高维数据聚类方法综述 *

贺 玲[1a], 蔡益朝[1b], 杨 征[2]
(1. 空军雷达学院 四系 a. 计算机教研室; b. 自动化教研室, 武汉 430019; 2. 国防科学技术大学 信息系统与管理学院, 长沙 410073)

摘 要: 总结了高维数据聚类算法的研究现状,分析比较了算法性能的主要差异,并指出其今后的发展趋势,即在子空间聚类过程中融入其他传统聚类方法的思想,以提高聚类性能。

关键词: 高维数据; 聚类; 子空间
中图分类号: TP392 **文献标志码**: A **文章编号**: 1001-3695(2010)01-0023-04
doi:10.3969/j.issn.1001-3695.2010.01.006

Survey of clustering algorithms for high-dimensional data

HE Ling[1a], CAI Yi-chao[1b], YANG Zheng[2]
(1. *a. Computer Science Teaching & Research Section, b. Automation Teaching & Research Section, Fourth Department, Air Force Radar Academy, Wuhan* 430019, *China*; 2. *School of Information System & Management, National University of Defense Technology, Changsha* 410073, *China*)

Abstract: This paper provided a survey of current clustering algorithms for high-dimensional data at first, then made a comparison among them and identifized the new direction in the future, which was the combination of subspace clustering and other typical clustering methods.

Key words: high-dimensional data; clustering; subspace

聚类是一种重要的数据分析手段,它按照一定的要求和规律对数据集中的数据对象进行区分和分类,进而把一个没有类别标记的数据集按照某种准则划分成若干个子集(类),并使相似的数据对象尽可能地归为一类、不相似的数据对象尽可能地划分到不同的类中。通过聚类分析,能有效地发现隐含在数据集中的数据分布特性,从而为进一步充分、有效地利用数据奠定良好的基础。与此同时,随着信息技术的迅猛发展,聚类所面临的不仅是数据量越来越大的问题,更重要的还是数据的高维度问题。换句话说,由于数据来源的丰富多样,图文声像甚至视频都逐渐成为聚类处理的目标对象,这些特殊对象的属性信息往往要从数十个甚至数百个方面来表现,其每一个属性都成为数据对象的一个维,对高维数据的聚类分析,已成为众多领域研究方向之一。

在与高维数据相关的应用领域,维度灾难(curse of dimensionality)是一个非常普遍的现象。这一术语最先由 Bellman 提出,它泛指在数据分析中遇到的由于变量(属性)过多而引起的一系列问题。此后又有很多研究者做了大量的研究致力于减小甚至消除维度灾难对高维数据处理的影响[1,2]。本文即以此为出发点,分析比较了高维数据聚类方法的研究现状,总结了其中存在的问题,并指出了今后的发展趋势。

1 现有的高维聚类方法

数据挖掘领域对聚类算法的研究已经取得了一定的成果,很多传统的聚类算法在对一般的低维数据进行聚类处理时,通常能获得较为准确的结果,而对于高维数据,由于维度灾难的影响,若采用这些传统的算法进行聚类,往往得不到所期望的结果。为了满足不同应用领域中众多用户的需求,研究者们提出了很多针对高维数据的聚类方法,本文将它们分成基于降维的聚类、基于超图的聚类、子空间聚类和联合聚类,如图 1 所示。

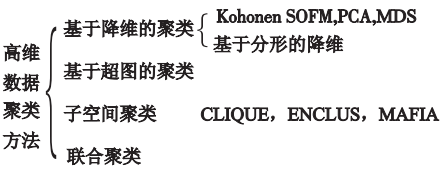

图1 高维数据聚类方法分类示意图

1.1 基于降维的聚类

在很多需要处理高维数据的应用领域,对高维数据进行降维是常用的方法之一。直观地讲,降维就是通过把数据点映射到更低维的空间上以寻求数据的紧凑表示的一种技术,这种低维空间的紧凑表示将有利于对数据的进一步处理。

降维问题的模型(S,M)可定义如下:

$S=\{x_i\}_{i=1}^N$是 D 维空间中的数据集合;

降维映射 $M:S\to L$

$$x\to y=M(x)$$

称 y 为 x 的降维表示。其中:L 是 d 维空间的一个子集,且有 $d<<D$。

降维作为目前很多研究领域的重要研究分支之一,其方法本身就多种多样,根据降维方法的不同,产生了很多基于降维的聚类方法,如 Kohonen 自组织特征映射(self-organizing feature map, SOFM)[3,4]、主成分分析(principle component analysis, PCA)[5]、多维缩放(multidimensional scaling, MDS)[6]等。此外

收稿日期: 2009-05-23; **修回日期**: 2009-11-12 **基金项目**: 国家自然科学基金资助项目(60802080)

作者简介:贺玲(1976-),女,讲师,博士,主要研究方向为多媒体信息系统、数据挖掘(heling6159@163.com);蔡益朝(1976-),讲师,博士,主要研究方向为仿真与智能决策;杨征(1978-),讲师,博士,主要研究方向为虚拟现实.

还有一种特殊的降维聚类方法,即基于分形的降维[7,8]。

Kohonen 自组织特征映射是一种基于神经元网络的方法,它在保留数据的近似关系的前提下,寻求高维数据的低维特征映射,基于该方法对高维数据进行聚类处理,是一类典型的投影聚类方法。在 Kohonen 自组织特征映射中,竞争层的每一个神经元都要相互竞争,胜出的神经元及其近邻神经元则更新它们的权值向量,以使其与输入数据尽可能相似。对神经网络进行训练之后,每个高维数据都将根据与神经元的权值向量的匹配情况投影到这些神经元上。

SOFM 的缺点在于它没有提供一个具体的准则来评价从高维到低维转换的优劣。而且对于很高维的数据而言,神经网络的训练过程收敛会很慢。

主成分分析也是应用较为广泛的降维方法之一。对于一个包含 n 个 m 维数据的数据集,PCA 方法首先计算一个 $m \times m$ 阶的协方差矩阵;然后计算该矩阵的 k 个主导的特征向量,这 k 个特征向量代表了原始数据的主要特征。在此基础上,即可把原始的高维数据沿着 k 个特征向量代表的方向进行投影。由于投影之后的数据具有相对很低的维度,则可以利用传统的聚类算法进行聚类处理。

PCA 虽然提供了一些方法来确定上述 k 值,但不同的方法所确定的 k 值相差很大,因此还是很难找到正确合理的 k 值。k 取值太小,会丢掉原始数据的重要特征;而 k 取值过大,虽然能保留绝大部分原始信息,但投影之后的数据维度依然会很高,聚类处理仍然会很困难。PCA 的另一个缺陷在于,其空间复杂度是 $O(m^2)$,时间复杂度是一个取决于特征值的数量并且大于 $O(m^2)$ 的值。为了将 PCA 的成熟思想更好地应用于非线性降维领域,又有研究者对线性 PCA 进行了扩展,从而产生了核 PCA(kernel PCA)[9]。

多维缩放也是把高维数据映射到低维空间的一种方法,其映射过程保留了数据点之间的差异性(或相似性),即在原始数据集中相近的点仍然靠在一起,而远离的点仍然远离。该类算法的基本出发点是数据点之间的相似性(或差异性)描述。由于降维的目的就是寻求保持数据集感兴趣特性的低维数据集,通过低维数据的分析来获得相应的高维数据特性,从而达到简化分析、获取数据有效特征以及可视化数据的目标。因此,只要最大限度地保持数据间的差异性,便可获得有效的低维表示。MDS 的缺陷在于,首先它没有提供一个好的原则来确定究竟将数据降到多少维;此外,大多数该类方法的时间复杂度都是 $O(n^2)$。其中 n 为数据集的规模。

如果一个数据集在所有的观察尺度下都具有自相似性,即一个数据集的部分分布有着与整体分布相似的结构或属性,则称该数据集是分形的。与分形相对应的,分形维则体现了数据集的固有特征。

基于分形的降维是近年来才得到关注的一类方法。采用分形的思想,首先可以较为准确地估计出数据的本征维,从而为进一步降维提供指导性的参考。与其他方法对本征维的估计所不同的是,基于分形的方法能得到非整数值的本征维,即通常所说的分数维。关于分数维的定义,也有多种不同的描述,其中应用较广泛的是计盒维(box-counting dimension)和相关维(correlation dimension)。基于对这些相应维的估计,产生了一系列不同的方法,它们都为降维处理奠定了良好的基础。

降维作为高维数据研究领域中一个应用极为广泛的课题方向,有很多研究者提出了很多种具体的方法和算法[10~12],从降维所采用的基本思想出发,这些算法不外乎四种类型,即基于数据低维投影的降维、基于神经网络的降维、基于数据间相似度的降维和基于分形的降维。

总而言之,无论基于什么样的降维方法对高维数据进行聚类处理,其基本目的都是先根据相应的方法寻求高维数据等价的低维表示,然后再利用已有的传统聚类算法对降维后的数据进行聚类处理,即用数据在相对低维空间中的聚类结果来表示高维数据的聚类特性。不同的降维方法,它们寻求高维数据的低维表示的方式不同,降维之后的数据与原始数据的近似程度也不同,从而它们的聚类性能也各不相同。

1.2　基于超图的聚类

超图是对常规图的扩展,图中的每条边可以连接多个顶点,称为超边。基于超图的聚类方法把高维数据间的关系映射到一个超图上,图中的每一条超边表达某些数据的关系,边上的权值则表示相应关系的密切程度。在此基础上,基于超图的聚类方法实际上就是寻找图顶点的一个划分,并使得处于同一个划分中的数据尽可能地相关。

基于超图划分的聚类步骤可简单地描述如下:

a)通过超图定义一个点(作为图的顶点)与其他若干点相连的条件;

b)定义图中连接权重的度量;

c)根据一定的图划分算法,寻找权重最小的超边并从中断开连接,从而将超图划分为两个部分,每个部分作为一个簇(类);

d)重复上述划分,直至划分出的簇达到某个特定的值,或所产生的新的划分质量低于预设的阈值。

文献[13]中所提出的聚类方法就是一种典型的聚类方法。该方法针对购物篮数据库中的客户交易数据,用频繁集项来构造加权超图。每个频繁集项作为超图中的一条边,其权值由从该项集出发的所有可能的关联规则的平均置信度确定。在这些基本的数据表示工作完成之后,聚类算法根据特定的超图划分算法对所有项(商品)进行划分,以使得由于划分而被断开的超边权值之和最小,划分的结果就是交易记录中同时出现的项,最后可以用这些项簇来作为聚类的描述,并使用一个度量来客户交易指派给最佳的项簇。

总的来说,基于超图划分的聚类算法的关键思想在于,把高维数据空间中的数据处理问题转换为图划分问题,通过构造特定超图的最小生成树来寻求高维数据的聚类。该方法最大的优点在于它在聚类的过程中不用显示地计算高维数据之间的相似度,因此算法的时间复杂度仅为 $O(ndk)$。其中:n 为数据集的规模;d 为数据的维度;k 为聚类的个数。针对不同的应用领域和应用背景,研究者们也提出了很多基于超图的聚类方法[14,15]。

1.3　子空间聚类

子空间聚类又称特征选择,它把原始数据空间划分为不同的子空间,只在那些相关的子空间上考察聚类的存在。这类算法一般使用贪心策略等搜索方法搜索不同的特征子空间,然后使用一些标准来评价这些子空间,从而找到所需的簇。

典型的子空间聚类算法有 CLIQUE (clustering in quest)[16]、ENCLUS (entropy-based clustering)[17] 和 MAFIA (merging of adaptive finite intervals algorithm)[18] 等。该类算法都使用 Apriori 策略[19] 来查找和合并某度量大于给定阈值的网格,产生候选子空间,并将这些候选子空间按其覆盖即子空间中点数量的大小排序;随后利用最小描述长度准则将规模较低的子空间剪枝。此类算法在理论上可以找到任意数量维中

任意类型和形状的簇,其结果由一组不同子空间的簇组成,并可由一个析取范式表达式所表示,且事先无须确定维数量。

CLIQUE 算法是较早尝试在数据子空间中查找簇的算法,它综合了基于密度和基于网格的聚类算法思想。CLIQUE 算法利用从 $k-1$ 维空间发现的密集单元来推断 k 维空间的候选的密集单元,并将这些候选单元按其覆盖的大小排序。这里单元 S 的覆盖是单元中点的数量。随后利用 MDL 策略删去覆盖较小的单元而只保留密集单元。最后算法以深度优先策略搜索与每一个密集单元邻近的密集单元,并用贪心策略合并这些单元形成聚类。CLIQUE 算法需要两个参数,即网格尺寸和密度阈值,它们的取值对聚类结果的质量有很大影响。如果参数设置不合适,在剪枝阶段很可能会删去一些重要的聚类。但是对一个指定的数据集来说,要确定这些参数非常困难。

与 CLIQUE 相比,ENCLUS 算法使用不同的准则来选择子空间,即它不是用密度或覆盖,而是使用一个中间值熵来查找子空间的簇。子空间的聚类能力有三个标准,即密度、覆盖和维度的相关性。ENCLUS 使用熵来代替密度,因为熵能同时衡量这三个标准。

首先,在一定条件下,某子空间覆盖增加则其熵值减小;其次,单元的密度增加时,其熵值减小,因此熵能衡量簇的密度;第三,维之间的相关性可使用兴趣度来衡量。兴趣度定义为子空间中各维的熵之和与该子空间熵之差。兴趣度越大,维之间的相关性越强。如果兴趣度为 0,则各维是独立的;当且仅当兴趣度超越某给定阈值时,维之间才相关。

ENCLUS 算法的主要依据是:可形成簇的子空间的熵值一般低于无法形成簇的子空间的熵值。该算法也需要三个参数,即熵的阈值 ω、信息增益阈值 ε' 和网格尺寸 Δ。与 CLIQUE 一样,算法结果对这些参数高度敏感。ENCLUS 算法的伸缩性也与 CLIQUE 完全相同。

MAFIA 算法中的网格大小是根据数据的分布动态调整的,而不是固定的,这样可以提高算法的效率和结果的质量。MAFIA 还引入并行处理来增强其伸缩性。

首先,算法扫描一遍数据,针对每一维,根据数据的分布建立直方图,然后合并相似密度的相邻箱(密度之差小于给定的差异阈值),形成窗口,并删除低于密度阈值的箱,从而动态确定网格的边界。最终这种动态网格单元能比固定大小的网格单元更精确地描述簇集的边界并减少了计算量。随后 MAFIA 就在这些候选动态网格上采用与 CLIQUE 一样的方法进行聚类。MAFIA 采用并行方法进行聚类处理,使得执行效率更高。

MAFIA 需要一个密度阈值参数,并为相邻窗口指定差异阈值。如果某一个维的分布基本是均匀的,MAFIA 还需要输入一个默认的网格单元大小作为缺省值。在这些维中,数据集被分割为一些固定尺寸的较小区间。尽管其他网格的大小是自动调整的,但算法结果对这些参数值更加敏感。

由于算法的改进和使用了并行处理,在类似的数据集中,MAFIA 的执行速度比 CLIQUE 快很多倍,其执行时间也与数据集中实例或维数呈线性关系,并且与输出簇的维数呈指数关系。

此外,以信号处理为基础的 Wave Cluster 算法也属于基于该方法的范畴。它是一种多分辨率聚类算法,首先通过在数据空间上强加一个多维网格结构来汇总数据;然后采用一种小波变换来变换原特征空间,在变换后的空间中进行聚类。小波变化是一种信号处理技术,它将一个信号分解为不同频率的子波段。在进行小波变换时,数据被变换以在不同的分辨率层次保留对象间的相互距离。这使得数据的自然聚类易于区别,从而通过在新的数据空间中寻找高密度区域以确定聚类。

Wave Cluster 的显著优点主要体现在以下几个方面:

a)能够获得较高质量的聚类;

b)具有较好的处理高维空间数据的能力;

c)能够很好地处理聚类中的异常点;

d)其时间复杂度为 $O(N)$,N 为数据集的规模。

关于子空间聚类的更多算法可参照文献[20~22]。

1.4　联合聚类

联合聚类的思想来源于 OLAP 中对多维数据的向上钻取分析。在 OLAP 中,每一次上钻取都可以看成是寻求某一组属性的代表值。联合聚类的一般思想就是先将聚类数据集的属性分成若干组,然后针对每个属性组提出一个新的属性来代表该属性组,继而针对若干派生出来的属性进行高维数据聚类。

联合聚类实际上是同时对数据点和其属性进行聚类。因此用该方法进行聚类时会出现这样的情形:数据集聚类质量的提高依赖于其属性的聚类,而对属性进行聚类也必须要依赖于相应的数据集。对于数据集和其属性而言,所有的数据和它们的属性描述可以看成是一个矩阵 $\boldsymbol{X}$。目前为止,很多文献处理更多的只是对矩阵 $\boldsymbol{X}$ 的行分组,如果要考虑对 $\boldsymbol{X}$ 的列分组,就需要利用数据点—属性这种数据表示形式中所包含的规范的二元性。

在相关文献中,联合聚类又被称做同时聚类[23]、双维聚类[24]、块聚类[25]、分配聚类[26]等。文献[27~29]分别针对不同的应用背景提出了各自具体的联合聚类方法。

2　现有典型高维聚类算法性能分析

从上述对现有高维聚类算法的总结分析不难看出,这些聚类方法的一个共同点在于,与传统聚类算法相比,它们都从不同方面提升了算法在处理高维数据时的能力,因此都在不同程度上适用于相对高维的数据。但这些聚类方法也各有优劣。

基于降维的高维聚类方法是对高维数据进行聚类处理的最为直观的方法之一,其优点是易于理解、实现简单,但其缺陷也是显而易见的:首先,数据集中噪声数据的存在是影响降维聚类效果的关键因素。在通过降维将原始高维数据映射到低维空间的过程中,同时也会缩小噪声数据与“干净”数据之间的距离,从而不可避免地降低聚类的质量。而在很多应用领域中,通常很难在预处理过程清除噪声数据的影响。

此外,基于降维的聚类从根本上说都是以数据之间的距离或相似度评价为聚类依据,当数据的维数不是很高时,这些方法效果较好,但当数据维度增高,聚类处理将很难达到预期的效果。原因在于:a)在一个很高维的空间中定义一个距离度量本身就是一个很困难的事情;b)基于距离的方法通常需要计算各个聚类之间的距离均值,当数据的维度很高时,不同聚类之间的距离差异将会变得很小。

基于超图的聚类方法的优点主要体现在两个方面:a)通过该方法,可以在聚类的过程中回避对高维数据之间相似度的计算,从而减小了维度灾难对高维聚类的影响;b)利用该方法还可以根据特定用户或领域的需求来控制聚类的质量,原因在于,利用 Apriori 算法中最小支持度的不同层次,超图模型所表达的数据间的关系可以进行适当的调整,较高的支持度值对应包含数据点较少的更高质量的聚类,较低的支持度值则对应包含数据点较多的粗糙的聚类。但是不容忽视的是,该算法聚类

效果的好坏与相应参数的选取有很大的关系。首先,在寻找频繁集时,支持度层次的确定与具体的应用领域密切相关;其次,对于连续变量,必须要对其离散化之后才能应用该算法进行处理。而对连续属性的离散化处理必然会导致数据间的某些关系的丢失,从而使得聚类结果与实际情况会偏差很大。

子空间聚类从某种程度上来讲与基于降维的聚类有些类似,但后者是通过直接的降维来对高维数据进行预处理,即在降维之后的某一个特定的低维空间中进行聚类处理;而前者是把高维数据划分成若干不同的子空间,再根据需要在不同的子空间中寻求数据的聚类。利用子空间聚类的思想,可以从多个角度、综合考虑多方面的属性来寻求数据的聚类。但是在这类算法中,子空间的划分和选取也是一个值得深入研究的重要问题。子空间划分太多,不仅计算复杂度会很大,聚类的结果也会过于繁杂;子空间划分太粗糙,则不能很好地避免维度灾难对聚类的影响。

联合聚类尝试了对数据和其属性同时进行聚类,该方法在提高聚类效果的同时,也不可避免地增加了聚类的时间复杂度,而且该类算法目前主要集中在对文本进行聚类处理,鲜有对其他类型的数据集进行聚类的研究。

3 结束语

维度灾难一直是处理高维数据时面临的一个关键问题,也是促使该领域内的研究不断向前推进的动力。为了解决这个问题,很多研究者提出了一系列的方法和算法。本文从高维聚类算法所采用的基本思想出发,将它们分为基于降维的聚类、基于超图的聚类、基于子空间的聚类和联合聚类。如果深入考虑聚类中涉及到更多更细节的内容,还可以有更详细的分类,这也是笔者下一步要进行的工作。

如本文所述,现有的这些高维聚类算法分别从各自的角度出发,从一定程度上减小了维度灾难对高维聚类的影响,但它们各自也都存在一些特定的问题有待解决。

总的来说,笔者认为,在上述四类高维聚类方法中,以子空间聚类为主体,有效结合利用其他聚类方法,甚至是适用于低维数据的传统聚类算法思想(如基于密度的聚类等),将是一个很有意义的、值得关注的方向。

此外,对于高维聚类而言,还有一些共性的问题有待进一步研究:

a)聚类结果的评价。聚类结果准确与否,需要有一个恰当的准则来评价,因此评价准则的全面、合理程度直接决定了对聚类结果正确性的判断。

b)聚类个数的确定。聚类个数是很多聚类算法在聚类之前必须确定下来的一个重要参数,它对聚类过程的进行起着重要的指导作用。

c)异常点的处理。对异常点的处理是数据聚类中必须面临的重要问题之一。如何在保留数据集整体特性的前提下尽可能地消除异常点对聚类的影响,是获取高质量聚类所必须解决的一个关键问题。

d)数据相似度的评价。简单地说,聚类的目的是把相似的数据放在同一类中,不相似的数据则分属不同的类。那么,如何判断、评价数据是否相似、相似程度有多高?这显然是聚类过程中必须处理的一个基础性问题。如上所述,虽然有些算法(如基于超图的聚类)在聚类过程中可通过一定的方式回避对数据相似度的直接计算,但从本质上说,它只是把相似度的计算形式进行了转换,因此仍然摆脱不了对相似度的评价。而这个评价准则恰当与否,对聚类的结果起着举足轻重的作用。

本文下一步的工作主要包括两个方面的任务:a)进一步深入分析各类高维聚类算法的差异性和共同特性,探讨更细致的分类表述;b)在上述分析比较的基础上,探讨将不同的聚类思想统一于同一个聚类问题中,从而合理利用不同算法的优势,扬长避短,最大限度地提高聚类的质量。

参考文献:

[1] ERTOZ L, STEINBACH M, KUMAR V. Finding clusters of different sizes, shapes and densities in noisy high-dimensional data[R]. Minnesota: Department of Computer Science, University of Minnesota, 2002.

[2] HAM J H, LEE D D, SAUL L K. Learning high-dimensional correspondences from low dimensional manifolds[C]//Proc of ICML Workshop on the Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining. Washington:[s. n.], 2003:34-41.

[3] KOHONEN T. Self-organization and associated memory[M]. [S. l.]: Springer-Verlag, 1988.

[4] KOHONEN T. Self-organizing maps[M]. New York: Spinger-Verlag, 2001.

[5] MINKA T P. Automatic choice of dimensionality for PCA[C]//Proc of International Conference on Advances in Neural Information Processing Systems. Cambridge:[s. n.], 2001:598-604.

[6] GRIFFITHS T L, KALISH M L. A multidimensional scaling approach to mental multiplication[J]. Memory & Cognition, 2002,30(1):97-106.

[7] CAMASTRA F, VINCIARELLI A. Estimating the intrinsic dimension of data with a fractal-based method[J]. IEEE Trans on Pattern Anal Mach Intell, 2002,24(10):1404-1407.

[8] CAMASTRA F. Data dimension estimation methods: a survey[J]. Pattern Recognition, 2003, 36:2945-2954.

[9] SCHOLKOPF B, SMOLA A, MULLER K. Nonlinear component analysis as a kernel eigenvalue problem[J]. Neural Computation, 1998,10(5):1299-1319.

[10] TENENBAUM J B, De SILVA V, LANGFORD J C. A global geometric framework for nonlinear dimensionality reduction[J]. Science, 2000,290(5500):2319-2323.

[11] BALASUBRAMANIAN M, SCHWARTZ E L, TENENBAUM J B, *et al*. The Isomap algorithm and topological stability[J]. Science, 2002,295(5552):7.

[12] IFARRAGNERRI A, CHANG C I. Unsupervised hyperspectral image analysis with projection pursuit[J]. IEEE Trans on Geosci Remote Sensing, 2000,38(6):2529-2538.

[13] HAN E H, KARYPIS G, KUMAR V, *et al*. Clustering based on association rule hypergraphs[C]//Proc of Workshop on Research Issues on Data Mining and Knowledge Discovery. 1997:9-13.

[14] WANG Bo, ZHANG Ming-wei, ZHANG Bin. An effective hypergraph clustering in multi-stage data mining of traditional Chinese medicine syndrome differentiation[C]//Proc of the 6th IEEE International Conference on Data Mining Workshops. 2006:848-852.

[15] HU Tian-ming, XIONG Hui, ZHOU Wen-jun. Hypergraph partitioning for document clustering[C]//Proc of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval Table of Contents. 2008:871-872.

[16] AGRAWAL R, GEHRKE J, GUNOPULOS D. Automatic subspace clustering of high dimensional data for data mining applications[C]//Proc of ACM SIGMOD Conference. Seattle, WA:[s. n.], 1998:94-105.

(下转第 31 页)

Coding. [S. l.]: Springer-Verlag, 2001:360-363.

[7] HORWITZ J, LYNN B. Toward hierarchical identity-based encryption [C]//Proc of Advances in Cryptology Eurocrypt' 02. [S. l.]: Springer-Verlag, 2002:466-481.

[8] GENTRY C, SILVERBERG A. Hierarchical ID-based cryptography [C]//Proc of Advances in Cryptology-Asiacrypt' 02. [S. l.]: Springer-Verlag, 2002:548-566.

[9] LYNN B. Authenticated ID-based encryption cryptology[R]. ePrint Archive Report 2002/072. 2002.

[10] CANETTI R, HALEVI S, KATZ J. A forward-secure public-key encryption scheme[C]//Proc of Advances in Crptology-Eurocrypt'03. [S. l.]:Springer-Verlag, 2003:255-271.

[11] CANETTI R, HALEVI S, KATZ J. Chosen-ciphertext security from identity based encryption[C]//Proc of Advances in Cryptology-Eurocrypt'04. [S. l.]: Springer-Verlag, 2004:207-222.

[12] BONEH D, BOYEN X. Efficient selective ID secure identity based encryption without random oracles[C]//Proc of Advances in Cryptology-Eurocrypt'04. [S. l.]: Springer-Verlag, 2004:223-238.

[13] BONEH D, BOYEN X. Secure identity based encryption without random oracles [C]//Proc of Advances in Cryptology-Crypto' 04. [S. l.]: Springer-Verlag, 2004: 443-459.

[14] WATERS B R. Efficient identity-based encryption without random oracles [C]//Proc of Advances in Cryptology-Eurocrypt'05. [S. l.]: Springer, 2005:114-127.

[15] BONEH D, GRESCENZO G D, OSTROVSKY R, *et al*. Public key encryption with keyword search[C]//Proc of Advances in Cryptology-Eurocrypt'04. [S. l.]: Springer-Verlag, 2004:506-522.

[16] SAHAI A, WATERS B. Fuzzy identity-based encryption[C]//Proc of Advances in Cryptology-Eurocrypt'05. [S. l.]: Springer, 2005: 457-473.

[17] NACCACHE D. Secure and practical identity-based encryption[J]. Information Security, 2007,1(2):59-64.

[18] BONEH D, BOYEN X, GOH E J. Hierarchical identity based encryption with constant size ciphertext[C]//Proc of Advances in Cryptology-Eurocrypt'05. Berlin: Springer-Verlag, 2005:440-456.

[19] ABDALLA M, CATALANO D, DENT A, *et al*. Identity-based encryption gone wild[C]//Proc of the 33rd International Colloquium Automata, Languages and Programming. [S. l.]: Springer-Verlag, 2006:300-311.

[20] BOYEN X, WATERS B. Anonymous hierarchical identity-based encryption(without random oracle)[C]//Proc of Advances in Cryptology. [S. l.]:Springer, 2006:290-307.

[21] GENTRY C. Practical identity-based encryption without random oracles[C]//Proc of Advances in Cryptology-Eurocrypt. [S. l.]:Springer-Verlag, 2006:183-189.

[22] BONEH D, GENTRY C, HAMBURG M. Space-efficient identity based encryption without pairings[EB/OL]. (2007). http://eprint. iacr. org/2007/177. pdf.

[23] BOYEN X,MARTIN L. Identity-based cryptography standard (IBCS) #1: supersingular curve implementations of the BF and BB1 cryptosystems [EB/OL]. (2007-12). http://www. ietf. org/rfc/rfc5091. txt.

[24] APPENZELLER G, MARTIN L. Identity-based encryption architecture and supporting data structures[EB/OL]. (2009-01). http://www. ietf. org/rfc/rfc5408. txt.

[25] MARTIN L, SCHERTLER M. Using the Boneh-Franklin and Boneh-Boyen identity-based encryption algorithms with the cryptographic message syntax(CMS)[EB/OL]. (2009-01). http://www. ietf. org/rfc/rfc5409. txt.

[26] WHYTE W, JOHNSON D B. Draft standard for identity-based public-key cryptography using pairings[EB/OL]. (2008-04). http://grouper. ieee. org/groups/1363/IBC/material/P1363. 3-D1-200805. pdf.

[27] BONEH D, BOYEN X. Short signatures without random oracles [C]//Proc of Advances in Cryptology-Eurocrypt' 04. [S. l.]: Springer-Verlag, 2004:56-73.

[28] CHEN L, KUDLA C. Identity based authenticated key agreement from pairings[R]. 2002.

[29] BAEK J, ZHENG Yu-liang. Identity-based threshold decryption[C]//Proc of Pracitce and Theory in Public Key Cryptography-PKC. [S. l.]: Springer-Verlag, 2004:262-276.

(上接第 26 页)

[17] CHENG C H, FU A W, ZHANG Yi. Entropy-based subspace clustering for mining numerical data[C]//Proc of the 5th ACM SIGKDD. San Diego, CA:[s. n.], 1999:84-93.

[18] NAGESH H, GOIL S, CHOUDHARY A. Adaptive grids for clustering massive data sets[C]//Proc of the 1st SIAM ICDM. Chicago, IL:[s. n.], 2001.

[19] 吴泉源, 刘江宁. 人工智能与专家系统[M]. 长沙:国防科技大学出版社,1995.

[20] KRIEGEL H P, KRÖGER P, ZIMEK A. Clustering high-dimensional data: a survey on subspace clustering, pattern-based clustering, and correlation clustering[J]. ACM Trans on Knowledge Discovery from Data, 2009,3(1):1-58.

[21] KRIEGEL H P, KRÖGER P, RENZ M. A generic framework for efficient subspace clustering of high-dimensional data[C]//Proc of International Conference on Data Mining. 2005.

[22] GAN Guo-jun, WU Jian-hong, YANG Zi-jiang. PARTCAT: a subspace clustering algorithm for high dimensional categorical data[C]//Proc of International Joint Conference on Neural Networks. 2006: 4406-4412.

[23] GOVAERT G. Simultaneous clustering of rows and columns[J]. Control and Cybernetics, 1995,24:437-458.

[24] MADEIRA S C, OLIVEIRA A L. Biclustering algorithms for biological data analysis: a survey[J]. IEEE/ACM Trans on Computational Biology and Bioinformatics, 2004,1(1):24-45.

[25] NADIF M, GOVAERT G. Block clustering with mixture model: comparison between different approaches[C]//Proc of International Symposium on Applied Stochastic Models and Data Analysis. Brest: [s. n.], 2005.

[26] BERKHIN P, BECHER J. Learning simple relations: theory and applications[C]//Proc of the 2nd SIAM ICDM. 2002:420-436.

[27] DHILLON I. Co-clustering documents and words using bipartite spectral graph partitioning[C]//Proc of the 7th ACM SIGKDD. San Francisco, CA:[s. n.], 2001:269-274.

[28] COSTA G, MANCO G, ORTALE R. A hierarchical model-based approach to co-clustering high-dimensional data[C]//Proc of ACM Symposium on Applied Computing. 2008: 886-890.

[29] TJHI W C, CHEN Li-hui. Robust fuzzy co-clustering algorithm[C]//Proc of the 6th International Conference on Information, Communications & Signal Processing. 2007:1-5.