Faculdade de Tecnologia Bandeirantes - BandTec 1^o semestre de 2015

Descobrindo segmentos de adolescentes em redes sociais - possível solução

Pós Graduação em Big Data Prof. Fabrício Jailson Barth

1 Introdução

Interagir com amigos em redes sociais, tais como Facebook e MySpace, tem se tornado um ritual entre os adolescentes do mundo todo. Estes adolescentes, ao mesmo tempo que interagem com os seus amigos, também estão sugeitos a propagandas de diversos produtos e empresas. É muito importante para as empresas, e até mesmo para os usuários de redes sociais, que tais propagandas sejam melhor direcionadas, ou seja, que sejam entregues para aqueles que realmente tem interesse naquele tipo de produto.

Dado o texto produzido por adolescentes em Serviços de Redes Sociais, acredita-se que podemos identificar grupos de adolescentes que compartilham interesses em comum, tais como esporte, religião ou música. Algoritmos de clustering podem automizar este processo de descoberta de segmentos naturais desta população.

O objetivo deste estudo é identificar agrupamentos de pessoas baseado no conteúdo gerado pelas mesmas.

2 Método

O método utilizado por esta análise é composto pelas seguintes fases: aquisição e pré-processamento dos dados; construção do modelo, e; análise do modelo.

2.1 Aquisição e pré-processamento dos dados

Para esta análise, nós vamos utilizar um dataset que representa uma amostra aleatória de 30.000 estudantes de *high school* americana que possuem perfil em uma rede social[1] em 2006. O dataset pode ser encontrado em [2].

Os atributos deste dataset são: **gradyear**: ano de graduação; **gender**: sexo, masculino ou feminino; **age**: idade representada por um número real; **friends**: quantidade de amigos na rede social; **demais atributos**: basketball, football, soccer, entre outros indicam classes onde os textos das pessoas foram classificados. Textos de usuários publicados na rede social foram processados para gerar esta tabela.

A aquisição e o pré-processamento dos dados são apresentados abaixo:

```
1 snsdata <- read.csv("data/snsdata.csv")
2 snsdata <- snsdata[,5:40]
3 sum(is.na(snsdata))
4 dados <- as.data.frame(lapply(snsdata , scale))
```

A linha 1 do código acima é utilizada para a leitura dos dados. Aqui, estamos considerando que o arquivo CSV já foi salvo na máquina local. A segunda linha filtra os atributos pessoais (gradyear, gender, age, friends). O objetivo desta análise é agrupar pessoas considerando apenas as categorias dos textos e posts das pessoas. A linha 3 do código verifica a quantidade de valores faltantes no dataset. O retorno da execução desta instrução é zero. A linha 4 do código faz o *rescaling* dos dados.

[1] Em inglês, Social Network System - SNS
[2] https://raw.githubusercontent.com/fbarth/posGraduacao/master/ExemplosClustering/data/snsdata.csv

2.2 Construção do modelo

Para identificar o melhor número de agrupamentos foi utilizado o método *Elbow*. O código que implementa a função *Elbow* é apresentado no código abaixo da linha 1 até 8. A chamada para a função é realizada nas linhas 10 e 11. O resultado da execução do método *Elbow* é apresentado na figura 1.

```
1  elbow <- function(dataset, title){
2    wss <- numeric(15)
3    for (i in 1:15)
4      wss[i] <- sum(kmeans(dataset,centers=i, nstart=100)$withinss)
5    plot(1:15, wss, type="b", main=paste(title),
6         xlab="Number of Clusters",
7         ylab="Within groups sum of squares", pch=8)
8  }
9
10 set.seed(1234)
11 elbow(dados, "Elbow com dados pré-processados com scale")
```

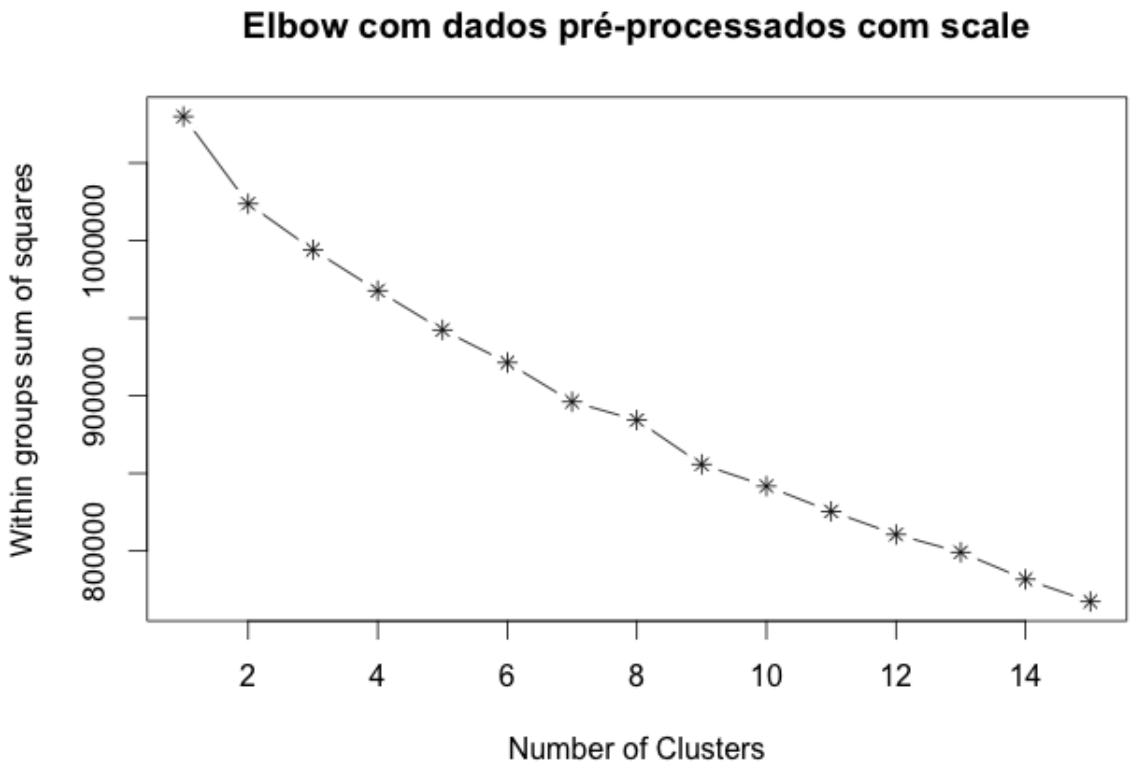

Figura 1: Resultado da execução do método Elbow para o dataset analisado. Apesar da curva apresentada neste gráfico não mostrar um "cotovelo" nítido, é possível verificar que o nível de coesão do modelo com 7 agrupamentos não é muito diferente do nível de coesão do modelo com 8 agrupamentos. Desta forma, a quantidade de agrupamentos escolhida para este dataset foi de 7 agrupamentos.

Após a execução do método *Elbow*, optou-se por construir um modelo com 7 (sete) agrupamentos. A construção do modelo está no código abaixo:

```
1  modelo <- kmeans(dados, centers = 7, nstart = 100)
```

2.3 Análise do modelo

O modelo gerado possui 7 agrupamentos. O tamanho e o nível de coesão de cada agrupamento podem ser vistos no trecho de código abaixo:

```
> modelo$withinss
[1]  33826.78 210504.98 53642.47 62909.45 223615.02 147190.80 169615.97
> modelo$size
[1]   584 20013   861   529  4583  2474   956
```

Para compreender exatamente o que cada agrupamento significa, o ponto central de cada agrupamento foi analisado. Foram identificados os atributos com maior valor em cada um dos sete pontos centrais. A forma utilizada para identificar os atributos com maior valor em cada ponto central está descrita no código abaixo:

```
> x <- as.data.frame(modelo$centers)
> sort(x[1,],decreasing = TRUE)[1:3]
  marching    band    music
1 5.286202 4.126345 0.5156067
> sort(x[2,],decreasing = TRUE)[1:3]
      blonde     tennis      jesus
2 -0.02927463 -0.04086648 -0.07479892
> sort(x[3,],decreasing = TRUE)[1:3]
  hollister abercrombie shopping
3  4.142464   3.988187 0.8049811
> sort(x[4,],decreasing = TRUE)[1:3]
     bible   jesus     god
4 5.225651 2.585811 2.528766
> sort(x[5,],decreasing = TRUE)[1:3]
   shopping     cute     mall
5 0.9052131 0.8046856 0.6759864
> sort(x[6,],decreasing = TRUE)[1:3]
  basketball football baseball
6   1.389166 1.260025 1.241445
> sort(x[7,],decreasing = TRUE)[1:3]
    kissed  drugs    hair
7 3.175714 2.88275 2.551358
```

No agrupamento de número 1, com 584 pessoas, os atributos com maior valor no ponto central são *marching*, *band* e *music*. Desta forma, é possível descrever este grupo como um grupo interessado em músicas e bandas militares.

No agrupamento de número 2, com 20.013 pessoas, os atributos do ponto central com maior valor possuem valores muito próximos da média. Desta forma, é difícil identificar quais são os tópicos de interesse deste agrupamento.

A análise segue desta forma para os demais agrupamentos e é sumarizada na tabela 1.

3 Considerações

Este tabalho identificou sete agrupamentos de pessoas baseando-se no conteúdo gerado pelas mesmas em uma rede social. Um resumo das características de cada agrupamento é apresentado na tabela 1. Uma descrição sobre cada agrupamento é apresentado abaixo:

- **Agrupamento 1**: agrupamento com 584 pessoas e com alta coesão. Os principais tópicos são *marching*, *band* e *music*. Pode-se definir este grupo como pessoas interessadas em músicas e bandas militares. O nível de confiança desta decisão é alta, pois a coesão do grupo é alta e os valores de *marching* e *band* também são altos.

- **Agrupamento 2**: agrupamento com 20.013 pessoas e com baixa coesão. Todos os tópicos possuem valores muito próximos de zero, ou seja, da média. Não pode-se afirmar nada sobre este agrupamento.
- **Agrupamento 3**: agrupamento com 861 pessoas e com alta coesão. Os principais tópicos são *hollister*, *abercrombie* e *shopping*. *Hollister* e *Abercrombie* são marcas de lojas de roupas dos EUA, sendo assim, pode-se concluir que este grupo é um grupo interessado em compras. O nível de confiança desta decisão também é alta devido ao nível de coesão do agrupamento e aos valores dos principais atributos.
- **Agrupamento 4**: agrupamento com 529 pessoas e com alta coesão. Claramente, é um grupo interessado em assuntos relacionados a igreja.
- **Agrupamento 5**: agrupamento com 4.583 pessoas e com baixa coesão. Os principais tópicos são: *shopping*, *cute* e *mall*. Pode-se dizer que este grupo de pessoas tem interesse em compras. No entanto, o nível de confiança desta decisão não é alto, pois o nível de coesão deste grupo é baixo e os valores dos atributos estão próximos de zero.
- **Agrupamento 6**: agrupamento com 2.474 pessoas e com coesão média. Os principais tópicos são tópicos relacionados com esportes. Ou seja, este grupo é de fato um grupo interessado em assuntos esportivos. O nível de confiança desta decisão é médio, pois o nível de coesão deste grupo é médio e os valores dos atributos são relativamente altos.
- **Agrupamento 7**: agrupamento com 965 pessoas e com coesão média. Os principais tópicos são *kissed*, *drugs* e *hair*. *Kissed* é o nome de um filme de 1996 com temática bem adolescente, envolvendo drogas e sexo. *Hair* é um musical da década de 60 que fala sobre drogas e amor livre. De certa forma, este agrupamento pode ser descrito como um agrupamento interessado em sexo e drogas. Esta decisão tem um nível de confiança igual a médio, pois o nível de coesão do grupo é médio e os valores dos atributos são relativamente altos.

	Agrupamentos						
Identificador	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Tamanho	584	20.013	861	529	4.583	2.474	965
Coesão	Alta	Baixa	Alta	Alta	Baixa	Média	Média
Principais tópicos	marching band music	?? ?? ??	hollister abercrombie shopping	bible jesus god	shopping cute mall	basketball football baseball	kissed drugs hair
Sumário	Músicas e bandas militares	??	Compras	Igreja	Compras	Esporte	Drogas e Sexo
Confiança	Alta	Nenhuma	Alta	Alta	Baixa	Média	Média

Tabela 1: Tabela que sumariza as informações dos agrupamentos identificados. Nesta tabela é possível visualiza o tamanho de cada agrupamento, o quão coeso cada grupo é e os principais tópicos relacionados ao agrupamento. Para facilitar a interpretação dos agrupamentos, o nível de coesão foi discretizado para Alta, Média e Baixa. A regra para a discretização é se o valor de withinss for menor que 100.000 então a coesão do agrupamento é alta, se o valor de withinss for entre 100.000 e 200.000 então a coesão é média e se o valor de withinss for maior que 200.000 então a coesão é baixa. Nesta tabela também é apresentado um sumário que descreve cada agrupamento e o nível de confiança desta decisão.