

Análise de agrupamento de escolas e Núcleos de Tecnologia Educacional: mineração na base de dados de avaliação do Programa Nacional de Informática na Educação

Maria de Fátima Ramos Brandão¹,

Carlos Renato dos Santos Ramos¹,

Bartholomeu T. Tróccoli²

¹Instituto de Ciências Exatas – Departamento de Ciência da Computação

²Instituto de Psicologia – Departamento de Psicologia Social e do Trabalho
Universidade de Brasília (UnB)

Caixa Postal 4500 – 70.910-900 - Brasília – DF – Brasil

Tel/Fax: 61- 273-3589 ramal :206

fatima@cic.unb.br, cr.ramos@ig.com.br, troccoli@unb.br

Resumo. O trabalho apresenta uma análise de agrupamento de escolas e Núcleos de Tecnologia Educacional (NTE) do Programa Nacional de Informática na Educação (ProInfo) segundo um modelo de avaliação definido por cinco padrões referenciais. Os escores de desempenho, obtidos pelas escolas e NTE nos cinco padrões, foram analisados utilizando-se mineração de dados com método de agrupamento. Os grupos identificados podem ser considerados como blocos experimentais em investigações futuras e subsidiar estudos sobre impactos e fatores críticos de sucesso de programas educacionais.

Palavras-chave: descoberta de conhecimento em bases de dados, mineração de dados, análise de agrupamento, avaliação.

Abstract. This work presents a cluster analysis of Schools and Educational Technology Centers (NTE) of the National Educational Computer Science Program (ProInfo). The analysis followed an evaluation model comprised of five reference standards. The schools and NTE performance scores obtained in these five standards were analyzed with the data mining clustering method. The groups that have been identified should be considered as experimental blocks in future investigations and should be used as the foundations for studies of impacts and critical success factors in educational programs.

Keywords: knowledge discovery in databases, data mining, clustering, evaluation.

1. Introdução

A análise de agrupamento é uma técnica multivariada cujo objetivo primário é agrupar objetos ou entidades com base em características comuns, utilizando um determinado critério de seleção. Os agrupamentos resultantes (*clusters*) apresentam alta homogeneidade interna e alta heterogeneidade externa [Hair & Black, 2000]. O trabalho apresenta análises de mineração de dados com os resultados obtidos com agrupamentos das escolas e NTE (Núcleos de Tecnologia Educacional) do Programa Nacional de Informática na Educação (ProInfo). Foram utilizados como critério de classificação cinco indicadores de desempenho, segundo os resultados da avaliação do programa [Brandão, Tróccoli & Guedea, 2003]. A mineração foi realizada utilizando a ferramenta WEKA [Witten & Frank, 1999].

A análise de agrupamento aplicada aos resultados de avaliação do ProInfo tem por objetivo identificar escolas e NTE que apresentam desempenhos semelhantes nos cinco indicadores de avaliação do uso pedagógico da tecnologia de informação e comunicação. Trata-se de identificar e segmentar em blocos experimentais a amostra estratificada nacional, segundo os seus valores de desempenho, para possibilitar análises posteriores de fatores de sucesso do Programa. Espera-se com esses resultados contribuir para uma melhor compreensão dos fenômenos e mudanças observadas nos contextos escolares desde a implantação do ProInfo.

2. A avaliação nacional do ProInfo

O Programa Nacional de Informática na Educação (ProInfo), criado em 1997, é uma iniciativa do Ministério da Educação que tem como metas a instalação de computadores nas escolas públicas e a capacitação de professores para trabalhar com recursos de telemática em sala de aula [MEC/SEED, 2002]. O ProInfo utiliza Núcleos de Tecnologia Educacional (NTE), que são estruturas descentralizadas onde professores multiplicadores e técnicos de informática dão suporte ao uso pedagógico da informática nas escolas de sua área de atuação.

Foram analisados dados obtidos em uma amostra nacional de 546 escolas que receberam recursos do ProInfo e 263 NTE instalados pelo programa até o ano de 2001. Foram aplicados sete tipos de questionários envolvendo diretores, coordenadores de informática, professores e alunos nas escolas, e ainda, coordenadores, professores multiplicadores e técnicos de informática nos NTE, totalizando 11.237 questionários (Tabela 1).

Tabela 1 – Número de questionários aplicados por grupo de respondentes

Questionário	Quantidade
Diretor de escola	551
Coordenador de Informática de escola	476
Coordenador de NTE	246
Professor Multiplicador de NTE	844
Técnico de Informática de NTE	213
Professor	3.541
Aluno	5.366
TOTAL	11.237

O tratamento censitário foi aplicado para todos os NTE e escolas situadas nas Unidades da Federação com menos de 30 escolas, enquanto que o tratamento por amostragem estratificada foi aplicado nas escolas das demais unidades da Federação [MEC/SEED/UnB, 2002]. Os dados coletados constituíram um banco de dados abrangente contendo 1.180 variáveis e 11.237 registros.

A avaliação do programa considerou um modelo que atribuiu às unidades pesquisadas escores de desempenho relativos a cinco indicadores de análise, denominados de padrões referenciais de: infraestrutura física, tecnológica e humana (PR1); direção e gestão administrativa e pedagógica (PR2); apropriação da informática e da cultura educativa (PR3); aperfeiçoamento e expansão do programa (PR4) e inovação e desenvolvimento social (PR5) [Brandão, Tróccoli & Guedea, 2003]. Os escores de

desempenho obtidos pelas escolas e NTE nos cinco indicadores, quantificados por valores entre 0 e 1, orientaram a análise da implantação do programa nas unidades pesquisadas e serviram de base para estudos exploratórios e qualitativos sobre o programa.

As médias, com seus desvios padrão, e as medianas dos escores de desempenho obtidos pelas escolas e NTEs nos cinco indicadores de padrões são apresentados na Tabela 2.

Tabela 2 – Desempenho médio, mediano e desvios padrão das Escolas e NTE

Escolas			NTE	
Padrão	Média (desvio padrão)	Mediana	Média (desvio padrão)	Mediana
PR1	0,14 (0,06)	0,13	0,10 (0,05)	0,10
PR2	0,18 (0,06)	0,18	0,07 (0,03)	0,07
PR3	0,12 (0,06)	0,12	0,10 (0,05)	0,10
PR4	0,07 (0,05)	0,06	0,13 (0,05)	0,11
PR5	0,04 (0,04)	0,03	0,07 (0,07)	0,06

Os valores apresentados na tabela 2 indicam que as médias das escolas diferem das médias dos NTE (as médias não diferem das medianas). Entretanto, os desvios padrão apresentam, de forma geral, baixa dispersão. Esses resultados evidenciam que escolas e NTE apresentam comportamentos distintos nos padrões e, portanto, não constituem um conjunto homogêneo. Gráficos de barra com os valores de desempenho das escolas e NTE são apresentados na Figura 1.

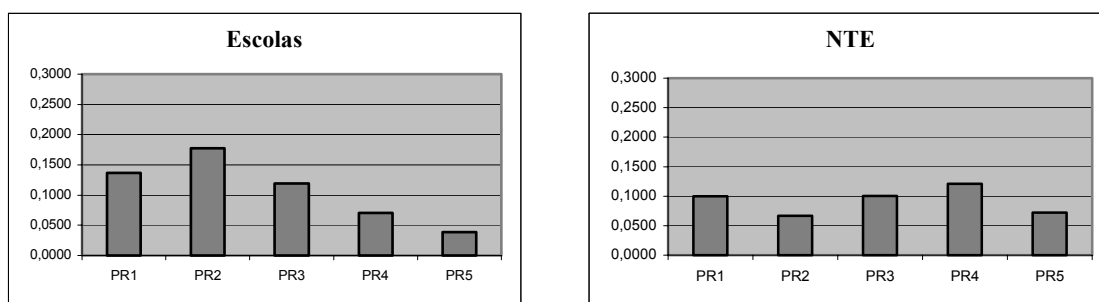


Figura 1 – Gráfico dos desempenhos médios das Escolas e NTE

Com exceção do padrão 1 (PR1), valores mais altos na Tabela 2 indicam maior percentual de implementações nos indicadores relativos aos padrões referenciais de PR2 a PR5. Os resultados das escolas e NTE no padrão 1 apresentam correlações negativas com as demais medidas dos padrões porque este tem características de situações críticas (e não de sucesso) no que diz respeito à infra-estrutura física, tecnológica e de recursos humanos. Assim, quanto mais baixo for o valor da medida de desempenho no indicador relativo ao padrão PR1, maior o percentual de implementação da infra-estrutura física, tecnológica e de recursos humanos. Para os demais padrões a interpretação é inversa [Brandão, Tróccoli & Guedea, 2003]

As escolas, em geral, apresentam valores mais altos no indicador relativo ao padrão de direção e gestão administrativa e pedagógica (PR2); aproximadamente iguais em infra-estrutura física, tecnológica e humana (PR1) e apropriação da informática e da cultura educativa (PR3); sucessivamente menores em aperfeiçoamento e expansão do programa (PR4) e na inovação e desenvolvimento social (PR5).

Os NTE apresentam valores mais altos no indicador de aperfeiçoamento e expansão do programa (PR4); aproximadamente iguais em infra-estrutura física, tecnológica e humana (PR1) e apropriação da informática e da cultura educativa (PR3); aproximadamente iguais em direção e gestão administrativa e pedagógica (PR2) e em inovação e desenvolvimento social (PR5).

Os resultados também indicam que o desempenho das escolas, em geral, está mais voltado para direção e gestão administrativa e pedagógica (PR2), enquanto que os NTE apresentam desempenho médio focado no aperfeiçoamento e expansão do programa (PR4). Essas análises, entretanto, não são suficientes para responder se esses valores médios, relativos aos indicadores de desempenho nos padrões, formam grupos homogêneos para escolas e NTE e ainda que fatores podem ser considerados críticos para o sucesso. A segmentação da amostra por agrupamentos semelhantes, no que diz respeito aos seus escores de desempenhos nos cinco padrões, permite investigar relações entre variáveis dentro de um mesmo grupo, possibilitando análises detalhadas e exploratórias da implantação do programa, por segmentos homogêneos, tanto de escolas quanto de NTE.

Este trabalho apresenta uma segmentação empírica obtida com a técnica de mineração por agrupamento, utilizando os valores de desempenho nos cinco padrões como critérios de seleção.

3. A mineração de dados

As técnicas de Descoberta de Conhecimento em Bases de Dados – DCBD (Knowledge Discovery in Databases – KDD), foram desenvolvidas visando identificar informações que estão presentes nas bases de dados, mas não são detectadas pelas técnicas usuais de análise. Consiste de um processo não trivial de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis nos dados. O processo de DCBD compreende vários passos que são repetidos em várias iterações.

Podemos considerar que os passos do processo de Descoberta de Conhecimento em Bases de Dados, prescritos por Fayyad, Piatetsky-Shapiro e Smyth (1996), podem ser agrupados em quatro fases: a) de definição das metas dentro do domínio do problema; b) de preparação dos dados por meio da seleção, limpeza ou pré-processamento e da redução ou projeção dos dados; c) de mineração propriamente dita com as definições do método, do algoritmo e a execução do processo de mineração nos dados; d) de análise, interpretação e consolidação do conhecimento descoberto a partir dos resultados obtidos. Independentemente do tipo de aprendizado envolvido na mineração de dados, esse processo tem como entrada instâncias e atributos e como saída uma descrição de conceito (Witten & Frank, 1999).

Existem vários algoritmos para cada método de mineração tais como classificação, regressão, agrupamento, sumarização, modelagem de dependências e detecção de desvios (Fayyad, Piatetsky-Shapiro & Smyth, 1996). De acordo com o tipo de dado dos seus atributos, as instâncias de um conceito estudado podem ou não serem tratadas por um algoritmo específico.

4. A análise de agrupamento

A análise de agrupamento procura identificar um conjunto finito de categorias para descrever os dados (Fayyad, Piatetsky-Shapiro & Smyth, 1996). As categorias podem ser mutuamente exclusivas ou consistir de ricas representações, como hierarquias ou categorias sobrepostas. A análise classifica objetos no grupo de acordo com um critério de seleção predeterminado. Os grupos resultantes devem exibir alta homogeneidade interna (dentro do grupo) e alta heterogeneidade externa (entre os grupos). Assim, os objetos de um mesmo grupo estarão próximos uns dos outros, quando representados geometricamente, e em grupos diferentes estarão separados (Hair Jr & Black, 2000).

O agrupamento é indicado para se obter uma classificação de grupos naturais (Witten & Frank, 1999). Essa técnica pode ser realizada por diversos algoritmos, sendo que o número de grupos desejados é, quase sempre, um dos parâmetros informados para o algoritmo escolhido. Os grupos resultantes podem ser exclusivos, sobrepostos, probabilísticos ou hierárquicos. Porém, uma das questões delicadas na análise de agrupamento é a determinação do número de grupos formados (Hair Jr & Black, 2000) e, para isso, utilizam-se critérios e diretrizes e não procedimentos objetivos de abordagem do problema.

Uma classe de regra examina a cada passo sucessivo de mineração alguma medida de similaridade, ou distância entre os grupos. O agrupamento final é definido quando a medida de similaridade excede um valor especificado, ou quando os sucessivos valores entre os passos de mineração apresentam um salto repentino. Outra classe de regra utiliza uma medida estatística, tal como a razão de

verossimilhança. O algoritmo EM (*Expectation-Maximization*) apresenta um valor de probabilidade para cada grupo, que reflete sua população relativa. É apresentada a distribuição normal de cada atributo das instâncias que pertencem ao grupo. Com base nas probabilidades das instâncias é calculada a verossimilhança, que mede a homogeneidade do agrupamento. O algoritmo EM, ao iniciar sua execução, estabelece para cada grupo valores iniciais para sua probabilidade e para a distribuição normal de suas instâncias, sendo sensível à ordem de apresentação das instâncias.

5. A mineração nos dados de avaliação do ProInfo

O banco de dados dos resultados da avaliação do ProInfo, segundo o modelo de padrões referenciais, contém atributos de desempenho que são representados pelos cinco indicadores para cada instância de escola e de NTE da amostra. Os cinco indicadores são consistentes para um processo de investigação por agrupamento uma vez que podem ser utilizados como critérios de identificação de grupos. Além disso, os grupos resultantes com esse critério representam escolas e NTE que apresentam características semelhantes quanto ao uso da tecnologia da informação e comunicação na educação. Esses resultados podem subsidiar investigações para uma maior compreensão do programa e, portanto, podem ser considerados como meta do processo de DCBD.

A etapa de preparação, seleção, limpeza e redução dos dados é responsável pela definição dos dados relevantes e pela validação desses dados de modo a evitar equívocos nos resultados. Normalmente, os dados brutos recebidos como entrada dessa etapa apresentam vários problemas que impedem seu uso na mineração de dados, tais como volume excessivo, armazenamento disperso, falta de padronização, valores incorretos ou inexistentes. O tratamento inicial dos dados resultou em uma base de dados com 535 escolas e 250 NTE que foi analisada utilizando a ferramenta WEKA (*Waikato Environment for Knowledge Analysis*, Witten & Frank, 1999).

As escolas e NTE apresentaram desempenhos médios diferentes, o que resultou em tratamentos separados. A ferramenta WEKA disponibiliza diversos algoritmos para classificação, regressão, agrupamento e associação. O algoritmo EM (*Expectation-Maximization*) para agrupamento tem a característica de apresentar uma medida de verossimilhança de cada agrupamento. Além disso, permite que o número de grupos desejados seja ou não informado, tornando-o adequado para o agrupamento de escolas e NTE do ProInfo. Por outro lado, a decisão sobre o número de grupos deve buscar equilíbrio entre dois objetivos: maior homogeneidade e menor número de grupos. A estratégia utilizada foi de testar o algoritmo com vários números de grupos e adotar aquele imediatamente antes de ocorrer um salto no valor que representa perda de verossimilhança (Hair Jr & Black, 2000).

Na primeira execução com a ferramenta WEKA não foi especificado o número de grupos desejado e o algoritmo sugeriu sete grupos para escolas e cinco grupos para os NTE. O valor de verossimilhança não diminuiu significativamente até quatro grupos, no caso das escolas, e também até quatro grupos, no caso dos NTE. No entanto, ao testar o algoritmo com três grupos, tanto no caso de escolas quanto no caso de NTE, o valor de verossimilhança diminuiu significativamente. A Tabela 3 apresenta os valores de verossimilhança encontrados em algumas minerações nos dados de escolas e NTE do ProInfo utilizando o algoritmo EM.

Como pode ser verificado na Tabela 3, há maior perda de homogeneidade quando o número de grupos cai para três. Esse fato nos remete ao agrupamento de tamanho quatro como o agrupamento recomendado tanto para escolas quanto para NTE.

Tabela 3 – Verossimilhança por número de grupos

Escolas		NTE	
Número de grupos	Verossimilhança	Número de grupos	Verossimilhança

10	10,79950	10	11,75774
7	10,69503	7	11,54141
6	10,67631	6	11,48635
5	10,61527	5	11,43135
4	10,50843	4	11,34303
3	8,68012	3	9,22941

Considerando ainda a possível influência da ordem de apresentação das instâncias pelo algoritmo EM, foram realizadas várias execuções do algoritmo para quatro grupos e alterando-se a ordem das instâncias nos arquivos. Nos agrupamentos de escolas não ocorreram alterações significativas de verossimilhança, que variou entre 10,50808 e 10,50843. Nos agrupamentos de NTE, as alterações foram maiores, ocorrendo entre 9,38438 e 11,34303. Entretanto, tanto para escolas como para NTE, foram considerados os agrupamentos de quatro grupos que apresentaram maior verossimilhança.

O algoritmo EM associa, para cada grupo obtido, uma probabilidade que reflete a população relativa do grupo e também representa a probabilidade de uma instância particular pertencer aquele dado grupo. A Tabela 4 apresenta as probabilidades de cada grupo e quantidade de instâncias nos grupos de escolas e NTE, respectivamente.

Tabela 4 – Probabilidade e número de instâncias por grupo

Escolas			NTEs		
Grupo	Probabilidade	Instâncias	Grupo	Probabilidade	Instâncias
E-1	0,3675	199	N-1	0,4385	107
E-2	0,2771	147	N-2	0,2815	73
E-3	0,2149	115	N-3	0,2320	58
E-4	0,1405	74	N-4	0,0480	12

6. Análise dos resultados

A análise de agrupamento das escolas e NTE reflete a situação do ProInfo no estágio de implantação do programa até 2001. Os resultados do processo de mineração de dados por agrupamento nas escolas e NTE evidenciaram, respectivamente, quatro grupos distintos em relação ao critério de desempenho nos padrões (PR1 a PR5).

A Tabela 5 apresenta um resumo do processo de mineração com as médias e os desvios padrão dos grupos de escolas e NTE encontrados.

Tabela 5 – Valor médio e desvio padrão por grupo de Escolas e NTE

Escolas			NTEs		
Grupo	Padrão	Média (desvio padrão)	Grupo	Padrão	Média (desvio padrão)

E-1	PR1	0,1511(0,0562)	N-1	PR1	0,0936(0,0426)
	PR2	0,1997(0,0416)		PR2	0,0788(0,0266)
	PR3	0,1373(0,0329)		PR3	0,1357(0,0388)
	PR4	0,0814(0,0246)		PR4	0,1604(0,0334)
	PR5	0,0526(0,0306)		PR5	0,1188(0,0626)
E-2	PR1	0,1454(0,0658)	N-2	PR1	0,1119(0,0394)
	PR2	0,1364(0,0372)		PR2	0,0640(0,0188)
	PR3	0,0756(0,0294)		PR3	0,0786(0,0297)
	PR4	0,0422(0,0199)		PR4	0,0958(0,0258)
	PR5	0,0313(0,0189)		PR5	0,0491(0,0296)
E-3	PR1	0,1079(0,0724)	N-3	PR1	0,1138(0,0511)
	PR2	0,1432(0,0684)		PR2	0,0580(0,0250)
	PR3	0,0808(0,0520)		PR3	0,0784(0,0344)
	PR4	0,0414(0,0322)		PR4	0,0840(0,0344)
	PR5	0,0000(0,0000)		PR5	0,0000(0,0000)
E-4	PR1	0,1283(0,0549)	N-4	PR1	0,0198(0,0075)
	PR2	0,2500(0,0410)		PR2	0,0105(0,0106)
	PR3	0,2133(0,0362)		PR3	0,0108(0,0074)
	PR4	0,1459(0,0440)		PR4	0,0832(0,0228)
	PR5	0,0766(0,0424)		PR5	0,1330(0,0587)

A Figura 2 apresenta um gráfico de linhas dos valores médios de desempenho nos cinco indicadores (PR1 a PR5) dos quatro grupos de escolas identificados.

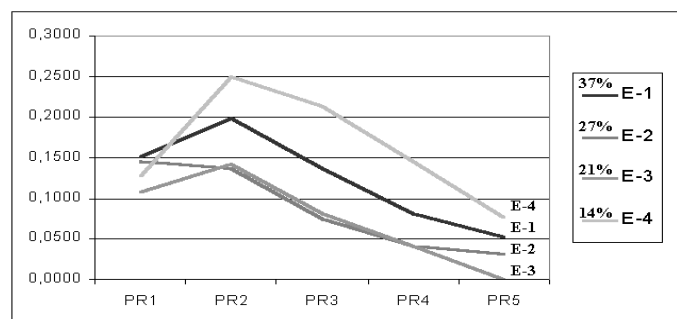


Figura 2 – Valores médios de desempenho nos padrões de acordo com os quatro agrupamentos de escolas

O grupo de escolas mais numeroso, nomeado E-1, agrupa 37% das instâncias. Seu desempenho é mais próximo dos valores médios de desempenho das escolas. O grupo E-2 aglutina 27% das instâncias e é o único grupo que apresenta valor em PR1 superior ao valor em PR2. Como os valores de PR1 tem correlação negativa, podemos supor que essas escolas enfrentam problemas quanto à infra-estrutura. O grupo E-3, com 21% das escolas, apresenta valor constante e nulo em PR5. Esse grupo esteve presente em todas as execuções do processo de mineração, com alta verossimilhança, evidenciando sua estabilidade no agrupamento. O grupo E-4, com 14% das escolas, é o grupo com maior desempenho. Esse grupo apresenta os maiores valores em PR2, PR3, PR4 e PR5, e o segundo valor mais baixo em PR1, o que é positivo uma vez que representa uma correlação negativa com os demais indicadores de desempenho.

De forma geral, podemos concluir que as escolas estão voltadas para as questões relativas aos processos gerenciais e pedagógicos (PR2). Caberia investigar fatores que explicariam os resultados de maior e menor desempenho nos diferentes grupos encontrados. Uma suposição plausível pode ser sobre o

tempo de implantação do programa nas escolas. Poderíamos esperar que escolas mais antigas no ProInfo tivessem superado seus problemas iniciais. Nesse caso, poderíamos esperar que as escolas melhorassem seu desempenho com a progressiva implantação do programa. Observamos, entretanto, que os desempenhos não apresentam valores crescentes nos padrões seguintes ao PR2, o que justificaria uma investigação mais detalhada em cada agrupamento.

A mineração de dados por agrupamento nas instâncias de NTE evidenciou também a existência de quatro grupos distintos. A Figura 3 apresenta um gráfico de linhas dos valores médios de desempenho nos cinco indicadores dos padrões (PR1 a PR5) dos quatro grupos identificados.

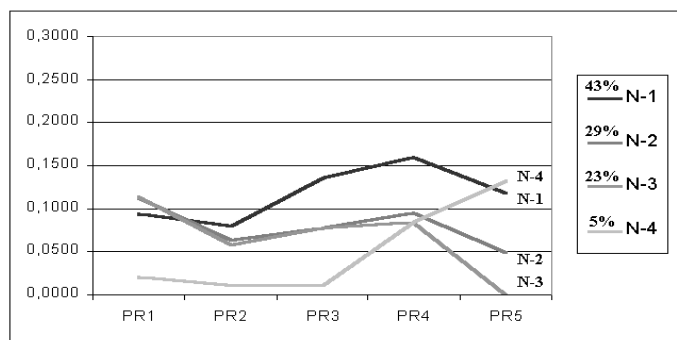


Figura 3 – Valores médios de desempenho nos padrões de acordo com os quatro agrupamentos dos NTE

O grupo N-1 foi o mais populoso, agrupando 43% das instâncias. Os NTE desse grupo apresentaram o melhor desempenho geral, tendo os valores mais altos em PR2, PR3 e PR4 e o segundo melhor desempenho em PR1 e PR5. Os grupos N-2 e N-3, com respectivamente 29% e 23% das instâncias, apresentaram desempenhos semelhantes, sendo que, o que os diferem entre si é o fato do grupo N-3 apresentar valor constante e nulo em PR5. O grupo N-3 mostrou-se estável nos agrupamentos, exceto quando a verossimilhança teve queda significativa no caso do agrupamento de três (Tabela 3). O grupo N-4, com apenas 5% das instâncias, apresentou valores baixos em PR1, PR2 e PR3, próximo ao valor médio dos NTE no padrão PR4 e alto em PR5. Esse grupo, apesar de pouco numeroso, apresentou um comportamento que parece evidenciar que um baixo desempenho em gestão administrativa e planejamento pedagógico não impedem um maior desempenho no aperfeiçoamento e expansão do programa (PR4), como também, na inovação e desenvolvimento social (PR5). Consideramos que o fato dos NTE apresentarem valores mais altos em aperfeiçoamento e expansão do programa (PR4) reflete o fato dos entrevistados serem multiplicadores, e terem sido capacitados pelo programa para exercerem as funções verificadas nas questões relativas ao indicador do padrão PR4.

7. Conclusões

O trabalho investigou a base dos dados gerados a partir das medidas de desempenho obtidas em função do modelo de avaliação de padrões referenciais, visando descobrir agrupamentos de escolas e NTE com desempenhos semelhantes. Foram realizadas as investigações utilizando o método de mineração por agrupamento com a ferramenta WEKA variando-se o número de grupos e, com base na verossimilhança, foram sugeridos como fator de agrupamento, quatro grupos de escolas e também quatro grupos de NTE do ProInfo. A população relativa dos grupos das escolas envolve 37, 27, 21 e 14% destas e dos grupos de NTE envolvem 43, 29, 23 e 5% destes.

Consideramos que esses resultados preliminares podem orientar pesquisas futuras no sentido de investigar relações que podem justificar ou explicar os diferentes desempenhos nos grupos encontrados. Podem, ainda, subsidiar estudos posteriores sobre fatores críticos ou de sucesso da implantação do ProInfo utilizando como blocos experimentais os grupos identificados de escolas e NTE.

Referências

- Brandão, M.F.R.; Tróccoli, B.T. & Guedea, M.T.D. (2003). “Um modelo de avaliação do Programa Nacional de Informática na Educação”. Em: Workshop de Informática na Escola. Anais do XXIII Congresso da SBC/WIE.
- Fayyad, U.; Piatetsky-Shapiro, G. & Smyth, P. (1996). “From Data Mining to Knowledge Discovery in Databases”. AI Magazine: Fall.
- Hair Jr, J.F & Black, W.C. (2000) “Cluster analysis”. In: Laurence G. Grimm & Paul R. Yarnold (Eds.). Reading and Understanding More Multivariate Statistics. Washington: American Psychological Association.
- MEC/SEED/UnB (2002) “ProInfo: perspectivas e desafios – relatório preliminar de avaliação”, Brasília.
- MEC/SEED (2002). “O que é o ProInfo”. <http://www.proinfo.mec.gov.br>, setembro.
- Witten, I.H & Frank, E. (1999). “Data Mining: Practical machine learning tools and techniques with Java implementations”. San Francisco: Morgan Kaufmann Publishers.