MINERAÇÃO DE DADOS EM SEQUÊNCIAS DE cDNA

WAGNER ARBEX[1]
LUIZ ALFREDO VIDAL DE CAVALHO[2]
MARCOS VINÍCIUS BARBOSA DA SILVA[3]

RESUMO: A alta taxa de crescimento da imensa quantidade de dados disponíveis nas diversas áreas de conhecimento aumenta a distância entre a geração desses dados e a interpretação deles, obrigando o desenvolvimento de técnicas, ferramentas ou procedimentos que busquem minimizar o problema da quantidade de dados em contraposição à capacidade de interpretá-los. Os resultados desse trabalho estão sendo organizados na área de mineração de dados e são muito utilizados em projetos de bioinformática nos quais, 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 high growth rates of the vast amount of data available within the various domains of knowledge increases the distance between the generation and the interpretation of data, which requires the development of techniques, tools and procedures aimed at minimizing the problem of such a huge amount of data as opposed to the ability of interpreting them. The results of this study are being organized into the data mining field and are widely used by projects in Bioinformatics where, usually, the large volume of data to be treated leads to an increase of the complexity of these projects.

KEYWORDS: data mining, machine learning, fuzzy inference system, fuzzyMorphic.pl, knowledge discovery in databases.

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 de tais dados. 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@cnpgl.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@cnpgl.embrapa.br

2. OBJETIVO

Este trabalho tem o objetivo de 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 neste 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. REFERENCIAL TEÓRICO

A proposta deste trabalho requer fundamentos de mineração de dados, de aprendizado de máquina e de aquisição de conhecimento por meio de inferência difusa, os quais são abordados nesta seção.

O termo mineração de dados se refere ao 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, desde que sejam realmente úteis, válidos e eficazes (Witten et al., 2005). 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, o qual ocorre "quando se altera um comportamento de maneira que este seja mais bem 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 pode, por exemplo, conduzir 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. Ainda, 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). Esse algoritmo tem a capacidade de classificar, de modo impreciso, as variáveis - que participam dos termos antecedentes e consequentes das regras -

[4] Esse texto não discute a capacidade, ou possibilidade, de sistemas de computação adquirirem, ou não, conhecimento. Assim, para os propósitos do mesmo, deve ser considerado que 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.

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 se reduza a precisão do resultado, o que não retira a credibilidade. Se as incertezas, quando consideras 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).

4. MODELO PROPOSTO

O modelo proposto, cuja implementação pode ser vista em Arbex (2009), fundamenta-se na mineração de dados e na inferência difusa e, a partir delas, originou um SID desenvolvido com o uso do fuzzyMorphic.pl. Essa ferramenta, desenvolvida em Perl, deve ser utilizada na modelagem e desenvolvimento de SIDs, para os quais fosse possível, no que tange à fuzzificação, representar as funções de pertinência sobre formatos de conjuntos padrão; no que diz respeito à máquina de inferência, utilizar os modelos de Mamdani ou de Larsen e; no concernente à 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: identificar o problema; preparar os dados, isto é, extrair, integrar, selecionar, complementar e eliminar dados; definir o modelo de análise; analisar os dados e descobrir informações e, ainda, a avaliar os resultados (Carvalho, 2005; Goldschmidt et al., 2005; Almeida et al., 2005).

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 são originadas as sequências e, ainda, quando é determinada a qualidade das bases dessas sequências. Essa etapa é feita pelo *pipeline* phredPhrap[5] e, quando concluída, gera 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*, sendo que 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 dos dados é feita pelos *scripts* parsepolyBayes.pl, parsepolyPhred.pl, parsephrapQuality.pl e joinparsersOut.pl (Arbex, 2009), que, ainda, estruturam 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 a análise de agrupamento sobre o conjunto de dados resultante do processamento do sistema de inferência, finalizando o protocolo de mineração de dados.

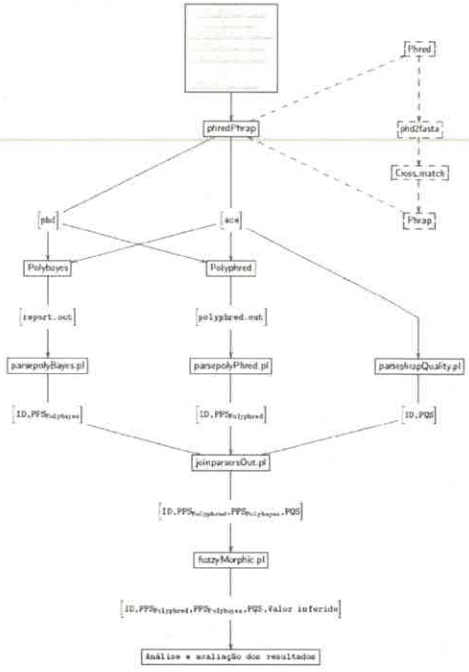


Figura 1: Síntese da estrutura funcional do modelo de mineração de dados.

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 (SNP_C), SNPs descartados (SNP_D) e SNPs não confirmados[6] (SNP_{NC}).

Estabelecer agrupamentos é uma tarefa complexa e de difícil implementação, pois procura-se

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

[6]Pontos sem elementos suficientes para uma definição conclusiva.

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 construídas 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_C, SNP_D e SNP_{NC} – e executar o agrupamento dos dados resultantes por um algoritmo não-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.

Uma ampla 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.

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.

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**. Cambrige, 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., GREFENSTETTE, 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.

TANSCHEIT, 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: pratical machine learning tools and techniques**. 2 ed. San Francisco, Morgan Kaufmann Publishers, 2005.

SBIAgro 2009

7º Congresso Brasileiro de Agroinformática

Agroinformática e a sustentabilidade do agronegócio e dos recursos naturais

Anais
Prefácio
SBIAGRO
Diretoria da SBIAgro
Comissão
Como citar
Créditos

21 a 25 de setembro de 2009
Universidade Federal de Viçosa • Viçosa/MG

SAIR

SBIAgro 2009

7º Congresso Brasileiro de Agroinformática

O sistema busca os artigos em Título, autor, palavra chave.
Ao selecionar as opções (Todos, Poster e oral) a busca será feita dentro item selecionado.

Wagner Arbex | Procurar | (•) Todos ( ) Poster ( ) Oral

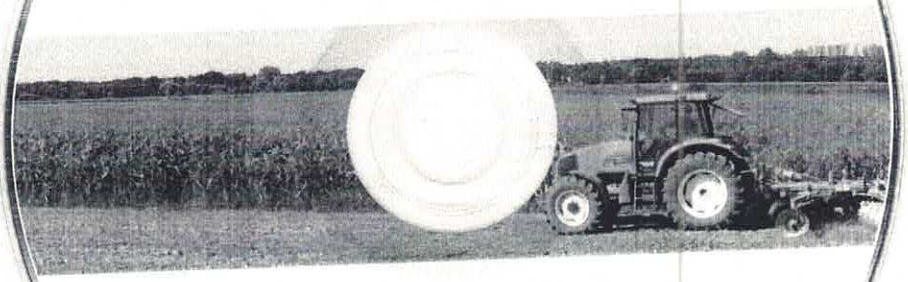
Foram encontrados 2 resultados.

Publicação
MINERAÇÃO DE DADOS EM SEQUÊNCIAS DE CDNA
MODELAGEM DIFUSA PARA SUPORTE À DECISÃO NA DESCOBERTA DE SNPS EM SEQUÊNCIAS DE CDNA

SBIAgro 2009

7º Congresso Brasileiro de Agroinformática

Agroinformática e a sustentabilidade do agronegócio e dos recursos naturais



21 a 25 de setembro de 2009

Universidade Federal de Viçosa • Viçosa/MG

Promoção: SB AGRO

Realização: UFV

Viçosa