基于特征权重量化的相似度计算方法

刘 铭[1),2)] 吴 冲[1)] 刘远超[2)] 孙承杰[2)]

[1)](哈尔滨工业大学管理学院 哈尔滨 150001)

[2)](哈尔滨工业大学计算机科学与技术学院 哈尔滨 150001)

摘 要 随着信息产业的迅猛发展,聚类的无监督特性使其成为一种极为有效的分析工具.而为获得良好的聚类结果,有效及准确的相似度计算方法是其必备的前提条件.事实上,在描述数据相似度时,不同的特征显然具有不同的作用,因此有必要借助一些先验知识,例如用户提供的限制数据,来衡量特征的重要性,并将其应用于相似度计算中以获取更加准确的计算结果.传统的特征权值量化方法均忽视了两点问题:(1)限制数据在特征空间中极有可能为非均匀分布;(2)限制数据可能包含不一致性.上述问题的存在使得传统的权值量化方法无法获得准确的结果甚至无法运行.基于此,文中提出了一种新颖的特征权值量化方法用以处理上述两点问题:(1)将限制数据划分为若干个等价类,进而通过计算参数"分布系数"来均匀化数据的分布;(2)将限制数据连接为无向图,进而通过计算参数"置信度"来衡量及弱化限制数据的不一致性.之后将这两个参数结合到特征权值量化函数中以获得准确的相似度计算结果.实验结果显示:该特征权值量化方法能够结合限制数据来获取不同特征对相似度计算的贡献能力,并能应用于任何聚类算法中以提高聚类的准确度.

关键词 限制数据;特征权重量化;分布系数;置信度

中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2015.01420

Similarity Calculation Based on Feature Weight Evaluation

LIU Ming[1),2)] WU Chong[1)] LIU Yuan-Chao[2)] SUN Cheng-Jie[2)]

[1)](*School of Management*, *Harbin Institute of Technology*, *Harbin* 150001)

[2)](*School of Computer Science and Technology*, *Harbin Institute of Technology*, *Harbin* 150001)

Abstract Along with high-speed advance of information technology, the unsupervised characteristic of clustering makes itself an effective implement for data analysis. To acquire high clustering performance, the effective and precise similarity calculation plays a prime and necessary role for clustering algorithms. Owing to the fact that different features have diverse contributions to describe similarity among data, it is necessary to assess feature's contribution by means of some transcendental knowledge (e.g. constrained data provided by users), and import it in similarity measurement to acquire more precise calculating results. Unfortunately, conventional weight evaluating methods all fail to consider two challenges: (1) high possibility of asymmetrical distribution of constrained data in feature space; (2) high possibility of inconsistency contained by constrained data. Previous two issues disable conventional weight evaluating methods to acquire high precision, and even make them unable to work. Hence, this paper proposes a novel constraint based weight evaluating method to deal with them. For the former one, constrained data are partitioned into several equivalent classes, and distributing parameters are assigned to them to balance their

收稿日期:2013-01-24;最终修改稿收到日期:2015-01-18.本课题得到国家自然科学基金青年基金(61300114)、高等学校博士学科点专项科研基金(20132302120047)、中国博士后科学基金特别项目(2014T70340)、中国博士后科学基金面上项目(2013M530156)、中央高校基本科研业务费专项资金(HIT.NSRIF.2013066)及 CCF-腾讯犀牛鸟基金资助.**刘 铭**,男,1981 年生,博士,讲师,中国计算机学会(CCF)会员,主要研究方向为文本聚类、文本挖掘.E-mail:liuming1981@hit.edu.cn.**吴 冲**,男,1971 年生,博士,教授,主要研究领域为数据管理、信息处理.**刘远超**,男,1971 年生,博士,副教授,主要研究方向为自然语言处理.**孙承杰**,男,1980 年生,博士,讲师,主要研究方向为信息推荐.

distributions. For the latter one, constrained data are connected to form an undirected graph, and belief values are thereby computed to measure and reduce their possibilities to be inconsistent. Finally, these two parameters are integrated in weight evaluating function to form an accurate similarity measurement. Experimental results demonstrate that, this weight evaluating method can combine constrained data to obtain diverse contributions of different features to similarity calculation, and can be applied in any clustering algorithm to improve its precision.

Keywords　constrained data; evaluation of weight of feature; distributing parameter; belief value

1　引　言

随着网络技术的迅猛发展，人们接触的信息量与日俱增，用户急需一种有效的信息分析工具以协助其日常工作. 如文献[1]所述，聚类即是一种有效的信息分析工具，其通过凝聚相似数据能够缩小用户的查找范围并加快用户寻找相关信息的速度.

聚类中最基本的要素就是数据间的相似度[2]，有效的相似度度量函数显然能够帮助聚类算法获得良好的聚类结果. 目前大多数聚类算法以向量空间模型组织数据，并通过计算不同数据间特征向量的夹角或距离来反映数据之间的相似度，例如欧式距离[3]、余弦相似度[4]. 此类相似度计算方法视所有特征对数据相似性的描述能力或对数据的划分能力是相同的[5]. 然而，现实中不同特征对数据的划分能力显然是不同的，因此有必要分析特征对相似度计算的贡献能力来为特征赋予不同的权值.

传统的聚类技术是一种无监督的学习方法，在算法运行前不需要获取任何先验知识. 然而，现实应用中，用户对于输入数据可能存在某些限制，而聚类结果显然要满足用户对于输入数据的限制. 目前最常使用的限制信息是文献[6]中提及的 must-link 和 can't-link 点对限制信息. 如果用户指定输入数据中的任两个数据位于同一类别中，则说明这两个数据或点对满足 must-link 关系，而 can't-link 关系正好相反. 此类限制信息刚好可以结合到特征权值量化中去以获得更为准确的相似度计算结果. 基于此，研究者们提出了多种特征权值量化方法[7-8]. 然而传统的基于限制数据的特征权值量化方法均无法处理以下两种情况：

(1) 用户指定的限制数据的数量通常远少于全部的输入数据，这使得限制数据经常是从整个特征空间中非均匀抽取的；

(2) 传统的特征权值量化方法认为用户提供的限制数据是准确无矛盾的，然而现实应用中用户提供的限制数据中某些满足 must-link 关系的数据对可能同时满足 can't-link 关系①.

当存在第 1 个问题时，非均匀分布的限制数据会使特征权值量化的结果出现“过适应”现象，即错误的将那些能够有效划分密集的限制数据的特征赋予较大的权值，而忽略了分布稀疏的限制数据对特征权值量化结果的影响. 针对此问题，本文引入参数“分布系数”来平衡限制数据的分布，降低密集分布的限制数据对特征权值量化结果的影响，同时提高稀疏分布的限制数据的作用，以防止出现“过适应”现象，具体介绍参见文中第 3 节.

当存在第 2 个问题时，传统的特征权值量化方法均无法对其进行处理. 针对此问题，本文引入参数“置信度”来衡量限制数据的不一致性，并对不一致的限制数据赋予较小的权值来降低其在特征权值量化中的作用. 具体介绍参见文中第 4 节.

2　基于限制数据的特征权值量化

引入限制数据的聚类方法可分为两类：一类是在聚类算法的有效性函数或搜索策略中融合限制数据引入惩罚函数，如 COP-K-means[9] 和 KCL[10]，另一类是根据限制数据修改算法的相似度计算函数以使类别划分结果满足用户的要求. 相比于第 2 类方法，第 1 类方法对算法的依赖性较大，针对不同的算法需要设计不同的优化函数或搜索策略，因此无法做到普适性. 第 2 类方法通常以特征的重要性为依据，依据特征对数据分布的影响对特征赋予不同的权值，进而去修改相似度计算函数[11-12]. 依据此想法，即可从用户指定的限制数据中挖掘能够有效凝聚相似数据以及区分不同数据的特征，并增大其在相似度计算中的作用，以使聚类结果满足用户的要

① 关于传递特性的介绍参加文中第 3 节式(9).

求.由上述分析可见:这种结合限制数据进行特征权值量化的方式不依赖于特定的算法,其可以应用于任何聚类算法中以提高聚类的准确性.

假设待聚类的数据集合为 D,设 D 的势为 n,则 D 中包含的可能点对数目为 $n(n-1)/2$,设存放这些点对的集合为 S.设集合 MCS 和 NCS 分别包含用户指定的满足 must-link 关系和 can't-link 关系的点对,其中 $MCS\in S$,$NCS\in S$.以 $(i,j)\in MCS$ 代表点对 (i,j) 满足 must-link 关系,同理 $(i,j)\in NCS$ 代表 (i,j) 满足 can't-link 关系.设 NDS 为非限制数据集合,其包含了 S 中除 MCS 和 NCS 外的其他点对.

本文以向量空间模型组织待聚类数据,并以式(1)计算输入数据之间的相似度,例如 p 与 q.式(1)被许多文献索引用来作为衡量数据之间的相似度的依据[13-14],其中 w_k 代表第 k 维特征在相似度计算中的重要性,$w_k\in[0,1]$.当 $w_k=1$ 时,所有特征的重要性相同,即为传统的欧式距离.

$$d_{pq}^{(w)}=\sum_{k=1}^{m}w_k^2(x_{pk}-x_{qk})^2 \tag{1}$$

式(1)的计算结果过于分散为 $[0,+\infty)$,不利于进行权值量化,因此本文对式(1)进行改变从而得到更为灵活的相似度计算函数 $-\rho_{pq}^{(w)}$,并通过参数 β 使 $\rho_{pq}^{(w)}$ 均匀分布于 0～1 之间.

$$\rho_{pq}^{(w)}=\frac{1}{1+\beta\times d_{pq}^{(w)}} \tag{2}$$

其中,参数 β 可由式(3)得到

$$\frac{1}{n\times(n-1)}\sum_{p,q\in D}\frac{1}{1+\beta\times d_{pq}^{(w)}}=0.5 \tag{3}$$

假设用户提供的限制数据是准确一致的①,那么显然位于 MCS 中的点对间的相似度应该较大,而位于 NCS 中的点对间的相似度应该较小.这样那些能够缩小 MCS 中点对间的相似度、而增大 NCS 中点对间的相似度的特征,对限制数据的划分能力较强,即上述特征在相似度计算中的作用应该较大,以此为依据即可得到下述的特征权值量化公式.当此公式达到最小值时对应的特征权值为最优的特征权值.

$$FW=\sum_{(i,j)\in MCS}\sum_{(k,l)\in NCS}\rho_{kl}^{(w)}\times\log\rho_{kl}^{(w)}+(1-\rho_{ij}^{(w)})\times\log(1-\rho_{ij}^{(w)}) \tag{4}$$

使用随机梯度下降算法优化特征的权值 w_k[15],即可得权值更新幅度 Δw_k:

$$\Delta w_k=\frac{\partial FW}{\partial w_k}=\sum_{(i,j)\in MCS}\sum_{(k,l)\in NCS}\times$$

$$\frac{\partial(\rho_{kl}^{(w)}\times\log\rho_{kl}^{(w)}+(1-\rho_{ij}^{(w)})\times\log(1-\rho_{ij}^{(w)}))}{\partial w_k}$$

$$=\sum_{(i,j)\in MCS}\sum_{(k,l)\in NCS}\frac{\partial\rho_{kl}^{(w)}}{\partial w_k}\times\log\rho_{kl}^{(w)}+\rho_{kl}^{(w)}\times\frac{1}{\rho_{kl}^{(w)}}\frac{\partial\rho_{kl}^{(w)}}{\partial w_k}-\frac{\partial\rho_{ij}^{(w)}}{\partial w_k}\times\log(1-\rho_{ij}^{(w)})-(1-\rho_{ij}^{(w)})\times\frac{1}{(1-\rho_{ij}^{(w)})}\times\frac{\partial\rho_{ij}^{(w)}}{\partial w_k} \tag{5}$$

其中,$\frac{\partial\rho_{ij}^{(w)}}{\partial w_k}$为

$$\frac{\partial\rho_{ij}^{(w)}}{\partial w_k}=\frac{\partial\frac{1}{1+\beta\times d_{ij}^{(w)}}}{\partial w_k}=-(1+\beta\times d_{ij}^{(w)})^{-2}\times\beta\times\frac{\partial d_{ij}^{(w)}}{\partial w_k}=-2\times(1+\beta\times d_{ij}^{(w)})^{-2}\times\beta\times w_k\times(x_{ik}-x_{jk})^2 \tag{6}$$

将式(2)、式(5)和式(6)综合在一起即可得 $t+1$ 时刻的特征权值:

$$w_k(t+1)=w_k(t)-\frac{1}{\sqrt{2\pi}\delta(t)}\times(e^{-\frac{\Delta w_k(t)^2}{\delta(t)^2}})\times(\Delta w_k(t)) \tag{7}$$

其中,以高斯函数控制梯度下降的步长,$\delta(t)$ 为线性时间衰减函数,以使下降的步长逐渐减小[15].

3 分布系数

式(4)～(7)仅考虑了限制数据在权值量化中的作用.事实上,用户提供的限制数据可能仅占全部数据的很少一部分,这样在按照式(4)进行权值量化时很可能会出现偏置现象,即特征对于限制数据过分表述,从而错误的描述了非限制数据间的相似关系,因此需要在权值量化中结合非限制数据.

如文献[16]所述:按式(2)计算数据间的相似度时,对于相似度趋近于 1 的数据对,其相似性较小,基本上不可能位于同一类别中,而对于相似度趋近于 0 的数据对正好相反,对于相似度趋近于 0.5 的数据对,其对于数据的类别分布的描述最为模糊,基本上无法辨别出数据的类别归属.如果输入数据中存在大量这样的数据会严重降低聚类算法的性能.因此可以通过优化特征的权值以使相似度小于 0.5 的数据对其相似度趋近于 0,而使相似度大于 0.5 的数据对其相似度趋于 1,以提高数据的可分性从

① 关于一致性的叙述详见文中第 4 节.

而提高聚类结果的准确性.

式(8)将非限制数据和限制数据结合在一起进行权值量化,其中 μ 为用户指定的参数,代表限制数据相对于非限制数据的重要程度.

$$FW=\mu\Big(\sum_{(i,j)\in MCS}\ \sum_{(k,l)\in NCS}\rho_{kl}^{(w)}\times\log\rho_{kl}^{(w)}+(1-\rho_{ij}^{(w)})\times\log(1-\rho_{ij}^{(w)})\Big)+(1-\mu)\times\Big(\sum_{m,n\in NDS\&\&\rho_{mn}^{(w)}<0.5}\ \sum_{r,t\in NDS\&\&\rho_{rt}^{(w)}>0.5}\rho_{mn}^{(w)}\times\log\rho_{mn}^{(w)}+(1-\rho_{rt}^{(w)})\times\log(1-\rho_{rt}^{(w)})\Big)\quad(8)$$

上述权值量化函数 FW 将数据分解为两部分:限制数据和非限制数据.但如前所述,用户指定的限制数据可能仅仅是全部数据的一个不均匀抽样,此时如果仍然对所有限制数据赋予同样的重要性,显然会在权值量化结果中引入误差,过分加大密集分布的限制数据在权值量化中的作用,从而忽略了分布稀疏的限制数据的影响.针对此问题,本文依据 must-link 关系将限制数据划分为多个等价类,然后据此计算"分布系数"以平衡限制数据的分布.

如文献[6]所述:must-link 关系满足自反、对称和传递特性,因此可将其视为等价关系,记为 R_{must},如式(9)所示①.

$$pR_{\text{must}}q;pR_{\text{must}}q\Leftrightarrow qR_{\text{must}}p;pR_{\text{must}}q,qR_{\text{must}}h\Rightarrow pR_{\text{must}}h\quad(9)$$

文献[6]将 can't-link 也看作为等价关系,但显然 can't-link 不满足传递特性,即 $pR_{\text{can't link}}q$,$qR_{\text{can't link}}h\not\Rightarrow pR_{\text{can't link}}h$.因此本文仅以 R_{must} 将满足 must-link 关系的点对集合 MCS 划分为若干个等价类,记这多个等价类的集合为 K_{must}.划分方式如算法 1 所示.

算法 1.

输入:满足 must-link 关系的点对集合 MCS,等价类 K,数据标记集合 C

输出:等价类集合 K_{must}

划分过程:

1. 从 MCS 中任选一个点对,设其为(a,b);
2. 清空 K 和 C;
3. 将(a,b)放入 K 中,并将(a,b)从 MCS 中删除;
4. 将 a 和 b 分别放入 C 中;
5. 从 C 中任选一个未标记数据,假设选择的是 a,标记 a 为已选择;
6. 将 MCS 中所有包含 a 的点对都放入 K 中,并将这些点对从 MCS 中删除;
7. 将每个点对中的数据无重复的放入 C 中(即如果某数据出现在 C 中,则不放入);
8. 如果 C 中的数据均为已标记则转步 9,否则转步 5;
9. 将 K 作为新的等价类放入 K_{must} 中;
10. 转步 1 直至 MCS 为空.

按照上述等价类的划分方式,每个等价类中包含的任一点对均满足 must-link 关系(某些 must-link 关系是根据传递特性推出的,比如点对(a,b)、(b,c)分别满足 must-link 关系,则点对(a,c)也满足 must-link 关系).由于满足 must-link 关系的点对为用户指定的位于同一类别中的数据,即相似的数据,因此这些数据位于数据空间中相对密集的区域内.这样即可使每个密集分布的区域在权值量化函数中拥有同样的重要性以平衡限制数据的非均匀分布.

然而,仅仅根据 R_{must} 无法确定 K_{must} 中不同的等价类是否可以合并为一个大的等价类,即无法确定两个不同的等价类中的数据是否分布于同一个密集区域内.而集合 NCS 恰好提供了这种信息:如果 K_{must} 中有两个等价类分别含有 NCS 中的一个点对中的两个数据,则显然这两个等价类中包含的数据隶属于不同的类别.按照如上方法即可根据 NCS 将 K_{must} 重新划分,将 K_{must} 中两个确定不在同一类别中的等价类放入到集合 T 中,并将它们从 K_{must} 中删除.对于 K_{must} 中剩余的、无法确定是否可以合并的等价类,如 P 和 Q,可以通过式(10)估算 P 和 Q 之间的相似度,并确定其是否可以合并.

$$V(P,Q)=\sum_{k\in P,l\in Q,\rho_{kl}^{(w)}>0.5}1-\sum_{i\in P,j\in Q,\rho_{ij}^{(w)}<0.5}1\quad(10)$$

如式(10)所述,当等价类 P 和 Q 之间的值大于 0 时,可以认为这两个等价类内的相似数据多于不相似数据,则这两个等价类应属于同一类别,此时将 P 和 Q 合并后插入到集合 T 中.反之认为 P 和 Q 属于不同类别,将它们分别插入到集合 T 中.

在将限制数据划分为多个密集区域(等价类)后,即可根据每个区域内包含的数据数分别对不同区域内的数据赋予不同的分布系数(λ_b),并将其结合到权值量化函数中去,结果如式(11)所述.其中 T_b 代表集合 T 中的第 b 个等价类,$|T|$ 代表集合 T 中等价类的数目.按照梯度下降算法对 FW 进行优化后,即可得到类似于式(7)的权值更新函数.

$$FW=u\sum_{T_b\in T}\lambda_b\Big(\sum_{i,j\in T_b}\ \sum_{(k\in T_b\&\&l\notin T_b)\|(k\notin T_b\&\&l\in T_b)}\rho_{kl}^{(w)}\times\log\rho_{kl}^{(w)}+(1-\rho_{ij}^{(w)})\times\log(1-\rho_{ij}^{(w)})\Big)+(1-u)\times\Big(1-\sum_{T_b\in T}\lambda_b\Big)\times$$

① 式(9)中等价关系的特性可参见文献[6,9],在此就不赘述了.

$$\Big(\sum_{m,n\in NDS\&\&\rho_{mn}^{(w)}<0.5}\ \sum_{r,t\in NDS\&\&\rho_{rt}^{(w)}>0.5}\rho_{mn}^{(w)}\times$$

$$\log\rho_{mn}^{(w)}+(1-\rho_{rt}^{(w)})\times\log(1-\rho_{rt}^{(w)})\Big)\quad(11)$$

式(12)为第 b 个等价类的分布系数 λ_b 的计算方法. 其中,T_x 代表 T 中的某个等价类,$|T_x|$ 代表集合的势,其他符号的含义如前所述. 由此公式可见,该参数平衡了分布不同的各等价类内的数据对特征权值量化结果的影响,其降低了密集分布(等价类规模较大)的数据对特征量化结果的影响,而提升了稀疏分布(等价类规模较小)的数据的作用.

$$\lambda_b=\frac{\sum_{x=1\&\&x\neq b}^{|T|}|T_x|}{|T|\Big(\sum_{x=1}^{|T|}|T_x|+|NDS|\Big)}\quad(12)$$

4 置信度

传统的特征权值量化方法认为用户提供的限制数据是一致的,即如果用户认为点对 (p,q) 在 MCS 中,即 (p,q) 满足 must-link 关系,那么 (p,q) 是不可能位于 NCS 中的,即 (p,q) 不满足 can't-link 关系. 然而在实际应用中,尤其是进行聚类反馈时,是无法保证用户指定的限制数据中不存在上述矛盾的. 即用户可能指定 (a,b)、(b,c) 分别满足 must-link 关系,这样根据传递特性,点对 (a,c) 也满足 must-link 关系,但是随着用户指定的点对数目的增多,用户不可能仔细检查每个点对,因此用户可能错误的指定 (a,c) 满足 can't-link 关系,此时即出现了不一致性. 这样即无法使用上述 FW 进行特征权值量化. 针对此问题,本文为 T 中的每个等价类内的数据点对提供置信度,以确定其满足用户指定的限制关系的可信性,并融合此置信度进行特征权值量化.

本文以边连接等价类中满足 must-link 关系(R_{must})且一致的点对. 显然对于一个不存在不一致性的点对的等价类,其应该对应于一个全连通图. 然而本文仅仅保留了 MCS 中包含的点对之间的边,而删除根据传递特性推导出的满足 must-link 关系的点对间的边. 上述做法可以理解为:与 MCS 中用户指定的满足 must-link 关系的点对相比,那些推导出的满足 must-link 关系的点对的可信度较低,因此删除连接上述点对之间的边.

设 B 中存储了存在不一致性的数据点对,即 $\forall(p,q)\in B;(p,q)\in MCS\&\&(p,q)\in NCS$. 假设 (p,q) 位于等价类 T_b 中,这样即可根据 T_b 内的其他数据,例如 i 和 j,与 p 和 q 间的距离来确定点对 (i,j) 是否是一致的置信度. 显然,如果 T 中的某个等价类(T_b)中不存在不一致的点对,则该等价类中的任意点对间的置信度均为 1. 如果 T 中的某个等价类中存在不一致的点对,设其为 (p,q),则对于 T_b 中的其他点对,例如 (i,j),如果 i 和 j 分别与 p 和 q 在连通图中距离较近,则 i 和 j 反映的信息分别与 p 和 q 相似,则 (i,j) 是不一致的可能性较大,即其置信度应该较低,反之,其置信度应该较高;当然,等价类中可能存在着多个不一致的点对,此时分别计算每个点对相对于每个不一致点对的置信度,并以其最小值作为该点对的置信度. 对于不一致的点对,其置信度显然为 0.5,即其满足 must-link 关系和 can't-link 关系的可能性均为 50%. 上述想法即可转化为式(13):

$$o_{ij}=\begin{cases}1; & \forall(i,j)\in T_b;(i,j)\notin B\\ \min\limits_{(p,q)\in B}\Big(\dfrac{1}{\min(path(i,p),path(i,q))}\times\dfrac{1}{\min(path(j,p),path(j,q))}\Big); & \text{If}(p,q)\in B;\exists(i,j),b;\\ & (i,j)\notin B\&\&(i,j),(p,q)\in T_b\\ 0.5; & \text{If}(i,j)\in B\end{cases}\quad(13)$$

式(13)中,$path(j,p)$ 对应于 j 与 p 之间相距的最少边数.

将置信度融合到权值量化函数中即可获得如式(14)所示的带有置信度的权值量化函数 FW:

$$FW=u\sum_{b\in T}\lambda_b\Big(\sum_{i,j\in T_b}\ \sum_{(k\in T_b\&\&l\notin T_b)\|(k\notin T_b\&\&l\in T_b)}o_{kl}\rho_{kl}^{(w)}\times\log\rho_{kl}^{(w)}+o_{ij}(1-\rho_{ij}^{(w)})\times\log(1-\rho_{ij}^{(w)})\Big)+(1-u)\Big(1-\sum_{b\in T}\lambda_b\Big)\times\Big(\sum_{m,n\in NDS\&\&\rho_{mn}^{(w)}<0.5}\ \sum_{r,t\in NDS\&\&\rho_{rt}^{(w)}>0.5}\rho_{mn}^{(w)}\times\log\rho_{mn}^{(w)}+(1-\rho_{rt}^{(w)})\times\log(1-\rho_{rt}^{(w)})\Big)\quad(14)$$

在采用梯度下降算法优化 FW 后,即可将特征的权值代入到式(2)中实现带有特征权值的相似度计算方法,并使其代替传统的数据相似度计算方法(例如欧式距离)应用于任意的聚类算法中.

5　实验及分析

本文采用 UCI 数据集[17]中的 Waveform、Pima 作为基准测试集，同时还采用了 Newsgroup 新闻语料和 863 文本分类语料来检测本文提出的权值量化方法在多特征、多类别下的表现情况. 表 1 中列出了各基准数据集的相关信息. 实验中分 3 种情况验证了本文所提出的权值量化函数的性能：

(1) 首先验证了在不存在不一致的限制数据的前提下，非均匀选择限制数据时，本文提出的权值量化函数的性能(主要证明“分布系数”的性能)；

(2) 接下来验证了在均匀选择限制数据的前提下，限制数据中存在不一致性时，本文提出的权值量化函数的性能(主要证明“置信度”的性能)；

(3) 最后验证了在非均匀选择限制数据且限制数据存在不一致性时，本文提出的权值量化函数的性能(主要证明“分布系数”和“置信度”可并存).

按照上述设计，本文需要在限制数据均匀分布和非均匀分布两种情况下测试权值量化函数的性能. 为实现上述测试，本文采用如下方法：

以数据集 Waveform 为例，其他数据集也采用类似方法. Waveform 数据集包含 3 个类别，每个类别中包含相似的数据，其意味着同一类别中的数据在数据空间中是紧密分布的. 为保证选择的限制数据是均匀分布的，可以从每个密集分布的区域内选择等数量的限制数据. 例如，如果需要选择 20 对均匀分布的限制数据完成测试(其他数量的限制数据采用类似方法)，且假设 Waveform 的 3 个类别分别标记为 A\B\C，则可以分别从 3 个类别中选择 6\7\7 对限制数据(由于 20 不可能为 3 整除，因此近似的选择 6\7\7 以保证限制数据接近均匀分布).

为使限制数据非均匀分布，可以从单一的类别中选择限制数据，比如对于 Waveform，可以仅从 A 类别中选择 20 对限制数据.

为在限制数据中包含一致性和不一致性下分别测试权值量化函数的性能，本文采用如下方法：

为从 Waveform 中选择 20 对一致的限制数据(其他数量和数据集也可按此种方式进行)，则可先从 Waveform 中随机选择 10 对满足 must-link 关系的点对，即选择位于同一类别中的数据组成点对. 然后在 Waveform 中随机选择 10 对满足 can't-link 关系的点对，即选择不在同一类别中的数据组成点对，也就是说从某个类别中选择一个数据，然后再从另一个类别中选择一个数据，将这两个数据作为满足 can't-link 关系的点对.

同样如果为使选择的限制数据中包含不一致性，可以将上述某些按照传递特性推导出的满足 must-link 关系的点对设定为满足 can't-link 关系. 比如，如果选择(a,b)和(a,c)为满足 must-link 关系的点对，则设定(a,c)为满足 can't-link 关系的点对，此时，限制数据中即包含不一致性.

实验中以无监督聚类算法 K-means 和半监督聚类算法 COP-K-means 和 KCL 作为基准的聚类算法，然后将由不同的权值量化函数得到的特征权值应用于式(2)中以得到不同的相似度计算方法，之后比较了使用不同的相似度计算方法后各聚类算法的性能以表明本文所提的特征权值量化函数的性能.

本文同时检测了在不同情况下，本权值量化函数和其他不同的权值量化函数在各基准聚类算法上的表现情况以验证本权值量化函数的性能.

表 1　数据集概述

数据集	数据数	类别数	属性数
Waveform	5000	3	21
Pima	768	2	8
Newsgroup	20000	20	62264
863	3800	38	53470

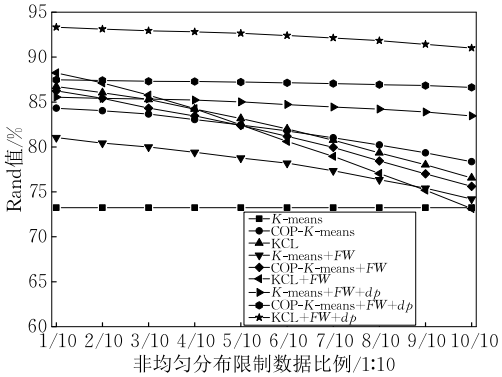
本文以纯度值[18]来衡量聚类结果的准确率. 在测试数据集中，已将待聚类数据划分到多个类别之中，当采用聚类算法进行处理时，仍然会将数据划分到多个类别中去. 这样可以使用 n^q 代表人工划分的第 q 个类别内包含的数据数，使用 n_r 代表聚类结果中的第 r 个类别内包含的数据数，使用 $n_r^q = n^q \bigcap n_r$ 代表聚类结果中的第 r 个类别内属于人工划分的第 q 个类别的数据的数目. 式(15)为第 r 个类别 C_r 的纯度值，其为具有最大数目的 n_r^q 与第 r 个类别的数据数 n_r 的比值. 式(16)为算法的纯度值，其为所有类别的纯度值的一个平均值，其中 c 代表类别的个数.

$$P(C_r) = \frac{1}{n_r} \max_{q=1}^{c}(n_r^q) \tag{15}$$

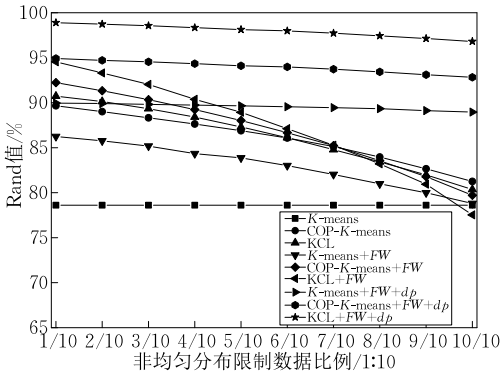
$$Purity = \sum_{r=1}^{c} \frac{c}{n_r} \max_{q=1}^{c}(n_r^q) \tag{16}$$

首先假设用户提供的限制数据是一致的，并且限制数据是从数据空间中非均匀选择的，在以上设定下，本文在图 1 中比较了 K-means、COP-K-means、KCL 在各基准数据集上的聚类准确率. 图 1 的横坐标代表聚类算法的准确率，纵坐标代表非均匀分布的限制数据占全部限制数据的比例①.

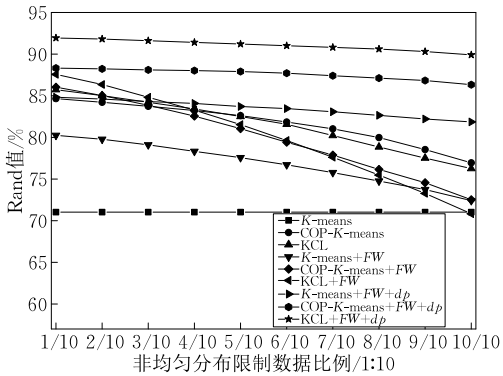
① 本文进行的所有实验中，均人工预先指定 100 对数据作为限制数据.



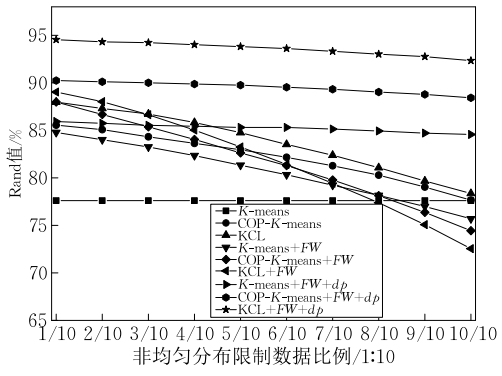
(a) Waveform



(b) Pima



(c) Newsgroup



(d) 863

图 1 当限制数据非均匀分布且一致时的聚类准确率

图中 FW 代表传统的权值量化函数(式(8)),$FW+dp$ 代表使用分布系数的权值量化函数(式(11)). K-means+FW 和 K-means+$FW+dp$ 代表使用基于特征权值的相似度计算方法的 K-means 算法,其中特征的权值分别由传统的权值量化函数和结合分布系数的权值量化函数得到.

由图 1 可知,假设用户提供的限制数据是一致的,则 K-means 的效果最差,这是因为 K-means 不结合限制数据并且视所有特征具有相同的类别划分能力,而不同特征对数据的划分能力显然是不同的,例如 Waveform 数据集中仅有 4 个特征是与数据分布相关的,其他特征均为噪声特征. 如果视所有特征具有相同的重要性,这些噪声特征显然会降低聚类算法对数据的划分能力. 相对于 K-means,COP-K-means 和 KCL 通过改变算法的收敛函数或调整搜索策略能够有效的保证数据的分割结果在满足用户指定的限制信息的前提下达到最优的划分[9-10],因此效果较好. 当非均匀分布的限制数据的数目较少时,KCL 的效果要好于 COP-K-means,然而随着非均匀分布的限制数据数目的增多,两种半监督聚类算法的性能逐渐下降,此时 KCL 下降的幅度要快于 COP-K-means. 性能下降的原因在于当非均匀分布的限制数据较少时,限制数据提供的有用信息能够帮助这两种算法改善聚类结果,然而当非均匀分布的限制数据增多时,非均匀分布引入的偏置会降低聚类的效果,同时由于 KCL 仅通过限制数据调整搜索策略,而 COP-K-means 同时结合了限制数据和非限制数据调整收敛函数,因此 KCL 受到的影响较大,性能波动幅度也较大.

由图 1 还可知,如果将传统的权值量化函数和 K-means 相结合后,其聚类效果有显著地提高. 这说明此函数在通过限制数据进行特征权值量化时,使那些能够有效划分限制数据的特征拥有较大的权值,从而使聚类结果能够满足用户提供的限制信息,同时该函数还结合了非限制数据以防止限制数据过少而引入的偏置,从而提高了算法对非限制数据的划分能力. 然而随着非均匀分布的限制数据数目的提升,所有结合传统的权值量化函数的聚类算法的性能均下降,并且此时结合此函数的半监督聚类算法的性能要低于不使用此函数的聚类算法的性能. 其原因在于,此权值量化函数结合了用户提供的限制数据来调整特征权值,因此会受到非均匀分布的限制数据的影响,从而出现性能波动. 同时还可见,相对于无监督聚类算法,结合传统的权值量化函数的半监督聚类算法受到的影响最大. 其原因在于,半监督聚类算法已经结合了限制数据进行类别划分,当与传统的权值量化函数相结合后,其受非均匀分

布的限制数据的影响显然也更大.

然而，当将聚类算法和使用分布系数的权值量化函数相结合后即可发现，无论是无监督聚类算法还是半监督聚类算法，其性能均保持了很好的平衡，受非均匀分布的限制数据的影响较小，并且各算法均维持了很高的准确率. 此种情况即可说明：分布系数能够平衡分布不均匀的限制数据对权值量化结果的影响.

如文中第 4 节所述，限制数据中很可能包含不一致性，因此本文在图 2 中比较了当限制数据中包含不一致性的前提下，使用不同的相似度计算方法后，各种聚类算法的准确率，此时设定限制数据为均匀分布的. 其中图 2 的横坐标代表聚类算法的准确率，纵坐标代表不一致的限制数据占全部限制数据的比例. 图中 FW 代表传统的权值量化函数(式(8))，$FW+bv$ 代表使用置信度的权值量化函数(式(14)中去掉参数 λ_b). K-means$+FW$ 和 K-means$+FW+bv$ 代表使用基于特征权值的相似度计算方法的 K-means 算法，其中特征的权值分别由传统的权值量化函数和结合置信度的权值量化函数得到.

对比图 2 和图 1 可知，图 2 中各种聚类算法的准确率曲线的分布和图 1 相差不多，其区别主要有：

(1) 与图 1 相比，除 K-means 外，图 2 中所有算法的准确率曲线均低于图 1 中的相应曲线.

为清晰展示图 2 和图 1 中各曲线的差距情况，本文将图 2 和图 1 中曲线的对比结果展示为表 2.

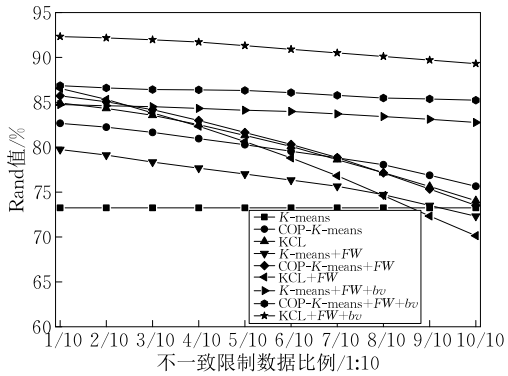
表 2 中，分别选择比例 1/10、5/10、10/10 来展示各聚类算法的效果. 由表中可以清晰的反映出：除 K-means 外，图 2 中所有算法的准确率曲线均低于图 1 中的曲线.

(2) 结合传统权值量化函数的半监督聚类算法的准确率曲线下降的幅度要大于图 1 中相应算法.

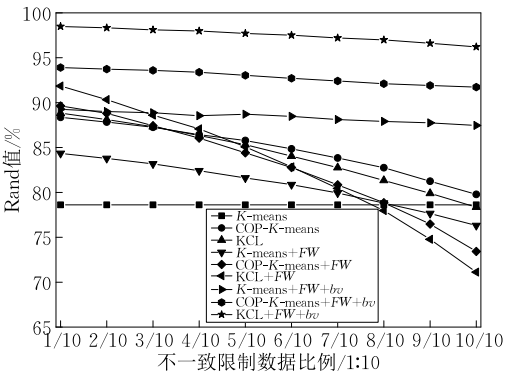
造成第 1 点的原因在于：相对于非均匀分布，限制数据包含不一致性时引入的误差要更大，因此除 K-means 外，各聚类算法的性能均下降. K-means 算法保持不变的原因在于该算法根本未考虑限制数据的影响，因此无论限制数据是否是不均匀分布还是包含不一致性时，该算法的性能均不受影响.

造成第 2 点的原因在于：由于半监督聚类算法已经结合了限制数据进行类别划分，因此当限制数据包含不一致性时，半监督聚类算法与传统的权值量化函数相结合后，其性能会降低的更多.

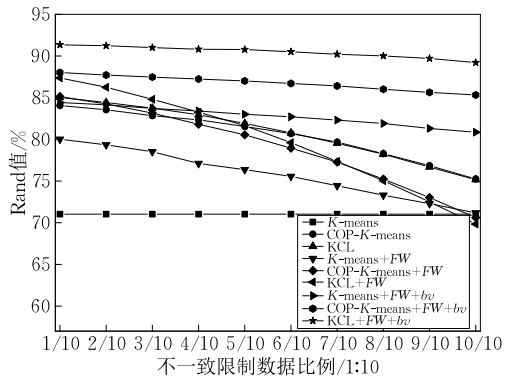
然而，当将所有的聚类算法和使用置信度的权值量化函数相结合后，各聚类算法的性能均有所提升，并且受不一致的限制数据的影响非常小. 此种情况即可说明：本文提出的置信度能够有效改善限制数据的不一致性对权值量化结果的影响，进而提高聚类算法的准确率.



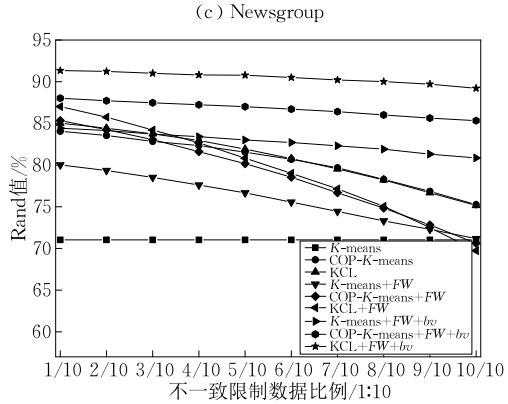
(a) Waveform



(b) Pima



(c) Newsgroup



(d) 863

图 2　当限制数据包含不一致且均匀分布时的聚类准确率

表 2 图表对比结果

图 1				图 2			
Waveform	1/10	5/10	10/10	1/10	5/10	10/10	Waveform
K-means	73.25	73.25	73.25	73.25	73.25	73.25	*K*-means
COP-*K*-means	84.32	82.45	78.35	82.66	80.28	75.65	COP-*K*-means
KCL	86.73	83.16	76.56	84.93	81.31	74.05	KCL
K-means+*FW*	81.05	78.77	74.21	79.75	77.01	72.31	*K*-means+*FW*
COP-*K*-means+*FW*	86.23	82.39	75.62	85.73	81.64	73.52	COP-*K*-means+*FW*
KCL+*FW*	88.25	82.46	73.14	86.55	80.63	70.14	KCL+*FW*
K-means+*FW*+*dp*	85.55	85.01	83.46	84.75	84.12	82.76	*K*-means+*FW*+*bv*
COP-*K*-means+*FW*+*dp*	87.46	87.23	86.63	86.86	86.32	85.23	COP-*K*-means+*FW*+*bv*
KCL+*FW*+*dp*	93.33	92.65	91.01	92.33	91.31	89.31	KCL+*FW*+*bv*
Pima	1/10	5/10	10/10	1/10	5/10	10/10	Pima
K-means	78.61	78.61	78.61	78.61	78.61	78.61	*K*-means
COP-*K*-means	89.66	86.89	81.25	88.36	85.78	79.78	COP-*K*-means
KCL	90.73	87.31	80.36	88.83	85.32	78.36	KCL
K-means+*FW*	86.25	83.87	78.81	84.35	81.61	76.31	*K*-means+*FW*
COP-*K*-means+*FW*	92.23	88.03	79.72	89.63	84.41	73.42	COP-*K*-means+*FW*
KCL+*FW*	94.55	88.94	77.54	91.85	85.07	71.12	KCL+*FW*
K-means+*FW*+*dp*	89.96	89.65	88.96	89.26	88.71	87.46	*K*-means+*FW*+*bv*
COP-*K*-means+*FW*+*dp*	94.91	94.11	92.83	93.91	93.04	91.73	COP-*K*-means+*FW*+*bv*
KCL+*FW*+*dp*	98.89	98.13	96.81	98.49	97.71	96.21	KCL+*FW*+*bv*
Newsgroup	1/10	5/10	10/10	1/10	5/10	10/10	Newsgroup
K-means	71.03	71.03	71.03	71.03	71.03	71.03	*K*-means
COP-*K*-means	84.66	82.62	76.95	84.06	81.55	75.25	COP-*K*-means
KCL	85.73	82.51	76.26	85.03	81.92	75.16	KCL
K-means+*FW*	80.25	77.57	72.41	80.02	76.67	71.18	*K*-means+*FW*
COP-*K*-means+*FW*	86.03	81.05	72.52	85.34	80.17	70.62	COP-*K*-means+*FW*
KCL+*FW*	87.57	81.54	70.84	87.02	80.83	69.74	KCL+*FW*
K-means+*FW*+*dp*	84.82	83.71	81.86	84.42	83.01	80.86	*K*-means+*FW*+*bv*
COP-*K*-means+*FW*+*dp*	88.32	87.91	86.33	88.02	87.01	85.33	COP-*K*-means+*FW*+*bv*
KCL+*FW*+*dp*	91.94	91.21	89.91	91.34	90.78	89.21	KCL+*FW*+*bv*
863	1/10	5/10	10/10	1/10	5/10	10/10	863
K-means	77.61	77.61	77.61	77.61	77.61	77.61	*K*-means
COP-*K*-means	85.55	82.96	77.65	85.12	82.17	76.45	COP-*K*-means
KCL	87.97	84.76	78.36	87.32	83.64	76.56	KCL
K-means+*FW*	84.78	81.33	75.69	84.05	80.38	74.11	*K*-means+*FW*
COP-*K*-means+*FW*	87.98	82.62	74.44	87.15	81.28	73.23	COP-*K*-means+*FW*
KCL+*FW*	89.03	83.25	72.55	88.33	81.92	70.75	KCL+*FW*
K-means+*FW*+*dp*	85.93	85.41	84.56	85.13	84.64	83.46	*K*-means+*FW*+*bv*
COP-*K*-means+*FW*+*dp*	90.24	89.76	88.43	90.04	89.27	87.93	COP-*K*-means+*FW*+*bv*
KCL+*FW*+*dp*	94.54	93.81	92.33	93.34	92.61	91.33	KCL+*FW*+*bv*

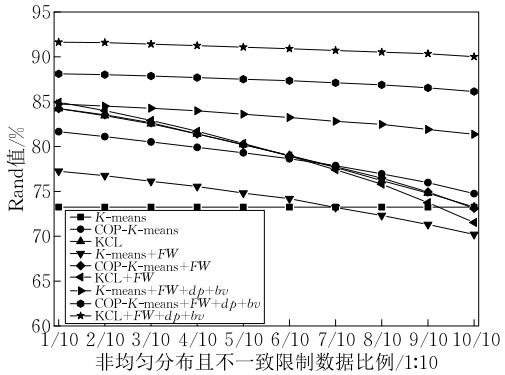
接下来在图 3 中比较了在限制数据中既存在非均匀分布又存在不一致性的情况下，使用不同的相似度计算方法后，各种聚类算法的准确率. 图 3 的横坐标代表聚类算法的准确率，纵坐标代表既是非均匀分布且是不一致的限制数据占全部限制数据的比例. 图中 FW 代表传统的权值量化函数(式(8))，$FW+dp+bv$ 代表使用分布系数和置信度的权值量化函数(式(14)). *K*-means+FW 和 *K*-means+$FW+dp+bv$ 代表使用基于特征权值的相似度计算方法的 *K*-means 算法，特征权值分别由传统的权值量化函数和结合分布系数和置信度的权值量化函数得到.

相比于图 1 和图 2，由图 3 可见，当限制数据同时存在非均匀分布和不一致性时，传统的半监督聚类算法和使用传统的特征权值量化函数的各种聚类算法的性能均急剧下降，并且使用传统的特征权值量化函数的半监督聚类算法的性能要远远低于不使用此函数的半监督聚类算法. 此种情况说明：非均匀分布和不一致性的同时存在使得依赖限制数据的聚类算法的性能受到极大的影响，其受影响程度远远大于上述两种情况单独存在时引入的影响. 然而，当将分布系数和置信度同时引入权值量化函数后，各种聚类算法的性能均有所提高，且随着非均匀分布和不一致的限制数据数目的增多，其性能下降的幅度减小. 此种情况说明分布系数和置信度可以同时存在来提高聚类算法的性能.

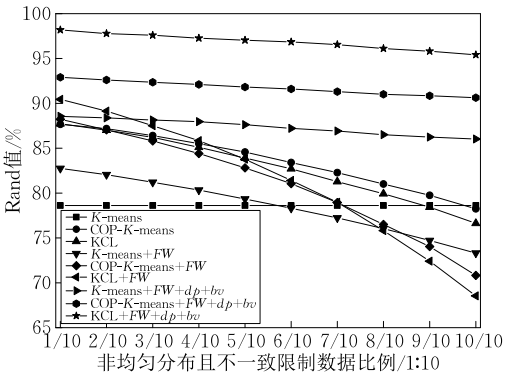
为显示本文提出的权值量化函数的性能，本文同时检测了在不同情况下，此权值量化函数和其他权值量化函数在各基准聚类算法上的表现情况. 作为比较的权值量化函数包括 $FWNB$[19]、$MWMR$[20]、$FWSA$[21]. 以 *K*-means 为例，将应用各权值量化函数的 *K*-means 算法标记为 *K*-means+$FWNB$、*K*-means+$MWMR$、*K*-means+$FWSA$，同时将应用“分布系数”和“置信度”的权值量化函数的 *K*-means 算法标记为 *K*-means+$FW+dp+bv$.

图 4 显示了当限制数据不存在不一致性，并且限制数据是从数据空间中非均匀选择的情况下，各权值量化函数和各聚类算法相结合后在各基准数据集上的聚类准确率.

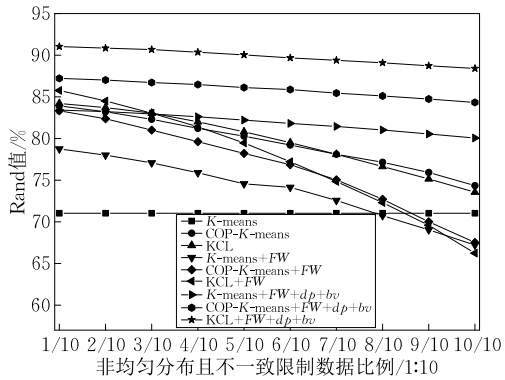
由图 4 可见，当限制数据为非均匀分布时，与未使用权值量化函数的聚类算法相比，除 $FWNB$ 和本文提出的权值量化函数外，$MWMR$ 和 $FWSA$ 与聚类算法相结合后，各聚类算法的性能均降低. 其意味着，相对于 $MWMR$ 和 $FWSA$，$FWNB$ 可以很好的适应非均匀分布的限制数据对权值量化结果的影



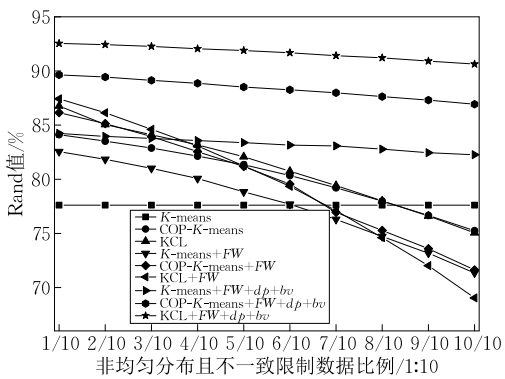
(a) Waveform



(b) Pima



(c) Newsgroup



(d) 863

图 3　当限制数据非均匀分布且不一致时的聚类准确率

响，其原因在于 $FWNB$ 使用数据的分布密度来融合限制数据进行特征权值量化，因此可以在一定程度上适应非均匀分布的限制数据对权值量化结果的影响. 然而，由于 $FWNB$ 需要预先设定数据的分布模型，这使得假设的分布模型与真实的数据分布存在一定的差距，从而降低了权值量化函数的准确率. 相应的，$MWMR$ 和 $FWSA$ 在进行权值量化前已经假设限制数据为均匀分布，因此权值量化的结果无法处理非均匀分布的限制数据. 相对于 $FWNB$、$MWMR$、$FWSA$，本文提出的权值量化函数($FW+dp+bv$)通过等价类划分可以很好的考虑非均匀分布的限制数据对权值量化结果的影响，因此当将本文提出的权值量化函数与各聚类算法相结合后，各聚类算法的性能均得到很大的提升.

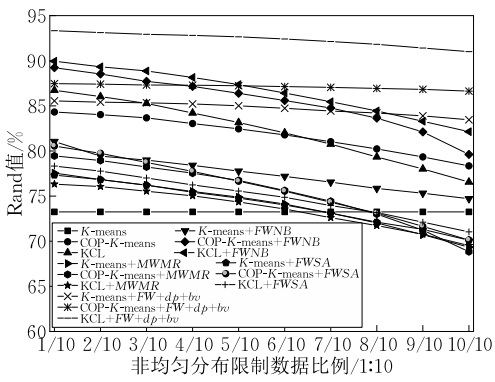
图 5 显示了当限制数据均匀分布时，且限制数据包含不一致性的情况下，各权值量化函数和各聚类算法相结合后在各基准数据集上的聚类准确率.

由图 5 可见，当限制数据包含不一致性时，与未使用权值量化函数的聚类算法相比，当使用 $FWNB$、$MWMR$、$FWSA$ 后，各聚类算法的性能均有所降低，其意味着上述的权值量化函数均无法对包含不一致性的限制数据进行处理. 相对来说，当使用本文提出的权值量化函数后，各聚类算法的性能均有所提升，并且随着不一致的限制数据数目的增多，各聚类算法的性能并没有下降得太多，这说明，相对于基准的权值量化函数($FWNB$、$MWMR$、$FWSA$)，本文提出的权值量化函数可以很好的处理限制数据中包含不一致性的情况.

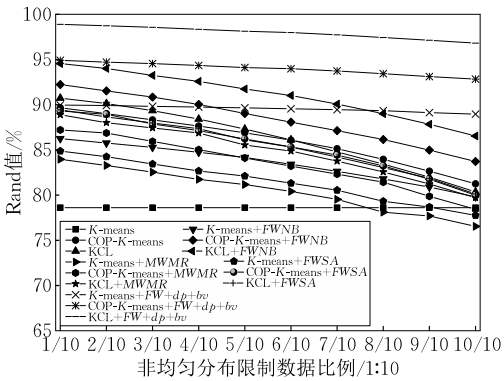
图 6 显示了当限制数据非均匀分布且包含不一致性的情况下，各权值量化函数和各聚类算法相结合后在各基准数据集上的聚类准确率.

由图 6 可见，当限制数据非均匀分布且包含不一致性的情况下，与未应用权值量化函数的聚类算法相比，将 $FWNB$、$MWMR$、$FWSA$ 和各聚类算法相结合后，各聚类算法的性能均下降. 与图 4 相反，与图 5 类似，当将聚类算法和 $FWNB$ 相结合后，各聚类算法的性能均低于未使用权值量化函数的聚类算法的性能，其意味着，非均匀分布且包含不一致性的两种情况会相互影响以更进一步的降低各基准的权值量化函数的性能. 相反，如果使用本文提出的权值量化函数，可以很好的解决这两点问题，使得各聚类算法的性能均有所提高.

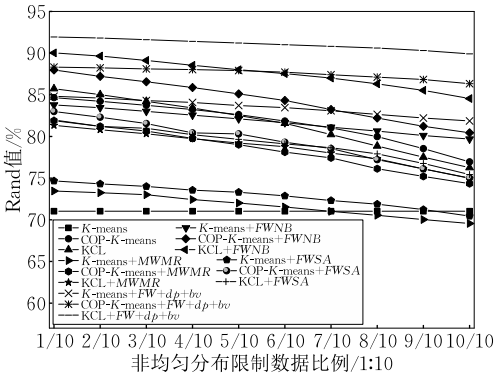
本文通过划分等价类来形成“分布系数”以改善限制数据的分布不平衡对权值量化结果带来的影响. 然而，如果按照传统的划分方式，即文中第 3 节所示的步骤来划分等价类，会忽略掉很多可能的等价类. 因此，本文引入式(10)来改善此种情况. 为显



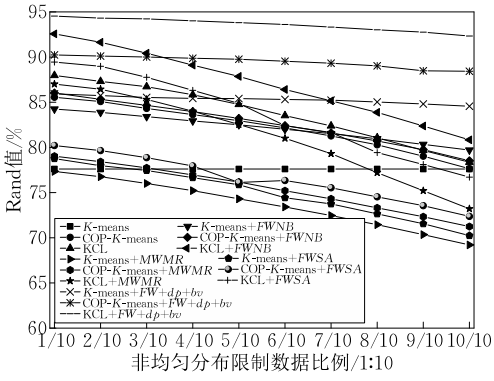
(a) Waveform



(b) Pima

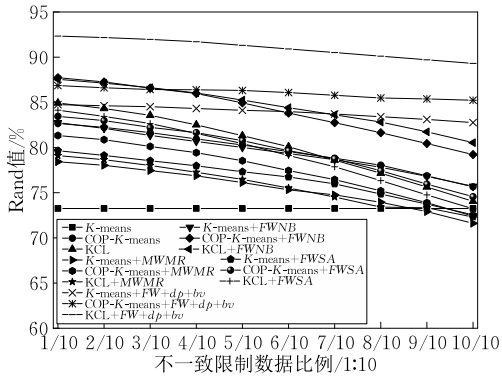


(c) Newsgroup

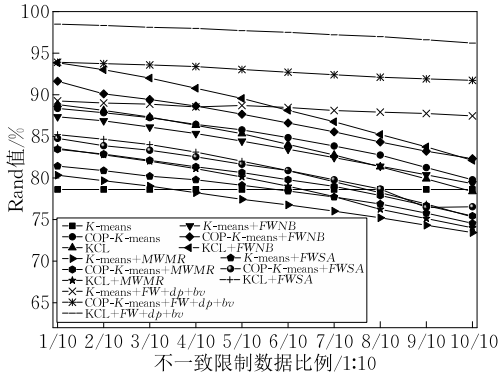


(d) 863

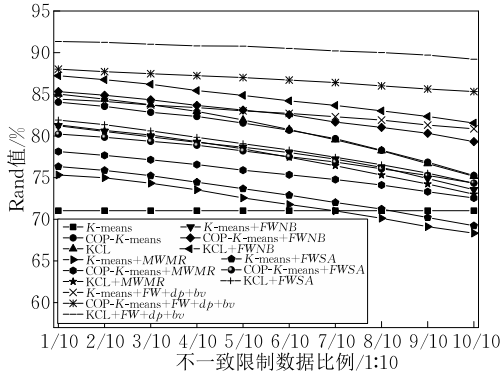
图 4 当限制数据非均匀分布且一致时应用各权值量化函数对应的聚类准确率



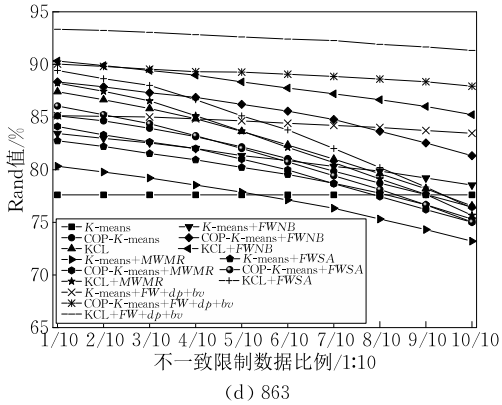
(a) Waveform



(b) Pima



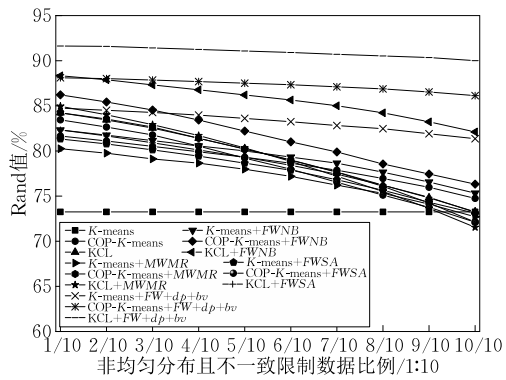
(c) Newsgroup



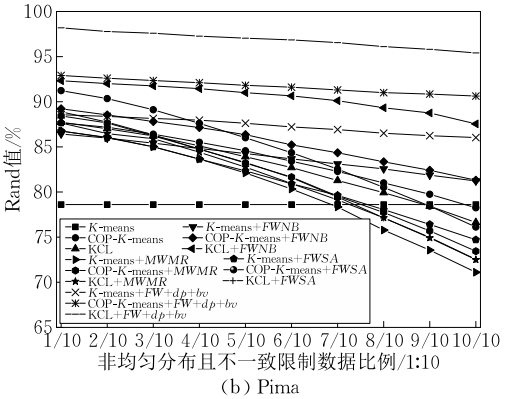
(d) 863

图 5 当限制数据包含不一致性且均匀分布时应用各权值量化函数对应的聚类准确率

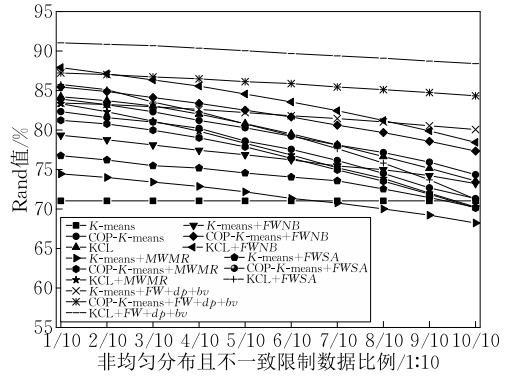
示该方法的效果,本文检测了在未使用式(10)和使用式(10)后的权值量化函数对非均匀分布的限制数据的处理能力. 图 7 中,$FW+dp1$ 和 $FW+dp2$ 均代表使用“分布系数”的权值量化函数,区别为置信度分别通过常规方法获得(文中第 3 节所示的划分步骤)和通过式(10)重新划分获得.



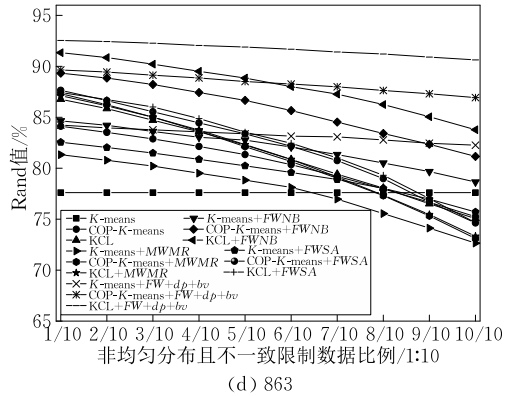
(a) Waveform



(b) Pima

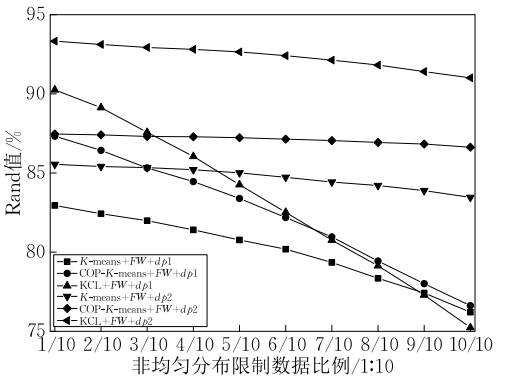


(c) Newsgroup

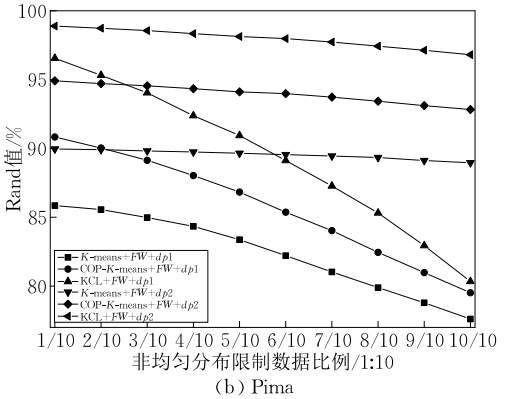


(d) 863

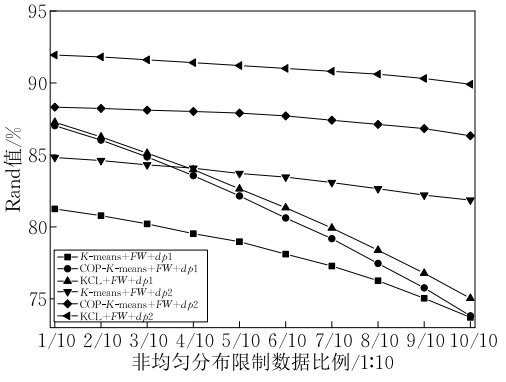
图 6　当限制数据非均匀分布且包含不一致性时应用各权值量化函数对应的聚类准确率



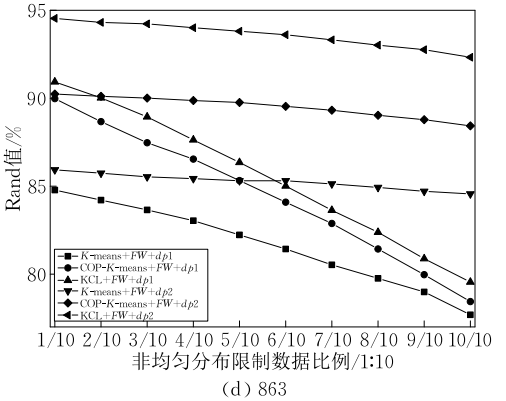
(a) Waveform



(b) Pima



(c) Newsgroup



(d) 863

图 7　当限制数据非均匀分布且一致时应用不同等价类划分方法对应的聚类准确率

由图 7 可见，当采用常规方法划分等价类得到参数“分布系数”后，使用结合此“分布系数”的权值量化函数($FW+dp1$)的各聚类算法的性能显然低于使用结合式(10)的权值量化函数($FW+dp2$)的各聚类算法的性能. 其意味着如果简单的根据 must-link 关系和 can't-link 关系划分等价类，显然

无法将限制数据的分布特性完全挖掘出来.例如,如果用户指定同一类别中的点对(a,b)和(c,d)均满足must-link关系,由于数据a、b、c、d位于同一类别中,因此点对(a,c)和(b,d)也满足must-link关系,但是按照传统的等价类划分方法会将(a,b)和(c,d)划分到不同的等价类中,即无法发现(a,c)和(b,d)之间的关系.而当采用式(10)后,通过计算不同等价类中的数据之间的相似性,可以很好的对等价类进行重新划分,即将点对(a,b)和(c,d)划分到同一等价类中.这说明,这种结合式(10)的权值量化函数可以更好的描述数据的分布特性,从而避免了不均匀分布的限制数据对权值量化结果带来的影响.

6 结 论

本文提出了一种结合限制数据的特征权值量化函数,该函数通过用户指定的限制数据进行特征权值量化,使得能够充分反映数据之间的相似性或能够有效划分数据的特征在相似度计算中具有较大的权值,并将相似度计算的结果应用于数据聚类之中.为了防止限制数据分布不均匀而形成训练结果的"过适应",此函数将限制数据划分为多个等价类,并为不同的等价类赋予不同的分布系数.针对用户指定的限制数据中可能包含不一致性的问题,此函数对不同的限制数据赋予不同的置信度并将其结合到权值量化函数中去.由实验结果可得,此权值量化函数独立于固定的聚类算法,可以应用于任何聚类算法中以改变相似度度量函数并提高聚类结果的准确率.

参 考 文 献

[1] Jain A K, Murty M N, Flynn P J. Data clustering: A review. ACM Computing Surveys, 1999, 31(3): 264-323

[2] Selim M, Erunc E. Combining multiple clusterings using similarity graph. Pattern Recognition, 2011, 44(3): 694-703

[3] Kweku M, Osei B. Towards supporting expert evaluation of clustering results using a data mining process model. Information Sciences, 2010, 180(3): 414-431

[4] Tata S, Patel J M. Estimating the selectivity of TF-IDF based cosine similarity predicates. ACM SIGMOD Record, 2007: 7-12

[5] Arias C E. Clustering based on pairwise distances when the data is of mixed dimensions. IEEE Transactions on Information Theory, 2011, 57(3): 1692-1706

[6] Wagstaff K, Cardie C. Clustering with instance-level constraints//Proceedings of the 17th International Conference on Machine Learning. Stanford, USA, 2000: 1103-1110

[7] Sugato B, Arindam B, Mooney R. Semi-supervised clustering by seeding//Proceedings of the 19th International Conference on Machine Learning. Sydney, Australia, 2002: 19-26

[8] Cheng Ying-Mei, Leu Sousen. Constraint-based clustering and its applications in construction management. Expert Systems with Applications, 2009, 36(3): 5761-5767

[9] Hu Guo-Biao, Zhou Shui-Geng, Guan Ji-Hong, et al. Towards effective document clustering: A constrained K-means based approach. Information Processing and Management, 2008, 44(4): 1397-1409

[10] He Zhen-Feng, Xiong Fan-Lun. A constrained partition model and K-Means algorithm. Journal of Software, 2005, 16(5): 799-809(in Chinese)
(何振峰,熊范纶. 结合限制的分隔模型及 K-Means 算法. 软件学报, 2005, 16(5): 799-809)

[11] Sugato B, Bilenko M, Mooney R J. Comparing and unifying search-based and similarity-based approaches to semi-supervised clustering//Proceedings of the ICML-2003 Workshop on the Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining Systems. Washington, USA, 2003: 42-49

[12] Peng Han-Chuan, Long Fu-Hui, Chris Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(8): 1226-1238

[13] Wang Xi-Zhao, Wang Ya-Dong, Wang Li-Juan. Improving fuzzy c-means clustering based on feature-weight learning. Pattern Recognition Letters, 2004, 25(10): 1123-1132

[14] Klein D, Kamvar S D, Manning C D. From instance-level constraints to space-level constraints making the most of prior knowledge in data clustering//Proceedings of the 19th International Conference on Machine Learning. Sydney, Australia, 2002: 307-314

[15] Chen Hai-Bo, Lv Xian-Qing, Qiao Yan-Song. Application of gradient descent method to the sedimentary grain-size distribution fitting. Journal of Computational and Applied Mathematics, 2009, 233(4): 1128-1138

[16] Wang Li-Juan, Guan Shou-Yi, Wang Xiao-Long, Wang Xi-Zhao. Fuzzy C Mean algorithm based on feature weights. Chinese Journal of Computers, 2006, 29(10): 1797-1803(in Chinese)
(王丽娟,关守义,王晓龙,王熙照. 基于属性权重的 Fuzzy C Mean 算法. 计算机学报, 2006, 29(10): 1797-1803)

[17] Zhou Zhi-Hua, Liu Xu-Ying. Training cost-sensitive neural networks with methods addressing the class imbalance problem. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(1): 63-77

[18] Xiong Xue-Jian, Chan Kap Luk, Tan Kian Lee. Similarity-driven cluster merging method for unsupervised fuzzy clustering// Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence. Banff, Canada, 2004: 611-618

[19] Chen Li-Fei, Wang Sheng-Rui. Automated feature weighting in naive bayes for high-dimensional data classification// Proceedings of the 21st ACM International Conference on Information and Knowledge Management. Sheraton, Mau Hawaii, USA, 2012: 1243-1252

[20] Wang Jian-Zhong, Wu Li-Shan, Kong Jun, et al. Maximum weight and minimum redundancy: A novel framework for feature subset selection. Pattern Recognition, 2013, 46(6): 1616-1627

[21] Tsai Chieh-Yuan, Chiu Chuang-Cheng. Developing a feature weight self-adjustment mechanism for a *K*-means clustering algorithm. Computational Statistics and Data Analysis, 2008, 52(10,15): 4658-4672

LIU Ming, born in 1981, Ph. D., lecturer. His research interests include text clustering and text mining.

WU Chong, born in 1971, Ph. D., professor. His research interests include data management and information processing.

LIU Yuan-Chao, born in 1971, Ph. D., associate professor. His research interests focus on natural language processing.

SUN Cheng-Jie, born in 1980, Ph. D., lecturer. His research interests focus on information recommendation.

Background

The research in this paper focuses on weight evaluation for features. This research topic is prevalent in research domain. Due to the significant growth of internet technology, automatic analysis tools become essential for web users to help them exploit useful information from large-scale web data. Apparently, accurate similarity measurement plays a decisive role for analysis tools to acquire high performance. Based on the fact that, different features contribute diversely to similarity calculation, it is necessary to utilize weight to measure feature's contribution and import it in similarity measurement. To accurately assign feature's weight, constrained data provided by users are usually imported in weight evaluation, whereas, conventional plans all fail to consider two challenges: (1) asymmetrical distribution of constrained data, (2) inconsistency contained by constrained data. When they happen, conventional weight evaluation methods are incompetent or even unable to work. Hence, this paper proposes a novel constraint based weight evaluation to deal with these two issues.

The research in this paper belongs to National Natural Science Foundation of China (No. 61300114). Its full name is "The Research on Fast Clustering and Information Evolution Analysis for Large-Scale and Dynamic Short-Texts". This fund aims at exploiting useful information concerned by users from large-scale web short-texts and comprehending information evolutional trend reflected by those web data. This fund intends to propose a fast and dynamic clustering algorithm for large-scale short-texts, which is hope to be applied to acquire information evolutional trend to reflect the transfer of user's attention. It's straightforward, to acquire high-quality clustering performance, the precise similarity measurement is most important. Unfortunately, for web data, since they are edited by users, they may include many noisy features. Those features certainly spoil similarity results. Thus, it's useful to propose a method to measure feature's weight and finally to import it in similarity measurement. Based on the research in this paper, feature's weight can be accurately evaluated via constrained data. Moreover, two problems of "asymmetrical distribution of constrained data" and "inconsistency contained by constrained data" often appearing in traditional weight evaluation methods can be tackled by this research. Therefore, it's obvious that the method proposed in this paper can describe similarities more exactly, and consequently improves clustering performance. Typically, this weight evaluation method is independent on any algorithm, and can be applied in any algorithm.