IMPLEMENTAÇÃO DE UMA REDE RBF COMO CLASSIFICADOR SEMI-SUPERVISIONADO BASEADO EM COP-KMEANS

Brayan Acevedo Jaimes*, Frank Sill Torres*, Cristiano Leite de Castro*, Antônio de Pádua Braga*

*Programa de Pós-Graduação em Engenharia Elétrica, Universidade Federal de Minas Gerais - Av. Antônio Carlos 6627, 31270-901, Belo Horizonte, MG, Brasil

Emails: payo@ufmg.br, franksill@ufmg.br, crislcastro@ufmg.br, apbraga@ufmg.br

Abstract— This work presents the implementation of a Radial Basis Function neural network (RBF) acting as classifier; it incorporates a COP-Kmeans algorithm to perform the centroids calculus of each cluster, within the context of semi-supervised clustering. The algorithm COP-Kmeans allows including a priori known informations over a subset of the data using restrictions intra-cluster or between different clusters. The RBF network contains an intermediate layer which has a Gaussian activation function, where the centroid value C is given by the COP-Kmeans implementation applied to training data. The method evaluation considers the classification of a data set partially labeled with non-linearly separable characteristics. Tests and simulations were performed using different numbers of centroids by each class in the RBF network. The obtained results are satisfactory and show that the COP-Kmeans algorithm provides properly the centroids in order to achieve an adequate classification of the implemented neural network.

Keywords— RNAs; COP-Kmeans; RBF; Must-Link; Cannot-Link

Resumo— Este trabalho descreve a implementação de uma rede RBF (Radial Basis Function) como classificador; incorporando nesta o algoritmo COP-Kmeans para o cálculo dos centroides de cada *cluster*, dentro do contexto de clustering semi-supervisionado. O algoritmo COP-Kmeans permite a inclusão de informações conhecidas a priori sobre uma parcela de um conjunto de dados, através de restrições de tipo intra *clusters* ou entre diferentes *clusters*. A rede RBF possui uma camada intermediária que tem uma função de ativação de tipo gaussiano, onde o valor do centroide C é fornecido pela implementação do COP-Kmeans aplicado nos dados de treinamento. A avaliação do método considera a classificação de um conjunto de dados parcialmente rotulados com características não linearmente separáveis. Testes e simulações foram realizados utilizando diferentes números de centroides por cada classe na rede RBF. Os resultados obtidos são satisfatórios e indicam que o algoritmo COP-Kmeans fornece centroides apropriados para que a rede RBF consiga uma adequada classificação.

Palavras-chave— RNAs; COP-Kmeans; RBF; Must-Link; Cannot-Link

1 Introdução

O agrupamento de dados é considerado uma abordagem de grande interesse para encontrar semelhanças entre conjuntos de amostras. Assim, o clustering particiona um conjunto de dados em vários grupos, de tal modo que a semelhança dos indivíduos dentro de um mesmo grupo é maior que entre indivíduos de diferentes grupos (Jang et al., 1997). Os métodos tradicionais de agrupamento de padrões são em geral usados de forma não-supervisionada. Portanto, o algoritmo toma como entrada um conjunto de dados, que é então particionado de acordo com os atributos que descrevem os padrões, através de alguma medida de similaridade. Mas não são fornecidos rótulos, ou qualquer informação que indique onde cada padrão deve ser alocado ou a que *cluster* deve pertencer. Um dos algoritmos mais utilizados para este tipo de agrupamento não-supervisionado é o algoritmo K-médias (Kanungo et al., 2002).

Em muitos casos, no entanto, possui-se algum tipo de informação sobre o problema ou seus dados, que pode ser útil no processo de agrupamento. Na maior parte das aplicações reais, tem-se 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 (Basu et al., 2002; Grira et al., 2004), que são capazes de lidar simultaneamente com dados rotulados e não-rotulados. Isto produz um particionamento mais eficiente, a partir de todas as informações disponíveis e extraídas dos dados. Os métodos de agrupamento semi-supervisionado propostos até então se dividem em dois tipos de abordagem, os métodos baseados em restrições e os métodos baseados em métricas de similaridade. O método baseado em restrições é aplicado no desenvolvimento deste trabalho. O algoritmo define um número de restrições de agrupamento entre pares de dados rotulados, que possam ser utilizadas para guiar o algoritmo na direção de uma classificação mais adequada dos dados.

Como trabalhos relacionados neste abordagem, os autores (Basu et al., 2002) propõem duas variações semi-supervisionadas do algoritmo Kmeans fazendo uso dos dados rotulados na geração de centroides iniciais. Os autores (Basu et al., 2003) visando facilitar o agrupamento semi-supervisionado através do aproveitamento de informação de pequenas quantidades de dados rotulados, apresentam um enfoque unificado baseado no algoritmo kmeans onde utilizam métodos de agrupamento supervisionados e métodos base-

ados em similaridade para guiar a busca dos melhores centroides. Finalmente, os autores (Bilenko et al., 2004) desenvolveram um método de agrupamento semi-supervisionado unificado, que juntam métodos baseados em restrições e métodos de aprendizado em função de distancia para melhorar o agrupamento de dados.

Este artigo tem como objetivo avaliar o desempenho da rede RBF como classificador quando se tem um conjunto de dados parcialmente rotulado com características não linearmente separáveis. Foi implementado o algoritmo de agrupamento COP-Kmeans (Wagstaff et al., 2001) para determinar os centroides de cada *cluster* no conjunto de dados que correspondem a cada neurônio da camada intermediária da rede RBF. Esta rede utiliza funções de tipo gaussianas onde o valor de ativação é dado em função da distância entre a entrada e o centro da gaussiana. Assim, conseguiu-se aproveitar o conhecimento extraído de padrões rotulados em um conjunto de dados para fornecer uma classificação com restrições entre classes. A diferença de quando é utilizado o algoritmo K-médias que não considera a informação de dados rotulados para gerar os centroides das classes.

2 COP-Kmeans

O algoritmo de COP-Kmeans é uma derivação semi - supervisionada do clássico algoritmo de k-médias, 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 (Texeira, 2011). Em problemas de agrupamento e classificação de dados, as restrições de agrupamento disponíveis são uma boa forma de aproveitar o conhecimento disponível *a priori*, no que diz respeito a quais pontos devem ou não ser colocados no mesmo grupo (MacQueen et al., 1967). Neste contexto, dois tipos simples e diretos de restrição de agrupamento entre pares de padrões foram definidos no método de COP-Kmeans:

- Ligação Obrigatória - *Must-Link* (LOB): determina que dois padrões em questão devem estar sempre no mesmo grupo.
- Ligação Proibida - *Cannot-Link* (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. 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 restrição alguma. O processo continua sequencialmente até que um grupo seja encontrado. Se nenhum grupo pode conter d_i legalmente, ele não será agrupado nessa iteração do programa. Em seguida são descritos os passos do algoritmo de COP-Kmeans, que toma como entrada o conjunto de dados D e as matrizes de restrição LOB e LPR, fornecendo na saída um agrupamento de D em k *clusters*.

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_o.
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 Rede RBF

Uma rede RBF é um tipo particular de rede neural que fornece interessantes resultados em sua utilização como um classificador não linear. Ao contrário das redes MLP, a abordagem das redes RBF é mais intuitiva; a rede faz classificação medindo a semelhança dos dados de entrada, onde cada neurônio da camada intermediária possui um centroide. Quando se quer classificar uma nova entrada, em cada neurônio da camada intermediária é calculada a distância euclidiana entre a entrada e o centroide de cada neurônio. Seguidamente, a camada de saída da rede realiza somas ponderadas sobre os sinais de ativação dos neurônios da camada intermediária, a fim de ativar ou inibir a saída correspondente a cada classe, obtendo a classificação da entrada. A Fig. 1 mostra a arquitetura típica de uma rede RBF, constituída por um vetor de entrada, uma camada intermediária com neurônios ativados por funções de base radial (função Gaussiana) e uma camada de saída que usa uma função de ativação linear, tipo Adaline (WIDROW et al., 1960).

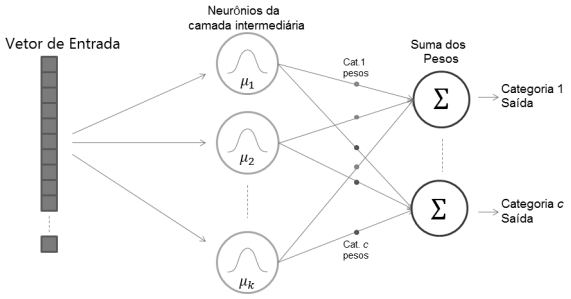

Figura 1: Arquitetura típica de Rede RBF.

O vetor de entrada são os dados a classificar, este vetor vai ser avaliado para cada um dos neurônios da camada intermediária. Cada neurônio armazena um dado protótipo que vai ser o centroide de cada *cluster* que pertence ao conjunto de dados de treinamento. Cada neurônio compara o vetor de entrada com seu centroide em termos de distância euclidiana. Esta é utilizada na função de ativação de base radial (gaussiana) retornando um valor entre 0 e 1, que representa uma medida de semelhança. Se a entrada foi igual ao centroide, então a saída desse neurônio na camada intermediária será igual a 1; à medida que a distância entre o dado de entrada e o centroide é maior, o valor da medida de semelhança vai diminuir exponencialmente em direção a 0. Este valor também é conhecido como valor de ativação. A forma da resposta do neurônio é uma forma de tipo gaussiana, ilustrado na Fig. 2.

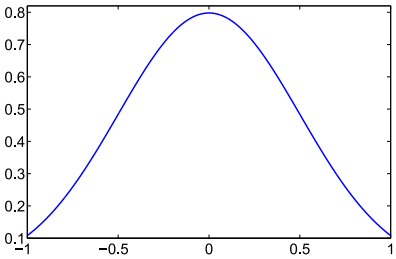

Figura 2: Formato de saída do neurônio da camada intermediária.

A camada de saída da rede é um conjunto de nodos de uma categoria que se pretende classificar, cada nodo de saída calcula pontuações de cada categoria associada, a decisão de classificação é feita atribuindo na saída a entrada com a maior pontuação. A pontuação é obtida tomando a soma ponderada dos valores de ativação de cada neurônio da camada intermediária. Como cada nodo de saída está calculando a pontuação para uma categoria diferente, cada um deles tem seu próprio conjunto de pesos. De modo que, na saída vai ter um peso positivo para os neurônios da camada intermediária que pertencem a sua categoria, e um peso negativo para os outros (Schwenker et al., 2001).

Cada neurônio da camada intermediária calcula uma medida da similaridade entre a entrada e seu centroide que é extraído do conjunto de treinamento. Os dados de entrada, que sejam mais próximos ao centroide retornam resultados mais próximos de 1. Em suma, existem diferentes escolhas possíveis de funções de similaridade, mas o mais popular é baseado na função de ativação Gaussiana descrita em (1). Onde x é cada um dos dados de entrada, μ é o centroide de cada neurônio na camada intermediária e σ corresponde ao desvio padrão.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{1}$$

A função de ativação dos neurônios da camada intermediária da rede RBF implementada, onde o centroide calculado pelo COP-Kmeans é o valor de μ; e está escrita como:

$$\varphi(x) = e^{-\beta\|x-\mu\|^2} \tag{2}$$

4 Implementação e Resultados

Como o objetivo é descrever o comportamento da rede RBF baseado no algoritmo COP-Kmeans e observar seu comportamento como classificador, foi aplicado e testado com um conjunto de dados chamado 2moons.dat que consiste em duas regiões semilunares com faces côncavas contrapostas, que possui características de dados não-linearmente separáveis. Esta base de dados sintética esta composta por 200 valores distribuídos entre as duas classes. A Fig. 3 mostra os dados de entrada do problema proposto. Para a avaliação do desempenho da rede RBF, a metodologia desenvolvida foi implementada em MATLAB 2014a sobre um processador Intel Core i5-4200U CPU 1.60GHz 2.30GHz com 8 GB de RAM.

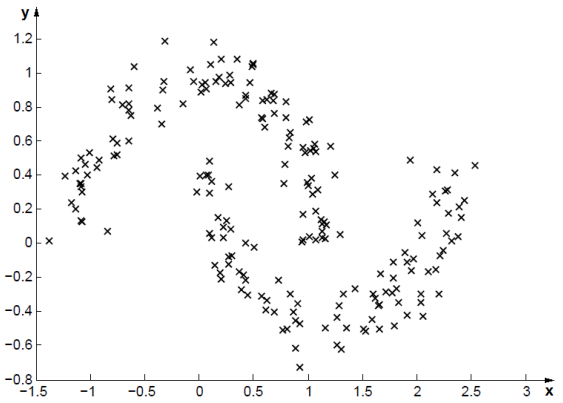

Figura 3: Data 2moons.

É abordada também a problemática de que na maioria dos casos de classificação de dados, nem sempre se tem a totalidade dos dados rotulados, tem-se apenas uma pequena parte de dados conhecidos o que faz que aumente a complexidade da classificação. Esta problemática foi abordada, e foram fornecidos 54 dados rotulados neste trabalho como é mostrado na Fig. 4. Os pontos de cor roxo indicam os dados rotulados, dos quais 30 dados pertencem a uma classe e os restantes 24 pertencem a outra classe. A ubicação desses

dados rotulados exemplificam a situação de maior complexidade onde os dados se encontram concentrados em uma região especifica ao em vez de estar uniformemente distribuídos sobre tudo o conjunto de dados.

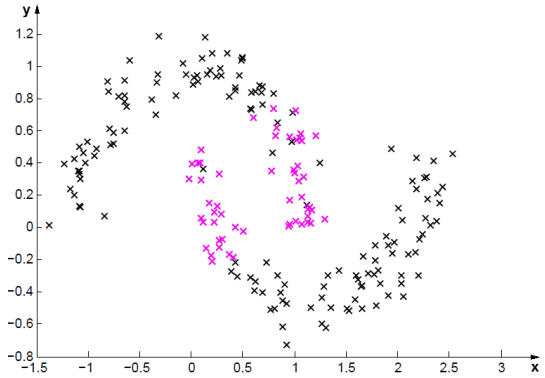

Figura 4: Problema duas luas com 54 dados rotulados.

Foram feitos vários testes com os dados rotulados, implementados em classificadores como: SVM (Support Vector Machine) usando o *kernel* RBF, onde se conseguiu fazer classificação somente dos dados rotulados. Assim, se não se tem conhecimento de uma boa porcentagem dos dados que fiquem perto da região de separação, não se consegue fazer uma boa classificação ao problema proposto. A Fig. 5 mostra a classificação feita por SVM com *kernel* RBF.

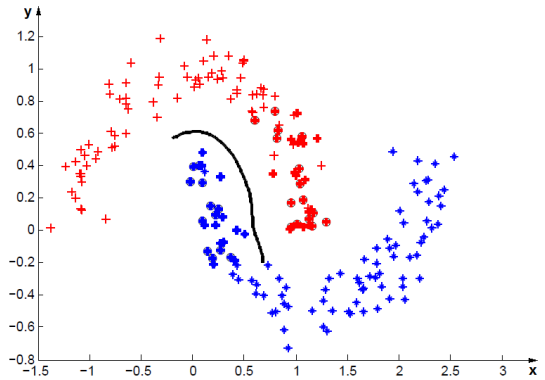

Figura 5: Classificação com *kernel* RBF - SVM.

Outro teste feito, foi a classificação dos dados usando a rede RBF com o algoritmo K-means, como se mostra na fig. 6. Neste só se consegue fazer classificação dos dados rotulados, a linha reta diagonal de cor azul, mostra a projeção da separação das duas classes classificadas, mostrando que para problemas onde não se tem uma boa porcentagem de dados rotulados, a rede não consegue classificar um problema não-linearmente separável.

A proposta de implementar o algoritmo COP-Kmeans em uma rede RBF para classificar um conjunto de dados não-linearmente separáveis onde não se tem um bom número de dados rotulados resulta uma solução factível. O algoritmo COP-Kmeans faz a escolha dos centroides usando a mistura de dados rotulados e não rotulados garantindo sempre a não violação das restrições es-

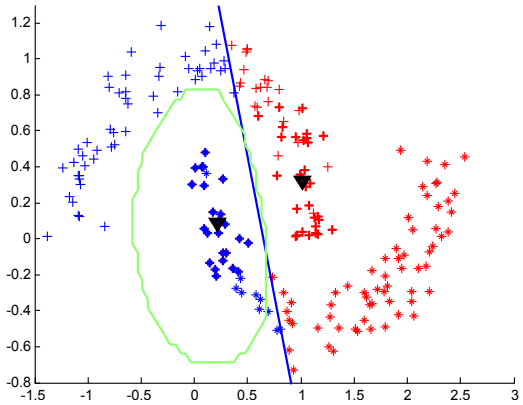

Figura 6: RBF com K-means.

tabelecidas para os dados rotulados. Então, pode-se fazer uma classificação de dados não rotulados, aproveitando o conhecimento *a priori* de dados rotulados. Na implementação da rede RBF com COP-Kmeans foram definidos para cada uma das iterações do algoritmo COP-Kmeans um número de 40 restrições entre *Cannot-Link* e *Must-Link* para os dados rotulados. Na Fig. 7 e Fig. 8 são mostradas como são atribuídas essas restrições em cada iteração.

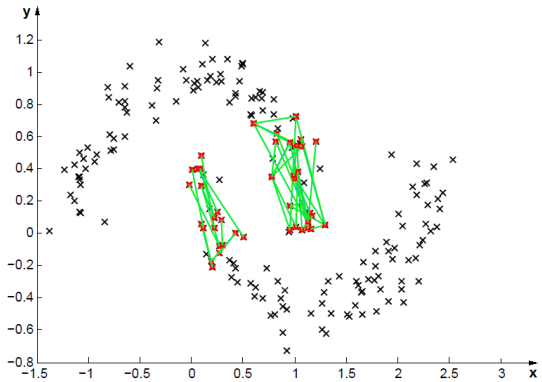

Figura 7: Restrições Must-Link.

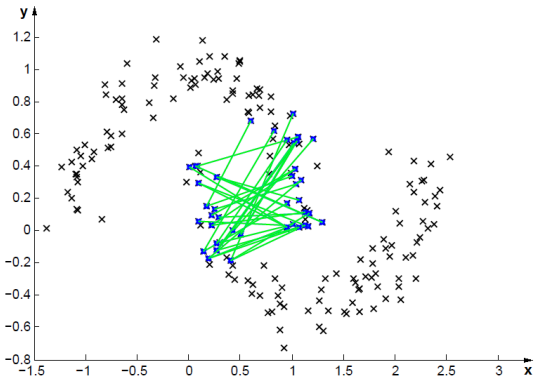

Figura 8: Restrições Cannot-Link.

O processo de treinamento da rede RBF consiste em selecionar três conjuntos de parâmetros estabelecidos que são característicos desse tipo de rede. Os centroides (μ) fornecidos pelo algoritmo COP-Kmeans, o coeficiente de beta (β) para cada um dos neurônios da camada intermediária da rede RBF, e a matriz de pesos de saída entre os neurônios da camada intermediária e os nodos de

entrada da camada de saída. Depois de usar o algoritmo COP-Kmeans para selecionar os centroides requeridos pela rede, o método para especificar os coeficientes do valor beta (β) é definir o valor de sigma (σ) igual à distância média entre todos os pontos que pertencem ao *cluster* e o centroide do *cluster*.

$$\sigma = \frac{1}{m}\sum_{i=1}^{m} \|x_i - \mu\| \tag{3}$$

O ajuste de pesos na camada de saída se converte em um problema linear, devido a que a camada intermediária tem o objetivo de transformar o conjunto de dados não linearmente separável em um conjunto linearmente separável. Foi aplicada a regra delta (WIDROW et al., 1960), como método de ajuste dos pesos desta camada de saída.

Foram feitos testes da rede RBF com COP-Kmeans para diferentes números de centroides por cada classe e foi observado que à medida que aumenta o número de centroides, a classificação obtida é melhorada consideravelmente. Na Fig. 9 é mostrada a classificação feita atribuindo só um centroide para cada classe, onde se conseguiu classificar 148 dados corretamente de 200 dados. A linha verde mostra a área de classificação feita pela rede RBF.

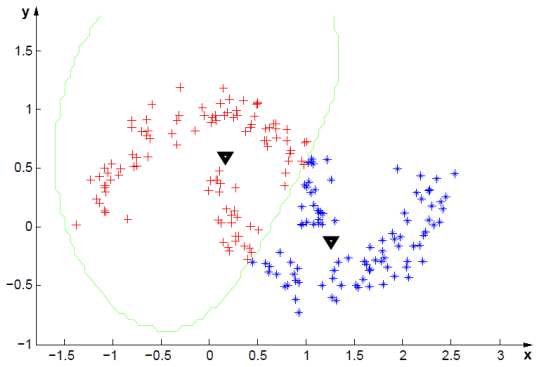

Figura 9: Rede RBF com COP-Kmeans (1 centroide).

Na Fig. 10 é mostrada a classificação feita atribuindo dois centroides para cada classe, onde se conseguiu classificar 187 dados corretamente de 200 dados. A linha verde mostra a área de classificação feita pela rede RBF.

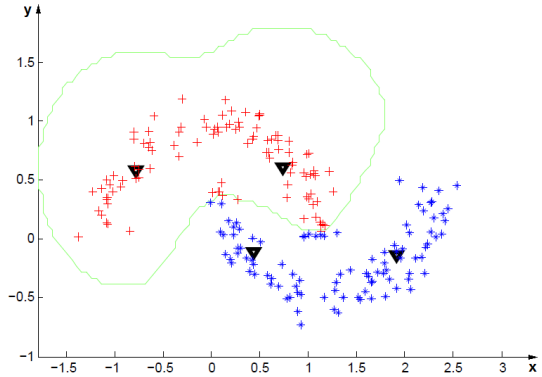

Figura 10: Rede RBF com COP-Kmeans (2 centroides).

Na Fig. 11 é mostrada a classificação feita para três centroides para cada classe, onde se conseguiu classificar 199 dados corretamente de 200 dados. A linha verde mostra a área de classificação feita pela rede RBF.

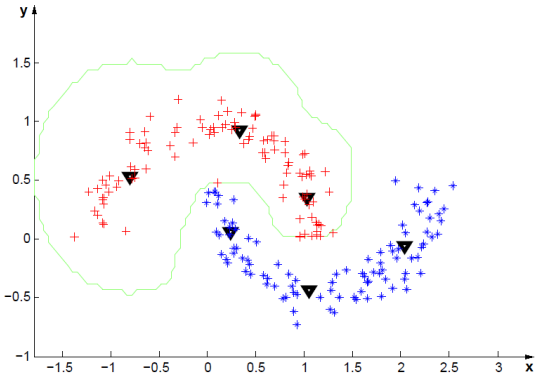

Figura 11: Rede RBF com COP-Kmeans (3 centroides).

Na Fig. 12 são mostradas as áreas de contorno das funções gaussianas para cada centroide calculado por classe, neste caso 4 centroides; onde as linhas de cor vermelho mais escuro representam as áreas com uma pontuação mais alta que os contornos com linhas mais claras, o mesmo sucede com a outra classe, o cor azul mais escuro representam as áreas com pontuação mais alta.

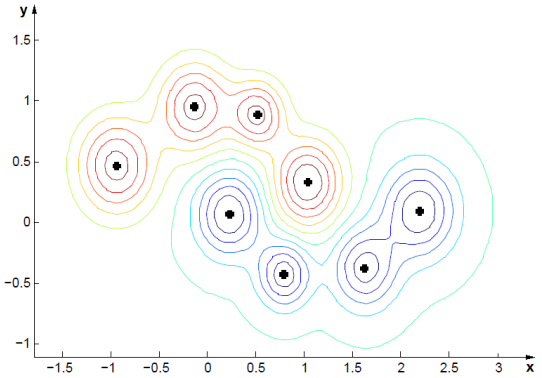

Figura 12: Áreas de contorno da função Gaussiana em cada centroide da Rede RBF com COP-Kmeans (4 centroides).

Na Fig. 13 é mostrada a classificação feita para quatro centroides para cada classe, onde se conseguiu classificar 200 dados corretamente de 200 dados. A linha verde mostra a área de classificação feita pela rede RBF.

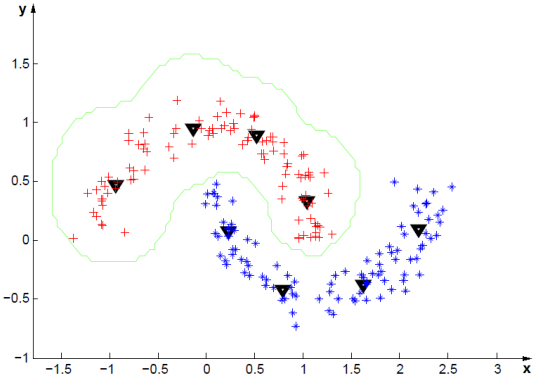

Figura 13: Rede RBF com COP-Kmeans (4 centroides).

Na tabela 1 são apresentados os valores de acurácia de diferentes centroides atribuídos a cada classe. Também, é mostrado o tempo de processamento que o classificador demora nos diferentes números de centroides utilizados por classe. É possível observar que à medida que o número de centroides aumenta, o nível de acurácia e a classificação dos dados são melhores. Além disso, o tempo de processamento aumenta progressivamente conforme aumenta o número de centroides, mas a diferença entre estes tempos não é significativa.

Tabela 1: Acurácia da rede RBF para diferentes centroides.

RBF com COP-Kmeans	Acurácia	Percentagem	Tempo de processamento (s)
1 Centroide	148/200	74%	0.3772
2 Centroides	187/200	93.5%	0.3824
3 Centroides	199/200	99.5%	0.4389
4 Centroides	200/200	100%	0.5289

5 Conclusões

Neste trabalho, o algoritmo COP-Kmeans foi descrito e implementado dentro do contexto de métodos semi - supervisionados de agrupamento de padrões baseados em restrições pontuais de ligação. Foi observado que a Rede RBF com o algoritmo COP-Kmeans avaliado é bastante versátil, intuitiva e capaz de obter níveis ótimos de acurácia.

Pode-se dizer que os níveis máximos de acurácia atingidos ao final foram satisfatórios para o problema de classificação em questão. Tudo isso qualifica esse algoritmo COP-Kmeans implementado na rede RBF como um bom complemento que ajuda consideravelmente na obtenção de uma boa aproximação na seleção de centroides nos neurônios da camada intermediária quando se tem um número limitado de dados rotulados no problema abordado. A combinação da rede RBF e COP-Kmeans resulta ser uma boa ferramenta de classificação semi-supervisionado de padrões quando o problema é não-linearmente separável.

Com o aumento de centroides por cada classe na Rede RBF, e fazendo uma boa aproximação dos centroides com a implementação do COP-Kmeans; pode-se obter uma classificação correta de conjuntos de dados onde as regiões de separação resultem ser complexas, porque com suficientes neurônios na camada intermediária é possível definir qualquer limite de decisão complexo e não-linearmente separável. Em outras palavras, sempre se pode melhorar a precisão de classificação usando mais neurônios na camada intermediária com os seus centroides respectivos.

Agradecimentos

O presente trabalho foi realizado com o apoio financeiro da FAPEMIG, CNPq e CAPES - Brasil.

Referências

Basu, S., Banerjee, A. and Mooney, R. (2002). Semi-supervised clustering by seeding, *In Proceedings of 19th International Conference on Machine Learning (ICML-2002*, Citeseer.

Basu, S., Bilenko, M. and Mooney, R. J. (2003). Comparing and unifying search-based and similarity-based approaches to semi-supervised clustering, *Proceedings of the ICML-2003 workshop on the continuum from labeled to unlabeled data in machine learning and data mining*, Citeseer, pp. 42–49.

Bilenko, M., Basu, S. and Mooney, R. J. (2004). Integrating constraints and metric learning in semi-supervised clustering, *Proceedings of the twenty-first international conference on Machine learning*, ACM, p. 11.

Grira, N., Crucianu, M. and Boujemaa, N. (2004). Unsupervised and semi-supervised clustering: a brief survey, *A Review of Machine Learning Techniques for Processing Multimedia Content* **1**: 9–16.

Jang, J.-S. R., Sun, C.-T. and Mizutani, E. (1997). Neuro-fuzzy and soft computing; a computational approach to learning and machine intelligence.

Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R. and Wu, A. Y. (2002). An efficient k-means clustering algorithm: Analysis and implementation, *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **24**(7): 881–892.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations, *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Vol. 1, Oakland, CA, USA., pp. 281–297.

Schwenker, F., Kestler, H. A. and Palm, G. (2001). Three learning phases for radial-basis-function networks, *Neural networks* **14**(4): 439–458.

Texeira, E. (2011). COP-Kmeans e Clustering Semi-Supervisionado Atraves de Restrições.

Wagstaff, K., Cardie, C., Rogers, S., Schrödl, S. et al. (2001). Constrained k-means clustering with background knowledge, *ICML*, Vol. 1, pp. 577–584.

WIDROW, B., HOFF, M. E. et al. (1960). Adaptive switching circuits.