智能数据挖掘作业 5

19200300004 黄铭瑞

实验目的

掌握使用层次聚类和 k 均值聚类的方法，并对 20newsgroups 数据集进行聚类。

实验原理

数据集

20newsgroups 数据集是用于文本分类、文本挖据和信息检索研究的国际标准数据集之一。数据集收集了大约 18000 个新闻组文档，均匀分为 20 个不同主题的新闻组集合。

层次聚类

层次的聚类方法（Hierarchical Clustering），字面理解，即层次化的聚类，最终结果是树状结构，换句话说，层次聚类通过计算不同类别数据点间的相似度，来创建一棵有层次的嵌套聚类树。

层次聚类又分为凝聚方法和分裂方法。

凝聚方法

凝聚方法是一种自下而上的方法，先将所有样本的每个点都看成一个簇，然后找出距离最小的两个簇进行合并，不断重复到预期簇或者其他终止条件。代表的有 Agnes 算法：

1. 初始化，每个样本当作一个簇。
2. 计算任意两个簇的距离，找到这两个簇，合并为一个簇。
3. 重复步骤 2，直到两个簇的距离超过阈值，或者簇的个数达到规定值，终止。

分裂方法

分裂方法是一种自上而下的方法，先将所有样本当作一整个簇，然后找出簇中距离最远的两个簇进行分裂，不断重复到预期簇或者其他终止条件。代表的有

Diana 算法：

1. 初始化，所有样本归为一类。
2. 同个簇中，计算任意两个样本的距离，找到距离最远的两个样本，作为两个簇的中心。
3. 计算原来簇中其余样本距离这两个中心点的距离，把它归为距离近的中心点所属的那一簇。
4. 重复步骤 2、3，直到两簇距离达不到分开的阈值，或者簇数量达到规定之，终止。

K-Means

K-Means 算法的思想很简单，对于给定的样本集，按照样本之间的距离大小，将样本集划分为 k 个簇。让簇内的点尽量紧密的连在一起，而让簇间的距离尽量的大。总的说，让数据与对应聚类中心的误差平方和最小，即：

$$\min(J = \sum_{i=1}^{K} J_i = \sum_{i=1}^{K} \sum_{x_i \in C_i} \left||x - m_i|\right|^2)$$

$$m_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$$

其中m_i是簇C_i的均值向量, J为平方误差和。

算法步骤：

1．初始化聚类中心$m_1, m_2, \dots, m_k$ 。

2．根据最小距离法则，把X划分到m_k为代表的C_k类中。

3．计算J，重新计算$m_1, m_2, \dots, m_k$ 。

4．重复 2、3，直到$m_1, m_2, \dots, m_k$ 不再变化，或者J不再减小。终止。

距离度量

类间距离

1. 单连锁：将两个簇中相距最近的两个点的距离作为两个簇的距离。
2. 全连锁：将两个簇中相距最远的两个点的距离作为两个簇的距离。
3. 平均连锁：将两个簇两两点距离的平均值作为这两个簇的距离。
4. Ward’s method：总是使得并类导致的类内误差平方和增量最小，但只能用于欧氏距离。

实验过程

大致过程：导入 20newsgroups 数据集，转为 DataFrame 格式，使用 nltk 里的停止词对数据进行清理，把清理后的文本数据转为 TF-IDF 向量，分别进行层

次聚类和 K 均值聚类。

数据集

导入数据集

使用 sklearn.datasets 里的 fetch_20newsgroups 进行数据集导入，但限于网络原因，使用的是本地下载好的数据集。

原本的数据集分有 20 个主题。

```
['alt.atheism',
 'comp.graphics',
 'comp.os.ms-windows.misc',
 'comp.sys.ibm.pc.hardware',
 'comp.sys.mac.hardware',
 'comp.windows.x',
 'misc.forsale',
 'rec.autos',
 'rec.motorcycles',
 'rec.sport.baseball',
 'rec.sport.hockey',
 'sci.crypt',
 'sci.electronics',
 'sci.med',
 'sci.space',
 'soc.religion.christian',
 'talk.politics.guns',
 'talk.politics.mideast',
 'talk.politics.misc',
 'talk.religion.misc']
```

选择其中的三个主题进行实验。

```
['comp.os.ms-windows.misc', 'rec.autos', 'sci.electronics']
```

转为 DataFrame

提取出其中的‘data’部分和‘target’部分，并重新命名为‘text’和‘category’，转为 DataFrame。0、1、2 分别为‘comp.os.ms-windows.misc’，‘rec.autos’，‘sci.electronics’的类标签。

	text	category
0	From: atom@netcom.com (Allen Tom)\nSubject: Re...	1
1	From: wtm@uhura.neoucom.edu (Bill Mayhew)\nSub...	2
2	From: corwin@igc.apc.org (Corwin Nichols)\nSub...	2
3	From: A.D.Bailey@lut.ac.uk\nSubject: Re: Utili...	0
4	From: mobasser@vu-vlsi.ee.vill.edu (Bijan Moba...	1
...	...	...
1771	From: dbd@icf.hrb.com (Douglas B. Dodson)\nSub...	0
1772	From: marshatt@feserve.cc.purdue.edu (Zauberer...	1
1773	From: hzhang@compstat.wharton.upenn.edu (Hao Z...	0
1774	From: ip02@ns1.cc.lehigh.edu (Danny Phornpraph...	1
1775	From: jeh@cmkrnl.com\nSubject: Electrical wiri...	2

1776 rows × 2 columns

停止词设置

使用 stopwords.words('english')获取英文停止词，list(string.printable)获取可打印字符，作为数据的筛选器。

```
['i',
 'me',
 'my',
 'myself',
 'we',
 'our',
 'ours',
 'ourselves',
 'you',
 'your',
 'yours',
 'yourself',
 'yourselves',
 'he',
 'him',
 'his',
 'himself',
 'she',
 'her',
 ...,
 '>',
 '?',
 '@',
 '[',
 '\\',
 ']',
 '^',
 '_',
 '`',
 '{',
 '|',
 '}',
 '~',
 ' ',
 '\t',
 '\n',
 '\r',
 '\x0b',
 '\x0c']
```

数据清洗

把不属于停止词的字符串提取出来，添加到新的列，列标签为‘cleaned_text’。

	text	category	cleaned_text
0	From: atom@netcom.com (Allen Tom)\nSubject: Re...	1	atom netcom com allen tom subject re dumb opti...
1	From: wtm@uhura.neoucom.edu (Bill Mayhew)\nSub...	2	wtm uhura neoucom edu bill mayhew subject re d...
2	From: corwin@igc.apc.org (Corwin Nichols)\nSub...	2	corwin igc apc org corwin nichols subject re f...
3	From: A.D.Bailey@lut.ac.uk\nSubject: Re: Utili...	0	bailey lut ac uk subject re utility updating w...
4	From: mobasser@vu-vlsi.ee.vill.edu (Bijan Moba...	1	mobasser vu vlsi ee vill edu bijan mobasseri s...
...	...	...	...
1771	From: dbd@icf.hrb.com (Douglas B. Dodson)\nSub...	0	dbd icf hrb com douglas dodson subject window ...
1772	From: marshatt@feserve.cc.purdue.edu (Zauberer...	1	marshatt feserve cc purdue edu zauberer subjec...
1773	From: hzhang@compstat.wharton.upenn.edu (Hao Z...	0	hzhang compstat wharton upenn edu hao zhang su...
1774	From: ip02@ns1.cc.lehigh.edu (Danny Phornpraph...	1	ip02 ns1 cc lehigh edu danny phornprapha subje...
1775	From: jeh@cmkrnl.com\nSubject: Electrical wiri...	2	jeh cmkrnl com subject electrical wiring faq q...

1776 rows × 3 columns

转为 TF-IDF 向量

对清理后的文本数据提取其中的 100 个 TF-IDF 特征，转为 TF-IDF 向量。

```
matrix([[0.        , 0.        , 0.        , ..., 0.        , 0.08423366,
         0.        ],
        [0.        , 0.        , 0.        , ..., 0.14888569, 0.        ,
         0.45667777],
        [0.        , 0.        , 0.        , ..., 0.        , 0.        ,
         0.        ],
        ...,
        [0.        , 0.        , 0.        , ..., 0.        , 0.        ,
         0.        ],
        [0.        , 0.        , 0.        , ..., 0.        , 0.        ,
         0.        ],
        [0.        , 0.        , 0.05576015, ..., 0.01875656, 0.        ,
         0.02876605]])
```

把得到的向量转为 DataFrame。

	0	1	2	3	4	5	6	7	8	9	...	90	91	
0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.00000	0.0
1	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.00000	0.0
2	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.00000	0.0
3	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.00000	0.0
4	0.0	0.0	0.399398	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.00000	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
1771	0.0	0.0	0.000000	0.139161	0.0	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.00000	0.0
1772	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.00000	0.4
1773	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.00000	0.0
1774	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	...	0.000000	0.00000	0.0
1775	0.0	0.0	0.055760	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	...	0.035639	0.06103	0.0

1776 rows × 100 columns

层次聚类

调用 scipy.cluster.hierarchy 里的 linkage()，计算欧氏距离，用 ward 方法使方差和最小化。使用 dendrogram()生成聚类树。绘制聚类树图案。

K-Means

调用 sklearn.cluster 里的 KMeans()，对向量数据进行计算，获取个样本所属的聚类中心‘obtained_clusters’。

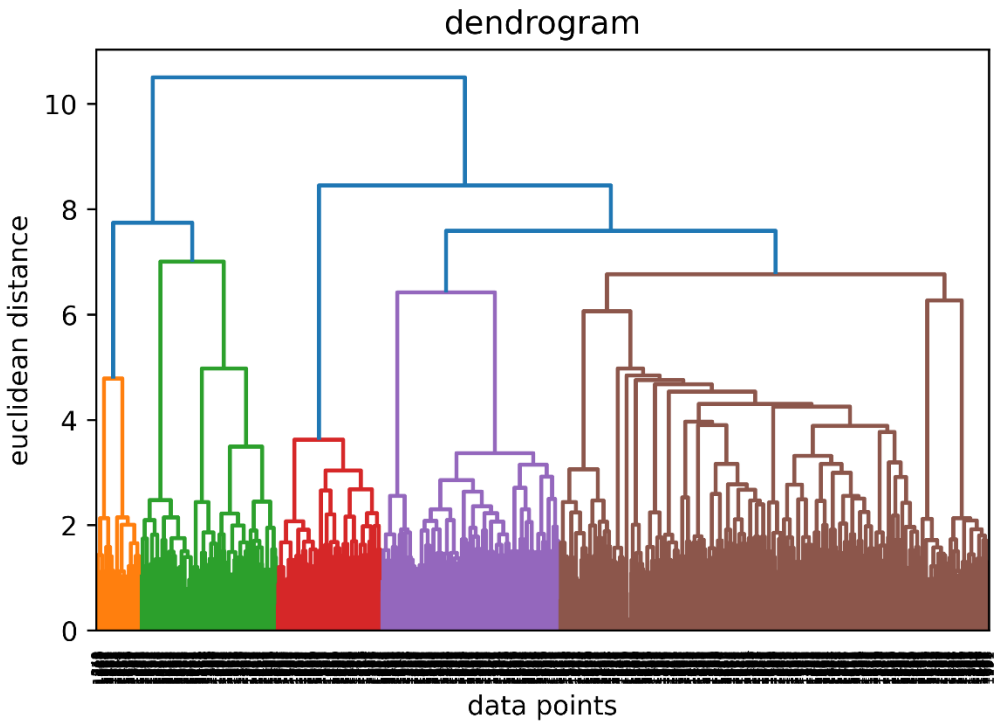
	text	category	cleaned_text	obtained_clusters
0	From: atom@netcom.com (Allen Tom)\nSubject: Re...	1	atom netcom com allen tom subject re dumb opti...	2
1	From: wtm@uhura.neoucom.edu (Bill Mayhew)\nSub...	2	wtm uhura neoucom edu bill mayhew subject re d...	1
2	From: corwin@igc.apc.org (Corwin Nichols)\nSub...	2	corwin igc apc org corwin nichols subject re f...	1
3	From: A.D.Bailey@lut.ac.uk\nSubject: Re: Utili...	0	bailey lut ac uk subject re utility updating w...	0
4	From: mobasser@vu-vlsi.ee.vill.edu (Bijan Moba...	1	mobasser vu vlsi ee vill edu bijan mobasseri s...	1
...	...	...	...	...
1771	From: dbd@icf.hrb.com (Douglas B. Dodson)\nSub...	0	dbd icf hrb com douglas dodson subject window ...	0
1772	From: marshatt@feserve.cc.purdue.edu (Zauberer...	1	marshatt feserve cc purdue edu zauberer subjec...	2
1773	From: hzhang@compstat.wharton.upenn.edu (Hao Z...	0	hzhang compstat wharton upenn edu hao zhang su...	1
1774	From: ip02@ns1.cc.lehigh.edu (Danny Phornpraph...	1	ip02 ns1 cc lehigh edu danny phornpraphа subje...	1
1775	From: jeh@cmkrnl.com\nSubject: Electrical wiri...	2	jeh cmkrnl com subject electrical wiring faq q...	1

1776 rows × 4 columns

使用交叉表显示出聚类结果。

实验结果 与分析

层次聚类



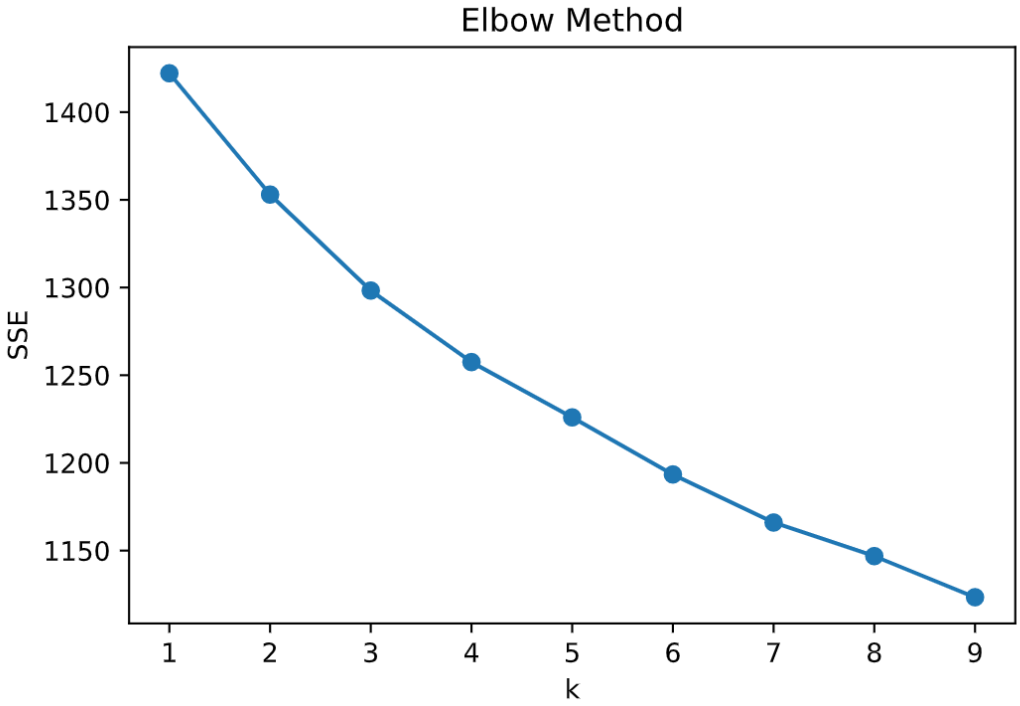
K-Means

obtained_clusters category	cluster1	cluster2	cluster3
comp.os.ms-windows.misc	43	165	383
rec.autos	367	223	4
sci.electronics	183	401	7

可以看出，第一个聚类簇里，大部分是属于‘rec.autos’这一主题的。第二个聚类簇里，大部分是属于‘sci.electronics’这个主题的。第三个聚类簇里，大部分是属于‘comp.os.ms-windows.misc’这个主题的。

在寻找最佳聚类数的时候，使用手肘法可以更快的定位到最适的 k 值。随着聚类数 k 的增大，样本划分会更加精细，每个簇的聚合程度会逐渐提高，那么误差平方和 SSE 自然会逐渐变小。当 k 小于真实聚类数时，由于 k 的增大会大幅增加每个簇的聚合程度，故 SSE 的下降幅度会很大，当 k 大于真实聚类簇数时，增加 k 所得到的聚合程度回报会迅速变小，所以 SSE 的下降幅度会骤减。可以

通过绘制 k 和 SSE 关系图，找出那个明显的拐折点，对应的 k 作为最佳聚类数。



可以看出，在 k=3 时，有拐点出现。选取 3 为最佳聚类数。

附录

20newsgroups.html