RBRH – Revista Brasileira de Recursos Hídricos Volume 17 n.1 - Jan/Mar 2012, 65-75

Uso de Técnicas de Mineração de Dados na Identificação de Áreas Hidrologicamente Homogêneas no Estado da Paraíba

Roberta Brito Nunes Diniz, Valéria Gonçalves Soares, Lucídio dos Anjos F. Cabral
Universidade Federal da Paraíba
roberta.diniz@terra.com.br, valeria@di.ufpb.br, lucidio@di.ufpb.br

Recebido: 19/04/10 - revisado: 16/05/11 - aceito: 26/09/11

RESUMO

A escassez de dados fluviométricos e a má qualidade dos dados existentes sobre os cursos d'água do Nordeste do Brasil têm obrigado os especialistas em hidrologia a buscar novos caminhos, ampliando assim os conhecimentos e metodologias sobre como desenvolver a região com base em suas limitações ambientais. Uma dessas metodologias consiste na utilização de técnicas de regionalização hidrológica, que possibilitam a transferência de dados e informações entre bacias com características similares. Nesse contexto, este trabalho visa identificar regiões hidrologicamente homogêneas no Estado da Paraíba, utilizando Mineração de Dados, através da técnica de Clusterização, possibilitando assim a identificação de padrões que permitam a transposição de dados de uma região para outra. Foram utilizados algoritmos com métodos baseados em partição, métodos hierárquicos, e métodos baseados em redes neurais, e aplicados índices de validação estatística nos agrupamentos gerados. De acordo com os resultados obtidos, o algoritmo Ward apresentou o melhor resultado em todos os índices de validação aplicados com a identificação de seis regiões hidrologicamente homogêneas no Estado da Paraíba.

Palavras-Chave: *mineração de dados; regionalização hidrológica; clusterização.*

INTRODUÇÃO

A escassez de dados para realizar estudos hidrológicos é uma realidade no nosso país. As comunidades acadêmica e profissional da área de hidrologia têm buscado, através de pesquisas, ampliar os conhecimentos e metodologias para minimizar as conseqüências dessa deficiência. Uma dessas metodologias consiste na utilização da técnica de regionalização hidrológica, a qual possibilita a transposição dos dados e informações entre bacias de características similares (Porto et al. 2004).

O desenvolvimento de métodos eficientes de regionalização hidrológica se torna cada vez mais necessário no Brasil, devido à implantação dos sistemas estaduais e federal de gerenciamento de recursos hídricos e da escassez de dados obtidos através de medições em campo. Dentre tais sistemas, destacam-se os subsistemas de outorga e cobrança pelo uso da água bruta, o licenciamento ambiental e a elaboração dos planos diretores de bacias, que necessitam estimar curvas de permanência de vazões, na maioria das vezes não disponíveis (Diniz 2008).

Uma nova área da Tecnologia da Informação – Descoberta de Conhecimento em Base de Dados (*Knowledge Discovery in Databases*) – surge como alternativa para atender, entre outras, essa necessidade. Esse processo permite analisar e utilizar de maneira útil os dados disponíveis para obter a informação desejada, a partir do uso de ferramentas de extração de conhecimento (Fayyad et al. 1996).

A mineração de dados, uma das etapas da descoberta de conhecimento em base de dados, tem avançado nestes últimos anos e tem despertado grande interesse junto às comunidades científica e industrial. Várias técnicas de mineração de dados aumentam as possibilidades de se desenvolver métodos mais adequados para os estudos de regionalização hidrológica, fornecendo, ao especialista do domínio, ferramentas que possibilitem uma maior segurança no processo de tomada de decisão (Rao et al. 2006; Porto et al. 2006).

Estudos desenvolvidos por pesquisadores mostraram que a análise de agrupamento (Clusterização) tem sido usada com sucesso na definição de regiões hidrologicamente homogêneas (Porto et al. 2004). A clusterização busca dividir um conjunto de objetos não rotulados em grupos, chamados de partição ou *cluster*, de maneira que os objetos pertencentes a um mesmo grupo apresentam alta similaridade entre si e baixa similaridade em relação aos objetos dos demais grupos. Esta técnica vem sendo

utilizada em tarefas de exploração de dados e extrações de padrões (Metz 2006).

Neste artigo será apresentada uma proposta de metodologia para a utilização de técnicas de mineração de dados na identificação de áreas hidrologicamente homogêneas no Estado da Paraíba.

O restante do artigo está organizado da seguinte maneira: a seção 2 apresenta algumas pesquisas desenvolvidas na área de regionalização hidrológica usando a técnica de clusterização; a seção 3 apresenta o estudo de caso, descrevendo o cenário onde o trabalho será desenvolvido e os dados coletados para a pesquisa; a seção 4 apresenta a metodologia proposta, descrevendo as etapas seguidas pela tarefa de clusterização, com o objetivo de identificar regiões hidrologicamente homogêneas; e a seção 5 apresenta as conclusões obtidas com os resultados apresentados pelos algoritmos de clusterização.

TRABALHOS RELACIONADOS

A carência de informação, um problema sempre presente nos estudos hidrológicos, tem motivado várias pesquisas na área de regionalização hidrológica.

Júnior et al. (2006) realizaram estudos para determinar regiões homogêneas quanto à distribuição de freqüência de chuvas no leste do Estado de Minas Gerais. Foram utilizados métodos estatísticos de análise multivariada (componentes principais rotacionados) e os algoritmos hierárquicos Single-Linkage, Complete-Linkage e Ward. Os estudos concluíram que o método Ward apresentou a melhor representação espacial das regiões hidroclimaticamente homogêneas caracterizadas pelos grupos, em relação aos demais métodos empregados.

Rao e Srinivas (2006) investigaram o potencial do método de cluster híbrido na regionalização de bacias para análise de freqüência de enchentes. Três algoritmos de cluster híbrido, que são uma combinação de algoritmos hierárquicos aglomerativo e algoritmo particional, foram utilizados com o objetivo de definir regiões homogêneas no estado de Indiana, USA. Os algoritmos hierárquicos utilizados foram o Single-Linkage, Complete-Linkage e Ward e o algoritmo particional, o K-means. Os resultados obtidos demonstraram que o desempenho geral do modelo híbrido, na otimização da função objetiva, foi melhor que os desempenhos dos algoritmos de clustering hierárquico e particional usados separadamente. Dos três modelos híbridos apresentados, a combinação dos algoritmos Ward e do K-means resultou numa boa estimativa de grupos de bacias sendo recomendado para regionalizar bacias hidrográficas.

Porto et al. (2004) aplicaram a técnica de clusterização para identificar sub-bacias com características físicas similares no estado do Ceará. Foi utilizada a abordagem hierárquica utilizando o método Ward. Os resultados demonstraram que a técnica de clusterização poderá ser de grande aplicabilidade na identificação de regiões hidrologicamente homogêneas subsidiando assim estudos de regionalização de dados hidrológicos.

ESTUDO DE CASO

A metodologia proposta é aplicada sobre os dados de 41 bacias do Estado da Paraíba. A motivação para a escolha desse cenário consiste no fato dos dados hidrológicos serem escassos, em algumas regiões do estado, dificultando assim o dimensionamento de obras hidráulicas, o dimensionamento do volume de reservatórios e o planejamento dos recursos hídricos, entre outros estudos.

Dados Hidrológicos

As características físicas e climatológicas de uma bacia são elementos importantes para descrever o seu comportamento hidrológico. As características que serão utilizadas no presente trabalho foram obtidas através de mapa digitalizado atualizado da hidrografia do Estado da Paraíba (TC/BR 2001) e do Modelo Digital de Elevação (MDE) gerado para todo o Estado através do software ArcView GIS do ESRI (Environmental Systems Research Institute). O Modelo Digital de Elevação (MDE) foi obtido a partir de pontos cotados, georeferenciados e com os valores de altitudes associados, oriundos da SRTM (Shuttle Radar Topography Mission – projeto internacional gerenciado pela agência nacional de Inteligência Geo-Espacial (NGA) e pela NASA, iniciado em 2000), cobrindo quase todo o globo com uma malha de 90 m x 90 m (Rabus et al. 2003). Estão apresentadas na Tabela 1, 12 das 32 características obtidas

Além das 12 características mencionadas na Tabela 1, também foram obtidas as seguintes características: linha de fundo (L), comprimento do curso d'água (Lt), Comprimento da rede de drenagem (Ld), largura média (Lm), índice de compacidade (Kc), ordem dos cursos de água (Or), índice de bifurcação (Rb), Índice das áreas (Ra), Coefic. de

Tabela 1 - Características da bacia hidrográfica.

Características das medidas lineares da bacia				
Parâmetro	Descrição	Unid	Valores	
			Min	Max
Área (A)	Área da bacia hidrográfica	km2	9,5	17.220
Perímetro (P)	Perímetro da bacia hidrográfica	km	11,4	892,30
Características da forma da bacia				
Índice de circularidade (Ke)	Relação entre a área de um círculo que tem o perímetro igual ao da bacia e a área da bacia.	adm	0,227	0,90
Fator de forma (Kf)	Relação entre a largura média e a linha de fundo da bacia.	adm	0,092	1,10
Características da rede de drenagem				
Índice dos comprimentos (RL)	Média geométrica das relações entre os cumprimentos médios dos talvegues de duas classes consecutivas.	adm	0,115	129,0
Densidade de drenagem (Dd)	Quociente entre o comprimento total da rede de drenagem e a área da bacia hidrográfica.	km/ km2	0,374	1,6
Características do relevo da bacia				
Declividade máxima (Imax)	Declividade máxima da bacia hidrográfica	m	0,018	85,1
Elevação média da bacia (Cmed)	Quociente entre a soma das elevações das sub-áreas entre curvas de nível e a área da bacia.	m/m	35,151	283,2
Características da capacidade de escoamento da bacia				
Lâmina Média Anual (L600)	Lâmina Média Anual escoada em uma bacia hidrográfica localizada na zona climática Sertão com precipitação anual média de 600 mm.	mm	22,20	73,6
Área do Espelho (AE)	Percentagem da área da bacia coberta por espelho de água.	adm	0,001	0,1
Características climatológicas da bacia				
Precipitação média (P)	Chuva média precipitada na bacia , determinada por interpolação entre vários postos da região.	mm	343,8	1.324,1
Evapo-transpiração média (E)	Soma entre a evaporação das superfícies e a transpiração da vegetação.	mm	1.580	1.968

equivalente (lr), Índice de declividade média da bacia (Ip), índice de declividade global (IG), desnível específico (DS), percentual do solo tipo 1 (SOLO1), percentual do solo tipo 2 (SOLO2), percentual do solo tipo 3 (SOLO3).

torrencialidade (Ct), índice de rugosidade (IR), extensão média do escoamento superficial (Le), sinuosidade do curso d´água (SIN), lado maior do retângulo equivalente (Lr), lado menor do retângulo.

METODOLOGIA PROPOSTA

A metodologia do trabalho é ilustrada na Figura 1. Os processos apresentados estão contidos nas principais etapas do processo de descoberta do conhecimento: pré-processamento, mineração de dados e pós-processamento.

A primeira etapa compreende desde a correção de dados inconsistentes até o ajuste da formatação dos mesmos para o algoritmo de mineração de dados a ser utilizado. Além disso, métodos de seleção de atributos são empregados com o objetivo de selecionar apenas os atributos que melhor descrevem e discriminam o conjunto de dados e suas estruturas latentes, o que conseqüentemente reduz a dimensionalidade dos dados, melhora a eficiência dos algoritmos em relação ao tempo de execução, e, em alguns casos, pode melhorar os resultados obtidos (Metz 2006).

Na segunda etapa são executados e avaliados alguns algoritmos de clusterização com o objetivo de identificar qual deles adéqua-se melhor aos estudos de regionalização hidrológica.

A terceira etapa determina o grau de significância dos resultados obtidos pelo algoritmo de clusterização.

No presente artigo são mostrados os resultados obtidos pelos algoritmos hierárquicos, *Single-Linkage*, *Complete-Linkage* e *Ward*, o particional *K-Means* e a Rede Neural de *Kohonen*.

O algoritmo *Single-Linkage* é um dos algoritmos de clustering hierárquico aglomerativo mais simples. Esse método utiliza a técnica do vizinho mais próximo (Nearest Neighbor Technique), na qual a distância entre dois grupos é determinada pela distância do par de exemplos mais próximo, sendo cada exemplo pertencente a um desses grupos(Larose 2005). A distância Euclidiana é usualmente utilizada para obter a matriz das distâncias dos elementos de dados a serem agrupados.
O algoritmo *Complete-Linkage* utiliza a técnica conhecida como Farthest Neighbor ou vizinho mais distante. Ao contrário do algoritmo Single-Linkage, esse algoritmo determina a distância entre dois grupos de acordo com a maior distância entre um par de exemplos, sendo cada exemplo pertencente a um grupo distinto (Metz 2006).

O método de Ward é um método de agrupamento onde os grupos de dados são formados em etapas e são sistematicamente reduzidos (n,n-1,n-2,...), considerando a união de todos os n(n-1)/2 possíveis pares e selecionando a união que tem um valor máximo para a função objetivo (Ward 1963).

O algoritmo K-Means é largamente utilizado na tarefa de clusterização e busca encontrar o melhor particionamento dos n objetos em k grupos (Larose 2005).

A rede neural de Kohonen é um tipo de rede neural artificial baseada em aprendizado competitivo e não supervisionado, capaz de mapear um conjunto de dados de entrada, em um conjunto finito de neurônios organizados numa grade regular de baixa dimensão, geralmente unidimensional ou bidimensional.

A organização dos neurônios fornece um mapa topográfico dos dados de entrada no qual as localizações espaciais (coordenadas) dos neurônios na grade são indicativas das características estatísticas intrínsecas aos dados de entrada (Haykin 2001).
Os números de grupos analisados no processo de clusterização foram 2, 3, 4 e 6 grupos.

Foram realizadas validações estatísticas nos grupos gerados utilizando o índice de Silhouette (Rao et al. 2006), índice Davies-Bouldin (Davies 1979) e índice Dunn (Rao et al. 2006). O índice de Silhouette é utilizado para determinar o melhor número de grupos a ser considerado, já que diversos agrupamentos foram obtidos. Os índices Davies-Bouldin e Dunn baseiam-se na idéia de identificar grupos compactos e bem separados.

Para a execução dos experimentos foi utilizado o software MATLAB versão 7.0 e para a validação estatística foi utilizado o programa Machaon CVE, (Clusteringand Validation Environment), que avalia a qualidade dos agrupamentos obtidos através de diferentes índices estatísticos (Bolshakova 2009).

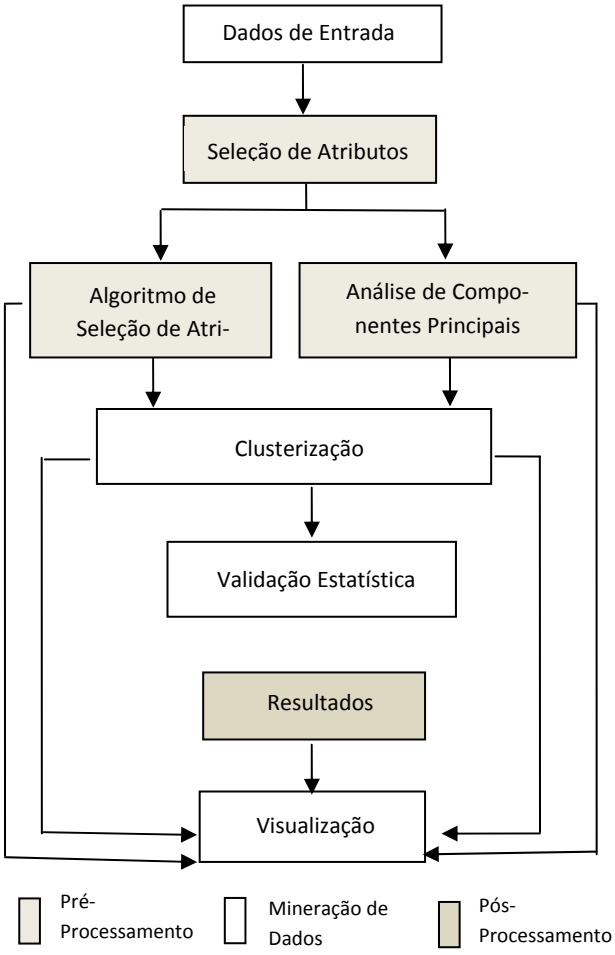

Figura 1 - Fluxograma da metodologia proposta e aplicada.

Etapa do Processo de Descoberta do Conhecimento

Pré-Processamento

Os dados utilizados foram primeiramente normalizados em virtude dos mesmos apresentarem escalas diferentes e, posteriormente submetidos a métodos de seleção dos atributos com o objetivo de

selecionar aqueles que melhor descrevem e discriminam o conjunto de dados.

Utilizou-se os seguintes métodos de seleção de atributos:

- **Algoritmo de seleção de atributos não supervisionado**

Nesta etapa foi utilizado o algoritmo proposto por Mitra et al. (2002), sendo o mesmo executado para vários valores de k (de 1 a 31), onde k representa o número de vizinhos mais próximos a serem considerados, já que o particionamento dos atributos realizado por esse algoritmo é baseado no algoritmo k-NN (Larose 2005) . Mitra et al (2002) descrevem o algoritmo nos seguintes passos:

- **Passo 1**: Escolher um valor inicial de k ≤ D – 1. Inicializar o conjunto reduzido de atributos R com o conjunto original de atributos R ← O.
- **Passo 2**: Para cada atributo $F_i \in R$, calcular r_i^k .
- **Passo 3**: Encontrar o atributo F_i para o qual r_i^k é mínimo. Reter este atributo em R e descartar os k atributos mais próximos de Fi´.
- (Nota: Fi´ representa o atributo para qual removendo os k vizinhos mais próximos irá causar o menor erro entre todos os atributos em R). Fazer Ɛ = r_i^k .
- **Passo** 4: Se k > cardinalidade (R) – 1: k = cardinalidade (R) – 1.
- **Passo 5**: Se k = 1: Ir para o passo 8.
- **Passo 6**: Enquanto $r_i^k > Ɛ$ faça: (a) k = k – 1.

Como o regime hidrológico de uma bacia hidrográfica é grandemente influenciado pelas suas características físicas, o critério de escolha do valor de k foi que os atributos escolhidos na iteração caracterizassem ao máximo a bacia hidrográfica.

Os valores de k escolhidos foram 16 e 18. O valor de k = 16 foi escolhido em virtude de nessa iteração serem selecionados atributos que representam uma maior diversidade de características das bacias hidrográficas, ficando as características de forma, rede de drenagem e relevo com várias representações. O valor de k = 18 foi escolhido pelo mesmo motivo, com o diferencial de que o número de atributos, que representam as características, ficou menor e as características linear, forma e relevo ficaram com apenas uma apresentação.

Com os resultados obtidos pela execução do algoritmo acima citado, foram criados os cenários I e II, descritos na Tabela 2.

- **Componentes Principais**

Os 32 atributos das bacias hidrográficas foram submetidos à análise de componentes principais A cada iteração foi feito um exame do ajuste de cada atributo, através da matriz anti-imagem de correlação, e os que apresentavam valores inferiores a 0,5 foram sendo eliminados. Ao final do processo, restaram 23 atributos. Os 23 componentes principais foram calculados e, após a análise dos autovalores, optou-se por selecionar os que apresentaram autovalores maiores que 1 e que explicam juntos 84,50% dos atributos originais. Como cada componente principal é uma combinação linear dos atributos originais, ele ganha um novo significado. A Tabela 3 resume a descrição dos cinco componentes principais.

Com os componentes principais gerados foi criado o cenário III, conforme Tabela 2.

Tabela 2 - Cenários para execução dos algoritmos de Clusterização

Cenário	Atributos	Nº bacias
I	A,Kc,Ke,Kf,Or,Rb, Rl,Ct,Ip,L600	41
II	A,Ke,Ra,IR,Imax	41
III	C1,C2,C3,C4,C5	41
IV	Todos os 32 atributos	41

Um quarto cenário foi formado com o conjunto completo de atributos das bacias hidrográficas, conforme Tabela 2.

Tabela 3 - Componentes Principais

Componentes Principais	**Significado**	**Variáveis Explicativas**
C1	Linear	A, Pr, L, Lt, Ld, Lm, Le e Lr
C2	Forma	Ke, Kc, IR
C3	Escoamento	SOLO1, L600
C4	Relevo	DS
C5	Drenagem	Dd

Clusterização

Todos os algoritmos foram executados utilizando os cenários descritos anteriormente e os índices de validação correspondentes calculados sobre os resultados obtidos.

Algoritmos Hierárquicos

Os resultados dos algoritmos hierárquicos são representados pelos dendogramas que são estruturas em formato de árvore, nas quais os elementos são dispostos no eixo horizontal, e a distância (ou a similaridade) com que os grupos são gerados, no eixo vertical. Desse modo, o dendograma é uma estrutura com toda a hierarquia dos agrupamentos gerados sobre o conjunto de elementos inicial.

A medida adotada para o cálculo da similaridade foi a distância Euclidiana. O coeficiente de correlação cofenético foi calculado para cada algoritmo hierárquico, em cada cenário (Tabela 4). O coeficiente de correlação cofenético é uma medida de validade para algoritmos de agrupamento hierárquico. Ele é usado para medir quão bem a estrutura hierárquica do dendograma representa os relacionamentos existentes nos dados de entrada (Rao et al. 2006). Mas conforme pode-se observar nos resultados apresentados na Tabela 4, o algoritmo que obteve o melhor valor para esse coeficiente não representou uma boa estrutura de grupos formados, confirmando assim as afirmações apresentadas na literatura, (Metz 2006); (Rao et al. 2006) de que a análise do coeficiente de correlação cofenético não é suficiente para identificar os melhores dendogramas nos modelos hierárquicos, sendo ainda necessária a inspeção visual dos mesmos.

Tabela 4 - Coeficiente de Correlação Cofenético.

	Cenário I		
Algoritmo	Single Linkage	Complete Linkage	Ward
Coeficiente Cofenético	0,71	0,57	0,54
	Cenário II		
	0,78	0,55	0,54
	Cenário III		
	0,70	0,51	0,54
	Cenário IV		
	0,71	0,52	0,57

- ***Single-Linkage***

Apesar de apresentarem os melhores valores do coeficiente de correlação cofenético, como mostra a Tabela 4, os dendogramas gerados por esse algoritmo demonstraram que não houve a formação de grupos bem definidos e separados em todos os quatro cenários, conforme ilustra a Figura 2. Nessa apresentação não há diferenças significativas no tamanho dos arcos que representam os grupos, o que dificulta a escolha da altura para o corte do dendograma.

- ***Complete-Linkage***

Os dendogramas gerados por esse algoritmo apresentaram uma melhor representação na estrutura dos grupos formados quando comparado com o resultado produzido pelo algoritmo Single-Linkage, portanto realça ainda uma grande individualidade das bacias, conforme ilustra a Figura 3. Diferenças significativas entre as alturas dos arcos aparecem desde as primeiras junções significando um grau de dispersão elevado entre os elementos dentro do grupo.

- ***Ward***

Os dendogramas gerados pelo algoritmo Ward, conforme ilustra a Figura 4, apresentaram uma melhor estrutura de grupos quando comparado com os dois algoritmos anteriores. Podemos observar a formação de grupos compactos e bem delimitados.

Observa-se através das Figuras 2, 3 e 4 que o algoritmo Ward apresentou o melhor resultado, com a formação de grupos bem definidos. Portanto os grupos gerados, em cada cenário por este algoritmo, foram selecionados para o cálculo dos índices de validação, a saber, índice de Silhouette (Rao et. al 2006), índice Davies-Bouldin (Davies 1979) e índice Dunn (Rao et al. 2006). O índice de Silhouette é utilizado para determinar o melhor número de grupos a ser considerado, já que diversos agrupamentos foram obtidos. Os índices *Davies-Bouldin* e *Dunn* baseiam-se na idéia de identificar grupos compactos e bem separados.

As técnicas de validação permitem comparar diversos algoritmos de agrupamento, comparar duas partições, determinar o valor mais apropriado de algum parâmetro do algoritmo, entre outros e são utilizadas para avaliar a qualidade dos agrupamentos obtidos.

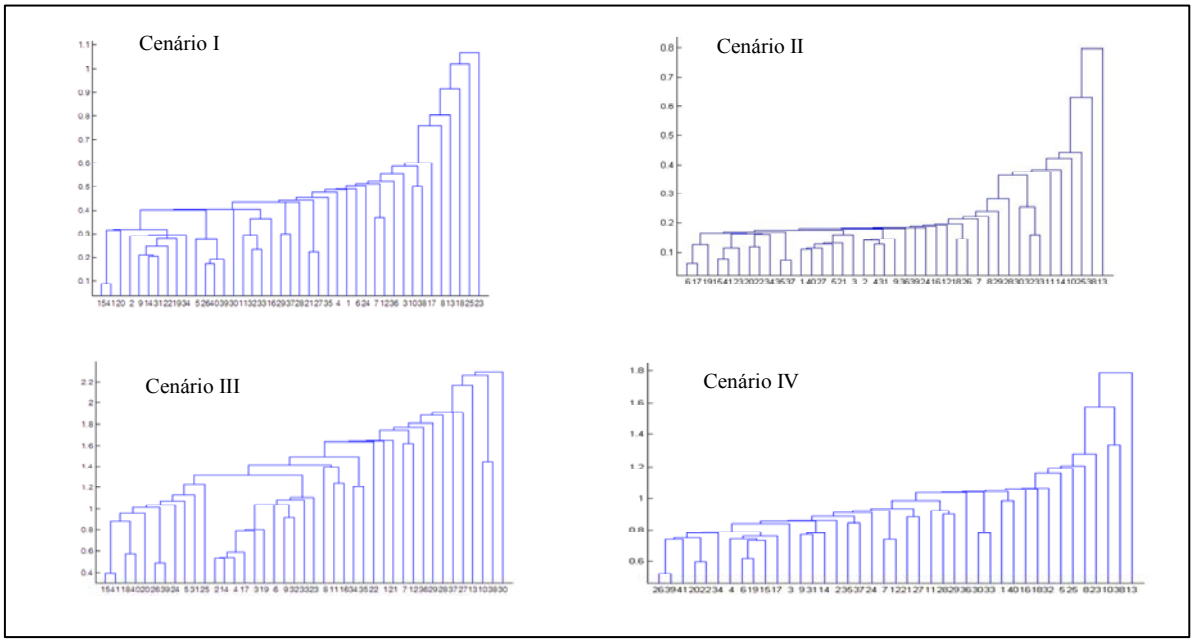

Figura 2 - Dendogramas gerados pelo Single-Linkage

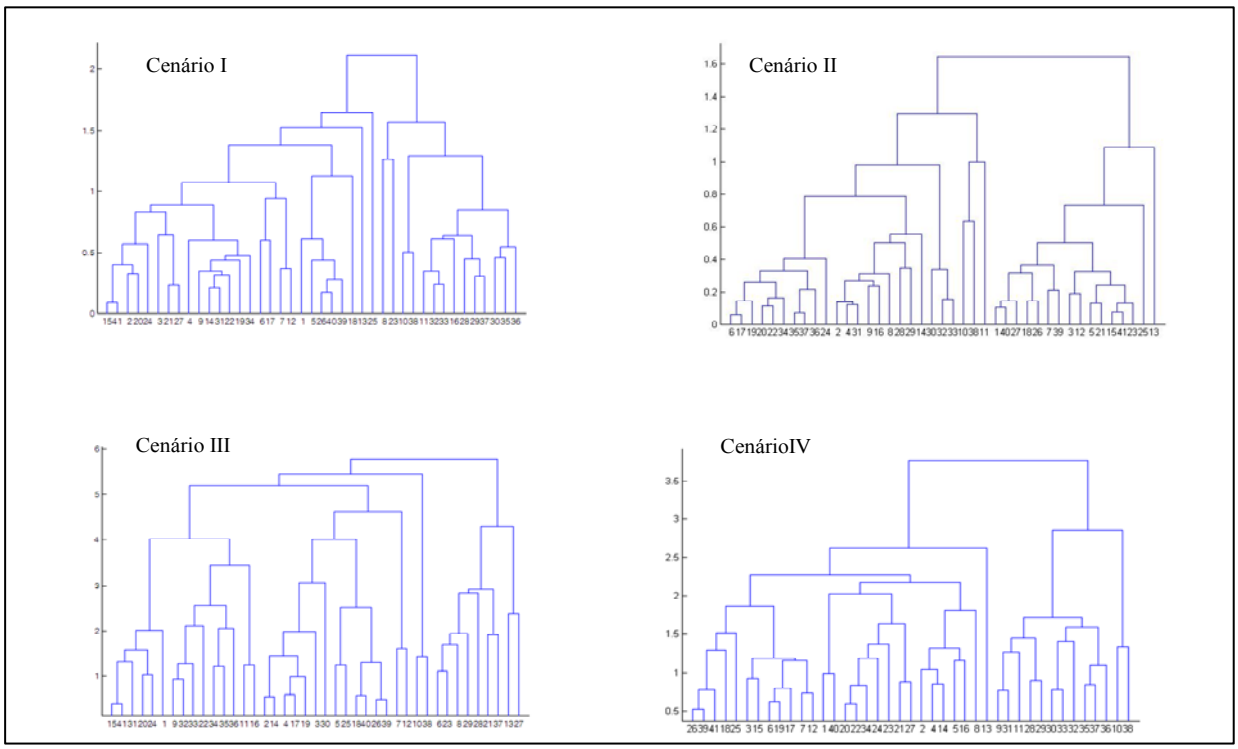

Figura 3 - Dendogramas gerados pelo *Complete-Linkage*

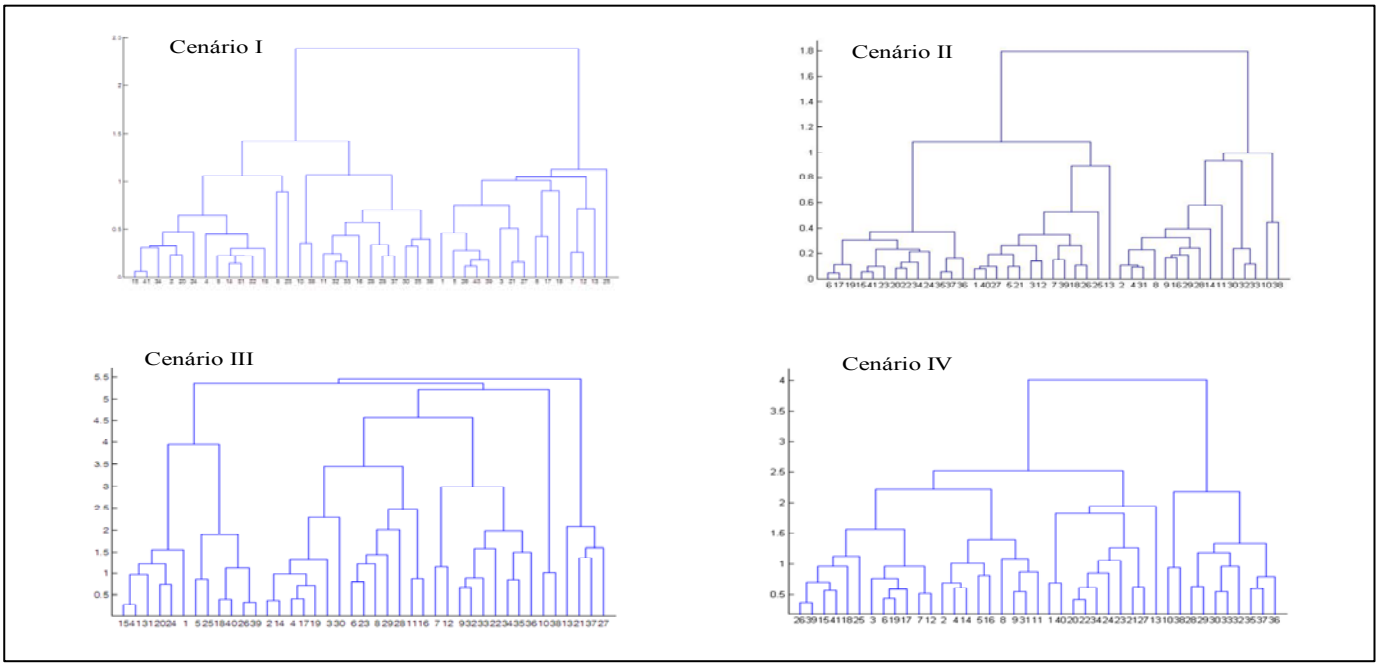

Figura 4 - Dendogramas gerados pelo *Ward*

Tabela 5 - Índices de Validação dos Agrupamentos gerados pelo *Ward*

Cenários	Nº de Grupos					
	Nº de Grupos = 2			Nº de Grupos = 3		
	S	*DB*	*D*	*S*	*DB*	*D*
I	0,264	1,516	1,285	0,165	1,681	0,949
II	0,340	1,387	1,439	0,219	1,345	0,997
III	0,144	1,815	1,089	0,179	1,756	0,935
IV	0,237	1,395	1,431	0,171	1,649	1,004
Cenários	Nº de Grupos					
	Nº de Grupos = 4			Nº de Grupos = 6		
	S	*DB*	*D*	*S*	*DB*	*D*
I	0,196	1,468	1,026	0,204	1,497	0,711
II	0,253	1,502	0,976	0,365	1,085	1,218
III	0,227	1,703	0,928	0,043	1,669	0,816
IV	0,141	1,661	0,840	0,184	1,446	1,084

S - Silhouette; DB – D.Bouldin; D - Dunn.

Com os resultados apresentados na Tabela 5, conclui-se que os índices de validação indicaram que os agrupamentos obtidos no cenárioII apresentam uma qualidade melhor que os obtidos nos demais cenários. Os índices de Silhouette e Dunn sinalizam um bom resultado quando o seu valor é maximizado. O índice de Davies- Bouldin se comporta de forma contrária aos demais, sinalizando, portanto um bom resultado quando o seu valor é minimizado.

Com os resultados apresentados na Tabela 5, conclui-se que os índices de validação indicaram que os agrupamentos obtidos no cenário II apresentam uma qualidade melhor que os obtidos nos demais cenários. Os índices de Silhouette e Dunn sinalizam um bom resultado quando o seu valor é maximizado. O índice deDavies- Bouldin se comporta de forma contrária aos demais, sinalizando, portanto um bom resultado quando o seu valor é minimizado.

Algoritmo Particional

- ***K-Means***

Na execução do algoritmo *K-Means*, foram utilizados como parâmetros: o número de grupos (k=2,3,4 e 6), a medida de distância

Euclidiana, a seleção dos k centróides iniciais de forma randômica e o número máximo de iterações igual a 100. Em cada cenário e para cada valor de k foram realizadas 10 simulações. Com os agrupamentos gerados foram calculados os índices de validação, já citados na seção anterior, e calculada a média para cada grupo de 10 simulações. O resultado pode ser observado na Tabela 6.

Pode-se concluir, com base no resultado os índices de validação, que os agrupamentos obtidos no cenário II apresentam uma melhor qualidade em relação aos obtidos nos demais cenários.

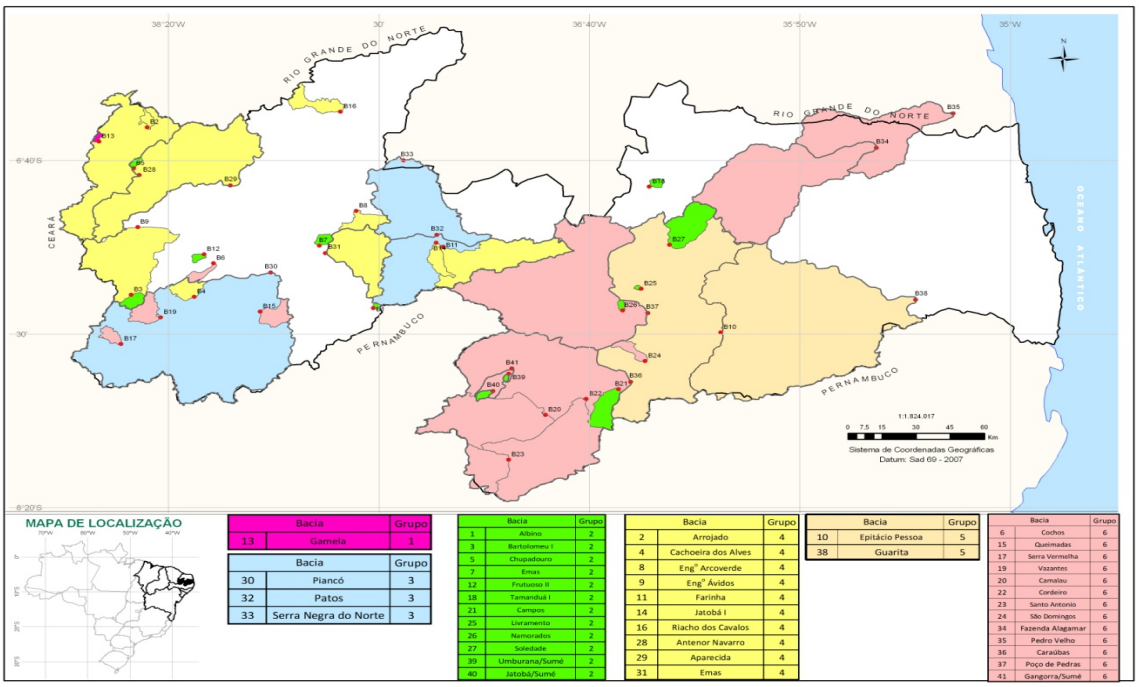

Figura 5 - Divisão das 41 bacias hidrográficas em 6 grupos

Tabela 6 - Índices de Validação dos Agrupamentos gerados pelo *K-Means*

Cenário	Nº de Grupos					
	Nº de Grupos = 2			Nº de Grupos = 3		
	S	*DB*	*D*	*S*	*DB*	*D*
I	0,261	1,516	1,285	0,224	1,604	1,002
II	0,309	1,380	1,396	0,295	1,453	1,046
III	0,176	1,818	1,015	0,177	1,822	0,943
IV	0,176	1,557	1,179	0,170	1,578	1,002
Cenário	Nº de Grupos					
	Nº de Grupos = 4			Nº de Grupos = 6		
	S	*DB*	*D*	*S*	*DB*	*D*
I	0,202	1,620	0,998	0,205	1,502	0,875
II	0,315	1,380	0,980	0,281	1,233	0,713
III	0,200	1,673	0,946	0,243	1,436	0,985
IV	0,147	1,610	0,880	0,143	1,50	0,844

S - Silhouette; DB – D.Bouldin; D - Dunn.

Rede Neural de *Kohonen*

Para a execução da rede neural de Kohonen foi utilizada a seguinte Configuração da rede: tamanho da rede [6x7], topologia retangular, função de vizinhança gaussiana e nº de épocas 30x120.

Com o resultado obtido do treinamento da rede neural de *Kohonen* em cada um dos cenários, foram executadas, para cada valor de *k* (2,3,4 e 6), 10 simulações do algoritmo *K-Means*. Com os agrupamentos resultantes foram calculados os índices de validação e obtidos a média e o desvio padrão. A Tabela 7 apresenta os valores obtidos de cada índice para cada valor de *k*.

Os resultados alcançados pela validação estatística confirmaram mais uma vez que a qualidade dos agrupamentos gerados no cenário II é superior aos gerados nos demais cenários.

Observando os números apresentados pode-se concluir que o algoritmo Ward se destacou na tarefa de clusterização das 41 bacias hidrográficas.O mesmo obteve os melhores resultados nos três índices de validação utilizados. Com relação ao número de grupos, tivemos o número de grupos igual a 6 obtendo os melhores resultados em dois dos índices de validação (Silhouette e Davies-Bouldin).

Portanto a Figura 5 ilustra o melhor resultado obtido neste trabalho que foi a divisão das 41 bacias hidrográficas em 6 grupos, obtidos através da aplicação do algoritmo hierárquico Ward com os dados do cenário II.

Tabela 7 - Índices de Validação dos Agrupamentos gerados pela Rede de *Kohonen*

Cenário	Nº de Grupos					
	Nº de Grupos = 2			Nº de Grupos = 3		
	S	*DB*	*D*	*S*	*DB*	*D*
I	0,260	1,481	1,262	0,204	1,722	0,983
II	0,306	1,358	1,439	0,256	1,571	1,164
III	0,231	1,853	1,042	0,172	1,805	0,972
IV	0,221	1,407	1,412	0,147	1,614	0,951
Cenário	Nº de Grupos					
	Nº de Grupos = 4			Nº de Grupos = 6		
	S	*DB*	*D*	*S*	*DB*	*D*
I	0,171	1,661	1,036	0,161	1,630	0,751
II	0,223	1,526	0,881	0,193	1,373	0,566
III	0,216	1,661	1,032	0,220	1,529	0,912
IV	0,120	1,600	0,885	0,111	1,625	0,761

S- Silhouette; DB – D.Bouldin; D - Dunn.

CONCLUSÃO

A avaliação dos resultados de uma clusterização não é tarefa trivial devido a mesma pertencer a classe de tarefas de aprendizado não supervisionado, ou seja, não se tem conhecimento a priori dos dados. Na literatura, recomenda-se a utilização de técnicas de validação que podem ser utilizadas para avaliar a qualidade dos agrupamentos obtidos. Os resultados apresentados na seção anterior permitem as seguintes conclusões:

- Os algoritmos aplicados obtiveram melhores resultados com os dados do cenário II. As medidas lineares e de declividade, que compõem esse cenário, apresentam a vantagem de serem facilmente determinadas a partir de mapas publicados para a maioria das regiões brasileiras com aceitável grau de confiabilidade;
- Com relação aos componentes principais, largamente utilizado em aplicações de clusterização (Demirel et. al 2009; Júnior et. al 2006; Llanillo et. al 2006), o seu uso não se mostrou eficiente do ponto de vista de prover uma boa classificação, para os dados das bacias hidrográficas levantadas neste trabalho;
- Com os resultados obtidos nos quatro cenários conclui-se que a etapa de pré-processamento, especificamente o processo de seleção de atributos, é de suma importância no processo de mineração de dados. Isto pôde ser observado no quarto cenário, composto pelos 32 atributos das bacias hidrográficas, que apresentou resultados inferiores aos obtidos nos cenários I e II;
    - A aplicação de índices de validação estatística nos agrupamentos gerados, normalmente ausentes em trabalhos na área de engenharia no Brasil, reduz o empirismo que tem caracterizado as análises e aplicações feitas em estudos afins.
- Os resultados da validação estatística apresentados pela Rede Neural de Kohonen foram inferiores aos apresentados pelos algoritmos Ward e K-Means. A pouca quantidade de dados utilizada na aplicação desse algoritmo, não permitiu uma exposição suficiente de dados de entrada para assegurar um melhor processo de auto-organização como recomenda a literatura, impossibilitando assim a geração de melhores resultados;
- Através dos grupos formados, pode-se observar que as bacias hidrográficas foram separadas sem a predominância da localização das mesmas, como normalmente se considera em estudos empíricos de separação de grupos;
- Os resultados obtidos neste trabalho constituem uma indicação de referência para estudos de regionalização hidrológica no Estado da Paraíba;
- A metodologia aqui proposta pode ser facilmente aplicada em outras regiões sem perdas da confiabilidade dos resultados, uma vez que para isto, só será necessário mudar apenas a base de dados das bacias hidrográficas e a consequente definição das regiões hidrologicamente homogêneas.

REFERÊNCIAS

BOLSHAKOVA, N. "*Machaon clustering and validation environment*". Disponível em: https://www.cs.tcd.ie/Nadia.Bolshakova/Machaon.html. Acessado em: 13/04/2009.

DAVIES , D. L.; BOULDIN, D. W. "*A cluster separation measure*". IEEE Transactions on Pattern Recognition and Machine Intelligence, 1:224–227, 1979.

DEMIREL M. C.; MARIANO A. J.; KAHYA E. *Performing k-means analysis to drought principal components of Turkish Rivers*. Hidrology Days 2007, http://hydrologydays.colostate.edu/Proceedings_2007.htm. Acessado em 20/01/2009.

DINIZ, L. S. "*Regionalização de parâmetros de modelos chuva-vazão usando redes neurais*". Tese de doutorado. Instituto de Pesquisas Hidráulicas da UFRGS, Porto Alegre-RS, 230 p, 2008.

FAYYAD, U. M., PIATETSKY-SHAPIRO, G., SMYTH P. "*From data mining to knowledge discovery in database*". AI Magazine, p.37-54, 1996.

JÚNIOR J. C. F. M.; SEDIYAMA G. C.; FERREIRA P. A.; LEAL B. G. "*Determinação de regiões homogêneas quanto à distribuição de frequência de chuvas no leste do Estado de Minas Gerais*". Revista Brasileira de Engenharia Agrícola e Ambiental, v.10, n.2, p.408–416, 2006.

HAYKIN, S. Redes Neurais – Princípios e Prática. 2ª Edição – Porto Alegre: Bookman, 2001.

LAROSE D. T. Discovering Knowledge in Data – An Introduction to Data Mining. Wiley-Interscience, p. 153-158. 2005.

LLANILLO, R. F, DEL GROSSI, M. E., SANTOS F. O., MUNHOS, P. D., GUIMARÃES M. F. "*Regionalização da agricultura do Estado do Paraná, Brasil*". Cienc. Rural vol.36 no.1 Santa Maria Jan./Feb. 2006.

METZ, J. "*Interpretação de Clusters Gerados por Algoritmos de Clustering Hierárquico*". Dissertação de Mestrado, USP, São Carlos - SP, 2006.

MITRA, P.; MURTHY, C. A.; PAL, S. K. (2002), "*Unsupervised Feature Selection Using Feature Similarity*". IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, nº 3, March 2002.

PORTO, M.M.; ANDRADE, E.M.; COSTA, R.N.T.; FILHO, L.C.A.L.; MEIRELES, M. "*Identificação de Bacias Hidrográficas Com Características Físicas Similares no Estado do Ceará*", Brasil. Revista Ciência Agronômica, vol. 35, nº. 1, 17-25, 2004.

RABUS, B., EINEDER, M., ROTH, A., BAMLER, R. "*The shuttle radar topography mission - A new class of digital elevation models acquired by spaceborne radar*". Journal of Photogrammetry & Remote Sensing, v. 57, p. 241-262, 2003.

RAO, A. R., SRINIVAS, V. V. "*Regionalization of watersheds by hybrid-cluster analysis*", Journal of Hydrology 318, p. 37-56, 2006.

TC/BR - Riverside Technology Inc. (2001). Plano Diretor de Recursos Hídricos da Bacia Hidrográfica do Rio Paraíba, Tomo I, Relatório Final de Diagnóstico, Paraíba, 142p.

WARD J. H. Hierarchical Grouping to Optimize an Objetive Function. American Statistical Association Journal, March 1963.

Use of Data Mining Techniques to Identify Hydrologically Homogeneous Areas in the State of Paraíba

ABSTRACT

The lack of fluviometric data and the bad quality of existing data related to watercourses in the Brazilian northeast have obliged hydrologists to seek new solutions, increasing knowledge and methodologies to develop the region according to its environment limitations. Through hydrologic regionalization techniques it is possible to transfer data and information among similar watersheds. In this context, the purpose of this work is to identify hydrologically similar regions in the State of Paraiba using Clustering - a kind of data mining technique - to find patterns that allow data transposition from one region to another. Algorithms were used with methods based on partition, hierarchical methods, and methods based on neural networks, and applied indexes of statistical validation in the generated groupings. In agreement with the results obtained, the Ward algorithm presented the best result of all the applied validation indexes with the identification of six hydrologically similar regions in the State of Paraíba.

Key-words: *Data mining; hydrological regionalization; clustering.*