基于标签树的生长干部招生计划聚类方法*

钱高祥[1,2],张 翀[1],张维明[1]
(1.国防科技大学 信息系统工程重点实验室,湖南 长沙 410073;
2.总政治部 干部部,北京 100031)

摘 要:生长干部招生计划是规定军队干部补充来源渠道及数量规模的重要依据,对干部队伍建设与发展具有重要的意义。随着招生计划每年拟制与积累,分析与评估计划将会起到辅助决策的作用,对招生工作的实施带来深远影响。本文从数据挖掘的角度,以聚类的方法研究分析招生计划,这为分析历年计划的波动性并理解评价计划与政策的贴合度提供了定量分析的手段。首先分析了招生计划的特点,进而提出以标签树量化招生计划的解决思路,通过抽取标签树中的特征子树作为聚类中的度量特征,并采用共现的方法实施"先形成核心,再依次分类"的步骤完成聚类。实验表明该方法在合成数据集和真实数据集上聚类效果较好、效率较高,对分析招生计划具有一定理论意义。

关键词:招生计划;标签树;聚类
中图分类号:TP311 **文献标志码**:A **文章编号**:1001-2486(2013)03-0024-06

Clustering cadet recruiting plans based on labeled trees

QIAN Gaoxiang[1,2], *ZHANG Chong*[1], *ZHANG Weiming*[1]
(1. Science and Technology on Information Systems Engineering Laboratory, National University of Defense Technology, Changsha 410073, China;
2. Officer Department, General Political Department, Beijing 100031, China)

Abstract: Cadets recruiting plan is an essential measurement to prescribe source and quantity of officers, which is significant for the construction and development of officer. With the accumulation of recruiting plans, it is meaningful to make analysis and evaluation of them. By utilizing the clustering technique, this research analyzed recruiting plans in the past years, and provided a quantitative approach to look into the fluctuation of plans and to understand and evaluate closeness of plans with policies. Firstly, the paper summarized characteristics of recruiting plans. Secondly, it proposed a solution which utilized labeled tree to represent recruiting plan and extracted representative subtrees to participate in clustering. Finally, it used co-occurrence idea of "forming clustering cores first, classifying the plans then" to finish the clustering process. Experimental results reveal that the method can provide better clustering results and is efficient.

Key words: recruiting plan; labeled trees; clustering

生长干部是我军干部补充的重要渠道,生长干部的培训规模、生源质量直接决定了我军干部队伍发展的水平。生长干部招生计划规定了本年度生源渠道和数量以及分配到各个军事院校、各专业、各生源地、男女性别等方面的指标,是生长干部补充规模、岗位数量分配、专业比例控制的重要依据,也是各级招生部门遵照实施的规定,对军队干部队伍的建设与发展起到直接影响的作用,是军队人才建设工程中不可或缺的法规依据。因此,对生长干部招生计划的评估分析具有十分重大的战略意义。

本文的研究背景是分析历年招生计划的波动性,从而:第一,可以更好地理解当年的招生相关政策,甚至可以有助于理解整个国家政治经济、人才培养等大的政策方针;第二,可以检验招生计划变化的合理性;第三,可以增加对招生计划评估的洞察力;第四,新政策出台时,有助于指导本年度招生计划拟制工作。本文以独特的视角提出采用聚类方法进行波动性分析,聚类将相似的个体聚集成簇,孤立点随之产生,从而可以洞察招生计划相似的某些年份,也可分析那些变化较大的计划的原因,进而考量评估相应的招生计划。

1 问题描述

生长干部招生计划以表格形式拟制,包含了生源渠道、招生数量以及分解到各个承训院校的指标、各专业的指标、性别指标和各生源地指标,

* **收稿日期**:2011-06-01
基金项目:国家自然科学基金资助项目(60172012)
作者简介:钱高祥(1971—),男,江苏高邮人,博士研究生,E-mail:early4932@163.com;
张维明(通信作者),男,教授,博士,博士生导师,wmzhang@nudt.edu.cn

其中生源渠道有青年学生(从高中报考军校)、士兵(现役士兵报考军校)、国防生等。以青年学生招生为例,在确定青年学生总数之后,将数量分解至各个军事院校,再将各院校的总数量分解到各个专业,然后再分解到男女性别,最后分解到各省份(注意,以上简化了分解指标,实际操作中分解的层次要复杂)。表 1 为招生计划的一个例子。

表 1　招生计划举例

Tab. 1　Example of recruiting plan

院校	专业	性别	省份				……
			北京市	天津市	河北省	山西省	
院校 A	专业 A	男	10	5	12	5	
		女	4	6	2	7	
	专业 B	男	5	6	8	12	
		女	12	4	7	8	
	……	……					
院校 B	专业 A	……					
	……	……					
……	……	……					

由以上描述可见,招生计划拟制过程是一种层层分解的过程,可将招生计划抽象为一种层次树结构,即根节点为补充总数量,继续分解为各个生源渠道,再分解为各院校数量,然后分解到各专业、性别以及生源地,如图 1 所示。

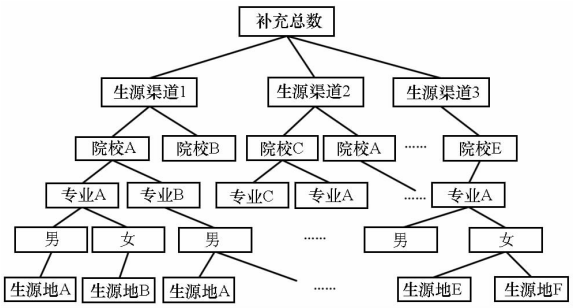


图 1　招生计划的树状表示

Fig. 1　Tree form of recruiting plan

本文认为单纯地考虑总数量或某个指标的数量不足以整体分析历年的招生计划,要从招生计划形成的整个层次树结构上进行度量,要以每年的招生计划所形成的结构为个体参与聚类,这样聚类结果才能够反映出历年招生计划的变化,才能更具有科学性。

2　相关工作

基于树结构的聚类技术以 XML 聚类为代表,相关工作分为两部分,一部分介绍目前 XML 聚类技术的发展现状,第二部分介绍树聚类中相似性度量的方法 p,q-gram 距离。

2.1　XML 聚类技术总结

文献[1]将整个 XML 文档集看作图,图的结点是 XML 文档,图的边是 XML 之间的 XLink 链接,XML 的聚类就转换为图的分割,这种方法易于引起 NP 问题[1]。

文献[2-5]将每个 XML 文档看作一棵树,再基于 1989 年 Zhang 等提出度量树之间相似性的 Editing Distance[6]理论,把度量两个 XML 文档之间的相似性转换为计算两棵树之间的编辑代价,对于这类方法,文献[7]证明其算法复杂度为 $O(|A||B|)$,其中 $|A|$ 和 $|B|$ 分别表示两个 XML 文档的元素数量,另外,文献[8]中举例说明这种方法有时不能如实度量 XML 文档之间的相似性,导致聚类结果不正确。

文献[9]对树的每个结点编码,遍历整个 XML 文档形成时域序列,再利用离散傅里叶变换将其转换为频域序列,通过比较频域序列检测出 XML 文档之间的差异;文献[8]将 XML 文档的父子结点进行共同提取,作为结构特征,并采用 VSM 表示 XML 文档,再利用 ROCK 算法进行聚类;文献[10]将 XML 文档中从根结点到叶结点的序列(即路径 Path)作为特征,采用 VSM 表示 XML 文档进行聚类。文献[11]指出这些方法构造出的向量普遍存在稀疏现象,会影响聚类过程的计算效率,因此应该采用降维的方法提高聚类的效率。基于此,文献[12]采用主成分分析(PCA)的方法降低维度;文献[13]将 XML 文档的路径作为特征,采用 VSM 表示 XML 文档,再根据最小支持度进行特征过滤,从而降低维度。

2.2　p,q-gram 距离

文献[14]中指出路径、边和节点不能准确地度量 XML 结构之间的相似性,因为它们不能反映出兄弟节点之间的关系。p,q-gram 距离[15]是度量 XML 结构相似性的工作中较新的研究成果。在文献[15]中,作者已经证明,p,q-gram 距离满足度量空间中距离的特性,并且它是编辑距离的下限,也就是说,采用 p,q-gram 距离度量 XML 结构相似性的精准程度与编辑距离相差无几。更具意义的是,采用 p,q-gram 距离进行度量的效率为 $O(n\log n)$,这大大提升了采用编辑距离带来的低效 $O(n^3)$。p,q-gram 距离的核心思想是从 XML 结构中提取子树作为特征,采用子树作为特征可以反映兄弟结点之间的关系,这比采用路径作为特征进行度量要精确。

3 基于特征共现的标签树聚类算法

本文提出基于特征共现的标签树聚类算法——COLES（Character cOoccurrence based Labeled trEe cluStering），首先将招生计划量化为标签树，再对标签树进行特征子树提取，由于招生计划分解项目和项目对应的分类值数量较多，从标签树中提取的特征子树数量庞大，这对聚类效率造成了压力。本文采用的主要思路就是利用具有代表性的特征指导历年招生计划的聚类。首先将招生计划转换为标签树，通过对标签树进行特征子树提取并根据共现相关度确定具有代表性的特征子树，本文称这样的特征子树为候选特征子树，通过对候选特征子树进行聚类，从而指导特征子树聚类，最后将特征子树的聚类结果指导整个招生计划的聚类，从而完成整个聚类过程。

3.1 标签树

第1节给出了招生计划可以表达为一种层次结构，这种层次结构在本文中以标签树的形式进行量化。

定义1 标签树。标签树具有一个根结点，每个结点都被一个标签所标记，并且标签的形式以 name-value 的方式进行表达。

图2示意了以标签树表示招生计划的形式。

3.2 提取特征子树

对于一个标签树，采用 p，q-gram 度量方法可提取出所有的 p，q-gram，这些 p，q-gram 即是该标签树的子树。使用大量的子树将导致聚类效率降低，本文采用的方式是将出现频率低和区别度低的子树进行过滤。提取特征子树过程如下：

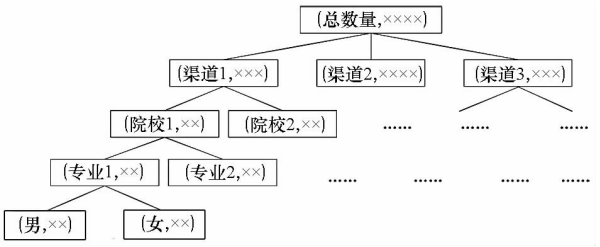


图2 招生计划以标签树形式表示

Fig. 2 Labeled tree of recruiting plan

1）设标签树集合 $Set_{LT}=\{LT_i \mid 1\leqslant i\leqslant M\}$，将所有标签树进行子树提取，去掉重复，形成 N 个不同的子树 $\{ST_j \mid 1\leqslant j\leqslant N\}$；

2）统计包含子树 ST_j 的标签树数量 OC_j，计算 ST_j 的频率 FC_j 为

$$FC_j = OC_j / M \tag{1}$$

其中 $M=|Set_{LT}|$；

3）计算每个子树 ST_j 在每个标签树 LT_i 中出现的次数 $OC_{i,j}$，ST_j 在 LT_i 中出现的频率 $FD_{i,j}$ 为

$$FD_{i,j} = OC_{i,j} / DP_i \tag{2}$$

其中 DP_i 表示 LT_i 中子树的总数量；

4）若某个子树 ST_j 满足 $FC_j<\theta$，其中 θ 称为最小支持度，则说明出现频率过低，在整个集合中重要性较低，应将 ST_j 过滤掉；

5）若某个 ST_j 满足

$$\max_{t,s\in M}(FD_{t,j}-FD_{s,j})/\max_{i\in M}(FD_{i,j})<\psi \tag{3}$$

其中 ψ 称为最小区别度，则表明 ST_j 对于标签树区别能力较低，体现不出特征，应将 ST_j 过滤掉；

6）剩余的子树即为特征子树。

3.3 构造共现相关图

共现相关度就是任意两个特征子树在标签树集合中同时出现的频繁程度。这种程度由式(4)进行度量：

$$CR_{s,t}=\frac{OC_{s\cap t}}{\max(OC_s,OC_t)} \tag{4}$$

其中 $OC_{s\cap t}$ 指特征路径 ST_s 与特征路径 ST_t 在整个标签树集合中共同出现的次数（在同一标签树中有多个 ST_s 或多个 ST_t，仅记为共现1次）。通过此式可计算出任意两个特征子树的共现相关度，从而可构造出共现相关度矩阵（上三角阵）

$$MCR=\begin{bmatrix} * & CR_{1,2} & \cdots & CR_{1,Q-1} & CR_{1,Q} \\ & * & CR_{2,3} & \cdots & CR_{2,Q} \\ & & \ddots & \ddots & \vdots \\ & & & * & CR_{Q-1,Q} \\ * & & & & * \end{bmatrix} \tag{5}$$

将小于一定值的 $CR_{s,t}$ 置为0，即当 ST_s 和 ST_t 满足 $CR_{s,t}<\dfrac{\sum CR_{s,t}}{Q(Q-1)/2}$ 时，可将对应的 $CR_{s,t}$ 置为0，其中 Q 表示特征子树的数量。经过这一过程，可形成关于特征子树的共现相关图，其中图的节点代表每一个特征子树，若两节点之间存在一条边，则表明对应的特征子树的 $CR_{s,t}\geqslant\dfrac{\sum CR_{s,t}}{Q(Q-1)/2}$，且此边的权重为 $CR_{s,t}$。

3.4 确定候选特征子树

这一过程是选定在共现相关图中具有代表性的候选特征子树。候选特征子树的作用在于在进行特征子树的聚类时，以候选特征子树为中心展开聚类。为此，本文定义特征子树的共现相关图中节点 ST_j 的权重

$$W_j = FC_j + \sum_{d=1}^{m} CR_{j,d}/m \tag{6}$$

其中 m 为节点 ST_j 的度。

若某个节点 ST_j 满足 $W_j > \sum_{j=1}^{Q} W_j/Q$,则选择其为候选特征子树,其中 Q 表示特征子树的数量,即共现相关图中节点数量。

3.5　候选特征子树聚类

这一过程为对提取出的候选特征子树进行聚类,采用的聚类方法为层次聚类法。假设所有的候选特征子树自成一类,合并距离相近(即共现相关度高)的候选特征子树,继续这一过程,直至使得聚类熵达到最小,计算聚类熵的方法如下:

$$En = \left(\sum_{j=1}^{k}\sum_{i=1}^{n_j} dist(ST_i^j, c^j)\right) - \sum dist(c^s, c^t) \tag{7}$$

其中式(7) 右边第一项为类内每一个点到该类中心的距离,第二项为每个类之间中心的距离,当此值为最小时,表明类内距离最小,类间距离为最大。

3.6　特征子树聚类

经过上一过程后,形成了候选特征子树的聚类,接下来进行特征子树的聚类。对于每个非候选特征子树 ST_r,计算其到每个聚类 k 的合成距离

$$CD_{r,k} = \sum_{j=0}^{S_k} dist(ST_r, ST_j^k) \tag{8}$$

其中 ST_j^k 表示第 k 类中的第 j 个点,S_k 表示第 k 类中所有点的数量。

找到最小的合成距离对应的聚类,将 ST_r 归到此聚类中。这样将每个非候选特征路径都进行归类,从而完成整个特征路径聚类。

3.7　标签树聚类

经过上述过程,就形成了特征子树的聚类。每个特征子树的聚类都可以表示为一个特征向量,每个标签树经过提取特征子树后也可以用一个特征向量来表示,这样计算标签树和每个特征子树聚类的距离,将标签树划分到与之距离最小的聚类中,这样就完成了整个标签树集合,即历年所有的招生计划的聚类。

4　实验与结果分析

全部实验都在一台计算机上进行。计算机硬件环境为 Intel(R) Core(TM)2 Quad Q9300@2.50GHz,2GB 内存;操作系统为 Windows XP Professional SP3,实验程序采用 Java 编制,开发包为 JDom[16] v1.1。

4.1　数据集与参数描述

实验中使用了两种数据集:合成数据集和真实数据集。考虑到 XML 也是一种标签树,因此合成数据集使用 XML 进行方法的有效性验证,采用不同的 DTD 产生不同数量的 XML,数据集如表 2 所示。真实数据集采用自 1980 年至今的招生计划,依据政策调整变化,招生计划相似的年份为:1980 年至 1990 年,1991 年至 2003 年,2004 年至 2010 年以及 2011 年至 2012 年。合成数据集中平均每篇文档中含有 XML 结点数量为 567.5,并且 DTD 之间有相互的重叠,所生成的 XML 种类中 ACM OrdinaryIssuePage, ACM ProceedingsPage 和 ACM SigmodRecord 区别不明显。

表 2　数据集文档数量分布

Tab.2　Distribution of data set

XML 结构种类	XML 文档数量
INEX Wikipedia	550
ACM OrdinaryIssuePage	300
ACM ProceedingsPage	400
ACM SigmodRecord	50
Shakespeare	600
Religion	800
总数量	2700

实验中最小支持度 θ 默认为 15%,对于合成数据集,参与聚类的文档数量默认为 2700,对于真实数据集,参与聚类的招生计划份数为 32。从 2 个方面验证所提方法的有效性:验证聚类效率与聚类质量。与本文进行对比的方法如下:

1)文献[1]中的 XLink 方法;

2)文献[8]中的 S-GRACE 方法;

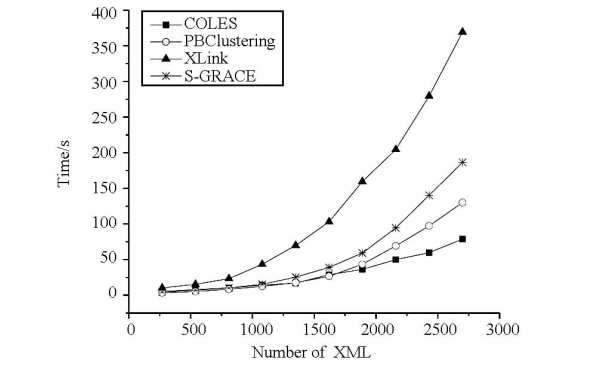
3)文献[13]中的 PBClustering 方法。

4.2　聚类效率测试

改变测试数据的数量,测试 4 种聚类算法所耗时间,每次每种类别增加该类总数的 10%,每次测试取连续 3 次消耗时间的平均值。

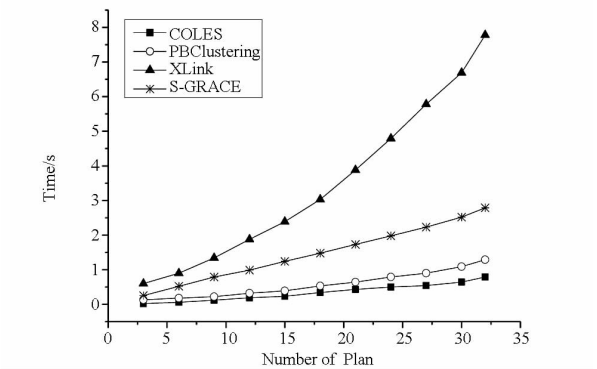
图 3(a)显示了 4 种聚类算法在合成数据集上的效率对比情况,可见 XLink 所采用的图分割算法效率的可扩展性较差,随着参与聚类文档数量的增多,所消耗的时间迅速增长;其次是 S-GRACE 算法,S-GRACE 算法提取 XML 的边为特征进行聚类,由于以边作为特征会产生较大的特征集合,因此这会严重影响算法的效率;接下来是 PBClustering,PBClustering 采用路径作为特征并采用支持度进行过滤,因此比以上 2 种算法效率要高;COLES 算法的效率要优于以上 3 种聚类算法,这是因为 COLES 算法使用了特征子树指导分类完成聚类而大大加速了聚类的效率。图 3(b)显示了在真实数据集上 4 种算法效率对比,与在合成数据

集上的效率对比类似,COLES 算法的执行时间和效率的可扩展性都要优于其他 3 种算法。



(a)合成数据集

(a) Synthetic dataset



(b)真实数据集

(b) Real dataset

图 3 聚类算法效率对比

Fig. 3 Comparison of clustering efficiency

4.3 聚类质量测试

本文提出综合正确率来度量聚类结果的合理性。利用解决指派问题的方法(如匈牙利法)识别聚类结果与原始数据集中种类的对应关系,构造出综合正确率度量矩阵,该矩阵中的元素 $e_{i,j}$ 表示原始种类 i 中有 $e_{i,j}$ 个文件分配给了簇 j,该矩阵对角线上的元素即为划分正确的文件数量,综合正确率(compositive accuracy)的计算方法为

$$\text{综合正确率} = \frac{\sum_{i=1}^{T}\left(\frac{e_{i,i}}{\sum_{k=1}^{T} e_{i,k}} \times \frac{e_{i,i}}{\sum_{k=1}^{T} e_{k,i}}\right)}{T} \quad (9)$$

其中 T 为原始数据集的种类数量,$e_{i,i}/\sum_{k=1}^{T} e_{i,k}$ 的意义是第 i 个簇中划分正确的文件数量占原始种类数量的比例,$e_{i,i}/\sum_{k=1}^{T} e_{k,i}$ 的意义是第 i 个簇中划分正确的文件数量占该簇中所有文件数量的比例。这个指标既考虑了文件是否划分到正确的类别,也考虑了簇的纯度。

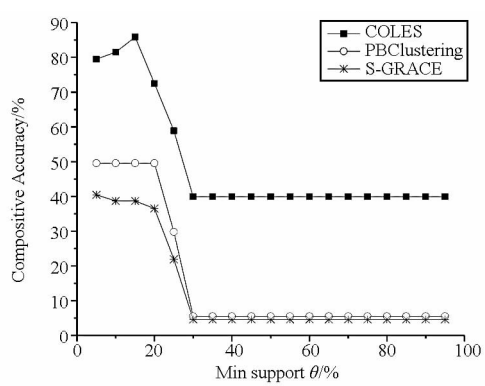
本部分实验改变最小支持度和种类数量分别进行了如下实验(由于 XLink 中没有采用最小支持度的概念,因此在变化最小支持度的实验中不进行对比):

1)顺序读取全部文件,逐渐增大最小支持度 θ,测试综合正确率;

2)首先取表 2(合成数据集)中前 2 种的全部 XML 文档,按照表 2 的顺序依次增加 XML 文档种类,直至全部增加完毕,取 $\theta = 15\%$,测试综合正确率。对于真实数据集的测试类似。

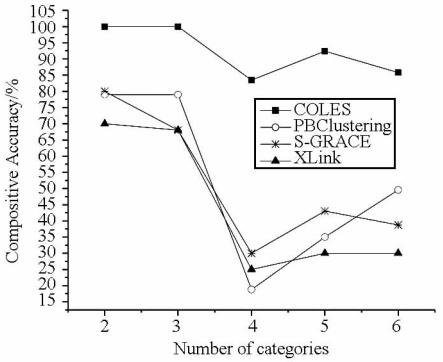
图 4(a)显示了合成数据集中,随着最小支持度的逐渐增大,COLES 比 PBClustering 和 S-GRACE 的聚类质量都要高。这是因为:1)COLES 采用子树作为特征进行聚类,这比采用路径(PBClustering)和边(S-GRACE)作为特征进行相似性度量的方法更加准确,更加从本质上反映标签树之间的相近与区别;2) PBClustering 和 S-GRACE 使用的是凝聚层次聚类法,对聚类过程中所形成的正交向量组只能按照就近的次序合并文件或簇,而 COLES 使用特征子树对正交向量组有很好的指导作用,使聚类结果更自然,效果更好。在 $\theta = 15\%$ 时,COLES 聚类的准确度达到了 85.8%,大大超过了 PBClustering(49.53%)和 S-GRACE(38.7%)。在 $\theta \geq 30\%$ 时,由于提取出的特征数量为0,所以3 种算法的综合正确率全部下降。

图 4(b)显示了随着合成数据种类数量的增加,COLES 算法的聚类准确度始终高于其他 3 种算法。在种类个数等于 4 时,增加了 ACM SigmodRecord 种类,从表 2 可知,这种类型只有 50 个文件,致使文档分布不均衡,造成了聚类质量下降,但 COLES 算法的准确度降幅不大(16.5 个百分点),小于 S-GRACE(38 个百分点),XLink(43 个百分点)和 PBClustering(60.25 个百分点)的降幅,可见在文档分布不均衡条件下,COLES 仍然能够保证一定的聚类质量。



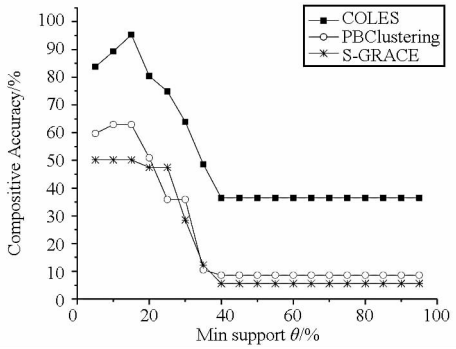
(a)合成数据集最小支持度 vs 综合正确率

(a) Min support vs compositive accuracy on synthetic dataset



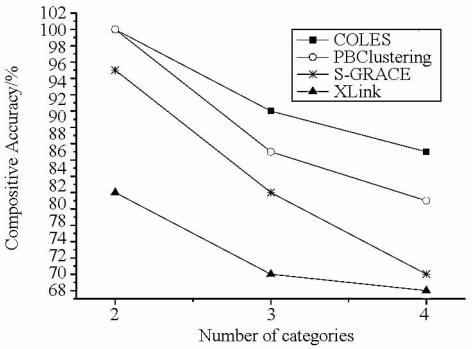
(b)合成数据集种类数量 vs 综合正确率

(b) Number of categories vs compositive accuracy on synthetic dataset



(c)真实数据集最小支持度 vs 综合正确率

(c) Min support vs compositive accuracy on real dataset



(d)真实数据集种类数量 vs 综合正确率

(d) Number of categories vs compositive accuracy on real dataset

图 4　聚类质量对比

Fig. 4　Comparison of quality of clustering

图 4(c)和(d)显示在真实数据集上,COLES 算法在聚类准确性方面仍然要优于其他 3 种算法。

5　总结与下一步工作

本文以军队生长干部招生计划分析评估为研究背景,采用聚类的方法支持计划的波动性分析。在分析了招生计划的特点基础上,提出将计划量化为标签树,再基于特征共现进行标签树的聚类方法。该方法充分利用了特征具有较强指导性这一特点,运用形成核心再指导分类的策略达到了聚类效率高、聚类质量好的效果。实验从合成数据和真实数据两方面证明了方法的有效性。

下一步工作将结合本项研究开展对招生计划的波动性分析,在实践中不断改进该方法。

参考文献(References)

[1] Guillaune D, Murtagh F. Clustering of XML documents[J]. Computer Physics Communications,2000,127(2-3):215-227.

[2] Cobena G, Abiteboul S, Marian A. Detecting changes in XML document[C]//Proceedings of the 18th International Conference on Data Engineering (IEEE ICDE'02). 2002:41-52.

[3] Wang Y, De Witt D, Cai J. X-Diff: A fast change detection algorithm for XML documents[C]//Proceedings of the 19th International Conference on Data Engineering (IEEE ICDE'03),2003:519-530.

[4] Chawathe S, Rajaraman A, Garcia-Molina H, et al. Change detection in hierarchically structured information [C]// Proceedings of ACM SIGMOD International Conference on Management of Data(SIGMOD'96),1996:493-504.

[5] Betino E, Guerrini G, Mesiti M. A matching algorithm for measuring the structural similarity between an XML document and a DTD and its applications[J]. Information Systems, 2004,29(1):23-46.

[6] Zhang K, Shasha D. Simple fast algorithms for the editing distance between trees and related problems[J]. SIAM J. Computing, 1989, 18(6):1245-1262.

[7] Nierman A, Jagadish H V. Evaluating structural similarity in XML documents [C]//Proceedings of Fifth International Workshop Web and Databases,2002.

[8] LianW, Cheung D W, Mamoulis N, et al. An efficient and scalable algorithm for clustering XML documents by structure [J]. IEEE Transactions on Knowledge and Data Engineering, 2004,16(1):82-96.

[9] Flesca S, Manco G, Masciari E, et al. Fast detection of XML structural similarity[J]. IEEE Transactions on Knowledge and Data Engineering,2005,17(2):160-175.

[10] Yoon J P, Raghavan V, Chakilam V. Bitmap indexing-based clustering and retrieval of XML documents[C]//Proceedings of ACM SIGIR Workshop on Mathematical /Formal Methods in Information Retrieval,2001.

[11] Kozielski M. Improving the results and performance of clustering bit-encoded XML documents[C]//Proceedings of the 6th IEEE International Conference on Data Mining-Workshops(ICDMW'06),2006.

[12] Liu J, et al. XML clustering by principal component analysis [C]//Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence(ICTAI'04),2004.

[13] Leung H P, Chung F L, Chan S C F, et al. XML document clustering using common XPath[C]//Proceedings of the 2005 International Workshop on Challenges in Web Information Retrieval and Integration(IEEE WIRI'05),2005.

[14] Tekli J,Chbeir R,Yetongnon K. An overview on XML similarity: background, current trends and future directions [J]. Computer Science Review, 2009,3(3):151-173.

[15] Augsten N, Bohlen M, Gramper J. The pq-Gram distance between ordered labeled trees [J]. ACM Transactions on Database Systems, 2009.

[16] JDom[EB/OL]. http:// jdom. org.