基于 seeds 集和频繁项集挖掘的半监督聚类算法

赵 倩,尚学群,王 淼
ZHAO Qian,SHANG Xue-qun,WANG Miao

西北工业大学 计算机学院,西安 710072
School of Computer,Northwestern Polytechnical University,Xi'an 710072,China
E-mail:zhaoqian_qiezi@126.com

ZHAO Qian,SHANG Xue-qun,WANG Miao.Semi_supervised clustering algorithm based on seeds set and frequent itemset mining.Computer Engineering and Applications,2010,46(8):123-126.

Abstract: Semi_supervised clustering makes use of few supervised information in unsupervised clustering to boost the clustering performance.This paper proposes a semi_supervised clustering algorithm based on seeds set and frequent itemset mining,which mines frequent itemsets in the beginning seeds set and the enlarged seeds set for eliminating the noise data and correcting the mislabeled data to improve the quality of seeds set and enhance the performance of clustering.A weighted χ^2 measure,as a classification rule evaluation measure,is used to label unlabeled data and they are added into the initial seeds set to enlarge the scale.The experimental results show that the proposed approach effectively reduces the noise data,and not only makes the results more correct but also makes the performance of clustering more better.

Key words: semi_supervised clustering;frequent itemset mining;weighted χ^2 measure;seeds set

摘 要: 半监督聚类在无监督学习中通过对少量监督信息的有效利用提高聚类性能。提出一种基于 seeds 集的半监督聚类算法,它采用 Apiori 算法对初始 seeds 集和扩大规模后 seeds 集的数据进行频繁项集挖掘,使得数据中存在的噪音数据和误标记数据得到净化、修正,以改善 seeds 集质量,提高聚类性能。该算法使用带权 χ^2 测试这一数学模型作为分类规则度量指标,以对无标记数据进行类标签值预测。实验结果显示,所提出的结合了频繁项集挖掘和带权 χ^2 测试的基于 seeds 集的半监督聚类算法不仅改善了 seeds 集质量,也提高了预测结果的精确度,优化了聚类性能。

关键词: 半监督聚类;频繁项集挖掘;带权 χ^2 测试;seeds 集
DOI: 10.3778/j.issn.1002-8331.2010.08.035 **文章编号:** 1002-8331(2010)08-0123-04 **文献标识码:** A **中图分类号:** TP311

1 引言

监督学习需要大量带标签数据作为训练集以保证泛化能力。但在文本处理、生物信息学和网页分类等实际应用中,对数据进行人工标记的代价很高,容易获得的是大量无标记数据。无监督学习属于无任何监督信息的自动学习,虽然不需要带标记数据,但所得模型却不够精确。因此,将少量带标记数据和大量无标记数据结合的半监督学习成为机器学习的研究热点。按照学习任务的不同,半监督学习分为半监督聚类和半监督分类[1]。

无监督聚类通过优化反映聚类结果优良性的目标函数将无标签数据分成若干组,组内数据相似,组间数据相异。半监督聚类使用一些监督信息作为无监督聚类的辅助信息以提高聚类性能,这些监督信息可以是数据的类别标记也可以是一对数据是否属于同一类或不同类的约束关系。按照监督信息的不同,半监督聚类的方法可分为两种:基于距离的半监督聚类和基于约束条件的半监督聚类[2]。

Basu 等人在文献[3]中提出的 Generative 模型结合 EM 理论支持的 Seeded-*K*-均值和 Constrained-*K*-均值算法是一种基于约束条件的半监督聚类方法,这两种算法是基于 seeds 集的,它们使用少量带标记数据形成 seeds 集来改善 *K*-均值聚类的初始化效果,进而提高整个数据集的聚类性能。同时,Basu 等人也通过实验表明:这两种基于 seeds 集的半监督聚类算法对 seeds 集的规模和噪声都十分敏感,规模大、噪声小的 seeds 集将显著地提高聚类性能[4]。另外,在文献[4]中提出了一种在 seeds 集的基础上结合半监督分类方法 Tri-Training 和数据剪辑技术的半监督聚类算法,其实验结果在显示了所提出算法有效性的同时进一步说明了 seeds 集规模和质量对聚类性能提高的重要性。

基于 seeds 集的半监督聚类算法受 seeds 集规模和质量的影响显著,而在实际应用中得到大量带标记的数据需花费很大

基金项目: 陕西省自然科学基金(the Natural Science Foundation of Shaanxi Province of China under Grant No.2007F27);西北工业大学研究生创新实验室项目(No.07046)。

作者简介: 赵倩(1985-),硕士研究生,研究方向:数据管理技术;尚学群(1973-),工学博士,硕士研究生导师,主要研究方向:数据库技术,数据挖掘,生物信息学等;王淼(1981-),博士研究生,主要研究方向:数据挖掘,生物信息学等。

收稿日期: 2008-09-18 **修回日期:** 2008-12-04

的代价，鉴于此矛盾将频繁项集挖掘以及带权χ^2测试这一分类规则度量方法运用到半监督聚类中，提出基于 seeds 集和频繁项集挖掘的半监督聚类算法。频繁项集挖掘根据给定的阈值挖掘频繁出现在数据集中的属性集，去除不满足最小阈值的属性项，可达到消除噪音、改善 seeds 集质量的目的。带权χ^2测试同时考虑规则的相关度、支持度和置信度，将三者结合起来构造统一度量作为分类规则的评价指标，在用挖掘出的分类规则对无标签数据进行标签值预测时该算法使用这一方法度量一组分类规则的联合效果以提高预测精确度，为聚类提供低噪声、大规模的 seeds 集，进而提高聚类整体性能。

2 知识点介绍

2.1 半监督聚类

在文献[5]中给出聚类的定义如下：给定数据样本集$X\{X_1, X_2, \cdots, X_n\}$，根据数据点间的相似程度将数据集分成 k 簇：$\{C_1, C_2, \cdots, C_k\}$的过程称为聚类，$C_i=\{X_i\cdots\}$，$C_i \cap C_j=\phi$，$i \neq j$，相似样本在同一簇中，相异样本在不同簇中。当数据样本集中的数据所属类信息未知时可使用聚类方法对数据进行处理。因此聚类是无监督学习的一部分[6]。在文献[6]中有关于聚类分析核心概念及代表性聚类方法的介绍。

在实际应用中有时可以得到关于一些数据的监督信息如：数据间的连接约束信息或数据的类标签信息。除了简单地将这些已知信息用于评价聚类性能，也可将它们用于指导或调整聚类过程，辅助于单纯使用相似度信息的无监督聚类[6]。这种将少量含已知信息的数据和大量无标记数据结合的聚类方法即为半监督聚类。和无监督聚类不同的是半监督聚类的研究历史不长，到目前为止提出的半监督聚类算法还很少[6]。按照监督信息的使用方式不同，半监督聚类方法可分为基于约束条件的和基于距离的两种。基于约束条件的半监督聚类方法根据用户提供的类标签信息或约束信息指导聚类过程，最终得到一个更合适的数据划分，比如：根据约束信息调整目标函数，根据已知类标签信息初始化并约束整个聚类过程。而基于距离的半监督聚类方法首先训练得到满足监督性数据含有的已知类标签信息或约束信息的距离度量公式，再根据此度量公式进行聚类[2]。基于约束条件的方法目前研究应用较多，文献[3-4]提出的基于seeds集的不同半监督聚类算法都属于基于约束条件的方法。

2.2 基于关联规则的分类

分类主要是通过分析训练数据样本，产生关于类别的精确描述，即分类器，它代表了这类数据的整体信息。根据分类器可以对未知分类的数据进行预测。目前分类算法很多，例如决策树、规则学习、朴素贝叶斯网络以及统计方法等。由于关联规则由若干个属性组成、具有高置信度关联特性并且和分类结果有关，利用这些规则进行分类可以弥补决策树分类方法一次只能处理一个属性的弊端[7]。在所有训练样本的数据属性中，并不是所有属性都最终和分类结果有关，找出和分类结果相关的属性才能构造出有效的分类器。基于关联规则的分类所挖掘出的规则中的属性都和分类结果相关，因此可以剔除和分类无关的属性，从而提高分类效率。

在基于关联规则的分类方法中，如何找出最有效的规则对未知数据进行预测是关键问题[7]。在实际应用中有多种方法用于解决这个问题，文献[8]较全面地介绍了分类规则选择的各种方法。其中，带权χ^2测试综合考虑分类规则的相关度、支持度、置信度度量一组关联规则的联合效果。文献[7-8]都通过实验说明了带权χ^2测试是一种相对于大多数其他度量公式来说可得到更精确预测结果的分类规则度量方法。关于带权χ^2测试文献[8]给出了详细定义和公式描述。

2.3 Seeding

在文献[3]中，Basu 等人给出 Seeding 的定义：给定数据集 x，对于 x 的子集合 S，当 S 中的所有数据含有满足以下条件的监督信息时称 S 为 seeds 集，这个必须满足的条件为：由于聚类将整个数据集划分为若干个 cluster，对于任意 $x_i \in S$，它一定包含自己属于某一 cluster 的监督信息。把加入了 seeds 集划分过程的聚类称为 seed clustering。

Basu 等人在文献[3]中提出的 Seeded-K-均值和 Constrained-K-均值算法用少量带标记的数据作为初始 seeds 集，并按标记将 seeds 集划分为 k 个聚类，由此计算 k 个聚类初始中心[4]。两种算法都是基于 seeds 集进行聚类的，不同之处在于：Constrained-K-均值算法保持 seeds 集数据在整个聚类过程中标签值不变，只有非 seed 数据被重新估值；而在 Seeded-K-均值算法执行过程中不仅非 seed 数据被重新估值，seed 数据即已标记数据的类标签值也会被重新估值。

seeds 集中的带标记数据会提供关于隐藏类标签值条件分布的先验信息，并且这些少量的带标记数据辅助于并不断调整无标记数据的聚类过程，文献[3]通过数学方法证明了将 seeds 集应用于 K-Means 算法的半监督聚类的有效性。文献[3-4]提出了采用不同方法的基于 seeds 集的半监督聚类算法，都通过实验说明了 seeds 集规模和质量对整个算法聚类性能的影响，如果能在增大 seeds 集规模的同时提高 seeds 集质量，算法性能的提高将十分显著。

3 基于 seeds 集和频繁项集挖掘的半监督聚类算法

针对完全是无监督的学习过程的聚类，利用现实中易得的含已知类标签信息的数据，可构成半监督聚类。在实际应用中，初始带标签数据集的规模很小，可能不足以代表规模庞大的无标签数据集的完整聚类结构，此时采用半监督聚类优于半监督分类。基于 seeds 集的半监督聚类算法在用标记后的无标签数据扩大 seeds 集规模时可能会引入大量误标记噪声数据，进而影响整个聚类算法的性能，而且虽然初始带标签数据集的规模很小，也会存在噪声数据。为了消除、净化这些噪声数据和误标记数据，进而提高聚类结果精度，优化聚类性能，文中使用 Apriori 算法对 seeds 集数据进行频繁项集挖掘处理，也为以后对未标记数据的类标签值预测产生分类规则，并保留全部满足最小支持度的关联规则。

使用挖掘出的关联规则对未标记数据进行类标签值预测时，若与未标记数据匹配的所有规则含有同样的类标签值，可直接将此类标签值赋给未标记数据；若匹配规则的类标签值不相同，则将所有匹配的规则按类标签值分为若干组，同组分类规则含有相同的类标签值，不同组分类规则含有不同的类标签值。为了减少误标记数据的产生，算法使用带权χ^2测试这一数学模型度量一组关联规则的联合效果，带权χ^2值最大的那组关联规则对应的类标签值即为未知数据的类标签预测值。将标记后的数据放入 seeds 集以扩大其规模，增大聚类所需的监督信息。

以上过程迭代进行，直到满足目标函数或聚类总迭代次数

达到预定允许的最大值。

算法总流程：

输入：已标记数据集 *LabeledDataSet*，未标记数据集 *UnlabeledDataSet*，类标签值个数 *LabNum*，Apriori 算法的最小支持度 *Support*，目标函数类标签预测值变化率阈值 *Precision_Ratio* 和最大聚类循环次数 *MaxCount*。

输出：对 *UnlabeledDataSet* 所有数据的聚类结果。

方法：

LabeledDataSet 构成初始 seeds 集，按类标签值将 seeds 集分为 *LabNum* 个 Cluster，同一 Cluster 的数据类标签值相同，不同 Cluster 的数据类标签值不同；

聚类循环次数 i=1；

do

{

for seeds 集的每个 Cluster

Apriori(Cluster，Support)；

计算每个关联规则的 χ^2 值和 MCS(最大 χ^2)值；

for *UnlabeledDataSet* 的每个数据

for 每个 Cluster 中与这个未标记数据匹配的关联规则集合

计算 WCS(带权 χ^2)值；

标记此无标签数据为最大 WCS 值对应的类标签值；并将此数据放入这个 Cluster 中；

if(i==1)

LabChangeRatio=1.00；

else

LabChangeRatio=*ObjectFunction*(*UnlabeledDataSet*)；

i++；

}

while(*LabChangeRatio*>*Precision_Ratio* 且 i<*MaxCount*)

其中，文献[9]给出了 Apriori 算法的详细流程，目标函数 *ObjectFunction*(*UnlabeledDataSet*)的实现如下：

ObjectFunction 过程：

输入：*UnlabeledDataSet*。

输出：类标签变化率 *ChangeRatio*。

方法：

MoveClusterNum=0；

for *UnlabeledDataSet* 的每个数据

if（本次聚类将数据从一个 Cluster 移入另一个 Cluster）

MoveCluster++；

ChangeRatio=*MoveCluster*/*UnlabeledDataSet* 数据总个数

此算法流程在迭代聚类的过程中只对初始未标记数据进行类标签值预测而始终保持初始已标记数据的类标签值不变，是 Constrained 半监督聚类算法。若每次迭代聚类时都对初始已标记数据进行类标签值的重新预测，则为 Seeded 半监督聚类算法。

4　实验及结果分析

实验的硬件环境是联想电脑：Pentium ® 2.6 GHz 处理机，512 MB 内存；软件环境是：微软 Windows XP SP2 操作系统，算法编程及运行环境为 Microsoft Visual C++6.0 SP6。

4.1　实验数据

实验使用了 4 个数据集 Inlinks、Ancurl、Alt 和 Caption。其中数据集 Inlinks 是从大学网站上收集的一部分 Web 网页，并被人工分成 Course 和 Non-Course 两类；数据集 Ancurl、Alt 和 Caption 是互联网网页上广告的集合，并被人工分成 Advertisement 和 Non-Advertisement 两类。为了使数据能够用一般值表示，4 个数据集中的文本信息都经过了预处理。经过处理的 4 个数据集的数据个数和属性个数如表 1 所示，属性是网页上的文本信息，每个数据在各个属性上的取值为 0 或 1，0 表示此网页数据不包含对应属性值表示的文本信息，1 表示此网页数据包含对应属性值表示的文本信息。

表 1　实验数据

数据集	数据个数	属性个数	类标签个数
Inlinks	133	334	2
Ancurl	228	472	2
Alt	94	111	2
Caption	25	19	2

4.2　实验结果

实验结果见表 2~表 4。

表 2　实验结果 1

数据集	算法 1		算法 2		最小支持度/(%)
	$\wedge^{(M)}$值	正确率	$\wedge^{(M)}$值	正确率	
Inlinks	0.272 648	0.933	0.155 454	0.876	5
Ancurl	0.150 520	0.928	0.138 596	0.888	5
Alt	0.159 057	0.906	0.092 894	0.890	7
Caption	0.148 526	0.882	0.148 526	0.882	25

表 3　实验结果 2

数据集	算法 1		算法 2		最小支持度/(%)
	$\wedge^{(M)}$值	正确率	$\wedge^{(M)}$值	正确率	
Inlinks	0.231 233	0.910	0.196 283	0.888	10
Ancurl	0.073 017	0.895	0.027 245	0.862	10
Alt	0.115 175	0.859	0.064 512	0.828	15
Caption	0.148 526	0.882	0.148 526	0.882	50

表 4　算法 3 的实验结果

数据集	Inlinks	Ancurl	Alt	Caption
$\wedge^{(M)}$值	0.149 264	0.000 160	0.133 499	0.104 723
正确率	0.854	0.421	0.891	0.824

算法 1　基于 seeds 集和频繁项集挖掘的半监督聚类算法。

算法 2　没有使用带权 χ^2 测试度量分类规则，所有分类规则权重相同，比较无标签数据与各个 Cluster 频繁项集的匹配率，用最高匹配率对应的 Cluster 的标签值标记此无标签数据，其他条件与算法 1 相同。

算法 3　文献[3]中的 Constrained-K-Means 算法，使用 Euclidean distance 作为距离计算公式。

把每个数据集随机的 1/3 数据作为已标记数据，其余 2/3 数据作为未标记数据，设定聚类最大循环次数为 8，目标函数类标签预测值变化率阈值为 0.02。聚类性能评价指标采用正确率和正则化互信息(Normalized Mutual Information，NMI)指标，它们分别从不同的角度评价聚类性能。NMI 的定义方式有很多种，文中采用文献[10]给出的定义，其计算公式为：

$$\wedge^{(M)}(k,\lambda)=\frac{1}{n}\sum_{e=1}^{k}\sum_{h=1}^{g}n_e^{(h)}\frac{\log\left(\frac{n_e^{(h)}\cdot n}{\sum_{i=1}^{k}n_i^{(h)}\sum_{i=1}^{g}n_e^{(i)}}\right)}{\log(k\cdot g)}$$

其中，k 为 Cluster 的个数，g 为实际标签的个数，n 为无标签数据总个数，$n_e^{(h)}$ 为实际标签为 h 被聚类到 Cluster c_e 中的数据个数。

NMI 可确定聚类算法在测试集上的类别标记结果与实际结果之间的共有信息量,NMI 值越大,则聚类效果越好。在实验中,4 个数据集的 k 和 g 的值都为 2,其中第一个标签值为 0,第二个标签值为 1,第一个 Cluster 对应的标签值为 0,第二个 Cluster 对应的标签值为 1。

4.3 实验结果分析

表 2、表 3 和表 4 给出了该文算法(算法 1)、整个实验进行过程中的前期算法(算法 2)和基于 seeds 集的半监督 *K*_Means 算法(算法 3)在 4 个数据集上的实验结果数据。其中,表 2 和表 3 是分别在两组不同支持度阈值下得到的算法 1 和算法 2 的实验结果比较,表 4 给出了算法 3 在 4 个数据集上运行的实验结果。由表 2、表 3 和表 4 的数据可知:无论从正确率还是 NMI 值的比较上来看,该文算法的聚类性能明显优于其他两个算法。从算法 1 和算法 2 的结果比较中得到的结论是:第一,关联规则的度量方法对整个算法的聚类性能有很大的影响;第二,支持度越低聚类结果精度越高、聚类性能越好。从算法 1 和算法 3 的结果比较中得到的结论是:在 seeds 集思想的基础上,采用频繁项集挖掘并使用带权 χ^2 测试方法度量关联规则可得到比采用 Euclidean distance 作为距离计算公式的 *K*_Means 算法更好的聚类性能。

另外,在实验进行过程中得到了这样一组数据,如表 5 所示。$n_2^{(1)}$ 是实际标签为 0 却被放在了 1 标签值对应的 Cluster 中的数据个数,$n_1^{(2)}$ 为实际标签为 1 却被放在了 0 标签值对应的 Cluster 中的数据个数。表 5 数据显示:$n_1^{(2)}$ 是使用该文算法对类标签值进行预测的主要错误来源,是影响聚类性能进一步提高的主要因素。针对此情况作者对数据集进行进一步分析时发现:0 标签值对应的涵义为非课程(Non-Course)/非广告(Non-Advertisement)网页,1 标签值为课程(Course)/广告(Advertisement)网页。频繁项集挖掘就是找出满足一定条件下数据间的共有信息。其实,无论是非课程网页数据和课程网页数据之间还是非广告网页数据和广告网页数据之间并不含有平等的共有信息量,显然非广告或非课程网页数据的共有信息量相对较少。而该文算法分别对两个标签值对应的 Cluster 进行频繁项集挖掘时使用了相同的支持度阈值,造成了相对较多实际标签为 1 的数据的类标签值被错误地预测为 0。由此可得,合适支持度阈值的选取是影响算法性能的另一重要因素。但从表 2 和表 3 的实验结果可以看出该文算法仍然得到了较好的聚类性能。不同类标签值对应的 Cluster 使用相同的支持度阈值也是造成表 2 的实验结果优于表 3 实验结果的一个原因:支持度阈值设置得越低,保留尽可能多的关联规则让带权 χ^2 测试方法度量其好坏,可得到更高的聚类结果精度、更好的聚类性能。但是支持度阈值降低时算法的运行时间无疑会增加,因此在现实生活使用该文算法时,应酌情考虑各种情况对运行时间和聚类结果精度进行折衷以尽量满足用户需求。

表 5 算法 1 实验结果数据

数据集	$n_2^{(1)}$	$n_1^{(2)}$	无标签数据总数	支持度/(%)
Inlinks	1	5	89	5
Ancurl	4	7	152	5
Alt	2	4	64	7
Caption	1	1	17	50
Ancurl	4	12	152	10

在实际应用中,Seeded 半监督聚类相对于 Constrained 半监督聚类抗噪性稍强。关于这一点文献[3-4]都通过实验进行了验证,并且为了考察和充分体现新算法中频繁项集挖掘和带权 χ^2 测试的作用,实验中不再考虑 Seeded 半监督聚类算法的性能评价,算法 1、算法 2 和算法 3 都是 Constrained 半监督聚类。

5 结束语

提出了基于 seeds 集和频繁项集挖掘的半监督聚类算法。该算法用频繁项集挖掘处理 seeds 集数据,并使用带权 χ^2 测试方法度量挖掘出的分类规则以对未标记数据进行类标签值预测,然后用新标记的数据扩大 seeds 集规模。实验结果显示:频繁项集挖掘能有效修正、净化 seeds 集中存在的噪声数据以及标签值预测时产生的误标记噪声数据;使用带权 χ^2 测试方法作为分类规则的度量指标可提高预测精确度,从而使得在 seeds 集规模迭代增大的同时改善 seeds 集质量,最终达到提高聚类性能的效果。

此外文中也存在着一些不足之处有待进一步改进,如:由于所处理的数据量不大,故采用 Apriori 算法进行频繁项集挖掘,而 Apriori 算法在处理大规模数据集时存在着明显的缺陷,这时应采用更有效的频繁项集挖掘算法;另外,用带权 χ^2 测试方法对关联规则进行度量前可对关联规则进行更有效的剪枝,以减少内存消耗量、缩短运行时间,使得在不影响聚类性能的前提下进一步提高算法的可扩展性。对挖掘出的关联规则进行有效的剪枝也是今后需要深入研究的一个重要方向。从实验结果可以看出关联规则的度量方法是提高算法预测精度的重要因素之一,在今后的研究中有待找出更合适的数学模型对关联规则进行更好的度量。

参考文献:

[1] Olivier C,Benhard S,Alexander Z.Semi-supervised learning[M].Cambridge:MIT Press,2006.

[2] Han Jia-wei,Kamber M.Data mining concepts and techniques[M].Beijing:China Machine Press,2006.

[3] Basu S,Banerjee A,Mooney R.Semi-supervised clustering by seeding[C]//Proceedings of the 19th International Conference on Machine Learning(ICML-2002),Sydney,Australia,July 2002:19-26.

[4] 邓超,郭茂祖.基于 Tri-Training 和数据剪辑的半监督聚类算法[J].软件学报,2008,19(3):663-673.

[5] 李远敏,林锦章.基于分治递归的层次聚类算法实现[EB/OL].http://xb.hbvtc.edu.cn/050320.htm.

[6] Grira N,Crucianu M,Boujemaa N.Unsupervised and semi-supervised clustering:A brief survey[R].A Review of Machine Learning Techniques for Processing Multimedia Content,Report of the MUSCLE European Network of Excellence(EP6).

[7] Li Wen-min,Han Jia-wei,Pei Jian.CMAR:Accurate and efficient classification based on multiple class-association rules[C]//Proc ICDM 2001:369-376.

[8] Coenen F,Leng P.An evaluation of approaches to classification rule selection[C]//Proceedings of the 4th IEEE International Conference on Data Mining,ICDM-04,November 2004:369-376.

[9] Agrawal R,Imielinski T,Swami A.Mining association rules between sets of items in large database[C]//ACM SIGM OD Conference,1993.

(下转 176 页)