Análise da Influência de Métricas de Distância no Algoritmo Semi-Supervisionado de Competição e Cooperação entre Partículas

Lucas Guerreiro, Fabricio Aparecido Breve
UNESP
Rio Claro, Brasil
guerreiroluc@gmail.com, fabricio@rc.unesp.br

***Resumo*— O Aprendizado de Máquina é uma área que vem crescendo nos últimos anos e é um dos destaques dentro do campo de Inteligência Artificial. O algoritmo de competição e cooperação entre partículas é uma das técnicas deste domínio, que sempre se utilizou da distância Euclidiana para medir a similaridade entre dados e construir o grafo. Este trabalho tem por objetivo implementar o algoritmo e aplicar nele outras medidas de distância, aplicadas em diferentes bases de dados. Com isso, os resultados de todas as métricas são comparados, identificando quais métricas melhor se adaptam à diferentes bases de dados.**

Área: Matemática e Inteligência Computacional.

I. INTRODUÇÃO

Aprendizado Semi-Supervisionado é uma das categorias de Aprendizado de Máquina [1][2], na qual 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. CONCEITOS E TÉCNICAS

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 na Tabela 1.

TABELA 1 – DISTÂNCIAS UTILIZADAS NO EXPERIMENTO

Distância	***Fórmula***
Euclidiana [5]	$d(x,y) = \sum_i \sqrt{(x_i - y_i)^2}$
Mahalanobis [5]	$d(\vec{x},\vec{y}) = \sqrt{(\vec{x}-\vec{y})^T S^{-1} (\vec{x}-\vec{y})}$
City Block [5]	$d(x,y) = \sum_{i=1}^{n} \lvert x_i - y_i \rvert$
Chebyshev [6]	$d(x,y) = max \lvert x_i - y_i \rvert$
Minkowski [5]	$d(x,y) = \sqrt[p]{\sum_{i=1}^{n} \lvert x_i - y_i \rvert^p}$
Bray-Curtis [7]	$d(x,y) = \frac{\sum \lvert x_i - y_i \rvert}{\sum (x_i + y_i)}$
Canberra [8]	$d(x,y) = \sum_{i=1}^{n} \frac{\lvert x_i - y_i \rvert}{\lvert x_i \rvert + \lvert y_i \rvert}$

V Workshop do Programa de Pós-Graduação em Ciência da Computação: "Qualidade na Pós-graduação", Bauru, 25 e 26 de junho de 2015.

III. METODOLOGIA DE DESENVOLVIMENTO

A proximidade entre os nós no algoritmo de similaridade define as conexões do grafo que é formado. Para avaliar a influência de uma alteração no cálculo das distâncias, 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].

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. Foram efetuados testes com valores absolutos e normalizados das bases de dados. A normalização é feita ao se decrementar a média aritmética de cada atributo e dividir cada instância por seu desvio padrão.

O fator p da medida Minkowski para o experimento foi fixado em $p = 4$.

IV. RESULTADOS PRELIMINARES

Observa-se na Tabela 2 os resultados obtidos no experimento.

TABELA 2 – RESULTADOS PRELIMINARES DO EXPERIMENTO

	Base de Dados		
Distância	***Iris***	***Wine***	***Banknote***
Euclidiana	89,44% (4,69%)	63,58% (7,02%)	95,48% (2,80%)
Euclidiana*	77,33% (10,74%)	91,50% (7,19%)	**96,71% (0,86%)**
Mahalanobis	**91,17% (4,06%)**	64,73% (5,54%)	68,63% (4,99%)
Mahalanobis*	70,93% (10,97%)	90,14% (12,01%)	79,69% (1,94%)
City Block	90,37% (5,51%)	65,40% (7,30%)	95,84% (2,46%)
City Block*	80,36 (12,68%)	**93,68% (8,02%)**	96,49% (0,66%)
Chebyshev	85,28% (8,87%)	63,31% (7,28%)	94,01% (2,97%)
Chebyshev*	70,93% (13,29%)	84,53% (15,28%)	94,81% (1,00%)
Minkowski	87,16% (5,64%)	64,40% (6,08%)	94,85% (2,18%)
Minkowski*	62,13% (11,60%)	86,53% (10,18%)	95,95% (0,81%)
Bray-Curtis	89,32% (8,27%)	65,67% (6,53%)	64,82% (19,48%)
Bray-Curtis*	35,27% (7,93%)	34,56% (6,30%)	51,76% (3,90%)
Canberra	81,87% (5,20%)	80,09% (19,06%)	52,04% (4,85%)
Canberra*	31,95% (4,25%)	33,72% (3,12%)	51,47% (3,41%)

**representa experimentos com base de dados normalizada*

Destaca-se da tabela acima a melhor métrica para cada base, sendo: a) Mahalanobis, para a base Iris; b) City Block, para a base Wine; c) Euclidiana, para a base Banknote Authentication. Observa-se ainda que a normalização melhorou os resultados quando aplicada sobre as bases Wine e Banknote Authentication, porém os valores absolutos dos parâmetros para a base Iris ainda obtiveram melhores taxas de classificação.

V. CONSIDERAÇÕES FINAIS

O foco deste trabalho foi a avaliação de métricas de distância 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. 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. Busca-se ainda entender sob quais condições determinada métrica atinge melhor performance, ou até mesmo a viabilidade de mesclar diferentes métricas de distância como forma de obter uma melhor taxa de classificação

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. Optmizing 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] SHYAM, R.; SINGH, Y. N. Evaluation of Eigenface and Fisherfaces using Bray Curtis Dissimilarity Metric, 9th International Conference on Industrial and Information Systems (ICIIS), pp.1-6, 2014.

[8] KOKARE, M.; CHATTERJI, B. N.; BISWAS, P. K. Comparison of Similarity Metrics for Texture Image Retrieval, Conference on Convergent Technologies for the Asia-Pacific Region, v. 2, pp.571-575, 2003.

[9] BACHE, K.; LICHMAN, M. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA. University of California, School of Information and Computer Science, 2013.