Linha de pesquisa: Aprendizagem de Máquina e Mineração

Tema de Pesquisa: Desenvolvimento e aplicações de mineração de dados e data science

Título da Proposta de Projeto: Abordagem evolucionária para *exceptional model mining* em domínios de alta dimensionalidade.

Proponente: Juliana Barcellos Mattos

1 Introdução

A mineração de dados propõe o desenvolvimento de técnicas computacionais com objetivo de extrair informações implícitas, desconhecidas e potencialmente úteis [1] a partir de grandes volumes de dados. As técnicas de mineração podem ser aplicadas sob duas diferentes perspectivas [2] [3]: p*redictive induction*, que objetiva extrair conhecimento dos dados com o intuito de prever ou classificar determinado valor de classe de um exemplo desconhecido; e d*escriptive induction*, que objetiva descobrir conhecimento interessante a respeito dos dados em forma de padrões [4].

São conhecidas como *Supervised Descriptive Rule Induction* as técnicas – tais como *contrast set mining* (CSM), *emerging pattern mining* (EPM) e *subgroup discovery* (SD) – que combinam ambas perspectivas. Novak et al. [3] unificam essas três áreas como padrões discriminativos, que consiste na tarefa da mineração que tem o objetivo de identificar conjuntos dos dados que distinguem um grupo alvo dos demais [5]. *Subgroup Discovery*, em específico, tem o intuito de extrair relações entre diferentes variáveis com respeito a uma determinada propriedade de interesse conhecida como variável alvo, e pode ser entendida como um caso especial de uma tarefa mais geral de aprendizado de regras [2] [3] [4]. No entanto, SD torna-se inviável em situações em que se deseja encontrar subgrupos que apresentem algum tipo de excepcionalidade em relação a uma propriedade de interesse que não possa ser expressa por uma única variável [6].

Nesse contexto, *Exceptional Model Mining* (EMM) é uma técnica de análise exploratória de dados que pode ser compreendida como uma generalização do SD, porém, ao invés de utilizar uma única variável alvo, utiliza um conceito mais complexo [7]. EMM estende a ideia a alvos que são algum tipo de modelo, buscando subgrupos que apresentem uma distribuição não-usual das variáveis desse modelo, ao invés de uma única variável [8] [9]. Objetiva-se encontrar subgrupos nos quais seu modelo ajustado (*fitted model*) seja substancialmente diferente do mesmo modelo quando aplicado ao universo completo de dados [7] [8]. Isto posto, quando comparado ao SD, EMM consiste em uma abordagem mais ampla de busca por excepcionalidades e, uma vez formulado um modelo específico para a propriedade de interesse em questão, tal método é capaz de obter subgrupos contendo maior quantidade de informações a respeito dos dados [6]. No entanto, a tarefa de EMM tem, em geral, um custo computacional mais elevado uma vez que lida com grandes conjuntos de dados numéricos e indução de modelos.

Vários métodos têm sido desenvolvidos para minerar padrões discriminativos, no entanto, a alta dimensionalidade dos atuais conjuntos de dados consiste em um desafio à maioria dos algoritmos existentes, devido, sobretudo, à natureza combinatória dos dados [5]. Há, ainda, problemas como a quantidade de padrões retornados, a redundância de padrões minerados, e aspectos computacionais, como tempo de execução e uso de memória, que vêm sendo abordados sob diferentes estratégias. Nesse contexto, abordagens exaustivas tornam-se, muitas vezes, inviáveis. Estratégias heurísticas, portanto, caracterizam uma alternativa de aplicação viável, uma vez que seus métodos, em geral, têm um desempenho mais eficiente que métodos exaustivos, gerando o mesmo número de padrões significativos [10]. Além disso, apesar da

grande quantidade de aplicações na literatura, pouca atenção tem sido dada à mineração de padrões discriminativos em domínios de alta dimensionalidade [5].

Portanto, o dado contexto motiva a seguinte pergunta de pesquisa: *é possível desenvolver um novo método heurístico eficiente para a mineração de modelos excepcionais (EMM) associada a dados de alta dimensionalidade?*

2 Motivação e Justificativa

Há na literatura várias contribuições, em diversas áreas de pesquisa, que buscam conhecimento descritivo dos dados associado a uma propriedade de interesse. As aplicações já existentes abrangem áreas de domínio médico, de bioinformática, de aplicações industriais em análise de faltas técnicas, de redes sociais, de *marketing* e de *e-learning* [2] [11]. Além disso, trabalhos direcionados a *exceptional model mining* [7] [8] [12] expandem ainda mais as possibilidades de aplicações [11].

Diversos algoritmos – exaustivos e heurísticos – para a mineração de padrões discriminativos são encontrados na literatura [2] [4] [10] [13]. Dentre as abordagens heurísticas, destacam-se algoritmos baseados em *beam search* e em computação evolucionária. Embora *beam search* seja a técnica mais comumente utilizada, apresenta como grande desvantagem a falta de diversidade, uma vez que o método enfoca apenas bons itens. Nesse contexto, algoritmos evolucionários constituem uma boa abordagem e têm sido amplamente estudados em aplicações para *subgroup discovery*. Os métodos existentes têm obtido bons resultados tanto no que se refere a aspectos computacionais quanto ao que se refere à qualidade dos padrões minerados, e têm se mostrado uma alternativa competitiva na mineração de padrões discriminativos. No entanto, a maioria das técnicas presentes na literatura foram desenvolvidas para dados de baixa dimensionalidade e, além disso, não foi encontrada nenhuma literatura a respeito de abordagens evolucionárias direcionadas a *exceptional model mining*.

Um método específico para mineração de modelos excepcionais (EMM) em dados de alta dimensionalidade abre novas possibilidades de aplicações em áreas de pesquisa médica, em bioinformática, e em diversas outras áreas. A boa aplicabilidade e os bons resultados de algoritmos evolucionários na mineração de padrões discriminativos indicam que abordar tal questão sob a ótica evolucionária possa ocasionar boas contribuições.

3 Objetivo

A presente proposta objetiva o desenvolvimento de um algoritmo evolucionário orientado ao *Exceptional Model Mining* com foco em aplicações em domínios de alta dimensionalidade.

4 Revisão Bibliográfica

Buscar subgrupos de dados que apresentem alguma excepcionalidade em relação a um atributo de interesse é uma ferramenta importante da análise exploratória de dados. Para um bom desempenho dos métodos computacionais, a estratégia empregada na busca de tais subgrupos é uma questão essencial, uma vez que os espaços de busca são exponenciais em relação à quantidade de atributos. Vários algoritmos têm sido pensados e desenvolvidos para minerar padrões discriminativos de forma eficiente, dentre eles, as estratégias mais utilizadas são: a busca exaustiva, o *beam search* e algoritmos genéticos. Embora a abordagem exaustiva garanta a descoberta das melhores soluções, para dados de alta dimensionalidade, o espaço de hipóteses cresce exponencialmente, tornando inviável tal tipo de busca [6] [9] [11]. Soluções para viabilizar a aplicação de buscas exaustivas geralmente restringem os atributos a serem nominais e impõem restrição anti-monotonicidade às medidas de qualidade. Uma vez que EMM objetiva capturar qualquer conceito e excepcionalidade e, portanto, é importante a manipulação de

atributos numéricos e de qualquer tipo de medida de qualidade [6], há a motivação de se investigar métodos heurísticos. A presente seção propõe uma breve revisão bibliográfica das principais estratégias heurísticas de busca aplicadas para mineração de padrões discriminativos e dos principais algoritmos existentes. Para uma abordagem mais completa a respeito, são referenciados os trabalhos [4] e [13].

Beam Search é a técnica heurística mais comumente utilizada para mineração de padrões discriminativos. Nela, são considerados candidatos apenas um número predefinido (determinado por um parâmetro de *beam size*) dentre as melhores soluções parciais, e, a cada nível de busca, novos candidatos são gerados a partir dos melhores candidatos anteriores. Tal técnica restringe o uso de memória explorando apenas parte do espaço de busca e, para tal, são utilizadas técnicas de poda, o que pode ocasionar eliminação de candidatos significativos e redundância de regras. Uma das grandes desvantagens dos algoritmos que aplicam essa técnica é a falta de diversidade, uma vez que, geralmente, são levados em conta apenas bons candidatos, podendo resultar em padrões já conhecidos [5]. Alguns algoritmos utilizam esquema de cobertura ponderada para aumentar a diversidade dos subgrupos gerados e tratar redundâncias. Os algoritmos mais populares existentes que aplicam *beam search* são *SubgroupMiner* [14], *SD* [15], *CN2-SD* [16], *RSD* [17] e *DSSD* [9].

Algoritmo genético é um método de busca heurística que acompanha o processo natural de evolução, e é utilizado na extração de soluções para diferentes otimizações e processos de busca. Cada solução é composta por várias variáveis e equipada de uma função de avaliação, e apenas às soluções melhores avaliadas é dada a oportunidade de evoluir [4]. Essa heurística realiza buscas globais com a habilidade de explorar grandes espaços de busca, de forma a não ser necessário o uso de técnicas de poda e facilitando a detecção de redundância de regras. Além disso, algoritmos evolucionários multiobjetivos permitem otimizar soluções levando em consideração mais de uma medida de qualidade. É interessante ressaltar, ainda, que algoritmos evolucionários proporcionam uma ampla flexibilidade de representação e permitem refletir a interação entre variáveis de forma adequada, o que é um fator importante em processos de aprendizado de regras [13]. Dentre os principais algoritmos evolucionários para mineração de padrões discriminativos, destacam-se: o SDIGA [18], que é uma abordagem mono-objetiva que usa busca global realizada por algoritmo genético seguida de busca local, via Hill Climbing; o CGBA-SD [19] que utiliza programação evolucionária e contexto gramatical para mineração de padrões discriminativos, e inclui mecanismos para adaptação da diversidade da população através da adaptação automática das taxas de mutação e *crossover*; e os algoritmos evolucionários multiobjetivos MESDIF [20] e NMEEF-SD [21]. O primeiro utiliza elitismo e o conceito de Frente de Pareto em sua estratégia de busca, enquanto o segundo utiliza um operador para reiniciar a população [4] [5]. O NMEEF-SD tem sido uma das abordagens mais competitivas e, quando comparado a outros algoritmos com abordagens exaustivas e de *beam search*, mostrou-se mais eficiente em termos de tempo de execução e seus subgrupos resultantes apresentaram maior qualidade [4].

Apesar da grande quantidade de algoritmos evolucionários para mineração de padrões discriminativos, poucas são as abordagens com foco em domínios de alta dimensionalidade. Nesse contexto, o algoritmo SSDP [5] [22] apresenta uma abordagem evolucionária mono-objetiva para mineração de padrões discriminativos e é caracterizado, principalmente, por ser adaptado a dados de alta dimensionalidade e por apresentar poucos parâmetros facilmente ajustáveis. Em [5], o desempenho do algoritmo foi avaliado utilizando conjuntos de dados tradicionais e de alta dimensionalidade e foi comparado a outras abordagens, incluindo algoritmos baseados em *beam search* e computação evolucionária. Os resultados obtidos demonstraram que, no contexto de alta dimensionalidade, o SSDP obteve resultados

estatisticamente melhores em relação à qualidade dos padrões minerados e, para banco de dados tradicionais, mostrou-se competitivo sem a necessidade de ajustes em parâmetros.

Em se tratando da tarefa de EMM, o desenvolvimento de algoritmos capazes de produzir resultados de alta qualidade e com baixo tempo de execução consiste em uma tarefa difícil. Isso porque, para o cenário de EMM, até métodos heurísticos rápidos que reduzem consideravelmente o espaço de busca – como é o caso do *beam search* – podem falhar em entregar um tempo de resposta rápido, uma vez que cada etapa de busca envolve uma etapa de indução de modelo que pode ser computacionalmente custosa [23]. Nesse contexto, diferentes heurísticas têm sido estudadas. O algoritmo EMDM [24] propõe o uso de uma estratégia de busca que explora estruturas tanto no espaço descritivo quanto no espaço de modelos, começando com um subgrupo candidato e aprimorando-o. Cada iteração consiste em uma etapa de *Exception Maximization* (EM) e uma de *Description Minimisation* (DM), tendo como resultado modelos excepcionais com descrições mínimas. Já o algoritmo TCGA (*tree-constrained gradient ascent*) [25] consiste em uma estratégia heurística de busca para EMM que explora informações sobre a contribuição de registros individuais para a qualidade do subgrupo e garante que o subgrupo possa ser descrito de forma concisa. Há, ainda, abordagem baseada em busca exaustiva [26] – através da adaptação do algoritmo GP-Growth – e abordagem que consiste em adaptação de técnica de amostragem direta de padrões para aplicação de EMM [23].

Embora o uso de algoritmos evolucionários venha sendo amplamente explorado para mineração de padrões discriminativos, e apesar dos bons resultados que essa abordagem tem obtido, não foi encontrada na literatura nenhum algoritmo evolucionário desenvolvido para *exceptional model mining*. Uma ferramenta computacional eficiente capaz de minerar padrões mais complexos que exigem medidas de qualidade mais custosas, viabilizaria novas perspectivas de análises em conjuntos de dados de *big data*, como os de domínios médicos e de bioinformática. Nesse mesmo contexto, a crescente disponibilidade de dados de alta dimensionalidade alinhada à pouca disponibilidade de algoritmos orientados a esse tipo de mineração evidencia um campo fértil para novas contribuições.

5 Metodologia

Abaixo, as atividades propostas para o período de realização do mestrado. O Quadro 1 apresenta o cronograma proposto.

Quadro 1 - Cronograma proposto de atividades

Ativ.	Meses																							
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1																								
2																								
3																								
4																								
5																								
6																								
7																								
8																								

FONTE: próprio autor

1. Estudo de novas literaturas e dos métodos computacionais existentes para mineração de padrões discriminativos e mineração em domínios de alta dimensionalidade;

2. Estudo e aprendizagem das técnicas computacionais para desenvolvimento do algoritmo proposto;
3. Desenvolvimento do algoritmo evolucionário para EMM em domínios de alta dimensionalidade;
4. Análise de desempenho do algoritmo frente às demais abordagens existentes;
5. Estudo de viabilidade para expandir a abordagem através da aplicação de processamento paralelo e outras técnicas;
6. Estudo de oportunidades de aplicações em áreas de pesquisa com bancos de dados reais;
7. Elaboração de artigos;
8. Desenvolvimento da dissertação.

Referências

[1] U. Fayyd, G. Shapiro e P. Smyth, "From data mining to knowledge discovery: An overview," em *Advances in knowledge discovery and data mining*, AAAI/MIT Press, 1996, p. 1–34.

[2] F. Herrera, C. Carmona, P. González e M. Del Jesus, "An overview on subgroup discovery: foundations and applications," *Knowledge and information systems,* nº 29, pp. 495-525, 2011.

[3] P. Novak, N. Lavrač e G. Webb, " Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining," *Journal of Machine Learning Research,* nº 10(Feb), pp. 377-403, 2009.

[4] S. Helal, "Subgroup discovery algorithms: a survey and empirical evaluation," *Journal of Computer Science and Technology,* vol. 31(3), pp. 561-576, 2016.

[5] T. Lucas, T. Silva, R. Vimieiro e T. Ludermir, "A new evolutionary algorithm for mining top-k discriminative patterns in high dimensional data," *Applied Soft Computing,* vol. 59, pp. 487-499.

[6] W. Duivesteijn, A. J. Feelders e A. Knobbe, "Exceptional Model Mining: Supervised Descriptive Local Pattern Mining with Complex Target Concepts," *Data Mining and Knowledge Discovery,* nº 30(1), pp. 47-98, 2016.

[7] W. Duivesteijn, A. Feelders e A. Knobbe, "Different slopes for different folks: mining for exceptional regression models with cook's distance," *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining,* pp. 868-876, August 2012.

[8] D. Leman, A. Feelders e A. Knobbe, "Exceptional model mining," *Joint European Conference on Machine Learning and Knowledge Discovery in Databases,* pp. 1-16, September 2008.

[9] M. van Leeuwen e A. Knobbe, "Diverse subgroup set discovery," *Data Mining and Knowledge Discovery,* nº 25(2), pp. 208-242, 2012.

[10] X. Liu, J. Wu, F. Gu, J. Wang e Z. He, "Discriminative pattern mining and its applications in bioinformatics," *Briefings in bioinformatics,* nº 16(5), pp. 884-900, 2014.

[11] M. Atzmueller, "Subgroup discovery," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* nº 5(1), pp. 35-49, 2015.

[12] W. Duivesteijn, A. Knobbe, A. Feelders e M. van Leeuwen, "Subgroup discovery meets Bayesian networks - an exceptional model mining approach," *In Data Mining (ICDM), 2010 IEEE 10th International Conference ,* pp. 158-167, December 2010.

[13] C. Carmona, P. González, M. del Jesus e F. Herrera, "Overview on evolutionary subgroup discovery: analysis of the suitability and potential of the search performed by evolutionary algorithms," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,*, vol. 4(2), pp. 87-103, 2014.

[14] W. Klösgen e M. May, "Census data mining — an application," *Proceedings of the 6th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD),* 2002.

[15] D. Gamberger e N. Lavrac, "Expert-guided subgroup discovery: Methodology and application," *Journal of Artificial Intelligence Research,* nº 17, pp. 501-527, 2002.

[16] N. Lavrač, B. Kavšek, P. Flach e L. Todorovski, "Subgroup discovery with CN2-SD," *Journal of Machine Learning Research,* nº 5(Feb), pp. 153-188, 2004.

[17] N. Lavrač, F. Železný e P. Flach, "RSD: Relational subgroup discovery through first-order feature construction," *International Conference on Inductive Logic Programming,* pp. 149-165, July 2002.

[18] M. Del Jesus, P. González, F. Herrera e M. Mesonero, "Evolutionary fuzzy rule induction process for subgroup discovery: a case study in marketing," *IEEE Transactions on Fuzzy Systems,* nº 15(4), pp. 578-592, 2007.

[19] J. Luna, J. Romero, C. Romero e S. Ventura, "On the use of genetic programming for mining comprehensible rules in subgroup discovery," *IEEE transactions on cybernetics,* nº 44(12), pp. 2329-2341, 2014.

[20] M. del Jesus, P. González e F. Herrera, "Multiobjective genetic algorithm for extracting subgroup discovery fuzzy rules," *Computational Intelligence in Multicriteria Decision Making, IEEE Symposium,* pp. 50-57, April 2007.

[21] C. Carmona, P. González, M. del Jesus e F. Herrera, "NMEEF-SD: Non-dominated multiobjective evolutionary algorithm for extracting fuzzy rules in subgroup discovery," *IEEE Transactions on Fuzzy Systems,* nº 18(5), pp. 958-970, 2010.

[22] T. Pontes, R. Vimieiro e T. Ludermir, "SSDP: a simple evolutionary approach for top-k discriminative patterns in high dimensional databases," *Intelligent Systems (BRACIS), 2016 5th Brazilian Conference,* pp. 361-366, October 2016.

[23] S. Moens e M. Boley, "Instant exceptional model mining using weighted controlled pattern sampling," *International Symposium on Intelligent Data Analysis,* pp. 203-214, October 2014.

[24] M. van Leeuwen, "Maximal exceptions with minimal descriptions," *Data Mining and Knowledge Discovery,* nº 21(2), pp. 259-276, 2010.

[25] T. Krak e A. Feelders, "Exceptional model mining with tree-constrained gradient ascent," *Proceedings of the 2015 SIAM International Conference on Data Mining,* pp. 487-495, June 2015.

[26] F. Lemmerich, M. Becker e M. Atzmueller, "Generic pattern trees for exhaustive exceptional model mining," *Joint European Conference on Machine Learning and Knowledge Discovery in Databases,* pp. 277-292, September 2012.

[27] W. Klösgen, "Explora: A multipattern and multistrategy discovery assistant," em *Advances in knowledge discovery and data mining,* American Association for Artificial Intelligence, 1996, pp. 249-271.

[28] S. Wrobel, "An algorithm for multi-relational discovery of subgroups," em *European Symposium on Principles of Data Mining and Knowledge Discovery,* Berlin, Heidelberg, Springer, 1997, pp. 78-87.