基于语义距离的 *K*-最近邻分类方法*

杨 立[1,2+], 左 春[1], 王裕国[1]

[1](中国科学院 软件研究所,北京 100080)

[2](中国科学院 研究生院,北京 100049)

K-Nearest Neighbor Classification Based on Semantic Distance

YANG Li[1,2+], ZUO Chun[1], WANG Yu-Guo[1]

[1](Institute of Software, The Chinese Academy of Sciences, Beijing 100080, China)

[2](Graduate School, The Chinese Academy of Sciences, Beijing 100049, China)

+ Corresponding author: Phn: +86-10-82523259, Fax: +86-10-82523227, E-mail: yangli@sinosoft.com.cn, http://www.ios.ac.cn

Received 2004-04-09; Accepted 2005-01-24

Yang L, Zuo C, Wang YG. *K*-Nearest neighbor classification based on semantic distance. *Journal of Software*, 2005,16(12):2054−2062. DOI: 10.1360/jos162054

Abstract: Most research on distance metric of *k*NN classification is focused on how to integrate the differences caused by various attributes, and the semantic difference between values of the same attribute is ignored. In addition, classification accuracy of the traditional approaches is very sensitive to the incomplete data described on different abstract levels. In this paper, a novel *k*NN approach based on semantic distance——SD*k*NN (semantic distance based *k*-nearest neighbor) is presented, which solves the two problems mentioned above. This approach analyzes the semantic difference between values of an attribute and presents how to calculate the semantic distance based on domain ontologies, and the semantic distance is then used to improve the traditional *k*NN methods. Experiments on the UCI (University of California, Irvine) machine learning repository and real application datasets show that the overall performance of SD*k*NN outperforms the traditional one, especially when the data is incomplete. SD*k*NN also has the desirable application value in practice.

Key words: ontology; semantic distance; *k*NN; classification

摘 要: 最近邻分类方法中对距离机制的研究大都集中在根据何种计算方法将不同属性取值的差异集中起来,而未考虑到同一属性间取值的语义差异所带来的影响;而且传统算法的分类准确率对于不同抽象层次描述的数据集带来的数据不完整性相当敏感.针对这两个问题,提出一种基于语义距离的最近邻分类方法SD*k*NN(semantic distance based *k*-nearest neighbor).该方法分析了同一属性内取值的语义差异,说明了如何基于领域本体计算语义距离,并将其应用到 *k*NN 算法中.经过在 UCI 数据集以及实际应用数据集中验证,SD*k*NN 的整

* Supported by the National High-Tech Research and Development Plan of China under Grant No.2003AA112050 (国家高技术研究发展计划(863)); the Key Science-Technology Project of the National 'Tenth Five-Year-Plan' of China under Grant No.2001BA102A05-02 (国家"十五"科技攻关计划)

作者简介: 杨立(1978−),男,江西南昌人,博士生,主要研究领域为数据仓库,数据挖掘,知识管理;左春(1959−),男,研究员,主要研究领域为数据库,网络工程;王裕国(1941−),男,研究员,博士生导师,主要研究领域为知识管理,多媒体应用技术.

体性能要优于传统方法,在数据不完整的情况下效果更为明显,实践证明,SD*k*NN 有较好的应用价值.

关键词: 本体;语义距离;最近邻;分类
中图法分类号: TP18 文献标识码: A

最近邻方法(*k*-nearest neighbor,简称 *k*NN)是一种简洁而有效的非参数分类方法.它的工作原理是首先找到被分类对象在训练数据集中的 k 个最近的邻居,然后根据这些邻居的分类属性进行投票,将得出的预测值赋给被分类对象的分类属性.这种方法也被称为延迟学习(lazy learning),其优点包括:

1) 无须事先知道属性值分布.大多数其他分类方法(如:Bayesian 分类法)则要求事先知道属性值分布.

2) 非常适合增量学习的情况,例如对数据流的分类.

3) 由于并不要求得出显式的规则,一般来说,*k*NN 的分类准确率要高于其他分类方法.

基于以上特点,自从 *k*NN 方法被提出以来,就在文本分类[1,2]、模式识别、图像及空间分类等领域得到广泛的应用.在实际应用中,人们又提出了很多改进的方法,主要分为以下几种:针对 k 值选择问题,文献[1]提出了根据上下文动态调整 k 值的选择;针对特征属性选择问题,文献[2]提出了基于权值的特征属性选择方法;针对距离机制问题,以往常见的距离度量包括 Minkowski Distance,Euclidean Distance,Mahalanobis Distance,Manhattan Distance,Cosine Angle Distance[2]等.文献[3]提出了基于梯度下降法生成权值来集成属性之间的距离,这种方法相对其他距离度量来说,对规模较小和分布不均匀的数据集有着更好的适应性.

距离机制是最近邻分类方法中的关键部分,传统上在对距离机制的讨论中,往往集中在根据什么计算方法来将各个属性取值的差异集成起来,不同属性取值差异对距离的贡献通过权值来反映.而对于如何计算同一属性内取值的差异则讨论得不多,往往都被考虑得相对简单,对离散型属性的取值,认为其间不存在任何语义关系,因而,在差异(或距离)计算时采取完全相同的处理方法.在以数值型数据为主的领域内(如:图像和空间分类),传统方法取得了良好的分类效果,但是当应用到以离散型数据为主的企业应用领域时就会存在性能下降的问题.如在企业财产保险的承保过程中,在填写企业行业类别时,其属性取值可能是“造纸”、“家具制造”、“机械零件制造”等.我们知道,“造纸”和“家具制造”由于是以木材为主要原料的加工,这些工厂发生火灾的概率是比较接近的,而且都要远远高于“机械零件制造”工厂发生火灾的概率.这种行业之间的语义关系对保险公司判断保险标的的风险程度来说是相当重要的.而在利用 Traditional *k*NN 进行风险预测分析时,对上述 3 个取值之间的差异(或距离)就无法进行区分,从而影响了预测的准确率.应该指出的是,离散型的数据在实际应用中是特别常见的,以保险领域为例,仅在评价一个企业的风险状况时,就可能遇到行业类别、企业性质、管理水平等多项离散型属性.另外一个重要问题是数据描述的粒度问题.在实际应用中,对数据的描述往往是处在不同抽象层次上的.如在企业填写承保单时,行业类别可能被笼统地描述为“五级工业”,也可能被描述得更为详细,如“造纸”或“家具制造”等.在保险理赔领域中,对事故现场勘查的描述可能也有详细程度的不同.显然,对属性值的描述程度越详细,越能精确反映对象的原貌,否则就会带来描述的模糊性,这属于对对象的部分描述或者数据不完整.在实践中我们发现,传统的最近邻分类算法通常假定属性都是用同一抽象层次的数据来描述,而缺乏对不同抽象层次数据描述的支持,分类准确率随着最细节数据缺失程度的增加而严重下降.针对这两个问题,本文对同一属性内不同取值之间的差异作了分析和研究,提出一种基于语义距离的 *k*NN 方法——SD*k*NN(semantic distance based *k*-nearest neighbor),该方法引入了领域本体,可针对不同抽象层次上的属性取值计算出语义距离,改进了传统方法中对距离机制的定义,同时解决了上述两个问题,实验证明了该方法的有效性.

本文第 1 节介绍了最近邻方法的定义并说明传统方法在实际应用中的不足.第 2 节给出基于本体计算语义距离的方法.第 3 节给出 SD*k*NN 的具体框架描述.第 4 节描述实验和结果分析,最后是结论及未来工作方向.

1 最近邻方法

训练集 $X=\{x_1,x_2,\ldots,x_{|F|}\}$是多维空间中的点集,$F$ 是特征属性集.F 中的属性可以为离散的或是连续的.分类属性为 l,它是一个离散性变量,其值域为 L.分类的目标是最小化分类误差(misclassification error)M_j,即对于每一

个取值 $l_j \in L$,

$$M_j = \sum_{l_{j'} \in L} R_{l_j l_{j'}} p(l_{j'} \mid q) \tag{1}$$

其中 $R_{l_j l_{j'}}$ 为将取值 l_j 分类为 $l_{j'}(j \neq j')$ 造成的误差,q 为被预测点, $p(l_{j'} \mid q)$ 为将 q 分类为 $l_{j'}$ 的概率.通常来说,kNN 假定所有的误分类都具有相同的误差,即

$$R_{l_j l_{j'}} = \begin{cases} 0, & j = j' \\ 1, & j \neq j' \end{cases} \tag{2}$$

而 kNN 方法并不能精确地预测出 q 的分类属性值,而是给出最有可能的预测值

$$kNN(q) = \underset{l_j \in L}{\operatorname{argmax}}\, p(l_j \mid q) \tag{3}$$

其中 $kNN(q)$表示 kNN 方法对于被预测点 q 的预测结果.

kNN 方法与其他分类方法相比,在定义先验概率的方式上有所不同

$$p(l_j \mid q) = \frac{\sum_{x \in N_q} \begin{cases} K(d(x,q)), & x_l = l_j \\ 0, & \text{otherwise} \end{cases}}{\sum_{x \in N_q} K(d(x,q))} \tag{4}$$

其中 N_q 为 q 在训练数据集 X 中根据距离函数 $d(x,q)$决定的最近邻集合,K 是一个核函数,定义为

$$K(d(x,q)) = \frac{1}{d(x,q)} \tag{5}$$

对于每一个 $x \in X$,在 kNN 中用下式计算相对于 q 的距离 $d(x,q)$:

$$d(x,q) = \left(\sum_{f \in F} w(f) \cdot \xi(x_f, q_f)^r \right)^{\frac{1}{r}} \tag{6}$$

当 r=2 时,$d(x,q)$就是最常见的 Euclidean distance,函数 ξ 定义了对于同一个属性 f 来说,取值的不同对于距离计算的影响.

$$\xi(x_f, q_f) = \begin{cases} |x_f - q_f|, & f \text{ is continuous} \\ 0, & f \text{ is discrete and } x_f = q_f \\ 1, & f \text{ is discrete and } x_f \neq q_f \end{cases} \tag{7}$$

$w(f)$定义了属性 f 在 kNN 中的权值,在计算中也可记为 w_f,满足

$$\sum_{f \in F} w(f) = s \tag{8}$$

其中,s 为一个常量,通常设为 1.

最近邻分类方法的步骤为:首先确定一个合适的距离机制(通常为 euclidean distance),对于测试集中的每一个数据点 s,在训练集中根据距离机制找到 s 的 k 个最近的邻居,根据 k 个最近邻的分类属性取值投票决定被预测点的分类属性.预测完成后,根据式(1)来确定分类误差或分类准确率.

在实际应用中,属性的取值很可能来源于某个领域本体[7],我们注意到,在式(7)中,对于离散型属性 f,如果两个属性值不同,ξ 值均为 1,这并不能反映事物的客观状况,特别是当两个属性值之间存在着非常紧密的语义关系时,如下义关系或同时与另一个值具有下义关系.在 kNN 方法中,当选取最近邻时,应该优先选取那些与被预测点具有更紧密语义关系的数据进行投票.Traditional kNN 未考虑属性值之间的语义关系,这将会严重影响到 kNN 在实际应用中的分类准确率.而正确处理这些语义关系将可以提高 kNN 方法的分类准确率,因此我们提出了基于语义距离的 kNN 方法 SDkNN,该方法的关键是如何基于本体计算属性值之间的语义距离.

2 基于本体的语义距离计算方法

2.1 本 体

本体(ontology)最初是一个哲学上的概念,从构成上说,本体是构成应用领域中词汇的基本术语和关系,以及结合这些术语和关系扩展新词汇的基本规则的有机组合体[4].从物理含义上说,本体是一种某一应用领域中概念的显示说明,即领域知识的概念化表达[5].

在许多应用环境中,概念之间的关系通过语义进行关联[6].多数文献建议,本体可以作为表达这种语义关联的框架[7].在本体中,我们可以直接知道某些概念间的语义关联,包括同义(synonymy)、反义(antonymy)、下义(hyponymy)、部分-整体关系(meronymy)、方式关系(troponomy)、必要条件关系(entailment)等.在本体中,最常见的是下义关系.在本文中,我们也将重点讨论这种关系.

定义 **1**(本体)[8].本体 O 是一个 5 元组 $O:=(V,F,C,H,Root)$,其中 V 为一组词汇集,C 为一组概念,F 为一个参照函数 $F:2^V \mapsto 2^C$.将一个词汇集 $\{V_i\} \subset V$ 映射到一个概念集.通常来说,多个词汇可以映射到一个概念,而一个词汇也可以映射到多个概念.H 是层次关系 $H \subseteq C \times C$,$H(c_1,c_2)$代表 c_1 是 c_2 的子概念,H 是有向的、无环的、传递的、自反的;$Root$ 是一个根概念,$\forall c \in C, H(c,Root)$ 成立,在一个本体中,有且只有一个 $Root$.在本文中,如未明确指出,一般认为 $H(c_1,c_2)$关系中,$c_1 \neq c_2$.图 1 描述了一个保险领域内行业分类的本体.

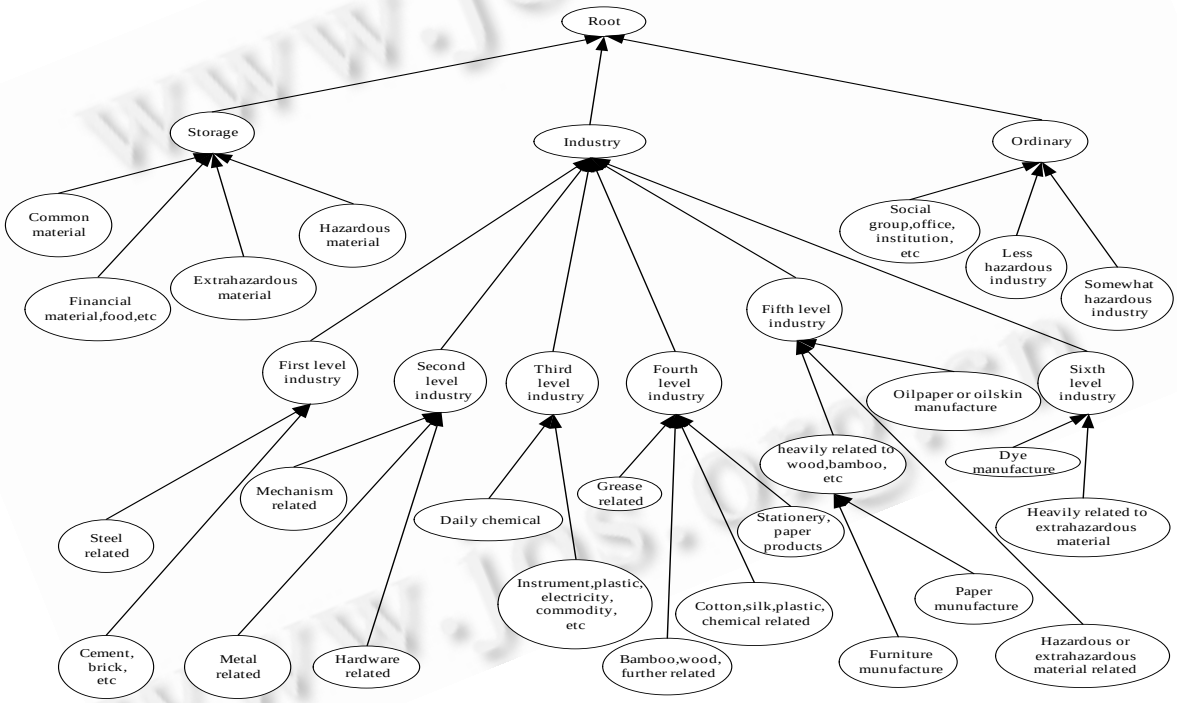

Fig.1 An ontology for industry taxonomy in insurance domain

图 1 保险领域内关于行业分类的本体

2.2 语义距离

定义 1 中的 H 关系只描述了一部分语义关系,除此以外,我们更关心的是与同一个概念相关联的概念之间是否有语义关联.例如:图 1 中的“造纸”行业和“家具”制造业都是“五级工业”的一种,在某种程度上我们可以认为它们是语义类似的.在这种情况下,语义关系的比较演变成了语义邻居的比较,这样,本体可以对单纯基于统计的属性取值分布做语义上的补充,而原有的统计模型可以利用本体提供的概念空间来提供更全面的统计证据,下面我们来具体讨论语义距离的计算方法.

基于本体的语义距离主要测量本体中概念间连接边的长度[9],概念间的语义关联程度通过几何度量来表征.显然,两个概念在本体中的连接路径越短,它们就越类似,每条连接边的长度由其在训练数据集中包含的信息量来决定.

设 $p(c)$为概念 c 在整个概念集中的发生概率.计算方法为:设 $count(c)$为概念 c 在数据集 D 中的出现次数,$count(D)$为数据集 D 的记录总数,考虑到概念的出现可能以不同抽象层次的形式出现,计算概念的总出现次数时应累加其所有子概念的出现次数.

$$p(c)=\frac{count(c)+\sum_{H(c',c)}count(c')}{count(D)} \tag{9}$$

不难得出, $p(c)$随着 c 所在层次的增大是单调增加的,且 $p(Root)=1$.

设 c 为本体中的概念,$parent(c)$为 c 在本体中的父亲集合,即

$$parent(c)=\{c'\mid H(c,c'),\neg\exists c'',\text{s.t.}\,H(c,c''),H(c'',c')\} \tag{10}$$

根据 H 关系的无环性,易知:

$$\forall c\in C,\text{if}\ \ parent(c)\neq\varnothing,\text{then}\mid parent(c)\mid=1 \tag{11}$$

其中,$|X|$代表集合 X 中包含的元素个数.

式(11)说明,如果 c 存在父亲,则其父亲是唯一的,以下我们用 $parent(c)$直接表示 c 的父亲.

根据信息论的知识,概念 c 所包含的信息量为

$$I(c)=-\log(p(c)) \tag{12}$$

连接边 $c\rightarrow parent(c)$ 包含的信息量为

$$\begin{aligned}I(c\rightarrow parent(c))&=-\log(p(c\rightarrow parent(c)))=-\log(p(c\mid parent(c)))\\&=-\log\left(\frac{p(c)}{p(parent(c))}\right)=I(c)-I(parent(c))\end{aligned} \tag{13}$$

连接边 $c\rightarrow parent(c)$ 的长度应正比于它所包含的信息量

$$length(c\rightarrow parent(c))\propto I(c\rightarrow parent(c)) \tag{14}$$

由式(13)、式(14)得到:

$$length(c\rightarrow parent(c))=\alpha\cdot(I(c)-I(parent(c))) \tag{15}$$

其中 α 为一个常量.为了方便表达,不失一般性,在本文中,令 $\alpha=1$,得到

$$length(c\rightarrow parent(c))=I(c)-I(parent(c)) \tag{16}$$

设 c_1 和 c_2 是属性 f 的两个取值,O 为属性 f 对应的领域本体,由于 $\forall c\in C,H(c,Root)$ 成立,故 c_1 和 c_2 肯定存在共同的祖先,设 $msa(c_1,c_2)$为 c_1 和 c_2 在本体 O 中的最小祖先(most specific ancestor).即

$$msa(c_1,c_2)=\{c\mid H(c_1,c),H(c_2,c),\neg\exists c',\text{s.t.}\,H(c_1,c'),H(c_2,c'),H(c',c)\} \tag{17}$$

定义 **2**(父子链). 若 c_1 和 c_2 满足 $H(c_1,c_2)$,则 c_1 到 c_2 的父子链记做 $pcc(c_1,c_2)$,定义为

$$pcc(c_1,c_2)=\{c_1^0,c_1^1,c_1^2,...,c_1^{k-1},c_1^k\mid c_1=c_1^0,c_2=c_1^k,k\geq1,\forall i,0\leq i\leq k-1,c_1^{i+1}=parent(c_1^i)\} \tag{18}$$

根据 H 关系的无环性,若 c_1 和 c_2 满足 $H(c_1,c_2)$,那么 $pcc(c_1,c_2)$是唯一的.

定义 **3**(连接路径). 根据 $\forall c\in C,H(c,Root)$ 以及 H 关系的无环性,c_1 和 c_2 在本体 O 中的连接路径有且只有一条,记做 $path(c_1,c_2)$,而 c_1 到 $msa(c_1,c_2)$的父子链与 c_2 到 $msa(c_1,c_2)$的父子链的并集即为这样的连接路径,定义如下:

$$path(c_1,c_2)=pcc(c_1,msa(c_1,c_2))\cup pcc(c_2,msa(c_1,c_2)) \tag{19}$$

定义 **4**(语义距离). 基于定义 3,我们将语义距离定义为

$$\begin{aligned}semantic\_distance(c_1,c_2)&=\sum_{c\in\{path(c_1,c_2)\text{-}msa(c_1,c_2)\}}length(c\rightarrow parent(c))\\&=\sum_{c\in\{path(c_1,c_2)\text{-}msa(c_1,c_2)\}}I(c)-I(parent(c))\\&=(I(c_1)-I(msa(c_1,c_2))+(I(c_2)-I(msa(c_1,c_2))\\&=2\log(p(msa(c_1,c_2)))-(\log(p(c_1))+\log(p(c_2)))\end{aligned} \tag{20}$$

我们注意到,式(20)中的语义距离定义实际上包含两方面的信息,一方面领域本体的构成决定了 $msa(c_1,c_2)$ 的位置,另一方面,$p(msa(c_1,c_2))$,$p(c_1)$,$p(c_2)$的值来自于数据集本身的统计信息.同一个属性相对于不同的领域本体,其取值之间的语义距离是不同的,而不同的数据集相对于同一个领域本体,其取值之间的语义距离也是不同的.这种计算方法综合了概念间的语义关系以及客观发生的统计信息,有助于我们更准确地模拟客观世界的原貌,并发现其中隐含的规律或模式.

3 基于语义距离的最近邻方法——SD*k*NN

在 SD*k*NN 中,用语义距离 *semantic_distance* 来构成距离函数,将式(7)重新定义:

$$\xi(x_f,q_f)=\begin{cases}|x_f-q_f|, & f \text{ is continuous}\\ 0, & f \text{ is discrete and } x_f=q_f\\ semantic\_distance(x_f,q_f), & f \text{ is discrete and } x_f\neq q_f\end{cases} \tag{21}$$

3.1 数据预处理——计算w_f

在 SD*k*NN 中,属性的权值是根据信息增益[10]来决定的,首先利用分类中常用的离散化方法[10]将连续性变量离散化.

假设属性 f 有 m 个不同取值,根据该属性的取值可将训练集 X 分为 m 个子集:$X_1,X_2,\ldots,X_m$.划分后所获得的信息增益记为 $gain_f$,由下式计算:

$$gain_f=Entropy(T)-\sum_{i=1}^{m}\frac{|X_i|}{|X|}Entropy(X_i),\ Entropy(D)=-\sum_{l_j\in L}\frac{count(l_j,D)}{|D|}\cdot\log\frac{count(l_j,D)}{|D|} \tag{22}$$

其中,$Entropy(D)$是数据集 D 的信息熵,$count(l_j,D)$为数据集 D 中分类取值为 l_j 的记录数.

属性权值 w_f 由 $gain_f$ 进行归一化处理得到,如下式:

$$w_f=\frac{gain_f}{\sum_{f\in F}gain_f} \tag{23}$$

3.2 算法描述

SD*k*NN 的框架描述如下:

```
Algorithm SDkNN(T:test dataset)
Begin
  Get weight of each attribute w_f using Eq.(23);
  For each node q in T Do
      Begin
          Compute distance between q and each node x in the training set X by calling function Distance(x,q);
          Choose the k nearest nodes as a neighbor set Nq;
          Take a majority voting among Nq to decide the class label of q;
          If no label is better than others then choose the most possible label in X;
      End
  For each label  l_j ∈ L  Do Compute the misclassification error M_j using Eq.(1);
  classification_accuracy = Σ_{l_j∈L} (1 − M_j) p(l_j) ;
End
Function Distance(x,q:node):real
Begin
    For each attribute f∈F Do
      Begin
        Compute  ξ(x_f, q_f)  using Eq.(21); /* where semantic_distance(x_f,q_f) is computed using formula
                                                (20), p() is computed using formula (9)   */
      End
      Compute d(x,q) from all  ξ(x_f, q_f)  and w_f using distance Eq.(6);
```

Return $d(x,q)$;
End

4 实 验

4.1 实验设计

为了评价 SD*k*NN 的整体性能,我们基于 Java 语言实现了 SD*k*NN,并通过实验与传统的 *k*NN 方法 Traditional *k*NN 和文献[2]中提出的算法 WA*k*NN 进行比较,其中 Traditional *k*NN 算法从http://www-2.cs.cmu.edu/afs/cs.cmu.edu/user/zhuxj/www/courseproject/knndemo/KNN.java处获得,基于 Java 自行实现了 WAkNN.实验环境为 P4 1.7G/256M RAM 的 PC 机,操作系统为 Microsoft Windows 2000 Server,数据库为 Microsoft SQL Server2000.在 UCI 数据集[11]中挑选了 Mushroom Toxicology 和 Nursery 进行实验,Mushroom Toxicology 数据集包含 8 124 条记录,每条记录都包含 22 个属性.其中 4 208 条记录的分类属性为 edible,剩下的 3 916 条记录为 poisonous.采用文献[12]中的本体,在 22 个属性中,有 17 个属性对应着 3 层以上的本体.

Nursery 数据集包含 12 960 条记录.每条记录有 8 个离散型的属性,分类属性包含 5 种可能取值.其中用到的本体来源于http://www.cs.iastate.edu/~cs573x/labs/lab2/04lab2.html.

我们还选取了国内某财产保险公司的理赔数据(记为 insurance)进行测试,该数据集包含 32 991 条记录,每条记录包含有行业类别、企业性质等 12 个属性,分类属性为风险程度,根据赔付率的大小分为高、中和低,采用保险公司内部应用的属性值分类作为本体,并采用 10-fold cross validation 计算分类准确率.

对于前两个数据集 Mushroom Toxicology 和 Nursery,每个数据集中随机挑选了一部分最细节数据替换成更高抽象层次的数据,替换比率从 10%~50%,用来进一步考察 SD*k*NN 对于这种数据不完整情况下的性能,在实验中取 k=1.

4.2 实验结果分析

由图 2~图 4 中可以看出,SD*k*NN 在 3 个数据集上的分类准确率均要优于 WA*k*NN 和 Traditional *k*NN,这是因为 SD*k*NN 较好地反映了离散属性值之间的语义关系,从而能更准确地模拟出数据的客观分布.而且从图 4 中可以看出,对于实际应用数据集 insurance,由于数据中存在着不同层次的数据,SD*k*NN 的性能优势更为明显.

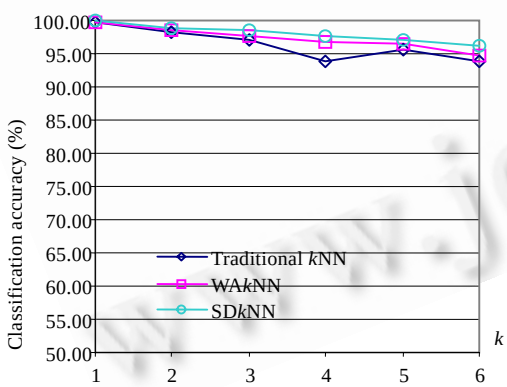

Fig.2 Classification accuracy comparison on mushroom dataset

图 2 在 mushroom 数据集上的分类准确率比较

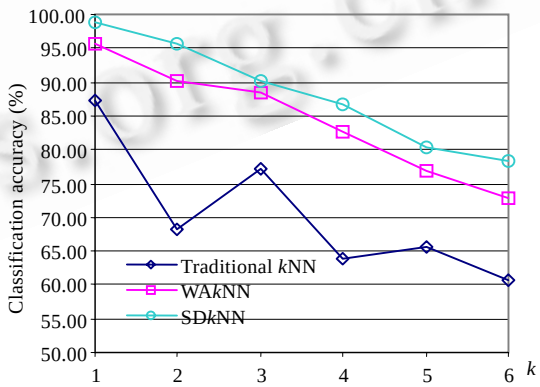

Fig.3 Classification accuracy comparison on nursery dataset

图 3 在 nursery 数据集上的分类准确率比较

从图 5,图 6 显示的实验结果可以看出,Traditional *k*NN 和 WA*k*NN 对这种数据不完整的情况相当敏感,准确率随着最细节数据缺失比率的增加有明显下降,而 SD*k*NN 可以较好地适应这种情况,分类准确率并未随着数据缺失比率的增加而大幅下降,这进一步验证了在存在多层次细节数据描述的情况下,基于语义距离的 SD*k*NN 要明显优于 WA*k*NN 和 Traditional *k*NN.

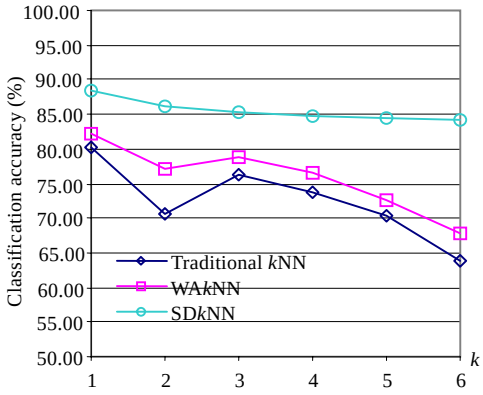

Fig.4 Classification accuracy comparison on insurance dataset

图 4 在 insurance 数据集上的分类准确率比较

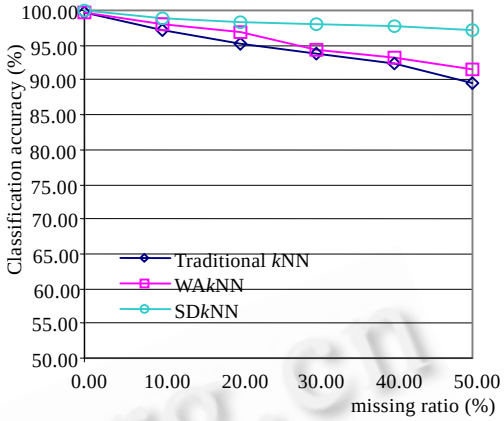

Fig.5 Classification accuracy comparison by missing ratio on mushroom dataset

图 5 在 mushroom 数据集上分类准确率随数据缺失率变化比较

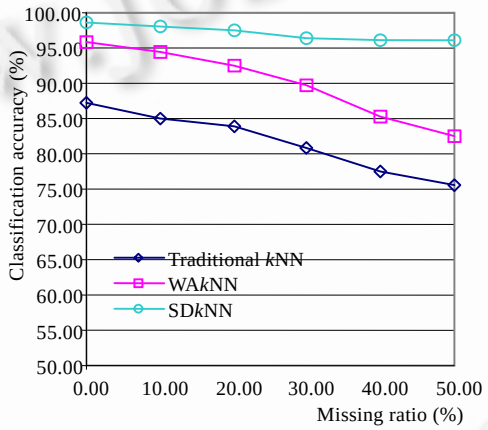

Fig.6 Classification accuracy comparison by missing ratio on nursery dataset

图 6 在 nursery 数据集上分类准确率随数据缺失率变化比较

4.3 讨 论

在实验中我们发现,式(9)~式(16)计算 $length(c \rightarrow parent(c))$时是通过概念在训练集的出现次数来决定的,在不太复杂的领域内,如保险领域中的企业财产保险范围内,文中的计算方法是合理的.而在相对复杂的领域,如整个财产保险领域,仅计算概念的出现次数就不足以代表 $length(c \rightarrow parent(c))$,还需要加上某些特征量的因素,如在评价不同险种的收入和风险状况时,就需要分别以累计保费和累计保额来代替出现次数.也可以采取“分而治之”的方法,将复杂领域分为相对简单的子领域,通过本文的方法计算出各个子领域的评价指标(如分类准确率),然后利用领域相关的特征量将它们集成起来.上述讨论的重点实际上涉及到如何融入领域特征问题.近年来,有关知识发现过程的领域相关性正逐渐得到重视[13],领域知识(如领域本体)与现有算法的融合是提高知识发现算法性能和实用性的重要手段.就我们所知,这类研究还处于起步阶段.本文为类似研究提供了新的思路.

5 结论及未来工作方向

本文首先分析了现有 *k*NN 方法在分析离散型属性和多抽象层次描述数据集时的不足,然后提出一种基于语义距离的最近邻分类方法 SD*k*NN,描述了一个形式化的本体,说明了如何利用本体建立概念间的语义联系及计算概念间的语义距离;利用语义距离改进了传统的最近邻分类算法.实验结果表明,该方法可有效提高最近邻

分类方法分析离散型属性的性能,针对多抽象层次描述的数据时,其效果尤为明显.我们还在国家“十五”科技攻关项目“财产保险防灾减损技术研究”中的财产保险损失预测模型中实现了该方法.在中国人民财产保险股份有限公司试点及推广的过程中,证明该方法可较好地适应保险数据按不同层次语义描述的特点,切实提高保险损失的预测效果,有效指导保险公司进行防灾减损工作,带来了广泛的社会效益和经济效益.

在本文中,语义距离只是被用来提高分类的准确率.我们将继续研究如何利用语义距离提高最近邻集合的搜索性能.

References:

[1] Li BL, Yu SW, Qin Lu. An improved *k*-nearest neighbor algorithm for text categorization. In: Sun MS, Yao TS, Yuan CF, eds. Proc. of the 20th Int'l Conf. on Computer Processing of Oriental Languages. Beijing: Tsinghua University Press, 2003.

[2] Han EH, Karypis G, Kumar V. Text categorization using weight adjusted k-nearest neighbor classification. In: Cheung D, Williams GJ, Li Q, eds. Proc. of the 5th Pacific-Asia Conf. Springer-Verlag, 2001. 53−65.

[3] Paredes R, Vidal E. A nearest neighbor weighted measure in classification problems. In: Sanfeliu A, Torres MI, eds. Proc. of the 7th SNRFAI. IOS Press, 2000. 44−50.

[4] Nechies R, Fikes R, Finin T, Gruber T, Patil R, Senator T, Swartout WR. Enabling technology for knowledge sharing. AI Magazine, 1991,12(3):36−56.

[5] Gruber TR. A translation approach to portable ontology specification. Knowledge Acquisition, 1993,5(2):199−220.

[6] Meersman R. An essay on the role and evolution of data(base) semantics. In: Meersman R, Mark L, eds. Proc. of the DataBase Application Semantics. London: Chapman Hall, 1997.

[7] Tastets MAR. Assessing semantic similarity among spatial entity classes [Ph.D. Thesis]. Department of Spatial Information Science and Engineering, University of Maine Orono, 2000.

[8] Hotho A, Maedche A, Staab S. Ontology-Based text document clustering. In: Klopotek MA, Wierzchon ST, Trojanowski K, eds. Proc. of the Conf. on Intelligent Information Systems. Zakopane: Springer-Verlag, 2003.

[9] Jiang J, Conrath D. Semantic similarity based on corpus statistics and lexical taxonomy. In: Proc. of the Int'l Conf. on Research in Computational Linguistics. 1997. 19−33.

[10] Han JW, Kamber M. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, 2001.

[11] Blake CL, Merz CJ. UCI repository of machine learning databases. Department of Information and Computer Science, University of California, 1998. http://www.ics.uci.edu/~mlearn/MLRepository.html

[12] Taylor MG, Stoffel K, Hendler JA. Ontology-Based induction of high level classification rules. In: Proc. of the ACM SIGMOD Data Mining and Knowledge Discovery Workshop. 1997.

[13] Kremer S, Smolnik S, Kolbe L. Towards knowledge discovery through context explication. In: Ralph H, Jr Sprague, eds. Proc. of the 37th Hawaii Int'l Conf.e on System Sciences. Hawaii: IEEE Computer Society, 2004.