Abordagem de Aprendizado Ativo para Classificação de Dados Biomédicos

Guilherme Camargo*, Pedro Henrique Bugatti*, Priscila Tiemi Maeda Saito*†
*Departamento de Computação, Universidade Tecnológica Federal do Paraná (UTFPR-CP)
†Instituto de Computação, Universidade Estadual de Campinas (UNICAMP)
Email: gcamargo@alunos.utfpr.edu.br, {pbugatti, psaito}@utfpr.edu.br

***Abstract*—A huge volume of biomedical data (images, genes, among others) is daily generated. The analysis of such data is a complex task that demands specialized knowledge, and the level of expertise directly impacts the diagnosis. Besides, due to the volume of data such task becomes extremely tiresome, and hence highly susceptible to errors. Trying to solve this problem, machine learning approaches have been proposed in the literature to perform automatic classification of such data. Despite the several proposed techniques, the great majority strictly focus just on the effectiveness, and relegate the efficiency of the classification. This paper presents a novel learning approach capable to obtain high accuracies, as well as maintaining a minimal involvement of the expert and interactive computational time during the learning process. To do so, the proposed approach exploits the active learning paradigm, in order to reduce, organize and select the most informative samples to the learning process of the pattern classifier.**

***Resumo*—Um grande volume de dados biomédicos (imagens, genes, entre outros) é gerado diariamente. A análise de dados biomédicos é uma tarefa complexa realizada por especialistas, e na qual o nível de especialidade pode afetar o diagnóstico. Além disso, devido ao grande volume de dados, tal tarefa torna-se um processo cansativo e suscetível a erros. Neste contexto, abordagens de aprendizado têm sido propostas na literatura para classificação automática. No entanto, muitas delas negligenciam a necessidade de técnicas mais efetivas e eficientes para classificação de dados biomédicos. Este trabalho apresenta uma abordagem de aprendizado de forma a obter acurácias elevadas na classificação, envolvimento mínimo do especialista e tempos computacionais interativos durante o processo de aprendizado. Para tanto, a abordagem proposta explora aprendizado ativo, de forma a reduzir, organizar e selecionar amostras mais informativas ao processo de aprendizado do classificador de padrões.**

***Keywords*-aprendizado ativo; análise de imagens; classificação; imagens biomédicas, floresta de caminhos ótimos**

I. Introdução

Devido aos avanços tecnológicos com relação aos dispositivos de aquisição, aumento da capacidade de armazenamento e velocidade de comunicação, uma grande quantidade de dados biomédicos (imagens, genes, entre outros) é gerada diariamente, produzindo grandes conjuntos de dados.

A análise de dados biomédicos é uma tarefa complexa realizada por especialistas, e em que o nível de especialidade pode afetar o diagnóstico. Além disso, dada a quantidade de dados, tal tarefa torna-se um processo extremamente cansativo e altamente suscetível a erros.

O desenvolvimento de técnicas para realizar o processamento, organização e classificação dos dados torna-se imprescindível para posterior manipulação e recuperação de informações. Uma maneira de anotar uma base de dados é associar uma informação textual (rótulo) às amostras, atribuindo-na a uma dada classe, de acordo com o contexto da aplicação.

Técnicas de aprendizado de máquina têm sido desenvolvidas e utilizadas para treinar e construir um classificador de padrões robusto, viabilizando o processo de anotação.

Muitos trabalhos na literatura consideram todas as amostras do conjunto de dados para o treinamento do classificador. No entanto, muitas amostras do conjunto podem fornecer informações redundantes, bem como algumas podem fornecer informações mais relevantes em relação às outras.

Neste contexto, técnicas de aprendizado ativo têm sido exploradas e bem sucedidas para selecionar uma quantidade reduzida de amostras mais significativas para o aprendizado do classificador.

Apesar dos esforços, a maioria das estratégias propostas não levam em consideração os requisitos de tempos computacionais interativos exigidos em aplicações reais. A abordagem de aprendizado ativo tradicional requer, a cada iteração do aprendizado, os processos de classificação, organização e seleção das amostras mais informativas para o treinamento do classificador, considerando todo o conjunto de dados.

Contribuições: Este trabalho propõe uma técnica de aprendizado ativo mais efetiva e eficiente com base em um paradigma de aprendizado ativo, o qual considera a redução e organização de um sub-conjunto de amostras mais significativas, para posterior seleção durante o processo de aprendizado.

II. Conceitos Relacionados

Para a construção de um classificador, algumas dificuldades são encontradas tanto no processo de seleção de amostras significativas para a etapa de aprendizado, quanto na quantidade de amostras que devem ser selecionadas.

A simples utilização de um método *totalmente randômico* pode não ser o ideal, pois amostras são selecionadas aleatoriamente do conjunto de treinamento sem nenhuma restrição ou pré-processamento.

É importante que o conjunto de amostras selecionadas represente todo o problema, reduzindo redundância de informação e diminuindo o tempo de processamento e o envolvimento do especialista.

Uma estratégia bastante utilizada para seleção dos dados é escolher as amostras mais representativas ou mais diversas, de forma a fornecer ao classificador o conhecimento de todas as classes mais rapidamente ao longo das iterações do aprendizado. Outra estratégia amplamente utilizada é a escolha das amostras mais informativas, incertas ou mais difíceis de serem classificadas, que encontram-se mais próximas à fronteira de classificação.

Trabalhos indicam que um melhor desempenho pode ser obtido levando-se em consideração o conhecimento a priori da distribuição dos dados. Nestes casos, técnicas de agrupamento têm sido incorporadas ao aprendizado ativo [1]–[6]. A partir do agrupamento, amostras mais representativas de cada classe podem ser obtidas por meio das raízes dos *clusters* e as amostras mais informativas por meio das amostras de fronteira entre *clusters*.

No entanto, a maioria dos trabalhos [3], [5]–[9] realiza o agrupamento a cada iteração do aprendizado, bem como utilizam todas as amostras do conjunto de dados para realizar os agrupamentos. Alguns trabalhos [10] propõem uma abordagem de aprendizado ativo, a qual realiza o agrupamento uma única vez e considera apenas um conjunto reduzido de amostras para realizar o agrupamento. Tal abordagem será utilizada como base neste trabalho, no entanto com uma proposta de organização e seleção das amostras do conjunto reduzido (descrita na Seção III).

III. Estratégia de Aprendizado Ativo Proposta

A estratégia de aprendizado ativo proposta realiza um pré-processamento, o qual consiste nos processos de redução e organização do conjunto reduzido de amostras mais informativas para o aprendizado do classificador, diferentemente das abordagens tradicionais. Posteriormente, a partir da pré-organização, o processo de seleção das amostras torna-se mais rápido, uma vez que não requer a classificação e re-organização de todas as amostras do conjunto de dados.

A. Estratégia de Redução

A estratégia de redução adotada é realizada por meio do agrupamento das amostras do conjunto de dados. O conjunto reduzido é composto por amostras raízes de cada *cluster* e amostras de fronteira entre *clusters*. Uma amostra é considerada de fronteira entre *clusters*, uma vez que, considerando uma dada vizinhança, exista pelo menos uma amostra k-vizinha mais próxima de *cluster* distinto da amostra em questão.

Após realizar o agrupamento das amostras do conjunto, bem como analisar quais amostras são de fronteira, são armazenados os pares de amostras de fronteira entre *clusters* distintos.

B. Estratégia de Organização

Muitas amostras de fronteira podem ser obtidas após o processo de redução. Para aumentar a possibilidade de selecionar amostras de classes distintas e de fato as mais significativas mais rapidamente, os pares de amostras de fronteira são pré-organizados.

O critério de organização consiste em calcular as distâncias entre as amostras de cada par e organizá-las em ordem crescente de distância. Ao final, tem-se uma lista de arestas de fronteira pré-organizadas em ordem crescente de distância.

A ideia é priorizar amostras mais similares e que sejam de classes distintas, as quais seriam as mais difíceis de serem classificadas. Para tanto, uma estratégia de seleção deve ser considerada e é apresentada na subseção III-C.

C. Estratégia de Seleção

Dado o conjunto reduzido e organizado previamente, tem-se o processo de seleção de amostras durante o ciclo de aprendizado. Diferentemente das abordagens de aprendizado ativo tradicionais, o processo de seleção não requer a classificação e re-organização de todas as amostras do conjunto de dados. Dessa forma, a seleção torna-se mais rápida até mesmo para grandes conjuntos de dados.

Na primeira iteração do aprendizado, as amostras correspondentes às raízes dos *clusters* obtidos a partir do agrupamento são exibidas ao especialista, o qual realiza a anotação dos rótulos para essas amostras. Tais amostras anotadas constituem o conjunto de treinamento da primeira instância do classificador.

Durante o ciclo de aprendizado, as amostras na lista ordenada de arestas são analisadas. Uma aresta por vez é obtida e as amostras que constituem tal aresta são rotuladas pela instância atual do classificador. Tais amostras, caso recebam rótulos distintos, são selecionadas para serem exibidas na iteração seguinte. Caso contrário, a próxima aresta na lista é obtida e rotulada pelo classificador. Note que o classificador auxilia no processo de seleção de amostras para o seu aprendizado.

Após a obtenção de uma quantidade de amostras a serem selecionadas por iteração, estas são exibidas ao especialista. A partir da primeira iteração, o especialista apenas confirma ou corrige os rótulos das amostras rotuladas incorretamente pelo classificador. Em seguida, após as confirmações e/ou correções do especialista, tais amostras são incorporadas ao conjunto de treinamento da iteração anterior, e uma nova instância do classificador é gerada. O processo de aprendizado continua até que o especialista esteja satisfeito com as acurácias, de acordo com o contexto da aplicação.

IV. Experimentos

Foram realizados experimentos com três conjuntos distintos de dados biomédicos disponíveis publicamente [11]–[13]. Os conjuntos de dados referem-se a pacientes com Leucemia Linfoblástica Aguda (*Acute Lymphoblastic Leukemia - ALL*), tipo de câncer que ataca as células brancas (leucócitos) do sangue. Possui sua origem na medula óssea e se espalha para outras partes do corpo. Foi utilizado também um conjunto de dados médicos referente a Leucemias de Linhagem Mista (*Mixed-Lineage Leukemia - MLL*), tipo de câncer que afeta principalmente crianças. Diferente da *ALL*, o paciente costuma apresentar altas taxas de glóbulos brancos. Informações de cada um dos conjuntos estão presentes na Tabela I.

Para cada um dos conjuntos de dados, a estratégia de aprendizado ativo proposta (denotada como AFC – Arestas

Tabela I
DESCRIÇÃO DOS CONJUNTOS DE DADOS.

Nome do conjunto	Quantidade de amostras	Quantidade de atributos	Quantidade de classes
AML ALL [11]	72	7129	2
MLL [12]	72	12582	3
Subtypes of ALL [13]	327	12558	7

de Fronteira Crescentes) foi avaliada com a estratégia de seleção aleatória (denotada como Rand – randômica) a partir do conjunto de dados completo.

Considerando cada conjunto de dados como um conjunto $\mathcal{Z}$, o qual é dividido aleatoriamente em 80% de amostras para o conjunto de aprendizado $\mathcal{Z}_2$ e em 20% de amostras para o conjunto de teste $\mathcal{Z}_3$. O conjunto de treinamento $\mathcal{Z}_1$ aumenta a cada iteração do aprendizado, quando o classificador seleciona novas amostras a partir de amostras de fronteira que foram obtidas de $\mathcal{Z}_2$ e o especialista confirma/corrige os rótulos das amostras selecionadas. O tamanho do conjunto selecionado a cada iteração é o mesmo para todos os métodos comparados, e corresponde ao número de classes c do conjunto.

É importante destacar que diferentes métodos de agrupamento (tais como, k-means, k-medoids, *OPF*) podem certamente ser utilizados no processo de redução dos dados. Neste trabalho, foi utilizado o k-means [14], sendo o valor de k definido como o número de classes c do conjunto. Diferentes métodos de classificação também podem ser utilizados, neste trabalho foram utilizados: o classificador *Support Vector Machines* (*SVM*) [15] e o classificador baseado em floresta de caminhos ótimos (*OPF*) [16]. Para facilitar a comparação entre os métodos, quando aplicável, tais métodos foram denominados como uma tripla, consistindo de método de agrupamento, método de aprendizado ativo e método de classificação. Os métodos foram denotados como: k-means_AFC_OPF, k-means_AFC_SVM, Rand_OPF e Rand_SVM.

Foram mensuradas as médias dos valores de acurácias obtidos por cada instância do classificador a cada iteração do aprendizado, quando aplicada ao conjunto de teste $\mathcal{Z}_3$. Para obtenção das médias, foram realizadas 10 execuções para cada um dos métodos comparados. Além das acurácias, foram mensurados os tempos computacionais para treinamento, seleção e teste para cada um dos métodos e conjuntos utilizados.

Para execução dos experimentos, foi utilizado um computador com processador *Intel Core i3* de 2.27 GHz e 3 GB de memória RAM, com sistema operacional *elementary OS*.

V. RESULTADOS

As Tabelas II-IV apresentam as acurácias médias obtidas pelos métodos para os conjuntos *AML ALL*, *MLL* e *Subtypes of ALL*, respectivamente.

De uma forma geral, as técnicas de aprendizado ativo (k-means_AFC_OPF e k-means_AFC_SVM), utilizando os classificadores *OPF* e *SVM*, apresentaram melhores resultados em relação aos métodos aleatórios (RAND_OPF e RAND_SVM). Comparando a utilização dos classificadores, o *SVM* apresentou um melhor resultado para o conjunto de dados *AMLALL* e o *OPF* apresentou um melhor resultado para os conjuntos de dados *MLL* e *Subtypes of ALL*.

Tabela II
MÉDIA DE ACURÁCIAS DOS MÉTODOS PARA O CONJUNTO DE DADOS *AMLALL*.

Iteração	k-means_AFC OPF (%)	k-means_AFC SVM (%)	RAND OPF (%)	RAND SVM (%)
1	50.00	68.75	50.00	66.25
2	58.89	72.22	50.00	68.33
3	65.78	78.50	56.22	70.50
4	67.33	79.09	62.33	74.09
5	70.89	82.08	66.33	76.25

Tabela III
MÉDIA DE ACURÁCIAS DOS MÉTODOS PARA O CONJUNTO DE DADOS *MLL*.

Iteração	k-means_AFC OPF (%)	k-means_AFC SVM (%)	RAND OPF (%)	RAND SVM (%)
1	79.73	64.44	62.17	50.56
2	84.40	70.95	79.03	66.19
3	87.95	75.00	84.60	74.58
4	87.47	79.63	87.99	77.78
5	86.81	79.00	89.88	81.33

Tabela IV
MÉDIA DE ACURÁCIAS DOS MÉTODOS PARA O CONJUNTO DE DADOS *Subtypes of ALL*.

Iteração	k-means_AFC OPF (%)	k-means_AFC SVM (%)	RAND OPF (%)	RAND SVM (%)
1	60.14	31.92	60.81	28.90
2	63.40	40.38	64.04	46.13
3	65.18	48.05	67.88	48.16
4	65.72	54.79	68.36	55.32
5	67.80	59.01	69.16	62.97

Tabela V
MÉDIA DOS TEMPOS DE EXECUÇÃO PARA TESTE, TREINAMENTO E SELEÇÃO DE AMOSTRAS PARA CADA UM DOS MÉTODOS E CONJUNTOS DE DADOS UTILIZADOS.

Conjunto de dados	Método	Tempo Teste (s)	Tempo Treino (s)	Tempo Seleção (s)
AMLALL	k-means_AFC_OPF	0,001936	0,002738	0,046727
	k-means_AFC_SVM	0,173788	0,035915	0,045989
	RAND_OPF	0,003384	0,012609	0,000020
	RAND_SVM	0,232777	0,050608	0,000020
MLL	k-means_AFC_OPF	0,005083	0,010059	0,162513
	k-means_AFC_SVM	0,380726	0,107149	0,162513
	RAND_OPF	0,005033	0,018921	0,000023
	RAND_SVM	0,285145	0,078135	0,000023
Subtypes of ALL	k-means_AFC_OPF	0,162658	0,352797	3,989391
	k-means_AFC_SVM	2,473569	1,335083	4,002085
	RAND_OPF	0,151755	0,479670	0,000036
	RAND_SVM	2,405556	1,350664	0,000036

Com relação ao tempo computacional obtido para teste, treinamento e seleção das amostras para cada um dos métodos e conjuntos utilizados (Tabela V), é possível observar que, de uma forma geral, o classificador *OPF* conseguiu realizar as etapas de treinamento e de teste mais rapidamente que o classificador *SVM*.

Com relação ao tempo para seleção das amostras, a diferença ocorre devido à utilização ou não de um método de

seleção com base em uma pré-organização. Embora o método aleatório tenha apresentado menor tempo para seleção, vale destacar que a técnica proposta também apresenta tempos computacionais interativos para seleção. Além disso, apresenta o conhecimento de todas as classes do conjunto desde a primeira iteração, diferentemente do método aleatório que obtém conhecimento após algumas iterações de aprendizado.

VI. Conclusão

Neste trabalho foi apresentada uma estratégia de aprendizado ativo para classificação de dados biomédicos. A estratégia de redução de dados adotada realiza uma diminuição significativa dos conjuntos de dados. As raízes obtidas a partir do agrupamento aumentam a possibilidade de selecionar amostras a partir de todas as classes desde a primeira iteração do aprendizado e, as amostras de fronteira possibilitam obter as mais difíceis para classificação.

As estratégias de organização e seleção propostas permitem a priorização de amostras mais informativas dentre as amostras de fronteira previamente analisadas. A estratégia proposta mostra-se mais adequada para lidar com grandes conjuntos de dados obtidos a partir de aplicações reais, as quais podem requerer tempos computacionais interativos, menor envolvimento do especialista e acurácias elevadas mais rapidamente ao longo do processo de aprendizado.

Dado que os processos de redução e organização são realizados por agrupamento de amostras, a escolha do método de agrupamento é fundamental. Para trabalhos futuros, experimentos devem ser realizados utilizando outros métodos, tais como k-medoid e *OPF*. Outros trabalhos envolvem o desenvolvimento de outras estratégias para ordenação das amostras. Além disso, pretende-se avaliar a combinação de técnicas de aprendizado ativo com técnicas de aprendizado semi-supervisionado.

Agradecimentos

Os autores gostariam de agradecer à CAPES, CNPq, Fundação Araucária, UTFPR e SETI.

Referências

[1] A. Cardoso-Cachopo and A. L. Oliveira, "Semi-supervised single-label text categorization using centroid-based classifiers," in *ACM Symposium on Applied Computing (SAC)*, New York, NY, USA, 2007, pp. 844–851.

[2] E. Lughofer, "Hybrid active learning for reducing the annotation effort of operators in classification systems," *Pattern Recognition*, vol. 45, no. 2, pp. 884–896, 2012.

[3] H. T. Nguyen and A. Smeulders, "Active learning using pre-clustering," in *International Conference on Machine Learning (ICML)*. New York, NY, USA: ACM, 2004, pp. 79–86.

[4] B. M. Nogueira, A. M. Jorge, and S. O. Rezende, "Hierarchical confidence-based active clustering," in *Proceedings of the 27th Annual ACM Symposium on Applied Computing (SAC)*, 2012, pp. 216–219.

[5] X. Shen and C. Zhai, "Active feedback - uiuc trec-2003 hard experiments," in *Text REtrieval Conference (TREC)*, 2003, pp. 662–666.

[6] Z. Xu, K. Yu, V. Tresp, X. Xu, and J. Wang, "Representative sampling for text classification using support vector machines," in *European Conference on IR Research (ECIR)*. Berlin, Heidelberg: Springer-Verlag, 2003, pp. 393–407.

[7] L.-J. Chien, C.-C. Chang, and Y.-J. Lee, "Variant methods of reduced set selection for reduced support vector machines," *Journal of Information Science and Engineering*, pp. 183–196, 2010.

[8] S. Ertekin, J. Huang, L. Bottou, and L. Giles, "Learning on the border: active learning in imbalanced data classification," in *16th ACM International Conference on Information and Knowledge Management (CIKM)*, New York, NY, USA, 2007, pp. 127–136.

[9] Y. J. Lee and O. L. Mangasarian, "RSVM: Reduced support vector machines," Computer Science, Univ. of Wisconsin, Tech. Rep. Technical Report 00-07 (First SIAM ICDM, Chicago, 2001), July 2000.

[10] P. T. M. Saito, P. J. de Rezende, A. X. Falcão, C. T. N. Suzuki, and J. F. Gomes, "Improving active learning with sharp data reduction," in *WSCG Communication Proceedings of 20th International Conference on Computer Graphics, Visualization and Computer Vision (WSCG)*, 2012, pp. 27–34.

[11] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring." *Science*, vol. 286, no. 5439, pp. 531–537, Oct. 1999.

[12] S. A. Armstrong, J. E. Staunton, L. B. Silverman, R. Pieters, M. L. den Boer, M. D. Minden, S. E. Sallan, E. S. Lander, T. R. Golub, and S. J. Korsmeyer, "Mll translocations specify a distinct gene expression profile that distinguishes a unique leukemia," *Nature Genetics*, vol. 30, no. 1, pp. 41–47, 2002.

[13] L. J, L. H, D. JR, Y. AE, and W. L, "Simple rules underlying gene expression profiles of more than six subtypes of acute lymphoblastic leukemia (all) patients," *Bioinformatics*, 2003.

[14] J. Wu, *Advances in K-means Clustering: A Data Mining Thinking*. Springer Publishing Company, Incorporated, 2012.

[15] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[16] J. P. Papa, A. X. Falcão, V. H. C. Albuquerque, and J. M. R. S. Tavares, "Efficient supervised optimum-path forest classification for large datasets," *Pattern Recognition*, vol. 45, no. 1, pp. 512–520, 2012.