半监督学习在研究生调剂中的应用①

黄树成，曲亚辉

(江苏科技大学 计算机科学与工程学院，镇江 212003)

摘 要：研究生调剂是研究生招生中的重要环节。传统的调剂方法都是通过手工操作的，考生很难从往年大量的调剂数据中分析出规律，选报合适的学校。提出了基于半监督学习的数据挖掘方法，也即是从已知类别的训练样本提取出其中的关联规则作为分类的监督信息，并结合非监督学习方法中的 K-mean 聚类算法，对大量未标识样本进行分类的算法，此方法克服了研究生调剂涉及因素繁多，无法准确填报的弊端。该方法实现过程简单，分类准确，可推广性较强。

关键词：半监督学习；K-mean 算法；关联规则；聚类

Application of Semi-Supervised Learning to Graduate Adjusting

HUANG Shu-Cheng, QU Ya-Hui

(School of Computer Science & Engineering, Jangsu University of Science & Technology, Zhenjiang 212003, China)

Abstract: Graduate Adjusting is an important step for Graduate Admission. The traditional adjusting methods which are all manual, make it very hard for students to choose a proper school from a huge number of data. This paper proposes a data-mining method based on semi-supervised study. Using the association rules, which are extracted from the labeled training samples, as supervised information, and combining with the K-mean algorithm in non-supervised study method, this paper elaborates on the semi-supervised study algorithm by classifying a large number of unlabeled data. This method overcomes the defects of inaccuracy in traditional methods which are influenced by a large number of factors. The method is simple to implement, has high accuracy, and can be widely used.

Keywords: semi-supervised learning; K-mean algorithm; association rules; clustering

1 引言

我国研究生招生采用考生在考试前填报志愿的方式，考生在考试中发挥失误、对自己水平估计错误、复习期间发生意外等情况而没有被录取到所报考学校的都需要调剂。目前还没有规范的调剂志愿填报系统为考生提供准确的参考，本文使用数据挖掘技术对研究生调剂的相关数据进行深入分析，得出分类关联规则，从而对考生填报志愿进行预测性的指导分析。研究生调剂数据不只是一方面的，它涉及到分数、第一志愿学校、专业课不同等方面，要想对其进行准确的分析只利用决策树一种分类算法进行是不可能做到的，并且同一个分数段内可供调剂的学校有很多个，不能说每个学生只能调剂唯一合适的学校。传统的方法都是基于大量带标签的数据，而实际上，能够标定的数量是非常有限的，本文利用半监督学习方法，来实现数据分析。具体方法是：首先我们利用往年调剂数据作为已标示数据，从中挖掘出关联规则作为监督信息，利用聚类算法的 K-mean 聚类算法对大量的未标示数据进行归类，对运用此算法不能合理归类的样本数据利用正、反分类器的方法进行准确分类。

① 基金项目:江苏省高校自然科学研究计划(2008DX065J)
收稿时间:2010-08-19;收到修改稿时间:2010-10-15

2 半监督学习算法

基于数据挖掘的学习方法分为监督学习和非监督学习两大类。监督学习是在具有带标示数据训练样本集的前提下，致力于构造最优模型，使它与已知知识之间的误差最小，是一个寻优过程，目标单一确切，准确度较高，一般适用于少量的容易标示的数据；但是需要对进行分析的数据的每个样本进行标示，不但开销大而且部分数据很难准确的标示。非监督学习正好相反，是在大量的未标示数据里根据对象自身的特征寻求有意义的模式，准确度相对较低[1]一般适用于大量的难标示数据。半监督学习方法是将两者结合起来，既可以灵活学习，又可提高准确度。

目前半监督学习算法很多，比较有代表性的有：

1) EM(Expectation Maximization)进行标识和参数估计，然后利用获得的完整样本集再进行模型的学习[2]，就是从基于未发现的潜在变量的模型中找出最大可能性的估计，其实是出于一种很自然的想法。2)另一种具有代表性的半监督学习方法是协同训练(Co-training)算法[3,4]. 3)i-training 算法[5]，通过利用集成学习来提高泛化能力，Tri-training 算法是将 Co-raining 算法进行进一步扩展。该算法采用三个分类器进行学习，其中两个分类器的功能跟 Co-raining 算法一样采用这两个分类器对无标识数据进行预测，预测结果一致则加入第三个分类器的有标识样本进行训练最后运用集成学习对测试样本进行预测。

基于正、反分类器的半监督学习算法：

该算法是对 Co-training 算法的扩展，但也不同于 Tri-training 算法，他不是建立三个分类器，而是两个分类器。算法以样本类别的差异为视角，利用少量有标识数据中的各类数据分别建立不同的单分类器，不同单分类器通过对相同的无标识样本进行测试，建立这些样本的置信度，选择置信度高的样本对单类分类器已建立的分类面进行调整。最终被识别出来的无标识数据和有标识数据集合在一起训练一个基分类器，多个基分类器集成在一起对测试样本进行测试。对于单类分类问题，已有许多解决方法[6]，比较经典的有三类。1)密度估计法。2)边界法。3)重建法。本文选用聚类算法中的 K-mean 算法结合关联规则挖掘算法抽取分类关联规则作为监督信息建立研究生调剂数据的判决模型，对难于归类的样本采用正、反分类器进行归类。

2.1 分类关联规则的抽取

Apriori 算法[7]是一种非常经典的监督学习算法，该算法通过在设定的最小支持度(Support)和最小置信度(Confidence)下对数据样本属性依次进行统计、剪枝、连接，最后找出频繁项集和提取满足最小置信度的关联规则。在数据样本不是很大的时候，Apriori 算法是非常有效的。本文首先应用 Apriori 算法从研究生调剂数据库中抽取分类关联规则，总的分两步抽取：1)找出所有频繁谓词集;2)由频繁谓词集产生强关联规则；下面具体介绍：

最小支持度指的是在大量数据中统计出指定的属性或者规则的事务数量与所有的属性或规则的事务数量比值，可以记为 $Support(A => B)$，其中 A、B 都是项集。

最小置信度指的是在包含 A 事务同时也包含 B 可能性大小，即条件概率，可以记为 $Confidence(A => B)$。可以用下面的式子表达

最小支持度 $Support(A => B) = P(AUB)$ (1)

最小置信度 $Condse(A => B) = P(B \mid A)$ (2)

分类关联规则：$Condset => y$，其中 $Condset$ 是项(或属性-值对)的集合，y 是类标号，由于分类规则右端是类标号，所以这种规则本身可以应用于分类。满足最小支持度的规则是频繁的，满足最小置信度的规则是精确的强规则。从聚类的角度看，假如每个样本本身是一个簇，那么满足最小支持度与最小置信度的分类关联规则可以近似地看作类中心。这是因为这种簇的密度比较大，由于它是频繁的；可以在高密度里充当中心点，因为它满足最小置信度。

故从已知的训练样本中提取分类关联规则对聚类具有重要的指导意义。

具体算法描述如下：

算法 1：抽取分类关联规则

输入：由 n 个属性$\{A_1, A_2, \cdots, A_n\}$构成的研究生调剂数据库 D，其中 A_n 为类标号属性；

最小支持度阈值 min_sup

出：分类关联规则

方法：

1) 统计样本总数,记为 SUM（A）。

2)记属性 A_r 有 N_r 个不同值，并 A_r 把各种属性值记为 $A_r[j]$，其中 $1 \le r \le n$； j 是每一种属性值的下标且 $1 \le j \le n$； j 与 r 均为整数。

3) 统计 $A_r[j]$ 下属性值的数量，记为 $NA_r[j]$，$1 \le r \le n$，$1 \le j \le n$。

4) For ($j=1; j \le n_r; j++$) if ($NA_r[j]/sum < \min\_\sup$)delete $A_r[j]$。

5) 留下来的 $A_r[j]$ 与 $NA_r[j]$ 组合，分别统计各种可能组合下的数量，把不满足最小支持度的组合项删除，即按照步骤(4)的方法进行剪枝，最后得到两个属性组合下的候选频繁项集。

6) 按照以上方法，进行逐层搜索迭代，由两个属性的组合逐步扩展到所有属性的组合，最后就会得到包含分类属性 A_M 的频繁项集。

7) If ($p(A_1 \cap A_2 \ldots A_{M-1} / A_M) \geqslant \min\_conf$)输出分类规则 $A_1 \cap A_2 \ldots A_{M-1} -> A_M$。

2.2 K-mean 聚类算法

K-mean 算法[8]是聚类分析比较经典的算法之一，K-mean 算法把对象集合 D 划分成一组聚类 $\{C_1, C_2, \cdots, C_K\}$，这里 $\bigcup_{r=1}^{k} C_r = D$，其中 k 是要得到的聚类个数。聚类的结果可以用一个隶属矩阵 $W = \{W_{ij},\ 1<i<n,\ 1 \le j \le k\}$ 来表示。K-mean 的聚类目标函数为 $\sum_{i=1}^{k}\sum_{j=1}^{m} w_{ij} d(x_i z_j)$，$x_i$ 是第 i 个对象；z_j 是第 j 个聚类的中心。为找到使目标函数值最小的聚类中心和隶属矩阵，K-mean 采用爬山算法[9]。

首先随机选取 k 个初始聚类中心，把每个对象分配给离它最近的据点，从而得到一组聚类。然后重新计算每个聚类的中心并把它作为新的聚点，把每个对象重新分配到最近的聚点。如果满足终止条件则算法结束，否则用新聚类代替原聚类。文中采用最常用的欧氏距离作为研究生调剂数据库中样本相似性度量的标准，模式样本向量 x 与中心 z 的距离为:

$$D = d(x_i, z_j) = \sqrt{\sum_{k=1}^{\infty} |x_{ik} - z_{jk}|^2}$$

3 于研究生调剂的半监督学习方法

对需要调剂的大学分成六个类型，考生只用合理找出自己属于哪一类学校的调剂范围，然后根据此类中每个学校的报考情况跟个人爱好选择合适的学校，根据基于单分类器的半监督学习算法来找到适合自己条件所属的分类。具体分析：K-mean 聚类实际上是以 k 个中心点来进行划分，算法基于最小化所有对象与其参照点之间的相异度之和的原则来执行。算法的优点是简单、易于执行，并且能探测到孤立点。

为了提高分类器的准确性和降低聚类的复杂度，可以利用分类关联规则作为指导信息。从已知类别的训练样本中抽取出来的分类关联规则实际上是频繁出现的高可信度的带类标签的簇，可以近似地看作类中心。利用这些带标签的簇充当监督信息进行聚类划分，就能对 K-mean 算法起到很好的作用。因此，本文选用 K-mean 算法结合关联规则挖掘算法抽取分类关联规则作为监督信息建立研究生调剂数据的判决模型。

对于难以识别的数据我们通过正、反分类器识别，具体就是：对于每一种学校的分类都建立正、反两个分类器，正类是“属于”，反类是“不属于”，首先把样本加入到正、反两个目标分类里面，然后对正、反两个目标分类里的有标示数据和无标示数据重新进行学习，根据重新学习的结果判断样本更合适属于哪个目标分类，从而判断目标样本是否属于此类学校。对于这样判断出现的一个样本属于多个学校分类或一个样本不属于任何分类的情况我们再根据不同的正分类器判断样本最适合哪一个分类。

假定研究生调剂数据库中的所有样本都独立，每个样本就是一个子集，具体建模步骤如下：

1) 首先对学校进行目标分类（可以用决策树算法对学校进行目标分类，此分类方法本文不做介绍，本文假定已经对学校做了合理的分类），根据部分已经调剂的考生挖掘出分类关联规则作为带标签的簇，每条规则作为初始聚类中心。

2) 对每一类学校建立正、反分类器，正、反类中平均选取除了本类外的各类的已标示的样本数据若干(最好跟正目标分类的数据量相当)。如图 1 所示：

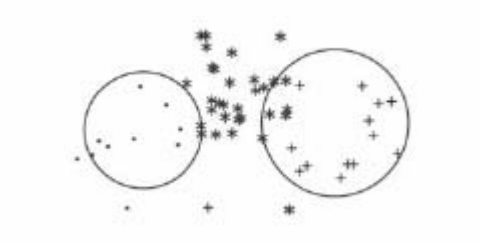

图 1 正、反分类器原理（“•”为正类样本,“+”为反类样本,”“*”为无标示样本）

3) 通过 K-mean 计算各未知样本与初始聚类中心的距离，并反复进行比较、调整。样本与聚类中心距离的计算公式：

$$D=d(x_i,z_j)=\sqrt{\sum_{k=1}^{\infty}|x_{ik}-z_{jk}|^2}$$

其中 x_i 是样本，z_j 是聚类的中心，$1<i<n,1<j<k$。

4) 通过比较寻找与各聚类中心距离最小的带标签的簇，把该未知样本类别标签为该簇对应的类别。

5) 对通过上面的聚类算法无法标示的样本，通过单分类器进行分类，对于通过单分类器产生的一个样本属于多个分类的情况通过比较所属目标分类的正类选择最合适的分类对于一个样本不属于任何分类的情况通过比较所有单分类器的正类选择最合适的分类。

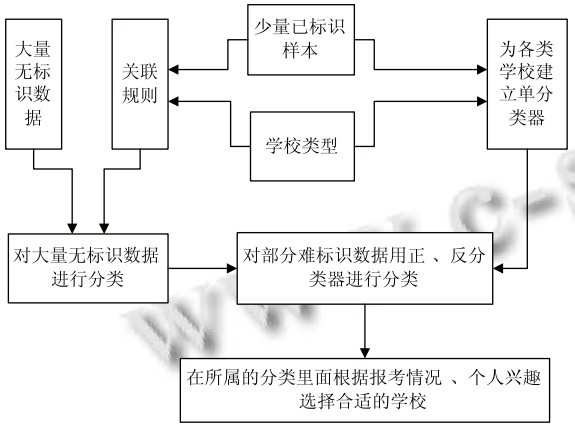

图 2 研究生调剂系统模型

4 实验与结果分析

4.1 实验数据

本文主要数据来源于自己身边的同学还有自己在网上搜集的部分学校学生调剂情况，以及部分高校的条件标准，调剂复试分数线等，本文采用样本毕业院校类型、第一志愿报考院校类型、考研分数三个指标作为评价因子。

根据学校实力，名声等把学校分成六个类型，以工科为例，1 型(清华大学、北京大学、上海交通大学、浙江大学、中科院研究生院)2 型(西安交通大学、南开大学、南京航空航天大学、北京理工大学等)3 型(郑州大学、河南大学、江南大学、上海大学、江苏大学等)4 型(河南科技大学、江苏科技大学、武汉科技大学等)5 型(华北水利水电学院、中北大学等)6 型(中原工学院等一些刚招生研究生的学校)表 1 是所选取样本学校，及所选学校实际调剂学生情况及实际调剂学生数占总学生数的比例。

表 1 调剂学校及调剂学生比例情况

等 级	学校样本数	调剂学生数	占总学生比例
等级一	5	0	0.00%
等级二	10	36	1.87%
等级三	20	127	6.61%
等级四	30	308	16.03%
等级五	40	506	26.34%
等级六	50	944	49.15%

4.2 试验分析及结果

采用半监督学习算法，在随机选取不同数量的已知类别的训练样本下得到的不同的带标签的簇，然后不断计算未知样本与聚类中心的距离，反复比较确定，最后得出准确的分类。先把往年的部分调剂数据作为已标示数据，挖掘出关联规则，然后对大量的往年调剂数据作为未标示数据，对其进行分类，对部分往年报考很少突然有一年报考较多，而考生考分较高落选的需要根据单分类器进行分类。

我们假设小明是江苏科技大学计算机学院的学生，去年报考的华中科技大学计算机应用专业，他去年研究生入学考试成绩是 368，最后复试没有被录取，需要调剂，根据本系统他需要进入本系统填写相关数据(具体不再详述)，我们只具体介绍一下系统内部的工作原理：

先根据系统内部已表示数据算出支持度和置信度，找出所有频谓词集，把分数按 20 分划一个段；一志愿报考院校、调剂院校 分上面三个类型，我们设定最小支持度为 20%，最小置信度为 60%，按照 1.1 的方法抽取的其中一个关联规则为：

$record(x,350)\Lambda firstschool(x,类型2)\Rightarrow result(x,类型4)$

$record(x,370)\Lambda firstschool(x,类型2)\Rightarrow result(x,类型3)$

然后根据：$D=d\ (x_i,z_j)=\sqrt{\sum_{k=1}^{\infty}|x_{ik}-z_{jk}|^2}$

计算样本属性值与中心属性值的差异 $\sqrt{\sum_{k=1}^{\infty}|368-350|^2}$ ，$>\sqrt{\sum_{k=1}^{\infty}|368-370|^2}$ 所以样本离聚类 370 分的中心更近点，选择关联规则 $record(x,370)\Lambda firstschool(x,类型2)\Rightarrow result(x,类型3)$，然后用同样的方法抽取出所有的关联规则，以此作为聚类算法的中心，然后利用本文 3 的方法计算样本属性值与中心属性值的差异，根据差异再进行准确的对样本进行归类，如果再无法进行分类就用建立的单分类器进行分类，然后通过各个关联规则的分类情况选择最多的分类，也即是最合适的分类。通过计算我们发现小明的情况最适合报考类型 3 的学校，所以小明可以在 3 型学校里面选择自己喜欢的学校进行调剂。图 3 显示了正反分类器的工作流程模型：

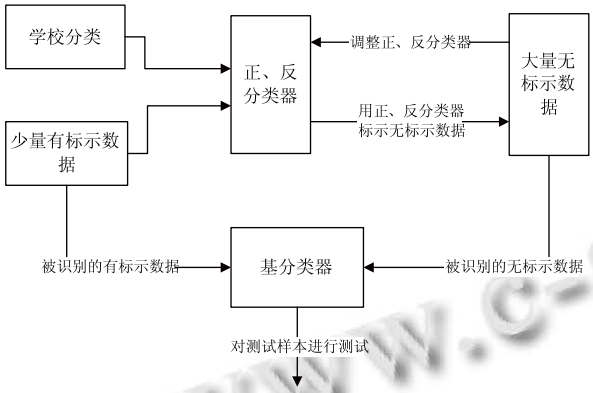

图 3　正、反分类器的工作模型

表 2 表明根据设计模型对已经样本进行预测所得的调剂结果，根据两个表的对比可以看出来根据单分类器的半监督学习方法设计的模型预测的调剂结果和实际情况基本符合。并且能让学生在自己现有的条件下选择尽量好的学校。

表 2　根据预测所得的调剂学校分布

等　级	学校样本数	调剂学生数	占总学生比例
等级一	5	2	0.10%
等级二	10	38	1.98%
等级三	20	102	5.31%
等级四	30	267	13.90%
等级五	40	486	25.30%
等级六	50	1026	53.41%

5　总结

本文利用数据挖掘算法从研究生调剂数据中挖掘出不同的分类关联规则作为监督信息，结合聚类分析中 K-mean 算法对样本数据进行归类，对部分难于归类的样本利用正、反分类器进行归类。该方法充分利用监督学习分类准确率高和非监督学习无需标识训练样本的优点，既可以灵活学习，又可提高准确度。只需利用少量的训练样本，以较高的分类准确率确定大量数据的类别，克服了传统模型中人为因素多的影响。实验结果表明，利用此方法得到的结果很接近实际结果。

参考文献

1 陈安,陈宁,周龙骧,等.数据挖掘技术及应用.北京:科学出版社,2006.

2 张钦礼,王士同.基于 expectation maximization 算法的 Mamdani-Larsen 模糊系统及其在时间序列预测中的应用.物理学报,MachineLearning,2009,1:7−8.

3 尹哲峰,崔荣一.协同训练在教师评估中的应用.延边大学学,2009,2:167−170.

4 Cheng JT. A Variational expectional-maximization method for the inverse black body radiation problem. Journal of Computational Mathematics, 2008,6:6−7.

5 Zhou ZH, Li M. Tritraining: exploiting unlabeled data using three classifiers. IEEE Trans. on Knowledge and Data Engneering, 2005,17(11):529−540.

6 Tax DMJ. One Class Classification: Concept Learning in the Absence of Counter Examples [Ph.D Dissertation]. Netherlands Delft University of Technology. Faculty of Information Technologand Systems, 2001.

7 雷鸣,朱祖平,姚立纲.基于 Apriori 算法的顾客需求自动映射研究.机械设计与制造,2009,2:91−93.

8 索红光,王玉伟.基于参考区域的 k-means 文本聚类算法.计算机工程与设计,2009,2:1−3.

9 单冬冬,吕强,李亚飞,王磊.贝叶斯网学习中一种有效的爬山算法.小型微型计算机系统,2009,12:32−33.