximos a 0,2, medianos, entre 0,5 e 0,8, e outros com valores mais altos, orbitando em torno de 0,9.

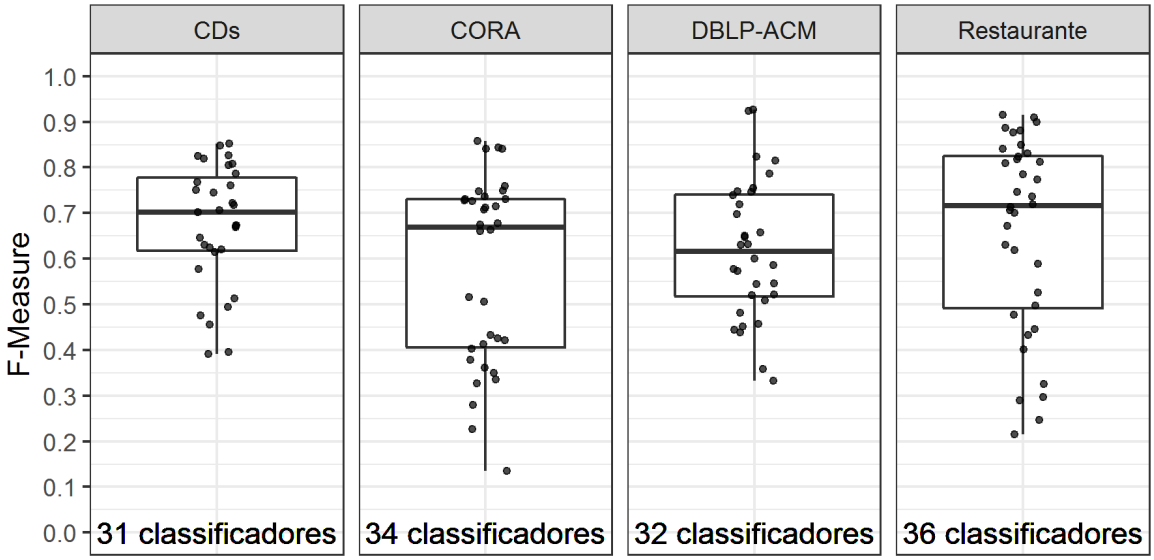


Figura 6.1: Distribuições de valores de F1 para os classificadores não-supervisionados utilizados nos experimentos.

Todos os experimentos foram realizados sobre o Sistema Operacional Ubuntu 16.04 LTS em computador com processador Intel i7 (3.60GHz) e 16GB de memória RAM. Os progra-

[17] https://scikit-learn.org/stable/

[18] https://hpi.de/naumann/projects/data-quality-and-cleansing/dude-duplicate-detection.html

mas utilizados e bases de dados utilizados estão disponíveis para acesso público[19].

Para treinar os classificadores supervisionados utilizados nos experimentos, foram realizadas intervenções sobre os dados disponibilizados como entrada para esses classificadores, a fim de melhorar o resultado por eles produzido. Tais intervenções dizem respeito a tarefas de engenharia dos dados[20], que buscam fazer com que os algoritmos de AM possam extrair melhores informações sobre os dados fornecidos [7; 27]. Neste trabalho a engenharia de dados consistiu, para cada base de dados, na criação de novos atributos a partir dos já existentes e da aplicação de pesos distintos para esses atributos.

Por exemplo, considerando a base de dados de CDs e o problema de RE abordado, é razoável considerar que o título de um CD deva possuir maior peso se comparado ao nome do artista, pelo fato de que o mesmo artista pode ter gravado inúmeros CDs, o que tende a dificultar o processo de identificação de duplicatas. Pode-se, ainda, criar um novo atributo "artista-título"que relacione um artista ao título de um CD, reduzindo a redundância dos dados e facilitando o processo de aprendizagem dos classificadores supervisionados.

Dessa forma, na Tabela 6.2 são apresentados os pesos atribuídos aos atributos e novos atributos resultantes da engenharia de dados realizada para cada base de dados utilizada nos experimentos deste trabalho.

Tabela 6.2: Engenharia de dados realizada sobre as bases de dados utilizadas.

Base de dados	**Atributos resultantes da engenharia de dados**	
CDS	**soma-pesos**: (artista*1 + título*2 + faixa1*0,8 + faixa2*0,8)/4,6	**artista-título**: (artista*título)
CORA	**soma-pesos**: (autor*0,7 + título*0,8 + jornal*0,8 + volume*0,4 + páginas*0,5 + data*0,3)/3,5	
DBLP-ACM	**soma-pesos**: (autores*0,5 + título*2 + local-do-evento*0,5 + ano*1)/4,0	
Restaurante	**soma-pesos**: (nome*1 + endereço*0,7 + cidade*0,5 + telefone*1,5 + tipo*0,5)/4,2	

6.4 Procedimentos para Análise dos Dados

Para cada experimento, foram geradas quinhentas combinações aleatórias distintas de resultados de RE, provenientes dos classificadores não-supervisionados utilizados por cada abordagem de geração de gabaritos estudada (*Annealing Standard* e GTGenERAL). A partir

[19] `https://github.com/DiegoFernandesAraujo/Master-SKYAM`

[20] Em inglês, *feature engineering.*

das combinações obtidas, aplicou-se o método de reamostragem *bootstrap* [18] com 10.000 permutações, para gerar intervalos de confiança de 95% sobre a mediana das inspeções manuais necessárias para compor os gabaritos, assim como sobre a mediana dos respectivos valores de F1 obtidos.

6.5 Estudo dos Parâmetros de STERSWin

Nos experimentos realizados nesta seção, são avaliados fatores relacionados à estratégia STERSWin, a fim de identificar aqueles mais adequados a proporcionarem a geração de conjuntos de treinamento informativos com acesso reduzido ao oráculo.

6.5.1 Tamanho das Janelas Utilizadas

A fim de identificar parâmetros adequados para aplicação da estratégia STERSWin com vistas a selecionar um conjunto de treinamento pequeno, porém suficiente para treinar bons classificadores supervisionados, buscou-se estudar a influência do tamanho da janela utilizada para agrupar pares de registros para rotulação pelo oráculo.

Assim, esta subseção tem por objetivo responder à questão de pesquisa **QP1**: *Qual o impacto do tamanho das janelas utilizado em STERSWin na qualidade do conjunto de treinamento e na quantidade de inspeções manuais necessárias?*

6.5.1.1 Desenho Experimental

O desenho deste experimento consiste em executar a estratégia de AA STERSWin com tamanhos variados de janelas e verificar a quantidade de rotulações manuais realizadas em cada combinação. Além disso, o experimento busca verificar a eficácia, em termos de F1, alcançada pelos classificadores *Random Forest* e *SVM* ao serem treinados com os conjuntos de treinamento gerados por STERSWin para os diferentes tamanhos de janela.

Dado que o conjunto PC possui tamanho variável, em função da quantidade de registros nas bases de dados e dos classificadores não-supervisionados aplicados sobre elas, a utilização de valores absolutos para o tamanho das janelas não se mostra adequada, devido aos diferentes tamanhos de conjuntos de pares conflitantes que podem haver.

Nos experimentos realizados, por exemplo, o menor conjunto de pares conflitantes (PC_1) obtido para a base de dados Restaurante continha 47 pares, enquanto o maior conjunto de pares conflitantes (PC_2) obtido para a base de dados CORA possuía 161.713 pares. Dessa forma, a utilização de uma janela de tamanho único para ambos os conjuntos percorreria uma maior região de PC_1, tendo acesso a uma maior diversidade de pares, enquanto para PC_2 apenas pares muito próximos aos extremos inferior e superior desse conjunto seriam rotulados. Assim, faz-se necessária a utilização de valores proporcionais ao tamanho de cada conjunto PC.

Para garantir que os conjuntos de treinamento gerados nos experimentos possuísem o tamanho mínimo pré-estabelecido para dar início ao treinamento dos classificadores, foram utilizados os seguintes tamanhos proporcionais de janela: 1%, 3% e 5%. Além disso, com o intuito de verificar se a comparação entre os valores de $\#clasf$ traz algum ganho ao processo de seleção de pares de registros para rotulação, foi utilizado um valor de janela constante de tamanho 1, o que obriga o oráculo a rotular todos os pares vizinhos subsequentemente até que a condição de parada (orçamento = 0 ou F1 atual < F1 anterior) seja alcançada.

6.5.1.2 Resultados

Os resultados obtidos no experimento são apresentados nos gráficos das Figuras 6.2 e 6.3. Ambos os gráficos são divididos em subgráficos que acomodam os resultados obtidos quando utilizados, respectivamente, 5, 15 e 25 classificadores não-supervisionados para gerar o conjunto PC.

No gráfico apresentado na Figura 6.2, a quantidade de inspeções manuais está disposta no eixo vertical, enquanto que o eixo horizontal representa o agrupamento dos tamanhos de janelas para cada base de dados.

O eixo horizontal do gráfico da Figura 6.3 também representa o agrupamento dos tamanhos de janelas para cada base de dados, enquanto que o eixo vertical acomoda valores de F1 no intervalo de 0,0 a 1,0. Além disso, o gráfico é dividido horizontalmente em subgráficos que acomodam resultados referentes aos classificadores *Random Forest* e *SVM*.

6.5.1.3 Discussão

Como ilustrado na Figura 6.2, a quantidade de pares rotulados tende a ser inversamente proporcional ao tamanho da janela em termos percentuais. Como exemplo, a maioria das janelas de tamanho 5% apresentaram a menor quantidade de inspeções manuais. Isso se deve ao fato de que, quanto maior a janela, menor a quantidade de blocos do conjunto PC, compreendidos pelos limites da janela, percorridos por ela, dos quais são extraídos um único par para rotulação. Em relação ao tamanho de janela 1, esse tende a apresentar uma quantidade de inspeções manuais semelhante à quantidade da janela de tamanho 1%, com algumas exceções.

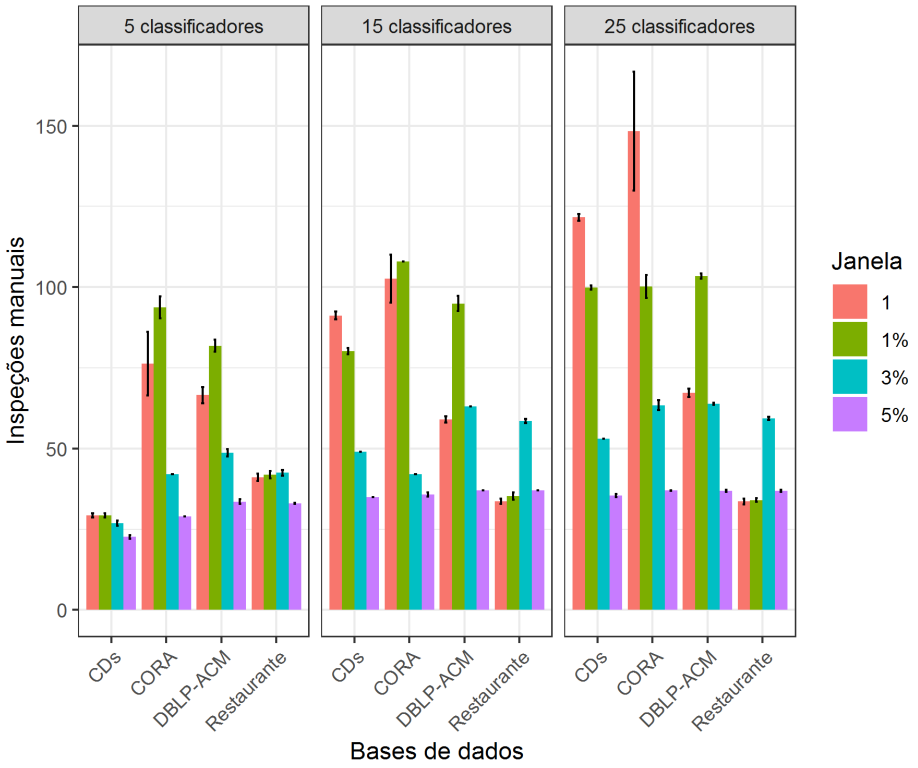


Figura 6.2: Inspeções manuais para a estratégia de AA STERSWin para diferentes tamanhos de janela.

Como explicado na Subseção 5.2, o conjunto PC é particionado a partir de um limiar k, em PC^d e PC^{nd}, para que sejam deslizadas duas janelas w^d e w^{nd}, respectivamente. Desse modo, com dois conjuntos nos quais as janelas são deslizadas, a quantidade máxima de rotulações para cada tamanho de janela é dobrada. Assim, com o tamanho 5%, por exemplo, podem ser realizadas até 40 consultas ao oráculo dado o tamanho mínimo de $T = 20$

utilizado nos experimentos.

Quanto à eficácia, observa-se na Figura 6.3 que os valores de F1 obtidos pelos classificadores treinados a partir dos conjuntos gerados pela estratégia de AA STERSWin com os tamanhos de janela 1 e 1% são bastante próximos. Para as janelas de tamanho 1%, é observada uma ligeira vantagem em alguns casos, provavelmente pelo fato de essas janelas percorrerem mais regiões do conjunto de pares conflitantes, possibilitando uma maior diversidade de pares selecionados e, por consequência, tornando o conjunto de treinamento menos redundante e mais informativo.

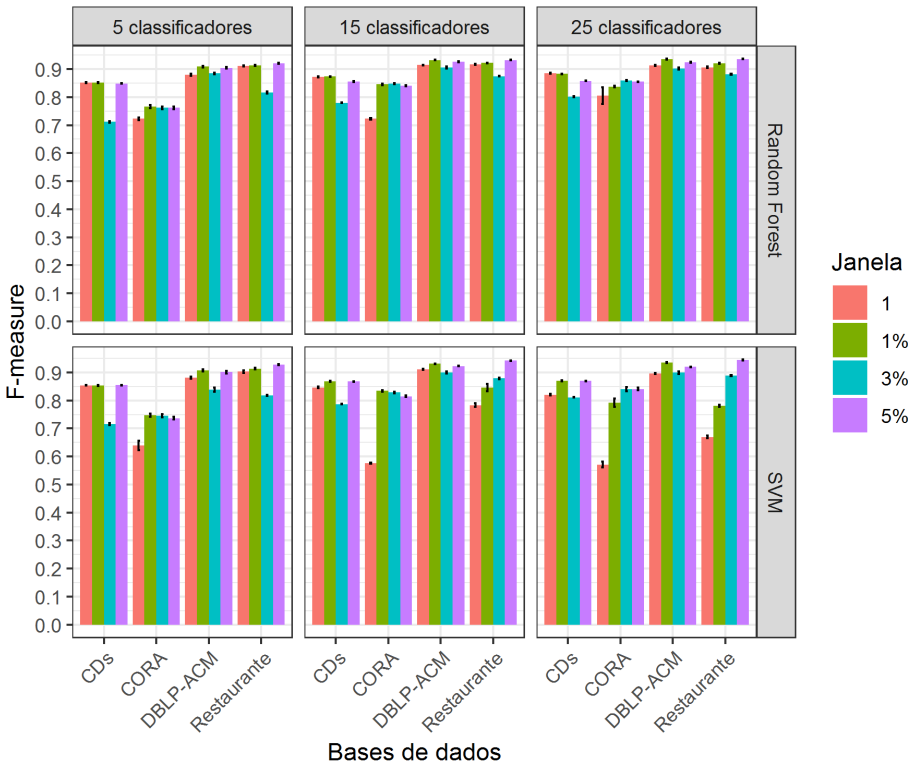


Figura 6.3: Valores de F1 para a estratégia de AA STERSWin com diferentes tamanhos de janela.

6.5.2 Ponto de Partida das Janelas

Como discutido na Subseção 5.2, apenas a monotonicidade não garante que pares informativos sejam selecionados para compor um conjunto de treinamento. Isso motivou considerar também a divergência entre os resultados dos classificadores que identificaram cada

par de registro como possível duplicado dentro de cada janela.

Ainda assim, considerando o fato de a janela w^{nd} iniciar seu deslizamento a partir do extremo inferior do conjunto de pares conflitantes PC^{nd} e da janela w^{d} partir do extremo superior do conjunto de pares conflitantes PC^{d}, é provável que os pares de registros selecionados inicialmente para rotulação sejam fáceis de classificar, o que pode influenciar no nível de informatividade do conjunto de treinamento gerado por STERSWin.

Dessa forma, buscou-se responder nesta subseção a questão de pesquisa **QP2**: *O ponto de partida para deslizamento das janelas influencia na qualidade do conjunto de treinamento gerado e na quantidade de inspeções manuais necessárias?*

6.5.2.1 Desenho Experimental

Para realizar essa avaliação, foram definidos dois pontos de partida para as janelas utilizadas: i) Extremos, em que w^{nd} e w^{d} partem, respectivamente, do extremo inferior do conjunto PC^{nd} e extremo superior do conjunto PC^{d}. Pela monotonicidade, espera-se que seja mais fácil identificar a classe a qual pertencem os pares presentes nesses extremos; e ii) Região de Incerteza, composta por pares de registros que ficam próximos ao limiar k que divide PC, onde não há fortes evidências das classes as quais esses pares pertencem, de modo que esses pares poderiam contribuir para o ganho de informatividade do conjunto de treinamento. Assim, para considerar esses pares no início do processo, w^{nd} deve deslizar a partir do extremo superior de PC^{nd} e w^{d} a partir do extremo inferior de PC^{nd}.

Com o intuito também de reforçar as evidências observadas nos experimentos discutidos na Subseção 6.5.1, a influência do ponto de partida das janelas foi avaliada considerando os tamanhos de janela que apresentaram melhores resultados na maior parte dos experimentos: 1 e 1%.

6.5.2.2 Resultados

Nas Figuras 6.4 e 6.5, são apresentados os resultados obtidos no experimento. Assim como para os experimentos apresentados na Subseção 6.5.1, os gráficos são divididos em subgráficos que acomodam os resultados obtidos quando utilizados, respectivamente, 5, 15 e 25 classificadores para gerar o conjunto PC.

No eixo horizontal do gráfico da Figura 6.4, é representado o agrupamento dos tamanhos

de janelas 1 e 1% para cada base de dados, enquanto que no eixo vertical é representada a quantidade de inspeções manuais realizadas quando aplicada a estratégia de AA STERSWin para gerar conjuntos de treinamento. Além disso, o eixo vertical é dividido em subgráficos que acomodam resultados referentes às regiões "Extremos"e "Incerteza".

Nos gráficos da Figura 6.5, o eixo vertical representa valores de F1 no intervalo de 0,0 a 1,0, referentes à classificação de pares de registros realizada com classificadores (*Random Forest* e *SVM*) treinados com conjuntos de treinamento gerados por STERSWin. No eixo horizontal, para cada base de dados, são agrupados os resultados referentes a cada região ("Extremos"e "Incerteza") utilizada por STERSWin para iniciar o deslizamento das janelas.

6.5.2.3 Discussão

Na Figura 6.4, observa-se que, de maneira geral, quando as janelas partem da Região de Incerteza, o processo exige um maior número de inspeções manuais. Isso se deve, provavelmente, à busca por mais exemplos que representem ambas as classes, o que não acontece quando o deslizamento se inicia nos Extremos, onde se pode encontrar mais facilmente exemplos distintos de cada classe. Em alguns poucos casos, pode-se observar também que o tamanho da janela 1 leva a uma quantidade de inspeções manuais mais elevada, alcançando de 25 até 100 inspeções a mais, principalmente quando se parte da Região de Incerteza.

Em relação à eficácia dos resultados, em termos de F1, considerando-se o tamanho da janela, observa-se na Figura 6.5 que janelas com tamanho 1% apresentam resultados equivalentes para ambas as regiões de onde partem. Esse comportamento não se observa quando utilizadas janelas de tamanho 1, devido ao fato de essas janelas considerarem apenas pares imediatamente vizinhos, o que tende a dificultar a seleção de pares mais distintos para compor o conjunto de treinamento, dependendo da região de onde as janelas partem (Extremos ou Incerteza). Esse comportamento pode ser potencializado no caso de conjuntos de pares conflitantes muito grandes, como aqueles provenientes da base de dados CORA com mais de 100.000 pares.

Em relação à região de partida das janelas, quando se partiu da Região de Incerteza foram obtidos valores de F1 semelhantes ou até melhores àqueles alcançados quando as janelas partiram dos Extremos. Por exemplo, como se pode ver na Figura 6.5, para a base de dados Restaurante, quando partindo-se da Região de Incerteza, os valores de F1 obtidos chegaram

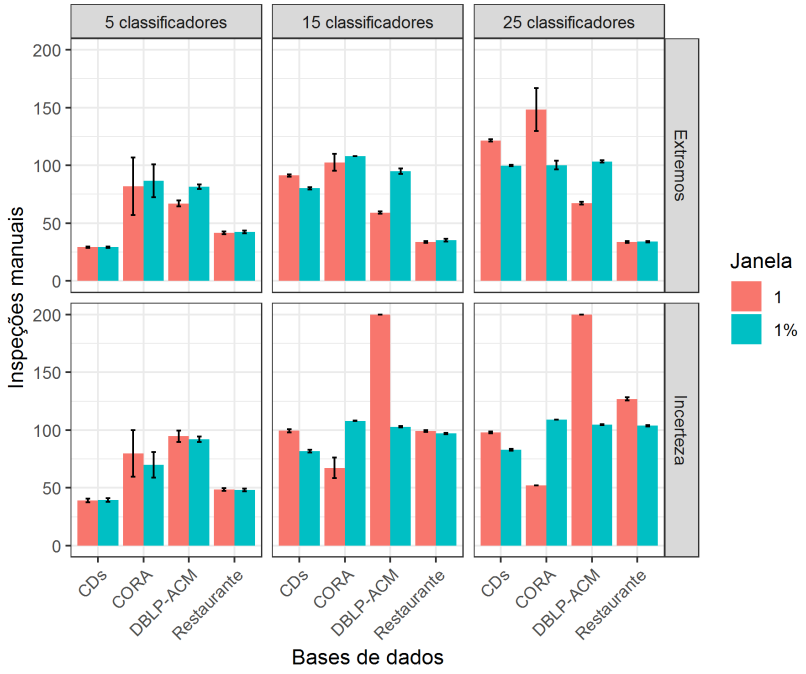


Figura 6.4: Quantidade de inspeções manuais para a estratégia de AA STERSWin considerando diferentes pontos de partida para as janelas.

bastante próximos de 1,0, considerando 15 ou 25 classificadores.

Porém, quando essa região foi considerada sobre conjuntos resultantes do conflito de cinco classificadores não-supervisionados, em vários casos, foram gerados conjuntos de treinamento altamente desbalanceados. Esses casos, que representam mais de 50% das observações, apresentaram enorme quantidade de pares de registros duplicados, contendo de 70% a 90% desses pares. Em alguns casos mais extremos, conjuntos foram gerados com a presença apenas de pares de registros duplicados. Com isso, classificadores treinados com esses conjuntos apresentaram valores de F1 muito baixos, entre 0,0 e 0,2.

Considerando o limiar $k = 2$ utilizado nos experimentos para particionar PC, torna-se difícil assumir que um par de registros (r_i, r_j) apontado por um ou dois classificadores como duplicado seja um falso positivo, quando apenas outros três dos cinco classificadores não o consideraram um par duplicado. Assim, torna-se provável que PC^{nd} possua poucos exemplos de pares não-duplicados, os quais tendem a estar concentrados em seu extremo inferior.

Dessa forma, ao iniciar o deslizamento das janelas a partir da Região de Incerteza, nesses casos, a tendência é que os pares selecionados para rotulação sejam apenas exemplos de

pares duplicados, fazendo com que a estratégia de AA STERSWin convirja mais rápido, resultando em um conjunto de treinamento povoado por um grande número de pares de uma única classe, o que leva a um conjunto pouco informativo.

6.6 Estudo da Abordagem GTGenERAL

Partindo dos resultados discutidos nos experimentos da Seção 6.5, são adotados os seguintes parâmetros para a estratégia de AA STERSWin nos demais experimentos: tamanho de janela de 1% e a Região dos Extremos como ponto de partida para o deslizamento das janelas. Além desses, são considerados os parâmetros pré-fixados no início deste capítulo (tamanho mínimo de $T = 20$ e limiar $k = 2$), quando não explicitados outros parâmetros.

Nos experimentos que seguem, são avaliados diversos fatores relacionados à abordagem GTGenERAL, a qual é testada utilizando tanto a estratégia de AA STERSWin, quanto a estratégia AdInTDS, que não assume monotonicidade. Ambos os casos são confrontados com a abordagem do estado da arte para geração de gabaritos para RE, *Annealing Standard*.

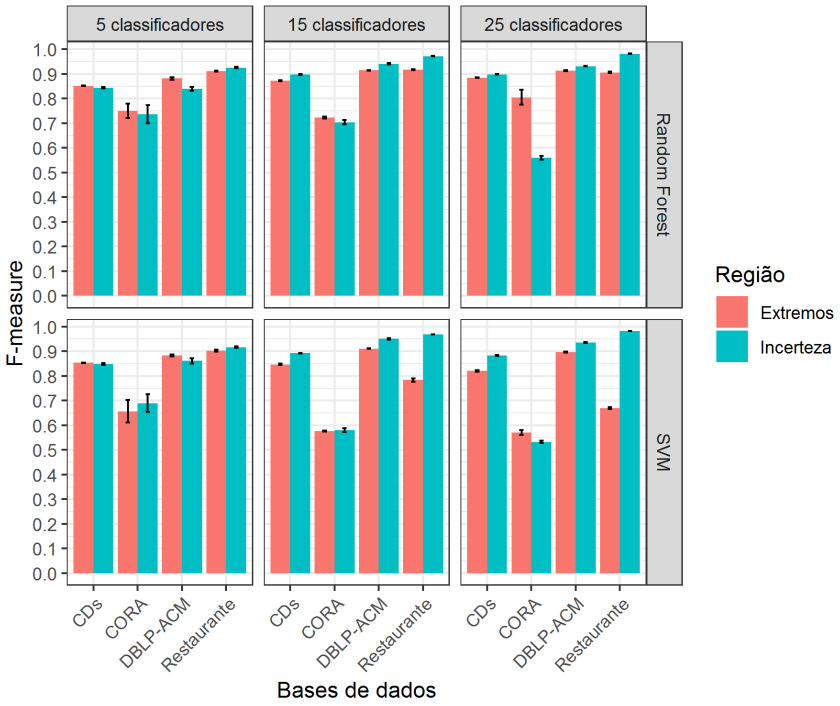
6.6.1 Quantidade de Classificadores utilizados na Geração de PC

A fim de verificar se a quantidade dos classificadores não-supervisionados utilizados para compor PC tem influência no resultado do processo de geração do gabarito, foram delimitados grupos de 5, 15 e 25 classificadores, para cada base de dados, de forma que seus resultados pudessem ser comparados.

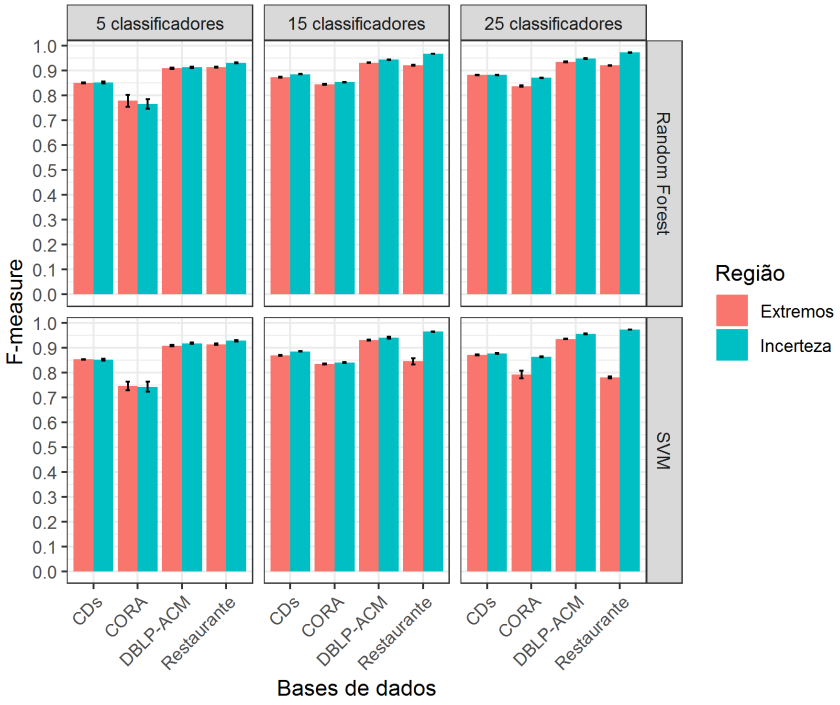
Com isso, busca-se responder nesta subseção a questão de pesquisa **QP3**: *A quantidade dos classificadores não-supervisionados influencia o resultado do processo de geração do gabarito, em termos de esforço manual empreendido e de eficácia?*

6.6.1.1 Desenho Experimental

O desenho experimental desta avaliação consiste em i) gerar conjuntos de pares conflitantes resultantes da divergência entre 5, 15 e 25 classificadores não-supervisionados e ii) utilizar esses conjuntos como entrada para as abordagens GTGenERAL (com STERSWin e



(a) Janela de tamanho 1.



(b) Janela de tamanho 1%.

Figura 6.5: Valores de F1 para a estratégia de AA STERSWin com diferentes pontos de partida para as janelas.

AdInTDS) e *Annealing Standard*, a fim de comparar seus resultados em termos de eficácia e esforço manual demandado para geração de gabaritos para RE.

6.6.1.2 Resultados

Nas Figuras 6.6 e 6.7 são apresentados os resultados obtidos no experimento. Os gráficos são divididos em subgráficos que acomodam os resultados obtidos quando utilizados, respectivamente, 5, 15 e 25 classificadores para gerar o conjunto PC, principais variáveis independentes consideradas nesta avaliação.

No gráfico da Figura 6.6, o eixo horizontal representa o agrupamento dos resultados das abordagens GTGenERAL e *Annealing Standard* para cada base de dados, enquanto que no eixo vertical é representada a quantidade de inspeções manuais realizadas durante a execução dessas abordagens. Por haver casos em que o número de inspeções manuais é bastante elevado, o eixo vertical apresenta-se em escala logarítmica.

Na Figura 6.7, o eixo horizontal representa também o agrupamento dos resultados das abordagens GTGenERAL e *Annealing Standard* para cada base de dados. O eixo vertical dos gráficos, por sua vez, representa valores de F1 no intervalo de 0,0 a 1,0, referentes aos pares de registros classificados e adicionados aos gabaritos pelas abordagens GTGenERAL e *Annealing Standard*.

Para esse experimento e os seguintes, os valores de F1 de *Annealing Standard* (que não faz uso de AM) são exibidos nos gráficos que abordam a eficácia alcançada pelos gabaritos gerados nas linhas referentes a *Random Forest* e *SVM*, apenas para fins de comparação com GTGenERAL que faz uso desses classificadores.

6.6.1.3 Discussão

Com relação à quantidade de esforço manual empreendido no processo, o número máximo de rotulações para a abordagem GTGenERAL corresponde ao orçamento estipulado para a estratégia de AA utilizada. No caso dos experimentos realizados, cujo orçamento corresponde a 200, esse valor só foi alcançado (e algumas vezes superado) quando utilizada a estratégia de AA AdInTDS, cuja heurística permite que o orçamento possa ser ultrapassado enquanto se busca compor o conjunto de treinamento ótimo.

Apesar disso, os valores de inspeção manual da abordagem GTGenERAL foram semelhantes para todas as quantidades de classificadores, como se pode observar na Figura 6.6. A respeito das rotulações para a abordagem *AS*, como são realizadas sobre todos os pares conflitantes existentes, esse número tende a aumentar conforme cresce o número de classificadores e os consequentes conflitos existentes entre seus resultados, tendo atingido nos experimentos mais de 140.000 para a base de dados CORA, quando utilizados 25 classificadores no processo.

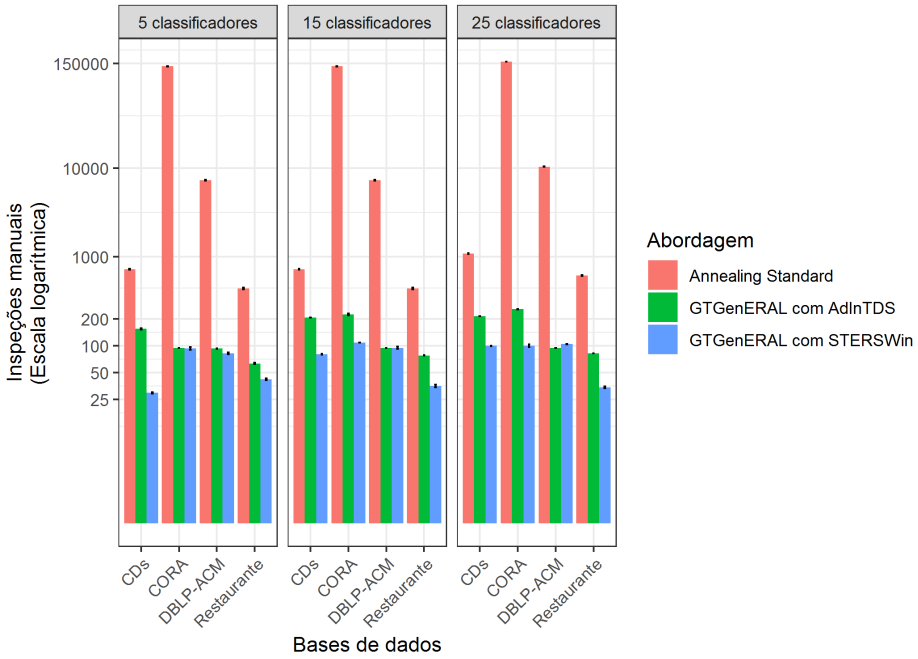
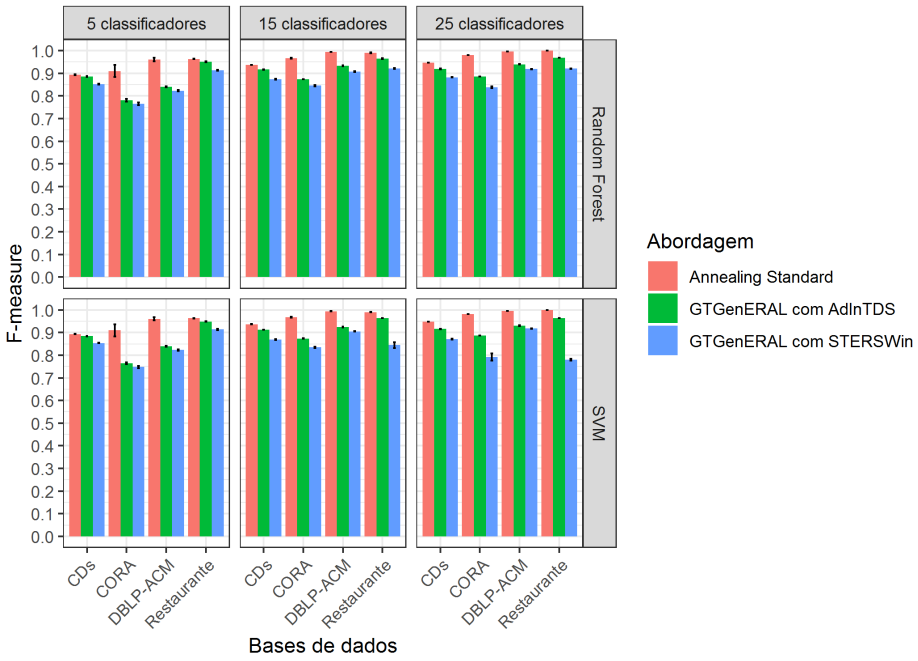


Figura 6.6: Efeito da quantidade de classificadores no esforço manual empreendido na abordagem GTGenERAL.

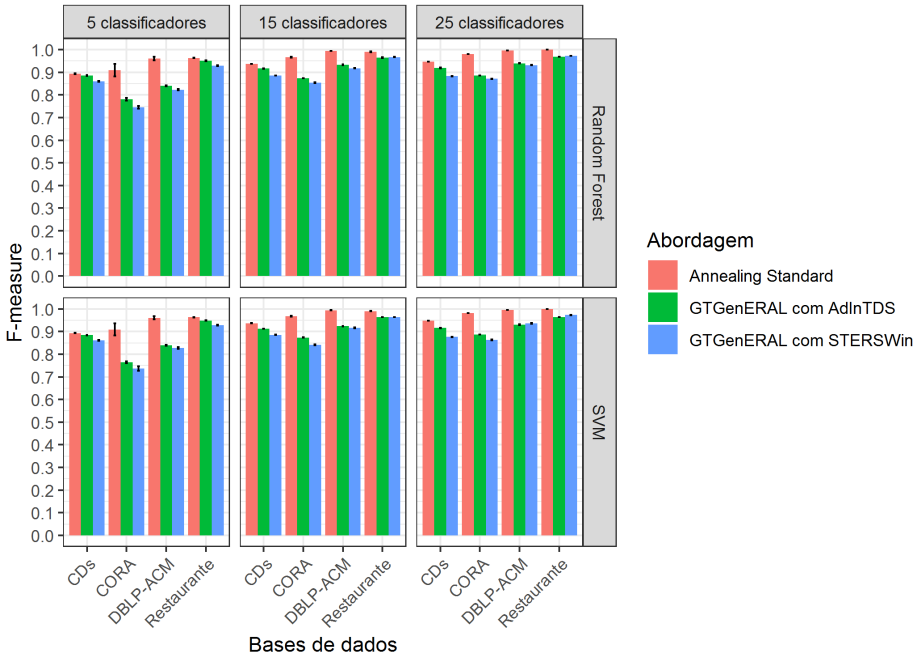
Em relação à eficácia, como se pode ver na Figura 6.7a, a abordagem *AS* sempre apresenta valor de F1 superior à abordagem GTGenERAL, devido ao fato de a rotulação dos pares no processo dessa abordagem ser totalmente realizada de forma manual por especialistas no domínio da(s) base(s) de dados. Apesar disso, mesmo com um número bastante reduzido de inspeções manuais, GTGenERAL alcança valores de F1 próximos aos de *AS*, entre 0,05 e 0,02, em especial quando é utilizada a estratégia de AA AdInTDS.

Se considerada a Região de Incerteza para a estratégia de AA STERSWin utilizada na abordagem GTGenERAL (Figura 6.7b), os resultados melhoram consideravelmente. Entretanto, como discutido no experimento da Subseção 6.5.2, nesses casos há a possibilidade de

geração de conjuntos de treinamento desbalanceados, contendo um grande número de pares de registros duplicados, quando utilizados poucos classificadores não-supervisionados.



(a) Extremos



(b) Região de Incerteza

Figura 6.7: Efeito da quantidade de classificadores na eficácia (F1) da abordagem GTGenERAL.

6.6.2 Qualidade dos Classificadores utilizados na Geração do Conjunto PC

Em todos os experimentos apresentados até então, os resultados dos classificadores não-supervisionados utilizados para compor as observações foram selecionados aleatoriamente, de modo que resultados de classificadores com alto valor de F1 podiam ser combinados com resultados de classificadores com baixo F1.

Neste experimento, buscou-se verificar se a qualidade dos classificadores utilizados, em termos de F1, tem influência sobre o gabarito resultante da abordagem GTGenERAL.

Dessa forma, esta subseção tem por objetivo responder a questão de pesquisa **QP4**: *A qualidade dos classificadores não-supervisionados utilizados influencia o resultado do processo de geração do gabarito, em termos de eficácia e esforço manual empreendido?*

6.6.2.1 Desenho Experimental

O desenho experimental desta avaliação consiste em i) gerar conjuntos de pares conflitantes resultantes da divergência entre resultados de classificadores não-supervisionados bons, assim como gerar conjuntos de pares conflitantes resultantes da divergência entre resultados de classificadores não-supervisionados ruins, e ii) utilizar esses conjuntos como entrada para as abordagens GTGenERAL (com STERSWin e AdInTDS) e *Annealing Standard*, a fim de comparar seus resultados em termos de eficácia e esforço manual demandado para geração de gabaritos para RE.

Para realizar a divisão desses grupos, podem ser considerados os valores de F1 obtidos pelos classificadores aplicados sobre cada base de dados. Entretanto, não há um valor único que possa caracterizar um classificador como bom ou ruim, para todas as bases de dados. Desta forma, a divisão dos classificadores utilizados baseou-se na mediana da distribuição de F1 apresentada na Figura 6.1 para cada base de dados. Os classificadores com F1 igual ou abaixo do valor mediano são considerados ruins, enquanto aqueles acima desse valor são tidos como classificadores bons.

6.6.2.2 Resultados

Os resultados obtidos no experimento são apresentados nas Figuras 6.8 e 6.9. Em ambas as figuras, os gráficos são divididos em subgráficos que acomodam os resultados obtidos quando utilizados, respectivamente, classificadores bons e classificadores ruins para gerar o conjunto PC.

No gráfico da Figura 6.8, o eixo vertical representa a quantidade de inspeções manuais realizadas durante a execução das abordagens GTGenERAL e *Annealing Standard*. Por haver casos em que o número de inspeções manuais é bastante elevado, o eixo vertical apresenta-se em escala logarítmica. No eixo horizontal do gráfico, é representado, para cada base de dados, o agrupamento dos resultados das abordagens GTGenERAL e *Annealing Standard*.

O eixo horizontal da Figura 6.7 representa também o agrupamento dos resultados das abordagens GTGenERAL e *Annealing Standard* para cada base de dados, enquanto que o eixo vertical representa valores de F1 no intervalo de 0,0 a 1,0, referentes aos pares de registros classificados e adicionados aos gabaritos pelas abordagens GTGenERAL e *Annealing Standard*.

6.6.2.3 Discussão

O gráfico da Figura 6.8, em escala logarítimica, mostra que o número de inspeções manuais realizadas quando são considerados apenas classificadores ruins apresenta-se maior, para ambas as abordagens (*Annealing Standard* e GTGenERAL). Isso ocorre devido a maior quantidade de pares de registros erroenamente classificados resultantes desses classificadores, o que gera uma maior quantidade de conflitos.

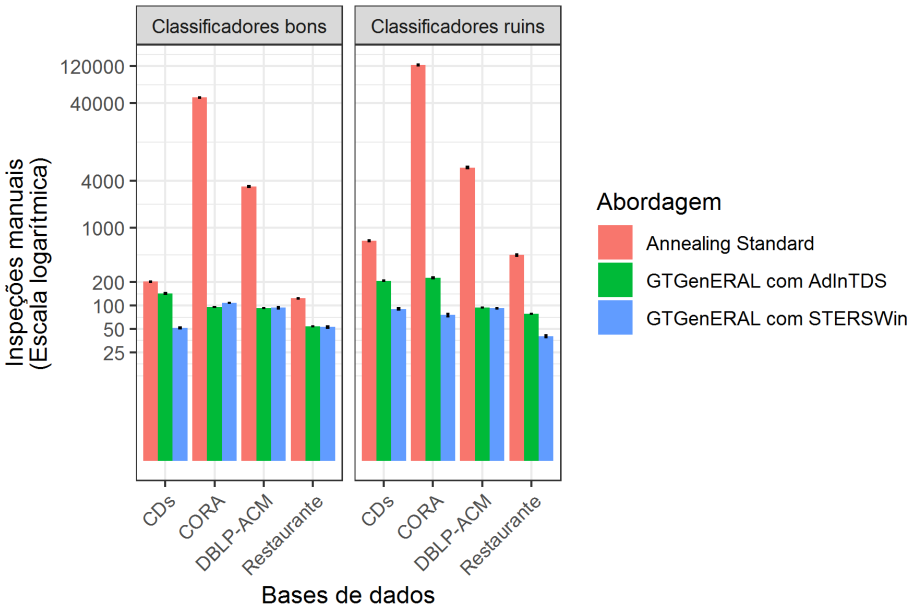


Figura 6.8: Efeito da qualidade dos classificadores no esforço manual empreendido na abordagem GTGenERAL.

Considerando o caso em que um classificador ruim possui alto valor de *Recall* $(0,8;\ \cdots;1,0)$ e baixo valor de *Precision* $(0,0;\ \cdots;2.0)$, por exemplo, muitos pares de registros tendem a ser classificados erroneamente e entrarem em conflito com os resultados de outros classificadores, o que para a abordagem AS significa um aumento significativo do número de inspeções manuais. Para a abordagem GTGenERAL, esse aumento, em alguns casos, representa o dobro das inspeções manuais quando utilizados classificadores bons, entretanto, como já destacado na Subseção 6.6.1, esse número é insignificante devido ao orçamento utilizado.

Com relação à eficácia obtida por ambos os grupos de classificadores, observa-se na Figura 6.9 que a abordagem *AS* tende a apresentar melhores valores de F1 quando utilizados apenas classificadores bons. Esse comportamento é justificado pelo fato de os conflitos entre esses classificadores apresentarem maior quantidade de pares de registros duplicados. Assim, quando rotulados com todos os demais pares, são computados como verdadeiro positivos, proporcionando a obtenção de valores de F1 mais elevados. Essa particularidade dos conjuntos de pares conflitantes gerados por classificadores bons também influencia positivamente os valores de F1 para a abordagem GTGenERAL, que em alguns casos aproxima-se bastante dos resultados obtidos por *AS*.

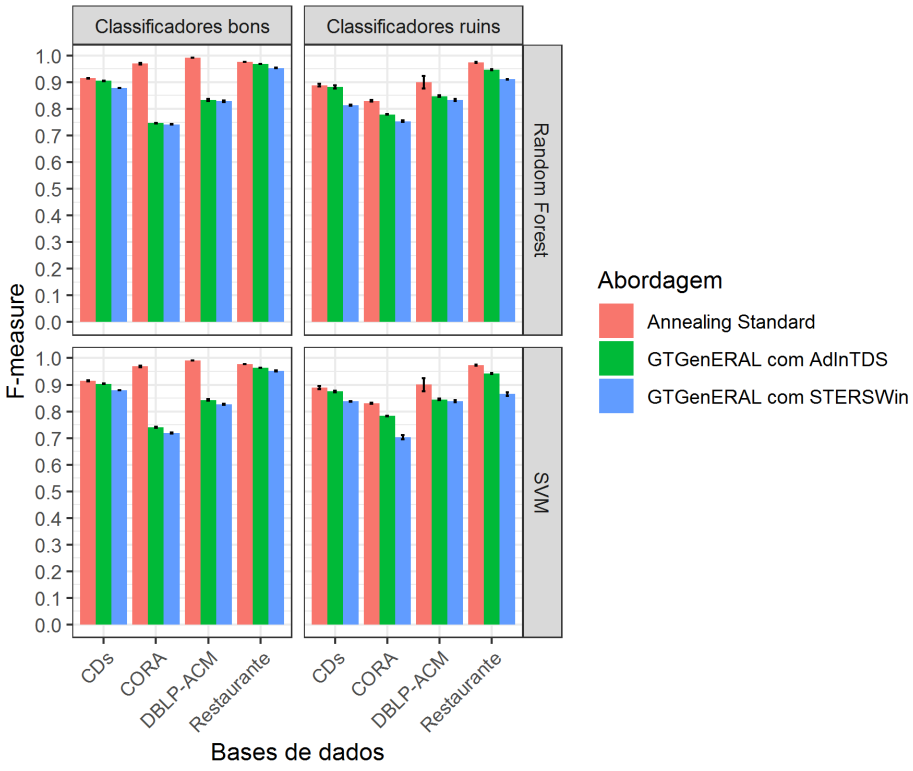


Figura 6.9: Efeito da qualidade dos classificadores na eficácia (F1) obtida com a abordagem GTGenERAL.

6.6.3 Conjunto de Treinamento: Seleção Randômica vs. Seleção com AA

Para diminuir o problema de desbalanceamento entre classes de pares de registros próprio do problema de RE, a fim de possibilitar a geração de um conjunto de treinamento informativo, a abordagem GTGenERAL explora o conflito entre resultados de classificadores, em que se espera que haja uma maior presença de pares duplicados.

Nesse experimento, buscou-se averiguar se apenas a seleção randômica de pares desse conjunto (possivelmente balanceado) seria suficiente para gerar conjuntos de treinamento bons para classificadores supervisionados.

Dessa maneira, esta subseção tem por objetivo responder a questão de pesquisa **QP5**: *Qual o efeito da seleção randômica de pares, ao invés da aplicação de abordagens de Aprendizagem Ativa, para geração do conjunto de treinamento em termos de eficácia?*

6.6.3.1 Desenho Experimental

O desenho experimental desta avaliação consiste em i) gerar conjuntos de pares conflitantes resultantes da divergência entre resultados de classificadores não-supervisionados; e ii) utilizar esses conjuntos como entrada para as abordagens *Annealing Standard* e GTGenERAL, essa última utilizando para geração de conjuntos de treinamento tanto estratégias de AA (STERSWin e AdInTDS) quanto seleção randômica. Com isso, busca-se comparar os resultados de ambas as abordagens para geração de gabarito para RE, em termos de eficácia e esforço manual demandado.

Para a estratégia randômica, foram estipulados quatro limites distintos para seleção de pares (50, 100, 150 e 200), baseados no tamanho médio dos conjuntos gerados pelas estratégias de AA nos experimentos previamente discutidos.

6.6.3.2 Resultados

Os resultados obtidos no experimento são apresentados nas Figuras 6.10 e 6.11. Os gráficos são divididos em subgráficos que acomodam os resultados obtidos quando utilizados, respectivamente, 5, 15 e 25 classificadores para gerar o conjunto PC.

No gráfico da Figura 6.10, o eixo vertical representa a quantidade de inspeções manuais realizadas durante a execução das abordagens GTGenERAL e *Annealing Standard*. O eixo vertical apresenta-se em escala logarítmica, por haver casos em que o número de inspeções manuais é bastante elevado. No eixo horizontal do gráfico, é representado, para cada base de dados, o agrupamento dos resultados das abordagens GTGenERAL, *Annealing Standard* e dos classificadores supervisionados treinados com pares selecionados randomicamente.

O eixo horizontal da Figura 6.11 representa também o agrupamento dos resultados das abordagens GTGenERAL e *Annealing Standard* para cada base de dados. O eixo vertical representa valores de F1 no intervalo de 0,0 a 1,0, referentes aos pares de registros classificados e adicionados aos gabaritos pelas abordagens GTGenERAL e *Annealing Standard*.

6.6.3.3 Discussão

A partir da Figura 6.10, verifica-se o mesmo comportamento observado nos experimentos anteriores, em que *AS* apresenta um número elevado de inspeções manuais, proveniente da

rotulação de todos os pares conflitantes gerados pelos resultados dos classificadores não-supervisionados.

Da mesma maneira, a abordagem GTGenERAL, quando utilizadas as estratégias de AA, apresenta um baixo número de inspeções manuais limitado pelo orçamento pré-estabelecido. Isso também ocorre quando utilizada a estratégia de seleção randômica, que igualmente faz uso de tamanhos pré-estabelecidos (50, 100, 150 e 200) para os conjuntos de treinamento a serem gerados.

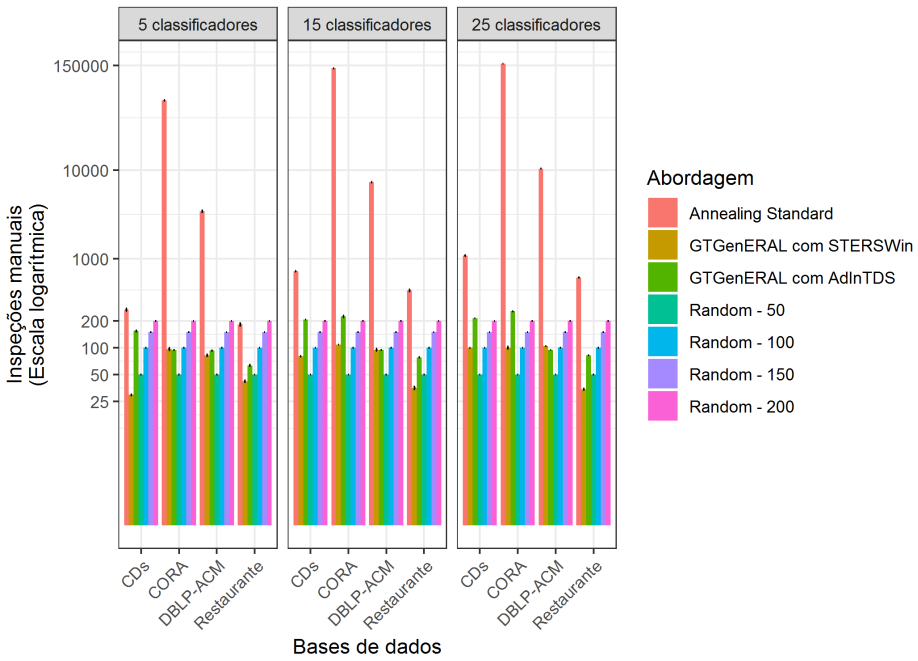


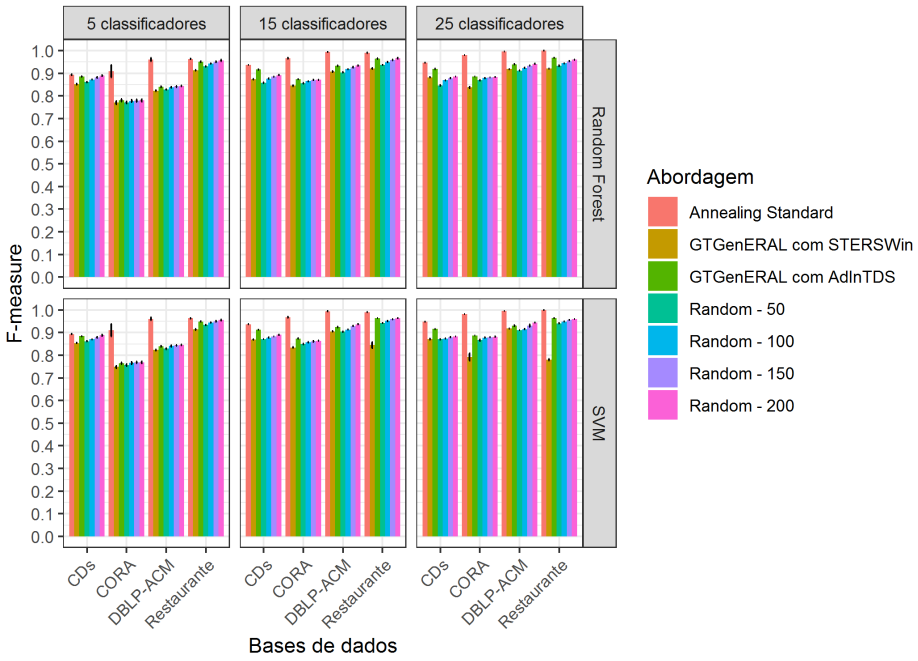
Figura 6.10: Efeito da seleção randômica de T vs. seleção através de AA no esforço manual empreendido na abordagem GTGenERAL.

Com relação aos valores de F1 obtidos, apresentados na Figura 6.11a, observa-se que os resultados da abordagem GTGenERAL quando utilizada seleção randômica são próximos aos da abordagem GTGenERAL quando aplicada a estratégia AdInTDS. Em alguns casos, esses valores chegam a se equiparar, quando os conjuntos gerados randomicamente possuem tamanhos 150 e 200.

Em muitos casos em que a estratégia de AA STERSWin é utilizada, os valores de F1 obtidos ficam aquém daqueles obtidos pelos conjuntos gerados randomicamente, como, por exemplo, quando utilizado *Random Forest* sobre os pares resultantes do conflito entre 25 classificadores para a base de dados CORA. Entretanto, quando considerada a Região de In-

certeza por STERSWin, a abordagem GTGenERAL apresenta valores próximos aos obtidos quando utilizada a estratégia AdInTDS (Figura 6.11b).

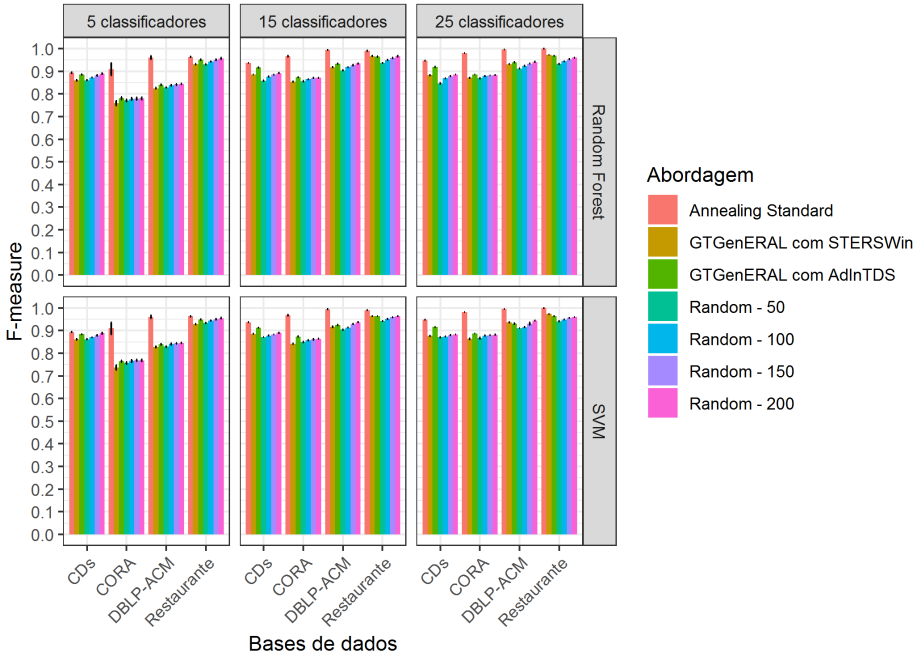
A partir dessas informações, é possível perceber que o propósito do agrupamento de conflitos entre resultados de classificadores não-supervisionados mostra-se eficaz, o que possibilita a obtenção de bons resultados de F1 provenientes da utilização de conjuntos gerados randomicamente. Ainda assim, a aplicação de estratégias de AA sobre o conjunto PC justifica-se pelo fato de ter produzido valores de F1 semelhantes utilizando em média metade da quantidade de inspeções manuais realizadas quando selecionados pares randomicamente.



(a) Extremos

6.7 Considerações Finais

Os resultados dos experimentos e avaliações apresentados neste capítulo, e sintetizados na Tabela 6.3, mostraram, a partir dos estudos gerados com base em **QP1** e **QP2**, que a estratégia de AA STERSWin, explorando a monotonicidade característica das bases de dados em que há registros duplicados, é capaz de gerar conjuntos de treinamento bons para treinar algoritmos de AM para RE. No entanto, ao considerar a Região de Incerteza, STERSWin não foi capaz de lidar com conjuntos de pares gerados pelo conflito de poucos classificadores



(b) Região de Incerteza

Figura 6.11: Efeito da seleção randômica de T no valor de F1 obtido através de GTGenERAL.

não-supervisionados.

Também foi possível observar a partir das questões de pesquisa **QP3**, **QP4** e **QP5** que a abordagem GTGenERAL, com grande redução do esforço manual, é capaz de gerar gabaritos próximos àqueles obtidos pela abordagem do estado da arte *Annealing Standard*, independentemente do tipo de estratégia de AA utilizada e sendo pouco influenciada por fatores como quantidade e qualidade de classificadores utilizados no processo.

Entretanto, GTGenERAL não se mostrou capaz de superar a abordagem *Annealing Standard* em termos de eficácia, devido ao fato de essa última realizar inspeção manual para todos os pares resultantes do conflito entre classificadores de RE.

Tabela 6.3: Síntese dos experimentos realizados

Questão de pesquisa	Conclusões obtidas
QP1: Qual o impacto do tamanho das janelas utilizado em STERSWin na qualidade do conjunto de treinamento e na quantidade de inspeções manuais necessárias?	As janelas de tamanhos 1 e 1% apresentaram melhor eficácia, demandando, para isso, um número de inspeções manuais maior que os demais tamanhos.
QP2: O ponto de partida (Extremos ou Região de Incerteza) para deslizamento das janelas influencia na qualidade do conjunto de treinamento gerado e na quantidade de inspeções manuais necessárias?	Com um número ligeiramente maior de inspeções manuais, se comparada à Região dos Extremos, a Região de Incerteza apresentou melhores resultados em relação à eficácia. Entretanto, quando considerados conjuntos de pares conflitantes resultantes do conflito entre apenas 5 classificadores, a eficácia obtida apresentou-se muito baixa.
QP3: A quantidade de classificadores não-supervisionados utilizados influencia o resultado do processo de geração do gabarito, em termos de esforço manual empreendido e de eficácia?	Observou-se que quanto maior a quantidade de classificadores não-supervisionados utilizados, maior a quantidade de inspeções manuais e a eficácia do gabarito gerado. Entretanto esses aumentos foram sutis, indicando que GTGenERAL é capaz de gerar bons gabaritos independentemente da quantidade de classificadores não-supervisionados utilizados no processo.
QP4: A qualidade dos classificadores não-supervisionados utilizados influencia o resultado do processo de geração do gabarito, em termos de eficácia e esforço manual empreendido?	Classificadores não-supervisionados bons tendem a gerar conjuntos de pares conflitantes com maior presença de pares duplicados, levando à melhora na eficácia do gabarito gerado e diminuição de inspeções manuais necessárias ao processo.

QP5: Qual o efeito da utilização de seleção randômica de pares, ao invés da aplicação de abordagens de Aprendizagem Ativa, para geração do conjunto de treinamento em termos de eficácia?	Os valores de eficácia dos gabaritos obtidos quando considerada a seleção randômica apresentaram-se semelhantes à eficácia obtida por GTGenERAL quando utilizando AA. Isso indica que o fato de agrupar os conflitos entre resultados de classificadores não-supervisionados no início do processo de GTGenERAL mostra-se eficaz. Mesmo assim, a aplicação de estratégias de AA é justificada pelo fato de produzirem valores de F1 semelhantes à estratégia de seleção randômica, utilizando em média metade da quantidade de inspeções manuais necessária a esta última.

Capítulo 7

Conclusões e Trabalhos Futuros

7.1 Conclusões

Neste trabalho, foi apresentada uma abordagem (GTGenERAL) semiautomática para geração de gabaritos para a tarefa de RE com emprego de esforço manual reduzido. A abordagem combina resultados de múltiplos classificadores e AM. Para seleção do conjunto de treinamento utilizado no processo, GTGenERAL pode fazer uso tanto de estratégias de AA que assumem monotonicidade, como aquelas que não consideram essa propriedade.

Para verificar o comportamento de GTGenERAL com ambos os tipos de estratégias de AA, além da utilização de uma estratégia do estado da arte que não assume monotonicidade, foi proposta a estratégia STERSWin, a qual é baseada em monotonicidade e faz uso de janelas deslizantes aplicadas sobre um conjunto ordenado de pares de registros conflitantes para selecionar pares de registros para treinamento.

Com o intuito de analisar o impacto da aplicação de GTGenERAL (utilizando ambas as estratégias de AA), em termos de redução de esforço manual e eficácia (F1) obtidos na geração de gabaritos para a tarefa de RE, foram realizados experimentos em bases de dados reais. Os resultados de GTGenERAL foram comparados aos resultados obtidos pela abordagem do estado da arte *AS*.

Os resultados obtidos apontam que, com grande redução do esforço manual, GTGenERAL é capaz de gerar gabaritos próximos àqueles obtidos pela abordagem do estado da arte. Entretanto, em termos de F1, GTGenERAL não se mostrou capaz de superar a abordagem *AS*, devido ao fato de essa última realizar inspeção manual para todos os pares resultantes do

conflito entre classificadores de RE.

Quanto aos fatores que podem influenciar os resultados obtidos por GTGenERAL, observou-se pouca variação na eficácia quando considerada a quantidade de classificadores utilizada, havendo sutil melhora quando empregados 15 e 25 classificadores não-supervisionados. Com respeito à qualidade dos classificadores, os melhores valores de eficácia apresentaram-se quando aplicados apenas classificadores bons. Para ambos os fatores, os melhores resultados se deram pela maior presença de pares de registros duplicados provindos dos conflitos entre os resultados dos classificadores.

Em relação à STERSWin, seus resultados apresentaram-se semelhantes aos de AdInTDS, principalmente quando considerada a Região de Incerteza para iniciar o deslizamento das janelas com tamanho 1%. Esses resultados foram possíveis devido às características distintas dos pares residentes nessa região e ao fato de a janela de tamanho 1% percorrer um maior número de blocos do conjunto de pares conflitantes, possibilitando a geração de conjuntos de treinamento mais diversificados. No entanto, ao considerar a Região de Incerteza, STERSWin não foi capaz de lidar com conjuntos de pares gerados pelo conflito de poucos classificadores não-supervisionados.

7.2 Trabalhos Futuros

Nesta seção, são apresentadas perspectivas de extensão do trabalho desenvolvido nesta dissertação, sendo essas perspectivas apresentadas a seguir:

Ampliação dos experimentos com mais bases de dados reais e outras estratégias de AA. Para que as evidências constatadas a respeito da abordagem GTGenERAL e da estratégia STERSWin, propostas neste trabalho, sejam fortalecidas, é importante a realização de experimentos adicionais com bases de dados reais de tamanhos e domínios distintos. Tais bases de dados podem ser encontradas em repositórios de acesso público como, por exemplo, o repositório RIDDLE[21] da Universidade do Texas em Austin, ou o repositório do *Database Group Leipzig*[22] da Universidade de Leipzig. Além disso, com o intuito de estudar possíveis modificações no comportamento de GTGenERAL, a utilização de outras estratégias de AA

[21] `http://www.cs.utexas.edu/users/ml/riddle/`

[22] `https://dbs.uni-leipzig.de/en/research/projects/object_matching/fever/benchmark_datasets_for_entity_resolution`

dentro do processo da abordagem para geração de gabaritos se faz necessário.

Avaliação da utilidade do gabarito gerado para a tarefa de AM

Além da aplicação dos gabaritos gerados por GTGenERAL para avaliação da tarefa de RE, pode-se verificar sua utilidade para o treinamento de algoritmos de aprendizagem de máquina para classificação. Para tanto, devem também ser aplicadas técnicas de sintetização de classes, como *Oversampling* e *Undersampling*.

Ampliação da recuperação de pares duplicados no início do processo. Tanto o *AS* quanto GTGenERAL consideram apenas pares de registros que foram identificados por pelo menos um dos classificadores não-supervisionados. Desse modo, se nenhum desses classificadores identificar um par de registro (r_x, r_y) antes da etapa de acumulação de pares em GTGenERAL, tal par não poderá ser classificado nas etapas seguintes como duplicado ou não-duplicado. Assim, mesmo na melhor hipótese, em que GTGenERAL classifica corretamente todos os pares do conjunto PC, o gabarito resultante estaria incompleto. Com isso, podem ser estudadas formas eficazes de maximizar a quantidade de pares possivelmente duplicados acumulados no início da abordagem de GTGenERAL.

Modificação da estrutura de GTGenERAL com adição de etapas de AA. A inserção de estratégias de AA em outras partes do processo de GTGenERAL apresenta-se como uma possibilidade de alavancar a acurácia do gabarito gerado, a partir de um leve aumento da quantidade de intervenção humana demandada. Esse aumento na acurácia do gabarito é esperado pelos seguintes fatores: i) os novos pares rotulados devem incrementar o conjunto de treinamento final e, consequentemente, melhorar o classificador supervisionado treinado; e ii) os pares verdadeiro positivos, rotulados manualmente pelo especialista no domínio da(s) base(s) de dados, são adicionados diretamente ao gabarito, enquanto que o aumento dessa intervenção manual deve ser pequeno, dadas as características das estratégias de AA.

Utilização de comitê de classificadores supervisionados no processo de GTGenERAL. Outro direcionamento possível para este trabalho consiste em estudar maneiras adequadas de combinar os resultados de vários classificadores supervisionados e agregá-los ao processo para compor o gabarito. Por exemplo, poderia ser estudada a influência da utilização de um conjunto de classificadores supervisionados e a classificação dos pares remanescentes do conjunto PC apontada pela maioria desses classificadores na etapa 6 de GTGenERAL (Figura 5.1).

Bibliografia

[1] Arvind Arasu, Michaela Götz, and Raghav Kaushik. On active learning of record matching packages. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, pages 783–794. ACM, 2010.

[2] Carlo Batini, Monica Scannapieco, et al. *Data and information quality: dimensions, principles and techniques*. Springer, 2016.

[3] Kedar Bellare, Suresh Iyengar, Aditya G Parameswaran, and Vibhor Rastogi. Active sampling for entity matching. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1131–1139. ACM, 2012.

[4] Mikhail Bilenko and Raymond J Mooney. Adaptive duplicate detection using learnable string similarity measures. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 39–48. ACM, 2003.

[5] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[6] David Guy Brizan and Abdullah Uz Tansel. A. survey of entity resolution and record linkage methodologies. *Communications of the IIMA*, 6(3):5, 2006.

[7] Girish Chandrashekar and Ferat Sahin. A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16–28, 2014.

[8] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[9] Qingxuan Chen, Dequan Zheng, Tiejun Zhao, and Sheng Li. A fusion of multiple classifiers approach based on reliability function for text categorization. In *Fuzzy Systems and Knowledge Discovery, 2008. FSKD'08. Fifth International Conference on*, volume 2, pages 338–342. IEEE, 2008.

[10] Zhaoqi Chen, Dmitri V Kalashnikov, and Sharad Mehrotra. Exploiting context analysis for combining multiple entity resolution systems. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, pages 207–218. ACM, 2009.

[11] Peter Christen. *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*. Springer Science & Business Media, 2012.

[12] Peter Christen and Karl Goiser. Quality and complexity measures for data linkage and deduplication. In *Quality Measures in Data Mining*, pages 127–151. Springer, 2007.

[13] Peter Christen, Dinusha Vatsalan, and Qing Wang. Efficient entity resolution with adaptive and interactive training data selection. In *Data Mining (ICDM), 2015 IEEE International Conference on*, pages 727–732. IEEE, 2015.

[14] Guilherme Dal Bianco, Renata Galante, Marcos André Gonçalves, Sergio Canuto, and Carlos A Heuser. A practical and effective sampling selection strategy for large scale deduplication. *IEEE Transactions on Knowledge and Data Engineering*, 27(9):2305–2319, 2015.

[15] Junio De Freitas, Gisele L Pappa, Altigran S da Silva, Marcos A Gonc, Edleno Moura, Adriano Veloso, Alberto HF Laender, Moisés G de Carvalho, et al. Active learning genetic programming for record deduplication. In *Evolutionary Computation (CEC), 2010 IEEE Congress on*, pages 1–8. IEEE, 2010.

[16] Chenxiao Dou, Daniel Sun, Guoqiang Li, and Raymond K Wong. Active learning with density-initialized decision tree for record matching. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*, page 14. ACM, 2017.

[17] Uwe Draisbach and Felix Naumann. Dude: The duplicate detection toolkit. In *Proceedings of the International Workshop on Quality in Databases (QDB)*, 2010.

[18] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

[19] Ahmed K Elmagarmid, Panagiotis G Ipeirotis, and Vassilios S Verykios. Duplicate record detection: A survey. *IEEE Transactions on knowledge and data engineering*, 19(1):1–16, 2007.

[20] Rémi Foucard, Slim Essid, Mathieu Lagrange, and Gaël Richard. A regressive boosting approach to automatic audio tagging based on soft annotator fusion. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 73–76. IEEE, 2012.

[21] Yoav Freund, Robert E Schapire, et al. Experiments with a new boosting algorithm. In *Icml*, volume 96, pages 148–156. Citeseer, 1996.

[22] Basilis Gatos, Georgios Louloudis, Tim Causer, Kris Grint, Veronica Romero, Joan Andreu Sánchez, Alejandro H Toselli, and Enrique Vidal. Ground-truth production in the transcriptorium project. In *Document Analysis Systems (DAS), 2014 11th IAPR International Workshop on*, pages 237–241. IEEE, 2014.

[23] Lifang Gu and Rohan Baxter. Decision models for record linkage. In *Data Mining*, pages 146–160. Springer, 2006.

[24] Karim Hadjar and Rolf Ingold. Minimizing user annotations in the generation of layout ground-truthed data. In *2011 International Conference on Document Analysis and Recognition*, pages 703–707. IEEE, 2011.

[25] James A Hammerton, Michael Granitzer, Dan Harvey, Maya Hristakeva, and Kris Jack. On generating large-scale ground truth datasets for the deduplication of bibliographic records. In *Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics*, page 18. ACM, 2012.

[26] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing*, pages 878–887. Springer, 2005.

[27] David J Hand. Data mining. *Encyclopedia of Environmetrics*, 2, 2006.

[28] Thomas N Herzog, Fritz J Scheuren, and William E Winkler. *Data quality and record linkage techniques*. Springer Science & Business Media, 2007.

[29] Robert Isele and Christian Bizer. Active learning of expressive linkage rules using genetic programming. *Web Semantics: Science, Services and Agents on the World Wide Web*, 23:2–15, 2013.

[30] Kunal Jha, Michael Röder, and Axel-Cyrille Ngonga Ngomo. All that glitters is not gold–rule-based curation of reference datasets for named entity recognition and entity linking. In *European Semantic Web Conference*, pages 305–320. Springer, 2017.

[31] Hanna Köpcke and Erhard Rahm. Frameworks for entity matching: A comparison. *Data & Knowledge Engineering*, 69(2):197–210, 2010.

[32] Sotiris B Kotsiantis, I Zaharakis, and P Pintelas. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160:3–24, 2007.

[33] Martha Larson, Christoph Kofler, and Alan Hanjalic. Reading between the tags to predict real-world size-class for visually depicted objects in images. In *Proceedings of the 19th ACM international conference on Multimedia*, pages 273–282. ACM, 2011.

[34] Xiang Liu, Liyun Li, and Nasir Memon. A lightweight combinatorial approach for inferring the ground truth from multiple annotators. In *International Workshop on Machine Learning and Data Mining in Pattern Recognition*, pages 616–628. Springer, 2013.

[35] David Menestrina, Steven Euijong Whang, and Hector Garcia-Molina. Evaluating entity resolution results. *Proceedings of the VLDB Endowment*, 3(1-2):208–219, 2010.

[36] Tom M Mitchell. Machine learning. 1997. *Burr Ridge, IL: McGraw Hill*, 45(37):870–877, 1997.

[37] Felix Naumann and Melanie Herschel. An introduction to duplicate detection. *Synthesis Lectures on Data Management*, 2(1):1–87, 2010.

[38] Kun Qian, Lucian Popa, and Prithviraj Sen. Active learning for large-scale entity resolution. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1379–1388. ACM, 2017.

[39] Dietrich Rebholz-Schuhmann, AJ Yepes, Erik M van Mulligen, Ning Kang, Jan Kors, David Milward, Peter Corbett, and U Hahn. The calbc silver standard corpusharmonizing multiple semantic annotations in a large biomedical corpus. In *Proceedings of the Third International Symposium on Languages in Biology and Medicine; Jeju Island, South Korea*, pages 64–72, 2009.

[40] Halsey Lawrence Royden and Patrick Fitzpatrick. *Real analysis*, volume 32. Macmillan New York, 1988.

[41] Parnia Samimi, Sri Devi Ravana, and Yun Sing Koh. Effect of verbal comprehension skill and self-reported features on reliability of crowdsourced relevance judgments. *Computers in Human Behavior*, 64:793–804, 2016.

[42] Sunita Sarawagi and Anuradha Bhamidipaty. Interactive deduplication using active learning. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 269–278. ACM, 2002.

[43] Burr Settles. Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(1):1–114, 2012.

[44] Rodrigo M Silva, Marcos A Gonçalves, and Adriano Veloso. A two-stage active learning method for learning to rank. *Journal of the Association for Information Science and Technology*, 65(1):109–128, 2014.

[45] Michael R Smith, Tony Martinez, and Christophe Giraud-Carrier. An instance level analysis of data complexity. *Machine learning*, 95(2):225–256, 2014.

[46] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing*, pages 254–263. Association for Computational Linguistics, 2008.

[47] Rachele Sprugnoli, Giovanni Moretti, Luisa Bentivogli, and Diego Giuliani. Creating a ground truth multilingual dataset of news and talk show transcriptions through crowdsourcing. *Language Resources and Evaluation*, 51(2):283–317, 2017.

[48] Fabio Sulser, Ivan Giangreco, and Heiko Schuldt. Crowd-based semantic event detection and video annotation for sports videos. In *Proceedings of the 2014 International ACM Workshop on Crowdsourcing for Multimedia*, pages 63–68. ACM, 2014.

[49] Yufei Tao. Entity matching with active monotone classification. In *Proceedings of the 35th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, pages 49–62. ACM, 2018.

[50] Sheila Tejada, Craig A Knoblock, and Steven Minton. Learning domain-independent string transformation weights for high accuracy object identification. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 350–359. ACM, 2002.

[51] Tobias Vogel, Arvid Heise, Uwe Draisbach, Dustin Lange, and Felix Naumann. Reach for gold: An annealing standard to evaluate duplicate detection results. *Journal of Data and Information Quality (JDIQ)*, 5(1-2):5, 2014.

[52] Jeroen BP Vuurens and Arjen P De Vries. Obtaining high-quality relevance judgments using crowdsourcing. *IEEE Internet Computing*, 16(5):20–27, 2012.

Apêndice A

Annealing Standard

A abordagem para geração de gabaritos para RE proposta por Vogel et al. [51] se vale da aplicação conjunta de classificadores não-supervisionados com inspeções manuais. A seguir, é descrito um breve exemplo da execução da abordagem.

Considere um conjunto de registros $R = \{a, b, c, d, e, f, g, h\}$ que, pelo Produto cartesiano sobre todos os elementos gera 28 pares, e o resultado de dois algoritmos de RE que declararam como registros duplicados, respectivamente, $\{\langle a, b\rangle; \langle c, d\rangle\}$ e $\{\langle a, b\rangle; \langle e, f\rangle\}$, como mostra a Figura A.1.

Os pares $\langle a, b\rangle$ e $\langle c, d\rangle$, identificados pelo primeiro classificador não-supervisionado, compõem o *baseline* e são automaticamente adicionados ao conjunto de duplicados. Na segunda iteração, $\langle a, b\rangle$ é confirmado como duplicado, mantendo-se no mesmo conjunto de início. Por sua vez, os pares $\langle c, d\rangle$ e $\langle e, f\rangle$ representam divergências entre o primeiro e o segundo classificadores não-supervisionados, necessitando, por isso, serem submetidos à inspeção manual.

Após o especialista humano analisar esses pares, averigua-se, neste exemplo, que $\langle c, d\rangle$ é de fato um par de registros duplicados e que $\langle e, f\rangle$ trata-se de um par não-duplicado. Assim, ambos são inseridos no conjunto de pares classificados manualmente e assumidos como verdadeiros, de forma que resultados de novos classificadores que possam vir a discordar deles serão ignorados.

Dos 28 pares possíveis, aqueles 25 pares que não foram classificados como duplicados por quaisquer dos classificadores não-supervisionados são automaticamente considerados como não-duplicados. Dessa forma, tem-se ao fim da execução dois pares duplicados e 26

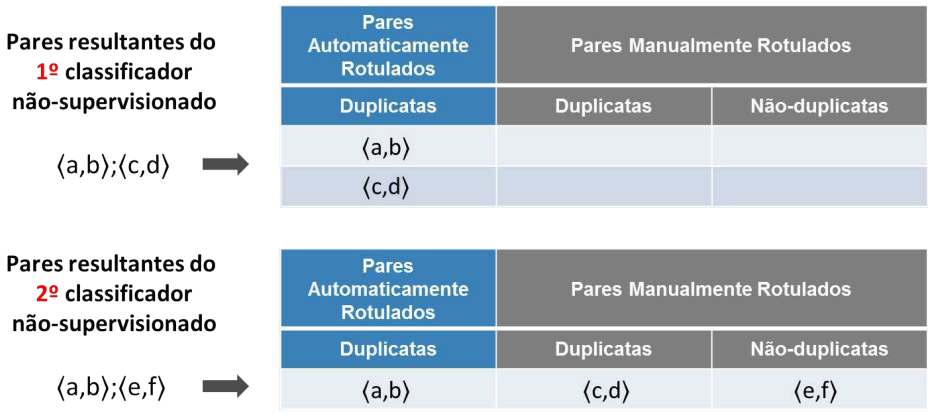


Figura A.1: Exemplo de execução do *AS*.

não-duplicados com apenas duas inspeções manuais, bem abaixo da quantidade de comparações necessárias quando utilizada a geração convencional de gabarito, que compara os pares gerados pelo Produto cartesiano sobre todos os elementos do conjunto.

Apêndice B

Algoritmo STERSWin

Algoritmo 2 STERSWin

Entrada:

- PC: lista de pares conflitantes na forma $\langle (r, s), sim(r, s), \#clasf \rangle$ ordenados crescentemente pelo valor *sim*
- ζ: oráculo humano, b: orçamento máximo
- $minT$: tamanho mínimo do conjunto de treinamento
- w_{ndup}: tamanho da janela a ser empregada sobre pares potencialmente não-duplicados
- w_{dup}: tamanho da janela a ser empregada sobre pares potencialmente duplicados
- d: limite inicial de w_{dup}, nd: limite inicial de w_{ndup}
- C^{sup}: conjunto de classificadores supervisionados

Saída: $Conjunto\ de\ treinamento\ (T)$

1: $prevF_1 \coloneqq 0$

2: $avgF_1 \coloneqq 10^{-5}$

3: $T \coloneqq \emptyset$

4: **for all** r in PC **do**
5: **if** $\#clasf <= k$ **then**
6: $PC^{nd} \coloneqq PC^{nd} \cup r$
7: **else**
8: $PC^{d} \coloneqq PC^{d} \cup r$
9: **end if**
10: **end for**
11: $d \coloneqq |PC^{d}|$
12: $nd \coloneqq 0$
13: **while** $avgF_1 > prevF_1$ **and** $budget(\zeta) \leq b$ **do**
14: $prevF_1 \coloneqq avgF_1$
15: $p_d \coloneqq argmin_{i=d\ down\ to\ d-wdup}\ \#clasf(PC_i^d)$
16: $p_{nd} \coloneqq argmax_{i=nd\ to\ nd+wndup}\ \#clasf(PC_i^{nd})$
17: $d \coloneqq d - wdup - 1$
18: $nd \coloneqq nd + wndup + 1$
19: $T \coloneqq T \cup \{\langle p_d, \zeta(p_d)\rangle, \langle p_{nd}, \zeta(p_{nd})\rangle\}$
20: **if** $T <= minT$ **then**
21: $sumF1 \coloneqq 0$
22: **for all** $classificador\ c^{sup} \in C^{sup}$ **do**
23: $sumF1 \coloneqq sumF1 + getF1(T, c^{sup})$
24: **end for**
25: **end if**
26: $avgF_1 \coloneqq \frac{sumF1}{|C^{sup}|}$
27: **end while**
28: **return** T