Sistema de Apoio à Decisão para Estimação do Sucesso do Aluno no Programa de Mestrado em Ciência da Computação da UFPE

Paulo J. L. Adeodato
UFPE - Recife,Brasil
NeuroTech, Recife, Brasil
paulo@neurotech.com.br

Adrian L. Arnaud
UFPE - Recife, Brasil
ala2@cin.ufpe.br

Claudia C. C. Salgues
UFPE - Recife, Brasil
cccs@cin.ufpe.br

Sílvio R. L. Meira
UFPE-Recife,Brasil
srlm@cin.ufpe.br

Abstract

The candidates for the Center for Computer Science postgraduate programme of UFPE, Brazil, are selected based on their CVs. Five attributes are linearly weighed to produce the candidate's score, which serves as ranking criterion. The postgraduate collegiate was concerned this process could be restricting access to the programme. This paper analyzes the profiles both of the 148 students and of the other 666 candidates who did not enter the programme in the last 4 years. This research concludes that the 5 attributes used could not discriminate between good and weak students. The 4 additional attributes made it feasible. Moreover, this paper has proposed an effective system and has shown that the previous system was indeed being restrictive.

Keywords*: Decision support systems, Data mining, Postgraduate education program, Higher education quality assessment, Knowledge discovery in databases.*

1. Introdução

Os custos de seleção de pessoal altamente qualificado são elevados. O processo de seleção de candidatos ao mestrado e a uma bolsa de estudos dentre os candidatos aprovados no Centro de Informática da UFPE têm-se baseado na análise curricular do candidato. Há alguns anos a avaliação foi sistematizada por meio de uma ponderação dos atributos curriculares do candidato.

Este processo de seleção vinha sendo questionado pelos professores do Centro, membros do colegiado, por poder não estar capturando aspectos relevantes para a definição do desempenho do aluno ao longo do curso. No processo seletivo de 2005, houve uma grande mudança que permitiu a aceitação de um total de 114 candidatos, provocando uma expansão de cerca de 35% do volume de alunos (antes, eram cerca de 85 novos alunos por ano, dos quais cerca de 65 se matriculavam).

Neste trabalho, objetivamos analisar a influência dos diversos aspectos do processo seletivo na qualidade do programa sob vários dos seus pontos de vista para podermos quantificar os impactos de cada decisão tomada pelo colegiado. Também queremos criar uma nova forma de pontuação para ordenar os alunos aceitos quanto à prioridade na alocação das bolsas de estudos.

O artigo está organizado em 9 seções. A seção-2 mostra a forma atual de cálculo da pontuação. A seção-3 detalha o trabalho da coleta à limpeza dos dados. A seção-4 explica as transformações dos dados. A seção-5 apresenta a análise de correlação das variáveis de entrada com as variáveis indicadoras de desempenho. A seção-6 analisa os modelos, otimiza os seus parâmetros pela regressão linear e gera uma solução não-linear baseada na regressão logística. A seção-7 induz as regras explicativas dos diversos perfis dos alunos. A seção-8 analisa os candidatos não aceitos e mostra que o sistema em vigor está sendo restritivo. A seção-9 apresenta as conclusões, avalia como o presente trabalho modifica a forma de analisar os impactos e faz sugestões para aprimorar o processo de seleção dos candidatos.

2. Cálculo da Pontuação

Atualmente, a pontuação de cada candidato é uma soma ponderada de cinco variáveis de entrada obtidas da sua documentação de inscrição. Entram nesse cálculo a nota do histórico (Hist) ponderada pela nota da universidade de graduação (FRed), a nota de cada uma das duas cartas de recomendação (Carta1 e Carta2), a nota de publicações (Publ) e a nota de experiência profissional (Exp) conforme a equação abaixo.

$$Y = a_1*Hist*FRed + a_2*Carta1 + a_3*Carta2 + a_4*Publ + a_5*Exp \quad \text{(Eq.1)}$$

onde os coeficientes a_i são constantes que somam um (1), de forma que a pontuação final é um valor entre 0 e 10.

3. Coleta, Digitação, Integração e Limpeza dos dados

Usualmente, esta etapa de um trabalho de descoberta de conhecimento a partir de dados (Knowledge Discovery in Databases – KDD) [1] consome até 80% do tempo gasto no projeto, mas neste caso, em particular, consumiu mais de 90% porque tivemos que digitar dados de fichas de inscrição e de relatórios impressos. Essa tarefa só foi possível em virtude da reduzida massa de dados: apenas 148 alunos dos que haviam ingressado no período haviam concluído o curso ou expirado o prazo até abril de 2004 e tinham todas informações *a posteriori* disponíveis para realizarmos a aprendizagem supervisionada (os rótulos). Mais 666 candidatos não aprovados também tiveram dados adicionais coletados, apenas das suas fichas de inscrição.

Os dados das cinco variáveis de entrada (informação *a priori*) de todos os candidatos já estavam em formato Excel®. Os indicadores de desempenho, porém, estavam dispersos em diversos documentos, alguns deles impressos. As notas estavam em documentos isolados, no formato Word®, organizados por aluno. A declaração de conclusão, também em formato Word®, continha o fato e a data de conclusão. Os relatórios da CAPES em formato PDF continham os artigos publicados pelos alunos, em conseqüência da pesquisa do mestrado. As variáveis adicionais foram extraídas das fichas de inscrição impressas. A integração e a limpeza desses dados tiveram que ser feitas ao longo do processo de coleta.

4. Descrição das Variáveis e Transformação dos Dados

As variáveis atualmente utilizadas no cálculo da pontuação de cada candidato são cinco e se encontram indicadas abaixo. Esta é a "visão atual":

1. Nota do histórico de graduação ponderada por pelo fator atribuído à universidade com base na avaliação da CAPES (Hist*FRed)
2. Nota da carta de recomendação de maior nota. A nota considera tanto a qualidade do candidato quanto o vínculo com o professor recomendante e a sua titulação (Carta1)
3. Nota da outra carta de recomendação (Carta2)
4. Nota de publicações onde o candidato é um dos autores (Publ)
5. Nota da experiência profissional em iniciação científica, monitoria ou emprego (Exp).

Os avaliadores das cartas de recomendação, publicações realizadas e experiência profissional têm liberdade para variar apenas cerca de 10% de uma nota em relação à do seu par no processo de seleção.

Além das variáveis acima, foram criadas outras indicadas abaixo. Estas compõem a "visão ampliada":

1. A nota do histórico de graduação foi decomposta nas suas duas variáveis originais (Nota do histórico original (Hist) e Fator de ponderação da universidade (FRed))
2. Número de candidaturas anteriores (NCand)
3. Sexo (Sexo)
4. Idade (Idade).

Fizemos entrevistas com o presidente da comissão de seleção do mestrado de 2005 e com coordenador e vice do programa de pós-graduação para validarem as definições adotadas quanto aos objetivos e medidas de desempenho do sistema, seguindo a recomendação da metodologia CRISP-DM [2] neste projeto.

Diante dos indicadores de desempenho levantados (informação *a posteriori*), houve unanimidade em considerar a conclusão no prazo como o objetivo mais importante do programa. Outros fatores também foram citados como relevantes e todos estão listados abaixo.

1. Ter concluído o curso no prazo definido (91 casos de sucesso na amostra)
2. Ter publicado artigos resultantes do mestrado
3. Não ter abandonado o curso
4. Não ter feito trancamento do curso
5. Não ter tido reprovação (uma nota D ou duas C)

É importante considerar que estamos definindo objetivos dicotômicos. Desses objetivos, utilizamos apenas os dois primeiros como alvos do sistema e uma combinação dicotômica de todos num alvo composto denominado "*desempenho*" que define que um bom aluno atende a pelo menos três dos objetivos acima. Na avaliação de desempenho, os impactos das decisões serão medidos sobre estes 3 objetivos e em outros mais globais.

Os valores dessas variáveis foram transformados pelas chances *p/(1-p)* (em inglês, odds) para se tornarem compatíveis com o classificador de regressão logística [3] e discretizados em faixas para permitir a indução das regras pelo método "*A priori*" [4].

5. Análise de Correlação

Tendo em vista a intenção de preservar o sistema linear já existente, tanto pela sua facilidade de interpretação quanto pelo conhecimento disseminado entre os avaliadores, a análise de correlação [5], é uma das formas mais adequadas de se medir o potencial de resolvermos o problema por meio de um sistema linear.

A matriz de correlação (abaixo) apresenta em cada eixo (linhas e colunas) as variáveis de entrada e de saída do sistema. Cada célula, em geral, contém o coeficiente de correlação entre as variáveis envolvidas, o nível de significância estatística da correlação medida e o tamanho

da amostra. Neste estudo utilizamos o coeficiente de correlação de Pearson [5] e as correlações estatisticamente significativas estão indicadas em negrito, para 0,01 de significância, e em itálico, para 0,05 de significância. Como a matriz é simétrica, deixamos em branco a sua parte triangular superior para facilitar a leitura. A figura abaixo mostra a matriz de correlação entre as variáveis da visão ampliada para a massa de 148 alunos com o desempenho conhecido.

Na *visão atual* do problema (variáveis de números 1 a 5 e alvos de números 10 e 11), vê-se que apenas a variável de entrada *nota do histórico* está correlacionada com apenas a resposta desejada (alvo) *titulação no prazo*, ainda, com baixa correlação e o menor grau de precisão. Ou seja, sistemas lineares não conseguem distinguir, de forma univariada, entre bons alunos e alunos fracos com base nessa visão do problema. Além disso, existe alguma correlação entre as variáveis de entrada.

Tabela 1. Matriz de correlação das variáveis.

		1	2	3	4	5	6	7	8	9	10	11	12
Hist.	1	1											
FRed.	2	,15	1										
Carta1	3	**,29**	*,18*	1									
Carta2	4	**,27**	*,19*	**,52**	1								
Exp.	5	*,18*	**,25**	**,24**	**,34**	1							
Publ.	6	,15	**,26**	**,22**	**,24**	*,18*	1						
NCand.	7	-,11	,00	,07	*-,18*	,06	,14	1					
Idade	8	-,15	*-,21*	**-,30**	**-,35**	-,08	**-,28**	,05	1				
Sexo	9	,10	,07	,02	,09	-,14	-,05	-,11	,00	1			
TPraz.	10	*,17*	-,05	,14	,11	-,01	,14	-,12	-,14	*,19*	1		
Public.	11	,01	,09	,09	,02	,05	,08	-,05	-,12	,14	**,28**	1	
Des.	12	*,16*	-,05	**,36**	**,28**	,06	,08	**-,22**	**-,26**	,12	**,50**	**,25**	1

Considerando os dois princípios mais importantes num processo binário de tomada de decisão

- Maximizar a "correlação" (similaridade) entre cada variável de entrada e a de saída, e
- Minimizar a "correlação" (similaridade) entre as variáveis de entrada,

vemos que o problema não permite, pelo menos de forma linear univariada, que a decisão seja feita com qualidade.

Na visão ampliada do problema, temos mais 4 variáveis de entrada e 1 de saída (variáveis de números 6 a 9 e alvo de número 12), que preserva a redundância das variáveis anteriores, mas agrega valor com a variável de saída — *desempenho* — criada na seção-4. O alvo *desempenho* apresenta uma correlação acima de 0,2 no maior nível de precisão, com 4 das variáveis de entrada, duas das quais fazem parte apenas da visão ampliada. O *número de candidaturas anteriores* e a *idade* são negativamente correlacionadas com o *desempenho*, enquanto as *cartas de recomendação* são positivamente correlacionadas. Assim, o alvo *desempenho* criado oferece a perspectiva de resolvermos o problema com sistemas lineares.

6. Soluções para o problema

Uma vez verificado que o sistema linear é capaz de discriminar entre bons alunos e alunos fracos, parametrizamos uma solução linear e outra não linear com o objetivo de minimizar o erro médio quadrático das decisões sobre a massa rotulada. A variável *desempenho,* criada aqui, e a visão ampliada de nove (9) variáveis de entrada foram usadas. Foi mantido o modelo original do especialista humano e foram produzidos, um modelo linear, pela clássica regressão linear [5] e, outro não linear, pela clássica regressão logística [3].

6.1. Métricas de desempenho das soluções

Em condições de pequena disponibilidade de dados, como aqui, são recomendados procedimentos de máximo aproveitamento dos dados para avaliação de desempenho e comparação entre técnicas de inferência. Utilizamos o procedimento leave-one-out [6] que consiste em realizar *N* vezes a parametrização do sistema com *N*-1 exemplos, avaliando o seu desempenho no exemplo restante. Ao final, temos o resultado que a técnica de inferência produziu sobre os *N* exemplos disponíveis. Neste artigo, todos gráficos apresentados sobre os dados rotulados foram produzidos de acordo com este procedimento. Por outro lado, para inferência sobre os candidatos não aceitos nas seleções (dados não rotulados), os resultados foram produzidos pela equação obtida com todos os 148 exemplos da amostra rotulada.

Como os modelos considerados oferecem decisão suave (variável de saída contínua – pontuação ou *escore*), medimos o desempenho por meio do teste Kolmogorov-Smirnov (KS-2) [7]. Em função de termos apenas 148 exemplos na massa de dados rotulada, os gráficos têm uma aparência "dentada" a cada novo elemento, ao longo da escala de pontuação. Outra métrica usada aqui é a tabela com os impactos dos valores do limiar de decisão numa regra binária baseada na variável *escore*.

O KS-2 é um teste não paramétrico criado para medir a aderência de dados a uma distribuição. Na Seção-8, fazemos esse uso do teste. Na Seção-6, por outro lado, e em sistemas decisórios em geral, ele serve para medir a separabilidade entre duas distribuições a partir da função de distribuição acumulada de cada uma. O KS-2 mede a máxima distância entre duas funções como uma medida de separação entre elas e pondera os erros pela probabilidade de ocorrência de cada classe.

Os impactos dos valores do limiar de decisão nas regras é medido pelos tradicionais métodos de avaliação de qualidade das regras: cobertura, confiança e *lift* [8].

6.2. Solução linear

A equação foi obtida a partir de todos os 148 exemplos da amostra rotulada pela regressão linear. Apenas os 4 maiores coeficientes (em negrito) foram estatisticamente significativos com grau de 5% de significância e bem diferente das demais variáveis.

Y = 0.15 + 0.97***Carta1** – 0.39***NCand** – 0.31***FRed** –
– 0.30***Idade** + 0.08*Sexo + 0.06*Hist +
+ 0.03*Carta2 + 0.02*Publ + 0.02*Exp (Eq.2)

Os valores absolutos dos coeficientes indicam a influência de cada variável na tomada de decisão e o sinal associado indica se a variável influencia positiva ou negativamente o sucesso do aluno. Comparando esses valores com os da fórmula do especialista, abaixo,

Y = *0.5*Hist*FRed + *0.15*Carta1 + *0.15*Carta2 +
+ *0.05*Publ + *0.15*Exp (Eq.3)

vemos que não há muita semelhança nos pesos, a não ser o do *Fator de ponderação da universidade* na Eq.2 com o mesmo ponderado pela *Nota do histórico* na Eq.3.

6.3. Solução com regressão logística

Em geral, sistemas não lineares podem gerar melhor desempenho que os lineares. Enquanto o sistema linear apenas resolve problemas lineares, a regressão logística aplica uma transformação não linear sobre cada variável antes de ponderá-la linearmente, tem alto poder de generalização e não precisa de hipóteses especiais sobre a distribuição estatística dos dados [3].

Utilizando a mesma amostra de dados do sistema linear, a regressão logística produziu a equação abaixo e os 5 coeficientes em negrito foram estatisticamente significativos com grau de 5% de significância.

Logit(p) = – 7.76 + 10,26***Carta1** – 5.80***FRed** –
– 3.13***Hist** – 2.88***Idade** – 2.63***NCand** +
+ 1.82* Carta2 + 0.80*Exp + 0.66*Sexo –
– 0.48*Publ (Eq.4)

A variável dependente é o logaritmo da razão entre as probabilidades dos dois possíveis resultados da variável de saída, {log[*p*/(1-*p*)]} e pode ser expressa como $P(Y=1) = [1+\exp(-\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k)]^{-1}$, onde Y é a variável dicotômica de saída, β_0 é a interseção e os β_i (i=1,...,k) são coeficientes de cada variável explicativa x_i.

Os valores absolutos dos coeficientes indicam a influência de cada variável transformada na tomada de decisão e o sinal associado indica se a variável influencia positiva ou negativamente o sucesso do aluno.

O resultado é similar ao da regressão linear tendo *Carta1, Fator de ponderação da universidade*, *Número de candidaturas anteriores* e *Idade* como variáveis de maior peso e com significância, comuns e com os mesmos sinais de correlação. Apenas a variável *Nota do histórico* foi acrescentada à lista das significantes.

Também, há similaridade com a matriz de correlação: em relação ao alvo, *Carta1* teve correlação positiva e *Número de candidaturas anteriores* e *Idade* tiveram correlação negativa, todas significantes e com alto valor.

6.4. Avaliação de desempenho das soluções

Os resultados de desempenho com o KS-2 sobre o conjunto de dados rotulado está ilustrado abaixo para os três modelos: especialista, regressão linear e regressão logística. A pontuação foi normalizada entre 0 e 1.

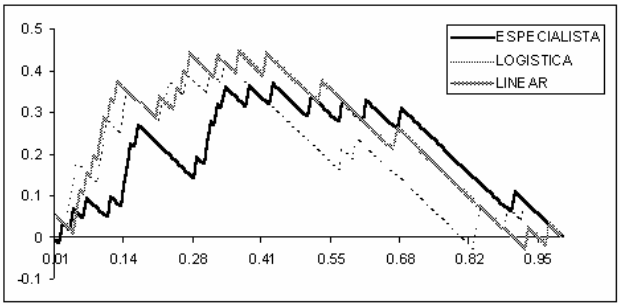

Figura 1. Curva de Kolmogorov-Smirnov (KS-2) para o desempenho dos três modelos avaliados.

A tabela abaixo mostra que, para o alvo *desempenho* do aluno, todas as técnicas produziram resultados estatisticamente significativos a 5% de significância. O resultado obtido pela regressão linear, além de superior aos demais, foi o de melhor confiança.

Tabela 2. KS-2 máximo para os três modelos.

Modelo	Especialista	Linear	Logística
$KS\text{-}2_{Máx}$	0,37	**0,45**	0,42
p-Value	$1{,}3 \times 10^{-2}$	$1{,}3 \times 10^{-3}$	$3{,}0 \times 10^{-2}$

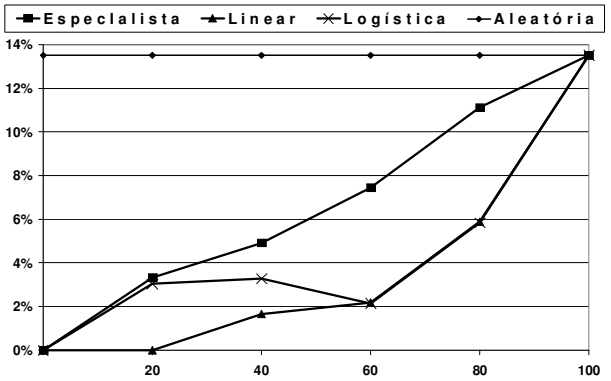

Figura 2. Percentual de insucesso dentre os alunos aceitos em função do percentual de aceitação.

A Figura 2 confirma que existe poder discriminante de todos os modelos com diversos limiares de aceitação

aplicados à massa de 148 alunos rotulados. A seleção aleatória serve de referencial do poder discriminante. Vê-se, também que a regressão linear tem melhor desempenho que a regressão logística que, por sua vez, é melhor que a fórmula vigente, definida pelo especialista e que todos os modelos têm poder discriminante em relação ao objetivo.

7. Perfis dos Alunos

Tendo validado os sistemas decisórios linear e logístico, podemos caracterizar os perfis dos alunos tanto que tenham alta chance de sucesso quanto alta chance de insucesso no curso. Utilizamos uma versão adaptada do algoritmo "*A Priori*" [4] para identificar as características desses nichos e listamos, abaixo, 3 das regras que melhor caracterizam cada categoria.

Tabela 3. Regras induzidas pelo algoritmo *apriori.*

Regra - Sucesso	Sup %	Conf %	*Lift*	Regra - Insucesso	Sup %	Conf %	*Lift*
Se Carta1 > 9,7 e 22 < Idade < 25	16,2	100	1,16	Se 0,95 < FRed < 1,01	52,0	19,5	1,44
Se 8,1 < Hist < 8,6 e 8,3 < Carta2 < 9	13,5	100	1,16	Se 0,93 < FRed < 0,95 e 22 < Idade < 25	10,1	20,0	1,48
Se 0,90 < Fred < 0,93	11,5	100	1,16	Se 7,22 < Exp < 8,33	15,5	17,4	1,29

Os perfis mais interessantes são caracterizados por meio das variáveis mais relevantes nos modelos. A variável *Idade*, descartada do modelo atualmente em uso, tem fundamental importância no sucesso do aluno. Essa mesma variável aparece com a mesma faixa etária associada a perfis de bom aluno e de aluno fraco, dependendo da outra variável da condição.

8. Caracterização de Perfis dos Candidatos que não Cursaram o Mestrado

Depois de validar os sistemas decisórios linear e logístico, avaliamos o impacto que eles teriam tido no contingente de alunos não selecionados. Dos 666 alunos que não ingressaram no curso, cerca de 45 candidatos aprovados, fizeram outra opção profissional. Eles correspondem a cerca de 30% dos 148 que ingressaram.

Submetendo estes 666 candidatos aos diversos sistemas pudemos medir o KS-2 entre as distribuições dos candidatos não aceitos e dos aceitos (alunos que cursaram o mestrado). Neste caso, o KS-2 mede a aderência entre as distribuições, como era seu propósito original na Estatística. Assim, quanto menor o KS-2 máximo, mais semelhantes são os perfis. O KS-2 de todos os modelos deu diferença a 5% de significância entre as distribuições. Os modelos linear e logístico foram os que menor discrepância apresentaram.

Vemos que, de fato, o perfil de pontuação dos candidatos selecionados (alunos) é muito superior ao dos que não cursaram, na sua maioria, não aceitos, uma vez que apenas 45 dos 666 haviam sido aceitos e desistido. O KS máximo indica que há uma diferença estatisticamente significativa entre as distribuições, como na tabela abaixo que resume as principais características quantificáveis.

Tabela 4. KS-2 máximo para os três modelos avaliados entre as distribuições dos candidatos que não cursaram o mestrado em relação aos que cursaram.

Modelo	Especialista	Linear	Logística
$KS\text{-}2_{Máx}$	0,43	0,19	0,19
p-Value	$2{,}4 \times 10^{-20}$	$2{,}3 \times 10^{-4}$	$3{,}7 \times 10^{-4}$

Se, por um lado, esses resultados confirmam a diferença na qualidade média dos selecionados, vários indivíduos não selecionados poderiam ter tido sucesso, caso tivessem ingressado no mestrado. A figura abaixo mostra o percentual de aceitação dentre todos os candidatos para cada modelo avaliado em função do ponto de corte definido sobre o percentual dos alunos (*quantis*) que efetivamente cursaram o mestrado. A reta diagonal indica qual seria o percentual, caso as distribuições dos candidatos que não cursaram fosse idêntica à dos alunos efetivos. Vemos que a regressão linear e a logística são bem mais similares à essa reta que a pontuação do especialista, como era de se esperar pelo valor do $KS\text{-}2_{Máx}$. A aparente estranheza de haver candidatos que não cursaram o mestrado nos *quantis* superiores na pontuação do modelo do especialista se deve aos candidatos aceitos que tomaram outro rumo e a pequenos ajustes da fórmula do especialista a cada ano do processo seletivo.

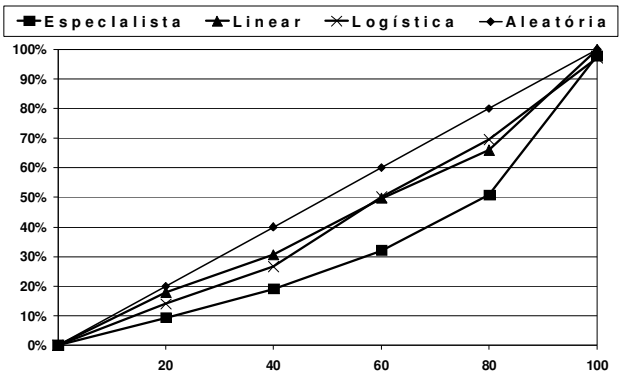

Figura 3. Percentual de aceitação dos candidatos em função do percentual de aceitação dos alunos.

Finalmente, apresentamos na tabela abaixo, os impactos da tomada de decisão com base na pontuação do modelo de regressão linear sobre os diversos universos considerados. Os resultados abaixo, porém, devem ser interpretados com cuidado, uma vez que o $KS\text{-}2_{Máx}$

mostrou uma diferença estatisticamente significativa entre as distribuições de alunos e de candidatos. A regressão linear foi utilizada na extrapolação abaixo porque apresentou menor discrepância e qualitativamente é bem superior às demais.

Tabela 5. Impactos do limiar de decisão sobre todos os candidatos e seu desempenho estimado pelo daqueles que cursaram o mestrado.

% Rot Aceitos	% Cand.	No. Cand.	% Insuces.	No. de Insuces.	No. de suces.
0%	0.0%	0	0.0%		
20%	17.9%	146	0.0%	0	146
40%	30.6%	249	1.7%	5	244
60%	49.8%	405	2.2%	9	396
80%	66.0%	537	5.9%	32	505
100%	100.0%	814	13.5%	110	704

9. Considerações Finais

Este artigo apresentou uma avaliação do processo de seleção de candidatos ao mestrado em Ciência da Computação da UFPE no período de 1999 a 2002 e propôs um modelo de inferência de melhor qualidade na estimação do sucesso dos candidatos como alunos programa. O grande impacto do modelo proposto seria aumentar o volume de candidatos aceitos preservando ou melhorando o nível de qualidade do programa.

Verificou-se ainda que, para os alunos analisados, o critério de seleção vigente na época não fazia diferença estatisticamente significativa entre os alunos que concluiriam no prazo e os que se atrasariam ou não o concluiriam. Quatro fatores podem ter contribuído para este resultado:

1. A reduzida visão sobre as variáveis de entrada consideradas na tomada de decisão que pudemos fundamentar pela melhora de desempenho do sistema com a visão expandida.
2. A reduzida capacidade discriminante de um sistema de decisão linear não adaptativo que foi confirmada pela melhora de desempenho obtida com as técnicas de regressão linear e logística.
3. O excesso de rigor no processo de seleção que pode estar aceitando apenas candidatos com perfil técnico muito bom.
4. As variáveis de entrada estarem restritas a aspectos exclusivamente técnicos que, para alunos muito bons, não seriam determinantes sobre o sucesso.

A hipótese 3, graças à ampliação de cerca de 35% da aceitação no processo seletivo para 2005, poderá estar sendo confirmada a partir março de 2007, quando os 114 candidatos deveriam ter concluído o curso. Assim, dos 114 candidatos aceitos, alguns poderiam estar abaixo do padrão permitindo que identificássemos o "joelho" da curva que indica o início da perda de qualidade.

A hipótese 4 teve indícios da sua validade com a maior influência da *Carta de aceitação 1* (normalmente escrita por um professor que acompanhou de perto o aluno na sua graduação ou iniciação científica). O sucesso dependeria muito mais da determinação pessoal, de percalços na vida pessoal etc. manifestados nas cartas de recomendação. Assim, uma natural extensão deste trabalho é considerar a nota de cada item das cartas como variáveis de entrada (vínculo com o candidato, atividade realizada, iniciativa, perseverança, comunicação escrita, comunicação verbal, trabalho em equipe etc.).

É importante considerar que o atual sistema de acompanhamento de desempenho do mestrado alterou o comportamento dos alunos, principalmente, no que diz respeito à sua responsabilidade diante da instituição, forçando-o a prestar contas semestralmente a avaliadores independentes.

O sistema proposto neste artigo já se constitui num avanço em relação ao vigente e ainda poderá ser muito melhorado a partir da informação obtida com a implantação do sistema de acompanhamento escolar em formato eletrônico pela "web" para a pós-graduação. O sistema aqui apresentado está em consideração pelo Colegiado da pós-graduação e poderá ser colocado em operação, diante do parecer.

10. Referências

[1] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "The KDD process for extracting useful knowledge from volumes of data", *Commun. ACM*, Vol. 39(11), 1996, pp. 27—34.

[2] R. Wirth, and J. Hipp, "CRISP-DM: Towards a standard process model for data mining", in *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining* (PADD00), 2000.

[3] Pampel, F.C., *Logistic regression — A primer*, Sage, Thousand Oaks, CA, 2000.

[4] Han J., Kamber M., *Data mining: concepts and techniques*, Morgan Kaufmann, San Francisco, CA, 2001.

[5] Johnson R.A., and Wichern D.W., *Applied multivariate statistical analysis*, 4th ed. Prentice hall, Upper Saddle River, New Jersey, 1998.

[6] A.K. Jain, R.P.W. Duin., J. Mao "Statistical pattern recognition: a review", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22(1), 2000, pp.4—37.

[7] Conover, W.J., *Practical Nonparametric Statistics*, Chap. 6, Third Edition, John Wiley & Sons, New York, 1999.

[8] Witten, I.H., and Frank, E., *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, M. Kaufmann, San Francisco, CA, 2000.