基于半监督聚类的文档敏感信息推导方法

苏赢彬[1,2] 杜学绘[1,2] 夏春涛[1,2] 曹利峰[1,2] 陈华成[3]
(解放军信息工程大学 郑州 450001)[1] (数学工程与先进计算国家重点实验室 郑州 450001)[2]
(解放军73503部队 福州 350018)[3]

摘 要 针对当前多文档聚合推导引起的敏感信息泄露问题存在风险大、隐蔽性高的特点,提出了一种基于半监督聚类的文档敏感信息推导方法。首先,为确保在较小的时间开销下获得高质量的约束信息,设计了一种新颖的二阶约束主动学习算法,它通过选择不确定性最大的样本点来生成信息量最大的约束闭包;然后,在引入约束信息的基础上结合DBSCAN提出一种新的半监督聚类算法,它能够有效解决DBSCAN算法存在的边界模糊问题,提高文档聚类准确性;最后,在半监督聚类结果的基础上,对相似文档进行敏感信息可能性测度。实验表明,半监督聚类算法准确率提升明显,推导方法能够有效推导出敏感信息。

关键词 半监督聚类,DBSCAN,主动学习,敏感信息,模糊数学,推导方法
中图法分类号 TP393 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.10.028

Sensitive Information Inference Method Based on Semi-supervised Document Clustering

SU Ying-bin[1,2] DU Xue-hui[1,2] XIA Chun-tao[1,2] CAO Li-feng[1,2] CHEN Hua-cheng[3]
(PLA Information Engineering University, Zhengzhou 450001, China)[1]
(State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou 450001, China)[2]
(73503 PLA Troop, Fuzhou 350018, China)[3]

Abstract For the problem that sensitive information leakage caused by multi-document clustering and inference has the features of high risk and high concealment, a sensitive information inference method based on semi-supervised document clustering was proposed. Firstly, a new second-order constraint active learning algorithm was designed, which can ensure to obtain high quality constraints with less time by choosing the most uncertain informative data. Then, a new semi-supervised clustering algorithm combining constraints and DBSCAN was proposed, which can effectively resolve fuzzy boundaries of DBSCAN and improve the precision of document clustering. Finally, possibility measure of sensitive information on similar documents was calculated based on the results of semi-supervise clustering. The experiments show that the precision of semi-supervised clustering improves significantly, and the inference method can infer sensitive information effectively.

Keywords Semi-supervised clustering, DBSCAN, Active learning, Sensitive information, Fuzzy math, Inference method

1 引言

随着互联网与信息技术的高速发展,办公终端逐渐成为政府、军队和企事业单位工作的主要平台,并通过加入互联网以减少信息孤岛的形成。由于办公终端中存储了大量办公文档,在处理和传输的过程中都存在着敏感信息泄露的风险,尤其是存在多个相似文档经过聚类后可能推导出新的或者更高级别的敏感信息的文档推理问题。例如相似文档集合 $D=\{d_1,d_2,d_3,d_4\}$,D 中文档最高密级 MAX(D)低于或等于其载体办公终端的密级 Class,但文档$\{d_1,d_3,d_4\}$或$\{d_1,d_2,d_4\}$都有推导出更高级别敏感信息的可能性,如果用户具有同时访问这些文件的权限,结合自身知识,经过推导就可能导致敏感信息泄露。由于这种情况下敏感信息泄露的隐蔽性很高,对于敏感信息检测来说是一个难点。

国外相关研究主要集中在对安全数据库的推理通道问题上,早在20世纪八九十年代,文献[1]就提到大量数据聚合后存在敏感信息泄露问题,并针对关系数据库进行了详细分析,提出了一个框架拒绝对某些资源的查询,以解决由于"聚合"导致的敏感信息泄露的问题,提高了数据库的安全性。文献[2]指出,推理问题在现代的数据使用管理中是一个很重要的问题,尤其是在以数据为中心的商业模式下;其详细分析了在社交网络下用户隐私存在通过推理被泄露的风险,并分析了推理问题,引入一个机制帮助用户完善隐私政策,减少推理。相对于国外,国内对这方面研究较少,文献[3]详细分析了安

到稿日期:2014-10-18 返修日期:2015-01-05 本文受国家高技术研究发展计划(863计划)项目(2012AA012704)资助。
苏赢彬(1989—),男,硕士生,主要研究方向为信息安全,E-mail:dr.suyingbin@foxmail.com;**杜学绘**(1968—),女,博士,教授,博士生导师,主要研究方向为网络安全;**夏春涛**(1979—),男,硕士,讲师,主要研究方向为网络安全;**曹利峰**(1981—),男,博士,副教授,主要研究方向为网络安全;**陈华成**(1986—),男,硕士,助理工程师,主要研究方向为网络安全。

全数据库中存在的推理通道问题，提出了动态泄露推理引擎，结合用户访问历史文件在新的查询被允许前对其进行相应评估，如果评估结果显示访问超过用户安全等级，则拒绝访问。文献[4]从多级安全的角度，对访问多个客体可以推导出更高级别信息的问题进行了分析，首先对客体进行聚类，提出了基于概念相似的递归客体聚类算法，最后推演相似客体推导出更高级别信息的可能性，但是由于客体与文档表示方式不同，其具体方法并不适用于对文档的敏感信息聚合推导问题。国内外研究主要集中于安全数据库和多级安全中客体的推理，针对多文档的敏感信息推导方法的研究较少。

文档之间的关系可以分为相似性和关联性两种。为了检测多个文档通过推导可能出现敏感信息泄露的问题，并且相似的文档由于可能具有相似主题、表达内容相近、存在信息互补的原因而更容易推导出新的或者更高级别的敏感信息，本文在分析文档之间关系的基础上，提出一种基于半监督聚类的文档敏感信息推导方法。首先提出一种结合约束主动学习的半监督聚类算法对文档进行聚类，将具有相似主题和内容的文档聚为类簇；然后引入模糊数学理论，通过对类簇的文档子集敏感度级别可能性测度计算文档推导出新的或者更高敏感度信息的可能性。

2 基于约束主动学习与 DBSCAN 的文档半监督聚类

少量先验知识用于指导聚类已被证明能够有效提高聚类的效果[5]。利用监督信息来辅助无监督聚类就是半监督聚类，常见的监督信息包括标签和约束等，根据使用方式的差异大致可以分为两大类：一类基于约束，通过修改聚类的目标函数将监督信息融合到聚类算法中，或者用监督信息约束聚类的收敛过程，使其能够满足约束条件；另一类基于距离，结合监督信息修改现有距离测度函数或者学习一种新的更合适的距离测度函数用于聚类。另外，也可以将两种方法结合使用。本文采用成对约束[6]（Pairwise Constraints）作为监督信息，它是由 Wagstaff K 于 2000 年提出的目前较为常用的约束形式，包含了 must-link 和 cannot-link 两种约束集合，其中元素为一个二元组(x,y)，当其属于 must-link 时，x,y 须分在一个聚类中，用于控制类簇的边界样本点；当属于 cannot-link 时，x,y 不能在一个聚类中，用于区分临近类簇重叠部分的样本点，并且具有以下性质：

$$[(x_i,x_j)\in ML]\wedge[(x_j,x_k)\in ML]\rightarrow[(x_i,x_k)\in ML]$$

$$[(x_i,x_j)\in ML]\wedge[(x_j,x_k)\in CL]\rightarrow[(x_i,x_k)\in CL]$$

文献[7]通过分析 DBSCAN（Density Based Spatial Clustering of Applications with Noise）算法，结合对约束的主动学习，提出了一种半监督聚类算法，但是其主动学习采用最远优先策略，只能确保每个聚类至少能发现一个核心点，没有进一步选取信息量最大的数据点学习约束。针对随机选取的约束所包含的信息量小、聚类效果欠佳等问题，首先对相似文档进行聚类，提出一种针对文档优化的结合约束主动学习与 DBSCAN 的半监督聚类算法（Active Learning of Constrains DBSCAN，ALC-DBSCAN），以提升文档聚类效果。

2.1 聚类文档预处理

为了减小聚类时间开销，提升聚类性能，需要对半监督聚类的文档进行预处理，主要包括对文档进行向量表示，解决高维度和稀疏矩阵问题以及定义相似度计算函数。

2.1.1 聚类文档表示及平滑

文档表示是影响聚类算法性能的重要因素之一[8]，特征项的选取直接会影响到聚类的准确性，好的特征项能够大大提升聚类性能。同时，为了计算文档之间的相似度，需要对文本进行表示。设文档集合 $D=\{d_1,d_2,\cdots,d_n\}$，D 中任意文档 d_i 经过中文分词、除去停用词等预处理后提取特征词项形成文档向量（Vector Space Model，VSM）。传统的向量模型采用 TF-IDF 表示文档，其采用简单统计词频和逆向文档频率的方法，这种传统的采用 TF-IDF 表示文档的向量模型往往忽略了文档的上下文信息[9]。本文采用文献[10]中的文档平滑模型，保留包括主题信息和结构信息在内的上下文信息，能更加真实地对文档建模，为聚类提供更准确的信息，减少由于文档信息丢失而引起的聚类偏移。对于任意文档 d_i，其经过平滑后表示为：

$$d_i=(t_{i1},\omega_{i1},t_{i2},\omega_{i2},\cdots,t_{im},\omega_{im}) \tag{1}$$

其中，t_{ij} 为提取出来的特征词项，ω_{ij} 为文档 d_i 中词项 t_{ij} 经过平滑后得到的权重。文档集合 D 最终形成的词项-文档矩阵 X 为：

$$X=\begin{bmatrix}x_{11} & x_{12} & \cdots & x_{1n}\\ x_{21} & x_{22} & \cdots & x_{2n}\\ \vdots & \vdots & \ddots & \vdots\\ x_{m1} & x_{m2} & \cdots & x_{mn}\end{bmatrix}_{m\times n}=[d_1 \quad d_2 \quad \cdots \quad d_n]=\begin{bmatrix}t_1\\ t_2\\ \vdots\\ t_m\end{bmatrix} \tag{2}$$

其中，$d_i(0<i\leqslant n)$是由词项表示的文档，$t_i(0<i\leqslant m)$是由文档表示的词项。

2.1.2 文档向量降维处理

由于向量表示的文档本身具有的高维度和稀疏数据的问题使得聚类时间和效果都表现不佳，因此需要对特征项进行选择，筛选出最具代表性的特征来对文档向量进行降维和特征提取，以解决数据稀疏问题。奇异值分解（Singular Value Decomposition，SVD）是降维技术中特征重构的一种，将文档从稀疏的高维词项空间映射到一个低维的向量空间，降低了词项和文档之间的语义模糊度，利用其对文本进行处理，能够起到降维、解决稀疏矩阵的作用。降到维度为 k 的潜在语义空间上（Latent Semantic Space，LSS），其分解公式为：

$$X=U_{m\times k}\sum_{k\times k}V^{T}_{k\times n} \tag{3}$$

通过上式分解后，U 和 V 都是正交矩阵，Σ为对角矩阵，其得到最大的 $k(k\leqslant r)$个奇异值最重要，r 为 X 的秩，较小的奇异值作为噪声除去，用$\sum_{k\times k}$近似表征词项-文档矩阵。

2.1.3 聚类相似度函数

定义 1（相似度函数） 降维后的向量空间较之前小很多，计算复杂度也大大降低，聚类时文档之间相似性度量采用欧氏距离来表征：

$$Sim(d_i,d_j)=\sqrt{\sum_k(x_{iy}-x_{jy})^2} \tag{4}$$

欧氏距离越近，相似度越大，两个文档越相似；反之亦然。最终得到相似度矩阵。

2.2 基于约束主动学习与 DBSCAN 的半监督聚类算法

2.2.1 设计思路

经过向量表示和降维等处理后的文档聚类能够大大降低聚类计算开销，并得到了可以直接用于聚类的相似度矩阵。本文提出一种基于约束主动学习与 DBSCAN 的半监督聚类

算法。监督信息采用成对约束，首先讨论约束闭包之间存在的关系，以便于约束表示；然后设计了一种约束二阶主动学习算法，一阶在较少的尝试次数下先确保每个聚类至少有一个样本被选取，二阶选择信息量最大的样本点判断 must-link/cannot-link 属性，结果返回约束闭包；最后将约束闭包引入 DBSCAN 聚类算法中，用约束闭包指导 DBSCAN 的聚类过程，形成的半监督聚类提高了边界的控制能力和聚类的准确性。

2.2.2 约束闭包及关系

约束闭包能够更加简洁明确地体现 must-link/cannot-link 关系，并且用很少的数据就可以获得大量的约束条件，例如已经确认 x 属于某一闭包，则可以推理出 x 与闭包中其他元素为 must-link 关系，与相异闭包中元素均为 cannot-link 关系，使其能够很方便地融合到聚类算法中。

约束闭包之间存在 3 种关系：相同闭包、相异闭包和无关闭包。相同闭包中元素两两具有 must-link 性质，相异闭包间元素具有 cannot-link 性质，属于不同闭包的元素属于不同类簇，即无关闭包，从约束来看不能判断其约束性质。

定义 2(闭包)　对任意集合 $A=\{a_1,a_2,a_3\}$，如果 $\{a_1,a_2\}$ 具有 must-link 关系，$\{a_2,a_3\}$ 具有 must-link 关系，则 $\{a_1,a_3\}$ 具有 must-link 关系，a_1,a_2,a_3 属于相同闭包(简称闭包)，A 为一个闭包，其中元素对 must-link 约束具有传递性。

图 1 中两个图实质一样，实线表示元素之间存在 must-link 关系，a、c 的关系可以通过 must-link 约束的传递关系推导出来，a、b、c 属于相同闭包。

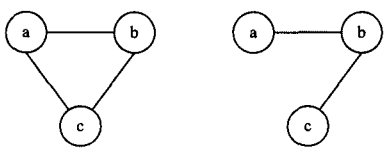

图 1　相同闭包(闭包)

定义 3(相异闭包)　$A=\{a_1,a_2,\cdots\}$，$B=\{b_1,b_2,\cdots\}$，A、B 为闭包，$\{a_i,b_j\}$ 属于 cannot-link，则 A、B 为相异闭包。

图 2 中都是相异闭包，虚线表示元素之间存在 cannot-link 关系，$\{a\}\{b\}\{c\}$ 两两为相异闭包。

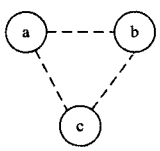

图 2　相异闭包

定义 4(无关闭包)　闭包 A、B 没有明确约束说明其之间的关系，其可能最终属于同一聚类，也可能为相异闭包。

图 3 中 $\{a\}\{b\}$ 或 $\{b\}\{c\}$ 为相异闭包，$\{a\}\{c\}$ 关系无法判断，出现这种未知情况的原因是：a、c 可能在聚类算法中能够明确关系，信息量小，不足以作为约束进行判断。

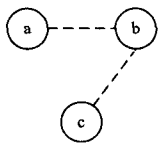

图 3　无关闭包

2.2.3 基于约束二阶主动学习的闭包生成方法

主动学习[11]最先在分类中使用十分广阔，在聚类中使用较少。由于分类解决的问题是将未知的数据划分到已经存在的类别中，主动学习能够从类别标签中获得足够的监督信息，大大提升了分类的效果和性能；但是聚类中不存在类别、标签等先验知识，只能靠人工标注等方式获取监督信息，效率低下。将主动学习方法引入到无监督的聚类中，能有效提升监督信息获取的效率。

现有主动学习能够自主地选择样本提交给专家进行标注，然后进行训练[12]，但是最终获得的监督信息存在质量不高的问题，比如得到一些聚类算法可以正确识别的约束，不仅不能提升聚类准确性，反而降低了算法效率。如何选择最大信息量[13](Most Informative)的约束是主动学习算法的关键点。

本文设计了一种约束二阶主动学习算法，通过筛选信息量大的样本点，减少约束集合大小和聚类时间消耗，这些点趋向于噪声、多个聚类边缘和重叠处，属于某一类簇不确定性最大的点，从而获得对聚类最有价值的约束闭包。

主动学习算法主要是为了获取聚类算法本身不能判别的样本与类簇之间的约束关系，通过筛选信息量大的样本来生成约束闭包。

定义 5(约束信息总量 AI)　指成对约束所包含有用信息的总量。不同的成对约束影响下产生的聚类结果差别很大，较大的 AI 能够使得聚类算法迭代次数减少，获得较好的聚类性能；太小的 AI 不足以起到监督聚类的作用，反而会增加算法冗余程度。但是 AI 太大会导致约束主动学习部分计算复杂，开销增大。

设成对约束集合为 PC，聚类算法为 E，P_D 为 D 在没有任何约束监督下得到的类簇，对 PC 信息总量评价为[14]：

$$INF_A(PC)=\frac{1}{|PC|}\sum_{pc\in PC} unsat(pc,P_D) \tag{5}$$

如果由类簇 P_D 不能得出成对约束 pc，则 $unsat(pc,P_D)=1$，否则为 0；$|PC|$ 为成对约束总量。

约束二阶主动学习算法能够在较小的计算代价下获得较大 AI 的约束信息。算法第一阶段依据最远优先策略选择样本点，以确保每个类簇至少有一个样本点被选取，利用成对约束性质生成闭包；第二阶段选取信息量最大的样本点添加到闭包中，直到满足约束数量。样本点属于一阶中生成的某一闭包的不确定性越大，能够预知其作用的期望越小，其信息量越大[13]。计算样本点的信息量首先需要估算其属于某一闭包的概率，设 $C=\{c_1,c_2,\cdots,c_r\}$ 为闭包集合：

$$p(d\in c_i)=\frac{\frac{1}{|c_i|}\sum_{d_j\in c_j} Sim(d,d_j)}{\sum_{k=1}^{l}\frac{1}{|c_k|}\sum_{d_j\in c_k} Sim(d,d_j)} \tag{6}$$

其中，$|c_i|$ 表示闭包 c_i 中包含样本点的数量，$Sim(d,d_j)$ 为相似度函数。

然后估算期望：

$$IE[q(d)]=\sum_{i=1}^{l} i * p(d\in c_j) \tag{7}$$

其中，$q(d)$ 表示需要确定 d 闭包关系的询问次数变量。

最后计算样本点最大信息量：

$$d_{maxinfo}=\arg\max_{d\in D}\frac{H(C|d)}{IE[q(d)]} \tag{8}$$

其中，$H(C|d)$为 d 闭包关系的熵，用于估算样本 d 的不确定性大小。

约束二阶主动学习算法详细描述如下：

输入：文档集合 D，一阶约束数量 M_1，约束总量 M_2

输出：约束闭包

1)StageI //确保每个类簇都有样本点被选取，生成初始闭包集合

2)初始化临时样本集合 CS 以及成对约束集合 PC

3)while 成对约束数量小于 M_1｛

4)　if (CS==null)

5)　｛随机选择样本点 d，添加到 CS｝

6)　else｛

7)　　根据公式$\arg\max_{y\in CS} dis_{min}(d,y)$选择 d

8)　　for (y in CS)

9)　　询问(d,y)约束类型，添加到成对约束集合 PC 中｝｝

10)根据成对约束性质生成初始闭包集合 C

11)StageII //选择信息量最大的样本点

12)while |PC|＜ M_2｛

13)　for(i=0;i＜|C|;i++)｛

14)　　估算不确定性 H(C|d)

15)　　根据式(7)估算 $p(d\in c_i)$

16)　　根据式(8)估算期望 IE[q(d)]｝

17)　根据式(9)选取信息量最大的样本点 $d_{maxinfo}$

18)　询问 $d_{maxinfo}$归属→添加到 c_k

19)Return 约束闭包 C

算法分为两个部分，1)－10)为第一阶段，采用最远优先策略选择样本点以确保在足够少的尝试次数下获得每个聚类的至少一个样本，约束数量为 M_1，根据成对约束性质形成初始闭包，为第二阶段计算样本信息量提供依据；8)－16)通过估算样本点的不确定性大小、期望得到信息量，选取剩下样本点中信息量最大的判断并添加到相应约束闭包中，以确保约束信息总量足够大。

2.2.4　文档半监督聚类 ALC-DBSCAN 算法

目前常用的聚类方法大致可以分为 3 类：1)基于划分的算法，代表算法为 K-means，其对十分庞大的数据集合聚类效果较好，但是需要预选 k 个初始划分作为聚类中心，如果 k 个初始值选择不佳，将会导致聚类结果偏差很大，并且其对噪声数据很敏感，只能发现圆形的簇。2)基于层次聚合的算法，代表算法为 AHC、BIRCH 等，它是对数据集合进行层次上的分解，细分又包括凝聚法和分裂法，分别对应自下而上和自上而下两种方法，其算法复杂度为 $O(n^2)$，只适合对小型数据集合聚类。3)基于密度的算法，代表算法为 DBSCAN，相比前两类算法，它具有可以发现任意形状聚类、对噪声不敏感等优点，能够获得全局最优解，其概念包括：核心对象、边界对象、密度可达等，其中两个重要参数为区域半径 Eps 及最小样本数目阈值 $MinPts$。以数据集合中每个点为中心统计半径区域中样本数量来进行聚类，如果样本数量大于 $MinPts$，则加入当前类簇或者新建类簇，否则标记为噪声点。这种算法可以在存在噪声的数据集合中发现任意形状的簇，但是参数 Eps、$MinPts$ 需要用户自行设定，人为因素导致聚类效果不稳定，并且其对聚类结果边界控制欠佳，尤其是对于多个类簇重叠部分，区分能力薄弱。

为解决 DBSCAN 的边界模糊问题以使得文档聚类结果更好，本文利用主动学习得到的约束闭包 $C=\{c_1,c_2,\cdots,c_r\}$ 监督聚类过程，算法具体思想为：输入区域半径 Eps、最小样本数目阈值 $MinPts$、约束闭包 C，判断聚类过程中的样本点是否属于某一闭包，通过闭包之间的关系，将整个闭包归为同一类簇，去除类簇中包含的相异闭包中的样本。通过约束条件的监督，可以矫正由于人为选取的 Eps 和 $MinPts$ 参数不合适而导致的聚类偏差，能提升聚类结果边界控制和区分能力以及聚类结果的准确性。

算法具体描述如下：

输入：文档集合 D，区域半径 Eps，最小样本数目阈值 MinPts，约束闭包 C，标记符号 flag

输出：若干类簇和噪声

1)初始化核对象集合 coreSet、边界对象集合 boundSet、样本数量 Sum

2)for (d in D)｛//遍历集合 D，逐个处理每一个元素。

3)　if (d 未被标记)｛//如果 d 未被标记，说明 d 未被处理

4)　　计算以 d 为中心、Eps 为半径的区域中的样本数量 Sum

5)　　if (Sum ＞ MinPts)｛

6)　　密度可达元素加入当前类簇(如果没有，新建并加入)，全部打上标记 flag｝

7)　　else｛

8)　　标记为噪声，打上标记 flag｝｝

9)　for (d in 当前聚类)｛

10)　　if (Sum of d ＞ MinPts)｛

11)　　加入 coreSet｝

12)　　else｛

13)　　加入 boundSet｝｝

14)　for (d′ in 当前聚类)｛

15)　　if (d′属于 coreSet &&. 属于闭包 ci)｛

16)　　闭包元素加入当前聚类，并用 flag 标记闭包中元素

17)　　if (存在边界对象 d_x 属于 c_i 相异闭包)｛

18)　　去除 d_x，去除 flag 标记｝｝

19)　　else if (d′属于 boundSet &&. 属于闭包 c_j &&. c_j 中存在核心对象)｛

20)　　去除 d′，去除 flag 标记｝｝｝

21)return 类簇及噪声

算法大致分为以下部分，1)－8)为基于 DBSCAN 的聚类，通过将计算得到的 Eps 半径中样本数量与阈值 $MinPts$ 进行比较，获得初始聚类或者噪声；9)－13)将当前类簇中样本分为核心对象和边界对象；14)－20)判断核心对象和边界对象的闭包情况，加入核心对象所属闭包，去除与类簇中有相异闭包关系的边界样本，最后返回聚类结果和噪声集合。算法在 DBSCAN 的基础上引入约束信息，能够很好地控制聚类边界，解决了两个或多个类簇由于重叠导致的边界对象归属模糊的问题。

3　聚类文档敏感信息推导方法

文档经过聚类之后形成具有相似关系的类簇，通过引入模糊集理论，对聚类结果进行敏感度区间模糊集可能性测度，评估聚类结果推导出更高级别敏感信息的可能性，高于阈值的类簇存在泄露敏感信息的风险，需要采用相应的措施防范于未然。

定义 6(敏感度) 敏感度是一个概率值 $P_s \in [0,1]$,用以表示对象 d 所含敏感信息的多少。$P_s(d)$越大,对象 d 的敏感度越高,泄露造成的损失的可能性就越大。对于任意集合 $D=(d_1,d_2,\cdots,d_n)$,D 的敏感度 $P(D)=\mathrm{Max}(P(d_i))(0\leqslant i\leqslant n)$。

定义 7(敏感级别) 根据敏感度大小将其泛化地划分为 5 个敏感级别 $Class=\{class_1,class_2,class_3,class_4,class_5\}$,$class_1=$(公开$|p=0$),$class_2=$(内文$|0<p\leqslant 0.3$);$class_3=$(秘密$|0.3<p\leqslant 0.7$);$class_4=$(机密$|0.7<p\leqslant 0.9$);$class_5=$(机密以上$|0.9<p\leqslant 1$)。允许存储对象 d 的最低的敏感级别称为 d 的安全等级。

模糊集合是模糊数学的基础,区别于普通集合元素归属非此即彼的特性,其所描述的事物本身含义具有不确定的性质,概念边界相对模糊。

定义 8(模糊集合)[15] 设在论域 U 上给定一个映射

$$A:U\rightarrow[0,1]$$
$$u \mapsto A(u) \tag{9}$$

则称 A 为 U 上的模糊集(Fuzzy Set),u 在 U 中取值,$A(u)$称为 A 的隶属函数(或者 u 对 A 的隶属度)。

定义 9(可能性分布) 设 X 是在论域 U 上取值的一个变量,Π_X是与变量 X 有关的可能性分布,π_X 表示与 X 有关的可能性分布函数(或Π_X的函数,其具体分布函数如何确定不在本文的讨论范围),并在数值上等于 B 的隶属度,即

$$\forall u\in U,\pi_X(u)=B(u) \tag{10}$$

式(10)表示的是 $X=u$ 的可能性,$\pi_X(u)$假定等于 $B(u)$。

定义 10(可能性测度) 设 B 为 U 上的模糊集,Π_X是与变量 X 有关联的可能性分布,X 在 U 中取值,则 B 的可能性测度定义为

$$P(X \text{ is } B)\overset{\triangle}{=}\pi(B)\overset{\triangle}{=}\bigvee_{u\in U}(B(u)\wedge\pi_X(u)) \tag{11}$$

符号"$\vee$"、"$\wedge$"在模糊集运算中分别表示取大和取小。

设经过聚类之后形成的类簇为 $Cluster=(cluster_1,cluster_2,\cdots,cluster_k)$,设论域 U 为 $cluster_i$ 所有子集组成的集合,$U=\{$子集$|$元素属于 $cluster_i\}$,X 在 U 中取值,则由命题"X 在其安全级别中"引入的可能性分布为

$$\Pi_X=\frac{p_1}{u_1}+\frac{p_2}{u_2}+\frac{p_3}{u_3}+\cdots \tag{12}$$

B(大于类簇 $cluster_i$ 的敏感等级)为 U 上的模糊集,则有

$$B=\frac{p_1'}{u_1}+\frac{p_2'}{u_2}+\frac{p_3'}{u_3}+\cdots \tag{13}$$

则聚类得到的类簇推导出更高级别敏感信息的推导函数为

$$P(X \text{ is } B)\overset{\triangle}{=}\bigvee_{u\in U}(B(u)\wedge\pi_X(u))\overset{\triangle}{=}\bigvee_{0\leqslant i\leqslant|U|}(p_i\wedge p_i') \tag{14}$$

$|U|$为论域 U 的大小。

设 τ 为能接受的最低的推导可能性阈值,当 $P(X \text{ is } B)\geqslant\tau$ 时,表示当前聚类中 X 文档子集推导出更高级别敏感信息的可能较大,存在聚合推导信息泄露的风险,应当禁止用户对此文档子集具有同时访问的权限。

4 仿真实验

为了验证本文提出的半监督聚类算法和敏感信息推导方法的效果,设计了一组实验进行检验,具体实验系统为 Windows 7(X86),处理器为 Pentium(R) Dual-core E5700 3.0 GHz,2GB 内存,实验环境为 Microsoft Visual Studio 2010。为了便于分析和统计实验结果,所用数据采用人工合成和实际文档的方式从本单位收集,共计 2000 份,分别模拟真实数据和将实际文档拆分为相似文档,由于拆分后的文档包含因数据细节不全而可能出现敏感度降低的情况,也为了便于测试推导方法的效果,实验数据采用 2.1 节所述方法进行预处理,并通过式(4)计算得到相似度。假设所有文档的标准聚类结果已知。

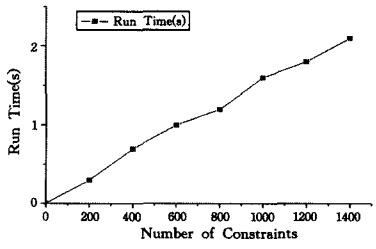

图 4 约束信息总量与时间开销的关系

主动学习得到的约束会直接影响聚类效果,适当的 AI 能够在较小的时间开销内获得较好的监督指导效果,参数 M_2 为约束信息总量,图 4 为约束信息总量与运算时间的关系。

通过图 4 可知,信息总量与开销成正相关,时间主要用于选择信息量最大的样本点。

本文采用归一化互信息(Normalized Mutual Information,NMI)对聚类算法进行评价,其表征的是聚类结果和数据真实划分的相似度。NMI 取值区间为[0,1],值越大表示聚类效果越好,反之亦然。约束信息对聚类结果影响的实验结果如图 5 所示。

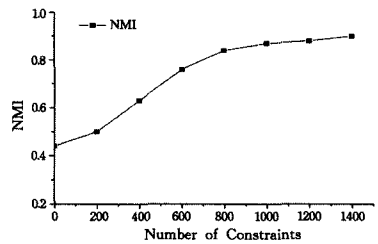

图 5 Number of Constraints-NMI 关系

从图 5 可知,少量的约束信息就能够大大提升聚类效果,随着约束量增加,其越来越逼近真实聚类,在总量为 800 后,约束再增加,但其提升的效果趋于平缓。

在约束总量为 800 的条件下,对比了 DBSCAN 和 ALC-DBSCAN 算法。图 6 是两种聚类算法在不同样本数量情况下聚类效果的实验结果。

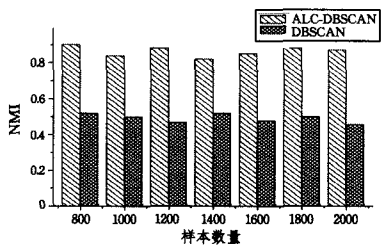

图 6 ALC-DBSCAN 与 DBSCAN 聚类性能的对比

实验结果表明,ALC-DBSCAN 聚类算法性能明显优于

无约束信息监督聚类的 DBSCAN。

推导方法的准确率与聚类算法以及阈值 τ 都有关系，在约束总量为 800 的前提下进行聚类，分别取不同的 τ 对推导方法的性能进行分析，用准确率(Precise Rate，PR)、错误率(Fault Rate，FR)来度量其性能好坏。PR 为实验中推导出来的敏感信息的文档数量与实际存在推导问题文档的总量的比值，FR 表示实验中检测出来的非推导问题文档数量与实际存在推导问题文档总数的比值，在 τ 分别取值 0.6，0.65，0.7，0.75，0.8 时推导方法的准确率、错误率如图 7 所示。

随着约束信息总量增加，约束二阶主动学习算法的时间开销也相应增大，其时间主要用于最大信息量样本点的选择；约束 AI 越大，其指导的聚类产生的结果也呈上升趋势，当约束总量达到某一个值后，聚类效果趋于稳定，是因为如果当前约束信息有足够的能力控制 DBSCAN 聚类中出现的边界模糊问题时，再增加约束信息也不能改善聚类算法的性能，达到峰值；在约束总量为 800 的前提下，对取不同阈值推导方法的准确率、错误率进行实验分析得出：合适的 τ 能得到很好的推导准确率，其过高或过低都会导致准确率下降、错误率上升的问题。

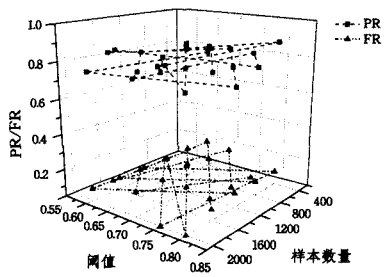

图 7　推导方法准确率(PR)及错误率(FR)的分布

与文献[2]中针对社交网络涉及的用户信息推理通道和文献[4]中针对多级安全中客体的推理分析相比，本文处理的文档所包含的内容更多、更广泛，处理过程具有更高的复杂性和综合性。较文献[2]中利用分散的社交信息，基于约束主动学习的半监督聚类能将具有相似信息的文档聚合起来，使得敏感推导方法具有更高的准确性。

结束语　本文通过分析多文档聚类后存在敏感信息推导的问题，提出一种基于半监督聚类的文档敏感信息推导发现方法。首先设计了一种约束二阶主动学习方法，在较少的时间开销内生成最有价值的约束闭包；将约束作为监督信息指导聚类，提出一种新的半监督聚类算法 ALC-DBSCAN，其提高了类簇边界区分和控制能力，改善了聚类性能；最后引入模糊数学理论，提出了多文档敏感信息推导方法。

下一步将更加深入地研究文档之间存在的关系和可能存在的推理通道，进一步降低敏感信息泄露风险。

参 考 文 献

[1] Motro A, Marks D G, Jajodia S. Aggregation in relational databases: Controlled disclosure of sensitive information[M]// Computer Security—ESORICS 94. Springer Berlin Heidelberg, 1994: 429-445

[2] Accorsi R, Müller G. Preventive inference control in data-centric business models[C]// 2013 IEEE Security and Privacy Workshops (SPW). IEEE, 2013: 28-33

[3] 冯婷. 安全数据库的推理通道问题研究[D]. 南京：南京航空航天大学，2010
Feng Ting. The study of the inference of security database[D]. Nanjing: Nanjing University of Aeronautics and Astronautics, 2010

[4] 曹利峰，陈性元，杜学绘，等. 基于聚类分析的客体聚合信息级别推演方法[J]. 电子与信息学报，2012，34(6)：1432-1437
Cao Li-feng, Chen Xing-yuan, Du Xue-hui, et al. A level inference method for aggregated information of objects based on clustering analysis[J]. Journal of Electronics and Information Technology, 2012, 34(6): 1432-1437

[5] 王玲，薄列峰，焦李成. 密度敏感的半监督谱聚类[J]. 软件学报，2007，18(10)：2412-2422
Wang Ling, Bo Lie-feng, Jiao Li-cheng. Density-Sensitive Smi-Supervised spectral clustering[J]. Journal of Software, 2007, 18(10): 2412-2422

[6] Wagstaff K, Cardie C. Clustering with instance-level constraints [C]// Proc. of the 17th Int'l Conf. on Machine Learning. 2000: 1103-1110

[7] 赵卫中，马慧芳，李志清，等. 一种结合主动学习的半监督文档聚类算法[J]. 软件学报，2012，23(6)：1486-1499
Zhao Wei-zhong, Ma Hui-fang, Li Zhi-qing, et al. Efficiently active learning for Smi-Supervised document clustering[J]. Journal of Software, 2012, 23(6): 1486-1499

[8] Jain A K. Data clustering: 50 years beyond K-means[J]. Pattern Recognition Letters, 2010, 31(8): 651-666

[9] 苏赢彬，杜学绘，夏春涛，等. 基于文档平滑和查询扩展的文档敏感信息检测方法[J]. 计算机应用，2014，34(9)：2639-2644
Su Ying-bin, Du Xue-hui, Xia Chun-tao, et al. Sensitive information detection approach for documents based on document smoothing and query expansion[J]. Journal of Computer Applications, 2014, 34(9): 2639-2644

[10] Goyal P, Behera L, Mcginnity T M. A novel neighborhood based document smoothing model for information retrieval[J]. Information retrieval, 2013, 16(3): 391-425

[11] Settles B. Active learning literature survey [R]. University of Wisconsin-Madison, 2010

[12] 龙军，殷建平，祝恩，等. 主动学习研究综述[J]. 计算机研究与发展，2008，45(z1)：300-304
Long Jun, Yin Jian-ping, Zhu En, et al. The research of active learning[J]. Journal of Computer Research and Development, 2008, 45(z1): 300-304

[13] Xiong S, Azimi J, Fern X Z. Active learning of constraints for semi-supervised clustering [J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(1): 43-54

[14] Davidson I, Wagstaff K. Measuring constraint-set utility for partitional clustering algorithms[M]// Lecture Notes in Computer Science, Vol 4213. Springer, 2006: 115-125

[15] 杨纶标，高英仪，等. 模糊数学原理及应用(第三版)[M]. 广州：华南理工大学出版社，2005：338-344
Yang Lun-biao, Gao Ying-yi, et al. The principle and application of fuzzy mathematics (third edition) [M]. Guangzhou: South China University of Technology Press, 2005: 338-344