文章编号: 1001- 2486(2010) 03- 0082- 07

一种面向混合数据集可视化的高效数据转换技术*

孙 扬, 封孝生, 周 城, 汤大权, 肖 卫东
(国防科技大学 信息系统与管理学院, 湖南 长沙 410073)

摘 要: 应用领域中存在大量多数据类型属性的混合数据集, 但是, 很多有效多变元可视化方法的适用范围都只局限于单一类型, 对于混合数据集可视化效果不甚理想。针对包含数值及分类型属性的多元混合数据集, 提出一种面向混合数据集可视化的数据转换技术, 首先对每一数值型属性使用聚类技术进行分类化, 然后应用对应分析算法量化所有分类型属性, 最后将转换后的混合数据集使用经典的数值型可视化方法——星形坐标法进行展现, 并且针对变元数量较多或分类型变元势较高的混合数据集, 在数据转换过程中提出一套降势策略, 减少参与计算的变元数量, 提高计算效率。实验表明, 该方法对混合数据集的可视化结果不仅易于理解, 而且有利于用户发现其中的隐性知识, 降势策略在提高内存及时间效率方面作用显著。

关键词: 混合数据可视化; 降势; 对应分析; 聚类; 数据转换技术; 星形坐标
中图分类号: TP391 文献标识码: A

An Efficient Data Transformation Technique for Mixed Data Visualization

SUN Yang, FENG Xiao-sheng, ZHOU Cheng, TANG Da-quan, XIAO Wei-dong
(College of Information System and Management, National Univ. of Defense Technology, Changsha 410073, China)

Abstract: There are abundant mixed data sets with various types of attributes in application fields. However, most multivariate data visualizations are only effective with simplex one data type. As for mixed data sets, the visualizations of them are usually dissatisfied. We present a data transformation technique for mixed data sets involving both numerical and categorical attributes. Firstly, every numerical attribute was categorized by clustering; then, all categorical attribute was quantified by Correspondence Analysis; finally, the transformed mixed data were presented in numerical data visualizations like Star Coordinates. Furthermore, aiming at those mixed data sets that have many attributes or the cardinality which is high, a set of cardinality reduction strategies were proposed to diminish the attributes number involved in computation to improve computational efficiency. Empirical studies show that the visualization of mixed data sets is easily-understandable and propitious for the user to discover the connotative information within; and that cardinality reduction strategies are highly memory-saving and time-efficient.

Key words: mixed data visualization; cardinality reduction; correspondence analysis; clustering; data transformation; star coordinates

随着信息获取及存储技术的飞速发展, 科学、工程及商业等应用领域产生多元信息的数据量越来越多, 使得对其进行全面深入研究分析的难度日益增大, 因此, 多变元可视化技术作为有效、直观、可交互的多元数据集展示工具被广泛应用于数据分析、数据挖掘、信息认知及知识发现过程中, 极大地减轻了用户的认知复杂性, 提高了用户认知能力[1]。但是, 大部分有效的多变元可视化方法只适用于单一数据类型, 如数值型或分类型(如, 颜色属性的分类值包括红、黄、蓝等) 等, 无法直接处理同时包含数值型和分类型变元(属性) 的多元混合数据集。

针对如何重用现有方法可视化混合数据集的问题, 至少存在两种解决思路。第一种是把数值型变元分裂成 bins[2], 即分类化数值型变元, 然后使用分类型可视化方法展示混合数据集; 第二种是为每一个分类型变元的分类值指定顺序及距离(数值与分类值的区别就在于前者存在自然的顺序和距离关

* 收稿日期: 2009- 10- 26
基金项目: 国家自然科学基金资助项目(60903225); 国防科技大学优秀研究生创新基金资助项目(B080503)
作者简介: 孙扬(1983—), 男, 博士生。

系),即量化分类型变元,然后使用数值型可视化方法展示混合数据集。但是,若采用人工方法分类化数值型变元或量化分类型变元,可视化效果会依赖于变元类型转换人员的专业知识,随意或简单地进行变元类型转换会在可视化过程中引入相关人员的主观臆断,产生对可视化结果的错误理解;更重要的是,分类型可视化方法通常只能表现数据集的统计信息,无法充分展示数据集的多元分布特征。因此,Johansson 等[3]使用对应分析(Correspondence Analysis, CA)[4]算法量化混合数据集中分类值,而后利用交互式数值型可视化工具对混合数据集进行分析。

Johansson 所用方法重点在于结果的交互分析过程,虽然他也提到为将数值型变元信息引入分类型变元量化过程,在量化前使用聚类算法分类化数值型变元使其也可以参与分类型变元量化计算,但对该过程没有详细描述,尤其是聚类技术的应用方式,而且,针对变元数量较多或分类型变元势较高(包含较多分类值)的混合数据集,未给出方案解决 CA 较严重的内存和时间消耗问题[5]。因此,针对上述不足,本文提出一种面向多元混合数据集可视化的高效数据转换技术,详细描述应用聚类技术分类化数值型变元,然后利用 CA 量化分类型变元,最后使用数值型星形坐标法(Star Coordinates)[6]可视化混合数据集的过程,并提出一套降势策略预处理变元数量较多或分类型变元势较高的混合数据集,以减少参与 CA 计算的变元数量或势。实验表明,本文提出的数据转换技术与任意量化分类型变元相比,可视化结果对混合数据集隐性信息的展现更加直观、准确;降势策略对降低CA 内存占用量的效果明显,对减少 CA 计算时间也起到一定作用。

1　相关工作

研究人员已提出若干种分类型数据可视化方法,其中最常用的一种实现思路是使用各种正比于分类值相应频率的可视元素表示分类型数据集。如,为可视化二元列联表[4]中分类型变元及分类值间的关系,Friendly 设计的 sieve diagrams 以列联表中的每一个单元格对应一个包含多个方格的长方形,长方形的宽正比于该列的边缘概率,高正比于该行的边缘概率,面积正比于相关分类值的期望概率,包含的方格数量正比于其观测概率,颜色标识两个概率的偏离方向[7]。mosaic displays[8]采用正比于列联表中单元格观测概率的长方形展现分类型数据集的统计信息。mosaic matrix 理论上能够用于可视化具有大量变元和分类值的数据集,但实际上,它和 sieve diagrams 具有相同的缺陷,而且在单一视图中也难以表达列联表中的全部信息。通过改进 Parallel Coordinates[9]得到的 Parallel sets[2]使用平行排列的正比于分类值观测概率的包围盒来可视化数据集的每一个分类变量,而且,将连续变量进行分段(bins),该方法勉强能可视化混合数据集,但其支持的变元数量较少,并且过多的分段会导致包围盒过于狭窄,使可视化结果退变得非常模糊。上述所有方法都适用于分析完全分类型数据集中各变元间的关系及关联信息,揭示其中的隐性知识,并且不需要对分类型变元进行映射操作,但是,这些方法在展现分类型数据集的统计和观察信息的同时,却丢失了数据集的多元分布信息。另外,它们都不能支持实际应用中常见的具有较多变元和分类值的数据集,也无法直接应用于混合数据集。

另一方面,很多有效的数值型可视化方法也取得了长足进步,如 Parallel Coordinates[9]、Star Coordinates[6]等,它们适于揭示多元数值型数据集中的奇异值、聚类及数据间的关系等隐性信息,但其存在共同缺陷,即,仅能处理数值型数据,如果数据集中存在分类型数据,这些方法只进行很简单的处理,如,任意指定分类值对应的数值,或按照分类值的字母顺序或其他变元(如时间)的值指定其顺序后再对应为一组数值等。显然,这些处理都会引入人为因素,产生对可视化效果的错误理解,且增加了分类值的等距性假设,导致量化结果无法体现分类值的相似性。为解决上述问题,Rosario 等使用 CA 技术基于分类值的间距和关联对分类变元进行量化[5],从而将分类型数据集输入数值型可视化工具,但他们的讨论只局限于单纯的分类型数据集,未涉及混合数据集的处理方法。

2 面向混合数据集可视化的数据转换技术

2.1 假设与符号

事实上,多元混合数据集包含的数据类型不只局限于数值型和分类型,但大体可分为基本类型(如整型、浮点型、双精度型、字符及字符串型等)和高级类型(如比率型、二元型、间隔型、序数型和分类型等)两类,且根据变元间的语义信息,不同数据类型都可映射转换为数值型和分类型[10],因此,本文只讨论包含上述两种通用数据类型的多元混合数据集,同时,为下文便于叙述引入如下假设及符号:

(1)多元混合数据集 $G(X)$ 中的对象 $\{X_1, X_2, \cdots, X_n\}$ 来自相同应用领域,由一组相同变元(属性) $A_1, A_2, \cdots, A_m$ 进行描述,变元 A_i 的值域记作 $DOM(A_i)$。如果 A_i 是数值型变元,则 $DOM(A_i)$ 是连续的;如果 A_i 是分类型变元,$DOM(A_i)$ 是一组有限而且无序的分类值,对于任意分类值 $a_{i1}, a_{i2} \in DOM(A_i)$,或者 $a_{i1} = a_{i2}$,或者 $a_{i1} \neq a_{i2}$,不存在其他关系;

(2) $G(X)$ 中任意对象 X 可逻辑表示为一个向量 $[x_1^c, x_2^c, \cdots, x_p^c, x_{p+1}^r, \cdots, x_m^r]$,其中,$x_j \in DOM(A_j)$,$1 \leqslant j \leqslant m$,前 p 个元素是分类型,其余的是数值型,并且假设 $G(X)$ 中的每个对象 X 都由完整的 m 个变元描述,不存在变元值缺失情况。

2.2 对应分析

对应分析(CA)是一种用于分析同一变元各分类值间差异及不同变元各分类值间对应关系的多元统计方法。CA 在原理上类似于主成分分析(PCA),但 CA 针对分类型变元,而 PCA 只作用于数值型变元。CA 可按其作用对象划分为适用于两个分类型变元的 Simple Correspondence Analysis(SCA),及适用于多分类型变元的 Multiple Correspondence Analysis(MCA)。SCA 的输入是二元列联表(Two-way contingency table),而 MCA 的输入是多元列联表(Multi-way Burt table)[4]。称列联表中列头中包含的变元为目标变元,行头中包含的变元为分析变元。

若将列联表的各列定义为多维空间的维度,则表中的样本就可视为该空间中的多维数据点。*CA* 通过提取多维空间的独立维对原空间进行降维,这样每一分类值可以几个主维度所代表的点在低维可视空间中直观展现,从而简洁明了地揭示各分类值间的相关信息。

2.3 数值型变元分类化

CA 实质在于充分利用样本集中隐含存在的统计特征量化全部分类型变元的所有分类值,使量化后的分类型变元能尽量反映及保持或加强该统计特征。而对于混合数据集,CA 原本只能处理分类型变元,但若能在 CA 过程中同时考虑数值型变元信息,则对分类型变元的量化将更加准确。因此,首先对混合数据集中数值型变元进行分类化,若使用手动交互方式,需要分析人员具有丰富的领域知识,而且,对于数值型变元较多的大型数据集,该方法也不切实际;因此,本文采用具有无监督特性的聚类技术,依据每一个数值型变元 A_k $(p < k \leqslant m)$ 对混合数据集所有对象进行聚类,然后自动根据不同聚类分类化数值型变元的方法,这样对于混合数据集就会产生形如表 1 的列联表,其中,$a_{i\cdot} \in DOM(A_i)$,$a_{j\cdot} \in DOM(A_j)$,$bin_{k\cdot}$ 表示数值型变元 A_k 聚类后形成的一个分类,$c_{i\cdot j\cdot}$ 表示 $G(X)$ 中同时包含 $a_{i\cdot}$、$a_{j\cdot}$ 的对象数量,$c_{i\cdot k\cdot}$ 表示 $G(X)$ 中既包含 $a_{i\cdot}$ 又属于 $bin_{k\cdot}$ 所代表聚类的对象数量。

鉴于 k-means 算法对大型数值型数据集聚类的高效性,本文选取该算法量化每一个数值型变元,当然,也可使用其他的数值型聚类算法代替 k-means,如 DBSCAN[11] 等,但是由于无监督聚类方法对应用领域的依赖程度较高,因此用户需依据目标数据集的特点及算法性能进行选择,此内容已超出本文范围,不予讨论。

2.4 分类型变元量化

所有数值型变元分类化后,对形如表 1 的列联表应用 MCA 量化所有分类型变元的分类值,为能使每一个分类值对应于一个数值,只取最大反映原空间特点的第一主维度,使用最优尺度法[4] 将每一个分类值映射为一个数值,这样,也就能够使用有效的数值型可视化工具对混合数据集进行展示。

表 1　数值型变元分类化后的列联表

Tab. 1　Burt table categorizing continuous variable

	a_{j1}	a_{j2}	…	bin_{k1}	bin_{k2}	…
a_{i1}	c_{i1j1}	c_{i1j2}	…	c_{i1k1}	c_{i1k2}	…
a_{i2}	c_{i2j1}	c_{i2j2}	…	c_{i2k1}	c_{i2k2}	…
…	…	…	…	…	…	…

表 2　RCCA 的列联表

Tab. 2　Burt table of RCCA

	a_{j1}	a_{j2}	…	$cluster^r_1$	$cluster^r_2$	…
a_{i1}	c_{i1j1}	c_{i1j2}	…	c_{i1r1}	c_{i1r2}	…
a_{i2}	c_{i2j1}	c_{i2j2}	…	c_{i2r1}	c_{i2r2}	…
…	…	…	…	…	…	…

3　降势策略

随着混合数据集变元数量的增加以及分类型变元势的升高,MCA 的内存和时间耗费问题越来越严重。Rosario 等提出使用 Focused Correspondence Analysis(FCA)[5] 替代 MCA 处理变元数量较多或势较高的分类型数据集。MCA 依据所有变元量化目标变元,而 FCA 只使用与目标变元最相关的 k 个分析变元量化目标变元,因此,FCA 与 MCA 相比更加节约内存,但是,FCA 无法同时分析所有变元,导致了 FCA 的时间效率比 MCA 略差,而且,由于 FCA 丢掉部分相关性低的分析变元,也就引入相应的信息损失。更重要的是,FCA 没有涉及混合数据集。因此,我们提出一套比 MCA 占用内存少、比 FCA 效率高的降势策略,以减少参与 CA 计算的变元数量,从而解决变元数量较多或势较高的混合数据集 CA 过程的内存和时间耗费问题。

上节提到的 k-means 分别作用于每个数值型变元,因此,列联表中分析变元数量与混合数据集变元数量相等。若混合数据集包含数值型变元较多,列联表会包含很多列。因此,在降势策略中将 k-means 直接作用于所有数值型变元产生一组聚类,然后直接以该组聚类作为所有数值型变元的分类值导入列联表,就减少了列联表的列数,如表 2。该策略称为 r-Clustering based Correspondence Analysis(RCCA)。

若混合数据集主要由分类型变元组成,则可利用分类型聚类算法减少列联表分析变元的数量,列联表中分类值由根据所有分类型变元产生的聚类所代替。虽然有很多分类型聚类算法可以使用,但考虑到算法的实现简单和高效性,本文选用 k-mode 算法[10]。这样,列联表就转换为表 3 的形式,这种策略称为 c-Clustering based Correspondence Analysis(CCCA)。由于 RCCA 和 CCCA 相互独立,所以可将它们进行组合,同时对数值型和分类型变元分别应用 k-means 和 k-mode 进行聚类,形成如表 4 的列联表,这就是 cr-Clustering based Correspondence Analysis(CRCA)策略。

表 3　CCCA 的列联表

Tab. 3　Burt table of CCCA

	$cluster^c_1$	$cluster^c_2$	…	bin_{k1}	bin_{k2}	…
a_{i1}	c_{i1c1}	c_{i1c2}	…	c_{i1k1}	c_{i1k2}	…
a_{i2}	c_{i2c1}	c_{i2c2}	…	c_{i2k1}	c_{i2k2}	…
…	…	…	…	…	…	…

表 4　CRCA 的列联表

Tab. 4　Burt table of CRCA

	$cluster^c_1$	$cluster^c_2$	…	$cluster^r_1$	$cluster^r_2$	…
a_{i1}	c_{i1c1}	c_{i1c2}	…	c_{i1r1}	c_{i1r2}	…
a_{i2}	c_{i2c1}	c_{i2c2}	…	c_{i2r1}	c_{i2r2}	…
…	…	…	…	…	…	…

既然单一类型聚类算法可分别同时应用于混合数据集以减少列联表列数,当然也可使用混合聚类算法同时对两种类型的变元进行聚类,如本文使用的 k-prototype[10],该策略称为 Mixed-Clustering based Correspondence Analysis(MCCA)。然而,混合聚类算法数量较少,且大都具有聚类结果不稳定、随机性大、准确性低等缺点[12],而新兴的聚类融合技术用若干独立的聚类器分别对原始数据进行聚类,然后对这些结果进行组合,最终获得对原始数据的聚类结果,从而屏蔽了数据集的原始特征,且具有良好的稳定性和处理不规则或噪声数据的能力,因此可应用于混合数据集的聚类[12]。该策略称为 Cluster Ensemble based Correspondence Analysis(CECA)。

4 实验结果

通过对比基于分类值任意量化(间距相等,任意排序)与基于 MCA 量化的混合数据集可视化结果验证本文提出的数据转换技术的有效性,并比较 MCA, FCA, RCCA, CCCA, CRCA, MCCA 和 CECA 针对相同数据集的内存使用及运行时间,证明降势策略在提高内存和时间效率方面的突出作用。我们在 MCA 和 FCA 原有定义的基础上,扩充分类化数值型变元,从而使 FCA 可处理混合数据集。鉴于在数据分析初期,相比平行坐标系,星形坐标系能更好辅助用户发掘数据集所隐含的趋势、奇异值和聚类信息,因此本文选其进行实验。

实验数据集使用真实的汽车和国旗数据集,汽车数据集包含 205 种品牌的汽车,每条记录包含 15 个数值型变元,10 个分类型变元和 1 个整数型变元(可映射为分类型变元)。国旗数据集包含 194 个国家的国旗,每条记录包含 2 个数值型变元,7 个分类型变元,1 个国家名称,8 个整数型变元和 12 个布尔型变元(最后三种数据类型变元都可以映射到分类型变元)。

4.1 可视化效果对比分析

由于汽车数据较典型且易于解释、理解,因此我们以其为基础比较两种量化方法的可视化效果。为使对比更加明显,选取分析其中 6 个分类型变元,即 make、body-style、drive-wheel、engine-type、numr-of-cylinders 和 fuel-system,以及 7 个数值型变元,即 wheel-base、highway-mpg、length、city-mpg、horsepower、price 和 engine-size。

图 1 是利用任意量化在没有进行任何交互的情况下最初的可视化效果,各代表点分布混乱,用户无法直观获得数据集的隐含信息。图 2 是相同阶段应用 MCA 进行量化的结果,用户首先能够直观发现任意变元所有分类值的相异(似)度,如,fuel-systems 属性中 2bbl 与 spdi, idi 与 mpfi 非常相近而 1bbl 与 mpfi 差距非常大,对于汽车品牌(make),图 3 清晰展示了其量化结果,所有汽车品牌明显地聚为 5 类,根据 make 的聚类情况对图 1、2 中的数据点进行标注,Mercedes-bens 类使用正三角形表示,Jaguar 类以点表示,Audi 类以倒三角表示,Nissan 类以方形表示,Mazda 类以圆点表示。观察图 2 可发现每一类汽车在星形坐标系中大都聚集在一起,进一步分析,可发掘数据集中更多的隐含信息(如:奇异点及聚类粒度等),而这些信息在图 1 中无法直观体现。(若在 MCA 之前应用降势策略则产生的可视化效果图与图 2 略有不同,但总体效果较为相近,因此不再赘述)。

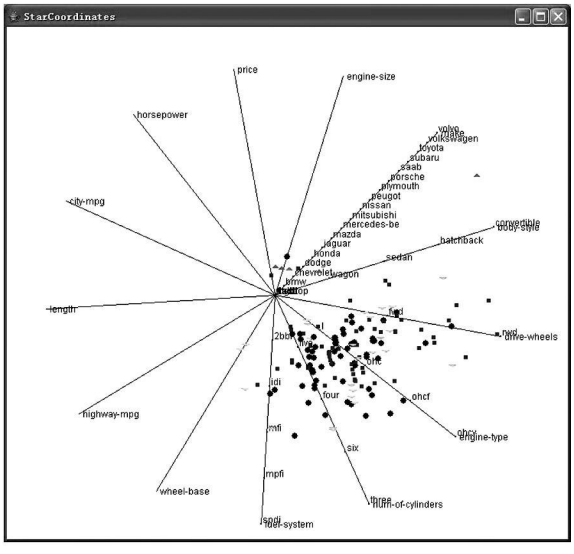

图 1 任意量化可视化
Fig. 1 Visualization of arbitrary quantification

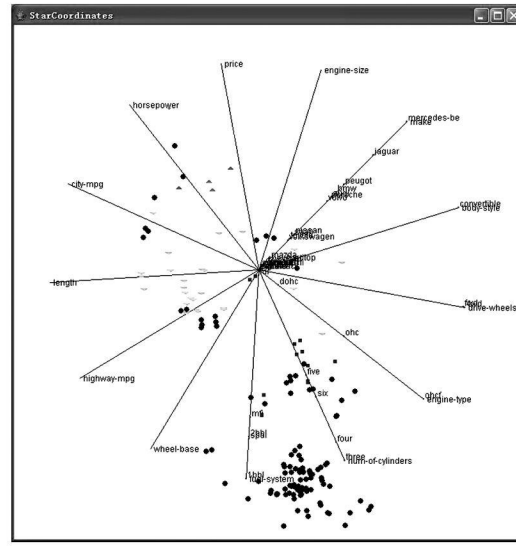

图 2 MCA 为基础的量化可视化
Fig. 2 Visualization of MCA-based quantification

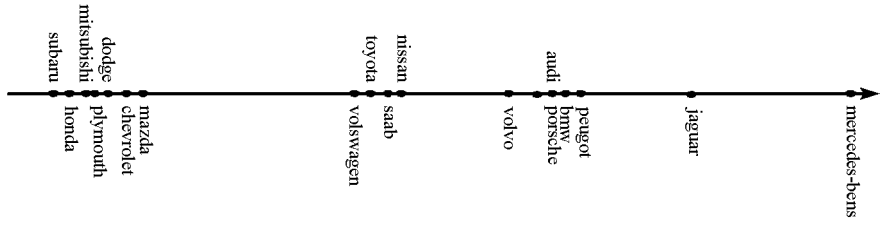

图 3　变元 make 的量化结果图
Fig. 3　Quantification of variable make

4.2　内存及时间性能比较

Rosario[5] 指出在其可视化系统中 CA 最为耗费内存, 因此, 我们对比每一种降势策以及 MCA、FCA 中 CA 所需内存。在忽略所有内存优化技术的前提下, MCA 在 CA 过程中使用最多内存, 因此使用公式 $\frac{\mathrm{FCA/RCCA/CCCA/CRCA/MCCA/CECA\_MemorySpace}}{\mathrm{MCA\_MemorySpace}} \times 100\%$ 归一化所有降势策略 CA 所需内存, 如图 4 所示。显然, 由于减少 CA 列联表中分析变元的个数, FCA 的内存占用率要优于 MCA, 而不同降势策略对不同种类数据集会产生不同的效果, 例如, RCCA 对国旗数据集的作用不十分明显, 甚至比 FCA 还差, 而 CCCA 对国旗数据集的效果却非常突出, 这是因为国旗数据集只包含两个数值型变元, 而大部分变元是分类型的。对于汽车数据集, 由于两类变元比例基本相同, 因此 RCCA/CCCA 的效果相似且都优于 FCA。CRCA 由于结合了 RCCA 和 CCCA 的优点, 因此, 它比二者效果都要好。另外, 与其它过程相比, MCCA 和 CECA 都节省了最多的内存。

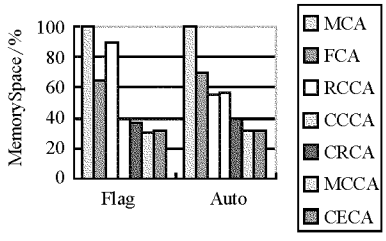

图 4　不同策略的内存使用情况
Fig. 4　Memory-using of different strategies

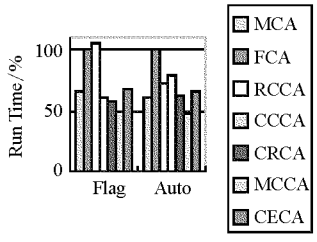

图 5　不同策略的运行时间
Fig. 5　Run times of different strategies

图 5 对比不同降势策略归一化运行时间。一般情况, FCA 无法同时分析所有分类型变元且其列联表中列较多, 导致其 CA 耗费更多时间, 因此, 使用 $\frac{\mathrm{MCA/RCCA/CCCA/CRCA/MCCA/CECA\_RunTime}}{\mathrm{FCA\_RunTime}} \times 100\%$ 归一化不同降势策略的运行时间。虽然本文提出的降势策略也无法同时分析所有分类型变元, 但由于其列联表中列数较少使得其 CA 时间较少, 且所使用聚类算法都是很高效的, 因此比 FCA 效率高。然而, 对于国旗数据集, 由于 RCCA 降势效果不明显, 其运行时间比 FCA 还要长。作为 RCCA 和 CCCA 的有效组合, CRCA 对两套数据集的时间效率都较高。MCCA 通过减少列联表列数抵消了其无法同时计算所有分类型变元的时间消耗, 因此在某些情况下比 MCA 时间效率更高。尽管 CECA 与 MCCA 输入的列联表形式相同, 但由于聚类融合的最优化过程比一般混合聚类算法耗费更多时间, 因此 CECA 比 MCCA 的效率略低。

虽然 MCCA 比其它降势策略的内存和时间效率高, 但由于上面提到的局限性, 因此不推荐使用 MCCA 进行降势, 用户最好根据数据集的结构组成选择降势策略, 或使用 CECA/CRCA 作为缺省选择。

5　结论与进一步工作

本文提出一种面向混合数据集可视化的高效数据转换技术。首先, 详细说明应用 CA 技术量化分类型变元的过程; 然后针对变元数量较多及部分分类型变元势较高的混合数据集提出了一套降势策略; 而后使用星形坐标系有效地对混合数据集进行可视化。实验结果表明, 本文提出的数据转换技术可视

化效果明显优于任意量化方法,降势策略对不同结构数据集的效果不尽相同,但在大部分情况下其时间及内存效率性能较为突出。

下一步对于降势策略将会进行更具体的评估以证明它的有效性及高效性。此外,在混合数据集中如何选取适当变元可视化能更准确地发掘隐含信息也是一个值得讨论的开放性问题。

参 考 文 献:

[1] Gershon N D, Eick S G. Information Visualization[J]. IEEE Computer Graphics and Applications, 1997, July/Augusts: 29- 31.

[2] Bendix F, Kosara R, Hauser H. Parallel Sets: Visual Analysis of Categorical Data[C]// Proc. IEEE Symposium on Information Visualization 2005. Los Alamitos, IEEE Computer Society, 2005: 133- 140.

[3] Johansson S, Jern M, Johansson J. Interactive Quantification of Categorical Variables in Mixed Data Sets[C]// Proc. the 12th International Conference Information Visualization. Washington, DC: IEEE, 2008: 3- 10.

[4] Greenacre M. Correspondence Analysis in Practice[M]. 2nd ed. Chapman & Hall, 2007.

[5] Rosario G E, Rundensteiner E A, Brown D C, et al. Mapping Nominal Values to Numbers for Effective Visualization[J]. Information Visualization, 2004, 3(2): 80- 95.

[6] Kandogan E. Visualizing Multi-dimensional Clusters, Trends, and Outliers Using Star Coordinates[C]// Proc. of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2001: 107- 116.

[7] Friendly M. Visualizing Categorical Data[M]. New York, SAS Publishing, 2000.

[8] Friendly M. Mosaic Displays for Multi-way Contingency Tables[J]. Journal of the American Statistical Association, 1994, 89:190- 200.

[9] Inselberg A. The Plane with Parallel Coordinates[J]. The Visual Computer, 1985, 1(2): 69- 91.

[10] Huang Z X. Extensions to the *k*-means Algorithm for Clustering Large Data Sets with Categorical Values[J]. Data Mining and Knowledge Discovery, 1998, 2(3):283- 304.

[11] 刘青宝,侯东风,邓苏,等. 基于相对密度的增量式聚类算法[J]. 国防科技大学学报, 2006, 28(5): 73- 79.

[12] 赵宇,李兵,李秀,等. 混合属性数据聚类融合算法[J]. 清华大学学报(自然科学版), 2006, 46(10): 1673- 1676.

(上接第 10 页)

参 考 文 献:

[1] 曲广吉,于登云,曾辛. 航天器空间对接动力学分析仿真软件 DODASS 及其工程应用[J]. 航天器工程, 1995, 4(4): 1- 10.

[2] 于登云,曲广吉,曾辛,等. 航天器对接接触过程撞击动力学分析[J]. 空间科学学报, 1998, 18(1): 62- 68.

[3] 关英姿,崔乃刚,刘育华. 航天器对接中接近至首次接触阶段的数值仿真[J]. 上海航天, 1999(1): 21- 26.

[4] 关英姿,崔乃刚,刘育华. 空间航天器对接动力学的数值仿真研究[J]. 哈尔滨工业大学学报, 1999, 31(4):121- 124.

[5] 关英姿,康为民,崔乃刚. 空间对接预捕获阶段的建模与仿真[J]. 系统仿真学报, 2000, 12(6):664- 667.

[6] 赵阳,王萍萍,田浩,等. 考虑惯性特性的对接机构缓冲系统特性研究[J]. 哈尔滨工业大学学报, 2004, 36(1):1- 3.

[7] 杨芳,曲广吉,杨雷. 空间对接中差动式缓冲阻尼机构的建模研究[J]. 中国空间科学技术, 1999(1):1- 8.

[8] 朱仁璋. 航天器交会对接技术[M]. 北京:国防工业出版社, 2007.

[9] Fehse W. Automated Rendezvous and Docking of Spacecraft[M]. Cambridge University Press, 2003.