文章编号:1671-4598(2017)08-0244-04 DOI:10.16526/j.cnki.11—4762/tp.2017.08.063 中图分类号:TQ028.1 文献标识码:A

基于数据挖掘的本体构建与重构技术研究

段妍羽，巩青歌，彭圳生
(武警工程大学 研究生管理信息工程系，西安 710086)

摘要：本体理论在知识工程领域得到广泛关注和普遍认可，构建完备且准确的领域本体已经越来越重要，同时，企业知识资源的更新与集成要求本体的不断进化与融合；针对目前本体构建与重构过程中数据处理效率低的问题，运用支持向量机分类及K-均值聚类的方法对本体构建数据进行处理，从文本数据中抽取关注的特定的信息，运用基于二叉树的多分类支持向量机以及支持向量机与K-均值融合的多样本聚类，总结基于分类与聚类的本体构建过程，并以离散型和连续型两种数据样本验证了方法的可行性；最后，在上述框架与理论研究的基础上，设计并开发了面向知识管理的本体工具平台，简单介绍系统的模块功能；实验结果表明，基于数据挖掘的本体构建与重构技术具有良好的应用效果。

关键词：本体构建与重构；文本处理；支持向量机；K-均值；分类；聚类

Researches on Ontology Construction and Reconstruction Based on Data Mining

Duan Yanyu, Gong Qingge, Peng Zhensheng
(Mangement Team of Postgraduate, Department of Information Engineering, Engineering University of PAP, Xi'an 710086, China)

Abstract: At present, ontology theory has attracted wide attention in the field of knowledge engineering. The construction of prefect and accurate domain ontology is getting more and more important, and at the same time, the update and integration of enterprise knowledge resource requires incessant evolution and merging of ontology. Aiming at the situation that process efficiencies and ontology integration is too slow, we use support vector machine classification and K- means clustering method to construct data processing. The thesis obtained specific information from the text data, and presented multiple-classification SVM and K-means clustering. Then, classification and clustering process was concluded for ontology construction and reconstruction, taking both discrete and continuous data sample as testing cases. The experimental results show that the proposed based on the ontology construction and reconstruction of data mining technology has good application effect.

Keywords: ontology construction and refactoring; text processing; support vector machines; K-means; classification; clustering

0 引言

随着科技的进步，各领域研究和应用的不断深入，针对相应领域的人和软件系统，基于数据挖掘的设计了一种通用全新的知识共享方式，其研究和应用已经延伸到多个领域，构建完备且准确的领域本体已经越来越重要[1]。本体理论研究不断走向成熟，本体构建方法也层出不穷，但目前而言，很多本体自动构建方法是基于某一特定语言的，大多是半自动的，距离完全自动构建还有一定差距，因此，如何自动化构建本体特别是中文本体，仍是一个需要不断改进的问题[2]。

自动化构建本体是为企业新领域知识服务的，随着本体技术的发展以及应用领域的推广，企业需要更多地考虑已有本体的更新以及重复利用，以支持企业知识的更快、更全面地共享[3]。但目前的重构技术应用十分有限，应用的领域比较集中，而且成本高，风险也大，因此需要通过重构技术规范本体，并通过实际的验证和应用来反映其应用价值[4]。

收稿日期:2017-03-07； **修回日期**:2017-03-15。

作者简介:段妍羽(1991-)，女，山东海阳人，硕士，主要从事大数据、数据挖掘方向的研究。

巩青歌(1967-)，女，陕西西安人，硕士，教授，主要从事虚拟现实和计算机仿真方向的研究。

本文针对目前知识管理中本体构建自动化程度低以及重用度低的问题，结合军用车辆设计领域研究了支持向量机、K-均值等挖掘算法等本体构建与重构中的关键技术一进行了深入研究。

1 文本构建与重构体系

本体构建是本体从无到有的过程，本体重构是对已存在的本体进行优化整合的过程。因此，知识管理的有效应用依赖于本体构建和重构两方面技术[5]。其中，本体构建方法研究是本体重构技术研究的基础和前期准备。通过本体构建方法的研究，深入理解领域概念及其语义关系在本体中的表现形式，本体重构技术可以更好挖掘本体建模元素以及他们之间的语义关系[6]。通过本体构建方法的研究可以构建语义关系明确，一致性较强的本体，以此支持本体重构技术研究。

针对本体构建方法对本体重构的影响，本文研究内容分为本体构建和本体重构两个研究阶段。第一，研究领域本体构建技术，利用已有工具并结合数据挖掘中数据处理方法，解决本体“从无到有”的问题；第二，研究本体重构技术，整理出本体重构总体流程，详细研究本体解析、数据处理和本体融合所需的关键技术。

本体构建主要包括本体规划、本体分析设计、本体评价确认、领域本体建立4个关键技术。本体的重构可以用于个体的完

善与更新，也可以是多本体的一个融合过程。该研究主要包括本体解析技术、数据处理技术、本体融合技术三个关键技术。

2　基于分类与聚类本体构建与重构技术

2.1　基于 SVM 的本体概念分类

基于线性可分情况下的思想，支持向量机是由最优分类面推论得出，核心的基本思想可用二维两类线性可分情况来说明[7]，具体如图 1 所示。图中两类不同的训练样本分别用实心点和空心点分别表述，其中 2 类没有错误地分开的分类线用 H 线表述。通过不同样本中距离分类线最近的点，同时平行于分类线 H 的直线，分别用 H1，H2 表述。两类的分类空隙或分类间隔具体指直线 H1 和 H2 之间的最短距离。通过定义最优分类线不但能将两类信息无错误地分开，而且能使两类的分类空隙最大［8］。前者的目的是为了保证经验风险最小，而后者的目标是使得分类空隙最大，实际上其本质就是使推广性界中具有最小化的置信范围，进一步降低真实风险。以此类推到到高维空间，最优分类线便构成了最优分类面。

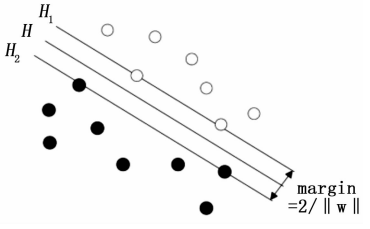

图 1　最优分类面的二维双类线性图

最优分类面的求解通常情况下可以分为两类：线性不可分、线性可分 2 种情况。而企业知识信息中提取的数据、术语等可能涉及多个领域，同一领域也可能涉及多个方面，根据分解重构法思想，一个复杂的多类问题可划分为多个两类问题来解决。采取决策树的组合分类策略已被证明是一种高效的多分类组合方法，利用 SVM 和决策树相结合的方法构造二叉树多级 SVM，从顶层开始，每一个包含多个类别的节点上的分类器将一个类别与其他类别分开从而实现了多类问题的分类。

本体的构建与重构首先要确定概念实例集的类别关系，而后再基于分类关系形成本体的机构框架，最后对实例、属性等进行修复得到较为完善的本体关系结构。本节重点描述基于 SVM 的有监督学习的概念实例类别划分过程。具体流程如下。

1）样本的选取：企业信息中已归类的概念样本，假设为 N 分类问题，训练样本为 $\varphi=\{X_1,X_2,\ldots,X_N\}$，且各树节点生成的最优分类面是将一类与其他类分开。

2）样本预处理：企业中的信息各式各样，其类别分布在多维空间，因此，需要选取适当的核函数，将训练样本向特征空间 H 中映射。

3）类间相对分离度计算：决策树构造中若分类错误越靠近树根节点，则对其性能的影响就越大。引入类之 间的相对分离度，可先将容易分的类分离出来，然后再分不容易分的类，从而达到较好的性能。

（1）计算特征空间 H 中各类的分类度 $F_i^H, i=1,2,\ldots,N$；

（2）将分离性测度按降序排列，设 $F_{m1}\geqslant F_{m2}\geqslant\ldots\geqslant F_{mN}$。

4）SVM 训练：

（1）设计数器 $k=1$；

（2）构造子分类器 SVM_k 的训练集 $\varphi=\sum 1+\sum 2$；其中：

$\sum 1=\{(X_{mk},+1)\}$，$\sum 2=\{(Y,-1)\mid y\in\{\varphi-X_{mk}\}\}$；

按两类问题构造分类器 SVM_k，计算过程如下：

当取适当的核函数 $K(x,x')$ 和适当的参数 C，在相应约束条件下，构造并求解最优化问题，得到最优解 $\alpha^*=(\alpha_1^*,\ldots,\alpha_n^*)^T$；选取 $\alpha*$ 的一个正分量 $\alpha j*>0$，并根据此计算出阈值 $b*$，其中 $b^*=y_j\left(1-\frac{\alpha_j^*}{C}\right)-\sum_{i=1}^{n}y_i\alpha_i^*(x_i\cdot x_j)$，构造决策函数 $f(x)$，$f(x)=\operatorname{sgn}\left(\sum_{i=1}^{l}\alpha_i^*y_iK(x_i\cdot x)+b*\right)$。

5）调整训练集和计数器：

其中，$\varphi=\varphi-\{X_{mk}\}$，$k=k+1$。

6）重复 4）和 5），直到构造完第 N—1 个子分类器 SVM N？1。

7）类别划分及评价：依据训练产生的规则，会产生一个新的分类结果，与样本对比，评价其准确性，同时，未知类别的样本可以通过学习规则，得到匹配的结果，其准确性与学习规则相一致。

8）生成最优或近优决策树：通过机器学习以及人为的辅助，提取的概念、样本集便得到各自的分类结果，并以树状形式展示。

2.2　基于 K－Means 的本体概念聚类

对于无学习样本的概念集，需要采用聚类的方式实现其类别划分，服务与本体的构建与重构，聚类过程与分类过程类似，区别只在于方法的选取，具体流程如下。

1）训练样本的选取：

选取企业信息中未归类的概念样本，训练样本为 $\varphi=\{X_1,X_2,\cdots,X_N\}$。

2）样本预处理：企业中的信息各式各样，其类别分布在多维空间，因此，需要选取适当的核函数，将训练样本向特征空间 H 中映射。

3）聚类计算步骤

（1）在随机情况下，确定 k 个沃罗诺伊集 K，其中 $k=1,\cdots,K$，L 个样本点的原样本集的子集表示为 Vk；

（2）针对每一个样本子集 Vk，采用线性规划下的支持向量机进行训练和计算；

（3）基于上一个步骤的结果，每个样本都会产生 k 个距离值，通过对比数值并且进一步重新分类，刷新替换每个 Vk 样本子集；

（4）在上一步骤的过程中，若每个样本 Vk 子集保持一致，则会出现聚类结果；否则转到第二个步骤继续训练。

4）聚类规则及结果：聚类过程中，机器会挖掘概念集之间的内在联系，产生聚类规则，并根据规则对样本进行归类，从而获得聚类结果。另外，如若有已分类的样本，可以二者对比，对聚类结果进行评价。

5）生成最优或近优聚类树：通过机器学习以及人为的辅助，提取的概念、样本集便得到各自的聚类结果。

3　算法设计与实验

3.1　基于 SVM 的本体概念分类实验

基于 SVM 的本体概念分类程序流程如图 2 所示。

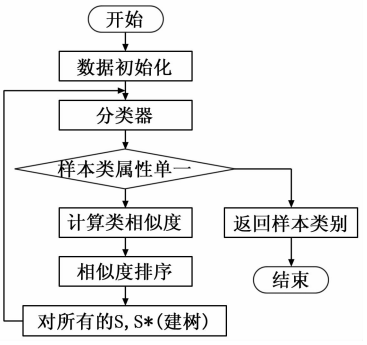

图 2 基于 SVM 的本体概念分类程序流程图

基于支持向量机的概念分类，其关键就是 SVM 分类器的构建。以下是其部分软件源代码：

```
Public void buildClassifier (Instances instances) throws Exception {
SVMTreeModelSelection modSelection=
new SVMTreeModelSelection (m minNum0bj, instances);
m root=new SVMTreeClassifierTree (modSelection);
m root. buildClassifier (instances);
}
```

ModelSelection 类是决定树的模型类。ClassifierSplitModel 对象的返回将由 SVMTreeModelSelection 类中的 selectModel 函数将根据系统指令执行，ClassifierSplitModel 本质上则是怎样分裂的模型。针对 SVMTreeModelSelection 类，其实由三个重要变量构成：

```
SVMTreeSplit [ ] currentModel;
SVMTreeSplit bestModel = null;
SVMTreeNosplit noSplitModel = null;
```

ClassifierSplitModel 被 SVMTreeNoSplit 和 SVMTreeSplit 继承，当样本均属于同一个样本时，系统不分裂，则 noSplitModel 对象被系统返回，若上述情况不发生，系统将针对第 j 个属性，调 currentModel [i] . buildClassifier 函数，根据 getErrors 的情况，系统最终选择具体的属性为最好的分裂属性。

属性值是缺失用公式表示为 treeIndex=－1，通过对每个子结点分开算计算其数值，然后累加起来。在不是缺失情况下，子结点为空，此时与上述子结点的计算方法保持一致，若情况不发生，则继续递归。当叶子结点发生下列情况：localModel 返回的是 ClassifierSplitModel 对象。则进一步调用 distributionForInstance，返回结果。

系统从有类别定义的样本中学习，得到样本的分类规则：

```
outlook = sunny
|humidity <= 75: yes(2.0)
|humidity > 75: no(3.0)
outlook = overcast: yes(4.0)
outlook = rainy
|windy = TRUE: no(4.0)
|windy = FALSE: yes(1.0)
```

系统从样本中学习了规则，系统会给出一个统计结果，用系统学习的规则对样本重新分类，然后再与原有样本比对，得到如下结果：

```
a b <-- classified as
7 2 | a = yes
1 4 | b = no
```

该结果表示：系统规则将 9 个原本类别为“ yes”的个体中的 7 个判为“yes”，而两个误判为“ no”，5 个原本为“no”一个判为“no”而又一个误判为“yes”，也就是说 14 个样本个体，11 个被正确判断、3 个误判，即准确率为 11/14。

3.2 基于 K－Means 的本体概念聚类实验

基于 K－Means 的本体概念聚类程序流程如图 3 所示。一共 4 个主要步骤：

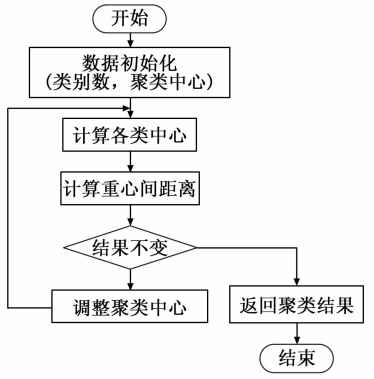

图 3 基于 K－Means 的本体概念聚类程序流程图

第一步，使用距离计算的最小平方法，统计从每个数据样例到群集中心（随机选中的数据行）的距离；第二部，通过计算，根据到每个群集中心的最短距离，将每个数据行分配给一个类集；第三步，通过每个类集的数的每列数据的平均数计算重心；第四步，统计计算所有数据样例与上述步骤创建重心之间的距离。当群集及群集数保持不变时，类集的创建工作完成。如果发生变化，则返回到第三步骤，重新开始并重复计算，直到保持稳定不再变化为止。

分类中训练一个分类器是用 buildClassifier ()，在聚类中学习一个 Clusterer 是用 buildCluster ()。分类中分类一个样本是用 classifyInstance，而在聚类中是用 clusterInstance。它继承自 RandomizableCluster，而 RandomizableCluster 又继承自 AbstactCluter，进入 AbstactCluster，它有三个比较重要的函数，buildCluster，clusterInstance，distributionForInstance。

聚类分析后，系统也是得到两类结果，一是样本的最优聚类中心；另一个则是样本中每个个体的的类别结果。

聚类中心即每一个类别的属性均值，在学习前，人为的定义类别的数量，如联轴器，我们已经知道列举的样本中包含的常用的 4 种类型，因此，系统会定义 4 个聚类中心，而对于类别数量未知的情况，只能通过系统的多次学习，比较结果中聚类中心哪个更合理，从而确定最优方案。

结果中统计了样本的所有属性，给给出了集合的属性均值以及类别数目，每个类集合展示了一种特征，专业人员根据经验分析，为每一个类别赋予定义：群集 0—凸缘联轴器，群集 1—弹性柱销联轴器，群集 2—弹性套柱销联轴器，群集 3—梅花形弹性联轴器。

聚类中心给出了每个类别的属性特性，系统学习的最终目的还是要得到每一个样本个体的类别，通过判断，得到详细聚类结果如图 4 所示。

图中每一个点代表了群集的一个样本个体，X 轴表示类别，Y 轴表示样本号，经过聚类训练后，原本分散在空间中的样本则有规则的堆积在一起，系统通过学习，发现了样本之前的内在关系，并通过这种关系进行聚类判断。因此，可以得

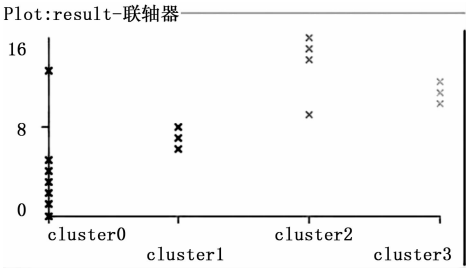

图 4　聚类结果

出，只要样本的属性间关系明确，便可以学习到准确率很高的聚类中心及结果。

4　结束语

在知识经济逐渐兴起，信息技术飞速发展，商业竞争日益加剧的背景下，知识管理得到越来越多企业的重视。为了解决知识管理中出现各种信息通信和知识共享问题，原本用于语义 Web 的本体论也被引入到知识管理中。

本文针对目前知识管理中本体特别是中文本体构建自动化程度低以及重用度低的问题，结合企业生产应用，提出了多分类支持向量机的本体设计方法和 K－均值聚类的本体设计方法流程，分析了支持向量机及统计学的基本原理与应用与 K－均值的基本原理与应用，实现了基于类间相对分类度的概念分类和基于类间相对分类度的概念聚类，并在此基础上，构建了本体关系框架，验证了方法的可行性。

参考文献：

[1] 李兴春. 计算机信息检索中的本体构建研究 [J]. 重庆文理学院学报，2013，3：87－91.

[2] 张　娟. 基于本体的可重构知识管理系统研究综述 [J]. 现代商贸工业，2009，21 (19)：59－60.

[3] 张　祥，李　星，温韵清，等. 语义网虚拟本体构建 [J]. 东南大学学报：自然科学版，2015，4：652－656.

[4] Dibike Y B，Solomatine D，Velickov S，et al. Model Induction with Support Vector Machines：Introduction and Applications [J]. Journal of Computing in Civil Engineering，2014，15 (3)：208－216.

[5] Ren H，Tian J，Wierzbicki A P，et al. Ontology Construction and Its Applications in Local Research Communities，Modeling for Decision Support in Network－Based Services [M]. Springer Berlin Heidelberg，2012：279－317.

[6] Xue S，Jing X，Sun S，et al. Binary－decision－tree－based multiclass Support Vector Machines [A]. 2014 14th International Symposium on Communications and Information Technologies (ISCIT) [C]. IEEE，2014：85－89.

[7] 任维武，胡　亮，赵　阔. 基于数据挖掘和本体的入侵警报关联模型 [J]. 吉林大学学报 (工学版)，2015 (3)：899－906.

[8] Balabantaray R C，Sarma C，Jha M. Document Clustering using K－Means and K－Medoids [J]. International Journal of Knowledge Based Computer System，2015，1 (1).

(上接第 243 页)

统计结果分析可以看出，采用本文方法进行数据加密设计，能更好地满足频数检验的要求，提高加密性能。图 2 给出了采用不同方法进行数据加密的抗攻击性效能对比，分析图 2 结果得知，采用本文方法进行 DES 数据加密，抗攻击能力较强，性能优于传统模型。

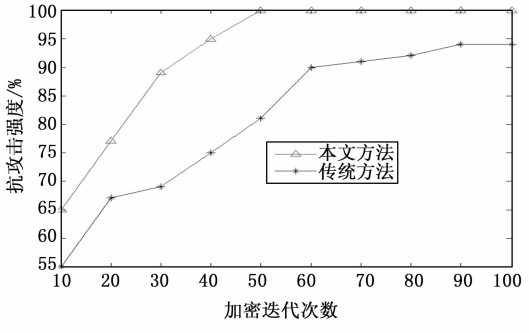

图 2　加密抗攻击性能对比

4　结束语

本文提出一种基于前向纠错编码的 DES 数据公钥加密技术，采用 Gram－Schmidt 正交化向量量化方法构建 DES 数据的 Turbo 码模型，通过三次重传机制产生密文序列，对密文序列进行前向纠错编码设计，结合差分演化方法进行频数检验，实现网络信息安全中 DES 数据加密密钥构造，选择二进制编码的公钥加密方案有效抵抗密文攻击。通过加密算法设计和实验分析表明，采用该加密技术进行 DES 数据加密的抗攻击能力较强，密钥置乱性较好，具有很高的安全性和实用价值。

参考文献：

[1] 林如磊，王　箭，杜　贺. 整数上的全同态加密方案的改进 [J]. 计算机应用研究，2013，30 (5)：1515－1519.

[2] 刘乐柱，张季谦，许贵霞，等. 一种基于混沌系统部分序列参数辨识的混沌保密通信方法 [J]. 物理学报，2014，13 (1)：010501.

[3] Lin Y P，Vaidyanathan P P. Theory and design of two－dimensional filter bank：A review [J]. Multidimensional System & Signal Processing，1996，7 (3)：263－330.

[4] Suzukit，Kudo H. Two－dimensional non－separable block－lifting structure and its application to M－channel perfect reconstruction filter banks for lossy－to－lossless image coding [J]. IEEE Transactions on Image Processing，2015，24 (12)：4943－4951.

[5] 徐光宪，徐山强，郭晓娟，等. DCT 变换与 DNA 运算相结合的图像压缩加密算法 [J]. 激光技术，2015，39 (6)：806－810.

[6] 龙昭华，龚　俊，王　波，等. 无线传感器网络中分簇安全路由协议保密通信方法的能效研究 [J]. 电子与信息学报，2015，37 (8)：2000－2006.

[7] 白恩健，朱俊杰. TinySBSec—新型轻量级 WSN 链路层加密算法 [J]. 哈尔滨工程大学学报，2014，35 (2)：1－6.

[8] 汤殿华，祝世雄，曹云飞. 一个较快速的整数上的全同态加密方案 [J]. 计算机工程与应用，2012，18 (28)：117－122

[9] 秦　怡，吕晓东，巩　琼，等. 利用附加密钥旋转在光学联合相关结构中实现多二值图像加密 [J]. 光学学报，2013，33 (3)：7－9.

[10] Shen L，Sun G，Huang Q，et al. Multi－level discriminative dictionary learning with application to large scale image classification [J]. IEEE Transactions on Image Processing，2015，24 (10)：3109－3123.

[11] Thiagarajan J J，Ramamurthy K N，Spanlas A. Learning stable multilevel dictionaries for space representations [J]. IEEE Transactions on Neural Networks & Learning Systems，2015，26 (9)：1913－1926.