Comparação de Técnicas de Aprendizado de Máquina Semi-Supervisionado na Bioinformática

Diego Henrique Negretto*,
Fabrício Aparecido Breve
Universidade Estadual Paulista "Júlio de Mesquita Filho"
Rio Claro, Brasil
dihnegretto@gmail.com

***Resumo*—A Bioinformática tem surgido como um novo campo de estudo a partir da união de áreas do conhecimento como Ciência da Computação, Biologia e Tecnologia da Informação, tendo como principal objetivo propiciar maneiras de se extrair conhecimentos úteis de grandes conjuntos de dados. O uso de Aprendizado de Máquina permite classificar dados de uma forma automatizada. Pensando nisso, o presente projeto propõe utilizar algoritmos de Aprendizado de Máquina Semi-Supervisionado e realizar uma avaliação comparativa entre esses algoritmos.**

Área: Matemática e Inteligência Computacional

I. Conceitos e técnicas

As pesquisas realizadas para Sequenciamento de Genomas, Proteômica, entre outros, geram uma quantidade de dados biológicos massiva, sendo assim, faz-se necessário o apoio de soluções computacionais para a análise desses dados. Assim, a utilização de técnicas de Aprendizado de Máquina, para a extração de conhecimentos úteis dessas grandes quantidades de dados tem sido amplamente discutida entre pesquisadores tanto da biologia como da computação. O processo para se rotular todos os dados, gerados pelas pesquisas biológicas, muitas vezes é difícil, caro e/ou demorado. Assim, buscar maneiras de se atingir uma grande acurácia com poucos dados rotulados torna-se uma tarefa importante e desafiadora. Dessa forma, pretende-se com esse trabalho empregar Aprendizado de Máquina Semi-Supervisionado e realizar uma avaliação comparativa entre alguns algoritmos na aplicação em dois *datasets* distintos relacionados à área de Análises da Proteômica.

A. Bioinformática e Aprendizado de Máquina

A Bioinformática refere-se a um campo de estudo interdisciplinar que combina as áreas de Computação, Biologia e Tecnologia da Informação. A Bioinformática engloba a análise de dados moleculares expressos sob a forma de aminoácidos, DNA, RNA, peptídeos e proteínas [1], sendo que, dado a enorme quantidade e amplitude de dados, é necessário o desenvolvimento de métodos eficientes para a extração de conhecimento, que possam lidar com o tamanho e complexidade dos dados acumulados.

O aprendizado de máquina trata do projeto e desenvolvimento de algoritmos que imitam o comportamento de aprendizagem humano, com um foco principal em aprender automaticamente a reconhecer padrões complexos e tomar decisões [2]. Dentre os diversos campos relacionados ao Aprendizado de Máquina pode-se citar: Inteligência Artificial, Mineração de Dados, Reconhecimento de Padrões, entre outros [2][3] e, dentre as principais categorias de Aprendizado de Máquina, pode-se citar: **Aprendizado Supervisionado** - deduzem uma função a partir dos dados de treinamento, que consistem em pares de exemplos de entradas e saídas [2]; **Aprendizado Não Supervisionado** - buscam determinar como os dados estão organizados, através de dados de treinamento que consistem apenas de exemplos de entrada e que não estão rotulados [2]; **Aprendizado Semi-Supervisionado** – está em um meio termo entre os das categorias de Aprendizado Supervisionado e Não Supervisionado. Assim, faz-se uso de dados rotulados e não rotulados para o treinamento [4]. O objetivo é obter uma boa classificação com uma menor quantidade de dados rotulados para o treinamento.

II. Metodologia de desenvolvimento

A. Algoritmos Semi-Supervisionados

Foi utilizado o algoritmo Competição e Cooperação entre Partículas (PCC) [5], bem como, os algoritmos Label Propagation (LP) [6], Linear Neighborhood Propagation (LNP) [7] e Local and Global Consistency (GLC) [8].

B. Conjuntos de Dados

A base de dados Yeast, presente no UCI Machine Learning Repository [9] refere-se à predição da localização celular de proteínas de levedura, e possui 1484 elementos e 8 atributos, sendo que, o objetivo é classificar nas 10 diferentes posições (classes) possíveis. Por outro lado, a base de dados E.coli [9] refere-se à predição da localização celular de proteínas na bactéria E.coli, e possui 336 elementos e 8 atributos. O objetivo é classificar nas 8 diferentes posições (classes) possíveis.

C. Conjuntos de Treinamento e Parametrização

Foram utilizadas as configurações de dados de treinamento contendo 3%, 5%, 10%, 15% e 20% de dados rotulados. Foram sorteados 10 subconjuntos de dados rotulados, para cada *dataset* e para cada configuração, garantindo que pelo menos um elemento de cada classe estivesse presente nesses subconjuntos. Os algoritmos tiveram seus parâmetros otimizados variando nos seguintes intervalos: PCC $1 \leq k \leq 100$; GLC e LP $0 < \sigma \leq 100$; LNP $1 \leq k \leq 100$.

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

* Bolsista FAPESP.

III. RESULTADOS PRELIMINARES

Para a base de dados E.coli, os algoritmos semi-supervisionados estudados, classificaram os dados corretamente no intervalo de aproximadamente 65% até 88%, conforme mostrado na Figura 1.

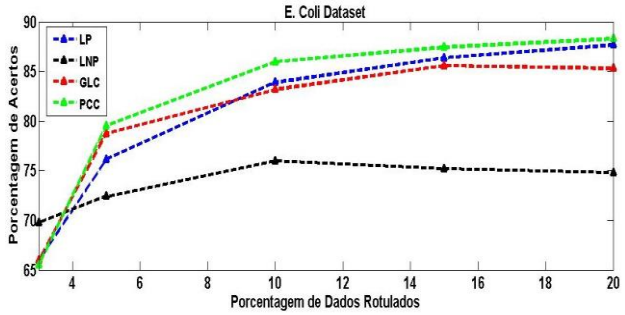

Fig. 1. Acertos E.coli Dataset.

Observa-se que o algoritmo LNP tem uma maior taxa de acerto para a configuração de 3%, porém, os algoritmos PCC, GLC e LP, possuem melhores resultados se comparados com o algoritmo LNP nas demais configurações, sendo que, o algoritmo PCC é o que apresenta melhores resultados. A Tabela 1 apresenta as médias e desvios padrão para os 10 subconjuntos em todas as configurações.

TABELA 1 – Médias Subconjuntos Ecoli Dataset.

E.coli Dataset				
Rotulados	***LP***	***LNP***	***GLC***	***PCC***
3%	66,01 (±15,75)	**69,73 (±11,05)**	65,95 (±14,15)	65,47 (±7,89)
5%	76,13 (±5,53)	72,40 (±5,82)	78,75 (±4,29)	**79,52 (±2,58)**
10%	83,90 (±2,59)	75,98 (±4,80)	83,18 (±3,84)	**85,95 (±0,66)**
15%	86,36 (±2,14)	75,20 (±5,52)	85,56 (±2,03)	**87,44 (±1,21)**
20%	87,64 (±2,76)	74,79 (±4,74)	85,35 (±2,72)	**88,27 (±0,78)**

Para a base de dados Yeast, os algoritmos classificaram os dados corretamente no intervalo de aproximadamente 43% até 66%, conforme mostrado na Figura 2.

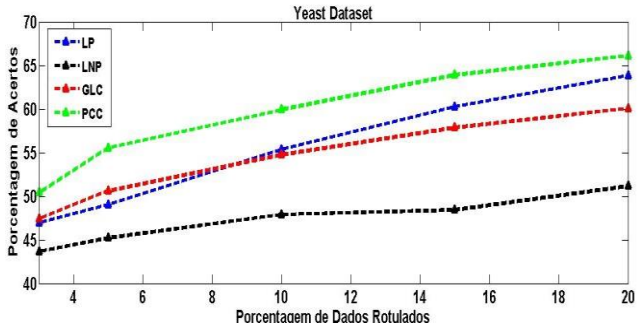

Fig. 2. Acertos Yeast Dataset.

Observa-se que o algoritmo PCC tem maiores taxas de acertos para todas as configurações, enquanto o algoritmo LNP possui as taxas de acertos mais baixas. A Tabela 2 apresenta as médias e desvios padrão para os 10 subconjuntos.

TABELA 2 – MÉDIAS SUBCONJUNTOS YEAST DATASET.

Yeast Dataset				
Rotulados	***LP***	***LNP***	***GLC***	***PCC***
3%	46,96 (±3,47)	43,65 (±3,72)	47,40 (±4,29)	**50,41 (±3,06)**
5%	49,04 (±3,75)	45,22 (±5,42)	50,57 (±4,28)	**55,51 (±2,49)**
10%	55,37 (±1,87)	47,91 (±3,88)	54,72 (±2,41)	**59,94 (±1,40)**
15%	60,28 (±1,92)	48,42 (±2,77)	57,87 (±1,66)	**63,90 (±0,64)**
20%	63,84 (±1,30)	51,13 (±2,59)	60,07 (±0,88)	**66,08 (±0,47)**

IV. CONSIDERAÇÕES FINAIS

Como observado nos resultados apontados por esse trabalho, os algoritmos semi-supervisionados apresentam taxas de classificação satisfatórias, sendo que, o algoritmo PCC consegue obter resultados semelhantes a algoritmos supervisionados com muito mais dados rotulados. Ressalta-se que com a possibilidade de classificação eficiente com os métodos semi-supervisionados, o trabalho de rotulagem de dados, realizado pelos especialistas de domínio, pode ser diminuído, o que contribui para a diminuição de gastos financeiros e de tempo. As próximas etapas desse trabalho consistem na definição de mais uma base de dados, em conjunto com biólogos especialistas de domínio, relacionada a enzimas presentes no processo de produção de biocombustível, bem como a análise de desempenho dos algoritmos nessa nova base.

REFERÊNCIAS

[1] Zhaoli, Machine Learning in Bioinformatics. In: *International Conference on Computer Science and Network Technology*, China, 2011.

[2] T. Mitchell, *Machine Learning*, McGraw Hill, 1997.

[3] E. Alpaydin, *Introduction to Machine Learning*, 2 ed. Cambridge: The Mit Press, 2010.

[4] O. Chapelle, B. Scholkopf, A. Zien, *Semi-Supervised Learning*, Cambridge: The Mit Press, 2010.

[5] F. Breve, L. Zhao, M. Quiles, W. Pedycz, J. Liu, Particle Competition and Cooperation in Networks for Semi-Supervised Learning. In: *Knowledge and Data Engineering, IEEE Transactions on 24(9).* 2012.

[6] X. Zhu, Z. Ghahramani, Learning from Labeled and Unlabeled Data with Label Propagation. In: *Technical Report CMU-CALD-02-107*, Carnegie Mellon University, Pittsburgh, 2002.

[7] F. Wang, C. Zhang, Label Propagation Through Linear Neighborhoods. In: *IEEE Transactions on Knowledge and Data Engineering*, v. 20, 2008.

[8] D. Zhou, O. Bousquet, T. N. Lan, J. Weston, B. Scholkopf, Learning with Local and Global Consistency. In: *Advances in Neural Information Processing Systems*, v. 16, pp, 321-328, MIT Press, 2004.

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