X Congresso Brasileiro de Inteligência Computacional (CBIC'2011), 8 a 11 de Novembro de 2011, Fortaleza, Ceará

COP-KMEANS E CLUSTERING SEMI-SUPERVISIONADO ATRAVÉS DE RESTRIÇÕES

Euler Teixeira, Antônio Braga, Hani Yehia
PPGEE-UFMG, DELT-UFMG, DELT-UFMG
euler.teixeira@cpdee.ufmg.br, apbraga@ufmg.br, hani@cefala.org

Resumo – O presente artigo descreve um aprimoramento do algoritmo de COP-Kmeans, dentro do contexto de *clustering* semi-supervisionado através de restrições. Esse método é uma derivação do popular algoritmo de agrupamento em k-médias, que permite a incorporação de informações conhecidas a priori sobre uma parcela do conjunto de dados, através de condições de restrição às ligações entre os padrões analisados. Serão definidos os fundamentos e conceitos básicos sobre o método, que também será implementado computacionalmente e usado para particionar alguns conjuntos de dados, com o objetivo de apontar as vantagens e desvantagens desse algoritmo sobre a sua versão tradicional não-supervisionada e também propor evoluções que possam corrigir suas principais fraquezas. Essas evoluções serão descritas ao final, definindo um algoritmo aprimorado para o método de COP-Kmeans, que será avaliado frente a conjuntos de dados amplamente utilizados na literatura da área.

Palavras-chave – COP-Kmeans, agrupamento semi-supervisionado.

Abstract – This paper describes an improvement on the COP-Kmeans algorithm, within the context of semi-supervised clustering based on constraints. This method is derived from the popular K-means clustering algorithm, allowing the incorporation of background information about the data, using linking constraints between the patterns. The basic concepts about this method will be defined, and it will be used to partition some data-sets, aiming to point out it's main virtues and problems. Then, evolutions will be proposed to improve it, and at the end, a new version of the algorithm will be presented and evaluated, based on the partition of benchmark data-sets.

Keywords – COP-Kmeans, semi-supervised clustering.

1 Introdução

Os métodos tradicionais de agrupamento de padrões são em geral usados de forma não-supervisionada. Seus algoritmos tomam como entrada um conjunto de dados, que é então particionado de acordo apenas com os atributos que descrevem os padrões, através de alguma medida de similaridade. Não são fornecidos rótulos, ou qualquer informação que indique onde cada padrão deve ser alocado. Em muitos casos, no entanto, se possui algum tipo de informação sobre o problema ou seus dados, que pode ser útil no processo de agrupamento. Como os algoritmos tradicionais não possibilitam uma forma direta de incorporação desse tipo de informação, houve uma busca por uma nova família de métodos que fosse capaz de se beneficiar desses dados. Na maior parte das aplicações reais, se possui uma grande quantidade de dados não rotulados e uma quantidade limitada de dados rotulados. Em virtude disso, existe grande interesse na definição de topologias de agrupamento semi-supervisionado, que são capazes de lidar simultaneamente com dados rotulados e não-rotulados, produzindo um particionamento mais eficiente, a partir de todas as informações disponíveis.

Pode-se dizer que os métodos de agrupamento semi-supervisionado propostos até então se dividem em dois tipos de abordagem [2]. Os métodos baseados em restrições e os métodos baseados em métricas de similaridade. Nos métodos baseados em restrições, o próprio algoritmo é modificado, de forma que restrições de agrupamento entre pares de padrões possam ser usadas para guiar o programa na direção de um particionamento mais adequado dos dados [2]. Isso é feito incluindo de alguma forma condições de satisfação das restrições definidas na função de objetivo do método. Já nos métodos baseados em métricas de similaridade, um algoritmo pré-existente de agrupamento não-supervisionado, que usa alguma métrica de distância ou similaridade, pode ser usado diretamente. No entanto, essa métrica é antes treinada, com base nos rótulos ou informações disponíveis sobre os padrões de supervisão, de forma a aumentar a eficiência do particionamento [2]. Esse trabalho será focado nos métodos de *clustering* semi-supervisionado baseados em restrições de agrupamento. Para isso, será estudado em detalhe um dos primeiros e mais populares representantes dessa categoria: o algoritmo de agrupamento COP-Kmeans (*constrained K-means*) [5], derivado do clássico algoritmo de k-médias.

2 Descrição do Método de COP-Kmeans

O algoritmo de COP-Kmeans é uma derivação semi-supervisionada do clássico algoritmo de k-médias [3], que permite a incorporação de informação conhecida a priori, sob a forma de restrições de agrupamento entre pares de padrões, durante o processo de particionamento do conjunto de dados de entrada. O algoritmo original de k-médias toma como entrada um conjunto

de dados D e fornece na saída um particionamento dos seus padrões em k grupos. Na maioria das aplicações, o valor de k é conhecido, e os valores iniciais para os centros dos grupos são tomados aleatoriamente dentre os padrões do conjunto de dados

2.1 Incorporação das Restrições de Agrupamento

Em problemas de agrupamento e particionamento de dados, restrições de agrupamento no nível dos padrões são uma boa forma de exprimir o conhecimento disponível a priori, no que diz respeito a quais pontos devem ou não ser colocados no mesmo grupo [5]. Nesse contexto, dois tipos simples e diretos de restrição de agrupamento entre pares de padrões foram definidos no método de COP-Kmeans:

- Ligação Obrigatória (LOB) : determina que os dois padrões em questão devem estar sempre no mesmo grupo.
- Ligação Proibida (LPR): determina que os dois padrões em questão nunca podem estar no mesmo grupo.

Cada um desses dois tipos de restrição pode ser representado por uma matriz binária NxN, onde N é o número de padrões de entrada e cada elemento (i, j) dessas duas matrizes indica se há uma condição de ligação obrigatória ou proibida entre o padrão d_i e o padrão d_j. Esses conjuntos de restrições são geralmente derivados da parcela rotulada dos dados, podendo também ser definidos a partir de algum outro tipo de conhecimento prévio sobre o problema.

2.2 O Algoritmo de COP-Kmeans

O algoritmo de COP-Kmeans possui uma estrutura similar ao do algoritmo de k-médias descrito anteriormente. A única diferença está no passo 2, que faz a atribuição dos padrões ao grupo de centro mais próximo. Essa etapa do programa deve agora também assegurar que nenhuma das restrições de ligação definidas seja violada, durante esse processo de atribuição.

O algoritmo tenta associar o padrão d_i ao grupo de centro C_j mais próximo. Se houver outro ponto d_l de ligação obrigatória a d_i, que não está associado a C_j, ou outro ponto d_l de ligação proibida a d_i, que está associado a C_j, será constituída uma violação de restrição e a atribuição não será feita. O algoritmo busca então o grupo de centro C_j mais próximo dentre os demais e testa se o padrão d_i pode ser associado a ele sem violar nenhuma restrição. O processo continua seqüencialmente, até que um grupo seja encontrado. Se nenhum grupo puder abrigar d_i legalmente, ele não será agrupado nessa iteração do programa (grupo = 0). A seguir são descritos em resumo os passos desse algoritmo de COP-Kmeans [5], que toma como entrada o conjunto de dados D e as matrizes de restrição LOB e LPR, fornecendo na saída um particionamento de D em k grupos.

1. Selecione k centros C_j iniciais para os grupos.
2. Associe cada padrão d_i do conjunto D ao grupo cujo centro está mais próximo dele, de forma que a Verificação-de-Violação-de-Restrições(d_i, C_j) seja negativa. Se tal grupo não existir, retorne C_0
3. Atualize cada centro de grupo C_j como sendo a média de todos os padrões d_i associados a ele.
4. Repita os passos 2 e 3 iterativamente, até que nenhum padrão troque mais de grupo.

Verificação-de-Violação-de-Restrições(d_i, C_j):

1. Para cada padrão $d_l, l = (1...N)$, se $LOB(d_i, d_l)$ é verdadeiro e d_l não está associado a C_j, retorne positivo.
2. Para cada padrão $d_l, l = (1...N)$, se $LPR(d_i, d_l)$ é verdadeiro e d_l está associado a C_j, retorne positivo.
3. Caso contrário, retorne negativo.

3 Implementação Computacional

O programa foi executado, fornecendo como entrada: o conjunto de dados D, o número k de grupos, e o número nR de restrições de ligação a serem geradas. Os centros dos grupos $(C_1...C_k)$ foram inicializados aleatoriamente. O conjunto inicial de dados D pode ser visto na Figura 1. Ele é constituído de 40 padrões, descritos por 2 atributos e distribuídos ao longo de 5 classes. As condições de restrição também foram geradas aleatoriamente, a partir dos rótulos dos dados, da seguinte forma:

1. Duas matrizes de zeros NxN, LOB e LPR, foram criadas.
2. Para cada uma das nR restrições a serem geradas, o programa tomou aleatoriamente dois pontos d_i e d_l.
3. Se o rótulo de d_i fosse igual ao rótulo de d_l, $LOB(d_i, d_l)$ e $LOB(d_l, d_i)$ eram feitos iguais a 1.
4. Se o rótulo de d_i fosse diferente do rótulo de d_l, $LPR(d_i, d_l)$ e $LPR(d_l, d_i)$ eram feitos iguais a 1.

As matrizes binárias LOB e LPR, geradas dessa forma, foram então usadas na etapa de verificação de violação de restrições do algoritmo, supervisionando o processo de agrupamento dos padrões, de acordo com a parcela dos dados que teve seus rótulos apresentados ao programa na etapa de inicialização das condições de restrição.

Como exemplo, a Figura 2 ilustra o resultado do agrupamento obtido pelo programa, em uma execução com $nR = 10$ que convergiu plenamente, alocando todos os padrões nos grupos corretos. Os marcadores em forma de cruz representam os centros dos grupos $(C_1...C_k)$, e os grandes círculos em torno deles englobam os padrões associados a cada um desses grupos.

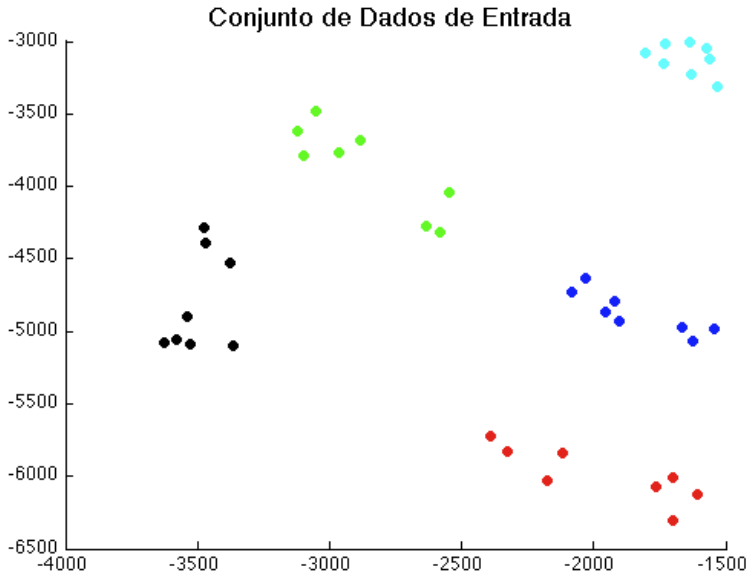

Figura 1: Conjunto bidimensional de dados D, com $N = 40$ e $k = 5$.

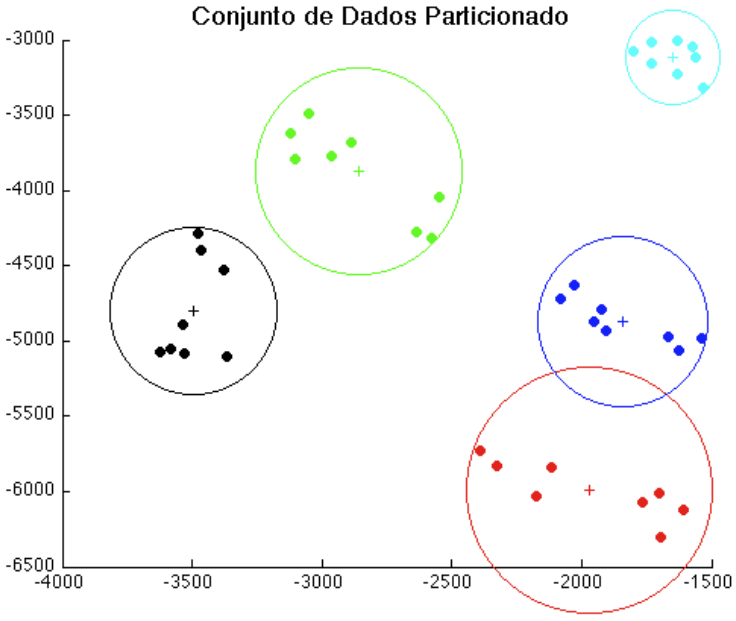

Figura 2: Resultado do agrupamento, em uma execução com $nR = 10$, que obteve 100% de acerto.

3.1 Índice de Desempenho

A acurácia desse algoritmo de agrupamento pode ser medida pelo índice Rand [4], que fornece uma taxa de concordância entre dois particionamentos P_1 e P_2 do mesmo conjunto de dados D. Nesse caso, P_1 representa a rotulação original dos dados e P_2 representa o agrupamento encontrado pelo algoritmo. Cada particionamento é visto como uma coleção de $(N^2 - N)/2$ decisões pareadas, conforme já foi descrito anteriormente. Para cada par de padrões (d_i, d_l), ou P os coloca no mesmo grupo, ou em grupos diferentes. Chamando de a o número de decisões onde d_i e d_l foram colocados no mesmo grupo, tanto por P_1 quanto por P_2, e chamando de b o número de decisões onde d_i e d_l foram colocados em grupos distintos, tanto por P_1 quanto por P_2, o índice Rand de acurácia percentual é dado por:

$$Rand(P_1, P_2) = \frac{a + b}{(N^2 - N)/2} * 100\%$$

Dessa forma, visando fazer uma análise de desempenho do método, o programa foi executado diversas vezes, com valores crescentes para o número nR de restrições definidas, e o índice Rand de acurácia percentual foi calculado em cada uma delas. A Figura 3 ilustra a relação resultante entre a acurácia obtida no agrupamento e o número de restrições empregadas pelo algoritmo. Como a inicialização dos centros dos grupos e a definição das restrições de ligação são feitas aleatoriamente, cada ponto do

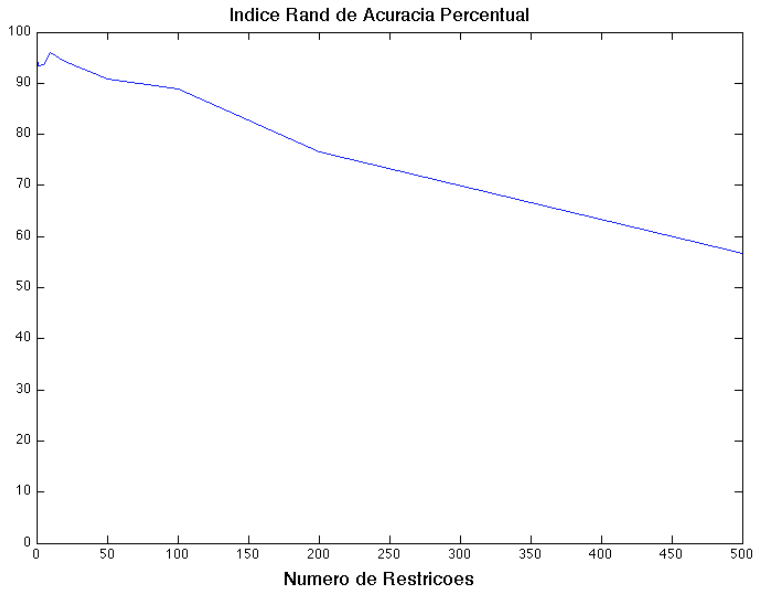

Figura 3: Índice Rand de acurácia percentual do agrupamento, em função do número nR de restrições de ligação empregadas.

gráfico foi obtido através do cálculo do índice Rand médio ao longo de 10 execuções do programa, de forma a minimizar a influência das condições iniciais.

4 Avaliação e Aprimoramento do Método de COP-Kmeans

Analisando a Figura 3 nota-se que a acurácia máxima atingida pelo algoritmo de COP-Kmeans foi obtida com o uso de 10 condições de restrição de ligação. Com o aumento do número de condições definidas, além desse ponto, foi observada uma queda monotônica e aproximadamente linear no índice Rand de acurácia. A partir de cerca de 100 condições de restrição, o algoritmo também apresentou certa instabilidade de convergência, o que contribuiu para a redução progressiva de sua acurácia média. Na região de 0 a 10 condições de restrição, o algoritmo exibiu, por outro lado, um aumento progressivo no índice de acurácia, que depois decaiu linearmente até cerca de 50 restrições e se manteve relativamente estável na região entre 50 e 100 restrições. Na região entre 0 e 100 condições de restrição, o algoritmo exibiu índices médios de acurácia superiores a 90%, com pico de 97%, o que indica um bom desempenho para esse problema nessas condições. Também é importante ressaltar que na condição de 0 restrições, o algoritmo se comporta exatamente como o algoritmo de k-médias original, e que nesse caso ele obteve uma acurácia de 94%, bem próxima do pico de desempenho alcançado.

Esse problema é relativamente simples, já que o conjunto de dados é pequeno, bidimensional e as classes estão razoavelmente separadas no espaço de parâmetros. Os dados de acurácia analisados sugerem que, nesse tipo de caso, o algoritmo não-supervisionado já é capaz de obter ótimos resultados e portanto, o estabelecimento de condições de restrição pouco tem a contribuir para o problema. Ao contrário, para um conjunto tão pequeno de dados, o estabelecimento de um número muito grande de condições de restrição acabou por ocasionar dificuldades de convergência, já que elas acarretam em um grande aumento da complexidade dos cálculos e reduzem a capacidade de movimentação dos padrões entre os grupos pelo programa, aumentando a sensibilidade às condições de inicialização. Mesmo assim, o estabelecimento de um número razoável de restrições possibilitou um aumento, ainda que pequeno, do índice de acurácia e, nesses casos, também da velocidade de convergência do programa, reduzindo em geral o número de iterações requeridas para o agrupamento.

Outro dado interessante de ser analisado é a porcentagem das execuções que obtiveram 100% de acerto no agrupamento, em função do número de condições de restrição definidas. Esses dados são ilustrados na Figura 4. É possível notar como o aumento do número de condições de restrição acarreta em uma forte queda desse parâmetro. Apesar de o aumento do número de restrições melhorar a acurácia média do programa na região entre 0 e 10 restrições, e pouco afetá-la na região entre 0 e 100 restrições, esse aumento reduz rapidamente a taxa de acerto pleno do algoritmo. Isso indica que as condições de restrição reduzem progressivamente a robustez do programa, mais uma comprovação de que essas restrições aumentam a sensibilidade do algoritmo às condições de inicialização. Como a acurácia média permanece alta, mas o programa parece ter maior dificuldade em obter soluções totalmente corretas, isso também indica a presença de anomalias e problemas pontuais no agrupamento.

Uma interpretação desse problema é que o algoritmo pode estar gerando imobilizações indesejadas de alguns padrões, em virtude das condições de restrição e das condições de inicialização. De fato, uma anomalia que pode gerar tal tipo de imobilização foi identificada, com base na análise do algoritmo. O passo 1 da rotina de verificação de violação de restrições impede que um padrão d_i seja alocado a um grupo C_j, se houver um outro padrão d_l, de ligação obrigatória a d_i, que não está alocado nesse mesmo grupo C_j. Isso pode gerar o seguinte problema: Se esses dois padrões de ligação obrigatória, d_i e d_l, forem inicializados no mesmo grupo, eles ficarão impossibilitados de trocar de grupo, já que apenas um deles pode se movimentar de cada vez e

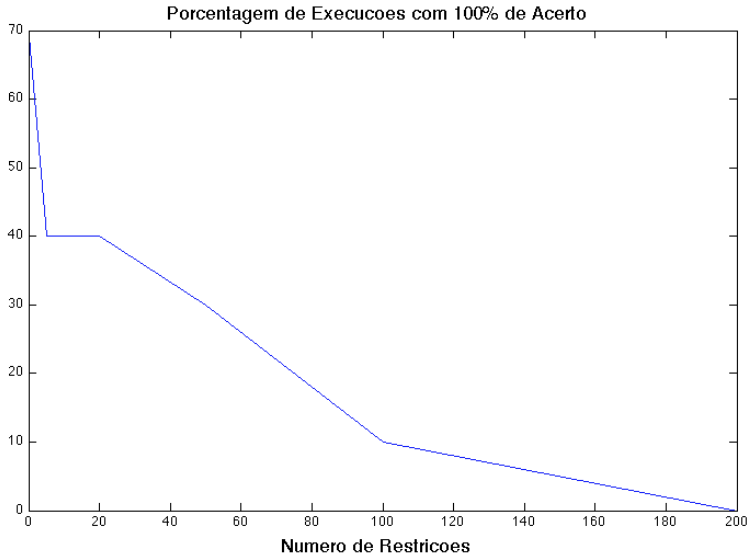

Figura 4: Porcentagem das execuções que obtiveram 100% de acerto no agrupamento, em função do número nR de restrições de ligação empregadas.

dessa forma nenhum outro grupo pode abrigá-los legalmente. Se o grupo ao qual eles foram inicialmente alocados não for o seu grupo correto de destino, eles ficarão imobilizados indefinidamente e não serão adequadamente agrupados ao final da execução.

Em suma, pode se dizer que o método de COP-Kmeans é bastante eficiente, possuindo como principal virtude a alta flexibilidade, já que ele possibilita o uso de todas as informações disponíveis sobre o problema no agrupamento, independente da parcela dos dados que dispõe de rótulos, ou do número de condições de restrição desejadas, bem como da forma como essas condições serão geradas. Com o número de condições de restrição escolhido adequadamente para o problema, altos índices de acurácia podem ser obtidos e o algoritmo é de fácil ajuste, a partir de poucos parâmetros. Sua principal deficiência está na sua grande sensibilidade às condições de inicialização, gerada em parte pela reduzida fluidez de movimentação dos padrões entre os grupos, em condições de muitas restrições, e também pelo método aleatório de inicialização. Essa deficiência pode gerar instabilidades de convergência e problemas pontuais de imobilização de padrões em etapas iniciais do programa, impedindo em muitos casos a obtenção de taxas plenas de acerto. A seguir serão propostos alguns aprimoramentos nesse algoritmo, no sentido de contornar os problemas descritos anteriormente, melhorando o desempenho e a robustez do método.

4.1 Propostas de Evolução ao Algoritmo

Como já foi descrito e exemplificado nessa seção, o algoritmo de COP-Kmeans é bastante sensível às condições de inicialização, o que pode gerar instabilidades de convergência e problemas pontuais de imobilização indesejada de alguns padrões. Aqui serão propostas 3 alterações no algoritmo, que buscam minimizar essas deficiências. Conforme explicado, o problema de imobilização indesejada de padrões decorre de uma inconsistência identificada no passo 1 da rotina de verificação de violação de restrições, que pode ancorar um par de padrões obrigatoriamente ligados a um grupo inadequado. Para solucionar esse problema, a seguinte modificação pode ser feita nessa etapa do algoritmo:

Verificação-de-Violação-de-Restrições(d_i, C_j):

1. Para cada padrão $d_l, l = (1...N)$, se $LOB(d_i, d_l)$ é verdadeiro e d_l não está associado a C_j:
   - Verifique se d_l também será associado a C_j na próxima iteração
   - Caso isso for ocorrer, retorne negativo
   - Caso isso não for ocorrer, retorne positivo

Dessa forma o algoritmo permite que dois padrões obrigatoriamente ligados se movimentem ao mesmo tempo, desde que seu grupo de destino seja o mesmo. Isso resolve uma boa parte dos problemas de ancoramento descritos, já que pontos obrigatoriamente ligados tendem a se mover para grupos iguais. Outro tipo de aprimoramento ao algoritmo, que pode ajudar a evitar esse tipo de problema de imobilização, bem como conferir uma maior fluidez global de movimentação entre grupos aos padrões, seria impedir que as condições de restrição atuem nas iterações iniciais do algoritmo. Nessas iterações iniciais os centros dos grupos se movem muito, já que ainda estão longe de seu destino final e muitos padrões serão re-alocados, exigindo uma boa fluidez de movimentação entre grupos, até que o agrupamento se aproxime a grosso modo de uma situação de equilíbrio. Passada essa etapa do processo, as condições de restrição poderiam ser ativadas, guiando o ajuste fino do agrupamento. O instante de início de atuação das restrições seria mais um parâmetro de ajuste, aumentando ainda mais a flexibilidade do método.

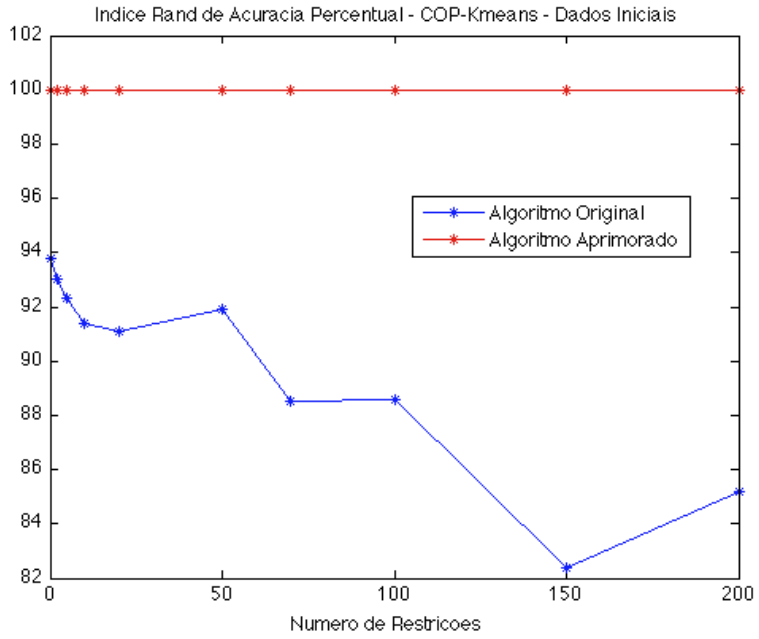

Figura 5: Índice Rand de acurácia percentual do agrupamento, em função do número de restrições empregadas - conjunto de dados inicial.

Essas duas propostas irão provavelmente reduzir consideravelmente a sensibilidade do algoritmo às condições iniciais, mas dentro do contexto de agrupamento semi-supervisionado, existe uma outra forma muito eficiente de torná-lo ainda mais robusto em relação à sua condição de inicialização [1]. Uma pequena parcela dos dados rotulados, de preferência os mesmos usados na geração das restrições, pode ser usada para fazer uma estimação inicial dos centros dos grupos. Desde que os rótulos estejam disponíveis, esse é um processo simples e que certamente pode acelerar muito a convergência do algoritmo, além de prevenir grande parte das instabilidades de execução identificadas. Espera-se que com essas alterações o desempenho e principalmente a robustez do algoritmo sejam melhorados consideravelmente, com base nos conceitos e dados apresentados até aqui.

5 Algoritmo de COP-Kmeans Aprimorado

As evoluções propostas na seção anterior foram então incorporadas ao algoritmo inicial de COP-Kmeans. A rotina de verificação de violação de restrições foi relaxada, conforme descrito anteriormente, sendo também apenas ativada a partir da terceira iteração do algoritmo. Além disso, foram escolhidos aleatoriamente três padrões de cada classe, para definir a posição inicial dos centros dos k grupos.

O algoritmo com esses aprimoramentos foi então aplicado a três conjuntos de dados do repositório da UCI, além do conjunto inicial descrito anteriormente. As bases de dados escolhidas foram a Iris (150 padrões, 4 atributos, 3 classes), a Soybean (47 padrões, 35 atributos, 4 classes) e a Wine (178 padrões, 13 atributos, 3 classes). Todas elas já foram amplamente usadas na literatura da área e por isso são bem representativas e adequadas para avaliar a eficiência do algoritmo de agrupamento. Elas também fornecem um bom grau de dificuldade para o agrupamento e diferentes panoramas de entrada para avaliar a adaptabilidade do algoritmo. As figuras de 5 a 8 ilustram os níveis de acurácia obtidos pelo algoritmo original de COP-Kmeans e pelo algoritmo aprimorado, em função do número de restrições definidas, para cada um desses conjuntos de dados. Esses níveis foram obtidos com o uso do índice de Rand, conforme descrito anteriormente no texto.

Analisando essas figuras, é possível observar que o aprimoramento no algoritmo de COP-Kmeans gerou uma grande melhora de acurácia no agrupamento, em todos os conjuntos de dados avaliados. A acurácia máxima subiu de 93% para 100% no conjunto inicial, de 86% para 94% na base Iris, de 84% para 97% na base Soybean e de 71% para 88% na base Wine. Além disso, é possível notar como, para o algoritmo aprimorado, os níveis de acurácia tenderam a subir progressivamente com o aumento do número de restrições definidas, até o ponto ótimo de operação, como seria de se esperar para um algoritmo semi-supervisionado. O mesmo não ocorreu para o algoritmo original, que apresentou um comportamento mais instável para essas curvas de acurácia, que em geral decresceram nessas mesmas regiões de operação, devido aos problemas apontados anteriormente no texto.

Mesmo na ausência de condições de restrição ($nR = 0$), o algoritmo aprimorado conseguiu um ganho significativo de acurácia. Isso pode ser observado na extremidade esquerda dos gráficos e ilustra como a estimação inicial dos centros dos grupos produz resultados melhores do que a sua inicialização aleatória, mesmo com uma estimação baseada em uma pequena parcela dos padrões de entrada. As curvas de acurácia do algoritmo aprimorado também se situaram acima das curvas do algoritmo original, ao longo de toda a região analisada, em todos os casos. Além disso, a introdução de condições de restrição no algoritmo aprimorado possibilitou, em todos os casos, um aumento significativo do nível máximo de acurácia obtido, o que não ocorreu para o algoritmo original.

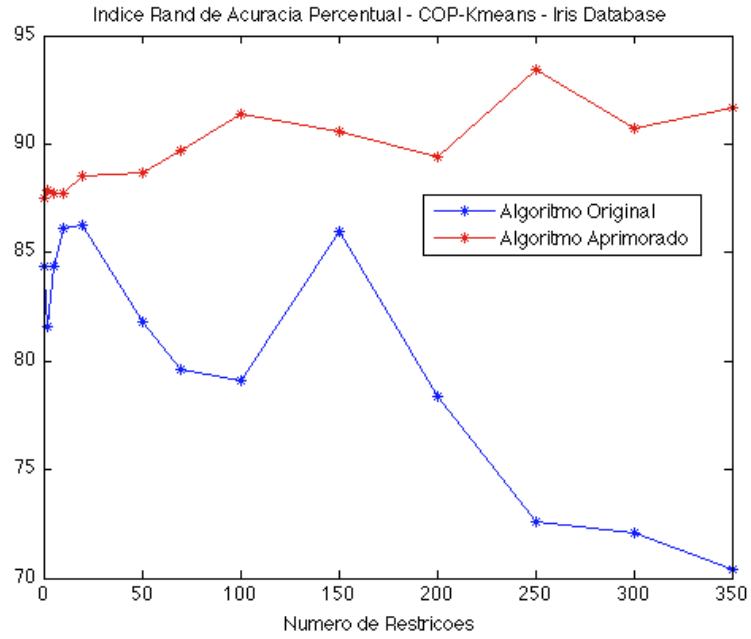

Figura 6: Índice Rand de acurácia percentual do agrupamento, em função do número de restrições empregadas - conjunto de dados Iris.

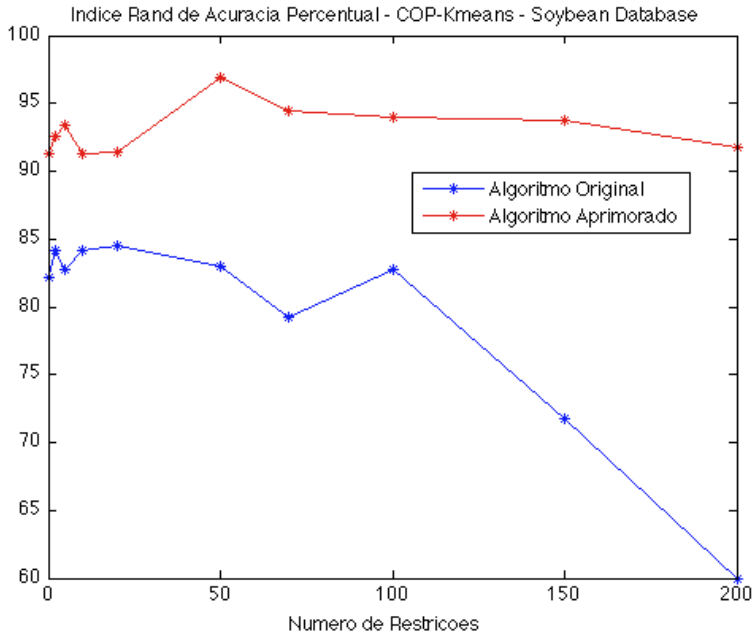

Figura 7: Índice Rand de acurácia percentual do agrupamento, em função do número de restrições empregadas - conjunto de dados Soybean.

Tudo isso comprova que o relaxamento proposto nas condições de restrição, além do método de inicialização, gerou um algoritmo mais robusto e eficiente de agrupamento. A maior melhora de acurácia foi observada no agrupamento do conjunto de dados Wine. O algoritmo aprimorado conseguiu nesse caso um aumento de 17% na acurácia máxima, em relação ao algoritmo original. Além disso, para o algoritmo aprimorado, a introdução de 500 condições de restrição gerou uma melhora de 16% na acurácia, em relação à condição de ausência de restrições. Essa é certamente a base de dados mais complexa, com o maior número de padrões e um grande número de atributos, onde os menores índices de acurácia foram obtidos, especialmente para o algoritmo original. Isso também demonstra como o aprimoramento realizado no algoritmo se mostrou especialmente eficiente em problemas difíceis, justamente onde os avanços são mais necessários.

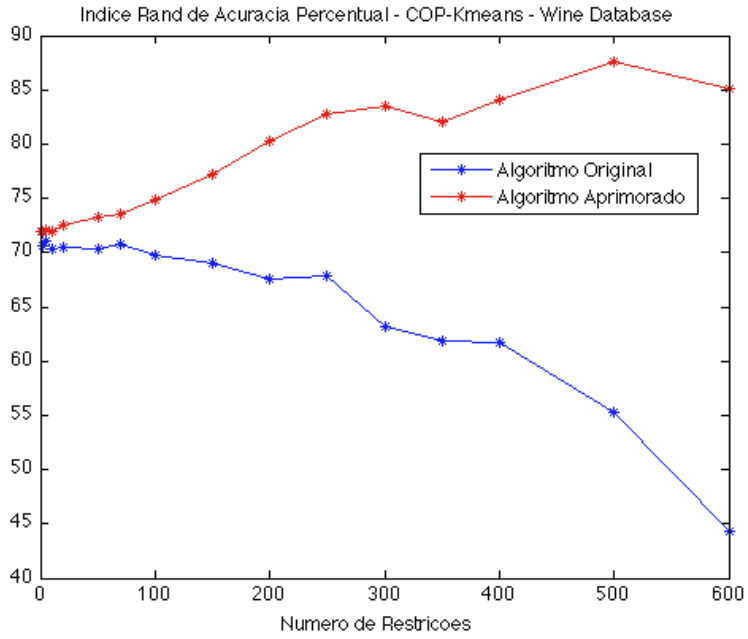

Figura 8: Índice Rand de acurácia percentual do agrupamento, em função do número de restrições empregadas - conjunto de dados Wine.

6 Conclusões

Foi observado que o algoritmo de COP-Kmeans avaliado inicialmente é bastante versátil, intuitivo e capaz de obter bons níveis de acurácia. Apesar disso, ele também apresenta grande sensibilidade às condições de inicialização e possui algumas inconsistências no método, o que pode gerar, em alguns casos, instabilidades de convergência e problemas pontuais de agrupamento. Três alterações foram então propostas ao algoritmo inicial para resolver esses problemas. Essas evoluções foram implementadas, gerando uma versão aprimorada do algoritmo de COP-Kmeans, com um relaxamento das condições de restrição e uma estimação inicial para os centros dos grupos. Finalmente, foram conduzidas avaliações completas de desempenho sobre esse novo algoritmo, utilizando conjuntos complexos de dados amplamente conhecidos na literatura da área, apontando até onde as modificações se refletiram nos índices de acurácia de agrupamento.

Os resultados obtidos mostraram um grande aumento de robustez e acurácia do algoritmo de agrupamento, devido às evoluções propostas, para todos os conjuntos de dados avaliados, especialmente nos problemas mais complexos. A estimação inicial dos centros dos grupos já produziu por si só um grande aumento de eficiência, mesmo com baixo nível de supervisão. O relaxamento das condições de restrição também melhorou significativamente o desempenho do algoritmo, gerando uma relação crescente e mais coerente entre a acurácia de agrupamento e a taxa de supervisão utilizada.

Em todos os problemas analisados, o algoritmo aprimorado de COP-Kmeans conseguiu resultados superiores aos obtidos pelo algoritmo inicial, também se beneficiando bem mais do conhecimento a priori incorporado sob a forma das condições de restrição. Além disso, pode se dizer que os níveis máximos de acurácia atingidos ao final foram muito bons para os problemas de agrupamento em questão, levando-se em conta os resultados de outros métodos disponíveis na literatura [2,5]. Tudo isso qualifica esse algoritmo aprimorado de COP-Kmeans como uma boa ferramenta de agrupamento semi-supervisionado de padrões, para os mais diversos problemas da área.

REFERÊNCIAS

[1] Basu, S. , Banerjee, A. , Mooney, R. *Semi-supervised clustering by seeding*, Proceedings of the 19th International Conference on Machine Learning, pp. 19-26, 2002.

[2] Bilenko, M. , Basu, S. , Mooney, R. *Integrating constraints and metric learning in semi-supervised clustering*, Proceedings of the 21st International Conference on Machine Learning, 2004.

[3] MacQueen, J. *Some methods for classification and analysis of multivariate observations*, Proceedings of the 5th Symposium on Math, Statistics and Probability, pp. 281-297, 1967.

[4] Rand, W. *Objective criteria for the evaluation of clustering methods*, Journal of the American Statistical Association, vol.66, pp. 846-850, 1971.

[5] Wagstaff, K. , Cardie, C. , Rogers, S. , Schroedl, S. *Constrained K-means clustering with background knowledge*, Proceedings of the 18th International Conference on Machine Learning, pp. 577-584, 2001.