基于Yelp评论数据的多标签分类问题

刘阳

复旦大学

13307130167@fudan.edu.cn

June 14, 2016

I. 摘要

这份报告基于Yelp提供的数据集（关注其中的餐饮评论数据），通过对文字评论的语义分析，提取出有关特征，构造多标签分类器来挖掘文字评论中对于餐饮服务的各方面的二值评价。报告中介绍了多标签分类问题(Multi-label classification)，比较了不同分类器的分类结果，并将分类器应用于实际的文字评论数据来观察几个餐厅的评论走势。

II. 简介

在互联网时代，人们的生活越来越多地受到网络信息的影响，所做的决策也或多或少的被网络的评论和观点所左右。在这样的背景下，孕育了类似大众点评和Yelp这种服务行业推荐和点评的网站。在这类服务点评网站上，一项服务的评价通常分为两类，一类是数值评分，另一类是文字评价。然而，我们也经常会碰到这样的情况：单一的数值评分并不能完整地表达一项服务各方面的好坏，而大部分时候顾客又没有足够的时间去浏览详细的文字评价来进行自我判断。选择的依据只有评分而文字评价却很少发挥作用。这种情况下，原本综合评分较高的服务，可能因为某方面的原因而让顾客的体验大打折扣。一个典型的例子就是国内的某些百年老店，做出来的食物美味可口，点评的评分很高，但是可能环境卫生做得差一些，导致有些顾客的就餐体验不佳。

针对这个问题，点评网站也相应的采取了措施，比如大众点评除了有综合评分，还针对“口味”，“环境”，“服务”这三个方面单独设置了评分项，如图1所示。

相对于显式的让用户来对服务的各个方面评分，这份数据挖掘项目在于构建一个分类器，能够通过对顾客文字评价的分析和挖掘，来判断该顾客对服务某些方面的评分。

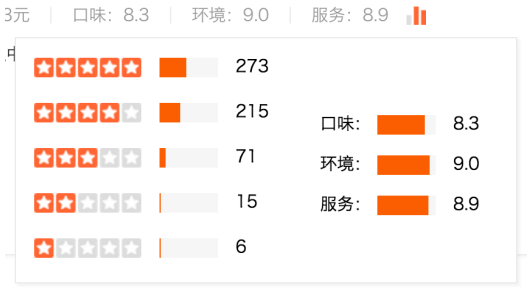

Figure 1: 大众点评评分数据

这个分类项目的应用前景在于不仅能够从文字评价中提取有用的信息为其他顾客提供直观的评分参考，而且也能为服务提供者有针对性地改进服务的某些方面提供参考。

本项目基于Yelp提供的数据[1]。报告的第三部分首先探索提供的数据，有一个直观的理解和感受。第四部分详细阐述研究的问题。第五部分针对分类问题清洗提供的数据，产生所需的训练集和测试集。第六部分解释多标签分类问题和采取的算法。第七部分进行不同分类算法的结果比较，以及最佳分类器在新的文字评论集上的应用。最后为结论和参考文献部分。

III. 数据探索

Yelp提供的全部数据集包括：business, user, review, tip。各个子数据集的数据量如下表1所示：

由于项目感兴趣的是顾客对于服务的文字评论，所以用到的数据集包括“*Business*"，“*Review*"。

Table 1: 子数据集数据量一览

数据集	Business	User	Review	Tip
数据量	77445	552339	2225213	591864

i. Business数据集

*Business*数据集包含的数据段有：business_id, full_address, hours, categories, city, review_count, name, neighborhood, longitude, latitude, stars, attribute等。
一条典型的商户数据如下（只包含感兴趣的数据段）：
{"business_id": "5UmKMjUEUNdYWqANhGckJw",
"categories": ["Fast Food", "Restaurants"],
"review_count": 4,}
通过对商户数据集的统计，发现一共有892个分类类别。对于所有分类进行商户计数，得到商户最多的前10个分类如下图2所示。

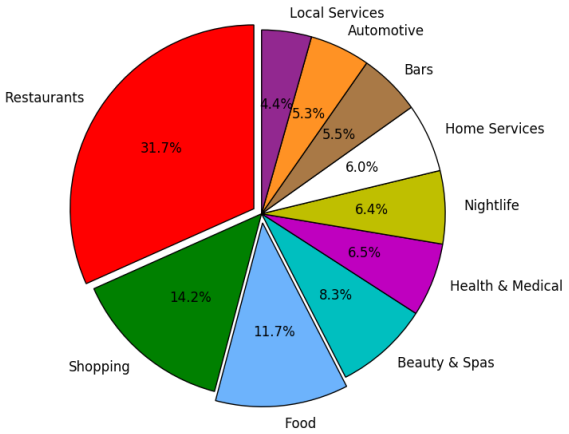

Figure 2: 商户数最多的前*10*个分类分布

从图2中可以看到Yelp的*Business*数据集主要集中在餐饮行业（Restaurant，Food分类），占到了前10个分类总数的31.7% + 11.7% = 43.4%，将近一半的数据量。
另外通过统计前10个分类下平均每个商户收到的文字评论数，可以得到如下图3。
从每个商户获得评论数的分布可以看出，餐饮行业不仅占据了商户的大部分，而且也是顾客参与评论最多的分类之一。利用数据集的这个特点，我们将多标签分类问题主要集中在餐饮行业的文字评论上。

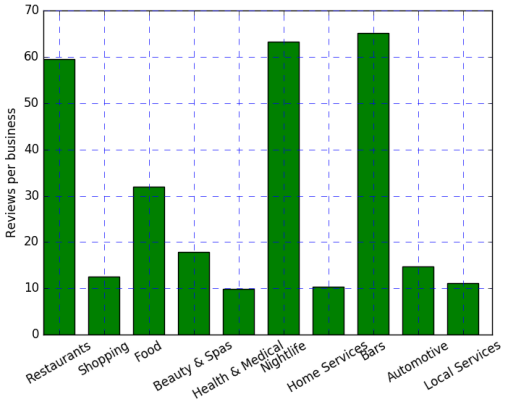

Figure 3: 前*10*个分类的每个商户获得评论数分布

ii. Review数据集

*Review*数据集包含的数据段有：votes, user_id, review_id, stars, date, text, type, business_id。
一条典型的评论数据如下（只包含感兴趣的数据段）：
{"stars": 5,
"date": "2014-02-13",
"text": "Excellent food. Superb customer service. I miss the mario machines they used to have, but it's still a great place steeped in tradition.",
"business_id": "5UmKMjUEUNdYWqANhGckJw"}
一条评论数据包含了数值评分*stars*，文字评价*text*。从上面这条评论可以看出顾客对于这家餐厅的评价是"口味好：*Excellent food*"，"服务好：*Superb customer service*"。这些信息正是我们的分类器想要提取出来的，辅之以数值评分，可以帮助顾客快速且准确地判断一家餐厅是否合意。

IV. 研究问题描述

基于对数据集的探索，本项目的目的在于从评论的文本中提取出有用信息，配合数值评分，来判断该评论对餐厅各方面的好坏评价。
具体来讲，我们将评价点限定为"口味"，"服务"，"氛围"，"折扣"，"性价比"。将文字评

论总结为这五个方面的好坏，可以方便顾客快速把握评论大意而不用逐一浏览评论。
同时我们注意到有别于通常的分类问题，这里的评论分类结果可以得到多个分类标签。例如上述*Review*数据集中的样例数据，既有“口味好”，又有“服务好”。所以我们将这个分类问题实现为多标签分类问题。

V. 数据清洗

有了明确的分类目标，我们开始从文字评论中提取分类特征，并将所有特征标准化(即所有特征都是二值特征，表示在某条评论有没有出现过)。

i. 特征提取

首先利用的是评论数据的数值评分，我们将数值评分分为三个档次，如下表2所示:

Table 2: 数值评分分档表示

数值评分	4-5 Stars	3 Stars	1-2 Stars
对应值	Good	Moderate	Bad

其次利用NLTK包的分词功能，提取出的文本特征[2]包含评论中的unigram, bigram, trigram。利用wordcloud包制作出分词的词云，用来直观观察文本特征。以下图4， 图5， 图6分别为unigram, bigram, trigram 的词云，词的大小代表出现频率，各挑选了出现频率最高的40个词。

Figure 4: *unigram*分词结果

从分词结果可以看出，大部分特征都是关于“口味”的，如“food”， “drink”， “the food is"。有部分特征关于“氛围”，如“happy hour”，部分特征关于“服务”，如“very friendly", "the service was"。

Figure 5: *bigram*分词结果

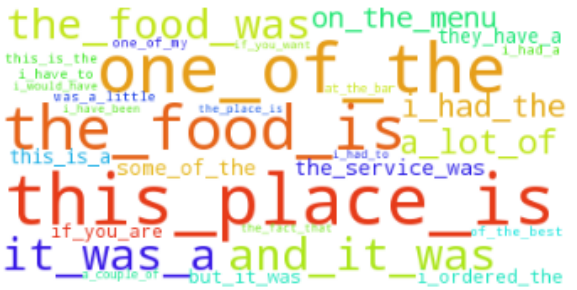

Figure 6: *trigram*分词结果

评论态度方面，词云中可以看出大部分都是积极的，如“great", "like", "very good", "go back"。
从分词结果中有一个直观的感受是trigram的特征没有明显的作用，这个观点将在之后的分类器评价阶段去检查。

ii. 标准化

从评论数据中共提取了671个特征，其中包括668个ngram特征(341个unigram, 208个bigram, 119个trigram)，3个数值评分特征(“IsRatingGood", "IsRatingModerate", "IsRatingBad")。
结果标签集为”IsFoodGood", "IsServiceGood", "IsAmbianceGood", "IsDealsGood", "IsPriceGood", 分别用来标识“口味”，“服务”，“氛围”，“折扣”，“性价比”的好坏。训练集中这5个标签的分布情况如图7所示。
标准化的结果是将每条餐饮评论数据构造成含有671个分类特征和5个结果标签的数据，每个特征对应的值若出现在评论数据中，则用1标识，否则用0标识。这样构造出来的数据特征都

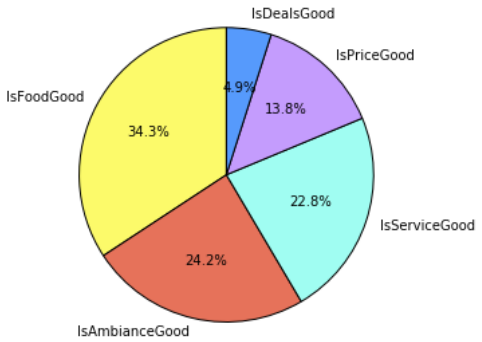

Figure 7: 训练集中标签分布图

是二值的。
从评论数据集中一共提取出了10,806条有效的数据，其中按照80%,20%的划分原则划分出8,846条数据作为分类训练集，1,960条数据作为分类测试集。

VI. 多标签分类问题

由于顾客的评论会涉及多方面的评价，所以上述将一条文字评论分类到多个评价点的问题是一个多标签分类问题。
通常的分类问题的输出是标量，而多标签分类问题形式化的定义是找到一个分类模型C，使得$C(\mathbf{X}) \rightarrow \mathbf{y}$, 其中 $\mathbf{X}$ 是输入的特征向量，$\mathbf{y}$ 是得到的标签向量。
有两种主要的途径可以用来解决多标签分类问题[3]，分别是问题转换方法(Problem transformation methods)，算法改编方法(algorithm adaptation methods)。问题转换方法将多标签问题转换成一系列的二元分类问题，然后使用二元分类算法解决。而算法改编方法则是改编分类算法来直接对多标签分类。
本项目主要采用的是问题转换方法。

i. 问题转换方法

多标签分类问题存在多种问题转换方法：基线方法[4](Baseline approach，又名二元关联方法(Binary relevance method))，标签幂集方法[3](Label power set) 。

二元关联方法

二元关联方法是最广泛使用的问题转换方法。二元关联方法将每个标签的分类转换成独立的二元分类任务。令结果标签集为L，对于每一个标签 $l_i \in L$，构造一个二元分类器 $C_i : C_i(\mathbf{X}) \rightarrow \{True, False\}$。由于要单独训练每一个二元分类器，所以我们将训练集和测试集都转换成$|L|$ 个训练集$\{Train_1, \ldots, Train_{|L|}\}$和测试集$\{Test_1, \ldots, Test_{|L|}\}$。针对每一个 l_i，若源数据集中出现该标签，而标识l_i为1，否则标识l_i为0。数据集转换的过程如图8所示。

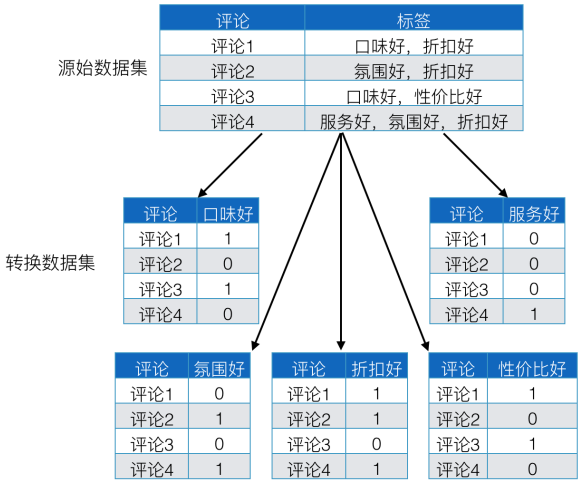

Figure 8: 二元关联方法数据转换示意

利用转换过后的训练集$Train_i$训练分类器C_i，最后的多标签分类结果由各个分类器C_i 汇总得到。如果$C_i(\mathbf{X}) = 1$, 则结果中$l_i = 1$，表示含有此标签，分类结果为 $\cup l_i, l_i \in L$。

标签幂集方法

标签幂集方法是另一种不太常用的问题转换方法。它将标签集合L的每个子集(除去空集)认为是一个单独的类，从而将多标签分类问题转换成多元分类问题。令标签集的幂集(除去空集)为$P(L)$，构造一个多元分类器 $C : C(\mathbf{X}) \rightarrow P(L)$。同二元关联方法一样，标签幂集方法也需要进行数据转换。将数据集中每条数据的标签转换成一个标签串，构成分类类别集合。具体过程示意如图9所示。

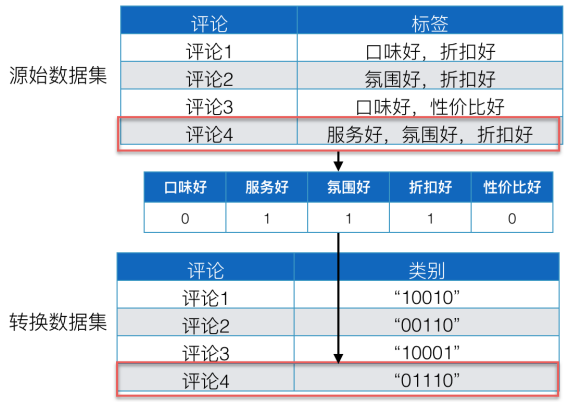

Figure 9: 标签幂集方法数据转换示意

ii. 优劣分析

二元关联方法将各个标签独立出来进行分类器的训练，而实际数据中各个标签之间可能存在一定的联系，并不完全相互独立。通过统计训练集各标签频率，可以得到表3。

Table 3: 各标签相关性统计

标签组合(l_1, l_2)	$f(l_1) * f(l_2)$	$f(l_1 \cap l_2)$
(口味，服务)	0.2035	0.2446
(口味，氛围)	0.2165	0.1579
(口味，折扣)	0.0439	0.0336
(口味，性价比)	0.1232	0.1175
(服务，氛围)	0.1440	0.1498
(服务，折扣)	0.0292	0.0288
(服务，性价比)	0.0819	0.0891
(氛围，折扣)	0.0311	0.0208
(氛围，性价比)	0.0872	0.0671
(折扣，性价比)	0.0177	0.0167

从表中可以看到差异最大的一组标签是(口味，氛围)，同时打了“口味好”和“氛围好”标签的评论频率小于单独打这两个标签的评论频率的积。可见这些标签之间还是存在一定的相关性的。而二元关联方法却没有考虑标签间的相关性。

标签幂集方法由于将标签集的每个子集认为是一个类别，所以这种方法考虑了标签之间的相关性。但是由于幂集可能会很大，容易导致每个类别的训练数据过少的问题。例如本项目中 $|L| = 5$, 除去空集不予考虑，有 $|P(L)| = 2^5 - 1 = 31$ 个子集。训练集大小为8846，平均每个子集可以有 $8846/31 \approx 286$ 条训练数据，而这样大小的训练数据显然是不够的。

iii. 折中方法

为了解决上述二元关联方法没有考虑相关性，标签幂集方法训练集过小的问题，我们在标签幂集方法上进行适当改进。相对于标签幂集方法将所有子集看作一个新类别，改进的方法仅考虑集合大小为2的子集，即表3中所有的“标签组合”。形式化地，对子集 $L'_i (L'_i \subset L, |L'_i| = 2)$，构造多元分类器 $C'_i : C'_i(\mathbf{X}) \rightarrow L'_i$。

最后多标签的分类结果由所有多元分类器 C' 的分类结果投票决定，取多数票作为最后的多标签分类结果。投票决定过程示意如图10所示。

	口味好	服务好	氛围好	折扣好	性价比好
(口味，服务)C'$_1$	0	1			
(口味，氛围)C'$_2$	1		0		
(口味，折扣)C'$_3$	1			1	
(口味，性价比)C'$_4$	0				0
(服务，氛围)C'$_5$		1	1		
(服务，折扣)C'$_6$		0		1	
(服务，性价比)C'$_7$		1			0
(氛围，折扣)C'$_8$			0	0	
(氛围，性价比)C'$_9$			1		1
(折扣，性价比)C'$_{10}$				1	1
投票	2	3	2	3	2
最终多标签分类结果	0	1	0	1	0

Figure 10: 投票示意图(深灰色表示没有分类结果)

计票过程中需要设置一个阈值，使得票数超过该阈值的标签值为1，出现在最终的多标签分类结果中，否则为0，不出现。通常这个阈值需要为所有标签单独设置，设置的阈值为标有该标签的评论数在总评论数中占比。如示意图10中，假设5个标签各占比20%，共有10个多元分类器参与投票，则某个标签的票数需超过2票才能被选为最终结果中的标签。

iv. 分类算法

本项目使用scikit-learn包的分类算法。实现了两类分类器，第一类是直接应用scikit-learn包中支持多标签分类算法的分类器[6]，第二类是实现上述二元关联方法的分类器。
第一类分类器包含以下三种分类器，并将在结果评价部分比较这三种分类器。

- 决策树 (Decision Tree)
- 随机森林 (Random Forest)
- K近邻 (K Nearest Neighbors)

第二类分类器包含以下三种分类器，并将在结果评价部分比较这三种分类器。

- 决策树 (Decision Tree)
- 随机森林 (Random Forest)
- 基于伯努利模型的朴素贝叶斯 (Bernoulli Naive Bayes)

VII. 多标签分类结果

i. 评价标准

在本项目中我们使用查准率(Precision)，查全率(Recall)，F1评分(F1 Score)作为评价多标签分类器的标准[5]。
令$\mathbf{X}$表示一条评论数据的特征向量，$\mathbf{y}$表示该条数据的标签向量，$\mathbf{y} \subseteq L$，C 为待评价的分类器，$C(\mathbf{X}) = \mathbf{p}$ 表示分类器的分类结果，则

$$Precision = \frac{|\mathbf{y} \cap \mathbf{p}|}{|\mathbf{p}|} \tag{1}$$

$$Recall = \frac{|\mathbf{y} \cap \mathbf{p}|}{|\mathbf{y}|} \tag{2}$$

$$F1\ Score = \frac{2}{\frac{1}{Precison} + \frac{1}{Recall}} \tag{3}$$

F1 Score是Precision和Recall的调和平均数。

ii. 评价结果

第一类分类器

将利用训练集训练得到的第一类三种分类器分别应用于测试集，得到如下图11所示的结果。
以F1 Score作为评价分类器的决定标准，则决策树是测试结果最好的分类器，但是和第一类分类器中其他两个分类器的差距不大，随机森林也是很不错的分类器。

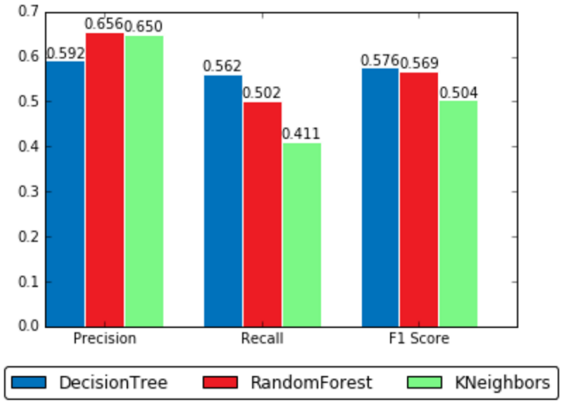

Figure 11: 第一类分类器在测试集上的分类结果

第二类分类器

利用第六部分提到的方法转换训练集，训练得到实现二元关联方法的第二类分类器，将其应用于测试集，得到如下图12所示的结果。

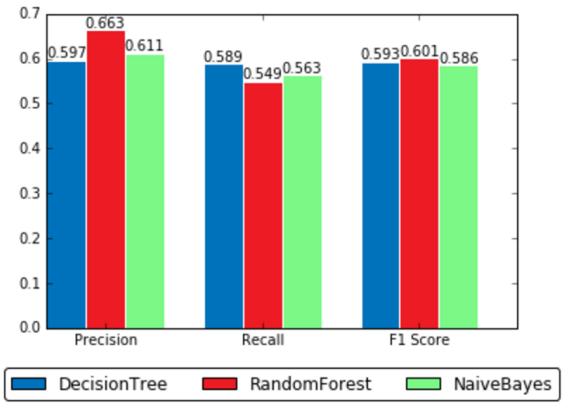

Figure 12: 第二类分类器在测试集上的分类结果

从图中可以看出第二类分类器中随机森林的表现稍好，但第二类分类器的三个分类器结果都在伯仲之间。
但是第二类分类器相较于第一类分类器，整体表现稍好。

检查trigram特征对分类的贡献

第五部分猜想trigram特征对分类没有明显的作用，这里我们利用第一类分类器，在除去了trigram特征的训练集上训练，并在相应调整了的测试集上测试，得到如下图13所示的结果。
将图13和图11相比，可以发现除去了trigram特

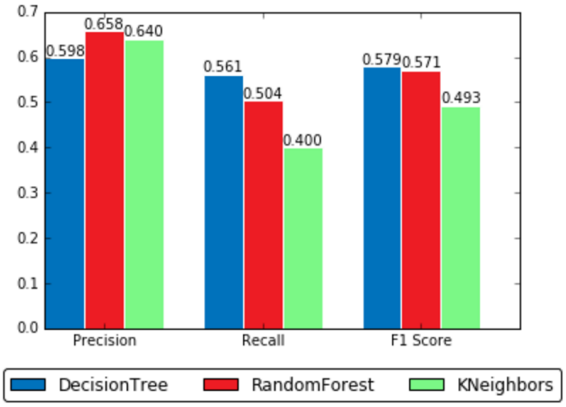

Figure 13: 去除*trigram*特征后的分类结果

征后，第一类分类器的评价结果几乎不会影响，由此也验证了之前的猜想，trigram特征对分类没有明显的贡献。

iii. 分类器应用

基于以上对第一类分类器和第二类分类器的比较，考虑到两者的结果相差无几，且第一类分类器利用scikit-learn直接实现，比较可靠且简便，所以在应用部分采用第一类分类器中的决策树分类器。

我们从数据集中抽取了评论数最多的三家餐厅，利用决策树分类器对餐厅的文字评价进行多标签分类，得到5个标签的分类结果。根据时间变化绘制出这三家餐厅5个方面的评价变化，来观察顾客对这三家餐厅的喜好变化，以及餐厅可以采取的针对性的改进措施。结果如图14，图15，图16所示。

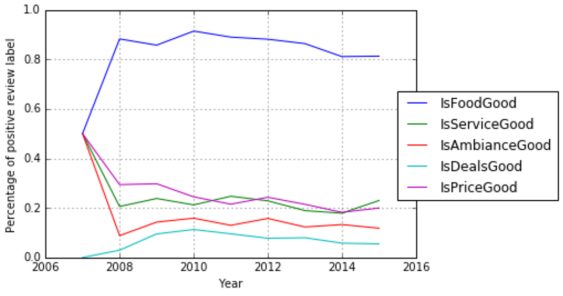

Figure 14: 分类器预测的对餐厅*1*的各标签评价变化

从这三家餐厅可以看出，大部分评论都认可“口味好”这一标签，都不是很认可“折扣好”这一标签，另外三个标签则没有明显的区别。这

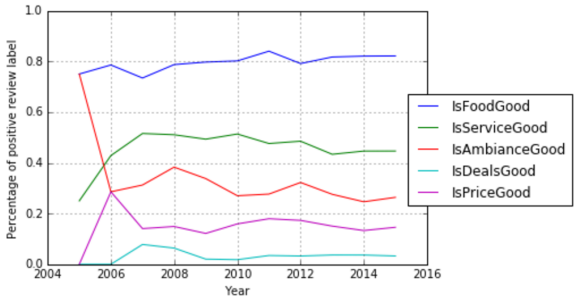

Figure 15: 分类器预测的对餐厅2的各标签评价变化

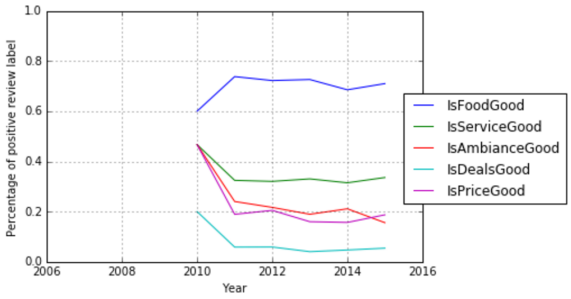

Figure 16: 分类器预测的对餐厅3的各标签评价变化(缺失部分表示此时该餐厅尚未开业)

样的评价分布也和真实的餐饮行业符合，餐饮行业”口味“是第一招牌，相对其他服务行业来说，“折扣”也相对少。同时可以观察到，评论最多的三家餐厅(也相应地意味着最受欢迎)的各方面评价都比较稳定。

VIII. 结论

点评网站的数值评分和文字评价都是有挖掘意义的数据。我们将文字评论抽取出语义特征，结合数值评分，利用多标签分类器得到餐厅各方面的评价，一方面可以帮助顾客快速准确地判断和筛选满意的餐厅，另一方面可以帮助营业者有针对性地解决经营问题。

在报告中我们展示了一个完整的数据挖掘流程，从粗略的数据探索，到精细的特征提取和标准化，再到算法的研究和讨论，最后比较结果的优劣并应用到实际的数据中。

利用二元关联方法解决多标签分类问题，我们最后得到了Precison在0.65左右，Recall在0.60左右，F1 Score在0.60左右的结果。但是也注意到各个方法的分类器差异并不明显，这个问题仍有改进的空间。

参考文献

[1] Yelp Dataset Challenge
`https://www.yelp.com/dataset_challenge`

[2] Sriram B, Fuhry D, Demir E, et al. *Short text classification in twitter to improve information filtering[C]* Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval. ACM, 2010: 841-842.

[3] Tsoumakas G, Katakis I. *Multi-label classification: An overview[J].* Dept. of Informatics, Aristotle University of Thessaloniki, Greece, 2006.

[4] Read J, Pfahringer B, Holmes G, et al. *Classifier chains for multi-label classification[J].* Machine learning, 2011, 85(3): 333-359.

[5] Godbole S, Sarawagi S. *Discriminative methods for multi-labeled classification[M]* Advances in Knowledge Discovery and Data Mining. Springer Berlin Heidelberg, 2004: 22-30.

[6] Multiclass and multilabel algorithms `http://scikit-learn.org/stable/modules/multiclass.html`