MÉTODO PARA O EMPREGO DE ALGORITMOS DE CLASSIFICAÇÃO NO APOIO ÀS DECISÕES ESTRATÉGICAS MILITARES

Mateus Felipe Tymburibá Ferreira[1], Éldman de Oliveira Nunes[2]

Resumo. Este trabalho tem o intuito de apresentar um método para o emprego de algoritmos de classificação no auxílio ao planejamento estratégico nas Forças Armadas. Através da previsão de padrões futuros, com o emprego da técnica de Mineração de Dados, pretende-se apoiar aos processos de tomada de decisão no âmbito militar, possibilitando um aprimoramento dos sistemas gerenciais rumo à almejada excelência. A Mineração de Dados, no contexto maior da Descoberta do Conhecimento, tem sido amplamente empregada em áreas como vendas, finanças, segurança e biomedicina, entre outras. Acredita-se que o método exposto neste artigo possibilite uma melhor aproximação entre a área de Sistemas de Apoio à Decisão e a esfera militar brasileira. É apresentado também um estudo de caso detalhado, como forma de demonstrar a aplicação do método e comprovar os benefícios decorrentes da sua utilização. O método proposto foi empregado nesse estudo de caso para a assistência ao planejamento pedagógico das instruções ministradas no Curso de Formação de Oficiais do Quadro Complementar de Oficiais do Exército Brasileiro. O resultado obtido confirma a capacidade de potencializar previsões, através da aplicação do método sugerido e aponta para uma oportunidade de melhoria dos sistemas de planejamento estratégico militares.

Palavras-chave: Planejamento Estratégico. Decisões Estratégicas. Descoberta do Conhecimento. Mineração de Dados. Algoritmos de Classificação. Sistemas de Apoio à Decisão.

Abstract. This paper aims at presenting a method for aplying classification algorithms to assist the Strategic Planing tasks in the Armed Forces. By the preview of future patherns, and through the Data Mining technic we intend to support the military Decision Making processes, making possible the improvement of the management systems to achieve the desired excellency. The Data Mining, as part of the Knowledge Discovery context, has been widely used in areas like sale, finances, security and biomedicine, among others. The method displayed in this article can make possible a better approach between the Decision Support Systems area and the brasilian military scope. It is also showed a detailed case study to demonstrate the method application and prove the benefits decurred of its use. The proposed

[1] Bacharelado em Ciência da Computação. Escola de Administração do Exército (EsAEx), Salvador, Brasil. mateustymbu@yahoo.com.br .
[2] Doutorado em Conputação. Escola de Administração do Exército (EsAEx), Salvador, Brasil. eldman@bol.com.br .

method was used in this case study to assist the pedagogic planning of the instructions given at the officers formation course of the complementary officers staff of the brazilian army. The reached result confirms the capability to get better previews by the application of the sugestted method and indicates an improvement opportunity for the military strategic planning systems.

Keywords: Strategic Planning. Strategic Decisions. Knowledge Discovery. Data Mining. Classification Algorithms. Decision Support Systems.

1 Introdução

Planejamento Estratégico pode ser conceituado como um processo gerencial que se volta para o alcance de resultados, através da antecipação sistemática de mudanças futuras, tirando vantagens das oportunidades que surgem, examinando os pontos fortes e fracos da organização, estabelecendo e corrigindo cursos de ação, propõe Oliveira (1991).

Apesar de a importância do processo decisório para a eficiência das organizações ter sido constatada e discutida ao longo das últimas décadas,

> as decisões envolvendo aspectos estratégicos de organizações estão falhando, em geral, por falta de instrumentos racionais e objetivos que permitam auxiliar os executivos na tomada de suas decisões. Paradoxalmente, enquanto lidam com todos os recursos computacionais de processamento e de gerenciamento, não absorveram ainda o uso de ferramentas de apoio à decisão (MURAKAMI, 2003).

De forma equivalente, os recursos de tratamento de dados para a extração de informações valiosas, oferecidos pelos sistemas computacionais, podem ser mais extensivamente explorados pelas forças armadas. O presente trabalho objetiva justamente desenvolver um método para o auxílio às tomadas de decisões estratégicas militares, através do emprego de técnicas de descoberta do conhecimento e mineração de dados.

Entre as decisões estratégicas rotineiramente tomadas por militares em função de comando nas Forças Armadas, e que podem ser auxiliadas por sistemas de apoio à decisão estão: a escolha de militares para comando de OM, para missões no exterior e missões de paz e para comissões especiais; a nomeação de instrutores em estabelecimentos de ensino e formação militar; a definição dos oficiais-generais a serem promovidos; a priorização orçamentária das diversas Organizações Militares e a previsão do resultado final dos alunos dos cursos de for-

mação, para aperfeiçoamento do processo de ensino e aprendizagem. Na realidade, o processo de descoberta do conhecimento se aplica a todas as situações em que existe uma missão a ser atribuída a um militar e, baseado em um perfil dos candidatos à missão, deve-se escolher o indivíduo mais indicado. Dessa forma, as possibilidades de utilização da metodologia desenvolvida neste trabalho não se esgotam nos exemplos acima, podendo surgir inúmeros novos casos para a aplicação dessa metodologia.

2 Referencial Teórico

2.1 Descoberta do Conhecimento e Mineração de Dados

À medida que passam os anos, as organizações acumulam em suas bases de dados uma extraordinária quantidade de informações, que poderiam tornar essas instituições mais competitivas, permitindo a elas detectar tendências e características escondidas, e que tenham uma reação mais rápida a eventos futuros. Contudo, conforme afirma Schneider (2002), são raros os estabelecimentos que exploram tal oportunidade, porque essas valiosas informações estão disfarçadas, implícitas nesses grandes conjuntos de dados, e não podem ser descobertas utilizando-se métodos tradicionais de gerenciamento de banco de dados.

Tan, Steinbach e Kumar (2006, p.769) definem como Descoberta do Conhecimento(*Knowledge Discovery in Databases* – KDD, sigla em inglês pelo qual é mais conhecido) justamente o processo global utilizado para converter dados tabulares em informações úteis. O processo consiste em uma série de etapas de transformação, desde o pré-processamento dos dados até o pós-processamento dos dados advindos da etapa de Mineração de Dados, considerada a etapa central e mais importante de todo o processo, representado na Figura 1.

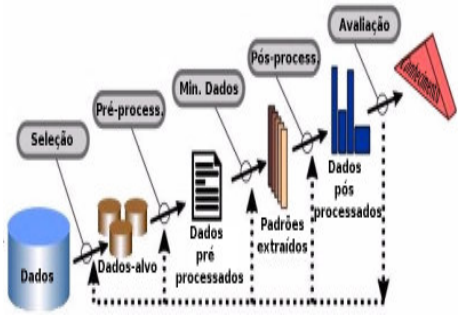

Figura 1 - Processo de Descoberta do Conhecimento (KDD)
Fonte: Moraes, 2002

O propósito da fase de pré-processamento é transformar os dados tabulares de entrada em um formato apropriado para a análise subseqüente. De forma análoga e contrária, a fase

de pós-processamento busca reformatar os dados em uma disposição adequada para a visualização, garantindo que apenas os resultados válidos e úteis sejam incorporados ao sistema de suporte a decisões.

A descoberta de conhecimento tem sido aplicada em uma ampla variedade de áreas e situações, dentre as quais destacam-se, segundo Moraes (2002), vendas, marketing, finanças, segurança, medicina, biologia molecular, astronomia, prospecção mineral, policial, formação de equipes e telecomunicações.

O termo Mineração de Dados refere-se à extração ou mineração de conhecimento a partir de grandes quantidades de dados. Apesar da Mineração de Dados ser considerada a fase essencial do processo de KDD, por ser a responsável pela descoberta dos padrões ocultos, deve-se lembrar que ela representa apenas uma etapa no processo completo de Descoberta do Conhecimento.

2.2 Algoritmos de Classificação

Os algoritmos de classificação são utilizados para prever o valor de um determinado atributo, com base nos dados de outras variáveis. A seguinte definição formal pode ser encontrada: "Classificação é a tarefa de aprender uma função-alvo ***f*** que mapeie cada conjunto de atributos ***x*** (objeto) em uma das classes ***y*** pré-definidas" (Tan, Steinbach e Kumar, 2006). Após a descoberta dessa função-alvo, o sistema é capaz de prever a classe de um registro desconhecido, ou seja, o algoritmo de classificação pode ser encarado como uma caixa preta que automaticamente assinala uma categoria quando recebe um conjunto de atributos de um registro desconhecido.

A abordagem geral para resolver problemas de classificação, representada na Figura 2, consiste em utilizar um conjunto de treinamento para obter-se a função-alvo. Esse conjunto de treinamento é formado por registros cujas classes já são conhecidas. Ele é usado para construir o modelo de classificação, que depois é aplicado ao conjunto de teste, que consiste em registros com classes desconhecidas.

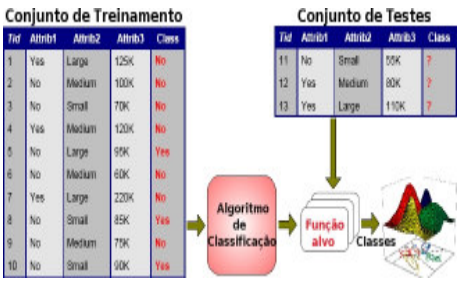

Figura 2 - Abordagem geral para a solução de problemas de classificação.
Fonte: Tan, Steinbach, Kumar (2006)

3 Modelo proposto

A fim de facilitar a compreensão do método aqui proposto e possibilitar o seu acompanhamento para emprego nas diversas áreas militares onde as tomadas de decisão tornam-se imprescindíveis, dividiu-se a seqüência de passos a ser seguida em cinco etapas: modelagem do problema, escolha do algoritmo, escolha da ferramenta, mineração das bases de dados e visualização dos resultados.

3.1 Modelagem do Problema

Uma das fases mais importantes no processo de Descoberta do Conhecimento é a etapa de pré-processamento de dados, pois é ela que permite a escolha e formatação das bases de dados que representam o mundo real para a sua utilização na fase de Mineração de Dados.

Na fase de pré-processamento dos dados, deve-se despender elevada atenção ao processo de seleção das variáveis, uma vez que um conjunto de atributos bem selecionado pode conduzir a modelos de conhecimento mais concisos e com maior precisão. Os atributos, também conhecidos na literatura pelos termos *características, variáveis, campos e dimensões*, são aspectos e propriedades de um objeto, que o identificam e distinguem dos demais, utilizados para representar itens do mundo real para fins de avaliação e tomada de decisão. A escolha de variáveis é feita com base no maior ganho de informação, isto é, na melhoria de qualidade de classificação que o atributo oferece.

Segundo Soares (2007), existem duas abordagens consagradas para a seleção de atributos: método independente de modelo - também conhecido por *filter* - e método dependente de modelo (também conhecido por *wrapper*).

> No primeiro método (*filter*), os atributos são selecionados por algum critério, e utilizados na etapa de mineração de dados, sem levar em consideração o algoritmo de classificação que será aplicado aos atributos selecionados. Já no segundo (*wraper*), um subconjunto de atributos é analisado por um algoritmo de classificação, que avalia o desempenho deste algoritmo com a seleção feita. Se um critério de desempenho mínimo não for atendido, um novo subconjunto de atributos é gerado, e nova avaliação é feita. Este processo iterativo se encerra quando o critério é atendido, e o último subconjunto de atributos gerados é a saída do método (SOARES, 2007).

Apesar de gerar um subconjunto de variáveis capaz de aumentar a precisão do processo de mineração de

dados, a técnica dependente de modelo é muito mais lenta, já que a busca pelo melhor subconjunto de atributos é um processo que demanda muito tempo de processamento. Além disso, a melhor configuração de características para um dado algoritmo de classificação pode não ser tão boa para um outro classificador, o que faz com que a seleção de campos realizada através dessa abordagem seja dependente do algoritmo utilizado. Como a metodologia aqui proposta prima pela flexibilidade e possibilidade de ampla aplicação no contexto militar, sugere-se a utilização da abordagem independente de modelo (*filter*), como forma de diversificar as possibilidades de emprego deste método.

Ressalte-se que o processo de escolha dos atributos representativos, durante a modelagem de um problema real, deve ser conduzido por pessoas bastante familiarizadas com o modelo de negócio a ser trabalhado. Esta é a forma de garantir uma boa proximidade entre a abstração fornecida pelo modelo e a realidade.

3.2 Escolha do Algoritmo

A escolha do algoritmo de classificação a ser empregado na etapa de mineração de dados deve ser pautada nas características específicas da base de dados a ser utilizada. Aspectos como tamanho dos conjuntos de treinamento e de teste, dependência entre variáveis, atributos intrinsecamente ordenados, e tipos de campos (discretos ou contínuos, categóricos ou numéricos) podem ser determinantes no desempenho dos diversos classificadores. Assim, torna-se crucial a identificação dessas características como forma de corroborar a escolha do algoritmo de classificação mais adequado a cada situação. Tan, Steinbach e Kumar (2006) dedicam uma subseção em cada uma das seções nas quais apresentam diversos algoritmos de classificação, onde discorrem detalhadamente sobre a aplicabilidade de cada um desses algoritmos, de acordo com as especificidades das bases de dados.

Além da adequação do algoritmo aos tipos de dados existentes nos conjuntos de treinamento e de teste, freqüentemente utilizam-se métricas de desempenho, como forma de medir e comparar a performance de algoritmos diferentes. Essa avaliação é baseada na contagem do número de registros de teste assinalados corretamente ou incorretamente às classes. A medida de acurácia, por exemplo, é definida por Tan, Steinbach e Kumar (2006) como dado pela equação (3.1).

$$Acurácia = \frac{Número\ de\ previsões\ corretas}{Número\ total\ de\ previsões} \quad (3.1)$$

Os algoritmos de classificação buscam funções-alvo que atinjam a máxima acurácia, quando aplicados ao conjunto de teste.

Como comentam Tan, Steinbach e Kumar (2006), a maioria dos métodos comumente utilizados para a avaliação do desempenho de algoritmos de classificação utilizam as medidas de acurácia como base para as comparações. Dentro desse universo, podem ser citados os métodos *Holdout*, *Random Subsampling* e *Cross-Validation*. Todos esses seguem o princípio básico de particionar o conjunto de treinamento, utilizando parte dos dados para o efetivo treinamento e construção da função-alvo, e o restante dos registros para a verificação da qualidade das respostas geradas pelo classificador. O método *Cross-Validation* foi escolhido por incorporar melhorias em relação aos métodos *Holdout e Random Subsampling,* mantendo contudo a mesma estratégia.

Conforme demonstram Tan, Steinbach e Kumar (2006), a proposta do método *Cross-Validation* é subdividir o conjunto de treinamento em alguns subconjuntos de mesmo tamanho e aplicar os diversos algoritmos de classificação nesses subconjuntos, utilizando um dos conjuntos como base de teste e os demais como base de treinamento. O processo deve ser repetido de forma que a cada iteração apenas um dos subconjuntos seja utilizado para os testes e todos os demais componham o conjunto de treinamento. A acurácia de um algoritmo é computada através da soma da acurácia obtida pelo algoritmo em cada uma das iterações. O classificador que apresentar o maior valor de acurácia total no final do processo, será o algoritmo escolhido.

Recomenda-se a utilização do caso especial em que o número de subconjuntos é igual ao número de registros existentes no conjunto de treinamento *(N).* Nessa abordagem, conhecida como *leave-one-out* (em língua portuguesa, deixe um fora), cada conjunto de teste contém apenas um registro. Essa estratégia tem a vantagem de explorar ao máximo os dados históricos no treinamento do classificador. Além disso, os conjuntos de teste são mutuamente exclusivos e cobrem efetivamente todo o conjunto de dados. Essas vantagens normalmente compensam o fato de que pode se tornar computacionalmente caro repetir o processo N vezes, quando o conjunto de dados for muito grande.

3.3 Escolha da Ferramenta

Para efetuar a computação das bases de dados e extrair os padrões desejados, deve-se optar por codificar os diversos algoritmos de classificação descritos na bibliografia existente ou utilizar uma das ferramentas de mineração de dados disponíveis. A opção de codificar algoritmos normalmente é adotada por pesquisadores que pretendem acrescentar técnicas aos algoritmos na tentativa de aprimorá-los ou adaptá-los para trabalhos específicos. Isso porque o processo de codificação dos algoritmos pode ser bastante complexo e dispendioso, demandando muito tempo e recursos.

A utilização de aplicações existentes, por sua vez, garante confiabilidade e eficiência ao usuário, uma vez que são amplamente testadas e aprimoradas ao longo do tempo, adquirindo maior robustez, desempenho e acurácia. Além disso, as ferramentas de mineração de dados proporcionam extrema flexibilidade, pois agregam as principais técnicas e possibilidades em um único ambiente. Como duas organizações ou dois conjuntos de dados nunca são iguais, não existe uma única técnica que atinja os melhores resultados para todas as situações. Ao invés de codificar um novo algoritmo a cada caso a ser trabalhado, uma instituição pode se aproveitar dessa ampla flexibilidade oferecida pelos sistemas de mineração de dados para aplicar a mesma ferramenta em um novo cenário. Por tudo isso, recomenda-se, no contexto das forças armadas, a adoção de uma ferramenta de mineração de dados.

Como ressalta Britos et al. (2006, p. 85-90), o número de aplicações de mineração de dados atualmente disponível é enorme, distribuindo-se entre ferramentas pagas ou de código livre. Torna-se necessário, portanto, um processo cuidadoso de escolha do sistema a ser utilizado por uma organização. Seguindo a orientação do governo federal para que os órgãos e entidades públicas adotem preferencialmente *softwares* de código livre, buscou-se estudar ferramentas de mineração de dados que atendessem a essa exigência. Dentre as diversas opções encontradas, destacam-se as seguintes: YALE, WEKA, Ratle e Orange (KDNUGGETS, 2007).

Após uma detida análise de diversas características de cada uma das ferramentas mencionadas, concluiu-se que o sistema WEKA (*Waikato Environment for Knowledge Analysis*) desponta como a aplicação mais adequada para o foco deste trabalho. Foram avaliados aspectos como facilidade de uso, acurácia, disponibilidade

das tarefas e dos algoritmos mais utilizados, adaptabilidade a tipos de dados, integração de ferramentas de visualização, documentação e suporte aos usuários.

O WEKA é uma coleção de algoritmos de mineração de dados que podem ser aplicados diretamente a bancos de dados ou podem ser importados como bibliotecas Java por outros sistemas. Ele contém, além dos algoritmos, diversas opções de ferramentas para pré-processamento e visualização dos resultados com relatórios estatísticos completos. Pode ser utilizado através de diversos tipos de interface gráfica e em quase todas as plataformas operacionais (U. WAIKATO, 2007). Possui ainda uma extensa documentação e uma comunidade de desenvolvimento bastante ativa, que contribui de forma expressiva para o crescimento das suas funcionalidades, tornando-o um dos sistemas de mineração de dados com o maior número de algoritmos implementados (SILVA, 2004). Por tudo isso, optou-se pelo emprego do WEKA no desenvolvimento do presente trabalho.

3.4 Mineração das Bases de Dados

Nesta etapa devem ser empregados os recursos computacionais para a construção da função-alvo e a predição das classes dos registros. Os dados históricos coletados e reunidos no conjunto de treinamento devem ser formatados em um arquivo compatível com a ferramenta a ser utilizada. Os formatos de arquivo de entrada reconhecidos pelo WEKA serão descritos na seção de avaliação dos resultados. Neste ponto, deve-se atentar para o fato de que os registros já deverão estar particionados nos subconjuntos descritos na seção 3.2, para que o processo de escolha do classificador possa ser efetuado. Feita a escolha do algoritmo mais adequado, parte-se para a mineração dos registros cujas classes pretende-se prever.

3.5 Visualização dos resultados

Tan, Steinbach e Kumar (2006) conceituam visualização como a exibição de informações em um formato gráfico ou tabular, e destacam que o objetivo dessa etapa é facilitar a interpretação das informações descobertas no processo de mineração de dados. Essa fase possibilita a formação de um modelo mental dos resultados. Para isso, os dados finais devem ser convertidos em um formato visual de forma que as características dos dados e os relacionamentos entre os atributos possam ser reconhecidos.

As técnicas de visualização são

normalmente específicas para o tipo de dado a ser analisado. Assim, novas técnicas de visualização e abordagens, bem como variações de abordagens existentes, estão continuamente sendo criadas em resposta aos novos tipos de dados e tarefas de visualização. As principais técnicas atualmente empregadas são: *Stem and leaf plot,* Histogramas, *Box plots, Pie charts, Empirical cumulative distribution functions* e *Scatter plots*, para a visualização de objetos com um número pequeno de atributos; *Countor plots, Surface plots, Vector Fields Plots, Lower-dimensional slices* e Animações, para a visualização de dados espaço-temporais; *Parallel coordinates, Star coordinates* e *Chernoff faces,* para a visualização de dados com várias dimensões, citam Tan, Steinbach e Kumar (2006). A maioria dessas técnicas pode ser encontrada nas diversas opções de interfaces gráficas disponíveis para o WEKA.

4 Avaliação dos resultados

Com o objetivo de demonstrar o emprego do modelo proposto e possibilitar a elucidação de dúvidas no aspecto prático da sua aplicação em problemas reais, desenvolveu-se um estudo de caso voltado para o apoio ao processo pedagógico em estabelecimentos de formação militar.

4.1 Auxílio Pedagógico ao CFO/ QCO

O ensino no Exército Brasileiro sempre se destacou pela qualidade, organização e seriedade com que é tratado pela Força e, sobretudo, pela sua característica de estar em constante processo de atualização. Em consonância com essa proposta, o presente trabalho volta-se para os estabelecimentos de ensino existentes na Força.

De acordo com idade e nível de escolaridade, existem várias opções para homens e mulheres ingressarem no Exército. Para o militar de carreira (oficial ou sargento), o ingresso só é possível mediante a aprovação em concurso público, de âmbito nacional, para uma das Escolas de Formação (BRASIL, 2007). Esse é justamente o escopo das escolas sobre as quais se pretende trabalhar, pelo fato de que os dados extraídos dos resultados desses concursos são de suma importância no processo que será descrito neste artigo. Para efeito de testes, foram utilizados dados colhidos na Escola de Administração do Exército.

A idéia central neste estudo de caso é aplicar o método aqui proposto para prever a classificação final de alunos

do Curso de Formação de Oficiais (CFO) do Quadro Complementar de Oficiais (QCO) do Exército Brasileiro. De posse dessa previsão, espera-se possibilitar uma adaptação das técnicas e processos de ensino, como forma de aprimorar os métodos de instrução empregados na qualificação dos oficiais do QCO. A previsão do resultado final do curso permite a identificação dos grupos de alunos com maior dificuldade de aprendizado e possibilita a formulação de propostas pedagógicas que atenuem essa dificuldade, elevando assim o padrão dos recursos humanos nas Forças Armadas.

4.2 Seleção dos Atributos do Modelo

Os atributos escolhidos para a representação dos alunos foram baseados nas características pessoais de cada instruendo bem como no desempenho desses militares no concurso de admissão. Procurou-se utilizar variáveis que aparentassem influenciar o desempenho dos alunos durante o curso, em especial aquelas que agem diretamente sob a motivação de cada indivíduo. A Tabela 1 representa os atributos empregados na modelagem.

Tabela 1 - Atributos da modelagem e seus respectivos domínios

	Atributos	Domínio
PESSOAIS	Sexo	M (Masculino), F (Feminino)
	Origem Militar	E (Exército), M (Marinha), A (Aeronáutica), C (Civil)
	Estado Civil	C (Casado), S (Solteiro)
	Idade	18-37
	Estado Federativo de Origem	AC, AL, AP, AM, BA, CE, DF, GO, ES, MA, MT, MS, MG, PA, PB, PR, PE, PI, RJ, RN, RS, RO, RR, SP, SC, SE, TO
CONCURSO	Área de Especialidade	ADM, CON, DIR, ECO, ENF, EST, INF, ESP, FIS, GEO, HIS, ING, MAT, POR, QUI, PED, PSI, CSO, VET
	Classificação na Área	1, 2, 3, ...
SAÍDA	Classificação final no Curso	A(1-10), B(11-20), C(21-30), D(31-40), E(41-50), ...

Fonte: do autor

4.3 Testes

A massa de dados utilizada na realização dos testes foi composta pelos registros dos alunos que concluíram o curso de formação de oficiais da Escola de Administração do Exército, nos anos de 2005 e 2006. Para a efetivação dos testes, esses registros foram particionados, resultando no conjunto de treinamento constituído por

89 registros de 2005; e no conjunto de teste, formado por 59 registros de 2006. A utilização de registros cujas classes finais fossem previamente conhecidas foi vital para a medição da qualidade do resultado gerado.

Partiu-se para a escolha do algoritmo de classificação a ser empregado. Foi utilizada a versão 3.4.11 do WEKA.Constatou-se que os algoritmos disponíveis na ferramenta mais adequados eram: *AdaBoostM1*, *MultiBoostAB* e *DecisionStump*. Foram priorizados os classificadores que apresentaram o melhor resultado no teste de aplicação do método *Cross-Validation*, com N = 89, conforme diretriz do modelo proposto. Além disso, observou-se a obrigatoriedade de emprego de algoritmos que trabalham com classes nominais, uma vez que a modelagem do problema impôs essa restrição. Finalmente, a escolha do algoritmo baseou-se no estudo de Caruana e Niculescu-Mizil (2006), onde os autores comparam diversas classes de classificadores e concluem que, em geral, os algoritmos baseados em Árvores-de-Decisão *(Decision Stump)* aprimorados *(Boosted)* apresentam os melhores índices de acurácia.

Dentre os algoritmos citados, optou-se por utilizar o *MultiBoostAB*, tendo em vista que ele reúne diversas técnicas para melhoria de desempenho de algoritmos de classificação, incluindo uma extensão do método *AdaBoost*, empregado no *AdaBoostM1* (WEBB, 2000). Além disso, ele permite a aplicação do método *DecisionStump* como algoritmo base para o seu funcionamento, sendo esse um de seus parâmetros de configuração no WEKA.

Os registros pertencentes aos conjuntos de treinamento e de teste foram consolidados em arquivos de texto com o formato *ARFF*, para que pudessem ser encaminhados como arquivos de entrada de dados no WEKA. A descrição detalhada desse formato pode ser encontrada no site da ferramenta (UNIVERSITY OFWAIKATO).

Para a execução das tarefas de mineração de dados no WEKA, é possível optar pela utilização de uma das suas interfaces gráficas (*Experimenter*, *Explorer* e *Knowledge Flow*) ou pela execução direta dos algoritmos através da linha de comandos (CLI – *Command Line Interface*). Optou-se pela utilização da interface denominada *Explorer* por motivos de simplicidade e ampla disponibilidade de material de consulta.

4.4 Resultados e Discussão

O algoritmo escolhido, aplicado ao conjunto de teste mediante treinamento com o conjunto de treinamento, obteve uma acurácia de 28,814%. O resultado, apesar de apresentar um valor distante de um padrão desejável para o seu pronto emprego em situações reais, demonstra a capacidade dos algoritmos de classificação em potencializar as chances de acerto em uma previsão, contribuindo para os processos de tomadas de decisão. Uma previsão aleatória, no estudo de caso em questão, teria chances de acerto de 1/9, ou seja, 11,111%.

Dentre os fatores que podem contribuir para a variação no desempenho dos algoritmos de classificação no processo de previsão de resultados, deve-se destacar a qualidade da modelagem e do conjunto de treinamento. A qualidade da modelagem pode ser influenciada pela representatividade dos atributos escolhidos e pela ausência de atributos representativos. A qualidade do conjunto de treinamento, por sua vez, é diretamente determinada pelo grau de correlação entre os registros existentes no conjunto de teste e aqueles que compõem o conjunto de treinamento. Além disso, a presença de registros com valores distantes do padrão normal seguido pela maioria dos registros pode distorcer a função-alvo gerada durante o treinamento. Finalmente, pode ser necessária a normalização de dados dos conjuntos de treinamento e de teste, para que alguns algoritmos possam interpretar corretamente as informações.

Neste estudo de caso, percebe-se a necessidade de inclusão de outros atributos possivelmente representativos. Durante a concepção do problema e elaboração da modelagem, cogitou-se a inclusão das seguintes variáveis ao modelo: número de filhos (0-...), número de vezes que prestou o concurso para a EsAEx (1-...), número de vezes que prestou outros concursos (1-...), possui pós-graduação (S, N), possui mestrado (S, N), possui doutorado (S, N), morava na EsAEx (S, N), nota na prova de conhecimentos gerais do concurso (5-10), nota total no concurso (5-10) e classificação geral no concurso (1-...). Esses atributos poderiam gerar uma representação mais fiel da realidade dos alunos do CFO/QCO. Contudo, não foi possível obter essas informações para a inclusão dessas variáveis no modelo. Por outro lado, alguns campos presentes na modelagem do problema, como a área de especialidade, podem apresentar pouca representatividade.

Em relação ao conjunto de treina-

mento, pode haver uma discrepância de perfis entre os formandos das turmas de 2005 e 2006, o que pode ter contribuído para perda de qualidade no processo de previsão do resultado. Ou seja, o padrão da turma de 2006 pode não coincidir com o padrão da turma de 2005. É possível que uma quantidade maior de dados permita caracterizar padrões de perfis. Além disso, a normalização dos dados referentes à classificação dos candidatos nas áreas do concurso poderia melhorar a fidelidade das informações, já que o número de candidatos e de vagas varia de um ano para o outro.

5 Conclusão

Este trabalho apresentou um método para o emprego de algoritmos de classificação no auxílio às tomadas-de-decisão estratégicas militares. É possível visualizar diversas oportunidades de aplicação do método aqui proposto no escopo militar. Decisões orçamentárias, referentes à movimentação de pessoal e escolha de militares para promoção são alguns exemplos que podem ser citados como beneficiários do emprego dessas técnicas de mineração de dados.

Um estudo de caso foi desenvolvido com o intuito de demonstrar a utilização do modelo proposto e confirmar os benefícios advindos do uso de algoritmos de classificação no apoio ao processo decisório. Foram detalhadas, nesse estudo de caso, as etapas a serem seguidas para que um resultado prático satisfatório seja atingido. Dessa forma, espera-se contribuir para o aprimoramento dos processos de tomada de decisões nas Forças Armadas Brasileiras.

Em trabalhos futuros, o modelo apresentado pode ser ainda mais detalhado, para que sejam apresentadas informações que acabaram sendo omitidas em função do escopo deste projeto. É possível também indicar caminhos alternativos às opções sugeridas, como forma de flexibilizar o modelo e possibilitar sua aplicação a um número maior de casos, além de permitir uma melhoria na sua eficiência. A extensão do modelo, através da criação de novas etapas ou do acréscimo de tarefas nas etapas, também pode contribuir para a melhoria do modelo. Finalmente, pode ser aprimorado o estudo de caso apresentado para que um desempenho mais expressivo seja alcançado ou mesmo desenvolvido um novo estudo de caso que permita atingir uma acurácia mais significativa.

Referências

BRITOS, P. et al. **Tool Selection Methodology in Data Mining**. Proceedings V Ibero- american Symposium On Software Engineering, Buenos Aires, p. 85-90. 2006. Disponível em: http://www.itba.edu.ar/capis/webcapis/RGMITBA/comunicacionesrgm/JIISIC-2006Tool-Selection-Methodology-in-Data-Mining.pdf>. Acesso em: 14 jul. 2007.

CARUANA, R.; NICULESCU-MIZIL, A. An empirical comparison of supervised learning algorithms. **Acm International Conference Proceeding Series:** Proceedings of the 23rd international conference on Machine learning, Pittsburgh, v. 148, p.161-168, 25 jun. 2006. Disponível em: <http://www.cs.cornell.edu/~caruana/ctp/ct.papers/caruana.icml06.pdf>. Acesso em: 27 jul. 2007.

EXÉRCITO BRASILEIRO. **Como ingressar no Exército**. Disponível em: <http://www.exercito.gov.br/02ingr/ingressar.htm>. Acesso em: 17 jun. 2007b.

KDNUGGETS. **Poll: Data Mining / Analytic Software Tools** . Disponível em: <http://www.kdnuggets.com/polls/2007/data_mining_software_tools.htm>. Acesso em: 14 jul. 2007.

MORAES, S. **Descoberta do Conhecimento em Base de Dados: uma breve visita.** UCB, Brasília, 2002. Disponível em: <http://www.fazenda.gov.br/ucp/pnafe/docs/p2-S%C3A9rgio.pps>. Acesso em: 24 Jun 2007.

MURAKAMI, M. **Decisão estratégica em TI: estudo de caso.** Dissertação de mestrado, USP, 2003. Disponível em:<http://www.teses.usp.br/teses/disponíveis/12/12139/tde-19112003-200926/>. Acesso em: 24 Jun 2007.

OLIVEIRA, D. P. R. **Planejamento estratégico: conceitos, metodologia e práticas.** 5 Ed. São Paulo: Atlas, 1991. 267p.

SCHNEIDER, L. F. **Mineração de Dados (Data Mining) – Conceitos**. Departamento de Agronomia, UFRS, Porto Alegre, 2002. Disponível em: <http://www.inf.ufrgs.br/~clesio/cmp151/cmp15120021/artigo_lfelipe.pdf>. Acesso em: 24 Jun 2007.

SILVA, M. P. S. **Mineração de Dados : Conceitos, Aplicações e Experimentos com Weka**. Livro da Escola Regional de Informática Rio de Janeiro - Espírito Santo. Porto Alegre: Sociedade Brasileira de Computação, 2004, v. 1, p. 1-20. Disponível em: <http://www.sbc.org.br/bibliotecadigital/download.php?paper=35> Acesso em: 14 jul. 2007.

SOARES, J. A. **Pré-Processamento em Mineração de Dados: Um Estudo Comparativo em Complementação**, Tese de Doutorado, UFRJ, 2007. Disponível em: <http://www.jsoares.net/artigos/TeseDScJAS.pdf>. Acesso em: 07 Jul 2007.

TAN, P. N.; STEINBACH, M.; KUMAR, V. **Introduction to Data Mining.** Boston: Pearson, 2006. 769p.

UNIVERSITY OF WAIKATO. **Weka 3 – Data Minig with Open Source Machine Learning Software in Java**. Disponível em: <http://www.cs.waikato.ac.nz/ml/weka>. Acesso em: 14 jul. 2007.

WEBB, G. I. MultiBoosting: A Technique for Combining Boosting and Wagging. **Machine Learning**, Boston, v. 40, n. 2, p.159-196, ago. 2000. Disponível em: <http://www.springerlink.com/index/G7K410V232R15363.pdf>. Acesso em: 27 jul. 2007.