NebFuzz: Um Novo Algoritmo de Agrupamento Semi-Supervisionado Baseado no *Fuzzy C-Means*

Valmir Macário[1], **Francisco de A. T. de Carvalho**[1]

[1]Centro de Informática, Universidade Federal de Pernambuco;
Av. Prof. Luiz Freire, s/n, Cidade Universitária, Recife-PE; Brazil; 50740-540

{vmf2,fatc}@cin.ufpe.br

Abstract. *Semi-supervised clustering uses unlabeled data, combined with the labeled data, to improve the algorithm performances. This paper presents a new algorithm for semi-supervised clustering based on Fuzzy C-Means algorithm. The new algorithm was evaluated and compared against two semi-supervised clustering algorithms in the context of learning from partially labeled data. The behavior of the proposed algorithm is discussed and the results are validated using accuracy rate, corrected rand index and a 95% confidence interval. Thus, it was possible to certify the better accuracy performance of the new algorithm when a few labeled data are available.*

Resumo. *Agrupamento semi-supervisionado utiliza dados não rotulados, juntamente com dados rotulados, com a finalidade de melhorar o desempenho dos algoritmos. Este trabalho apresenta um novo algoritmo de agrupamento semi-supervisionado baseado no algoritmo* Fuzzy C-Means*. O novo algoritmo é avaliado em relação à dois algoritmos de agrupamento semi-supervisionados a partir de dados parcialmente rotulados nas tarefas de classificação e de agrupamento. Além disso, o comportamento do algoritmo é discutido e os resultados validados com taxa de acerto, índice de Rand corrigido e intervalos com 95% de confiança. Desse modo, foi possível certificar que o novo algoritmo de agrupamento semi-supervisionado apresenta desempenho melhor quando há poucos dados rotulados disponíveis.*

1. Introdução

A aprendizagem semi-supervisionada [Chapelle et al. 2006][Zhu 2008] é uma abordagem intermediária entre a aprendizagem supervisionada e a aprendizagem não-supervisionada. A abordagem semi-supervisionada é importante devido a dificuldade de rotular dados, que pode ser uma tarefa extremamente complexa, cara, demorada e requerer especialistas humanos em algumas aplicações reais. Em contraponto, existe a facilidade de obtenção de dados não rotulados [Nigam et al. 2000]. Na aprendizagem semi-supervisionada são utilizados exemplos rotulados e não rotulados para guiar a aprendizagem com a finalidade de melhorar o desempenho dos algoritmos.

Uma das técnicas semi-supervisionadas, que despertaram grande interesse, foram as de agrupamento semi-supervisionado [Pedrycz and Waletzky 1997]. Essas técnicas são de grande interesse devido a possibilidade de incorporar dados rotulados nos algoritmos tradicionais de agrupamento não supervisionado. A tarefa de agrupamento tem sido aplicada em diversos problemas. Para citar alguns, temos a mineração de textos, agrupamento de expressões gênicas, processamento de imagens, entre outras.

O objetivo deste artigo é apresentar um novo algoritmo de agrupamento semi-supervisionado chamado de NebFuzz. O novo algoritmo demonstrou nos experimentos que é uma boa opção para tarefas com poucos dados rotulados. O algoritmo foi avaliado em duas tarefas: classificação e agrupamento. O algoritmo foi comparado com algoritmos totalmente supervisionados [Mitchel 1997] e com algoritmos de agrupamento semi-supervisionado consolidados na literatura [Pedrycz and Waletzky 1997][Bouchachia 2007].

Este artigo está estruturado da seguinte forma: a Seção 2 apresenta trabalhos relacionados de aprendizagem semi-supervisionada. A Seção 3 apresenta detalhadamente o algoritmo semi-supervisionado proposto neste trabalho. A Seção 4 descreve a metodologia dos experimentos utilizados neste trabalho. Depois, na Seção 5, os resultados dos experimentos são apresentados. Finalmente, a Seção 6 traz as conclusões e trabalhos futuros.

2. Trabalhos Relacionados

O agrupamento é a aglomeração de objetos (ou exemplos) em grupos, de modo que os objetos pertencentes ao mesmo grupo sejam mais similares entre si de acordo com alguma medida de similaridade [Hathaway et al. 2000], enquanto os objetos pertencentes a grupos diferentes tenham uma similaridade menor. O objetivo do processo de agrupamento é maximizar a homogeneidade dos objetos de um mesmo grupo enquanto maximiza a heterogeneidade entre objetos de grupos diferentes [Stepp and Michalski 1986]. Segundo Jain e Dubes [Jain and Dubes 1988], agrupamento é o estudo formal de algoritmos e métodos para agrupar exemplos que não estão rotulados numa classe correspondente.

Alguns algoritmos não supervisionados clássicos foram adaptados para levar em conta informações supervisionadas em seu treinamento. Assim, os algoritmos de agrupamento se tornaram semi-supervisionados, pois em seu treinamento utilizam informações supervisionadas e não supervisionadas. O algoritmo de agrupamento clássico que a maioria desses algoritmos se baseou foi o *Fuzzy C-Means* (FCM) [Bezdek 1981].

O Algoritmo de Pedrycz e Waletzky [Pedrycz and Waletzky 1997], que foi chamado de ISODATA com supervisão parcial (ISODATA-PS), foi um dos primeiros algoritmos que utilizou a abordagem semi-supervisionada para algoritmos de agrupamento. Pedrycz formulou um novo algoritmo modificando a função objetivo do algoritmo FCM. O termo adicionado à função objetivo do algoritmo FCM utiliza os rótulos disponíveis dos exemplos da base de dados para aumentar os graus de pertinência atribuídos pelo algoritmo para os grupos que representam a classe que o exemplo pertence. A utilização desse algoritmo foi numa base de sinais biológicos (eletrocardiogramas), onde havia poucos elementos rotulados.

O algoritmo proposto por Bouchachia e Pedrycz [Bouchachia and Pedrycz 2006] basicamente estende a função objetivo do algoritmo *Fuzzy C-Means*. O objetivo dessa mudança é capturar estruturas dos dados não rotulados e rotulados. As estruturas dos dados não rotulados são adquiridas pelo primeiro termo da função objetivo, que é igual a função objetivo do algoritmo FCM. O segundo termo leva em conta as estruturas refletidas pela avaliação dos rótulos disponíveis. Uma das vantagens desse algoritmo é que a função objetivo leva em conta possibilidade de mais de um grupo representar uma determinada classe. Aqui, esse algoritmo é identificado pela sigla SSC (*Semi-Supervised Clustering*).

3. Algoritmo NebFuzz

O algoritmo proposto neste trabalho é uma nova abordagem para algoritmos de agrupamento semi-supervisionado que acrescenta um segundo termo supervisionado à função objetivo do algoritmo FCM de Bezdek [Bezdek 1981]. A motivação para construir outro algoritmo de agrupamento semi-supervisionado foi a de que existe a hipótese que padrões pertencentes à mesma classe possuem alta similaridade. Como consequência, esses padrões estão próximos entre si, e também, o grau de pertinência do padrão que pertence ao grupo que representa sua classe tem quer ser similares. Assim, desenvolvemos uma função de otimização em que dois padrões são comparados entre si, e caso pertençam à mesma classe, a diferença entre os graus de pertinência para o grupo da classe que os dois pertencem é atenuada para que seus graus de pertinência se tornem similares. Essa manobra faz com que padrões da mesma classe possuam graus de pertinência parecidos para os C grupos. Desse modo, a função objetivo do novo algoritmo de agrupamento semi-supervisionado, dado que N padrões (inclui padrões rotulados e não rotulados) são agrupados em C grupos e existe P classes conhecidas é:

$$J(U,V) = \sum_{k=1}^{C}\sum_{i=1}^{N} u_{ik}^2 d_{ik}^2 + \alpha \sum_{k=1}^{C}\sum_{i=1}^{N}\sum_{j=1}^{N} b_i b_j t_{ij}(u_{ik} - u_{jk})^2 d_{ij}^2 \quad (1)$$

onde U representa a matriz de graus de pertinência u_{ik} do padrão $\mathbf{x_i}$ no grupo k. V representa o conjunto de protótipos $\mathbf{v_k}$ $(v_k = (v_{k1}, ..., v_{kz}))$ associado a cada grupo, onde z é o número de atributos. Existe um conjunto de dados representado pelo conjunto $X = \{\mathbf{x_1}, \mathbf{x_i}, ..., \mathbf{x_N}\}$, tal que cada padrão $\mathbf{x_i} = (x_{i1}, ..., x_{iz})$. A distância euclidiana entre o padrão $\mathbf{x_i}$ e o protótipo $\mathbf{v_k}$ é representada por d_{ik}. $b_i = 1$ e $b_j = 1$ se os padrões $\mathbf{x_i}$ e $\mathbf{x_j}$ são rotulados, e $b_i = 0$ e $b_j = 0$ caso contrário. O fator de balanceamento α indica a confiança nos rótulos disponíveis a priori. Caso os rótulos disponíveis não sejam confiáveis, esse valor deve ser próximo de 0, e se for totalmente confiável, deve ser 1. Caso α seja 0, a função objetivo se torna a mesma função do Fuzzy C-Means (FCM) com o termo de fuzzificação $m = 2$. O termo t_{ij} é uma função indicativa que fornece a informação do rótulo utilizada na função objetivo. O valor de t_{ij} é igual a 1 quando o padrão $\mathbf{x_i}$ e o padrão $\mathbf{x_j}$ pertencerem à mesma classe l, e $t_{il} = 0$, caso contrário.

$$t_{ij} = \begin{cases} 1 & se\ (\mathbf{x_i}\ and\ \mathbf{x_j})\ \in\ l \\ 0 & senão \end{cases}$$

A função objetivo obedece as restrições apresentadas na equação 2, que também são encontradas no algoritmo FCM.

$$\sum_{k=1}^{C} u_{ik} = 1\ \forall\ i,\ \ 0 < \sum_{i=1}^{N} u_{ik}\ <\ N\ \forall\ k \quad (2)$$

O multiplicador de lagrange é utilizado para otimizar a equação 1 com respeito à matriz U, obtendo as equações para o cálculo dos graus de pertinência u_{ik}, obtendo a equação 3:

$$u_{ik}^{(s)} = \frac{1 + \alpha b_i \left[\left(\sum_{h=1}^{C} \frac{\sum_{j=1}^{N} b_i b_j t_{ij}(u_{jk}^{(s-1)} - u_{jh}^{(s-1)}) d_{ij}^2}{(d_{ih}^{(s)})^2 + \alpha b_i \sum_{j=1}^{N} b_j t_{ij} d_{ij}^2}\right)\right]}{\sum_{h=1}^{C}\left[\frac{(d_{ik}^{(s)})^2 + \alpha b_i \sum_{j=1}^{N} b_j t_{ij} d_{ij}^2}{(d_{ih}^{(s)})^2 + \alpha b_i \sum_{j=1}^{N} b_j t_{ij} d_{ij}^2}\right]} \quad (3)$$

onde s é o contador de iterações. Uma das características desse cálculo é a necessidade do grau de pertinência de uma iteração anterior. Portanto, há necessidade de preencher a matriz U antes de utilizar esta fórmula pela primeira vez. A equação para o cálculo do protótipo é igual a equação utilizada no algoritmo base FCM dada pela equação 4:

$$\mathbf{v_k} = \frac{\sum_{i=1}^{N} u_{ik}^2 \mathbf{x_i}}{\sum_{i=1}^{N} u_{ik}^2} \tag{4}$$

A execução do algoritmo proposto segue o mesmo processo de algoritmos de agrupamento não supervisionados, como o FCM. Assim, o processo de agrupamento semi-supervisionado é descrito a partir do pseudo-código logo abaixo.

1 *Inicialização.* Iniciar os parâmetros do algoritmo: Fixar C, tal que $1 < C \leq N$; Fixar $MaxIter$ (número máximo de iterações); Iniciar contador de iterações $s = 0$; Atribuir um valor para ϵ, tal que $0 < \epsilon << 1$; Iniciar Matriz de Protótipos escolhendo C exemplos diferentes tal que $\mathbf{v_k} \in X = \{x_1, ..., x_N\}(k = 1, ..., C)$; Iniciar Matriz de Graus de Pertinência, incluindo todos os valores de pertinência conhecidos;

2 *Passo 1. Obtendo os melhores protótipos:* Calcular os protótipos $v_k^{(s)}$ da matriz V utilizando a equação 4;

3 *Passo 2. Obtendo as melhores partições:* Atualizar o valor de cada grau de pertinência dos padrões $\mathbf{x_i}$ a cada um dos C grupos: $u_i^{(s)} = (u_{i1}^{(s)}, ..., u_{iC}^{(s)})(i = 1, ..., N)$ da matriz de partição $U^{(s)}$ utilizando a equação 3;

4 *Critério de parada:* Calcular o valor da função objetivo $J^{(s)}$ utilizando a equação 1. **Se** $|J^{(s)} - J^{(s-1)}| \leq \epsilon$ ou $s \geq MaxIter$) e $(s > 1)$ **então** PARE. **Senão** Atualizar $s = s + 1$ e ir para 2 (Passo 1).;

Algoritmo 1: Algoritmo Semi Supervisionado NebFuzz

O algoritmo NebFuzz é uma nova abordagem de algoritmos de agrupamento semi-supervisionado. Uma das características do novo algoritmo é a otimização levando em conta apenas os padrões que são de uma mesma classe, e ainda, a utilização da distância entre estes dois padrões ao invés da distância entre o padrão e o protótipo do grupo. Ainda, o algoritmo leva em conta na função de otimização, que o número de grupos poder ser maior que o número de classes conhecidas previamente.

4. Avaliação do Algoritmo

4.1. Base de Dados

Neste artigo, apresentamos a avaliação dos algoritmos para 3 bases de dados, sendo 2 delas tiradas do repositório digital UCI (*http://archive.ics.uci.edu/ml/*) e uma base de dados sintética. As bases de dados retiradas do UCI foram: câncer de mama e iris. A base de dados sintética foi gerada utilizando uma distribuição normal com os parâmetros média e desvio padrão (μ, σ) apresentadas na Tabela 1. A primeira classe p_1 consiste de um grupo e a segunda classe p_2 consiste de dois grupos descritas por dois atributos. Cada grupo é formado por 100 padrões.

Tabela 1. Base de dados Sintética

Características		μ_1	μ_2	σ_1	σ_2
Classes	p_1	3.0	1.0	7.0	0.5
	p_2	4.0	4.5	1.0	1.0
		2.0	-2.5	1.0	1.0

4.2. Validação Cruzada

Um dos métodos para avaliar métodos de aprendizagem é a validação cruzada com k *folds* [Mitchel 1997]. Esse método consiste numa divisão da base de dados em k grupos de tamanhos aproximadamente iguais. Os padrões contidos nos grupos $k-1$ são utilizados para treinar o algoritmo e os padrões do grupo k são utilizados para testar o algoritmo. Esse processo é repetido para todas as combinações de $k-1$ grupos. O desempenho do algoritmo é medido através das médias das taxas de erro dos resultados das k repetições.

Na aprendizagem não supervisionada, quando existe uma classificação prévia disponível do conjunto de dados, a relevância estatística da diferença entre médias de taxa de erros de classificação do conjunto de testes também pode ser utilizada [Costa et al. 2003]. O conjunto de treinamento é apresentado ao método de agrupamento, o resultado é uma partição de treinamento. Depois, a técnica de centróide mais próximo é utilizada para construir um classificador a partir da partição de treinamento. A técnica de centróide mais próximo calcula a proximidade de cada padrão do conjunto de teste ao protótipo, ou o centro, de cada agrupamento formado na partição de treinamento. Desta forma, cada padrão do conjunto de teste é atribuído ao agrupamento cuja proximidade seja mínima. Depois, o conjunto de teste é comparado com a partição prévia utilizando um índice externo.

Para a aprendizagem semi-supervisionada com métodos baseados em agrupamento, adaptamos a metodologia da validação de métodos não supervisionados. A técnica de centróide mais próximo é adaptada. No trabalho original, proposto por Costa [Costa et al. 2003], o padrão pertence ao grupo cuja similaridade for maior. A similaridade é calculada utilizando a forma original do algoritmo, ou seja, pela equação do algoritmo que calcula a similaridade entre o exemplo e o protótipo. No entanto, devido a característica *fuzzy*, um exemplo pertence a todos os grupos, quantificado pelo grau de pertinência do exemplo para cada grupo. Assim, o exemplo é atribuído ao grupo cujo grau de pertinência for maior.

4.3. Taxa de Acerto

A saída obtida com métodos de agrupamento é composta por estruturas como partições ou hierarquias, que não podem ser diretamente utilizadas para classificar outros objetos. Assim, é importante saber qual grupo representa cada classe para confirmar se o algoritmo atribuí o padrão para a classe correta. Depois dos grupos formados pelo algoritmo, uma matriz de confusão no modelo da Tabela 2 é construída. Seja $V = v_1, ..., v_k, ..., v_C$ o conjunto de C grupos, $P = p_1, ..., p_l, ..., p_H$ o conjunto de H classes e a entrada n_{kl}, o número de objetos que estão tanto no grupo k quanto na classe p_l. O termo $n_{l\bullet}$ é a soma dos padrões contidos na classe p_l e $n\bullet k$ representa o número de objetos no grupo k e n é o número total de objetos nas partições. A matriz de confusão é caracterizada pelas colunas que representam o grupo atribuído pelo algoritmo e a linha, que representa o grupo ao qual

Tabela 2. Tabela de confusão

Classes	Grupos					
	v_1	$\cdots$	v_k	$\cdots$	v_C	$\sum$
p_1	n_{l1}	$\cdots$	n_{1k}	$\cdots$	n_{1C}	$n_{1\bullet} = \sum_{k=1}^{C} n_{1k}$
$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
p_l	n_{l1}	$\cdots$	n_{lk}	$\cdots$	n_{lC}	$n_{l\bullet} = \sum_{k=1}^{C} n_{lk}$
$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
p_H	n_{H1}	$\cdots$	n_{Hk}	$\cdots$	n_{HC}	$n_{H\bullet} = \sum_{k=1}^{C} n_{Hk}$
$\sum$	$n_{\bullet 1} = \sum_{l=1}^{H} n_{l1}$	$\cdots$	$n_{\bullet k} = \sum_{l=1}^{H} n_{lk}$	$\cdots$	$n_{\bullet C} = \sum_{l=1}^{H} n_{lC}$	$n = \sum_{l=1}^{H} \sum_{k=1}^{C} n_{lk}$

o padrão pertence originalmente. A taxa de acerto é obtida pela equação 5. O conjunto π_l contém todos os grupos que representam a classe p_l. Um grupo k representa apenas uma classe. A classe representada será aquela que tiver mais objetos neste grupo k.

$$Tx = \frac{\sum_{k \in \pi_l}^{C} \max_{1 \leq l \leq H} n_{lk}}{n} \tag{5}$$

4.4. Índice Externo - Índice de Rand Corrigido

É preciso medir a qualidade da partição gerada pelo algoritmo de agrupamento semi-supervisionado. Índices externos são utilizados para avaliar o grau de concordância entre duas partições (U e V), onde a partição U é o resultado do método de agrupamento e a partição V é gerada por uma informação prévia, independente da partição U, como uma classe[Jain and Dubes 1988].

O índice de Rand corrigido [Costa et al. 2003] possui seus valores corrigidos de acordo com acertos nas comparações entre as partições.O índice de Rand corrigido pode assumir valores em [-1,1], onde o valor 1 indica uma coesão perfeita entre as partições, enquanto valores próximos de 0 (ou negativos) correspondem a uma coesão encontrada por acaso. A equação 6 apresenta o índice corrigido de Rand levando em conta a matriz de confusão apresentada na Tabela 2.

$$CR = \frac{\sum_{l}^{H} \sum_{k}^{C} \binom{n_{lk}}{2} - \binom{n}{2}^{-1} \sum_{l}^{H} \binom{n_{l\bullet}}{2} \sum_{k}^{C} \binom{n_{\bullet k}}{2}}{\frac{1}{2}[\sum_{l}^{H} \binom{n_{l\bullet}}{2} + \sum_{k}^{C} \binom{n_{\bullet k}}{2}] - \binom{n}{2}^{-1} \sum_{l}^{H} \binom{n_{l\bullet}}{2} \sum_{k}^{C} \binom{n_{\bullet k}}{2}} \tag{6}$$

5. Experimentos e Resultados

Foram executados 2 tipos de experimentos com algoritmos de agrupamento semi-supervisionado nas tarefas de agrupamento e classificação. Os experimentos foram baseados nos trabalhos de Pedrycz e Waletzky [Pedrycz and Waletzky 1997] e de Amini e Gallinari [Amini and Gallinari 2005]. Nesses dois experimentos, os padrões da base de dados são divididos em 2 conjuntos, o conjunto de dados rotulados e o conjunto de dados não rotulados. O conjunto de dados rotulados contém uma certa porcentagem de exemplos rotulados que varia nas seguintes proporções: $(0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100)$, do total da base. O restante do conjunto de dados faz parte do segundo conjunto, o de dados não rotulados. Assim, para os valores respectivos de base rotulada, esses conjuntos utilizam as porcentagens $(100, 90, 80, 70, 60, 50, 40, 30, 20, 10, 0)$ do total da base. Os experimentos são realizados submetendo os dois conjuntos de dados ao algoritmo de

agrupamento semi-supervisionado com as devidas proporções de cada um dos conjuntos. Desse modo, observamos a influência da quantidade de dados rotulados e não rotulados no desempenho dos algoritmos. Os algoritmos comparados foram o algoritmo NebFuzz, o algoritmo ISODATA-PS [Pedrycz and Waletzky 1997] e o algoritmo SSC [Bouchachia and Pedrycz 2006].

O experimento de tarefa de classificação com algoritmos semi-supervisionados emprega a validação cruzada com 10 *folds* para avaliar os algoritmos estudados. A validação cruzada emprega dois conjuntos de dados separados, um para treinamento e outro para teste. Neste experimento foram avaliadas 3 bases de dados: câncer de mama, iris e sintética. A validação cruzada foi repetida 30 vezes. Para cada repetição, a inicialização dos padrões foi repetida 20 vezes para que os parâmetros de inicialização do algoritmo tenham pouca influência no resultado final do algoritmo. Dessas 20 repetições, o resultado escolhido é o da iteração que conseguir obter o menor valor da função objetivo. Desse modo, são gerados os valores dos quais obtemos a média e o desvio padrão para cada uma das configurações do experimento e construímos intervalos com 95% de confiança com a finalidade de comparar os algoritmos estudados.

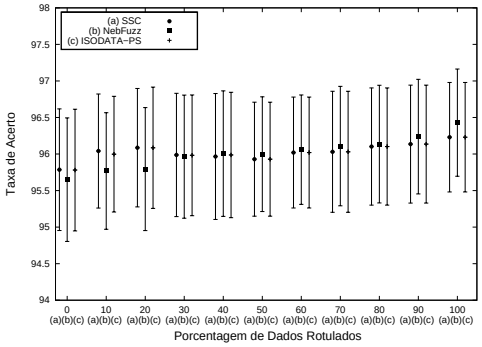
Os resultados dos experimentos para a tarefa de classificação são apresentados na Tabela 3. Na tarefa de classificação os 3 algoritmos tiveram desempenhos muito parecidos para a maioria das configurações de dados rotulados principalmente nas bases de dados câncer de mama e iris. Mesmo com mais dados rotulados, o aumento do desempenho não foi muito alto na base de dados de câncer de mama. Na base de dados iris, o algoritmo proposto NebFuzz obteve o maior ganho no aumento de dados rotulados, também teve o melhor desempenho com todos os dados rotulados. Os algoritmos SSC e ISODATA-PS tiveram um ganho de 2% com a adição de dados rotulados. Na base de dados sintética, o algoritmo proposto obteve os melhores resultados em todas configurações, exceto com 100%, onde não obteve a taxa de acerto máxima. Os algoritmos ISODATA-PS e SSC tiveram desempenhos parecidos nessa base de dados.

Tabela 3. Média da taxa de acerto na tarefa de classificação nas bases de dados câncer de mama, iris e sintética com porcentagens de 0% a 100% de dados rotulados

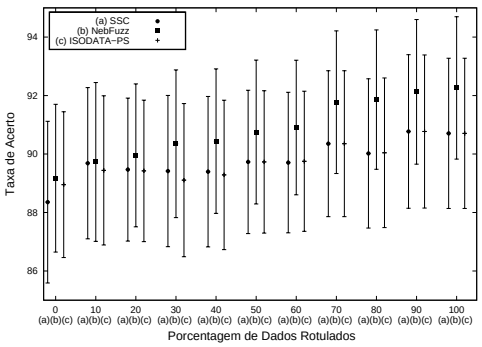
	Câncer			Iris			Sintética		
P	SSC	NebFuzz	ISODATA-PS	SSC	NebFuzz	ISODATA-PS	SSC	NebFuzz	ISODATA-PS
0	95.79	95.65	95.78	88.35	89.17	88.95	79.00	80.70	79.00
10	96.04	95.77	96.00	89.69	89.73	89.44	79.30	82.09	78.70
20	96.09	95.79	96.09	89.47	89.95	89.42	80.91	84.99	79.64
30	95.99	95.96	95.98	89.42	90.35	89.11	85.75	88.40	80.73
40	95.97	96.01	95.99	89.40	90.44	89.29	89.00	91.36	86.70
50	95.93	96.00	95.93	89.73	90.75	89.73	91.92	93.31	91.65
60	96.02	96.06	96.02	89.71	90.91	89.75	93.68	94.50	93.68
70	96.03	96.11	96.03	90.35	91.77	90.35	95.53	95.97	95.53
80	96.10	96.14	96.10	90.02	91.86	90.04	97.30	97.56	97.30
90	96.14	96.24	96.14	90.77	92.13	90.77	98.57	97.81	98.57
100	96.23	96.43	96.23	90.71	92.26	90.71	100.0	97.67	100.0

Os intervalos de confiança do experimento de classificação são apresentados na Figura 1. Podemos perceber quanto os resultados foram equilibrados para praticamente todas as configurações do experimento nas bases de dados de câncer de mama e iris. Na base de dados sintética, o algoritmo proposto obteve os melhores resultados com poucos dados rotulados até 60% de dados rotulados. Os algoritmos SSC e ISODATA-PS tive-

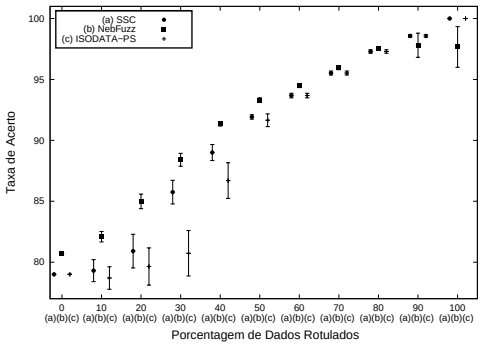
ram desempenhos parecidos, exceto para 30% de dados rotulados onde o algoritmo SSC obteve o segundo melhor desempenho.



(a) Intervalos com 95% de confiança para a base de dados câncer de mama



(b) Intervalos com 95% de confiança para a base de dados iris



(c) Intervalos com 95% de confiança para a base de dados sintética

Figura 1. Intervalos com 95% de confiança para as base de dados câncer de mama, iris e sintética na tarefa de classificação

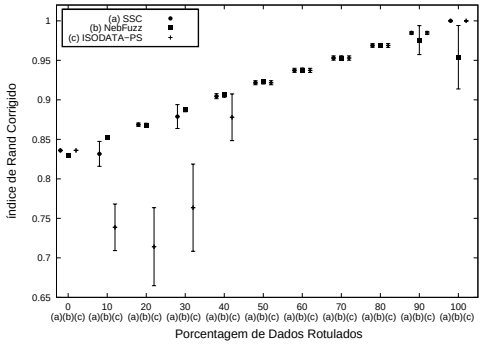
O experimento da tarefa de agrupamento possui como principal característica a utilização de toda a base de dados disponível para o treinamento e para o teste do algoritmo, respeitando as configurações já apresentadas dos conjuntos de dados rotulados e não rotulados. O experimento é repetido 100 vezes para obter uma significância estatística. Para cada iteração do experimento, a inicialização dos padrões foi repetida 20 vezes para que o melhor resultado dessas repetições seja selecionado da iteração cuja função objetivo obter o menor valor em relação aos demais. Aqui, utilizamos o índice de Rand corrigido para validar o desempenho dos algoritmos. Depois, intervalos com 95% de confiança são construídos para comparar o desempenho dos algoritmos de agrupamento semi-supervisionado na tarefa de agrupamento. Apresentamos apenas os resultados do índice de Rand corrigido para as bases de dados câncer de mama, iris e sintética desse experimento na Tabela 4.

O algoritmo proposto NebFuzz obteve os melhores resultados nas bases de dados iris e sintética principalmente quando havia poucos dados rotulados, 0% a 50%. Na base de dados de câncer de mama, os 3 algoritmos tiveram desempenhos muito parecidos para praticamente todas as configurações de dados rotulados do experimento. Na base de dados iris, o algoritmo SSC obteve um resultado melhor que o algoritmo ISODATA-PS.

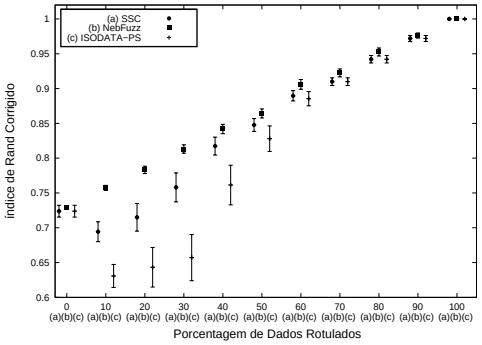
Tabela 4. Média do índice de Rand corrigido global na tarefa de agrupamento nas bases de dados câncer de mama, iris e sintética com porcentagens de 0% a 100% de dados rotulados

	Câncer			Iris			Sintética		
P	SSC	NebFuzz	ISODATA-PS	SSC	NebFuzz	ISODATA-PS	SSC	NebFuzz	ISODATA-PS
0	0.84	0.83	0.84	0.72	0.73	0.72	0.29	0.34	0.29
10	0.83	0.85	0.74	0.69	0.76	0.63	0.32	0.38	0.31
20	0.87	0.87	0.71	0.72	0.78	0.64	0.38	0.47	0.35
30	0.88	0.89	0.76	0.76	0.81	0.66	0.50	0.58	0.37
40	0.90	0.91	0.88	0.82	0.84	0.76	0.60	0.68	0.54
50	0.92	0.92	0.92	0.85	0.86	0.83	0.69	0.74	0.69
60	0.94	0.94	0.94	0.89	0.91	0.89	0.76	0.79	0.76
70	0.95	0.95	0.95	0.91	0.92	0.91	0.83	0.84	0.83
80	0.97	0.97	0.97	0.94	0.95	0.94	0.89	0.90	0.89
90	0.98	0.98	0.98	0.97	0.98	0.97	0.94	0.92	0.94
100	1.00	0.95	1.00	1.00	1.00	1.00	1.00	0.93	1.00

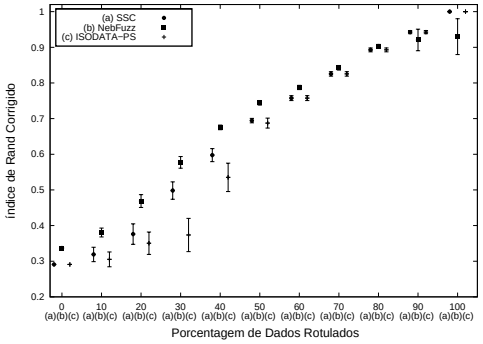
Na base de dados sintética, o algoritmo SSC obteve o segundo melhor desempenho até 40% de dados rotulados, depois os algoritmos ISODATA-PS e SSC tiveram desempenhos similares.



(a) Intervalos com 95% de confiança para a base de dados câncer de mama



(b) Intervalos com 95% de confiança para a base de dados iris



(c) Intervalos com 95% de confiança para a base de dados sintética

Figura 2. Intervalos com 95% de confiança para as base de dados câncer de mama, iris e sintética na tarefa de agrupamento

Intervalos com 95% de confiança para o índice de Rand corrigido são apresentados na Figura 2. Na base de dados de câncer de mama, note que exceto para 10% a 30% de dados rotulados onde o algoritmo ISODATA-PS obteve o pior desempenho, os 3 algo-

ritmos tiveram desempenhos similares. Para a base de dados iris, o algoritmo NebFuzz obteve o melhor resultado com poucos dados rotulados, 10% a 40% de dados rotulados, ficando o algoritmo SSC com o segundo melhor desempenho e o ISODATA-PS com o terceiro melhor. Nas outras configurações os 3 algoritmos empataram em termos de desempenho. Na base de dados sintética, o algoritmo NebFuzz se destacou quando havia poucos dados rotulados, até 60% de dados rotulados disponíveis. Com mais de 60% de dados rotulados, os 3 algoritmos empataram em termos de desempenho. Os algoritmos SSC e ISODATA-PS conseguiram desempenhos parecidos, exceto para 30% e 40% de dados rotulados, onde o algoritmo SSC foi melhor que o algoritmo ISODATA-PS.

Por fim, podemos destacar que o algoritmo NebFuzz de agrupamento semi-supervisionado proposto obteve bons resultados para a maioria das base de dados avaliadas nesse trabalho, principalmente quando havia poucos dados rotulados disponíveis. Também, podemos afirmar que o aumento de dados rotulados disponíveis no treinamento do algoritmo aumenta seu desempenho na grande maioria dos casos.

6. Conclusões

Este trabalho propôs um novo algoritmo de agrupamento semi-supervisionado, chamado NebFuzz, que alcança bons resultados com poucos dados rotulados. Os resultados alcançados pelo algoritmo NebFuzz foram bastante satisfatórios. Para poucos dados rotulados, o novo algoritmo de agrupamento semi-supervisionado obteve o melhor ou esteve entre os melhores desempenhos em relação aos outros algoritmos de agrupamento semi-supervisionado. À medida que mais dados eram disponibilizados para seu treinamento, o algoritmo melhorou seu desempenho na grande maioria dos casos, alcançando desempenho compatível com outros algoritmos semi-supervisionados. Assim, o algoritmo NebFuzz, fruto deste trabalho, pode ser utilizado em aplicações reais onde rotular dados seja custoso.

Como objetivo futuro, pretende-se realizar alterações nessa função de similaridade. As distâncias adaptativas permitem a construção de partições em diversos formatos, além do padrão circular, padrão formado pela distância euclidiana, e desse modo pode aprender estruturas mais complexas de distribuições de dados. Assim, pretende-se utilizar distâncias adaptativas com a finalidade de melhorar o desempenho do algoritmo proposto.

Agradecimentos.

Gostaríamos de agradecer pelo suporte financeiro das agências brasileiras CNPq e FACEPE.

Referências

Amini, M. R. and Gallinari, P. (2005). Semi-supervised learning with an imperfect supervisor. *Knowledge and Information Systems*, 8:385–413.

Bezdek, J. (1981). *Pattern Recognition With Fuzzy Objective Function Algorithms*. Plenum.

Bouchachia, A. (2007). Learning with parly data. *Neural Computing and application*, (16):267–293.

Bouchachia, A. and Pedrycz, W. (2006). Data clustering with partial supervision. *Data Mining and Knowledge Discovery*, (12):47–78.

Chapelle, O., Zien, A., and Scholkopf, B. (2006). *Semi-supervised learning*. MIT Press.

Costa, I. G., Carvalho, F. A. D., and de Souto, M. C. (2003). Comparative study on proximity indices for cluster analysis of gene expression time series. *Journal of Intelligent & Fuzzy Systems*, 13:133–142.

Hathaway, R. J., Bezdek, J., and Hu, Y. (2000). Generalized fuzzy c-means clustering strategies using lp-norm distances. *IEEE Transaction on Fuzzy Systems*, 8(5):576–582.

Jain, A. K. and Dubes, R. C. (1988). *Algorithms for clustering data*. Prentice Hall, New Jersey.

Mitchel, T. (1997). *Machine Learning*. McGraw Hill.

Nigam, K., McCallum, A. K., Thrun, S., and Mitchell, T. (2000). Text classification from labeled and unlabeled documents using em. *Machine Learning*, 39:103–134.

Pedrycz, W. and Waletzky, J. (1997). Fuzzy clustering with partial supervision. *IEEE transactions on system, man and cybernetics*, 27(5).

Stepp, R. E. and Michalski, R. S. (1986). *Machine Learning: An Artificial Intelligence Approach*, volume 2, chapter Conceptual Clustering: Inventing Goal-Oriented Classifictions of Structured Objects, pages 471–478. Morgan Kaufmann.

Zhu, X. (2008). *Semi-Supervised Learning Literature Survey*. Carnegie Mellon University.