Combinando técnicas não-supervisionadas e supervisionadas para explanação de agrupamentos

Hércules A. Prado*
Paulo M. Engel**

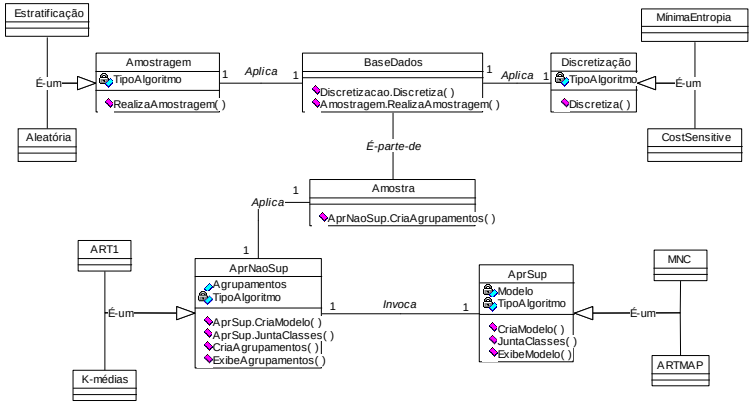
Resumo

A descoberta de conhecimento a partir de dados não classificados compreende duas tarefas principais: a identificação de "grupos naturais" e a análise destes grupos de modo a interpretar o seu significado. Estas tarefas são realizadas através do aprendizado não-supervisionado e supervisionado, respectivamente, correspondendo também às fases de taxonomia e explanação do processo de descoberta descrito por Langley [LAN 98]. A pesquisa em Descoberta de Conhecimento em Bases de Dados (DCBD) tem atacado estas duas fases segundo duas dimensões: (1) tornando os algoritmos de aprendizado aptos para executar bases de dados cada vez maiores, e (2) facilitando o processo completo de descoberta de conhecimento de maneira análoga ao que fazem as ferramentas CASE com relação à Engenharia de Software. Conforme descrito por Langley, o objetivo principal na descoberta de conhecimento em dados não classificados é a obtenção de descrições de potenciais classes no sub-espaço das dimensões dos objetos. Neste mesmo sentido, Wittgeinstein (*op. cit. in* [HAN 90]) advoga que os conceitos possuem uma natureza politética, i.e., eles podem ser descritos de muitas formas, compartilhando um número de características comuns, sem que nenhuma delas seja essencial para descrever o conceito. Outros trabalhos como [EAS 85] e [MUR 85] reforçam esta visão do processo de descoberta de conhecimento de dados não classificados. Pelas referências acima, fica evidente que as duas tarefas - taxonomia e explanação - compõem na verdade um único processo.

Numa primeira fase, a pesquisa em DCBD concentrou-se nas técnicas de aprendizado supervisionado, dando origem a diversos melhoramentos nos algoritmos, além de criar facilitadores para tarefas como coleta de dados, seleção de atributos e ajuste de parâmetros. Num segundo momento o foco das atenções deslocou-se para o aprendizado não-supervisionado [PRA 00c], levando a que os algoritmos pudessem processar bases de dados cada vez maiores e continuando o processo de facilitação das atividades envolvidas [TAL 98]. Nesta segunda fase foi proposto um conjunto de requisitos para os algoritmos de aprendizado não-supervisionado [BRA 98] que se tornou referência para a pesquisa neste tipo de algoritmo, sob a bandeira de DCBD. Um destes requisitos, de especial interesse para nós, é que o algoritmo consiga extrair padrões em uma única leitura da base de dados, uma vez que o enorme volume de dados não comportaria iteratividade.

* hercules@cpac.embrapa.br
** engel@inf.ufrgs.br

Bradley *et. al.* [BRA 98] apresentaram uma abordagem para implementação de algoritmos iterativos sem a necessidade de leitura da base de dados mais de uma vez. A idéia consiste em mapear a memória disponível em três zonas de dados: (1) dados que precisam ser retidos; (2) dados que podem ser descartados após a retenção de seus parâmetos estatísticos; e (3) dados comprimidos e sumarizados. A cada acesso à base de dados é carregada uma amostra e realizado o mapeamento acima, que libera mais memória para a próxima amostra. Foi reportada a implementação desta proposta para o algoritmo K-



médias [DUD 73], sendo adotada a distribuição gaussiana para a sumarização. Os resultados reportados mostraram-se equivalentes aos do algoritmo padrão.

Figura 1 – Estrutura para integração de algoritmos de aprendizado não-supervisionado e supervisionado

Considerando a natureza do processo de descoberta de dados não classificados, os requisitos de DCBD e os avanços já ocorridos nesta área, propomos a construção de uma estrutura de componentes (Figura 1) para obter descrições extensionais e intensionais de dados não classificados ([PRA 00a] e [PRA 00b]). A principal funcionalidade a ser provida por esta estrutura consiste em, a partir de um conjunto de dados não classificados, prover ao analista de dados, automaticamente, diferentes visões de agregação dos elementos representados, com respectivas descrições extensionais e intensionais, respeitados os requisitos de DCBD de se realizar a extração de padrões em uma única leitura da base de dados.

Os principais componentes desta estrutura são: (1) **BaseDados**: representando o conjunto de dados para análise; (2) **Amostra**: subconjunto próprio da base de dados que pode ser usado de acordo com a conveniência do analista; (3) **AprNaoSup** e **AprSup**: classes abstratas para especialização dos algoritmos de aprendizado não-supervisionado e supervisionado; (4) **Amostragem**: guarda as alternativas para a obtenção de amostras; e (5) **Discretização**: onde são especializados os algoritmos para discretização dos dados.

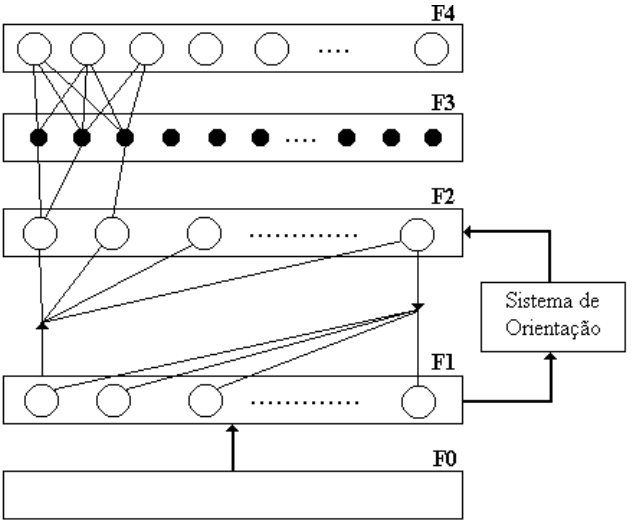


Figura 2 – Extendendo a estrutura para os algoritmos ART1 e CNM

No escopo deste trabalho será realizada a implementação integrada das redes neurais ART1 (*Adaptive Resonance Theory* [CAR 88]), para aprendizado não-supervisionado, e CNM (*Combinatorial Neural Model* ([MAC 89] e [MAC 92]) para aprendizado supervisionado, conforme Figura 2. As camadas F1 e F2 da figura correspondem à rede ART1. F1 recebe a codificação dos dados de entrada na estrutura, enquanto F2 representa os agrupamentos (*clusters*) identificados no processo. F2, F3 e F4 correspondem à rede CNM. Ambas as redes compartilham F2, já que os agrupamentos encontrados na parte não-supervisionada são tomados como classes na parte supervisionada. F3 no CNM contém as combinações encontradas durante o aprendizado enquanto F4 representa todos os atributos-valores existentes na base de dados, sendo cada um guardado em um neurônio.

Para validação da estrutura serão realizadas aplicações, sendo uma real, no domínio da pesquisa agropecuária, e outras em bases de dados do repositório público da UCI [BLA 00]. O desempenho do modelo integrado será comparado com a execução separada dos componentes de modo a avaliar as vantagens da estrutura.

Referências

[BLA 00] BLAKE, C.L.; e MERZ, C.J. UCI Repository of Machine Learning Databases [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science, 2000.

[BRA 98] BRADLEY, P., FAYYAD, U. e REINA, C. Scaling Clustering Algorithms to Large Databases. In: Proceedings of Fourth International Conference on Knowledge Discovery and Data mining. New York, NY: AIII Press, 1998.

[CAR 88] CARPENTER, G. e GROSSBERG, S. Neural Dynamics of Category Learning and Recognition: Attention, Memory, Consolidation, and Amnesia. In: Joel L. Davis (ed.), Brain structure, learning, and memory. AAAS Symposia Series, Boulder, CO: Westview Press, 1988. p.233-287.

[DUD 73] DUDA R.O e HART, P.E., Pattern Classification and Scene Analysis. New York: John Wiley and Sons. 1973.

[EAS 85] EASTERLIN, J.D, e LANGLEY, P.: A Framework for Concept Formation. In: Seventh Annual Conference of the Cognitive Science Society, Irvine, CA, 1985.

[HAN 90] HANSON, S.J. Conceptual Clustering and Categorization: Bridging the Gap Between Induction and Causal Models. In: KODRATOFF, Y., MICHALSKI, R. (Eds.). Machine Learning: An Artificial Intelligence Approach. San Mateo, CA: Morgan, 1990.

[LAN 98] LANGLEY, P. The Computer-Aided Discovery of Scientific Knowledge. In: Proceedings of the First International Conference on Discovery Science, Fukuoka, Japan, 1998.

[MAC 89] MACHADO, R. J. e ROCHA, A. F. **Handling Knowledge in High Order Neural Networks:** The Combinatorial Neural Network. Rio de Janeiro: IBM Rio Scientific Center, 1989. (Technical Report CCR076)

[MAC 92] MACHADO, R. J. e ROCHA, A. F. A Hybrid Architecture for Fuzzy Coneccionist Expert Systems. In: KANDEL, A.; LANGHOLZ, G. **Hybrids Architectures for Intelligent Systems**. Boca Raton: CRC Press, 1992.

[MUR 85] MURPHY, G. e MEDIN, D.: The Role of Theories in Conceptual Coherence. Psychological Review, 92(3):289-316, July, 1985.

[PRA 00a] PRADO, H.A. Explaining Cluster's Structures by Integrating Unsupervised and Supervised Learning Algorithms. In: FOURTH IEEE WORKSHOP ON DB & IS, 2000, Vilnius, Lituânia. **Proceedings …**, 2000.

[PRA 00b] PRADO, H.A., HIRTLE, S.C. e ENGEL, P.M. Scalable Model for Extensional and Intensional Descriptions of Unclassified Data. In: 3RD WORKSHOP ON HIGH PERFORMANCE DATA MINING, Cancún, México. **Proceedings …**, 2000.

[PRA 00c] PRADO, H.A., HIRTLE, S.C., e ENGEL, P.M. Clustering Algorithms for Data Mining. **Revista Tecnologia da Informação**, Editora Universa, Brasília, DF, 2000.

[TAL 98] TALAVERA, L. Exploring Efficient Attribute Prediction in Hierarchical Clustering. In: *VII Congreso Iberoamericano de Inteligencia Artificial, IBERAMIA98*, Lisboa, Portugal, 1998.