Recomendação de grupos heterogêneos utilizando estilos de aprendizagens

Daniel F. dos Santos[1], **Luziane F. de Mendonça**[2], **Marcello G. Teixeira** [2]

[1]PPGI, Programa de Pós-Graduação em Informática, UFRJ
C. P. 68530, Cep 21941-590, Rio de Janeiro, RJ
E-mail: sfdanielrg@gmail.com

[2]DCC, Departamento de Ciência da Computação, UFRJ
C. P. 68530, Cep 21941-590, Rio de Janeiro, RJ
E-mail: luziane@dcc.ufrj.br; marcellogt@dcc.ufrj.br

Abstract. *In this work, clustering heuristics combined with PCA (Principal Component Analysis) technique are proposed in order to create recomendation systems of study groups. The composition of these groups is built in such way to ensure a minimal heterogeneity among the characteristics (learning style) of their members, in order to promote better efficiency in collaborative learning.*

Resumo. *Neste trabalho são propostas heurísticas de agrupamento combinadas com a técnica PCA (Principal Component Analysis) para a criação de um sistema de recomendação de grupos de estudos. A composição de tais grupos é construída de modo a respeitar uma heterogeneidade mínima entre as características (estilos de aprendizagem) de seus integrantes, com o intuito de garantir maior eficiência no objetivo da aprendizagem colaborativa.*

1. INTRODUÇÃO

Os seres humanos têm diferentes estilos de aprendizagem, ou seja, características e preferências quanto à forma de se apropriar das informações, processá-las e construir novos conhecimentos. A competência em uma determinada atividade depende, muitas vezes, da habilidade em dosar esses diferentes estilos. Por exemplo, enquanto há profissionais que são inovadores e absorvem a realidade de uma forma quase aleatória, outros tendem a ser metódicos, observadores e reflexivos. No entanto, desenvolver o equilíbrio entre estilos antagônicos de aprendizagem é uma forma de proporcionar maiores chances de adaptação às situações do dia a dia ou às exigências dos estudos e do trabalho [Felder 1995].

Os critérios adotados para a formação dos grupos influenciam diretamente no resultado da atividade de aprendizagem. Por exemplo, Melser [Melser 1999] mediu a auto-estima de alunos superdotados em grupos homogêneos e os comparou com aqueles que trabalham em grupos heterogêneos. A avaliação da auto-estima dos dois grupos difere significativamente quando comparados. Enquanto os alunos superdotados trabalhando em grupo heterogêneo tiveram um aumento da auto-estima, os alunos superdotados que trabalharam em um grupo homogêneo tiveram uma diminuição da mesma.

O melhoramento da auto-estima pode ser verificado na formação de outros grupos de aprendizagem, não somente com superdotados. Além disso, um grupo de estudo heterogêneo propicia uma maior troca de informações, gerando ganho de conhecimento por

parte dos seus membros e resultados mais criativos e completos ao final da atividade de aprendizagem (visão mais ampla sobre o tema abordado).

2. Agrupamento

A análise de agrupamento é uma ferramenta útil para a análise de dados em muitas situações diferentes. Esta técnica pode ser usada para reduzir a dimensão de um conjunto de dados, reduzindo uma ampla gama de objetos à informação do centro do seu conjunto. Tendo em vista que agrupamento, ou *clustering*, é uma técnica de aprendizado não supervisionado (quando o aprendizado é supervisionado, o processo é denominado de classificação), pode servir também para extrair características escondidas dos dados e desenvolver as hipóteses a respeito de sua natureza.

A análise de agrupamento é o nome dado para o grupo de técnicas computacionais cujo propósito consiste em separar n objetos em grupos, baseando-se nas p características que estes objetos possuem [Everitt 1974].

O critério baseia-se normalmente em uma função de dissimilaridade (métrica), função esta que recebe dois objetos e retorna a discrepância entre eles. Podemos citar como exemplos funções baseadas na norma Euclidiana, distância de Mahalanobis [Mahalanobis 1936], coeficiênte de Gower [Gower 1986], coeficiente de similaridade de Cattel [Johnson], coeficiente de Camberra [Bussab 1990], etc..

Nas técnicas mais tradicionais, os grupos determinados por uma métrica de qualidade devem apresentar alta homogeneidade interna e alto índice de separação (heterogeneidade externa). Isto quer dizer que os elementos de um determinado conjunto devem ser mutuamente similares e, preferencialmente, muito diferentes dos elementos de outros conjuntos.

Por exemplo, uma técnica bastante conhecida é o $k - means$ [Pimentel 2003]. O *K-means* é uma técnica que usa o algoritmo de agrupamento de dados por K-médias (*K-means clustering*); ou seja, é uma heurística de agrupamento não hierárquico que busca minimizar a distância dos elementos a um conjunto de k centros de forma iterativa. A distância entre um ponto p_i e um conjunto de *clusters* é definida como sendo a distância do ponto ao centro mais próximo dele.

Existem muitos trabalhos recentes na literatura especializada que versam sobre as várias técnicas de agrupamento com a finalizadade de aprendizagem [Wolf 2003]. Entretanto, há consenso de que bons resultados no processo de aprendizagem surgem com grande frequência em indivíduos pertencentes a agrupamentos que levam em consideração a heterogeneidade das características que podem ser aferidas de cada membro (área de formação, experiência prévia no assunto abordado, capacidade de liderança e trabalho em grupo, entre outros)[Wolf 2003].

As técnicas tradicionais de agrupamento não se mostram adequadas para o problema em estudo, uma vez que este necessita formar grupos de elementos heterogêneos. Uma possibilidade seria executar uma etapa inicial, formando grupos homogêneos com quantidade de membros igual ao número de elementos que desejamos; posteriormente, um elemento de cada grupo homogêneos poderia ser "sorteado" formando assim um grupo heterogêneo; esse procedimento de sorteio poderia se repetir até que todos os elementos tivesse sido selecionados. Entretanto, embora o procedimento descrito forme

agrupamentos heterogêneos, não há possibilidade de garantir que tais grupos sejam homogêneos entre si, a menos que sejam utilizadas estratégias adicionais durante esse "sorteio".

Neste trabalho são apresentadas heurísticas para a realização de agrupamentos de indivíduos, priorizando a heterogeneidade de seus componentes.

As características coletadas de cada indivíduo[1] permitirão sua representação com um vetor do $I\!R^n$, fazendo com que todo o universo de alunos possa ser representado como uma nuvem de pontos (subconjunto discreto) do $I\!R^n$.

As heurísticas de agrupamento serão desenvolvidas tomando por base a representação desses pontos em uma base ortonormal obtida por meio da técnica PCA (*Principal Component Analysis*). Os vetores geradores dessa base são determinados pelos autovetores da matriz de covariância dos pontos originais, conforme será detalhado na seção a seguir.

3. Algoritmo Proposto

3.1. PCA

O PCA foi originalmente descrito por *Karl Pearson* [Pearson 1901], sendo posteriormente consolidado por *Hotelling* [Hotelling 1933] com o propósito particular de analisar estruturas de correlações. Aproximadamente 30 anos mais tarde (anos 60), esse tipo de análise foi introduzido na Química por *Malinowski* [Malinowski]. Como visto, a abordagem PCA não é nova, mas somente com a difusão nas últimas décadas do uso massivo de computadores está ocorrendo um rápido crescimento de sua utilização, como por exemplo o uso dessa técnica para auxiliar o reconhecimento facial em imagens [Turk 1991].

A abordagem PCA é uma técnica estatística que permite identificar padrões em um conjunto de dados e expressar esse conjunto de dados de forma a salientar suas similaridades e diferenças utilizando uma transformação ortogonal para alterar a base em que os mesmos estão representados. O objetivo do PCA é rotacionar rigidamente os eixos do espaço original $n-$dimensional para novas posições de maior variância (componentes principais)[2]

Os componentes principais podem ser obtidos como os autovetores da matriz de coraviância associada às características de todos os indivíduos[3].

[1]Entre os vários modelos de estilos de aprendizagem apresentados em publicações sobre psicologia educacional [Felder 1995], podemos destacar o de Felder e Silverman [Felder 1988], que classifica os aprendizes em cinco dimensões: ativos/reflexivos; sensoriais/intuitivos; visuais/verbais; indutivos/dedutivos; seqüenciais/globais. A partir deste modelo, Felder e Soloman [Felder 1991] desenvolveram um instrumento denominado Índice de Estilos de Aprendizagem (*Index of Learning Styles - ILS*), que classifica os estudantes em quatro das dimensões acima citadas.

[2]Nesse sentido, a dimensão dos dados pode ser reduzida utilizando os componentes principais mais representativos, o que permite o uso de tal técnica na compressão de dados.

[3]O PCA utiliza um conjunto de dados representados por uma matriz de m indivíduos e n características que podem estar correlacionados. A correlação entre os dados pode ser aferida utilizando a matriz de covariância dos mesmos. Os vetores ortonormais que apontam para as direções de maior variação da massa de dados correspondem aos autovetores da matriz de covariância, e indicam as direções, a partir do ponto médio do conjunto de dados, nas quais há maior discrêpância entre os mesmos.

3.2. Formação dos Grupos

Após o uso da técnica PCA para representar as características de todos os indivíduos em uma nova base ortonormal, os algoritmos estudados neste trabalho tem por objetivo realizar agrupamentos heterogêneos.

A medida de heterogeneidade proposta é a distância de Mahalanobis [Bussab 1990]; ela é uma métrica que difere da distância Euclidiana por levar em consideração a correlação entre os conjutos de dados. A fórmula para distância de Mahalanobis entre dois vetores $\vec{x}$ e $\vec{y}$ da mesma distribuição que possuam uma matriz de covariância Σ é dada por:

$$d(\vec{x}, \vec{y}) = \sqrt{(\vec{x} - \vec{y})^T \Sigma^{-1} (\vec{x} - \vec{y})}.$$

Se a matriz de covariância é a matriz identidade, esta fórmula se reduz à distância euclidiana. Caso contrário, ela leva em consideração a correlação entre as características. A distância de Mahalanobis consegue portanto capturar melhor a relação de correção entre os dados.

A ideia mais intuitiva para realizar o agrupamento seria selecionar para formar o primeiro grupo os indivíduos que, dois-a-dois, tenham os maiores valores para a medida de dissimilaridade. Entretanto, ao selecionar sempre os indivíduos de maior distância para formar os grupos, invariavelmente o último grupo formado será homogêneo. Ou seja, o uso dessa ideia maximiza a heterogeneidade do primeiro grupo, mas ao finalizar a formação dos grupos, essa forma de selecionar converge para formação de grupos homogêneos, o que não é desejado neste trabalho.

Dessa forma, apresentamos a seguir o primeiro algoritmo para agrupamento heterogêneo proposto. A idéia principal consiste em calcular, para o conjunto total de elementos ainda não agrupados, as características principais (eixos ortonormais de maior variação) via técnica PCA. Então, escolhe-se para formar um grupo os indivíduos que possuem suas características mais próximas às características principais.

Para tanto, calcula-se os vetores diretores dos dados partindo do ponto centróide a cada um dos pontos que representam os indivíduos e armazena-se os ângulos entre cada vetor e os eixos ortonormais. Descobrindo o ângulo mais próximo de 0^0 e o mais próximo de 180^0, os pontos correspondentes são selecionados para formar o novo grupo.

Ou seja, diferentemente da ideia mais intuitiva, a discrepância entre os elementos de um mesmo grupo será determinada pela proximidade dos mesmos aos eixos ortonormais. Cabe observar que esse algoritmo irá permitir o agrupamento de no máximo $2p$ (número de semi-eixos ortonormais) indivíduos por grupo, sendo p o número de características de cada indivíduo.

Algoritmo 1.1 - Considere que NA é o número de alunos, NC é o número de características, e $X = \{x_1, ..., x_{NA}\}$ o conjunto dos vetores de características.
Passo 1 - Obtenha a matriz de covariância C dos dados armazenados em X, e calcule os seus autovetores $\mu_i, ..., \mu_{NC}$.

Passo 1.1 - Calcule $y_i = x_i - \mu, \ \ \forall i = 1, ..., NA$, onde $\mu = \dfrac{\left(\sum_{i=1}^{NA} x_i\right)}{NA}$.

Passo 2 - Para todo $i = 1, ..., NA$, $j = 1, ..., NA$ calcule os ângulos dos vetores

$$\theta_{ij} = \arccos\left(\frac{y_i^T \mu_j}{\|y_i\|.\|\mu_j\|}\right)$$

Passo 3 - Para todo $j = 1, ..., NC$, obtenha o vetor $\vec{y_i}$ que forma o menor (analogamente, o maior) ângulo com μ_j, e insira o vetor associado no grupo:

$$grupo = grupo \cup \{\vec{x_i}\}.$$

Passo 4 - Atualização dos pontos restantes, fazendo $X = X - grupo$. Se $X = \emptyset$, pare a execução. Caso contrário, retorne para Passo 1.

Na Figura 1, é apresentado o resultado da execução do algoritmo para um conjunto de 500 pontos, obtendo 3 grupos cujos elementos são marcados com cores diferentes. Podemos ver, em iterações distintas (1a, 30a, 60a, 90a, 120a e última iteração), a seleção de quatro indivíduos assinalados em preto que estão mais próximos dos eixos em azul (características principais), em relação aos ângulos; tais indivíduos irão compor um grupo.

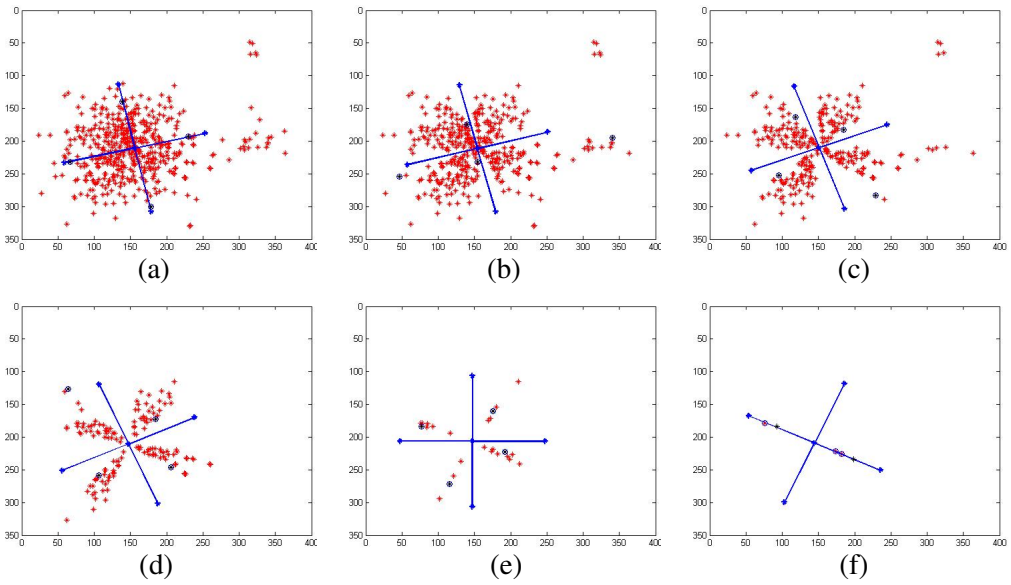

Figura 1. Algoritmo proposto - formação dos grupos: (a) 1o.; (b) 30o.; (c) 60o.; (d) 90o.; (e) 120o.; (f) 125o.

4. CONCLUSÃO

As técnicas de agrupamento presentes na literatura promovem o surgimento de grupo de indivíduos homogêneos entre si; a heterogeneidade é levada em consideração apenas entre os grupos. Entretanto, o problema de estudo deste trabalho requer a formação de grupos cujos indivíduos sejam heterogêneos entre si; quanto à comparação entre os grupos, nada impede que haja similaridade.

Diante da dificuldade de resolver o problema abordado utilizando as técnicas já conhecidas, neste trabalho temos o intuito de apresentar e analizar o comportamento de algoritmos desenhados para este fim. O algoritmo aqui apresentado faz uso do tamanho do

ângulo entre os ponto-centróide e os eixos ortonormais (de maior variância) como medida de seleção dos indivíduos. Essa estratégia assegura a heterogeneidade do início ao fim da seleção dos grupos mantendo uma diferença mínima entre seus elementos, segundo os testes numéricos realizados com caracteríticas dos indivíduos geradas aleatoriamente.

Tais resultados permitem, na sequência da pesquisa, analisar a eficiência dos grupos recomendados em atividades práticas de aprendizagem, assim como a proposição de novos algoritmos e a análise dos mesmos quanto às especificidades dos grupos formados e sua adequação aos vários cenários cabíveis.

Referências

Bussab, W. O., M. E. S. e. A. D. F. (1990). *Introdução à Analise de Agrupamentos*. IME-USP.

Everitt, B. (1974). *Cluster Analisys*. Heinemann Educational Books.

Felder, R. M. e Soloman, B. A. (1991). Index of learning styles questionnaire. *North Carolina State University*.

Felder, R. M. e Henriques, E. M. (1995). Learning and teaching styles in foreign and second language education. 28:21–31.

Felder, R. M. e Silverman, L. K. (1988). Learning and teaching styles in engineering education,. *Engr. Education*, 78(7):674–681.

Gower, J. C. e Legendre, P. (1986). Metric and euclidian properties of dissimilarity coefficients. *Journal of Classifications*, 3:5 – 48.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal component. *Educ. Psychol*, 24:417–441.

Johnson, R. A. e Wichern, D. W. *Applied multivariate statistical analysis*. 4th edition edition.

Mahalanobis, P. C. (1936). On the generalised distância in statistics. *Proceedings of the National Institute of Sciences of India*, 2:49 – 55.

Malinowski, E. R. *Factor Analysis in Chemistry, 3rd ed.* Wiley.

Melser, A. N. (1999). *Gifted students and cooperative learning: A study of grouping strategies.* Roeper Rev.

Pearson, K. (1901). On lines and planes of closest fitto systems of points in space. *Phil. Mag*, 6:559–72.

Pimentel, E. P., F. V. F. e. O. N. (2003). A identificação de grupos de aprendizer no ensino presencial utilizando técnicas de clusterização. *Anais do Simpósio Brasileiro de Informática na Educação*.

Turk, M. e Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3:71–86.

Wolf, S. (2003). Interações sociais em grupos homogêneos e heterogêneos em relação à formação profissional. Master's thesis, Universidade Federal de Santa Catarina.