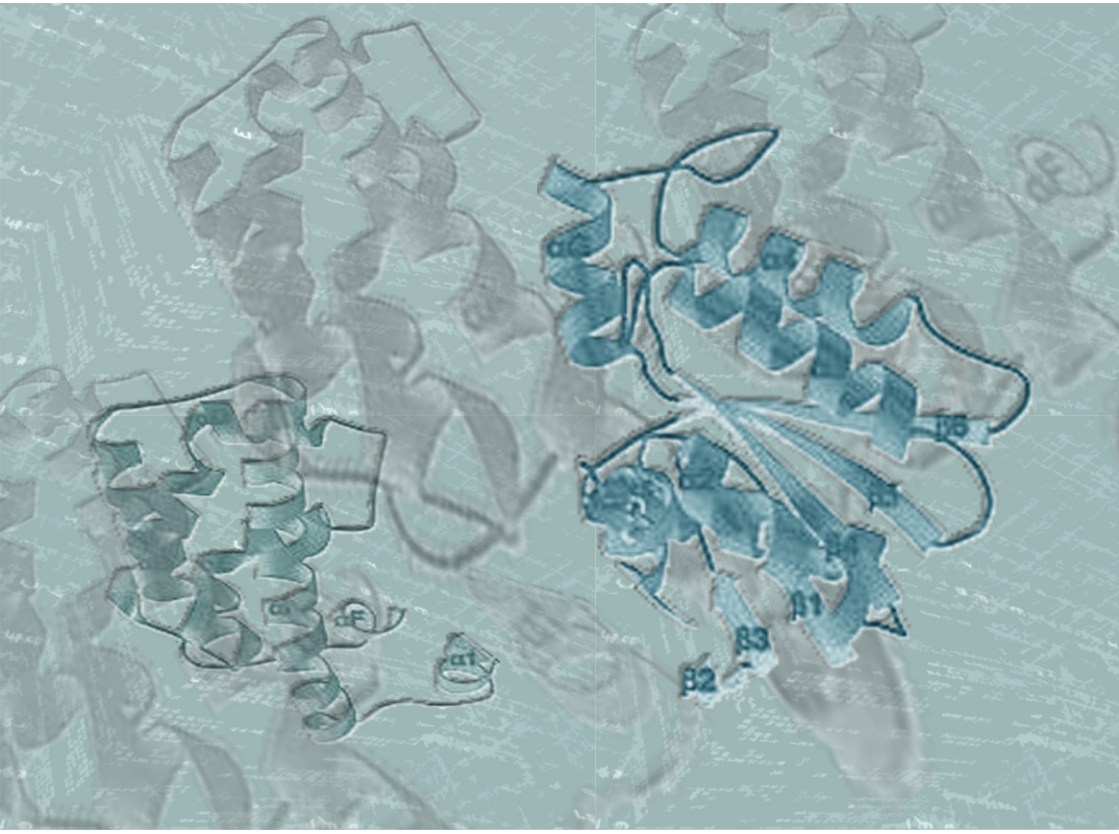
Boletim de Pesquisa 14
e Desenvolvimento
Setembro, 2006

ISSN 1677-9266

# Uma Metodologia para Seleção de Parâmetros em Modelos de Classificação de Proteínas



Embrapa

Embrapa

*Empresa Brasileira de Pesquisa Agropecuária*
*Embrapa Informática Agropecuária*
*Ministério da Agricultura, Pecuária e Abastecimento*

*ISSN 1677-9266*
*Setembro, 2006*

# *Boletim de Pesquisa e Desenvolvimento* 14

## Uma Metodologia para Seleção de Parâmetros em Modelos de Classificação de Proteínas

Stanley Robson de Medeiros Oliveira
Michel Eduardo Beleza Yamagishi
Luiz César Borro
Paula Regina Kuser Falcão
Edgard Henrique dos Santos
Fábio Danilo Vieira
Ivan Mazoni
José Gilberto Jardine
Goran Neshich

Campinas, SP
2006

**Embrapa Informática Agropecuária**
**Área de Comunicação e Negócios (ACN)**
Av. André Tosello, 209
Cidade Universitária "Zeferino Vaz" Barão Geraldo
Caixa Postal 6041
13083-970 - Campinas, SP
Telefone (19) 3789-5743 Fax (19) 3289-9594
URL: http://www.cnptia.embrapa.br
*e-mail*: sac@cnptia.embrapa.br



Supervisor editorial: *Ivanilde Dispato*
Normalização bibliográfica: *Marcia Izabel Fugisawa Souza*
Editoração eletrônica: *Área de Comunicação e Negócios (ACN)*

**1ª. edição on-line - 2006**


Uma metodologia para seleção de parâmetros em modelos de classificação de proteínas / Stanley Robson de Medeiros Oliveira [et al.]. — Campinas : Embrapa Informática Agropecuária, 2006.
18 p. : il. — (Boletim de Pesquisa e Desenvolvimento / Embrapa Informática Agropecuária ; 14).

ISSN 1677-9266

1.Bioinformática. 2. Classificação de proteínas. 3. Mineiração de Dados. I. Oliveira, Stanley Robson de Medeiros. II. Série.

CDD - 570.285 (21st. ed.)
006.33

# Sumário

# Uma Metodologia para Seleção de Parâmetros em Modelos de Classificação de Proteínas

*Stanley Robson de Medeiros Oliveira* [1]
*Michel Eduardo Beleza Yamagishi* [2]
*Luiz César Borro* [3]
*Paula Regina Kuser Falcão* [4]
*Edgard Henrique dos Santos* [5]
*Fábio Danilo Vieira* [6]
*Ivan Mazoni* [7]
*José Gilberto Jardine* [8]
*Goran Neshich* [9]

## Resumo

Os principais desafios relacionados ao problema de classificação de enzimas em banco de dados de estruturas de proteínas são: 1) o ruído presente nos dados; 2) o grande número de variáveis; e 3) o número não-balanceado de membros por classe. Para abordar esses desafios, apresenta-se uma metodologia para seleção de parâmetros, que combina recursos da matemática (ex.: Transformada Discreta do Cosseno) e da estatística (ex.g., correlação de variáveis e amostragem com reposição). A metodologia foi validada considerando-se os três principais métodos de classificação da literatura, a saber: árvore de decisão, classificação Bayesiana e redes neurais. Os experimentos demonstram que essa metodologia é simples, eficiente e alcança resultados semelhantes àqueles obtidos pelas principais técnicas para seleção de parâmetros na literatura.

Termos para indexação: classificação de enzimas, predição de função de proteínas, estruturas de proteínas, banco de dados de proteínas, seleção de parâmetros, métodos para classificação de dados.

[1] Doutor em Ciência da Computação, Pesquisador da Embrapa Informática Agropecuária, Caixa Postal 6041, Barão Geraldo - 13083-970 - Campinas, SP. (e-mail: stanley@cnptia.embrapa.br).
[2] Doutor em Matemática, Pesquisador da Embrapa Informática Agropecuária, Caixa Postal 6041, Barão Geraldo - 13083-970 - Campinas, SP. (e-mail: michel@cbi.cnptia.embrapa.br).
[3] Bolsista CNPq, Embrapa Informática Agropecuária, Caixa Postal 6041, Barão Geraldo - 13083-970 - Campinas, SP.
[4] Doutora em Cristalografia de Proteínas, Pesquisadora da Embrapa Informática Agropecuária, Caixa Postal 6041, Barão Geraldo - 13083-970 - Campinas, SP. (e-mail: paula@cbi.cnptia.embrapa.br).
[5] Bacharel em Ciência da Computação, Técnico de Nível Superior da Embrapa Informática Agropecuária, Caixa Postal 6041, Barão Geraldo - 13083-97 - Campinas, SP. (e-mail: edgard@cbi.cnptia.embrapa.br).
[6] Tecnólogo em Processamento de Dados, Técnico de Nível Superior da Embrapa Informática Agropecuária, Caixa Postal 6041, Barão Geraldo - 13083-970 - Campinas, SP. (e-mail: fabio@cbi.cnptia.embrapa.br).
[7] Tecnólogo em Processamento de Dados, Técnico de Nível Superior da Embrapa Informática Agropecuária, Caixa Postal 6041, Barão Geraldo - 13083-970 - Campinas, SP. (e-mail: ivan@cbi.cnptia.embrapa.br).
[8] Doutor em Engenharia de Alimentos, Pesquisador da Embrapa Informática Agropecuária, Caixa Postal 6041, Barão Geraldo - 13083-970 - Campinas, SP. (e-mail: jardine@cnptia.embrapa.br).
[9] Doutor em Biofísica, Pesquisador da Embrapa Informática Agropecuária, Caixa Postal 6041, Barão Geraldo - 13083-970 - Campinas, SP. (e-mail: neshich@cbi.cnptia.embrapa.br).

# A Methodology for Feature Selection in Protein Classification Models


## Abstract

The major challenges related to the enzyme classification problem in protein databases are: 1) the noise in the data; 2) the large number of parameters; and 3) the class-imbalanced protein datasets. To address these challenges, we introduce a methodology for parameter selection that combines the strength of mathematics (e.g., Discrete cosine transform) and Statistics (e.g., correlation analysis and resampling). The methodology was validated taking into account three of the most known classification methods in the literature, as follows: decision trees, Bayesian classification and neural networks. Our experiments demonstrate that this methodology is simple, efficient and yields results similar to those obtained by feature selection methods available in the literature.



Index terms: enzyme classification, predicting protein functions, protein databases, parameter selection, data classification methods.

# Introdução

Um dos principais objetivos da bioinformática é a caracterização automatizada de um grande número de proteínas disponíveis em bancos de dados de estruturas de proteínas. O alvo primordial dessa caracterização é o entendimento detalhado das funções das proteínas e de suas interações com outras moléculas em processos bioquímicos (Radivojac et al., 2004).

Em particular, o debate sobre a relação entre estrutura e função de proteínas e, como inferir a função de uma proteína por meio de sua estrutura, vem crescendo cada vez mais (Shrager, 2003). A maioria dos métodos para classificar proteínas é baseada na detecção de similaridades de seqüências ou de estruturas. Infelizmente, isso não é sempre possível porque o número de novas proteínas sem seqüência homólogas ou sem similaridade estrutural vem crescendo a cada ano. Por exemplo, no sítio do Protein Data Bank - PDB (Berman et al., 2000), o número de estruturas de proteínas com funções desconhecidas vem crescendo muito rápido. Em 2000, o número de estruturas com funções desconhecidas era 10, em 2002 esse número passou para 83 e, em 2004, o número cresceu para 401.

Um trabalho recente, conduzido por Dobson & Doig (2005), apresentou um novo método para predizer classes de enzimas, a partir de dados estruturais de proteínas, que facilita o entendimento sobre a relação entre estrutura e função. A idéia do método proposto por Dobson e Doig é a partir de um grupo de atributos estruturais de proteínas, classificar cada proteína dentro de uma das seis classes protéicas, conhecidas como as superfamílias. Dobson e Doig utilizaram o banco de dados de proteínas ASTRAL SCOP 1.63 (Brenner et al., 2000; Chandonia et al., 2002). A precisão desse método chega a 35%, usando o classificador *support vector machine*.

Motivados pelo trabalho apresentado em Dobson & Doig (2005) e pelo fato da Embrapa possuir o STING_DB (Neshich et al., 2003, 2004, 2005), um dos maiores bancos de dados estruturais do mundo, o grupo de bioinformática da Embrapa Informática Agropecuária propôs uma metodologia para seleção de parâmetros em modelos de classificação de proteínas.

A metodologia proposta tem o objetivo de melhorar o processo de seleção de parâmetros para aumentar a precisão do modelo de classificação de proteínas, por meio de uma abordagem híbrida que contempla recursos da matemática (ex.: Transformada discreta do cosseno) e da estatística (ex.: correlação de variáveis e amostragem com reposição). A metodologia aborda os três desafios presentes em classificação de banco de dados de proteínas, a saber: 1) o ruído presente nos parâmetros; 2) o grande número de variáveis; e 3) o número não-balanceado de membros por classe.

Os experimentos demonstram que, a partir da adoção dessa metodologia, é possível se obter um modelo de classificação com precisão de aproximadamente 70%, um resultado significativo quando comparado com o método proposto em Dobson & Doig (2005).

# Material e Métodos

A classificação é uma tarefa de mineração de dados que tem por objetivo classificar itens de dados em uma entre diversas classes previamente definidas, com base em propriedades comuns, entre um conjunto de objetos no banco de dados.

A técnica de classificação utiliza um conjunto de exemplos para desenvolver um modelo, conhecido como conjunto de treinamento. Em geral, o conjunto de treinamento contém dois terços dos exemplos disponíveis em um banco de dados. Um terço dos exemplos é então usado para testar a precisão do modelo. Após a construção do modelo de classificação, esse é usado para predizer classes de novos casos que estão para ser inseridos no banco de dados. Um padrão de classificação é similar a um padrão de regressão, porém ele prediz o valor de um atributo nominal ou uma categoria de um valor real.

Na técnica de classificação, possui-se um conjunto de dados predeterminados para a classificação, isso caracteriza um método de aprendizado supervisionado, onde o algoritmo é controlado por parâmetros que são passados ao sistema.

Como aplicações das técnicas de classificação, pode-se citar o diagnóstico médico, a detecção de fraudes, a avaliação de riscos de empréstimos, a aprovação de créditos, a classificação de proteínas, entre outros.

Para o estudo de caso de classificação de proteínas, serão utilizados alguns dos principais métodos propostos na literatura, a saber: árvores de decisão, classificador bayesiano simples, redes neurais e *support vector machine*. Detalhes sobre a definição matemática e implementação desses classificadores podem sem encontrados em Han & Kamber (2001).

A escolha adequada de quais parâmetros devem ser mantidos e quais devem ser descartados é uma tarefa desafiadora na classificação de proteínas. O conjunto de todos os parâmetros usados para representar cada proteína pode não ser a melhor opção para construir um modelo de classificação de proteínas. Em geral, parâmetros altamente correlacionados agregam pouca informação ao modelo o que pode comprometer a precisão do modelo de classificação.

O alvo principal da seleção de parâmetros é remover o ruído dos dados, isto é, eliminar os parâmetros altamente correlacionados, sem comprometer a utilidade dos dados.

Existem vários métodos para seleção de parâmetros propostos na literatura como, por exemplo, as técnicas de inteligência artificial para seleção de atributos relevantes (Langley, 1994; Blum & Langley, 1997). O problema é que essas

Técnicas dependem da natureza dos dados e, portanto, não podem ser aplicadas em todo conjunto de parâmetros sob análise.

Uma alternativa seria lançar mão de métodos para seleção de parâmetros que usam algumas medidas estatísticas sobre o conteúdo dos parâmetros como, por exemplo, o coeficiente de correlação. A idéia fundamental desses métodos é que dois parâmetros altamente correlacionados contêm bastante redundância e, portanto, um deles poderia ser removido do conjunto de parâmetros associados ao modelo de classificação. Isto reduziria o ruído dos dados e, ao mesmo tempo, aumentaria a precisão do modelo.

Nesse trabalho enfatiza-se uma abordagem para seleção de parâmetros que fundamentam-se em técnicas estatísticas, matemáticas e de mineração de dados. A abordagem proposta é baseada nas seguintes etapas:

**a) correlação de parâmetros:** nesta etapa, uma matriz de correlação é calculada envolvendo todos os parâmetros para representar cada arquivo PDB (Tabela 1). A matriz de correlação irá identificar os pares de parâmetros altamente correlacionados. Em particular, foi definido o parâmetro de controle (coeficiente de correlação) com valor igual a 0.8. Em outras palavras, se um par de parâmetros tem coeficiente de correlação com valor igual ou maior que 0.8, um dos parâmetros pode ser removido, uma vez que a presença de ambos é desnecessária. O coeficiente de correlação poderia ainda ser avaliado assumindo valores menores ou maiores a 0.8. Entretanto, em nossos experimentos, 0.8 apresentou os melhores resultados.

**Tabela 1**. Parâmetros representativos do banco de dados STING_DB.

| | *STING_DB Parameters* |
|---|---|
| 1 | Cross Presence Order @ca |
| 2 | Cross Presence Order @cb |
| 3 | Cross Presence Order @lha |
| 4 | Cross Link Order @ca |
| 5 | Cross Link Order @cb |
| 6 | Cross Link Order @lha |
| 7 | Contacts Energy (Internal) |
| 8 | Unused Contacts Energy |
| 9 | Surface Accessibility Complex |
| 10 | Surface Accessibility Isolation |
| 11 | Ep @ca |
| 12 | Ep @lha |
| 13 | Ep Average |
| 14 | Ep @surface |
| 15 | Dist. N-Terminal |
| 16 | Dist. C-Terminal |
| 17 | Dist. Center of Gravity |
| 18 | Hydrophobicity IKD |
| 19 | Hydrophobicity IR |

**b) identificação de redundância:** depois de identificar os pares de parâmetros altamente correlacionados em cada arquivo de proteína, o alvo é verificar quais são os pares de parâmetros altamente correlacionados em todo o banco de dados STING_DB. Essa análise pode ser obtida por meio de uma tarefa de mineração de dados conhecida como associação (Han & Kamber, 2001). O objetivo desta tarefa é identificar associações e/ou relacionamentos entre parâmetros de um banco de dados. Nessa abordagem, o alvo é identificar grupos freqüentes de parâmetros altamente correlacionados em todo o STING_DB. Um par de parâmetros é chamado de freqüente se ele é encontrado em, pelo menos, uma certa porcentagem $\sigma$ (conhecida como suporte) em todo o STING_DB. Nos experimentos, o parâmetro $\sigma$ foi ajustado para 60%, o que é muito alto em aplicações de regras de associação.

**c) remoção de redundância:** uma vez que a redundância deve ser identificada na etapa anterior, alguns atributos podem agora ser removidos do conjunto de parâmetros usados para a classificação de proteínas. O processo de remoção leva em consideração os pares de parâmetros altamente correlacionados e que são freqüentes em todo o STING_DB. Para cada par de parâmetros altamente correlacionados, um dos parâmetros pode ser removido aleatoriamente com o propósito de reduzir o ruído ou redundância dos dados. No final, os atributos restantes são utilizados para o modelo de classificação de proteínas. A Tabela 2 contém os parâmetros depois da remoção de redundância (ruídos). O processo de remoção considerou o parâmetro  ajustado para 60%.

**Tabela 2.** Parâmetros selecionados depois da remoção de ruído.

| | *STING_DB Parameters* |
|---|---|
| 1 | Cross Presence Order @ca |
| 2 | Cross Link Order @ca |
| 3 | Contacts Energy (Internal) |
| 4 | Unused Contacts Energy |
| 5 | Surface Accessibility Isolation |
| 6 | Ep @ca |
| 7 | Ep @lha |
| 8 | Ep Average |
| 9 | Ep @surface |
| 10 | Dist. Center of Gravity |
| 11 | Hydrophobicity IKD |

**d) representação de uma proteína em forma de matriz:** depois da fase de seleção de parâmetros, cada proteína pode então ser representada por uma matriz, onde as linhas são os parâmetros e as colunas são os aminoácidos. O problema com essa representação é que o número de aminoácidos varia de proteína para proteína, o que resulta em matrizes com dimensões diferentes. Portanto, para construir o modelo de classificação, todas as matrizes devem ter o mesmo número de linhas e colunas. Para isto, cada linha da matriz pode ser representada por um sinal, de acordo com os fundamentos do processamento de sinais. Foi decidido o uso da Transformada Discreta do Cosseno - TDC (Ahmed et al., 1974) em cada linha (sinal) e truncou-se o número de coeficientes de expansão para 20, depois de alguns experimentos. A seleção da TDC foi escolhida por duas razões: 1) é uma transformação ortonormal que preserva as normas e os ângulos dos vetores; 2) é uma transformada para números reais, ao contrário da Transformada Discreta de Fourier, que é uma transformada definida sobre o corpo dos números complexos. A Fig. 1 mostra que 20 coeficientes da TDC são suficientes para se obter uma boa representação do sinal (parâmetro) e para se manter a representação geral do sinal original. Cada proteína pode agora ser representada por 241 parâmetros, onde 240 representam o produto das 11 variáveis da Tabela 2, multiplicadas pelos 20 coeficientes de expansão totalizando 220 parâmetros. Além disso, adicionou-se a freqüência dos 20 aminoácidos presentes nas proteínas. O último parâmetro é a classe de enzima da proteína.

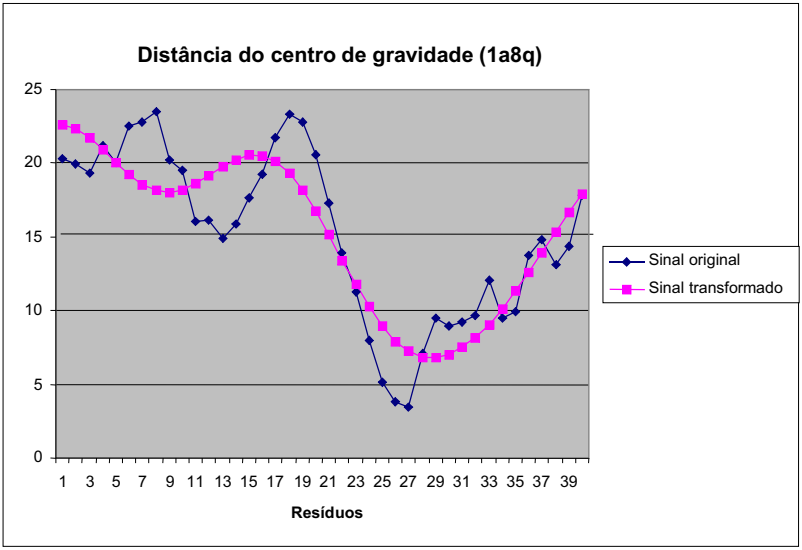



**Fig. 1.** O sinal transformado mantém a representação global do sinal original. As altas freqüências, como esperado, são eliminadas.

**e) balancear número de membros por classe:** a Tabela 3 mostra as seis classes de enzimas das proteínas e o número de membros por classe para o banco de dados ASTRAL SCOP, release 1.63, em nossos experimentos. No total, existem 492 estruturas de proteínas em seis diferentes classes. Observando-se a Tabela 3, é possível perceber que o número de membros por classe não é balanceado. Por exemplo, a classe Hidrolase tem 158 membros, enquanto a classe Ligase tem apenas 19. Se o modelo for construído considerando esse número de membros por classe, a precisão será comprometida devido a um possível enviesamento do modelo em direção às classes com um maior número de representantes. Para tentar amenizar esse problema, pode-se usar a técnica estatística de amostragem com reposição para balancear o número de membros por classe, na seleção de amostras para os conjuntos de treinamento e teste do modelo de classificação (Breiman, 1996). Duas restrições foram consideradas: a) o número total de elementos foi preservado; e b) uma amostra seguindo a distribuição Uniforme foi gerada.

**Tabela 3**. Classes de enzimas e seus respectivos números de membros (banco de dados Astral, *release 1.63*)

| *Classes de enzimas* | *Número de membros* |
|---|---|
| Oxidoredutase | 77 |
| Transferase | 127 |
| Hidrolase | 158 |
| Liase | 60 |
| Isomerase | 51 |
| Ligase | 19 |

As etapas de a, b, c e d, da metodologia, têm a finalidade de lidar com dois dos desafios em classificação de proteínas: reduzir o ruído presente nos parâmetros e selecionar o conjunto de variáveis para a construção do modelo de classificação. A etapa **e** é fundamental para balancear o número de membros por classe.

## Resultados e Discussão

Uma vez que a metodologia para seleção de parâmetros foi apresentada, serão avaliados os resultados obtidos durante a sua validação e apresenta-se as lições aprendidas nesse processo.

Conforme mencionado previamente, foi utilizado um conjunto de estruturas de proteínas do banco de dados ASTRAL SCOP, *release 1.63*, para a construção do

modelo de classificação de proteínas.

Convém ressaltar que, as estruturas de proteínas presentes no ASTRAL SCOP são um subconjunto daquelas presentes no banco de dados Sting_DB. Como mencionado, na etapa d de Material e Métodos, o número de parâmetros para representar cada proteína foi 241, onde 240 representam o produto das 11 variáveis disponíveis na Tabela 2, multiplicadas pelos 20 coeficientes de expansão totalizando 220 parâmetros. Além disso, adicionou-se a freqüência dos 20 aminoácidos presentes nas proteínas. O último parâmetro é a classe de enzima da proteína.

Para avaliar a metodologia de seleção de parâmetros em classificação de proteínas, foram utilizados os algoritmos do *software* Weka, versão 3.4.4 (Witten & Frank, 2005). Weka é um ambiente de *software* usado em problemas de descoberta do conhecimento, composto de uma coleção de algoritmos nas áreas de aprendizado de máquina e mineração de dados. Weka é um *software* livre que está disponível sob licença General Public License - GNU.

Além do *software* Weka, o Matlab foi utilizado e alguns programas foram desenvolvidos em linguagens C + + e Java para o tratamento dos dados.

A metodologia proposta para seleção de parâmetros foi avaliada utilizando-se os métodos de classificação descritos em Material e Métodos. Para cada método, um classificador foi escolhido, conforme segue:

- Árvore de Decisão (Random Forest)
- Modelo Bayesiano (Naïve Bayes)
- Redes Neurais (Multilayer Perceptron)
- Support Vector Machine (algoritmo de otimização proposto por John Platt)

Para cada método de classificação investigado nesse trabalho, foram consideradas duas abordagens: 1) *split-sample technique:* a técnica tradicional que divide os dados em dois grupos - dois terço para o conjunto de treinamento e um terço para o conjunto de teste; 2) *k-fold cross-validation:* nesse caso, os dados são divididos em k subconjuntos de tamanhos aproximadamente iguais. O modelo é então treinado k vezes, onde cada vez um dos subconjuntos é usado para teste e, os demais, para treinamento. Em Goutte (1997), pode-se observar que, para conjunto de dados pequenos, a técnica de validação cruzada (k-fold cross-validation) apresenta resultados melhores do que aqueles obtidos por meio da técnica de segmentação de amostras (*split-sample technique*).

A Fig. 2 mostra uma comparação dos resultados obtidos para os quatro métodos de classificação analisados, usando os 241 parâmetros. Em outras palavras, não foi aplicado nenhum filtro para redução de variáveis. O método baseado em árvore de decisão obteve os melhores resultados, próximos de 70% de precisão, seguido do método baseado em redes neurais, com precisão de

quase 65%. Pode-se notar ainda que, mesmo o método Bayesiano obteve precisão acima de 40%, o que supera os resultados publicados em Dobson & Doig (2005). Isso demonstra claramente que, usando a metodologia para seleção de parâmetros, é possível se obter um ganho significativo na precisão do modelo de classificação.

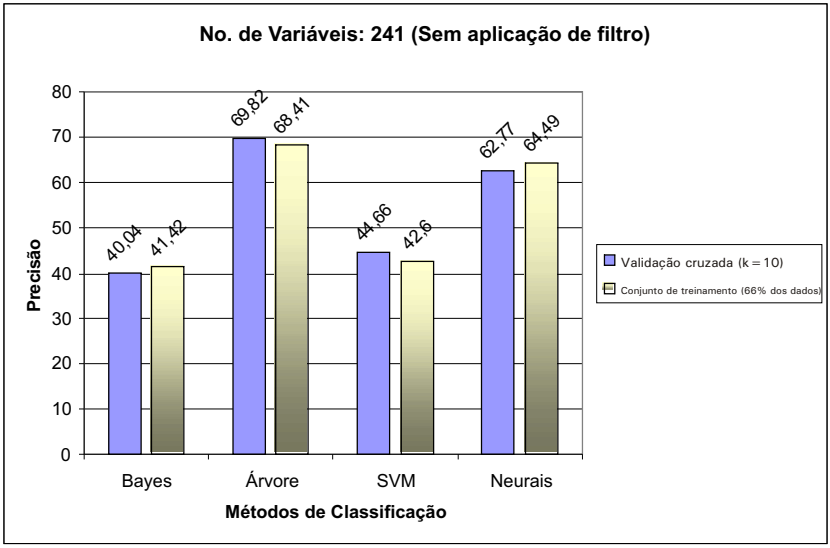



**Fig. 2.** Comparação dos métodos de classificação sem aplicação de filtro para redução de variáveis - 241 variáveis.

Depois da avaliação da metodologia, sem a aplicação de filtros para redução de variáveis, foi utilizada uma outra abordagem para redução de variáveis usando a técnica de análise de componentes principais PCA (do inglês Principal Component Analysis). Em síntese, PCA é um método que tem por finalidade básica, a redução de dados a partir da eliminação das combinações lineares das variáveis originais (Aitchison, 1986).

Na Fig. 3, pode-se observar a comparação dos resultados obtidos para os quatro métodos de classificação analisados, usando-se PCA. Nota-se que o número de parâmetros foi reduzido para 80. Nesse caso, o método baseado em árvore de decisão obteve os melhores resultados novamente, acima de 69% de precisão, seguido do método baseado em redes neurais, com precisão acima de 63%, enquanto *support vector machine* alcançou precisão de quase 60%. Pode-se notar ainda que, mesmo o método Bayesiano obteve precisão de quase 50%, o que é um resultado superior aquele obtido em Dobson & Doig (2005).

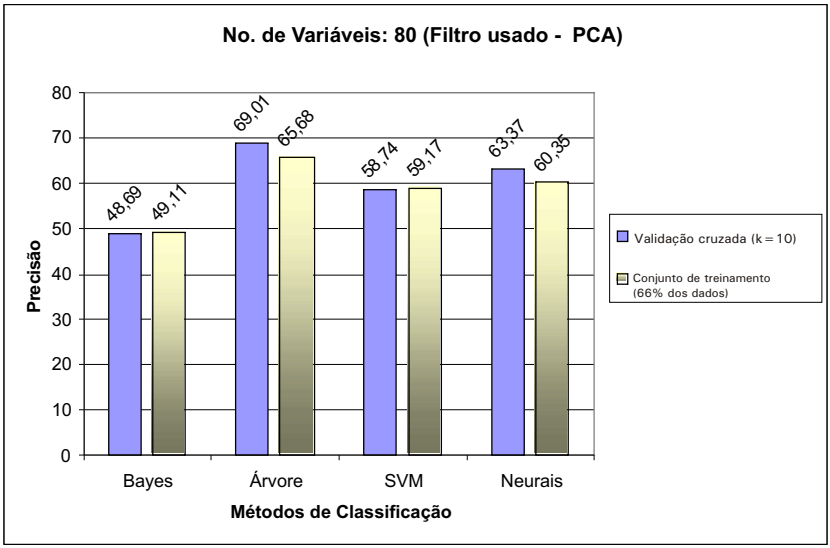



**Fig. 3.** Comparação dos métodos de classificação com aplicação de filtro para redução de variáveis (PCA) - 80 variáveis.

Convém notar que os resultados alcançados com e sem redução de filtros são próximos (Fig. 2 e 3). A razão principal é que as três primeiras fases da metodologia proposta tem o objetivo de eliminar a dependência de variáveis, o que também é obtido com o uso da técnica de PCA. Em particular, o modelo Bayesiano obteve um melhor resultado com o uso de PCA, pois a classificação através de modelos Bayesianos é favorecida quando as variáveis em análise são independentes.

Em muitos projetos de mineração de dados, a fase de pré-processamento pode consumir mais de 60% do tempo de execução. Contudo, as tarefas que demandam mais tempo nem sempre são limpeza, resolução de inconsistências e integração dos dados. No caso de classificação de proteínas, a lição aprendida foi que o tratamento dos dados antes da geração do modelo de classificação é a fase que mais demanda tempo.

Outras lições aprendidas nesse projeto de mineração de dados foram as seguintes:

- a seleção de parâmetros é a decisão crucial: quando não existe um tratamento adequado dos dados para a seleção de parâmetros, a precisão do modelo de classificação pode ser comprometida. Observou-se que nem sempre o maior número de parâmetros é a melhor opção para a geração do modelo. O ruído introduzido em alguns parâmetros, que são combinação linear de outros, precisa ser removido;

- a transformação é uma etapa com obstáculos: a mineração de dados é um projeto contínuo de busca de inteligência e inferência aplicada aos dados. Portanto, em muitos casos, a transformação de dados pode exigir: a) requisitos de análises complexas; b) verificação de tendências escondidas; c) verificação de hipóteses. Dificilmente alguém vai resolver tudo na primeira tentativa.

Finalmente, acredita-se que a regra 80:20, conhecida como Lei de Pareto (Pareto, 1897), pode ser uma alternativa para a construção do modelo de classificação de proteínas. Em outras palavras, 80% da precisão do modelo de classificação pode ser alcançada por meio de 20% dos parâmetros disponíveis (os parâmetros mais representativos). Esses parâmetros devem discriminar as estruturas de proteínas.

## Conclusões

Nesse trabalho foram abordados os três principais desafios relacionados ao problema de classificação de enzimas em banco de dados de estruturas de proteínas: 1) o ruído presente nos parâmetros; 2) o grande número de variáveis; e 3) o número não-balanceado de membros por classe.

Para abordar esses desafios, foi apresentada uma metodologia para seleção de parâmetros, que combina recursos da matemática (Transformada Discreta do Cosseno) e da estatística (correlação de variáveis e amostragem com reposição).

A metodologia foi validada utilizando-se o banco de dados ASTRAL SCOP, *release* 1.63, considerando-se quatro diferentes métodos de classificação propostos na literatura, a saber: árvores de decisão, classificador bayesiano simples, redes neurais e *support vector machine*. A metodologia também foi comparada com uma similar apresentada em Dobson & Doig (2005), que obteve uma precisão de até 35%, usando o classificador *support vector machine*.

Nos experimentos desse trabalho de pesquisa, foi demonstrado que a metodologia proposta obteve uma precisão de aproximadamente 70% para o método de árvore de decisão e, próximo de 65% para o método baseado em redes neurais. Isso demonstra claramente que, usando essa metodologia para seleção de parâmetros, é possível se obter um ganho significativo na precisão do modelo de classificação.

# Referências Bibliográficas

AHMED, A.; NATARAJAN, T.; RAO, K. R. Discrete Cosine Transform. **IEEE Trans. on Computers**, C-23, p. 90-93, 1974.

AITCHISON, J. **The statistical analysis of compositional data**. New York: Chapman & Hall, 1986. 416 p.

BERMAN, H. M.; WESTBROOK, J.; FENG, Z.; GILLILAND, G.; BHAT, T. N.; WEISSIG, H.; SHINDYALOV, I. N.; BOURNE, P. E. The Protein Data Bank. **Nucl. Acids Res**., v. 28, p. 235-242, 2000.

BLUM, A.; Langley, P. Selection of relevant features and examples in machine learning. **Artificial Intelligence**, v. 97, p. 245-271, 1997.

BREIMAN, L. Bagging predictors. **Machine Learning**, v. 26, p. 123-140, 1996.

BRENNER, S. E.; KOEHL, P.; LEVITT, M. The ASTRAL compendium for sequence and structural analysis. **Nucl. Acids Res**. v. 28, p. 254-256, 2000.

CHANDONIA, J. M.; WALKER, N. S.; CONTE, L. L.; KOEHL P.; LEVITT, M.; BRENNER, S. E. ASTRAL compendium enhancements. **Nucl. Acids Res**., v. 30, p. 260-263, 2002.

DOBSON, P. D.; DOIG, A. J. Predicting enzyme class from protein structure without alignment. **J. Mol. Biol**., v. 345, p. 187-199, 2005.

FELLER, W. **An introduction to probability theory and its applications**. 2nd ed. New York: John Wiley, 1971. v. 2.

GOUTTE, C. Note on free lunches and cross-validation. **Neural Computation**, v. 9, p. 1211-1215, 1997.

LANGLEY, P. Selection of relevant features in machine learning. In: AAAI FALL SYMPOSIUM ON RELEVANCE, 1994, New Orleans. **Proceedings**... New Orleans: AAAI Press, 1994.

HAN, J.; KAMBER, M. **Data mining**: concepts and techniques. San Francisco, Morgan Kaufmann, 2001.

NESHICH, G.; TOGAWA, R. C.; MANCINI, A.; KUSER, P. R.; YAMAGISHI, M. E. B.; PAPPAS, G.; TORRES, W. V.; CAMPOS, T. F. E.; BAUDET, C.; HIGA, R. H. STING Millennium: a web-based suite of programs for comprehensive and simultaneous analysis of protein structure and sequence. **Nucl. Acids Res**., v. 31, n. 13, p. 3386-3392, 2003.

NESHICH, G.; HIGA, R. H.; YAMAGISHI, M. E. B.; MANCINI, A.; TOGAWA, R. C. SMS: Integrated software for extensive analyses of 3D structures of proteins and their complexes. **BMC Bioinformatics**, v. 5, n. 1, p. 107, 2004.

NESHICH, G.; BORRO, L. C.; HIGA, R. H.; KUSER, P. R.; YAMAGISHI, M. E. B.; FRANCO, E. H.; KRAUCHENKO, J.; FILETO, R.; RIBEIRO, A. A.; BEZERRA G. B. P.; VELLUDO, T. M.; JIMENEZ, T. S.; FURUKAWA, N.; TESHIMA, H.; KITAJIMA, K.; BAVA, A.; SARAI, A.; TOGAWA, R. C.; MANCINI, A. The Diamond STING server. **Nucl. Acids Res**., v. 33 (Web Server issue), p. W29-W35, July 2005.

PARETO, V. **Cours d'economie politique**. Paris: F. Pichou, 1897. v. 2.

RADIVOJAC, P.; CHAWLA, N. V.; DUNKER, A. K.; OBRADOVIC, Z. Classification and knowledge discovery in protein databases. **Biomedical Informatics**, v. 37, p. 224.239, 2004.

*Informática Agropecuária*
*Ministério da Agricultura, Pecuária e Abastecimento*