数学学院案例分析报告

案例名称：药物关系数据探索

报告参与人：魏丹怡（202032164）

完成时间：2021.06.01

目　录

一、 研究背景

1、研究背景

在人们日常的用药过程中，同时服用多种药物的现象随着人们年龄的增长变得越来越普遍。在美国青年人群（年龄处于 18 岁至 44 岁）中，超过三分之一的人会同时服用两种以上的药物，其中百分之四的人同时服用了五种以上的药物；在美国中年人群（年龄处于 45 岁至 64 岁）中，接近四分之三的人会同时服用两种以上的药物，这其中有五分之一的人同时服用了五种以上的药物；而年龄在 65 岁以上的老年人群中同时服用两种以上药物的人数占 90.9%且其中 42.2%的人会同时使用超过五种药物。

2、研究意义

药物相互作用是指病人同时或在一定时间内由先后服用两种或两种以上药物后所产生的复合效应，可使药效加强或副作用减轻，也可使药效减弱或出现不应有的毒副作用。为了减少药物不良反应的发生，我们需要减少药物间发生相互作用的可能性，主要的方式是在临床医生为患者开具多种药物时，首先在药物数据库中查询所开药物之间可能发生的相互作用信息，对可能发生相互作用的药物进行替换或调整。这样可以避免或减少药物间的相互作用对人体产生负面影响。

二、数据来源和基本情况

1、数据来源

SemEval-2013 Task 9 数据集由来自 DrugBank 数据库中的 730 个文档和 MEDLINE 数据库中 142 个文献摘要信息组成。整个数据集中涵盖了 33508 条涉及药物间相互作用关系的样本，数据集被官方划分为训练集和测试集两个部分。训练集由 572 篇来自 DrugBank 数据库中的文本和 142 篇来自 MEDLINE 数据库中的摘要信息组成，涵盖了 19476 个样本；测试集由 158 篇来自 DrugBank 数据库中的文本和 33 篇来自 MEDLINE 数据库中的摘要信息组成，涵盖了 14032 个样本。

2、基本情况

（1）数据描述

在训练集数据中，每条数据包括药物之间的相互作用关系、关系类别以及两种药物名称，其中药物关系类别被划分为 other、effect、mechanism、advise 和 int 五个类别。对后四个类别注解如下：

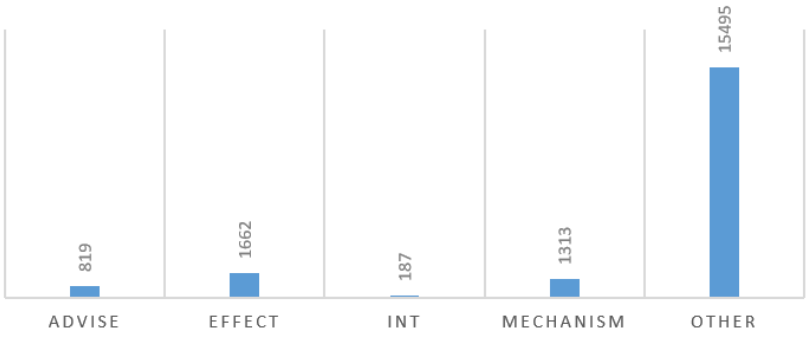
表 1　药物关系类别

类别	解释
effect	原始文本的描述中体现了药物实体之间的相互作用关系且描述了两种药物同时服用的影响或者药效机制。
mechanism	原始文本的描述中体现了药物实体之间的相互作用关系且描述了两种药物同时服用的药代动力学机制。
advice	原始文本的描述中体现了药物实体之间的相互作用关系且给出了在服用该两种药物时的用药建议。
int	原始文本的描述中体现了药物实体之间的相互作用关系，但没有做过多的额外描述。

（2）数据构成

	Label	DDI	名称1	名称2
count	19476	19476	19476	19476
unique	5	19021	2237	1959
top	other	- <e1> DRUG1 </e1> (e.g. , <e2> DRUG2 </e2>)...	alcohol	warfarin
freq	15495	20	207	257

（a）



（b）

```
alcohol                                207
phenytoin                              199
cimetidine                             172
antibiotics                            151
amiodarone                             148
                                      ...
pyrantel                                 1
streptomycin                             1
adrenergic beta-receptor blockers        1
thienobenzodiazepine derivative          1
antipsychotic medications                1
Name: 名称1, Length: 2237, dtype: int64
```

```
warfarin                257
phenytoin               226
theophylline            218
digoxin                 212
contraceptives          182
                       ...
ergocalciferol            1
Azulfidine                1
NORVIR                    1
interleukin-2             1
Furosemide                1
Name: 名称2, Length: 1959, dtype: int64
```

（c）

图 1　数据统计性描述

由上图可知，药物关系包括五种类型，其中出现频率最高的是 other 类，共有 15495 条数据；药物 1 包括 2237 种不同药物，出现频率最高的是 alcohol（酒精），为 207 次；药物 2 包括 1959 种不同药物，出现频率最高的是 warfarin（华法林），为 257 次。

三、数据关联性挖掘

1、关联性规则形式化描述

设$I=\{I_1,I_2,\ldots,I_m\}$是一个项集，m 为项的个数，其中I_i表示第 i 个项，对应于一个学生的数学题答题情况。事务t_i表示 I 的一个子集，对应于一个个订单。事务组成的集合记做 TID，每个事物中都包含若干个项。

关联性规则是形如$X \longrightarrow Y$的蕴含式，其中，X 和 Y 分别称为关联性规则的先导和后继。其中，关联规则 X Y，存在支持度和置信度，定义如下：

支持度$(X \longrightarrow Y)=\frac{\text{同时包含}X\text{和}Y\text{的事务数量}}{\text{所有事务数量}}$，理解为某一个项出现的概率。通常设置一个阈值 minsupport，当支持度不小于该值时认为是频繁项；

置信度$(X \longrightarrow Y)=\frac{\text{同时包含}X\text{和}Y\text{的事务数量}}{\text{包含}X\text{的事务数量}}$，理解为在 X 事务的基础上，X 和 Y 均出现的条件概率。

提升度$(X \longrightarrow Y)=\frac{\text{置信度}(X\longrightarrow Y)}{\text{包含}Y\text{的事务数量}}$，理解为规则的提升度的意义在于度量项集{X}和项集{Y}的独立性。

因此关联规则实际上包含两个子任务：

频繁模式发现：也称频繁模式挖掘、频繁项挖掘等，是指从一系列候选的项中选择频繁的部分，通常衡量频繁的程度可以是对每一项出现的频率，当超过某一阈值是则任务这个项是频繁的。

生成关联规则：在已经发现的最大频繁项目集中，寻找置信度不小于用户给定的 minconfidence 的关联规则。

2、Apriori 算法

Apriori 算法是经典的关联规则算法，其主要思路如下图所示：

图 2　Apriori 算法流程

算法流程：

（1）首先对数据库中进行一次扫描，统计每一个项出现的次数，形成候选 1-项集；

（2）根据 minsupport 阈值筛选出频繁 1-项集；

（3）将频繁 1-项集进行组合，形成候选 2-项集；

（4）对数据库进行第二次扫描，为每个候选 2-项集进行计数，并筛选出频繁 2-项集；

（5）重复上述流程，直到候选项集为空；

（6）根据生成的频繁项集，通过计算相应的置信度来生成管理规则。

Apriori 算法的特点：

简单且易于实现，是最具代表性的关联规则挖掘算法。随着数据集规模的不断增长，逐渐显现出一定的局限性：

（1）需多次扫描数据库，很大的 I/O 负载，算法的执行效率较低；

（2）产生大量的候选项目集，尤其是候选 2-项集占用内存非常大，会消耗大量的内存；

（3）对于每一趟扫描，只有当内存大小足够容纳需要进行计数的候选集时才能正确执行。如果内存不够大，要么使用一种空间复杂度更小的算法，要么只能对一个候选集进行多次扫描，否则将会出现“内存抖动”的情况，即在一趟扫描中页面频繁地移进移出内存（页面置换算法也无法避免内存抖动问题），造成运行时间的剧增。

3、实证研究

对药物关系数据集进行探索时，我们有必要考虑是否存在某种药物与其他药物之间存在固定的药物关系类别。为了探索药物与药物关系类别之间的潜在关系，我们通过调用 mlxtend.frequent_patterns 库中的 Apriori 包进行关联性分析。首先利用 Apriori 找出频繁项集，即出现频率较高的药物与药物关系类别的组合（支持度＞0.0005）；在频繁项集的基础上，使用关联规则算法找出其中物品的关联结果。

	support	itemsets
0	0.000513	(1,3-difluoroacetone)
1	0.000822	(18-MC)
2	0.000513	(3H-spiroperidol)
3	0.001643	(ACE inhibitors)
4	0.000667	(AEDs)
...	...	...
939	0.002516	(zidovudine, other)
940	0.001335	(zileuton, other)
941	0.001438	(zinc, other)
942	0.002208	(ziprasidone, other)
943	0.002259	(zonisamide, other)

1928 rows × 2 columns

从中筛选形式为药物与药物关系组合的项，并按照出现频率进行排序，结果如下所示：

	itemsets	support
0	frozenset({'phenytoin', 'other'})	0.021000
1	frozenset({'warfarin', 'other'})	0.017303
2	frozenset({'theophylline', 'other'})	0.014223
3	frozenset({'alcohol', 'other'})	0.013761
4	frozenset({'digoxin', 'other'})	0.013401
...	...	...
662	frozenset({'imidazoles', 'other'})	0.000513
663	frozenset({'mechanism', 'lomefloxacin'})	0.000513
664	frozenset({'phenothiazine derivatives', 'other'})	0.000513
665	frozenset({'progestogens', 'other'})	0.000513
666	frozenset({'sulfasalazine', 'other'})	0.000513

667 rows × 2 columns

其中，出现频率排名前 20 的药物与药物关系组合如下表所示：

表 2　高频率组合排名

itemsets	support	itemsets	support
phenytoin, other	0.021000	carbamazepine, other	0.00991
warfarin, other	0.017303	acetaminophen, other	0.009602
theophylline, other	0.014223	rifampin, other	0.008985
alcohol, other	0.013761	phenobarbital, other	0.008883
digoxin, other	0.013401	antihistamines, other	0.008113
cimetidine, other	0.013247	anesthetics, other	0.00801
contraceptives, other	0.012990	lithium, other	0.007804
amiodarone, other	0.010937	corticosteroids, other	0.007753
quinidine, other	0.010885	erythromycin, other	0.007702
antibiotics, other	0.010115	aspirin, other	0.007548

此外，上表中 20 种组合的出现频率的比例如下图所示：

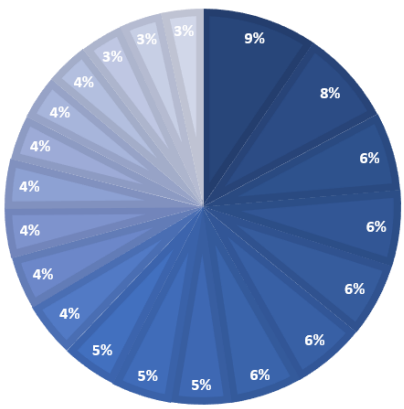


图 3　组合出现频率占比（前 20）

4、结论

我所进行探索的数据集中涵盖2698种不同药物及5种不同的药物关系类别,在Apriori算法种，可以形成的事务数量为16193个，由上述数据关联性探索可知，在不同种类药物和药物关系类别之间存在着一定联系。

其中，出现频率最高的药物与药物关系组合为（phenytoin,other），其支持度为0.021000206，则可以认为，phenytoin（苯妥英）这一药物与其他大多数药物之间的关系类别为“other”;同理,(warfarin,other)、(theophylline,other)、(alcohol,other)、(digoxin,other)、(cimetidine,other）和（contraceptives,other）这六个组合的支持度也相对较高，分别为0.017303348、0.014222633、0.013760526、0.01340111、0.013247074和0.012990346，也就是说，warfarin（华法林）、theophylline（茶碱）、alcohol（酒精）、digoxin（地高辛）、cimetidine（西米替丁）、contraceptives（避孕药）这六种药物与其他大多数药物的关系类别也为“other”。但由于数据集中，药物关系类别为“other”的样本占比过大，上述结论仅可以作为参考，在实际应用中，还需要进一步判断。

四、关于DDI实例数据探索

由于最后需要进行的目的是用过DDI描述对于药物的作用进行划分,所以通过先对其进行数据探索，并将发现的结果运用到分类之中。这里做的是针对于各个药物关系进行词频统计，发现一些在各个关系中具有独立性的高频词汇，这样在后续分类中可以将其这些词作为重点，赋予一定权重，使得分类的效果更好。

1、数据清洗

由于所给的数据中存在″<e1> DRUG1 </e1> ″、″<e2> DRUG2 </e2> ″以及″DRUG0 ″等一些不需要的高频出现的词，利用python中的replace函数将其替换掉，并利用re库中的compile以及sub函数对于一些不需要的标点符号进行处理。同时利用lower函数对于所有大写字母小写化，这样可以保证在后续去停用词以及词频统计时，不会因为其置于句首大写而将其看作是两个不同的词。

最后利用nltk库中的停用词库对于本次数据集的停用词进行词频统计，并画出热门停用词图，如下图1所示。可以看出其中the的频率最高有16000，其次是of，and等无意义的连接词。

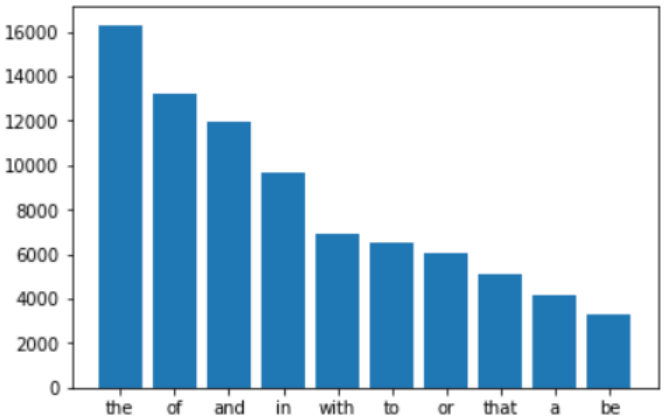


图 4　热门停用词

2、初步词频探索

主要思想是利用上述的 nltk 库中的停用词库，对于不在停用词库中的词语的频数进行统计，同时将数据依据药物效果不同进行分类，并对每一类分别进行词频统计，并选出出现频率最高的 21 个词，统计结果如下图所示：

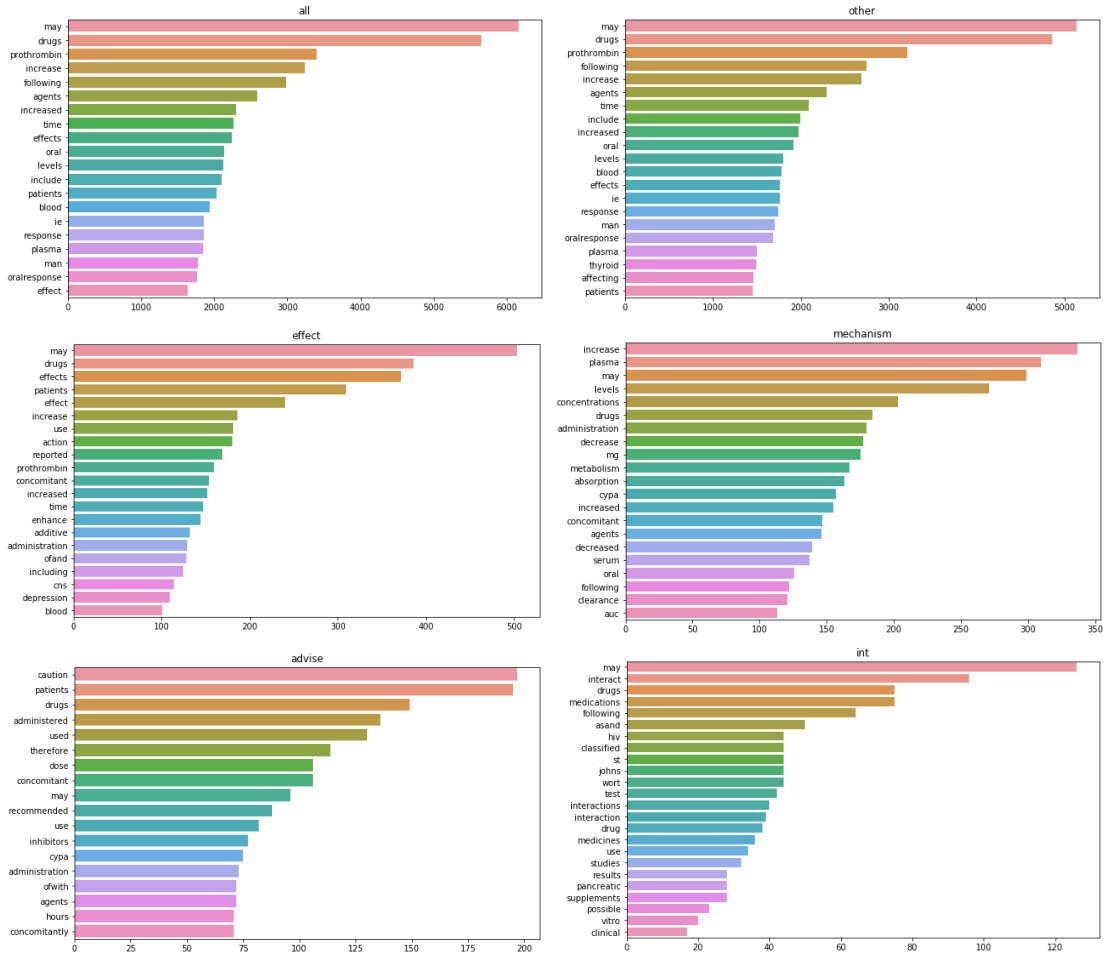


图 5　初步词频探索

由上图可以看出，对于所有的 DDI 描述文本进行统计发现“may”，“drug”出现频率最多大约五千到六千左右，其次是“prothrombin”（凝血酶原），“increase”、“following”等词。但并不能因为他们出现的频率高而武断的将其加入停用词中，因为要时刻明白这五种标签的样本数是极不平衡的，很可能这些词也是某个标签的特征词，只是因为该标签样本数多导致该词出现频率较高。

因此还需观察各个类别词频，如若不同种类的效果说明中都有同一个词出现的频率高，则需要将其放置入停用词库中。经过统计可以发现其中“drugs”和“may”每个标签都出现了并且其在“int”、“effect”以及“other”类中都占比极高，所以我应在停用词中加上这两个单词。我还注意到其中还有 73 个词至少两个标签中共同出现，但我考虑到可能有些词在特定一个标签中频繁出现，而在其他集合中出现频率并不算高，这样该词仍能表示这个集合，所以我暂不将这些词计入停用词中。

3、词频统计

通过第二节加入停用词后的各个标签下的词频进行探索，可以得到新的关于各个标签下词频统计图。这里不再展示完整图，只针对其中几个标签进行结果展示和说明。

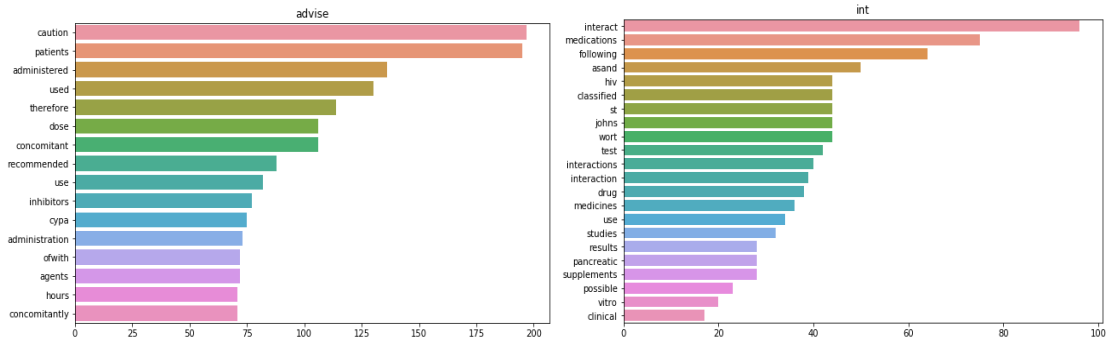


图 6　词频统计

由上图可以看出“advise”类别下“caution”以及“patients”这两个单词出现的频率最高均为 200 次左右，由于给出用药建议时需要患者谨慎用药，所以这两个单词在这里出现频率最高也十分合理。同时在“int”类别下，“interact”出现频率最高，因为“int”类别就是在探讨两个要药物间的相互作用，所以也非常合理。通过上述药物 DDI 文本数据词频探索可以在后续分类中给这些高频词加入权重从而提高分类的准确性。

五、数据不平衡问题

数据不平衡也可称作数据倾斜。在实际应用中，数据集的样本特别是分类问题上，不同标签的样本比例很可能是不均衡的。因此，如果直接使用算法训练进行分类，训练效果可能会很差。在本次实验所使用的训练集和测试集中，“other”类的频数很大，与其他类别的样本比例相差甚远，数据存在不均衡问题。类别构成如下所示：

```
other         15495
effect         1662
mechanism      1313
advise          819
int             187
Name: Label, dtype: int64
```

```
other         2925
effect         358
mechanism      298
advise         221
int             96
Name: Label, dtype: int64
```

图 7　类别构成

为了解决这一问题，我使用了下采样、上采样及上下采样相结合这三种方法，具体情况如下。

1、下采样

下采样则是从多数量的类别中随机抽取样本（抽取的样本数量与少数类别样本量一致）从而减少多数量的类别样本数据，使数据达到平衡的方式。缺点是会丢失多数类中的一些重要信息。

可以从图 7 可以看出，“int”类是频数最小的，训练集和测试集的频数分别为 187、96。利用下采样的方法将其他类别的频数都减少至与“int”类别的频数一致，从而达到数据平衡。训练集和测试集的下采样数据如下：

Out[69]:

	Label	DDI	名称1	名称2
1320	other	We investigated the effects of adenosine rece...	8-phenyltheophylline	DPCPX
8683	other	DRUG0 (CNS) drugs including DRUG0 , DRUG0 , <...	antihistamines	reserpine
5282	other	DRUG0 may interact with DRUG0 , DRUG0 s , DRU...	lithium	Aleve
2275	other	DRUG0 : DRUG0 should be used with caution in ...	receptor	propranolol
5719	other	- <e1> DRUG1 </e1> : DRUG0 blunts the increas...	Indomethacin	bumetanide
...	...	...	...	...
18085	int	The in vitro interaction between <e1> DRUG1 <...	nevirapine	warfarin
18151	int	DRUG0 : Immediate Release Capsules : Since th...	digoxin	nifedipine
18660	int	<e1> DRUG1 </e1> preparations are incompatibl...	Sulfacetamide	silver
18661	int	<e1> DRUG1 </e1> may interact with any of the...	Sulfapyridine	Acetaminophen
18662	int	<e1> DRUG1 </e1> may interact with any of the...	Sulfapyridine	Tylenol

935 rows × 4 columns

Out[105]:

	Label	DDI	名称1	名称2
3708	other	The use of DRUG0 , in combination with <e1> D...	heparin	aspirin
2688	other	While no formal drug interaction studies have...	acetylsalicylic acid	prednisone
162	other	Systemic and apparent oral DRUG0 clearance we...	cyclosporine	tacrolimus
1606	other	Although specific drug or food interactions w...	mifepristone	itraconazole
2534	other	A multiple dose drug-drug interaction study d...	ketoconazole	atazanavir
...	...	...	...	...
3801	int	<e1> DRUG1 </e1> may interact with DRUG0 or <...	Trilostane	mitotane
3870	int	<e1> DRUG1 </e1> can interact with the drugs ...	Vindesine	Phenytoin
3871	int	<e1> DRUG1 </e1> can interact with the drugs ...	Vindesine	Live virus vaccines
3873	int	<e1> DRUG1 </e1> can interact with the drugs ...	Vindesine	Mitomycin-C
3874	int	<e1> DRUG1 </e1> can interact with the drugs ...	Vindesine	Killed virus vaccines

480 rows × 4 columns

训练集和测试集中的“int”类别的频数分别为 187 和 96，故经过下采样后，数据量较少，训练集和测试集分别有 935 和 480 条数据。

2、上采样

上采样则是从少数量的类别中随机抽取样本（抽取的样本数量与多数类别样本量一致，或着增加一定的比例，可控制）从而增多少数量的类别样本数据，使数据达到平衡的方式。

可以从图7可以看出,“other类”的频数是最大的,训练集和测试集的频数分别为15495、2925。利用上采样的方法将其他类别的频数都增加至与“other”类别的频数一致，从而达到数据平衡。训练集和测试集的下采样数据如下：

Out[178]:

	Label	DDI	名称1	名称2
0	other	The <e1> DRUG1 </e1> are a rapidly growing cl...	fluoroquinolones	antibiotics
1	other	These agents , including <e1> DRUG1 </e1> , D...	norfloxacin	lomefloxacin
2	other	These agents , including DRUG0 , <e1> DRUG1 <...	ciprofloxacin	lomefloxacin
3	other	These agents , including DRUG0 , DRUG0 , <e1>...	ofloxacin	lomefloxacin
4	other	These agents , including DRUG0 , DRUG0 , DRUG...	enoxacin	lomefloxacin
...	...	...	...	...
14382	int	A possible interaction between <e1> DRUG1 </e...	glyburide	ciprofloxacin
2085	int	Data from in vitro studies of DRUG0 suggest a...	alprazolam	paroxetine
2091	int	Data from in vitro studies of <e1> DRUG1 </e1...	benzodiazepines	nicardipine
12180	int	<e1> DRUG1 </e1> may interact with the follow...	Etonogestrel	Theo-Dur
11606	int	There have been reports of interactions of <e...	erythromycin	cyclosporine

77475 rows × 4 columns

Out[216]:

	Label	DDI	名称1	名称2
0	other	In the present study , we tested whether the ...	3-[(2-methyl-1,3-thiazol-4-yl) ethynyl] pyridine	1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine
1	other	Weekly intramuscular <e1> DRUG1 </e1> injecti...	1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine	3-[(2-methyl-1,3-thiazol-4-yl) ethynyl] pyridine
2	other	Weekly intramuscular <e1> DRUG1 </e1> injecti...	1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine	3-[(2-methyl-1,3-thiazol-4-yl) ethynyl] pyridine
3	other	Weekly intramuscular DRUG0 injections (0.2-0....	3-[(2-methyl-1,3-thiazol-4-yl) ethynyl] pyridine	1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine
4	other	Weekly intramuscular DRUG0 injections (0.2-0....	3-[(2-methyl-1,3-thiazol-4-yl) ethynyl] pyridine	1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine
...	...	...	...	...
1701	int	Other drugs which may enhance the neuromuscul...	MIVACRON	sodium colistimethate
1133	int	<e1> DRUG1 </e1> may interact with DRUG0 (DR...	Methscopolamine	potassium chloride
3629	int	Interactions for <e1> DRUG1 </e1> (DRUG0) :...	Vitamin B1	Contraceptives
567	int	<e1> DRUG1 </e1> may interact with the follow...	Melatonin	progestin
566	int	<e1> DRUG1 </e1> may interact with the follow...	Melatonin	fluoxetine

14625 rows × 4 columns

训练集和测试集中的“other”类别的频数分别为 15495 和 2925，故经过上采样后，数据量增大，训练集和测试集分别有 77475 和 14625 条数据。

3、上、下采样结合

当正负样本差异非常极端的时候，上、下采样一起结合，少数量的样本重采样（复制）一定比例，多数量的类别样本随机减少一定比例样本，让两者达到平衡。

利用上下采样结合的方式，将“other”类的数据减少，其他 4 类的数据增加，5 类样

本数据的比例设置为 1:1:1:1:1，训练集和测试集中每个类别的数据分别有 4648、877 条。训练集和测试集的下采样数据如下：

Out[311]:

	Label	DDI	名称1	名称2
2843	effect	Drugs that reportedly may increase oral <e1> ...	anticoagulant	chloral hydrate
11220	effect	However , reports suggest that <e1> DRUG1 </e...	NSAIDs	ACE inhibitors
17850	effect	Certain drugs , including DRUG0 (<e1> DRUG1 ...	NSAIDs	antidiabetic drugs
11809	effect	The action of the <e1> DRUG1 </e1> may be pot...	benzodiazepines	phenothiazines
18174	effect	In Europe , <e1> DRUG1 </e1> was observed to ...	Nimotop	antihypertensive compounds
...	...	...	...	...
8033	other	DRUG0 / <e1> DRUG1 </e1> : The coadministrati...	Phenobarbital	calcitriol
3045	other	Drugs that reportedly may increase oral DRUG0...	aminosalicylic acid	warfarin sodium
8444	other	<e1> DRUG1 </e1> : In vitro and/or in vivo da...	Antibiotics	erythromycin
12866	other	The ratios of the AUCs of unbound DRUG0 to th...	valproate	Felbatol
9309	other	The use of <e1> DRUG1 </e1> should be conside...	antacids	SPRYCEL

23240 rows × 4 columns

Out[325]:

	Label	DDI	名称1	名称2
3048	effect	In addition to bleeding associated with DRUG0...	aspirin	Retavase
823	effect	<e1> DRUG1 </e1> antagonize the effects of <e...	Anticholinergics	antiglaucoma agents
2896	effect	DRUG0 : Concurrent use of <e1> DRUG1 </e1> an...	procaine hydrochloride	anticholinesterase agents
965	effect	<e1> DRUG1 </e1> may increase the responsiven...	Thiazide drugs	tubocurarine
1530	effect	The sedative effect of <e1> DRUG1 </e1> is ac...	VERSED Syrup	meperidine
...	...	...	...	...
1766	other	Other drugs which may enhance the neuromuscul...	magnesium	lithium
2122	other	DRUG0 : In 12 normal-weight subjects receivin...	orlistat	glyburide
2926	other	Because DRUG0 exhibits some monoamine oxidase...	sympathomimetic drugs	amitriptyline HCl
1635	other	Absorption of DRUG0 is impaired by <e1> DRUG1...	antacids	iron
3373	other	It may also interact with DRUG0 (increased th...	warfarin	phenytoin

4385 rows × 4 columns

训练集和测试集经过上下采样后，数据量分别为 23240 和 4385。

4、结果分析

为了测试我的方法对数据分类是否有益，将处理后的数据带入朴素贝叶斯模型进行预测并分析分类结果。

（1）原始数据

	precision	recall	f1-score	support
advise	0.60	0.05	0.10	221
effect	0.44	0.07	0.12	358
int	0.00	0.00	0.00	96
mechanism	0.00	0.00	0.00	298
other	0.76	0.99	0.86	2925
avg / total	0.64	0.75	0.66	3898

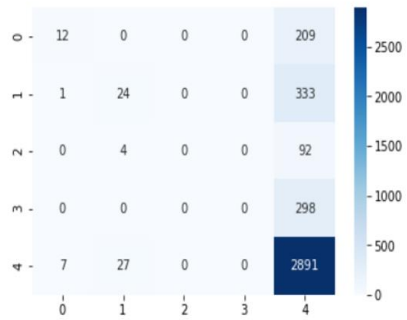
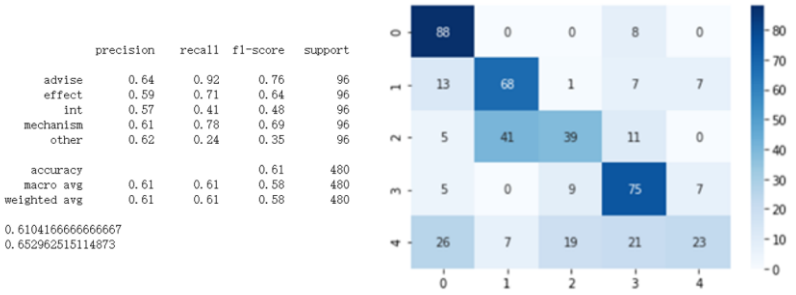


图 8　原始数据分类结果

由上图可以看出，采用原始数据通过朴素贝叶斯模型进行预测，得到的准确率为 66%，其中“other”类准确率为 76%。由于在原始数据中，“other”类所占比例很大，只要将药物分到“other”类中，就可能会分对，从混淆矩阵可以看出，其他 4 类分类正确的很少，大多数都误分到“other”类中。但是由于“other”类样本价值很小，因此这一分类结果意义不大。

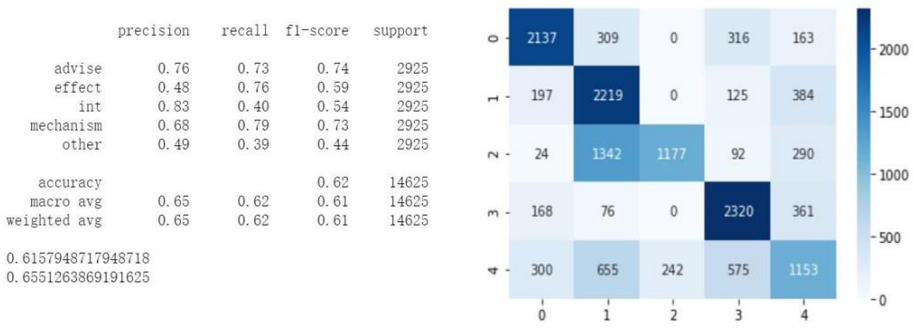
（2）处理后数据



	precision	recall	f1-score	support
advise	0.64	0.92	0.76	96
effect	0.59	0.71	0.64	96
int	0.57	0.41	0.48	96
mechanism	0.61	0.78	0.69	96
other	0.62	0.24	0.35	96
accuracy			0.61	480
macro avg	0.61	0.61	0.58	480
weighted avg	0.61	0.61	0.58	480

0.6104166666666667
0.652962515114873

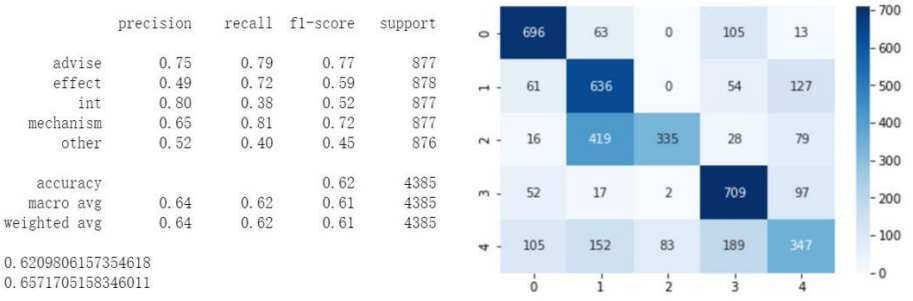
（a）下采样数据分类结果



	precision	recall	f1-score	support
advise	0.76	0.73	0.74	2925
effect	0.48	0.76	0.59	2925
int	0.83	0.40	0.54	2925
mechanism	0.68	0.79	0.73	2925
other	0.49	0.39	0.44	2925
accuracy			0.62	14625
macro avg	0.65	0.62	0.61	14625
weighted avg	0.65	0.62	0.61	14625

0.6157948717948718
0.6551263869191625

（b）上采样数据分类结果



	precision	recall	f1-score	support
advise	0.75	0.79	0.77	877
effect	0.49	0.72	0.59	878
int	0.80	0.38	0.52	877
mechanism	0.65	0.81	0.72	877
other	0.52	0.40	0.45	876
accuracy			0.62	4385
macro avg	0.64	0.62	0.61	4385
weighted avg	0.64	0.62	0.61	4385

0.6209806157354618
0.6571705158346011

（c）上下采样结合数据分类结果

图 9　处理后数据分类结果

由上图可以看出，使用下采样、上采样和上下采样后的的数据通过朴素贝叶斯模型进行预测，得到的准确率分别为 61%、62%、62%。相比于使用全部数据的结果，下采样、上采样、上下采样后的准确率略微降低。但是除去“other”类的其余四类分类准确率都在增加，并

不是一味地误分到“other”类中。相较于使用全部数据得到的结果，虽然总体准确率降低了，但是可以将药物分到各个类别中，使得分类会更加有意义。

六、总结

在本次实验中，首先进行了数据关联性挖掘，利用 Apriori 算法找到出现频率较高的药物与药物关系类别的组合，推断出了药物与药物关系类别之间存在的一定联系。

其次利用 python 进行药物解释文本数据高频词探索，得出每个标签中出现的高频词可以帮助们可以在后续分类中给这些高频词加入权重从而提高分类的准确性。最后，针对数据中存在的不平衡问题，通过下采样、上采样、上下采样三种方法解决此问题，并利用朴素贝叶斯模型对数据集进行预测。比较了处理前后共四组数据的分类结果，我认为在解决了数据不平衡问题后，得到的分类结果更有意义。

本次实验的不足在于，在词频探索中，有些高频词并不只有一个名词组成，还有可能是一个搭配或者一个短语，所以在后续研究中我可以利用 python 强大的自然语言处理功能，对于两个或多个单词的高频词进行探索，从而为我后续研究做好准备。当进行上下采样相结合的方法时，默认将训练集、测试集中所有类别的数据个数统一为 4648 条和 877 条，此做法是因为这一数值基本接近于样本总数除以类别数，但没有考虑若将数据个数设置为其他数值，是否会出现更好的结果，在之后的实验中，可以进行尝试。