

# Analysis of the Influence of Distance Metrics on the Semi-supervised Algorithm of Particle Competition and Cooperation

Lucas Guerreiro, Fabricio Aparecido Breve

Departamento de Estatística, Matemática Aplicada e Computação (DEMAC)  
Universidade Estadual Paulista (UNESP)  
Rio Claro, Brasil  
guerreioluc@gmail.com, fabricio@rc.unesp.br

**Abstract**— Machine Learning is an increasing area over the last few years and it is one of the highlights in Artificial Intelligence area. Nowadays, one of the most studied areas is Semi-supervised learning, mainly due to its characteristic of lower cost in labeling sample data. The most active category in this subarea is that of graph-based models. The Particle Competition and Cooperation in Networks algorithm is one of the techniques in this field, which has always used the Euclidean distance to measure the similarity between data and to build the graph. This project aims to implement the algorithm and apply other distance metrics in it, over different datasets. Thus, the results on these metrics are compared to analyze if there is such a metric that produces better results, or if different datasets require a different metric in order to obtain a better correct classification rate.

**Keywords**—artificial intelligence; semi-supervised learning; distance metrics; graphs

## I. INTRODUÇÃO

Aprendizado de máquina (AM) é um campo da inteligência artificial, aplicado em diversos segmentos. O AM busca desenvolver algoritmos e técnicas baseadas em conhecimento, de forma análoga a que os humanos fazem ao adquirir novos conhecimentos, com base em exemplos anteriores [1][2]. Dentro de AM podemos citar a categoria de aprendizado semi-supervisionado. Nesta classe de técnicas, faz-se uso de apenas uma pequena porção de dados rotulados, com a maioria das amostras consistindo de dados não rotulados. O objetivo é obter uma classificação eficiente fazendo uso de ambos os dados rotulados e não rotulados, visto que rotular dados costuma ser uma tarefa custosa e demorada. Dentro da categoria de algoritmos semi-supervisionados, podemos citar a classe de algoritmos baseados em grafos, sendo esta a área mais ativa neste tipo de aprendizado [3]. Nota-se nesta classe o algoritmo de Competição e Cooperação entre Partículas [4], o qual foi objeto de estudos deste trabalho. A premissa deste algoritmo é rotular uma pequena parte dos nós do grafo que é formado inicialmente e distribuir partículas por estes nós. Estas partículas passam a caminhar no grafo e propagam os rótulos das classes de seus respectivos nós de origem. Partículas de mesma classe cooperam para ganhar territórios, reforçando sua classe em nós pelos quais caminham. Partículas de classes diferentes competem por territórios, tentando reforçar o

domínio de sua classe nos nós do grafo e enfraquecendo o domínio das demais partículas sobre os nós em que caminham; ao visitar um nó que é dominado por outra classe, a partícula perde força, não sendo mais tão eficiente ao visitar um nó de outra classe, por outro lado a partícula ganha força ao visitar um nó de sua classe. Ao fim da execução das iterações do algoritmo, os nós são classificados pela classe que tem o maior domínio ganho pela caminhada das partículas sobre os mesmos. Assim como em diversos algoritmos baseados em grafos, a estrutura do grafo formado por este algoritmo depende da relação entre os nós, mais especificamente, da distância entre a representação de cada exemplo de dados. O algoritmo desde sua concepção se utilizou da distância Euclidiana para inferir tal estrutura.

O projeto tem por principal objetivo estudar como o uso de outras métricas de distância na formação do grafo pode influenciar nos resultados de classificação. São utilizadas diferentes bases de dados para analisar a influência do uso de cada medida em cada uma delas.

## II. MÉTRICAS DE DISTÂNCIA

Distância, neste contexto, representa o grau de similaridade entre dois elementos, quanto menor a distância entre dois elementos, mais similares estes são. Tais medidas podem ser aferidas de diferentes formas, gerando resultados diferentes. As métricas de distância aplicadas neste projeto são apresentadas abaixo.

### A. Distância Euclidiana

A distância Euclidiana é uma das distâncias mais comuns encontradas na literatura para uso em classificações, inclusive sendo a medida padrão utilizada no algoritmo de Competição e Cooperação entre Partículas. Com o uso de tal medida, o espaço euclidiano é convertido em espaço métrico [5]. Segue a fórmula da distância Euclidiana na equação 1.

$$d(x, y) = \sqrt{\sum_i^n (x_i - y_i)^2} \quad (1)$$

### B. Distância de Mahalanobis

A aplicação da distância de Mahalanobis visa minimizar possíveis distorções que podem ocorrer pelas correlações lineares entre as características dos exemplos de dados analisados. Esta medida de distância se utiliza da correlação entre as variáveis de cada exemplo de dado, mantendo cada característica com média zero e variância um.

A distância de Mahalanobis para um determinado conjunto de dados, em que  $x$  e  $y$  correspondem aos vetores de dados sendo comparados, sendo  $\mathbf{S}$  a matriz de covariância do conjunto de dados, pode ser definida como [5]:

$$d(\vec{x}, \vec{y}) = \sqrt{(\vec{x} - \vec{y})^T \mathbf{S}^{-1} (\vec{x} - \vec{y})} \quad (2)$$

### C. Distância City Block

A distância City Block (também conhecida como distância de Manhattan) parte do princípio de ser sempre maior ou igual a zero, sendo igual a zero quando os exemplos de dados são idênticos. É uma medida de cálculo de distância similar à Euclidiana, porém não elevamos as diferenças ao quadrado, permitindo que uma alta distância entre dois pontos (ou variáveis de um exemplo) únicos não tenham uma grande influência no valor de distância final [5]. A fórmula da distância City Block segue abaixo.

$$d(x, y) = \sum_{i=1}^n |x_i - y_i| \quad (3)$$

### D. Distância de Chebyshev

A distância de Chebyshev leva em consideração a maior distância entre um par de variáveis. Com isso, para cada par de amostra de dados, compara-se a diferença entre cada par de atributos, e o maior valor de diferença entre estes elementos é tomado como a distância entre as amostras de dados [6]. Segue fórmula que representa esta métrica de distância.

$$d(x, y) = \max_i |x_i - y_i| \quad (4)$$

### E. Distância de Minkowski

A distância de Minkowski recai no plano euclidiano, sendo uma forma mais genérica da própria distância euclidiana. Pode-se observar a fórmula na equação 5.

$$d(x, y) = \sqrt[p]{\sum_{i=1}^n |x_i - y_i|^p} \quad (5)$$

O fator que diferencia esta distância para a Euclidiana está no valor de  $p$ , que conforme for escolhido afeta o cálculo da distância. Observa-se ainda que para o caso de  $p = 2$  temos exatamente a fórmula da distância Euclidiana, e ainda quando  $p = 1$  temos a fórmula da distância de City Block. Além disso,

se  $p = \infty$  temos a distância de Chebyshev, pois a maior diferença dominaria a distância total [5].

### F. Distância de Bray-Curtis

A medida de distância de Bray-Curtis é largamente aplicada na área biológica para medir a dissimilaridade entre exemplos do meio ecológico vindos de duas localidades diferentes. O valor recai sempre entre 0 e 1, avaliando-se, no meio ecológico, que um valor de 0 representa total similaridade, ou seja, ambos locais têm as mesmas espécies e 1 quando não têm qualquer espécie em comum entre eles. Podemos utilizar esta métrica para calcular a distância entre demais exemplos de dados com múltiplas variáveis, sendo aplicado neste trabalho [7]. Nota-se a fórmula na equação 6.

$$d(x, y) = \frac{\sum |x_i - y_i|}{\sum (x_i + y_i)} \quad (6)$$

Basicamente, calcula-se a diferença entre cada par de variáveis e soma-se o valor destas diferenças (em módulo) e divide-se pela soma de todos os valores dos dados.

### G. Distância de Canberra

A distância de Canberra é aplicada em casos em que os dados estão dispersos da origem, sendo similar à distância de City Block, porém na distância de Canberra divide-se cada diferença pela soma dos valores das variáveis. Em comparação a City Block, a distância de Canberra é menos influenciada por variáveis de valores altos, devido ao denominador que reduz esse impacto [8]. Segue a fórmula desta métrica na equação 7.

$$d(x, y) = \sum_{i=1}^n \frac{|x_i - y_i|}{|x_i| + |y_i|} \quad (7)$$

## III. EXPERIMENTOS REALIZADOS

Este trabalho propõe o estudo da aplicação de novos cálculos de distância para a definição do grafo com os nós do conjunto de dados. O modelo de Competição e Cooperação de Partículas sempre se utilizou da distância Euclidiana para esta definição. Neste projeto é discutido o uso das distâncias descritas na seção anterior como alternativas. Conforme discutido, no passo de construção do grafo, a proximidade entre os nós e como essa similaridade é medida têm alta relação com a eficiência do algoritmo, já que altera a caminhada que as partículas virão a percorrer, e então competir e cooperar por territórios. Para avaliar a influência desta alteração, o algoritmo foi aplicado e comparado nas bases Iris, Wine e Banknote Authentication. A base Iris contém 150 instâncias, com 4 variáveis em cada e um total de 3 classes. A base Wine possui 178 instâncias, com 13 variáveis em cada e um total de 3 classes. A base Banknote Authentication possui 1372 instâncias, com 4 variáveis em cada e um total de 2

classes. Todos estes conjuntos de dados estão disponíveis no UCI Machine Learning Repository [9].

O primeiro passo do desenvolvimento do projeto foi a implementação do algoritmo [4], conforme descrito em sua totalidade, sem quaisquer alterações. O programa foi implementado em linguagem C. Como o algoritmo é estocástico devido à escolha das partículas e dos nós escolhidos nas caminhadas, para cada experimento, executa-se o programa 100 vezes como forma de diminuir este efeito, e a média dos resultados e seu desvio-padrão são anotados. Cada experimento consiste na escolha de uma base e uma métrica de distância, podendo inferir a resposta para o principal objetivo deste trabalho, o qual é encontrar em quais casos determinada métrica pode funcionar melhor.

A base Iris foi utilizada conforme fornecida, não necessitando de qualquer ajuste em seus parâmetros. Já as bases Wine e Banknote Authentication não estão normalizadas, sendo aplicado no algoritmo uma correção em cima de seus valores iniciais. A normalização se deu da seguinte forma: calculou-se a média aritmética dos valores em cada variável, considerando todas as amostras, bem como seu desvio padrão; aplicou-se então a normalização em cada variável de cada item de dados, subtraindo-se o valor do atributo pela média daquela variável e dividindo-se o resultado pelo desvio padrão, o valor resultante passa a representar aquele atributo no exemplo de dado. Contudo, os experimentos nas bases Wine e Banknote Authentication foram feitos tanto para os valores absolutos como os normalizados, como base de comparação.

Outro ponto a se ressaltar nos experimentos foi a métrica Minkowski, como descrito anteriormente, esta métrica é sensível ao fator  $p$ . Como já estamos aplicando  $p = 1$ ,  $p = 2$  e  $p = \infty$ , optou-se por utilizar  $p = 4$  como forma de observar os resultados nesta de distância. Portanto, sempre que citamos os resultados na medida Minkowski, nos referimos a aplicação da métrica com um fator de  $p = 4$ . O fundamento para uso de  $p = 4$  é de fins experimentais para avaliação de uma métrica diferente.

Por fim, destacamos que os resultados foram avaliados em termos de taxa de classificação e desvio-padrão, tomamos como melhor resultado aquele vindo da maior taxa de classificação no algoritmo. Porém, para diferentes fins, o desvio-padrão ou outro fator de comparação poderia ser considerado.

#### A. Parâmetros

Conforme descrito acima, o algoritmo se torna sensível às escolhas de parâmetros. Este trabalho tem por objetivo obter bons resultados de classificação e avaliar a influência de diferentes métricas de distâncias nestes resultados. Com isso, ao ter-se o algoritmo reproduzido, o primeiro passo foi a definição dos parâmetros. Para tal fim, foi escolhido o conjunto de dados Iris, por se tratar de um modelo genérico nos estudos de classificação, ser uma base pequena e com dados já normalizados; como o objetivo é a avaliar a influência da aplicação de diferentes métricas de distância, o uso de tal base teve apenas o objetivo demonstrativo para que os parâmetros fossem escolhidos. Para esta fase de definição de parâmetros, fez-se uso da métrica de distância Euclidiana por se tratar da

métrica padrão do algoritmo. Os parâmetros foram variados um por vez dentro de determinada escala e enquanto os demais parâmetros estavam fixados. Os principais parâmetros do algoritmo são a taxa de aprendizagem ( $\Delta v$ ), número de partículas ( $p$ ), número de vizinhos ( $k$ ), número de iterações ( $z$ ) e taxa de caminhada gulosa ( $p_{grad}$ ). O número de iterações é o de menor sensibilidade quando usado um número alto e reproduz uma boa classificação, então foi fixado um número de 100.000 iterações por partícula, sem perca de desempenho do programa. O número de partículas tem influência direta nos resultados finais, quanto maior seu número, melhor a classificação; porém, algoritmos semi-supervisionados focam em manter este número baixo o suficiente para diminuir o esforço humano e ainda rotular de forma eficiente, com isso e conforme sugerido na literatura, fixou-se o número de partículas em 10%, número razoável e de boa taxa de classificação. Restou-se assim a definição dos parâmetros  $\Delta v$ ,  $k$  e  $p_{grad}$ . Usando a técnica de variar um dos parâmetros e fixar-se os demais, os resultados e testes podem ser vistos nesta subseção.

O primeiro parâmetro a ser definido foi a taxa de aprendizagem, a mesma foi variada entre 0,05 e 0,50, com intervalos de 0,05, valores maiores não foram usados para evitar um aprendizado muito alto que descaracterizasse o modelo aplicado, além de se identificar de imediato que se passa a piorar a taxa de classificação com tais valores com o aumento da taxa. Fixou-se também, para isso, o número de vizinhos de cada nó em 10 e a taxa de caminhada gulosa em 0,5. Os resultados dos testes podem ser vistos na Fig. 1.

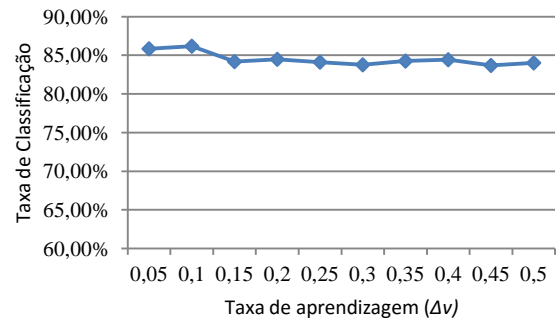


Fig. 1. Gráfico da classificação pela taxa de aprendizagem,  $\Delta v$ .

Observa-se que a taxa em que obtivemos os melhores resultados foi de 0,1. Nota-se ainda que a partir de 0,1 a taxa de classificação começa a decair, mostrando que conforme a fórmula de atualização de domínio, uma taxa alta tende a decrementar o novo valor de domínio no nó. Com base nisso a taxa de aprendizagem adotada foi de 0,1.

O próximo parâmetro a ser explorado foi o número de vizinhos. O número  $k$  variou entre 5 e 20, ou em termos de proporção sobre  $n$  variou entre 3% e 13%. Os demais parâmetros foram mantidos conforme o teste anterior. Os resultados deste teste podem ser vistos na Fig. 2.

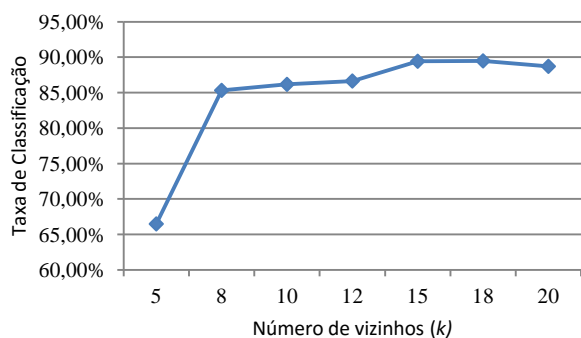


Fig. 2. Gráfico da classificação pelo número de vizinhos,  $k$ .

Nota-se que os melhores resultados ficaram por volta de  $k = n/10$ . Maiores valores não foram testados, para evitar muitas conexões com bases grandes. Apesar de  $k = 18$  ter tido um resultado um pouco melhor que  $k = 15$ , manteve-se o segundo pelo motivo discutido anteriormente, de restringir as conexões do grafo. Assim o valor escolhido foi de  $k = 10\%$ , que seria equivalente ao número de partículas.

O último parâmetro a ser definido foi a taxa de caminhada gulosa, que representa a probabilidade entre 0 e 1 de que a caminhada escolhida aleatoriamente seja gulosa. Conforme recomendado no algoritmo original, variou-se este valor em uma taxa próxima de 0,5. Os resultados são vistos logo abaixo.

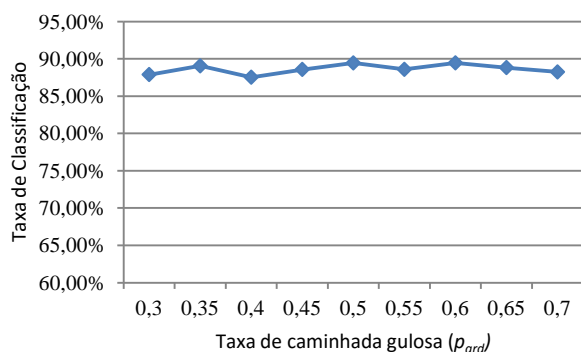


Fig. 3. Gráfico da classificação pela taxa de caminhada gulosa,  $p_{grd}$ .

Constata-se que o melhor valor foi o mesmo sugerido na referencia, 0,5. Porém outras taxas chegaram perto da taxa de classificação obtida com  $p_{grd}$  de 0,5, manteve-se então 0,5 como a taxa no experimento.

#### IV. RESULTADOS

Primeiramente, na Tabela I e na Fig. 4, observa-se os resultados de classificação e desvio-padrão obtidos na base Iris para as diferentes medidas de distância aplicadas. Não houve necessidade de aplicar normalização de dados, já que a base Iris já tem seus dados normalizados por padrão.

TABELA I. RESULTADOS PARA A BASE DE DADOS IRIS

Métrica de Distância	Taxa de Acertos	Desvio Padrão
Euclidiana	89,44%	4,69%
<b>Mahalanobis</b>	<b>91,17%</b>	<b>4,06%</b>
City Block	90,37%	5,51%
Chebyshev	85,28%	8,87%
Minkowski	87,16%	5,64%
Bray-Curtis	89,32%	8,27%
Canberra	81,87%	5,20%

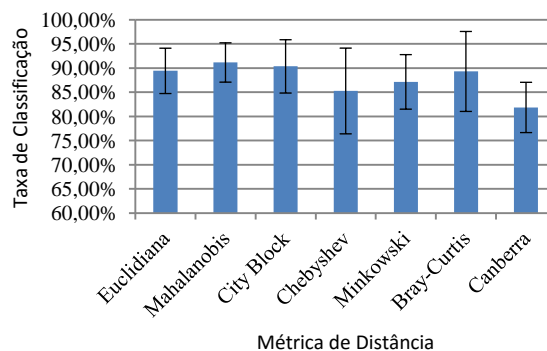


Fig. 4. Gráfico das taxas de acerto na base de dados Iris.

Nota-se então que a melhor métrica nesta base de dados foi a distância de Mahalanobis, com uma taxa de classificação de 91,17% dos dados não rotulados. Observa-se também que as distâncias City Block, Euclidiana e a de Bray-Curtis tiveram resultados próximos dos atingidos pela melhor.

TABELA II. RESULTADOS PARA A BASE DE DADOS WINE

Métrica de Distância	Taxa de Acertos	Desvio Padrão
Euclidiana	63,58%	7,02%
Euclidiana Normalizada	91,50%	7,19%
Mahalanobis	64,73%	5,54%
Mahalanobis Normalizada	90,14%	12,01%
City Block	65,40%	7,30%
<b>City Block Normalizada</b>	<b>93,68%</b>	<b>8,02%</b>
Chebyshev	63,31%	7,28%
Chebyshev Normalizada	84,53%	15,28%
Minkowski	64,40%	6,08%
Minkowski Normalizada	86,53%	10,18%
Bray-Curtis	65,67%	6,53%
Bray-Curtis Normalizada	34,56%	6,30%
Canberra	80,09%	19,06%
Canberra Normalizada	33,72%	3,12%

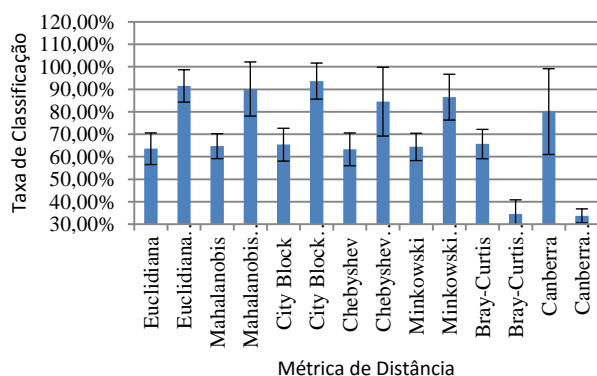


Fig. 5. Gráfico das taxas de acerto na base de dados Wine.

A segunda base na qual foi aplicado o experimento foi a base Wine. Esta não vem normalizada por padrão. Dessa forma os testes foram feitos tanto para os valores absolutos como para os normalizados em cada métrica. Pode-se observar os resultados na Tabela II e no gráfico da Figura 5.

Constata-se que a métrica City Block foi a que obteve os melhores resultados, com uma taxa de classificação de 93,68%. Como pode ser visto, em geral, a normalização melhorou os resultados, excetuando-se as métricas que não recaem no espaço euclidiano (Bray-Curtis e Canberra). Além disso, nota-se que a distância Euclidiana e a de Mahalanobis, chegaram perto do melhor resultado, sendo as únicas a ficarem acima de 90%.

TABELA III. RESULTADOS PARA A BASE DE DADOS BANKNOTE AUTHENTICATION

Métrica de Distância	Taxa de Acertos	Desvio Padrão
Euclidiana	95,48%	2,80%
<b>Euclidiana Normalizada</b>	<b>96,71%</b>	<b>0,86%</b>
Mahalanobis	68,63%	4,99%
Mahalanobis Normalizada	79,69%	1,94%
City Block	95,84%	2,46%
City Block Normalizada	96,49%	0,66%
Chebyshev	94,01%	2,97%
Chebyshev Normalizada	94,81%	1,00%
Minkowski	94,85%	2,18%
Minkowski Normalizada	95,95%	0,81%
Bray-Curtis	64,82%	19,48%
Bray-Curtis Normalizada	51,76%	3,90%
Canberra	52,04%	4,85%
Canberra Normalizada	51,47%	3,41%

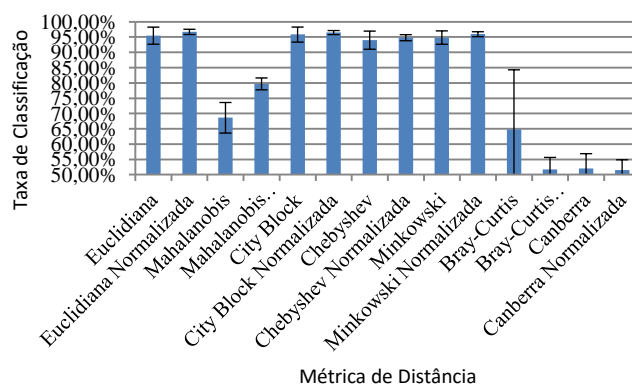


Fig. 6. Gráfico das taxas de acerto na base de dados Banknote Authentication.

Por último, os testes foram feitos numa base maior, de 1372 instâncias, a Banknote Authentication. Esta base também foi normalizada após os testes iniciais com valores absolutos. Portanto, assim como nos experimentos da base Wine, os testes foram feitos tanto para os valores recebidos por padrão da base, assim como para os valores normalizados. Os resultados são mostrados na Tabela III e na Figura 6.

Nesta base a distância Euclidiana (com dados normalizados) foi a que obteve o melhor resultado, com 96,71% de amostras não rotuladas classificadas corretamente. Novamente, a distância City Block teve bons resultados, estando próxima da taxa de acertos atingida pela distância Euclidiana. Por fim, nota-se que as distâncias Bray-Curtis e Canberra tiveram os piores resultados, tendo ainda um alto desvio padrão na distância Bray-Curtis sem normalização.

Portanto, nota-se que para as três bases de dados, obtivemos três métricas de distância diferentes como as de melhor taxa de classificação. Sendo, em síntese, Mahalanobis a melhor métrica para a base Iris, City Block Normalizada para a base Wine e Euclidiana Normalizada para a base Banknote Authentication. Contudo, excetuando-se os resultados da distância de Mahalanobis na base Banknote Authentication, as mesmas três métricas de distância tiveram bons resultados para todas as bases estudadas, podendo ser consideradas, com base nos testes feitos, boas opções para iniciar-se os testes quando se implementando o algoritmo. Observa-se ainda que a normalização trouxe resultados, em geral, positivos em comparação com a aplicação do algoritmo nas bases com dados originais.

## V. CONCLUSÃO

O foco deste trabalho foi a avaliação de métricas no Algoritmo de Competição e Cooperação entre Partículas. Para tanto, pudemos apresentar a análise de diferentes métricas encontradas na literatura, bem como o estudo de suas influências na construção do grafo.

Como resultado, pode-se observar que em duas das três bases que foram objeto de estudo neste trabalho, métricas diferentes da Euclidiana, que é a medida de distância usada por padrão no algoritmo, foram mais eficientes, o que demonstra que o algoritmo pode alcançar taxas de classificação ainda

melhores. Deve-se considerar ainda que nos experimentos, usamos parâmetros definidos na base Iris com uso da distância Euclidiana como “ótimos” para todos os demais experimentos com outras bases e outras distâncias. Porém, é importante ressaltar que os parâmetros poderiam ser definidos e diferentes para cada mudança no algoritmo, como troca de base ou de métrica de distância, abrindo espaço para outras melhorias na taxa de classificação, o qual não era o exato objetivo deste trabalho, mas permanece como ideia para trabalhos futuros.

Trabalhos futuros neste projeto incluem ainda análise de outras métricas de distância ou até mesmo alterações na medida de similaridade que estas calculam, como forma de generalizar tal medida e, com isso, melhorar a classificação correta de rótulos que o algoritmo atinge.

Destaca-se por fim a eficiência do algoritmo estudado e ainda a utilidade do aprendizado de máquina ou até mesmo da Inteligência Artificial em um contexto geral. Sendo tal campo de estudos fundamental nos dias atuais e futuros, tendo de acompanhar as tecnologias que são lançadas e propondo melhorias e automação nas mais diversas indústrias existentes. Deve ser notado também o uso de técnicas de aprendizado semi-supervisionado, os quais fazendo uso de dados não rotulados rotulam eficientemente os dados em questão com um esforço relativamente baixo se comparado a algoritmos supervisionados, por exemplo.

## REFERÊNCIAS

- [1] E. Alpaydin. *Introduction to Machine Learning*. MIT Press, 2004.
- [2] T. Mitchell. *Machine Learning*. McGraw Hill, 1997.
- [3] O. Chapelle, B. Scölkopf, A. Zien. *Semi-Supervised Learning, Adaptive Computation and Machine Learning*. MIT Press, 2006.
- [4] F. A. Breve, L. Zhao, M. G. Quiles, W. Pedrycz, J. Liu. Particle Competition and Cooperation in Networks for Semi-Supervised Learning. *IEEE Transactions on Knowledge and Data Engineering*, v. 24, p. 1686-1698, 2012.
- [5] Q. Liu, X. Chu, J. Xiao, H. Zhu. Optimizing Non-orthogonal Space Distance Using PSO in Software Cost Estimation. *IEEE Computer Software Applications Conference (COMPSAC)*, pp. 21-26, 2014.
- [6] Z. Yang, Y. Shufan, X. Yang, G. Liqun. High-Dimensional Statistical Distance for Object Tracking. *International Conference on Measuring Technology and Mechatronics Automation (ICMTMA)*, vol.2, pp.386-389, 2010.
- [7] R. Shyam, Y. N. Singh. Evaluation of Eigenface and Fisherfaces using Bray Curtis Dissimilarity Metric, *9th International Conference on Industrial and Information Systems (ICIIS)*, pp.1-6, 2014.
- [8] M. Kokare, B. N. Chatterji, P. K. Biswas. Comparison of Similarity Metrics for Texture Image Retrieval, *Conference on Convergent Technologies for the Asia-Pacific Region*, v. 2, pp.571-575, 2003.
- [9] K. Bache, M. Lichman. *UCI Machine Learning Repository* [<http://archive.ics.uci.edu/ml>]. Irvine, CA. University of California, School of Information and Computer Science, 2013.