Redes probabilísticas baseadas em similaridade na exploração de comunidades

Robson C. da Motta, Leonardo J. Almeida e Alneu de A. Lopes

Instituto de Ciências Matemáticas e de Computação - ICMC
Universidade de São Paulo/Campus de São Carlos - (USP São Carlos)
Caixa Postal 668, São Carlos, SP, Brasil

Abstract. Neste trabalho, apresentamos uma técnica para a modelagem de redes complexas probabilísticas baseadas na relação de similaridade entre os objetos. Também definimos os critérios para obtenção da melhor rede e sua divisão em comunidades temáticas ou agrupamentos. A abordagem proposta utiliza-se de um processo aglomerativo iterativo no qual o agrupamento de componentes é baseado em uma medida que procura maximizar a densidade de ligações intra-componentes e minimizar as ligações entre-componentes. O modelo proposto foi avaliado em um conjunto de dados textuais e um conjunto de dados numéricos. As comunidades encontradas foram avaliadas por medida de pureza demonstrando uma correspondência significativa com os assuntos ou classes das bases de dados.

Key words. redes complexas, similaridade, estrutura de comunidade, agrupamento.

1 Introdução

A enorme quantidade de dados armazenados e disponíveis em bases de dados e na *Web* tem impulsionado a pesquisa em descoberta de conhecimento em base de dados (*Knowledge Discovery in Database*) e em Mineração de Dados. Provavelmente quase 90% dos dados das empresas são armazenados em documentos textuais, tais como: cartas, e-mail, relatórios, contratos, patentes, etc, e a maior parte desses documentos tem sua versão digital. Desenvolvimento de técnicas que permitem análise, exploração e organização de grandes conjuntos de dados textuais devem, portanto, permanecer como uma parte fundamental na tecnologia de descoberta do conhecimento.

Tarefas típicas de mineração de dados, tais como classificação e análise de *cluster* comumente buscam encontrar padrões em dados representados por uma coleção de instâncias independentes de um único conceito. Entretanto, além dos atributos que caracterizam os objetos ou instâncias, em um grande número de aplicações existem também informações adicionais relativas às relações entre os objetos. Essas informações, em geral representadas em grafos, provêem uma visão adicional dos dados que pode ser usadas no processo de mineração e de descoberta de conhecimento [1–3].

Uma rede que represente adequadamente a relação de similaridade entre os dados deve ter grupos de objetos similares densamente conectados entre si e fracamente conectados a outros grupos. Supondo, no entanto, que se construa uma rede na qual todos os vértices com similaridade acima de um dado limiar sejam conectados, dois problemas podem ocorrer: se o limiar for alto, a rede gerada pode não ser conexa, com vários componentes desconectados; se for baixo suficiente para obter uma rede conexa, o número de arestas cresce demasiadamente, perdendo assim a característica de expressar relação entre os mais similares e aumentando a complexidade de diversas tarefas.

Dessa forma, neste trabalho propomos um algoritmo que tem como objetivos (i) construir uma rede conexa baseada na similaridade entre os vértices; (ii) usar pouca quantidade de arestas (grau médio $\langle k \rangle$ baixo), priorisando as conexões entre vértices mais similares; (iii) e que a vizinhança de um vértice, dada pelos seus vértices adjacentes, represente bem esse vértice. Os experimentos realizados indicam que tal rede representa apropriadamente a relação de vizinhança entre os objetos, permitindo identificar comunidades ou agrupamentos. Para a análise das redes obtidas utilizamos propriedades de Redes Complexas [4–6].

O restante do artigo está organizado como segue. Na Seção 2 são descritos os principais conceitos de redes complexas e técnicas de detecção de comunidade. Na Seção 3 é descrita a abordagem para criação da rede baseada em similaridade. Na Seção 4, avaliamos a abordagem proposta usando um corpus de artigos científicos e um conjunto de dados numéricos. Por último, na Seção 5, apresentamos as conclusões e trabalhos futuros.

2 Detecção de comunidades e redes complexas

Algumas abordagens utilizadas em *clustering* são baseadas em conceitos e propriedades da teoria dos grafos. Chameleon [7] é um algoritmo de *Clustering* Hierarquico Aglomerativo (*Hierarchical Agglomerative Clustering* - HAC) baseado em grafo de *k-nearest-neighbour*. O aspecto determinante desse algoritmo é seu processo de união de dois *clusters*. O algoritmo usa um critério que considera tanto a interconectividade entre os *clusters* como a proximidade entre eles. O fato de levar em conta a interconectividade aproxima-o deste artigo, entretanto nossa medida de interconectividade já reflete simultaneamente a proximidade e interconectividade dos componentes.

Hartuv e Shaamir [8] propuseram um algoritmo baseado em grafo de similaridade e agrupamentos são considerados como subgrafos altamente conectados. O algoritmo é apresentado com um tempo polinomial. Sharan e Shamir [9] apresentaram uma abordagem baseada no cálculo do corte de peso mínimo para a construção dos *clusters*. O grafo considerado pelos autores é ponderado e os pesos das arestas são atribuídos por uma combinação de probabilidade e teoria dos grafos.

Brinkmeier [10, 11] apresentou uma abordagem baseada na conectividade das arestas nos subgrafos. O autor considera a construção de uma árvore, na qual os

filhos de um nó são disjuntos e estão contidos no pai. O autor chama os subgrafos encontrados de "comunidades".

Newman [12] elaborou um algoritmo baseado na idéia de modularidade. Primeiramente foi criada uma medida para se determinar a qualidade de uma divisão particular, chamada medida de modularidade Q. Para uma divisão da rede a modularidade Q é expressa pela Equação 1, sendo e_{ii} a fração de ligações internas em um componente e e_{ij} a fração de ligações para outros componentes. Portanto, o método de identificação de comunidades [13] é iniciado com cada vértice sendo um único componente (nenhuma aresta), e iterativamente agrupa-se os dois componentes que indicarem maior acréscimo (ou menor decréscimo) na modularidade Q, inserindo as arestas existentes entre eles na rede original. Haverá então uma divisão da rede que possuirá o maior valor de modularidade, sugerindo assim a quantidade de comunidades ideal para a rede.

$$Q = \sum_i [e_{ii} - (\sum_j e_{ij})^2] \tag{1}$$

Outras duas grandezas comumente usadas para caracterizar redes complexas são o grau k e o coeficiente de agrupamento cc. O grau k de um vértice i é a quantidade de vértices adjacentes a i. O coeficiente de agrupamento expressa a probabilidade de dois vértices que estão conectados possuírem uma conexão em comum com um terceiro vértice. O coeficiente de agrupamento de um vértice i é expresso pela Equação 2, com e_i sendo o número de arestas existentes entre os vizinhos do vértices i, e $k_i(k_i - 1)/2$ é o número máximo possível de arestas entre esses vizinhos, sendo k_i o grau de i.

$$cc_i = \frac{2e_i}{k_i(k_i - 1)} \tag{2}$$

O grau médio $\langle k \rangle$ de uma rede é a média do grau de todos os vértices, assim como o coeficiente de agrupamento médio $\langle cc \rangle$ é a média do coeficiente de agrupamento de todos os vértices.

Uma medida comumente usada para avaliação das técnicas é a pureza de cada comunidade encontrada. A pureza é dada pela fração de exemplos da classe majoritária na comunidade, e a pureza global $\langle Pureza \rangle$ é a média ponderada da pureza de cada comunidade.

3 Explorando estrutura de comunidades em rede probabilística baseada em similaridade

O algoritmo proposto adota uma abordagem aglomerativa hierarquica, porém com duas diferenças fundamentais das técnicas de *clustering* aglomerativas usuais (i) a seleção dos vértices não é determinística; (ii) o processo de agrupamento de dois componentes é baseado em uma medida de interconectividade entre os componentes. Essa medida prioriza a união de componentes densamente conectados entre si e fracamente conectados a vértices de outros componentes. Essa medida é formalmente definida a frente.

A construção da rede inicia-se com cada vértice sendo um componente e com uma exigência elevada de similaridade para permitir a conexão entre vértices e a cada iteração esse limiar de similaridade é decrementado. O processo é repetido até formar uma rede conexa. Em experimentos com diferentes conjuntos de dados observamos que a similaridade entre os vértices, em diferentes conjuntos, pode variar muito. Assim, a similaridade inicial e o decaimento foram definidos de forma relativa, isto é, considerou-se a cada iteração um limiar de similaridade que permite a conexão de mais 5% dos vértices iniciais. O Algoritmo 1 detalha o processo de criação da rede.

Algoritmo 1 Aglomeração iterativa hierárquica

Entrada:
Conjunto de vértices: $V = v_1,...,v_n$
Grau médio: $grauMedio$
Matriz de similaridade entre exemplos: $similaridade$
Saída:
Rede gerada, sendo um conjunto de vértices e de arestas: (V,A)

Componentes $C \leftarrow V$
Arestas $A \leftarrow \emptyset$
$minSim \leftarrow$ similaridade para se obter os 5% pares de vértices mais similares
Enquanto ($\#C > 1$)
Enquanto ($\exists$ par (x,y) | similaridade$(x,y) \geq minSim$, $x \in C_i$, $C_i \in C$, $y \in C\text{-}C_i$)
Agrupamento dos componentes(C,A,$grauMedio$,$minSim$)
$minSim \leftarrow$ similaridade para se acrescentar 5% pares de vértices mais similares

Retorna (V,A)

Para o processo de agrupamento dos componentes (Algoritmo 2), selecionamos aleatoriamente pares de vértices (x, y) tal que x pertence a C_i e y pertence a qualquer um dos outros componentes da rede. Se a similaridade entre x e y é maior ou igual a similaridade mínima a aresta entre x e y é "pré-selecionada". Este processo repete-se até todos os pares serem avaliados ou ser pré-selecionada uma quantidade de arestas que faça com que o componente atinja o grau médio desejado. A quantidade de arestas para atingir o grau médio desejado é obtida considerando a quantidade de vértices e de arestas já existentes no componente, ou no minímo 1, para que a rede sempre fique conexa.

Com as arestas entre os pares de componentes pré-selecionadas é, então, definido quais são os melhores componentes a serem agrupados. O processo adotado foi definido de tal forma a maximizar as ligações intra-componentes e minimizar as entre-componentes, após a união dos componentes selecionados. Esse critério é obtido pelo argumento máximo da Equação 3.

$$interconectividade(C_i, C_j) = \frac{1}{\#C_i + \#C_j} \sum_{\substack{x \in C_i, y \in C_j, \\ \exists aresta(x,y)}} sim(x, y) \qquad (3)$$

Une-se então o par de componentes com maior medida de interconectividade. A medida reflete simultaneamente a contribuição das novas arestas à "densidade" de ligações internas após a união dos componentes e a "qualidade" dessas ligações (pois também leva em conta a similaridade entre os vértices). Esse critério de união de componentes diferencia nossa técnica das abordagens tradicionais de aglomeração hierárquica (*single-link*, *complete-link* e *average-link*).

Algoritmo 2 Agrupamento dos componentes

Entrada:
Conjunto de componentes: C
Conjunto de arestas: A
Grau médio: $grauMedio$
Limiar de similaridade: $minSim$
Saída:
Conjunto de componentes: C
Conjunto de arestas: A

arestasPreSelecionadas ← ∅
Para cada componente C_i de C
  $qtdDeArestas$ ← ($grauMedio$ * $\#C_i$ / 2) - $\#A(C_i)$
  Se (qtdDeArestas ≤ 0)
    qtdDeArestas ← 1)
  Para todo par(i,j) obtido aleatoriamente | $i \in C_i$ e $j \in C$ - C_i
    Se (similaridade(i,j) ≥ $minSim$)
      arestasPreSelecionadas(C_i,C_j)←arestasPreSelecionadas(C_i,C_j) ∪ (i,j)
      $qtdDeArestas$−−
      Se ($qtdDeArestas$ == 0)
        $break$
(C_a,C_b) ← max(interconectividade(C_i,C_j)) %componentes selecionados
C_a ← C_a ∪ C_b %união dos componentes
$A(C_a)$ ← $A(C_a)$ ∪ $A(C_b)$ ∪ arestasPreSelecionadas(C_a,C_b) %união das arestas dos componentes unidos
C ← C - C_b %remove componente que foi unido
A ← A - $A(C_b)$ %remove de A as arestas do componente removido

Retorna (C,A)

No algoritmo proposto, o grau médio definido leva naturalmente a diferentes redes. Uma rede baseada em similaridade apropriada para exploração de comunidades deve ter um compromisso entre três características: (i) alto valor para a relação entre a quantidade de ligações intra- e entre-componentes, expressa pela medida de modularidade Q (Equação 1); (ii) alta densidade das ligações

intra-componentes, expressa pelo coeficiente de agrupamento (Equação 2); e (iii) baixa quantidade de arestas, expressa pelo grau médio. Ou seja, queremos uma rede que maximize o compromisso expresso pela Equação 4.

$$Qualidade = \frac{Q * \langle cc \rangle}{\langle k \rangle} \tag{4}$$

As redes geradas a partir desse critério minimizam as ligações entre comunidades temáticas ou agrupamentos comparativamente à quantidade de ligações intra-comunidade, permitindo a visualização dos agrupamentos, conforme pode ser observado na Figura 1.

Em uma próxima etapa do trabalho essa característica da rede será usada para desenvolver uma nova técnica de detecção de comunidades, separando os sub-grafos identificados na rede.

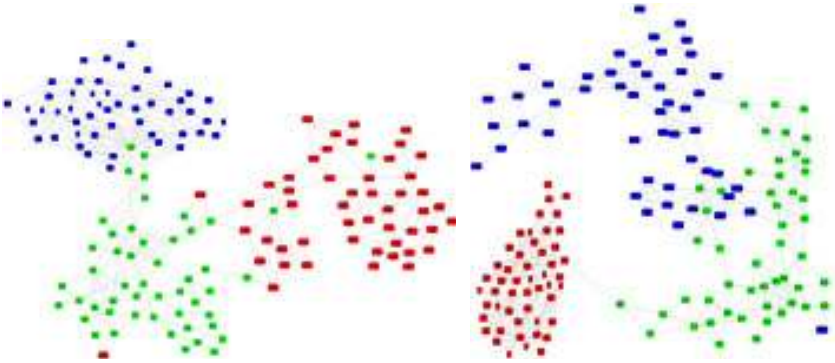
4 Experimentos

Para realização dos experimentos foram utilizadas duas bases de dados, sendo uma base textual e uma base numérica. A base de dados textual contém 574 documentos em inglês, são publicações obtidas na *LNAI* das áreas de *Cased Based Reasoning* (CBR) e *Inductive Logic Programming* (ILP), e documentos recuperados na *Web* sobre *Information Retrieval* (IR). Os documentos foram representados em um modelo *bag-of-word* após a eliminação de *stopwords* e aplicação do algoritmo *Porter's stemming* e de corte de Luhn. A base numérica é a de plantas Íris, contendo 150 exemplos distribuídos igualmente em três classes (iris-setosa, iris-versicolour e iris-virginica). Possui quatro atributos numéricos e nenhum valor ausente. Também foram realizados experimentos com outras bases, que se comportaram de forma semelhante.

Primeiramente foi definido o conjunto de grau médio a ser analisado (2, 3, 4, 5, 6, 7, 8, 9, 10, 20 e 30). Para cada grau médio foram gerados resultados utilizando *10-fold cross validation*, com as redes sendo geradas a partir de cada conjunto de treino. Para determinar o padrão médio de comportamento das redes, foram criadas 5 redes com cada conjunto de treino, obtendo uma média para o grau, a modularidade Q e o coeficiente de agrupamento.

O melhor resultado para a função de qualidade das duas bases foi para redes com grau médio próximo de 4 (para maioria dos experimentos observou-se que os melhores valores são para grau médio entre 3 e 5). Como exemplo, foram geradas duas redes com a base Íris, com grau médio 3,35 e 5,37, ilustradas na Figura 1.

Embora tenham sido definidos os graus médios de 2 a 30 as redes probabilísticas tornaram-se conexas com graus inferiores. Na Tabela 1 resumimos os valores encontrados, sendo que para identificação de comunidades foi utilizado o método proposto por Newman [13] baseado na modularidade Q. Nesta tabela comparamos as redes probabilísticas com uma rede determinística na qual se adotou o maior limiar de similaridade que mantinha a rede conexa. Observa-se que a quantidade de arestas ou grau médio são extremamente superiores nas



(a) (b)

Fig. 1. Redes formadas a partir da base Íris, (a) com grau médio 3,35 e (b) com grau médio 5,37.

Table 1. Características das redes

		$\langle k \rangle$	#arestas	$\langle cc \rangle$	modul. Q	qualid.	$\langle Pureza \rangle$	Pureza local
Íris	determinística	55,58	3751,90	0,90	0,40	0,13	0,87	0,88
	probabilística	2,14	144,28	0,03	0,82	0,24	0,91	0,93
		4,71	318,10	0,34	0,71	1,00	0,94	0,95
		7,61	513,72	0,53	0,65	0,90	0,94	0,95
CBR ILP IR	determinística	26,79	6924,24	0,48	0,58	0,97	0,88	0,96
	probabilística	2,03	524,23	0,01	0,91	0,32	0,79	0,89
		5,59	1444,06	0,10	0,61	1,00	0,88	0,94
		19,73	5099,51	0,36	0,57	0,96	0,88	0,96

redes determinísticas, e os melhores valores de pureza são próximos dos obtidos nas redes probabilísticas.

Nesta tabela, além das medidas características de redes complexas, adicionamos às medidas de qualidade e pureza de comunidades já apresentadas uma medida que reflete a fração dos vértices que estão conectados majoritariamente a vértices de sua classe, denominada pureza local. A pureza local de cada base foi normalizada entre 0 e 1, de acordo com o maior valor obtido.

5 Conclusões

A rede probabilística proposta permitiu a construção de uma rede conexa, com uma quantidade de aresta extremamente inferior a uma rede determinística. Com grau médio $\langle k \rangle$ com valores entre 3 e 5 obteve-se os melhores resultados em termos de pureza de componentes e na distribuição das arestas entre e

intra-componentes. As redes probabilística com grau médio baixo conseguiram representar bem os dados conforme se observa na Tabela 1 e na Figura 1. A medida de qualidade para a seleção do grau médio para a construção da rede que melhor representa os dados é compatível com a pureza observada nas redes com os mesmos graus médios. Em trabalhos futuros, utilizaremos essa estratégia de construção da rede em uma abordagem de classificação baseada em *links* e em novos métodos para identificação de comunidades.

6 Agradecimentos

O primeiro autor agradece a CAPES-Brasil pelo suporte financeiro a este trabalho.

References

1. Getoor, L.: Link mining: a new data mining challenge. SIGKDD Explorations **5**(1) (2003) 84–89
2. Senator, T.E.: Link mining applications: progress and challenges. SIGKDD Explorations **7**(2) (2005) 76–83
3. Getoor, L., Diehl, C.P.: Link mining: a survey. SIGKDD Explorations **7**(2) (2005) 3–12
4. Barabási, A.L., Albert, R., Jeong, H.: Mean-field theory for scale-free random networks. Physica A **272** (1999) 173–187
5. Barabási, A.L., Albert, R., Jeong., H.: Scale-free characteristics of random networks: the topology of the world-wide web. Physica **281**(1) (2000) 69–77
6. Newman, M.E.J.: The structure and function of complex networks. Society for Industrial and Applied Mathematics Review **45**(2) (2003) 167–256
7. Karypis, G., Han, E.H., Kumar, V.: Chameleon: Hierarchical clustering using dynamic modeling. IEEE Computer **32**(8) (1999) 68–75
8. Hartuv, E., Shamir, R.: A clustering algorithm based on graph connectivity. Information Processing Letters **76**(4-6) (2000) 175–181
9. Sharan, R., Shamir, R.: Click: A clustering algorithm with applications to gene expression analysis. In: Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology, AAAI Press (2000) 307–316
10. Brinkmeier, M.: Communities in graphs. Technical report, Technical University of Ilmenau (2002)
11. Brinkmeier, M., Werner, J., Recknagel, S.: Communities in graphs and hypergraphs. In: Proceedings of the sixteenth ACM Conference on information and knowledge management, New York, NY, USA (2007) 869–872
12. Newman, M.: Detecting community structure in networks. European Physical Journal B **38**(2) (2004) 321–330
13. Newman, M.: Fast algorithm for detecting community structure in networks. Physical Review E **69** (2004) 066133