

MINERAÇÃO DE DADOS EM SEQUÊNCIAS DE cDNA

WAGNER ARBEX¹

LUIZ ALFREDO VIDAL DE CAVALHO²

MARCOS VINÍCIUS BARBOSA DA SILVA³

RESUMO: A imensa quantidade de dados que cresce de forma extremamente rápida, aumenta a distância entre a geração dos dados e a interpretação desses, obrigando o desenvolvimento de técnicas, ferramentas ou procedimentos que buscam minimizar o problema da quantidade de dados em contraposição à capacidade de interpretá-los. Os resultados desses estudos estão sendo organizados na área de mineração de dados e são muito utilizados em projetos de bioinformática onde, em geral, o grande volume de dados a ser tratado provoca um aumento da complexidade desses projetos.

PALAVRAS-CHAVE: mineração de dados, aprendizado de máquina, sistema de inferência difusa, fuzzyMorphic.pl, descoberta de conhecimento em bases de dados

DATA MINING IN cDNA SEQUENCES

ABSTRACT: The vast amount of data that grows in extremely fast, increasing the distance between generation and interpretation of data, which requires the development of techniques, tools and procedures that find out to minimize the problem of the huge amount of data as opposed to ability to interpret them. The studies results are being organized in the data mining area and they are widely used in bioinformatics projects where, usually, the large volume of data to be treated leads increased complexity of these projects.

KEY-WORDS: data mining, machine learning, fuzzy inference system, fuzzyMorphic.pl, knowledge discovery in database.

1. INTRODUÇÃO

Em diversas áreas do conhecimento existe uma imensa quantidade de dados que cresce de forma rápida, ampliando a distância entre a capacidade de geração e de interpretação desses, assim, são pesquisados recursos que buscam minimizar o problema da enorme quantidade de dados em contraposição à capacidade de interpretá-los e muitas dessas descobertas se encontram nas áreas de mineração de dados e de aprendizado de máquina.

Mineração de dados é um nome estabelecido para aplicações de algoritmos de aprendizado de máquina em grande massas de dados e, na ciência da computação, é chamada de descoberta de conhecimento em bases de dados (*knowledge discovery in databases* - KDD) (Alpaydin, 2004; Carvalho, 2005). Entretanto, esses termos se diferenciam na compreensão de que aprendizado de máquina refere-se à disciplina na qual são desenvolvidos e estudados os algoritmos, técnicas e ferramentas que permitem o aprendizado. Por outro lado, a mineração de dados deve ser vista como o processo, em si, que aplica o aprendizado de máquina para a descoberta de “conhecimento”⁴.

¹ Doutor em Engenharia de Sistemas e Computação, Empresa Brasileira de Pesquisa Agropecuária, E-mail: arbex@cnpq.embrapa.br

² Doutor em Engenharia de Sistemas e Computação, Universidade Federal do Rio de Janeiro, E-mail: alfredo@cos.ufrj.br

³ Doutor em Genética e Melhoramento, Empresa Brasileira de Pesquisa Agropecuária, E-mail: marcos@cnpq.embrapa.br

⁴ Não cabe na proposta desse texto nenhuma discussão sobre a capacidade, ou possibilidade, de sistemas de computação adquirirem, ou não, conhecimento. Assim, para os propósitos do mesmo, deve ser considerado que

2. OBJETIVO

Apresentar e discutir um modelo de mineração de dados em sequências de cDNA, como um sistema de suporte à decisão fundamentado em aprendizado de máquina e implementado por meio do fuzzyMorphic.pl, uma ferramenta de modelagem e desenvolvimento de sistemas de inferência difusa (SIDs). O modelo proposto nesse artigo traz um exemplo de mineração de dados para busca não-supervisionada de polimorfismos de base única (*single nucleotide polymorphisms* – SNPs) em sequências de cDNA.

3. MATERIAIS E MÉTODOS

A proposta desse trabalho requer fundamentos de mineração de dados, de aprendizado de máquina e de aquisição de conhecimento por meio de inferência difusa, que são abordados nessa seção.

Mineração de dados é um conjunto de técnicas reunidas com o objetivo de descobrir conhecimento em grandes massas de dados ou, ainda, é um conjunto de processos de descobertas de padrões em grande quantidade de dados, mas que sejam realmente úteis, válidos e efetivos (Witten et al., 2005) e, por sua vez, o aprendizado de máquina busca a construção de sistemas computacionais que sejam capazes de adquirir conhecimento de forma automática (Rezende, 2005). A mineração de dados ainda está associada à aprendizagem de máquina sob o aspecto de que a identificação de padrões pode levar ao aprendizado, que ocorre “quando se altera um comportamento de maneira que este seja melhor executado no futuro” (Witten et al., 2005), como consequência de conhecimento prévio.

Por conseguinte, modelos computacionais de aprendizado de máquina são inerentes e complementares aos de mineração de dados, na busca de relações complexas ou correlações “escondidas” em massas de dados (Matukumalli et al., 2006). Ou seja, na descoberta de conhecimento, o aprendizado de máquina relaciona-se por duas vias com a mineração de dados: utilizando padrões descobertos para aprender e, como consequência, gerando novas informações que possibilitem a descoberta de novos padrões.

A subjetividade intrínseca ao raciocínio é capaz de lidar com situações complexas, baseadas em informações imprecisas, incertas ou aproximadas e, para tanto, a estratégia adotada é a de utilizar “operadores humanos” que são expressos por termos ou variáveis linguísticas. Essa perspectiva, de descrever ou tratar problemas, em geral, não permite uma solução em termos de números exatos, mas, por exemplo, conduz a solução a uma classificação, agrupamento ou agregação qualitativa em categorias ou possíveis conjuntos de soluções (Mitchell, 1997).

SIDs são adequados para representar a informação imprecisa, que pode ser expressa por um conjunto de regras linguísticas e, caso exista a possibilidade de que os operadores sejam organizados como um conjunto de regras da forma

se ANTECEDENTE então CONSEQUENTE

logo, o raciocínio subjetivo pode ser construído em um algoritmo computacionalmente executável (Tanscheit, 2007), com capacidade de classificar, de modo impreciso, as variáveis que participam dos termos antecedentes e consequentes das regras, em conceitos qualitativos, e não quantitativos, o que representa a ideia de variável linguística (Almeida et al., 2005).

Trabalhar com valores incertos possibilita a modelagem de sistemas complexos, mesmo que reduza a precisão do resultado, o que não retira a credibilidade. Se as incertezas, quando consideradas isoladamente, são indesejáveis, quando associadas a outras características dos sistemas a serem modelados, em geral, permitem a redução da complexidade do sistema e aumentam a credibilidade dos resultados obtidos (Klir et al., 1995).

um sistema de computação adquire e utiliza conhecimento caso seja capaz de utilizar-se de informações prévias, primárias ou derivadas dessas, para inferir novos resultados e novas informações.

4. MODELO PROPOSTO

O modelo proposto, cuja implementação pode ser vista em Arbex (2009), fundamenta-se em mineração de dados e inferência difusa, originou um SID desenvolvido com o uso do *fuzzyMorphic.pl*, uma ferramenta de modelagem e desenvolvimento de SIDs, desenvolvida em Perl, para os quais fosse possível, para a fuzzificação, representar as funções de pertinência sobre formatos de conjuntos padrão; para a máquina de inferência, utilizar os modelos de Mamdani ou de Larsen e, para a defuzzificação, representar a função de saída sobre formatos de conjuntos padrão e utilizar o centro dos máximos como método de defuzzificação. Assim, assumindo essas condições, a partir de um arquivo texto com diretivas de descrição dos dados de entrada e dos elementos do modelo do sistema, o SID pode fazer mineração dos dados e inferir conhecimento, a partir das regras de inferência descritas pelas diretivas.

Procedimentos de mineração de dados são feitos em etapas que estabelecem um protocolo, que pode variar em função do problema abordado pelo enfoque adotado por diferentes pesquisadores. Porém, em geral, um protocolo de mineração de dados compreende (Carvalho, 2005; Goldschmidt et al., 2005; Almeida et al., 2005): identificar o problema; preparar os dados, isto é, a extrair, integrar, selecionar, complementar e eliminar dados; definir o modelo de análise; analisar os dados e descobrir informações e, ainda, a avaliar os resultados.

Assim, o modelo proposto, que pode ser visto na Figura 1, estabelece etapas bem definidas, quais sejam:

1. O processamento inicial dos cromatogramas, gerados a partir de clones de cDNA, quando é feita a leitura das bases e originadas as sequências e, ainda, determina a qualidade das bases dessas sequências. Essa etapa é feita pelo *pipeline* *phredPhrap*⁵ e, com a sua conclusão são gerados diversos arquivos, entre eles, o arquivo formato *ace* e os diversos arquivos *phd*, um para cada sequência lida;
2. Em seguida, são executados o Polyphred (Ewing et al., 1998) e o Polybayes (Marth et al., 1999), sobre os arquivos *ace* e *phd*, e cada um desses programas, de acordo com a sua metodologia, identifica os pontos candidatos a SNPs e estabelece uma probabilidade para cada um desses pontos. Esses resultados são registrados nos arquivos *polyphred.out* e *report.out*, que serão utilizados como dados de entrada para o procedimento de mineração de dados;
3. Na etapa seguinte é feita a preparação dos dados, quando os dados oriundos do Phrap, do Polyphred e do Polybayes são extraídos e selecionados dos seus respectivos arquivos e, ainda, se necessário, complementados. Essa etapa de preparação de dos dados é feita pelos *scripts* *parsepolyBayes.pl*, *parsepolyPhred.pl*, *parsephrapQuality.pl* e *joinparsersOut.pl* (Arbex, 2009), que, ainda, estrutura o arquivo para que seja lido pelo *fuzzyMorphic.pl*;
4. No passo seguinte, com a execução do *fuzzyMorphic.pl*, é feito o procedimento de mineração de dados, implementado em um SID, que fornece como saída um arquivo com os mesmos dados de entrada, acrescentando o valor inferido sobre a característica investigada;
5. Para a última etapa de análise e avaliação dos resultados são utilizadas técnicas e ferramentas para verificação dos resultados inferidos, tais como, análise de agrupamento sobre o conjunto de dados resultante do processamento do sistema de inferência, finalizando o protocolo de mineração de dados.

⁵Maiores informações referentes ao *phredPhrap* e ao *Phrap* podem ser encontradas em <<http://www.phrap.org/>>.

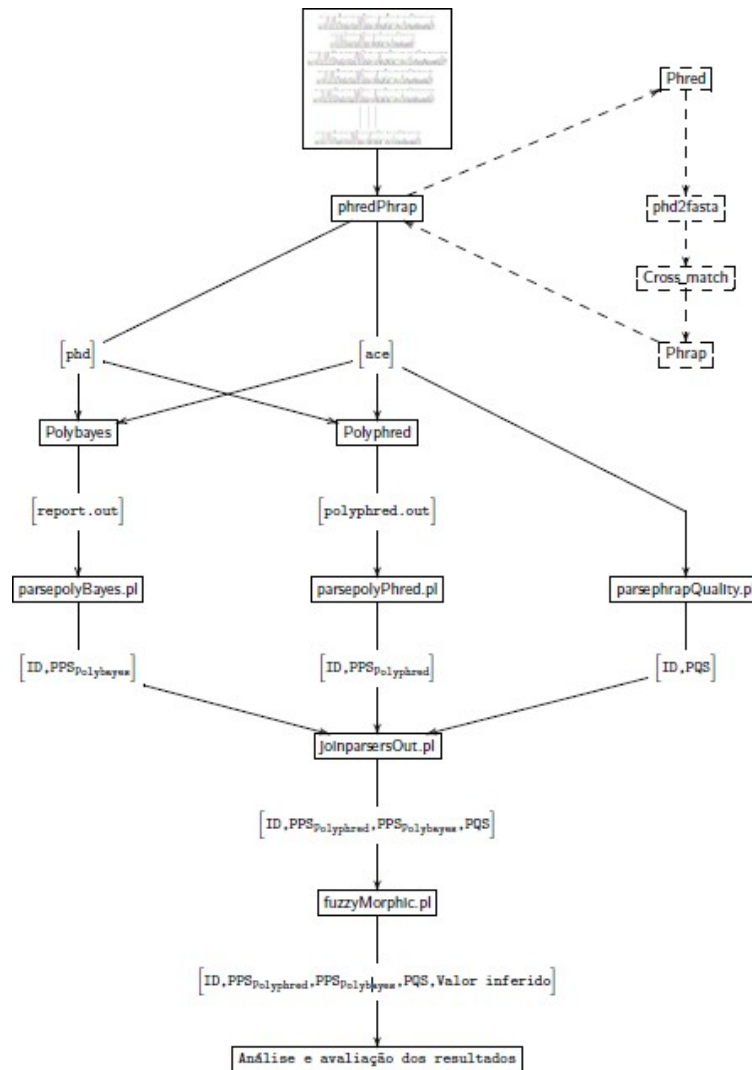


Figura 1: Síntese da estrutura funcional do modelo de mineração de dados.

5. RESULTADOS E DISCUSSÃO

A mineração dos dados investiga o conjunto originado a partir da junção das informações geradas pelo Polyphred e pelo Polybayes; avalia as probabilidades, estabelecidas por suas diferentes propostas, de cada elemento do conjunto e, então, determina, para cada um dos elementos, um novo atributo, que deve servir como uma referência na tentativa de agrupar o conjunto de dados em grupos de elementos que podem ser tratados como SNPs confirmados, SNPs descartados e SNPs não confirmados⁶.

Estabelecer agrupamentos é uma tarefa complexa e de difícil implementação, pois procura-se dizer como são e em quantas classes os dados se distribuem, sem que se tenha conhecimento prévio dos mesmos. As classes podem não existir, caso os elementos se distribuam equitativamente por todo o espaço, não caracterizando qualquer categoria, pois as classes são construídos com base na semelhança entre os elementos, cabendo a verificação das possíveis classes resultantes para avaliar a existência de algum significado útil (Carvalho, 2005).

Sob essa análise, o modelo proposto, estabelece um novo atributo, que permite agrupar os pontos dentro das três partições - SNP confirmado, SNP descartado e SNP não confirmado – e, o que se propõe, é executar o agrupamento dos dados resultantes por um algoritmo não-

⁶Pontos sem elementos suficientes para uma definição conclusiva.

supervisionado e com estabelecimento dinâmico do número de grupos, esperando que o resultado obtido confirme o particionamento do conjunto em três grupos, baseado no novo atributo.

6. CONCLUSÃO

O modelo de mineração substitui, mediante a inferência difusa, as medidas de probabilidade do Polyphred e do Polybayes, associadas à possibilidade de um ponto vir a ser um SNP, por um outro atributo, que permite agrupar os pontos dentro das três partições já explicadas.

Uma maior discussão sobre esse modelo de mineração de dados, pode ser encontrada em Arbex (2009), juntamente com a descrição da ferramenta fuzzyMorphic.pl, utilizada para o desenvolvimento e a implementação do mesmo.

7. REFERÊNCIAS

ALMEIDA, P. E. M.; EVSUKOFF, A. G. Sistemas fuzzy. **In:** REZENDE, S. O. (ed.). **Sistemas inteligentes: fundamentos e aplicações**. Barueri: Manole, 2005. p. 169–202.

ALPAYDIN, E. **Introduction to machine learning**. Cambridge, MIT Press, 2004.

ARBEX, W. **Modelos computacionais para identificação de informação genômica associada à resistência ao carrapato bovino**, 2009. Tese de doutorado. Universidade Federal do Rio de Janeiro. 200 p.

CARVALHO, L. A. V. **Datamining: a mineração de dados no marketing, medicina, economia, engenharia e administração**. Rio de Janeiro, Ciência Moderna, 2005.

EWING, B.; HILLIER, L.; WENDL, M. C.; et al. Base-calling of automated sequencer traces using Phred (I): Accuracy assessment, *Genome Research*, v. 8, p. 175-185, 1998.

GOLDSCHMIDT, R., PASSOS, E. **Data mining: um guia prático**. Rio de Janeiro: Elsevier, 2005.

KLIR, G. J.; YUAN, B. *Fuzzy sets and fuzzy logic: theory and applications*. Upper Saddle River: Prentice Hall, 1995. 592 p.

MARTH, G. T.; KORF, I.; YANDELL, M. D.; et al. A general approach to single-nucleotide polymorphism discovery, *Nature Genetics*, v. 23, p. 452-456, dec. 1999.

MATUKUMALLI, L. K., GREFFENSTETTE, J. J., HYTEN, D. L., et al. “Application of machine learning in SNP discovery”, *BMC Bioinformatics*, v. 7, n. 4, Jan. 2006.

MITCHELL, T. M. *Machine learning*. New York, 1997.

REZENDE, S. O., (ed). **Sistemas inteligentes: fundamentos e aplicações**. Barueri, Manole, 2005.

TANSCHKEIT, R. Sistemas fuzzy. **In:** OLIVEIRA JR., H. A. (ed.). **Inteligência computacional: aplicada à administração, economia e engenharia em Matlab**. São Paulo: Thomson Learning, 2007, pp. 229-264.

WITTEN, I. H., FRANK, E. **Data mining: practical machine learning tools and techniques**. 2 ed. San Francisco, Morgan Kaufmann Publishers, 2005.

s, v. SMC-3, p. 28-44, 1973.