Universidade Estadual Maringá
EAIC 21° 2°EAITI
21° ENCONTRO ANUAL DE INICIAÇÃO CIENTÍFICA
2° ENCONTRO ANUAL DE INICIAÇÃO TECNOLÓGICA E INOVAÇÃO

META-APRENDIZADO NO AUXÍLIO À SELEÇÃO DE ATRIBUTOS: UM ESTUDO PARA MEDIDAS DE CORRELAÇÃO E CONSISTÊNCIA

Antonio Rafael Sabino Parmezan (PIBIC/UNIOESTE), Wu Feng Chung, Huei Diana Lee (Orientadora), e-mail: hueidianalee@gmail.com.

Universidade Estadual do Oeste do Paraná/Centro de Engenharias e Ciências Exatas/Laboratório de Bioinformática/Foz do Iguaçu, PR.

Ciências Exatas e da Terra, Ciência da Computação.

Palavras-chave: pré-processamento de dados, caracterização de dados, aprendizado de máquina.

Resumo:

Na fase de pré-processamento, em Mineração de Dados, há uma considerável diversidade de algoritmos candidatos para selecionar, segundo algum critério, atributos importantes. A escolha adequada desses métodos pode potencializar a qualidade dos dados fornecidos como entrada para algoritmos de extração de padrões. Visando prover suporte a essa questão, neste trabalho é investigada a viabilidade do uso de Meta-Aprendizado, baseado em características intrínsecas aos conjuntos de dados e aos algoritmos de Seleção de Atributos avaliados empiricamente em trabalhos anteriores, como auxílio para a construção de uma arquitetura de recomendação de algoritmos de Seleção de Atributos apropriada.

Introdução

Um dos modos de se extrair conhecimento a partir de bases de dados é por meio do processo de Mineração de Dados (MD), o qual é dividido em três fases [5]: pré-processamento, extração de padrões e pós-processamento. A primeira fase é considerada como uma das mais custosas por consumir aproximadamente 80% de todo o processo e é de fundamental importância para assegurar que os dados sejam de qualidade. Nesse sentido, tarefas de pré-processamento como a Seleção de Atributos (SA), podem contribuir para, além da remoção de atributos redundantes e irrelevantes, uma melhor compreensibilidade dos resultados gerados [2].
Diversos Algoritmos de SA (AlgSA) têm sido propostos na literatura [2, 3]. Do ponto de vista prático, a escolha de um determinado AlgSA, ou um conjunto deles, deve ser conduzida de acordo com o conhecimento do domínio,

usualmente detido pelos especialistas do domínio, e o conhecimento técnico sobre os AlgSA disponíveis, geralmente detido por especialistas da área computacional. No entanto, a maior disponibilidade de AlgSA pode não ser aproveitada ao máximo, pois esse grande número de algoritmos, com diferentes características, aumenta a dificuldade em se determinar, *a priori*, qual ou quais seriam mais apropriados, de acordo com as características do problema (Conjuntos de Dados (CD)) e as características dos AlgSA, uma vez que nenhum algoritmo pode ser considerado o melhor independentemente do problema em questão.

Em geral, as abordagens tradicionais para a escolha de algoritmos, além de utilizar conhecimento especialista, envolve também procedimentos delongados de avaliação empírica. Meta-Aprendizado (MA) surge, nesse contexto, como uma possibilidade de solução mais efetiva, capaz de fornecer ao usuário um auxílio automático e sistemático para a escolha de algoritmos. O MA pode ser definido como um processo de construção de modelos computacionais, no nível meta, que associam o desempenho relativo de algoritmos com propriedades extraídas dos CD [1].

Neste trabalho, é apresentado um estudo sobre a viabilidade do uso de MA, baseado em características intrínsecas aos CD e aos AlgSA analisados experimentalmente em trabalhos anteriores [3], como auxílio para a construção de uma arquitetura de recomendação de AlgSA apropriada.

Materiais e Métodos

A arquitetura de recomendação de AlgSA via MA foi organizada em quatro etapas, como ilustrado na Figura 1.

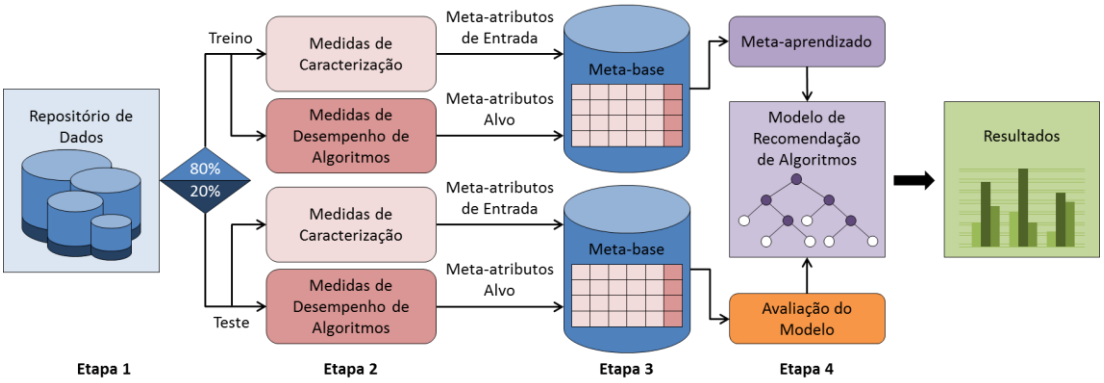

Figura 1 – Arquitetura de recomendação de AlgSA utilizando MA

Etapa 1: a partir do resultado de uma revisão sistemática de trabalhos publicados na área de MD, foram selecionados 17 CD naturais obtidos do

repositório de dados UCI[1]. Os atributos com valores desconhecidos concentrados em alguns poucos exemplos foram removidos. Após, cada um desses 17 CD foi particionado em dois, contendo, aproximadamente, 80% e 20% dos dados originais, gerando 17 exemplos para treino (conjunto de treino) e 17 exemplos para teste (conjunto de teste);

Etapa 2: cada conjunto de treino e de teste passou, inicialmente, pela extração de propriedades segundo as medidas dispostas na Tabela 1 e referidas como Meta-Atributos de Entrada (MAE).

Tabela 1 – Medidas de caracterização direta de CD

Medidas Estatísticas	Medidas de Informação
• Erro da classe majoritária; • Assimetria média dos atributos; • Curtose média dos atributos; • Correlação média dos atributos; • Dimensão fractal do CD.	• Entropia da classe; • Entropia média dos atributos; • Entropia condicionada média entre classe e atributos; • Informação mútua média entre classe e atributos; • Razão sinal/ruído.

Em seguida, foi realizada a SA utilizando as medidas de consistência e correlação (algoritmos *Consistency-based Filter* (CBF) e *Correlation-based Feature Selection* (CFS)). A partir dos subconjuntos de atributos selecionados, foram construídos modelos de árvores de decisão usando o algoritmo J48 [5]. Para cada conjunto investigado, foi identificado o algoritmo mais adequado para a SA com base na Taxa de Acerto (TA) do modelo induzido, estimada por meio de validação cruzada com dez partições estratificada. Desse processo, obteve-se um Meta-Atributo Alvo (MAA) nominal que pode assumir os valores “CBF”, “CFS” ou, no caso dos algoritmos selecionarem os mesmos atributos, “CBF_CFS”;

Etapa 3: foi construída uma meta-base que contempla o conjunto de treino e uma meta-base que contempla o conjunto de teste, ambas associando os respectivos MAE e MAA;

Etapa 4: o meta-modelo foi induzido a partir do conjunto de treino e usando novamente o algoritmo J48 (Modelo de Recomendação (MR)). Esse meta-modelo mapeia o meta-conhecimento embutido na meta-base de treino, e será utilizado, posteriormente, para auxiliar na recomendação de AlgSA. Ainda, a TA do MR foi estimada sobre o conjunto de teste;

Para auxiliar na aplicação do método apresentado foi construído um sistema computacional, desenvolvido em linguagem de programação Java[2], integrado com a ferramenta *Measure Distance Exponent* (MDE) [4] e as bibliotecas Weka [5] e Java/R *Interface*[3] (JRI).

[1] Frank A and Asuncion A. UCI Machine Learning Repository. Irvine: University of California, School of Information and Computer Science, 2010. Disponível em: <http://archive.ics.uci.edu/ml>. Acesso em: 25 jul. 2012.
[2] http://www.java.sun.com.
[3] http://www.rforge.net/JRI.

Resultados e Discussão

O MR de AlgSA apresentou três regras de classificação. Observou-se que do total de dez MAE, apenas um (entropia média dos atributos) foi escolhido como nó de decisão no meta-modelo. Esse fato pode ter sido causado devido ao considerável poder discriminatório desse meta-atributo. Quanto à TA, o MR obteve 59,94% para um erro da classe majoritária de 47,06%. Nesse caso, é importante notar que o modo de avaliação do meta-modelo visou abranger à proporção dos exemplos por atributos na representação dos CD. Adicionalmente, é interessante ressaltar que não há um consenso sobre que proporção seria adequada, porém, uma regra geral é que quanto maior essa proporção melhor deve ser essa representação.

Conclusões

Neste trabalho foi proposta uma arquitetura de recomendação de AlgSA via MA. Embora o MR tenha apresentado uma considerável TA, verificou-se que algumas das medidas de caracterização utilizadas podem não ter sido capazes de representar o conhecimento implícito dos CD. Isso significa que os MAE considerados podem não ter cooperado com a indicação adequada de determinados AlgSA. Trabalhos futuros incluem a avaliação de outros métodos para construção de sugestões e para indução de meta-modelos, bem como a análise de outras medidas de caracterização.

Agradecimentos

Ao Programa Institucional de Bolsas de IC (PIBIC) e ao LABI.

Referências

1. BRAZDIL, P. B.; GIRAUD-CARRIER, C.; SOARES, C.; VILALTA, R. **Metalearning**: applications to data mining. Springer, 2009.
2. LIU, H.; MOTODA, H. **Computational methods of feature selection**. Chapman & Hall/CRC data mining and knowledge discovery, 2008.
3. PARMEZAN, A. R. S.; WU, F. C.; LEE, H. D. Estudo de medidas de importância e algoritmos para seleção de atributos para mineração de dados. In: **Anais do 20º EAIC**, Ponta Grossa, PR, 2011.
4. TRAINA, C.; TRAINA, A. J. M; FALOUTSOS, C. **Measure distance exponent manual – MDE**. Internal document, 2003.
5. WITTEN, I. H.; FRANK, E. **Data mining**: practical machine learning tools and techniques. California: Morgan Kaufmann, 2005.