研究生学术就业大数据分析报告

基于weka的分类算法研究

大学生科创项目结项报告

分析部分

1. 综述

机器学习技术近年来蓬勃发展，而在教育数据领域的应用方兴未艾。此项目尝试使用机器学习技术分析中科院某所研究生大数据，探讨是否可以使用机器学习技术找到研究生学术就业与否的影响因素，并尝试根据分析结果给出预测方法。

在整个研究过程中，我们首先对数据进行了清洗，并创建了部分属性进行了离散化的数据集以满足部分分类算法的需要。此后使用聚类算法初步建立研究生人群画像，并结合属性选择算法和人工选择来找出对学术就业情况影响显著的因素。利用选择出的属性和全部属性分别测试各种常用分类算法的分类结果，选取结果比较突出的(C4.5为代表的各种决策树,贝叶斯)进行进一步分析，并尝试使用元算法和参数调整算法对分类算法进行优化。

实验结果给出了研究生学术就业倾向与研究生攻读学位水平的强烈联系，但是与其他因素的联系并没有特别紧密。实际影响学生就业的因素可能更为复杂且具有一定的随机性，因此通过其他因素预测研究生学术就业倾向的难度可能很高。

文章分为以下几个部分：

1. 综述 (第1段)
2. 数据预处理 (第2段)
3. 分析关键属性，进行属性选取 (第3，4，5段)
4. 分类算法的选取及调优 (第6，7段)
5. 引入元分类算法的尝试 (第8段)
6. 结语 (第9段)

2. 数据预处理

```
TODO: Paste 预处理 here.
```

数据预处理部分补充:

> 学分数据存在问题, 不同年份, 不同数据来源的学分采用不同的算法. 对这些数据, 我们在取得所有相同计算方式的学分之后, 按照学分的排名分布进行统计和量化, 计算出一名学生的相对学分和离散化相对学分, 以在之后的测试之中使用.

> 下面是数据分析部分：

基于预处理之后的研究生数据, 我们得到了以下的结果:

3. 聚类分析

在开始分类之前, 我们先尝试对处理过的数据进行初步的聚类处理, 来发掘数据可能的自然分类, 进而找出影响学生发展的主要因素(而不仅仅限于对于学生就业的影响). 这里的聚类分析, 我们使用了应用广泛的k均值(k-means)算法和最大期望(EM)聚类算法. 最大期望聚类算法通过交替进行期望计算和最大似然估计两个步骤来将数据聚合到几个集合之中. 我们希望通过聚类的结果, 来发现影响聚类的主要属性, 从而快速地找出影响学生是否选择学术界也得关键属性.

k-means聚类分析得出的影响聚类的主要属性如下(除去学术就业情况):

1. 学生入学方式(推免或是考研)
2. 学生类型(学硕或是专硕))
3. 入学年龄
4. 毕业年龄
5. 延迟毕业时间

EM聚类分析后影响分类的主要属性为:

1. 学生入学方式(推免或是考研)
2. 学生类型(学硕或是专硕))
3. 入学年龄
4. 毕业年龄
5. 延迟毕业时间
6. 导师职称
7. 学分
8. 就业情况(是否为学术就业)

以EM聚类算法的结果为例进行分析:

使用的参数如下:

```
weka.clusterers.EM -I 100 -N -1 -X 10 -max -1 -ll-cv 1.0E-6 -ll-iter
1.0E-6 -M 1.0E-6 -K 10 -num-slots 1 -S 100
```

EM算法给出了3个聚类, 数据所占的比重如下所示:

```
1       93 ( 12%)
2      178 ( 22%)
3      531 ( 66%)
```

三组数据的学术/非学术就业数据如下: 从左到右分别对应1,2,3三组:

```
EMPLOYMENT_UNIT_CLASS
  NO_EDUCATION_RESEARCH      136.592    95.345   372.063
  EDUCATION_RESEARCH         29.2626   97.6163   77.1211
  [total]                   165.8546  192.9613  449.1841
```

上面的结果中, 第二组的学生学术就业率达到了50%以上, 显著的高于其他组. 但是观察分类结果所涉及的各个属性可以发现, 这三组的数据特征大致可以归纳如下: 第1组为延毕的硕士研究生, 第二组为博士研究生, 第三组为正常毕业的硕士研究生. 显然: 学术就业与否与在攻读的学位水平有着明显的相关, 博士研究生的学术就业水平要明显的高于硕士研究生. 同时博士研究生与硕士研究生在"延毕年限"这一属性上有着显著的差异. 对于博士研究生, 延毕的可能性要远远高于硕士研究生. 尽管聚类算法仅仅揭示了学术就业与攻读学位水平之间的简单关系, 不过也为之后的实验设计提供了启示: 在下面的实验中可以分别就博士研究生和硕士研究生的学术就业情况进行分类, 来探究"攻读学位"属性之外其他因素对于学术就业情况的影响.

这里列出聚类算法生成的3个类的主要差异属性; 详细的聚类信息可以参见附件1.

```
Attribute                        1         2         3
                               (0.2)    (0.24)    (0.56)
=========================================================
RECRUITING_WAY
  MASTER_DOCTOR                7.7421  127.7561    7.5018
  COMMON_EXAM                       1        59         1
  UNIFIED_EXAM                97.4092    2.6508    111.94
  EXAM_FREE                   61.7033    4.5544  330.7423
  BACHELOR_DOCTOR                   1         2         1
  [total]                   168.8546  195.9613  452.1841
STYPE
  DOCTOR                       7.7421  186.7561    7.5018
  MASTER                      70.0039    5.5526  382.4435
  MASTER_SPECIALIZED          89.1086    1.6526   60.2388
  [total]                   166.8546  193.9613  450.1841
SEX
  MALE                      148.7906  159.6152  343.5941
  FEMALE                      17.064    33.346    105.59
  [total]                   165.8546  192.9613  449.1841
SAGE_IN
  mean                        22.515   24.6541   22.1784
  std. dev.                   1.1951    2.0992     0.925
SAGE_OUT
  mean                        25.435    28.626   24.9984
  std. dev.                   1.2062    2.3349    0.9093
DELAY
  DELAY                       15.8249  157.1751         1
  UNDELAY                   150.0297   35.7862  448.1841
  [total]                   165.8546  192.9613  449.1841
DALEY_YEAR
  mean                         0.063    1.1059         0
  std. dev.                   0.2211    0.8564         0

***
```

```
EMPLOYMENT_UNIT_CLASS
  NO_EDUCATION_RESEARCH      136.592    95.345   372.063
  EDUCATION_RESEARCH          29.2626   97.6163   77.1211
  [total]                    165.8546  192.9613  449.1841
```

在聚类的尝试中还有一些其他有趣的发现, 譬如聚类同样揭示了硕士和博士学位的在导师水平, 男女比例, 专业选择等方面的区别, 但是这些区别与研究的题目没有太大的相关性, 在此不再赘述.

---

4. 手动分析

在使用聚类算法分析影响聚类的因素之后, 我们将研究生的数据与博士生的数据分开考虑. 首先 使用weka.trees.UserClassifier进行人工分析, 查看学术就业情况是否对应其他显著的特征. 但直接观察散点图并不能得出有用的结果.

---

5. 选择属性

鉴于难以人工识别高影响力的属性, 我们尝试使用Weka提供的CfsSubsetEval来自动查找对于"学术就业"有较高影响的属性.

"CfsSubsetEval评估器评估每个属性的预测能力以及相互之间的冗余度，倾向于选择与类别属性相关度高，但是相互之间相关度低的属性. " 在每次迭代中, 这一评估器尝试引入余下的属性中与类别属性相关度最高的属性.

使用的参数和结果如下:

```
weka.attributeSelection.CfsSubsetEval -P 1 -E 1
```

```
Selected attributes: 1,2,3,12,20,21,33,34 : 8
                     YEAR_IN
                     RECRUITING_WAY
                     STYPE
                     SAGE_OUT
                     CREDIT
                     GPA
                     DALEY_YEAR
                     DELAY
```

综合CfsSubsetEval评估器的结果和聚类中得出的对整体聚类有较大影响的属性列表, 我们选择在这些属性上应用分类算法进行学术就业预测. 转而舍弃掉其他相关度比较小的属性, 来减少它们对分类难度和结果的影响.

除此之外，在属性选择上我们还参考了决策树桩分类器的结果。关于决策树桩使用的详细介绍在下文分类算法部分。

6. 不同分类算法的比较

接下来, 我们尝试观察不同的分类算法在对于"研究生毕业后是否选择学术就业"这个属性上的分类效果. 为了确定较为有效的分类算法, 这里我们不引入元分类算法, 而是比较各个单独的分类算法在常用参数下的分类效果. 这里我们使用了weka的experimenter功能实现以来批量运行相似的算法来提高效率. 一些算法有着特定的数据要求, 比如只接受离散化的数据. 对于这部分算法, 在运行之前数据格式已经按照算法的要求进行了调整.

下面的表格反映了不同分类算法在常用参数下应用于全部剩余属性数据集的实验结果. 所有的验证采用10折交叉验证来确定算法效果:

分类算法	**ned-ned**	**ed-ned**	**ned-ed**	**ed-ed**	**edTP**	**nedTP**	注记
贝叶斯	504	97	102	99	0.492537313	0.838602329	
朴素贝叶斯	515	86	117	84	0.417910448	0.856905158	
逻辑斯蒂回归	522	49	142	59	0.293532338	0.914185639	
反向传播	492	109	124	77	0.383084577	0.818635607	速度很慢
SGD随机梯度下降	568	33	157	44	0.218905473	0.945091514	nedTP很高但edTP很低
简单逻辑斯蒂	565	36	151	50	0.248756219	0.940099834	
SMO序列最小优化算法	582	19	163	38	0.189054726	0.968386023	
ibk k最近邻分类器	466	135	135	66	0.328358209	0.775374376	
kstar	483	118	136	65	0.323383085	0.803660566	稍慢
lwl局部加权学习	556	45	149	52	0.258706468	0.925124792	
DecisionTable	551	50	146	55	0.273631841	0.916805324	
Jrip	541	60	131	70	0.348258706	0.900166389	
OneR	545	56	171	30	0.149253731	0.906821963	kappa过低不考虑
PART	471	130	126	75	0.373134328	0.783693844	
决策树桩	522	79	128	73	0.36318408	0.868552413	
J48 C4.5决策树	534	67	139	62	0.308457711	0.888519135	
LMT逻辑斯蒂模型树	565	36	151	50	0.248756219	0.940099834	
随机森林	560	41	143	58	0.288557214	0.931780366	
随机树	460	141	136	65	0.323383085	0.765391015	kappa过低不考虑
REP树	544	57	147			0.90515807	

7. 表现(相对)良好的几种算法

综合上面的结果, 贝叶斯算法, C4.5决策树算法表现出了较好的效果, 对这些算法进行进一步的梳理和优化的结果如下, 如无特殊说明，分类准确度验证采取10折交叉验证:

详细的运行结果以及统计量可以在前表及附录中找到。

决策树桩

决策树是简单而常用的机器学习算法. 它通过贪婪地构建决策树来极小化目标函数, 进而取得分类的依据. 而决策树桩是单层简化版本的决策树. 它只通过极小化信息熵进行一轮划分来做出预测. 因此, 在不能直接观察到影响学术就业的属性的情况下, 我们采用决策树桩来找出影响比较大的属性. 与上文所提及的结果一致, 首先被选出的是"学生种类"属性, 亦即学生是否为博士研究生. 一个有趣的结论是, 这样朴素的方法与之后的方法在统计意义上具有相似的分类水平.

决策树桩算法可以直接调用:

```
weka.classifiers.trees.DecisionStump
```

决策树桩给出了简单但是明确的相关关系. 但是通过使用更加复杂的分类算法, 分类的准确程度还可以进一步提升.

贝叶斯

贝叶斯算法同样是机器学习领域的常见算法之一. 通过一定的先验假设(根据贝叶斯算法的不同变种, 这个假设会有所区别), 贝叶斯算法会学习这个数据集输入/输出之间的联合概率分布, 进而尝试进行分类. 对研究生数据集应用贝叶斯算法,可以得到略好于按博士生/非博士生的区别预测学术就业的准确率.

使用以下的参数运行贝叶斯算法:

```
Scheme:       weka.classifiers.bayes.BayesNet -D -Q
weka.classifiers.bayes.net.search.local.K2 -- -P 1 -S BAYES -E
weka.classifiers.bayes.net.estimate.SimpleEstimator -- -A 0.5
```

分类的结果如下:

```
Correctly Classified Instances         603               75.187  %
Incorrectly Classified Instances       199               24.813  %
Kappa statistic                          0.3339
Mean absolute error                      0.2683
Root mean squared error                  0.4803
Relative absolute error                 71.3496 %
Root relative squared error            110.8278 %
Total Number of Instances              802
```

```
=== Confusion Matrix ===

   a   b   <-- classified as
 504  97 |   a = NO_EDUCATION_RESEARCH
 102  99 |   b = EDUCATION_RESEARCH
```

Weka-J48(C4.5决策树)

Weka-J48决策树, 亦即C4.5决策树是决策树的一个亚种. 通过在树的生长操作中使用信息熵下降率而非信息熵减益作为生长依据, 在一些场合取得了比基本的决策树算法更好的效果. 同时在教育信息领域, 也有基于C4.5决策树进行分类预测的有力尝试.

> 忘了是哪篇论文了, 但是我们都看过就是了, 好像是上海某高校的

在分类中使用C4.5决策树的默认参数如下:

```
weka.classifiers.trees.J48 -C 0.25 -M 2
```

这是此分类器默认的参数设置, 对于J48分类器, 很少需要为获得良好的性能而更改这些参数. (袁梅宇. 数据挖掘与机器学习:WEKA应用技术与实践. 清华大学出版社.)

在研究生数据集上应用C4.5决策树的结果略好于决策树桩的结果.

---

8. 元分类算法分析

元分类算法

我们在这里尝试在元算法中组合贝叶斯和J48,决策树桩, 来尝试取得更好的分类效果. 在运行元算法时不是用之前人工选取的属性，而是使用`weka.AttributeSelectedClassifier`先进行相关属性选择来提高分类效果。在迭代的过程中，通过`weka.CVParameterSelection`来调整参数。但是原分类算法比起单独使用某种算法的分类结果提升非常有限(见表1)，详细结果参见附录5.

9. 结语

综上所述，研究生学术就业倾向与研究生攻读学位水平有着强烈的联系，博士研究生学位的攻读者学术就业的概率远远高于硕士研究生，且两者在人群画像上也有着显著的差异。但通过目前的数据集中的数据，我们尚且不能给出准确率特别高的学术就业情况预测方法。我们猜测就业结果可能有很大的随机成分且与复杂的因素有关。

尽管如此，依然有一些分类算法的结果在预测学术就业上可以作为参考。整体来说决策树类的算法体现出了比较好的效果。贝叶斯系的算法有类似的效果但是结果比较晦涩。如果能对算法的参数进行进一步调整的话，贝叶斯系算法可能会获得更好的实验效果。

反向传播算法在生成模型的过程中花费了过长的时间，因此不在此次测试的范围之内。但是通过对模型进行针对性优化和使用硬件加速，可能会取得比较理想的效果。在这一问题上反向传播算法的实用程度还有待研究。

通过使用元分类算法和自动参数调整，上述算法可能还有一定的优化空间。本文所使用的元算法仅仅是J4.5，贝叶斯，决策树桩(攻读学位属性)的简单叠加，通过精心设计元算法，预测的准确度应当可以进一步的提升。

10. 致谢

本文的大量成果基于Weka--怀卡托智能分析环境（Waikato Environment for Knowledge Analysis）的开源算法。