计算机科学
COMPUTER SCIENCE

基于三角不等式判定和局部策略的高效邻域覆盖模型

陈于思，艾志华，张清华

引用本文

陈于思，艾志华，张清华. 基于三角不等式判定和局部策略的高效邻域覆盖模型[J]. 计算机科学, 2022, 49(5): 152-158.

CHEN Yu-si, AI Zhi-hua, ZHANG Qing-hua. Efficient Neighborhood Covering Model Based on Triangle Inequality Checkand Local Strategy[J]. Computer Science, 2022, 49(5): 152-158.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210300302

基于三角不等式判定和局部策略的高效邻域覆盖模型

陈于思　艾志华　张清华
重庆邮电大学计算智能重庆市重点实验室　重庆 400065

摘　要　邻域覆盖模型由于其原理简单以及对复杂数据具有较好的处理能力,在分类任务中得到了广泛应用。然而,邻域覆盖模型普遍存在运行效率较低的问题,且缺乏相关研究工作。为解决此问题,在传统邻域覆盖模型中引入距离间的三角不等式关系以提升构建邻域的效率,同时引入局部策略,定义了局部邻域覆盖以提升构建邻域覆盖的效率。为提升运行效率,从两个角度对传统邻域覆盖模型进行了改进,提出了基于三角不等式判定和局部策略的邻域覆盖模型(Neighborhood Covering Model based on Triangle Inequality Check and Local Strategy,TI-LNC)。此外,当前基于邻域覆盖模型的分类算法通常仅根据邻域中心以及邻域半径对样本进行分类,缺乏对邻域内样本信息的使用,从而影响了分类精度。为提高邻域覆盖模型的分类精度,增加了对邻域内样本信息的考虑,并基于TI-LNC设计了新的分类算法。在10个UCI数据集上的实验结果表明,所提模型能达到较高的运行效率以及较好的分类精度,具有一定的合理性及有效性。

关键词:邻域粗糙集;邻域覆盖模型;局部邻域覆盖;三角不等式判定
中图法分类号　TP391.9

Efficient Neighborhood Covering Model Based on Triangle Inequality Check and Local Strategy

CHEN Yu-si,AI Zhi-hua and ZHANG Qing-hua
Chongqing Key Laboratory of Computational Intelligence,Chongqing University of Posts and Telecommunications,Chongqing 400065,China

Abstract　Neighborhood covering model is widely used in classification tasks for its simple mechanism and ability to handle complex data. However,the neighborhood covering model has the problem of low efficiency and lack of related research work. To solve this problem,triangle inequality between distances is introduced to improve the efficiency of constructing neighborhood. Meanwhile,local neighborhood covering is defined. The local strategy is used to improve the efficiency of constructing neighborhood covering. In summary,to improve the efficiency,traditional neighborhood covering model is improved from two perspectives,and a neighborhood covering model based on triangle inequality check and local strategy (TI-LNC) is proposed. In addition,current classification algorithms based on neighborhood covering models only classify samples based on neighborhood centers and neighborhood radius,and ignore the sample information in neighborhoods,which affects classification accuracy. To improve the classification accuracy of the neighborhood covering model,the consideration of sample information in the neighborhood is added,and a new classification algorithm based on TI-LNC is designed. The experimental results on 10 UCI data sets show that the proposed model which is reasonable and effective can achieve higher efficiency and better classification accuracy.

Keywords　Neighborhood rough set,Neighborhood covering model,Local neighborhood covering,Triangle inequality check

1　引言

通过将邻域关系引入经典粗糙集模型[1],邻域粗糙集[2-3]被提出、以拓展经典粗糙集的应用场景。经典粗糙集中使用等价关系来构建等价类的方式更多地被应用于由字符型特征刻画的数据中。邻域粗糙集中的邻域信息粒由样本间的邻域关系形成,可用于更好地处理由数值型特征刻画的数据。近年来,越来越多的学者提出基于邻域粗糙集的创新理论[4-6]。在不同的应用场景中,各种基于邻域粗糙集模型的分类算法以及特征选择算法被陆续提出[7-13]。邻域覆盖模型通过生成一组邻域内样本与标签的邻域以近似数据分布,因其原理简单和对复杂数据的鲁棒性,常被应用于数据挖掘任务中。2011年,Du等[14]提出邻域覆盖约简,用于对初始邻域覆盖进行约简,得到一个更加精简的邻域覆盖,应用于分类任务中。此后,更多的邻域覆盖模型相继被提出。Zhu等[15]提出一种基于随机特征选择与邻域覆盖的集成学习方法。Zhang等[16]结合相似性度量提出一种用于处理字符型数据的邻域覆盖模型并应用于分类中。Yue等[17]提出一种基于三分法的邻域覆盖约简算法,以更好地在噪声环境下完成邻域覆盖的构建与邻域覆盖的约简。随后,Yue等[18]将模糊集理论

到稿日期:2021-03-31　返修日期:2021-10-24
基金项目:国家自然科学基金(61876201)
This work was supported by the National Natural Science Foundation of China(61876201).
通信作者:陈于思(576131318@qq.com)

引入邻域覆盖模型，提出模糊邻域覆盖，并结合三支决策理论提出一种三支分类方法。进一步地，Yue 等[19]在模糊邻域覆盖的基础上引入阴影集理论，提出阴影邻域，提升了模型在处理具有不确定性数据时的分类效果。然而，现有研究主要聚焦于提升模型的分类能力，缺乏对邻域覆盖模型运行效率较低这一问题的研究。模型运行效率较低主要有两个方面的原因：第一，为样本构建邻域时，会涉及大量的距离计算，这是一个耗时的过程；第二，邻域覆盖的构建需要为训练集中的每个样本构建邻域，过程中会产生大量冗余邻域，并放到下一阶段进行约简，这个过程是低效的。

为提升邻域覆盖模型的运行效率，本文分别从邻域构建以及邻域覆盖构建两个方面对邻域覆盖模型进行改进。首先，在为样本构建邻域时，引入样本间距离的三角不等式关系。在文献[20]以及文献[21]中，基于距离间的三角不等式关系的近邻搜索策略被引入 KNN 以提升搜索近邻的效率。受文献[20-22]的启发，在本文中，首先，三角不等式判定被引入邻域构建中，以减少一些由不必要的距离计算带来的计算开销，提升邻域的构建速度。其次，本文定义了局部邻域覆盖并设计了构建局部邻域覆盖的具体策略。局部策略可以减少初始邻域覆盖中邻域的数量，从而提升构建邻域覆盖的效率。此外，基于所提邻域覆盖模型，本文进一步设计了新的分类算法，该算法引入了邻域内部样本的信息，使分类精度得到进一步提升。

本文第 2 节简要介绍了邻域覆盖约简；第 3 节提出了基于三角不等式判定和局部策略的邻域覆盖模型；第 4 节通过实验验证了所提算法的有效性及优势；最后总结全文。

2 相关基本概念

为了使本文思想更易于理解和表述，本节将介绍一些与本文研究相关的基本概念和定义。

定义 1(邻域)[3] 设 $U=\{x_1,x_2,\cdots,x_n\}$ 为样本集合，$O(x_i)=\{x_j\in U|\Delta(x_i,x_j)\leqslant r_i\}$ 为样本 x_i 的邻域，其中 r_i 为半径阈值，$\Delta(\cdot)$ 为距离函数。

定义 2(邻域覆盖)[14] 设 $U=\{x_1,x_2,\cdots,x_n\}$ 为样本集，邻域集合 $O_U=\{O(x_i)|x_i\in U\}$ 形成了论域 U 上的一个覆盖，$C=\langle U,O_U\rangle$ 为邻域覆盖近似空间。

初始的邻域覆盖为每个样本构建了邻域以近似样本空间，但是邻域之间可能存在相互重叠的现象，这种重叠的邻域有可能是冗余的。为得到一个更为精简的邻域覆盖，文献[8]定义了两种邻域覆盖约简，如定义 3 和定义 4 所示。

定义 3(邻域覆盖约简)[14] 设 $C=\langle U,O_U\rangle$ 为一个邻域覆盖近似空间，$\forall x_i\in U$，若 $\bigcup\limits_{x_j\in U-\{x_i\}}O(x_j)=U$，则 $O(x_i)$ 为可约简的，否则为不可约简的。若 C 中的所有邻域都为不可约简的，则 C 为不可约简的。

定义 4(相对邻域覆盖约简)[14] 设有邻域覆盖近似空间 $C=\langle U,O_U\rangle$，$X\subseteq U$ 为一个样本集合，且 $O(x_i)\in O_U$ 为一个邻域，有 $O(x_j)\in O_U$，$i\neq j$，使 $O(x_i)\subseteq O(x_j)\subseteq X$，则 $O(x_i)$ 对于 X 为相对可约简的邻域，否则 $O(x_i)$ 为相对不可约简的。若 C 中的所有邻域都为相对不可约简的，则 C 为相对不可约简的。

3 基于三角不等式判定与局部策略的邻域覆盖模型

传统邻域覆盖模型在面对维度较高的数据时显得较为低效。在构建邻域方面，为样本搜索近邻样本以求得半径阈值的过程中，需要多次通过距离函数计算样本间距离。距离的计算会在每个属性上产生计算开销。维度较高时，距离的计算比较耗时。减少邻域覆盖模型在构建邻域时的距离计算，提升邻域构建速度为本节的第一个目标。在邻域覆盖的构建方面，为得到精简的邻域覆盖，传统模型为每个样本构建邻域，然后通过邻域覆盖约简去除冗余邻域。生成所有邻域后再将冗余邻域约简的策略，不管是在邻域覆盖构建阶段还是约简阶段，都是低效的。在邻域覆盖构建阶段减少冗余邻域的产生而不是将所有冗余邻域都留到邻域覆盖约简阶段是本节的第二个目标。本节将结合三角不等式判定与局部策略分别实现上述两个目标。

3.1 基于三角不等式判定的邻域构建策略

本节将介绍一种基于三角不等式判定的邻域构建策略。

设有样本 x_i，r_i 为 x_i 的邻域半径阈值，x_i 的邻域为 $O(x_i)=\{x_j\in U|\Delta(x_i,x_j)<r_i\}$。为使处于同一邻域 $O(x_i)$ 的样本拥有相同的标签，可通过为 x_i 搜索最近邻异类样本来确定对应邻域的半径阈值 r_i。对于样本 x_i，$NM(x_i)$ 为其最近邻的异类样本(Nearest Miss)。邻域半径 r_i 可通过 $r_i=\varphi\times\Delta(x_i,NM(x_i))$ 计算得到，其中 $\varphi\in[0,1)$ 为常数参数，用于控制邻域的大小。

在为样本找寻最近邻异类样本的过程中，需要计算从样本到所有异类样本的距离。为减少距离计算带来的计算负担，模型引入了距离间的三角不等式关系。在计算距离前，模型先进行三角不等式判定，若满足条件，则无需计算距离，这样可以减少不必要的距离计算。为构建距离间的三角不等式关系，定义 5 定义了基于属性均值的类中心。

定义 5(类中心) 设 U 为样本空间，$D=\{d\}$ 为决策属性集，$U/D=\{X_1,X_2,\cdots,X_m\}$ 为根据决策属性对样本空间的划分，$X_l\in\{X_1,X_2,\cdots,X_m\}$ 为一个类，m 为类别数，类中心 $c(X_l)$ 定义为：

$$c(X_l)=\frac{1}{|X_l|}\sum_{x_i\in X_l}x_i \tag{1}$$

其中，$|X_l|$ 表示集合 X_l 中对象的数量。

邻域的构建主要分为两个步骤：1)邻域半径的计算；2)邻域的生成。

邻域半径的计算如图 1 所示，设样本 $x_i\in X_l$，$x_j\in X_l$，$c(X_l)$，$c(X_l)$ 分别为 X_l，X_l 的类中心，$d_{\min}$ 为 x_i 到当前所搜索到的最近的异类样本的距离，其中 X_l，$X_l\in\{X_1,X_2,\cdots,X_m\}$。通过三角不等式 $\Delta(x_i,c(X_l))<\Delta(x_i,x_j)+\Delta(x_j,c(X_l))$，可进行判定以减少距离计算。因此，给出如下两条判断策略。

(1)若 $d_{\min}\leqslant abs(\Delta(x_i,c(X_l))-\Delta(x_j,c(X_l)))$，则通过距离间的三角不等式判定可知必然存在不等式 $d_{\min}\leqslant abs(\Delta(x_i,c(X_l))-\Delta(x_j,c(X_l)))<\Delta(x_i,x_j)$，即必然有 $d_{\min}<\Delta(x_i,x_j)$，故无需计算 $\Delta(x_i,x_j)$，$abs(\cdot)$ 为绝对值。

(2)若 $d_{\min}>abs(\Delta(x_i,c(X_l))-\Delta(x_j,c(X_l)))$，则无法

通过三角不等式判定省略距离的计算，故仍需计算 $\Delta(x_i, x_j)$。若 $\Delta(x_i, x_j) < d_{\min}$，则将 $d_{\min}$ 的值替换为 $\Delta(x_i, x_j)$。若 $\Delta(x_i, x_j) \geqslant d_{\min}$ 则不做操作。

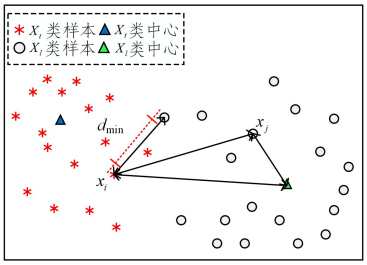

图 1 邻域半径的计算

Fig. 1 Computation of neighborhood radius

通过上述方式为 x_i 遍历所有异类样本，可找到最近异类样本 $NM(x_i)$ 以及到最近异类样本的距离 $\Delta(x_i, NM(x_i)) = d_{\min}$。通过 $r_i = \varphi \times \Delta(x_i, NM(x_i))$ 可计算样本 x_i 的邻域半径阈值 r_i。

通过半径阈值可以搜索同类样本以生成邻域。图 2 所示为 x_i 搜索其邻域内对象时，将邻域半径阈值 r_i 作为三角不等式判定所用到的阈值。设有 $x_z \in X_t$，其中 $z \neq i$。给出如下两条判断策略。

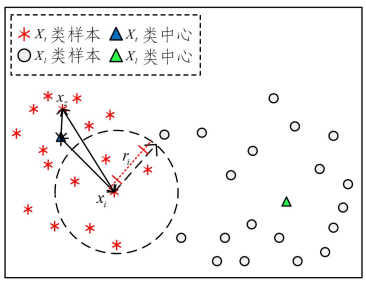

图 2 邻域生成

Fig. 2 Generation of neighborhood

(1)如果 $r_i \leqslant abs(\Delta(x_i, c(X_t)) - \Delta(x_z, c(X_t)))$，则通过距离间的三角不等式判定可知必然会存在不等式 $r_i \leqslant abs(\Delta(x_i, c(X_t)) - \Delta(x_z, c(X_t))) < \Delta(x_i, x_z)$，即必有不等式 $r_i < \Delta(x_i, x_z)$，故无需计算 $\Delta(x_i, x_z)$，可判断 x_z 不属于邻域 $O(x_i)$。

(2)若 $r_i > abs(\Delta(x_i, c(X_t)) - \Delta(x_z, c(X_t)))$，则仍需计算 $\Delta(x_i, x_z)$。若 $\Delta(x_i, x_z) \leqslant r_i$，则样本 x_z 属于邻域 $O(x_i)$，反之则不属于。

遍历完 x_i 的所有同类样本，即可完成 x_i 邻域的生成。

基于三角不等式判定的邻域构建策略需要计算每个样本到各个类中心的距离，故搜索前会初始化矩阵 $\boldsymbol{M}_c$ 与 $\boldsymbol{M}_s$，用于距离的储存，其中 $\boldsymbol{M}_c$ 用于储存样本到各类中心的距离，$\boldsymbol{M}_s$ 用于储存构建邻域过程中所计算的样本间距离。为方便叙述，定义 X_i^{heter} 为样本 x_i 的所有异类样本的集合，X_i^{homo} 为样本 x_i 的所有同类样本的集合。$idx(x)$ 为 x 所处类 X 的索引。基于三角不等式判定的邻域构建算法如算法 1 所示。

算法 1 基于三角不等式判定的邻域构建算法

输入：样本集合 U，待构建邻域样本集合 X⊆U，半径参数 φ

输出：一组邻域集合 O

```
1. for x_i in X do
2. 随机选择样本 x_r∈U
3. 计算 d_min=Δ(x_i,x_r)
4.    找到 X_i^heter,X_i^homo
5.    for x_j in X_i^heter do
6.       if d_min≤abs(Δ(x_i,c(X_idx(x_j)))−Δ(x_j,c(X_idx(x_j)))) then
7.          continue
8.       else if Δ(x_i,x_j)<d_min then
9.          d_min=Δ(x_i,x_j)
10.      end if
11.   end for
12.   r_i=ω×Δ(x_i,NM(x_i))
13.   for x_z in X_i^homo do
14.      if r_i≤abs(Δ(x_i,c(X_idx(x_z)))−Δ(x_z,c(X_idx(x_z)))) then
15.         continue
16.      else if Δ(x_i,x_z)<r_i then
20.         将 x_z 放入 O(x_i)中
21.      end if
22.   end for
23.   将 O(x_i)放入 O 中
22. end for
23. 返回 O
```

3.2 局部邻域覆盖

为了在初始邻域覆盖构建阶段减少邻域的重叠，而不是在邻域覆盖约简阶段再对重叠邻域进行约简，本节中定义了局部邻域覆盖，如定义 6 所示。

定义 6(局部邻域覆盖) 设样本集合 $U = \{x_1, x_2, \cdots, x_n\}$，有 $\bigcup_{x_i \in X_{\text{local}}} O(x_i) = U$，$X_{\text{local}} \subset U$，则存在邻域集合 $O_{\text{local}} = \{O(x_i) \mid x_i \in X_{\text{local}}\}$，形成论域 U 上的一个局部邻域覆盖。$C_{\text{local}} = \langle U, O_{\text{local}} \rangle$ 表示局部邻域覆盖近似空间。若 $X_{\text{local}} = U$，则局部邻域覆盖近似空间 C_{local} 退化为经典邻域覆盖近似空间[8]。

局部邻域覆盖使用更少的邻域覆盖样本空间，选择哪些样本为其构建邻域以覆盖样本空间是本节需要解决的另一个问题。为合理地选择样本以构建局部邻域覆盖，本节提出了一种基于内邻域的局部邻域覆盖构建方法。离邻域中心相对较近的样本与其对应邻域的邻域中心有较为相似的空间位置，可能生成相似的邻域。故在邻域中生成内邻域，以便在选择下一个样本为其构建邻域时跳过内邻域中的样本。通过上述策略，可以为初始邻域覆盖减少冗余邻域，更快地完成初始邻域覆盖的构建。且更少的冗余邻域进入邻域覆盖约简阶段可以更快地完成邻域覆盖约简。内邻域定义如定义 7 所示。

定义 7(内邻域) 设 $O(x_i)$ 为样本 x_i 的邻域，r_i^{inner} 为内邻域的半径，样本 x_i 的内邻域 $O_c(x_i)$ 定义如下：

$$O_c(x_i) = \{x_j \in O(x_i) \mid \Delta(x_i, x_j) \leqslant r_i^{\text{inner}}\} \tag{2}$$

其中，$r_i^{\text{inner}} = \dfrac{1}{|O(x_i)|} \sum_{x_j \in O(x_i)} \Delta(x_i, x_j)$。

在模型中，r_i^{inner} 通过计算邻域中样本到邻域中心距离的均值得到，考虑均值是为了适应邻域内的数据分布。若邻域内无除邻域中心以外的其他样本，则 $r_i^{\text{inner}} = 0$。为了更清楚地描述构建局部邻域覆盖的具体策略，我们在下文中分别给出了内邻域样本集和非内邻域样本集的定义，如定义 8 和定义 9 所示。

定义 8（内邻域样本集） 设已有邻域集合为 $O_{cur}=\{O(x_i)|x_i\in X_{cur}, X_{cur}\subseteq U\}$，$X_{cur}$ 为已为其生成了邻域的样本，内邻域样本集 X_{inner} 定义如下：

$$X_{inner}=\bigcup_{x_i\in X_{cur}} O_c(x_i) \tag{3}$$

定义 9（非内邻域样本集） 设 X_{inner} 为内邻域样本集，非内邻域样本集 X_{outer} 定义如下：

$$X_{outer}=U-X_{inner} \tag{4}$$

局部邻域覆盖构建策略：在构建邻域 $O(x_i)$ 后，为其生成内邻域 $O_c(x_i)$，并更新内邻域样本集 $X_{inner}=X_{inner}\cup O_c(x_i)$。在下次构建邻域时跳过 X_{inner} 中的样本，即选择样本为其构建邻域时，从 X_{outer} 中选择。由于算法是一个渐进的过程，因此需要确定邻域生成的顺序。在模型中，每次选择用来为其构建邻域的样本为当前离其对应类中心距离最小的非内邻域样本。通过这种方式可以避免离群点被优先建立邻域。对应的详细算法如算法 2 所示。TI-LNC 模型通过结合基于三角不等式判定的邻域构建算法与局部邻域覆盖约简算法得到。

算法 2 局部邻域覆盖约简算法

输入：训练数据集 U

输出：规则集合 R

1. 初始化 $X_w=U, X_{cur}=\emptyset$
2. while $X_w=\emptyset$ do
3. 从 X_w 中找出离对应类中心最近的样本 x_i
4. 根据算法 1 计算 $O(x_i)$
5. 将 $O(x_i)$ 放入 O_{local} 中
6. 将 x_i 放入 X_{cur}
7. 计算 $O_c(x_i)$
8. 获取 $X_{inner}=\bigcup_{x_i\in X_{cur}} O_c(x_i)$
9. 得到 $X_{outer}=U-X_{inner}$
10. 将 X_w 中的样本更新为 X_{outer} 中的样本
11. end
12. while($O_{local}\neq\emptyset$)do
13. 从 O_{local} 中选择一个覆盖样本最多的邻域 $O(x_i)$
14. 规则($x_i, r_i, r_i^{inner}, L(x_i)$)放入 R 中
15. if $O(x_j)\subseteq O(x_i)$ then
16. 将 $O(x_j)$ 移除
17. end
18. 在 R 中，根据邻域内样本数降序排列，返回前 k 个邻域

使用 TI-LNC 可以构建一个比传统的初始邻域覆盖更精简的初始邻域覆盖，因为所需构建的邻域更少，算法相对而言更加高效。更少的初始邻域使得 TI-LNC 在邻域覆盖约简阶段也更加高效。此外，由于在构建邻域时利用三角不等式进行判定，每次判定成功都会为模型节省距离的计算开销。综上所述，相比 NCR，TI-LNC 构建邻域覆盖的过程更加高效。

3.3 基于 TI-LNC 的分类策略

基于邻域覆盖模型的分类算法通过使用一组由标签标注的邻域对待预测样本进行分类，以完成分类任务。在 NCR[14] 中，分类策略的依据为邻域半径和邻域中心，对邻域内部的数据分布信息没有进行刻画和使用。本文提出的模型 TI-LNC 中加入了内邻域以刻画邻域内的数据分布。在本节中，所设计的分类策略会结合邻域中心、内邻域半径和邻域半径所提供的信息进行分类。在介绍具体策略前，定义样本到内邻域的距离，如定义 10 所示。

定义 10（样本到内邻域的距离） 设 O_c^i 为样本 x_i 的内邻域，r_i^{inner} 为内邻域的半径，样本 y 到内邻域 O_c^i 的距离 $\Delta(O_c^i, y)$ 定义如下：

$$\Delta(O_c^i,y)=\begin{cases}\Delta(x_i,y)-r_i^{inner}, & \text{if } \Delta(x_i,y)-r_i^{inner}>0\\ 0, & \text{otherwise}\end{cases} \tag{5}$$

设 y 为待分类样本，$O(x_i)$ 为训练样本 x_i 的邻域，样本定位的表示如下：

$\Delta(x_i,y)<r_i\Rightarrow y\in N_i$，

$\Delta(x_i,y)<r_i^{inner}\Rightarrow y\in CN_i$，

其中，N_i 为属于邻域 $O(x_i)$ 的测试样本的集合，CN_i 为处于样本 x_i 的内邻域 $O_c(x_i)$ 中的测试样本的集合。

综上所述，邻域索引集合可表示如下：

$N(y)=\{i|y\in N_i\}$

$CN(y)=\{i|y\in CN_i\}$

其中，$N(y)$ 为样本 y 所处邻域的索引集合，$CN(y)$ 为样本 y 所处内邻域的索引集合。另外，为了方便描述，设 $Index(O_r)$ 为规则集中所有邻域的索引集合。

分类时，对于不同的定位，测试样本有可能处于几种不同的分类条件中。对处于不同分类条件的测试样本，模型会采用不同的策略以完成分类任务。设 y 为待分类样本，$L(\cdot)$ 为样本的真实标签，$P(\cdot)$ 为样本的预测标签，具体分类策略设计如下：

(1)若 y 仅处于某一个内邻域中，即 $|CN(y)|=1$，将该邻域的标签赋予待分类样本 y，$P(y)=L(x_i), i\in CN(y)$；

(2)若 y 处于多个内邻域中，即 $|CN(y)|>1$，将邻域中心与 y 距离最近的邻域的标签赋予待分类样本 y，$P(y)=\arg\min_{L(x_i)}\Delta(x_i,y), i\in CN(y)$；

(3)若 y 仅处于某一个邻域中，且不处于任何内邻域中，即 $|CN(y)|=0, |N(y)|=1$，将该邻域的标签赋予待分类样本 y，$P(y)=L(x_i), i\in N(y)$；

(4)若 y 处于多个邻域中，且不处于任何内邻域中，即 $|CN(y)|=0, |N(y)|>1$，将与 y 距离最近的内邻域的邻域标签赋予待分类样本 y，$P(y)=\arg\min_{L(x_i)}\Delta(O_c^i,y), i\in N(y)$；

(5)若 y 不处于任何邻域中，即 $|CN(y)|=0, |N(y)|=0$，将与 y 距离最近的内邻域的标签赋予待分类样本 y，$P(y)=\arg\min_{L(x_i)}\Delta(O_c^i,y), i\in Index(O_r)$。

在分类阶段，TI-LNC 与 NCR 类似，时间复杂度为 $O(k)$，其中 k 为用于分类的邻域数。与传统的基于最近邻的算法（如 KNN）的时间复杂度 $O(n)$ 相比，因为 $k\ll n$，所以 TI-LNC 的时间复杂度更低，在分类时更高效。具体算法如算法 3 所示。

算法 3 基于 TI-LNC 的分类算法

输入：测试样本 y，规则集合 R

输出：样本 y 的标签 P(y)

1. 初始化 $N(y)=\emptyset, CN(y)=\emptyset$
2. 计算样本 y 到 R 中邻域的距离
3. 更新 N(y)，CN(y)
4. if $|CN(y)|=1$ then
5. $P(y)=L(x_i), i\in CN(y)$
6. else if $|CN(y)|>1$ then
7. $P(y)=\text{argmin}_{L(x_i)}\Delta(x_i,y), i\in CN(y)$
8. else if $|CN(y)|=0$ and $|N(y)|=1$ then

9. P(y)=L(x_i), i∈N(y)
10. else if |CN(y)|=0 and |N(y)|>1 then
11. $P(y)=\underset{L(x_i)}{\operatorname{argmin}}\Delta(O_t^i,y), i\in N(y)$
12. else if |CN(y)|=0 and |N(y)|=0 then
13. $P(y)=\underset{L(x_i)}{\operatorname{argmin}}\Delta(O_t^i,y), i\in \mathrm{Index}(O_r)$
14. end if
15. 返回 P(y)

4 实验分析

为进一步阐述所提 TI-LNC 模型的优势，分别在 10 个 UCI 数据集上进行实验，数据集详细信息如表 1 所列。用于实验对比的算法有 NCR，TNCR、朴素贝叶斯(NB)、k 最近邻算法(KNN)、决策树(CART)以及线性核支持向量机(LSVM)。实验采用五折交叉验证。实验环境为 Window10 操作系统，8 GB 内存以及 3.30 GHz 主频。编程语言采用 Python。实验分为 3 部分进行分析。第一部分为参数 φ 与 TI-LNC 模型的分类精度以及运行时间之间的关系分析，第二部分为模型的分类效果分析，第三部分为模型的效率分析。

表 1 实验数据集

Table 1 Data sets

数据集	属性个数	样本个数	类别个数
Pima	8	768	2
Parkinson	754	756	2
WDBC	31	569	2
Ionosphere	34	351	2
Iris	4	150	3
Sonar	61	208	2
Breast-w	10	699	2
Blood	4	748	2
German	24	1 000	2
Mammographic	6	961	2

在第一部分的实验中，模型在不同参数值下进行自对比实验分析。在第二部分的实验中，除与 NCR 以及 TNCR 对比分类精度外，与经典机器学习算法 NB，KNN，CART 以及 LSVM 也进行了分类精度指标上的对比，以进一步验证所提模型具有较好的分类能力。在第三部分的实验中，通过 TI-LNC 与 NCR 以及 TNCR 在邻域覆盖的整个构建过程中所用运行时间的对比，验证了 TI-LNC 相比相关经典模型在邻域覆盖构建阶段的高效性。实验中样本间的距离度量均采用欧氏距离[23-24]。

4.1 评价指标

本文采用广泛使用的分类精度评价指标来评估模型的分类能力。给定一个数据集，设 P 表示正例的个数，N 表示负例的个数，TP 表示被正确预测为正例的个数，FP 表示被错误预测为正例的个数，TN 表示被正确预测为负例的个数，FN 表示被错误预测为负例的个数，那么分类精度指所有被正确分类的样本占总样本的比例，即：

$$Accuracy=\frac{TP+TN}{P+N} \tag{6}$$

4.2 参数分析

参数 φ 对于模型 TI-LNC 的运行过程较为重要。参数 φ 用于控制模型 TI-LNC 中每次构建的邻域的大小，同时也影响着邻域覆盖中邻域的个数。我们设置步长搜索 TI-LNC 的阈值以控制邻域的大小，在本节中，步长被设置为 0.05 以遍历 0～1 之间的参数值，并以分类精度以及运行时间作为评价指标。实验均采用五折交叉验证方法。实验结果图 3 和图 4 所示。实验结果表明，在参数合适的条件下，TI-LNC 在运行效率与分类精度上通常都能达到较好的效果。从运行效率的角度看，随着参数 φ 的值增大，运行时间呈下降趋势。若每次生成邻域的半径较大，则有更多的样本可以进入邻域中，同时更多的样本就有机会进入内邻域中。此时，局部策略可发挥较大作用。当参数 φ 的值逐渐变小时，局部策略发挥的作用逐渐变小。此时，主要由基于三角不等式判定的加速策略为模型提供效率上的提升。从分类精度的角度看，当参数 φ 的值较小时，在多个数据集中分类精度会大幅下滑，这是因为参数 φ 的值较小时，邻域较小，测试样本很难落入任何邻域中，因而只能通过寻找最近邻邻域中心来进行分类。当参数值位于[0.95,1)时，模型的分类精度更稳定。综合考虑，推荐参数 φ 的取值范围为[0.95,1)。为了实验的公平性，在第二部分与第三部分的实验中，参数 φ 的值均设置为 0.99。

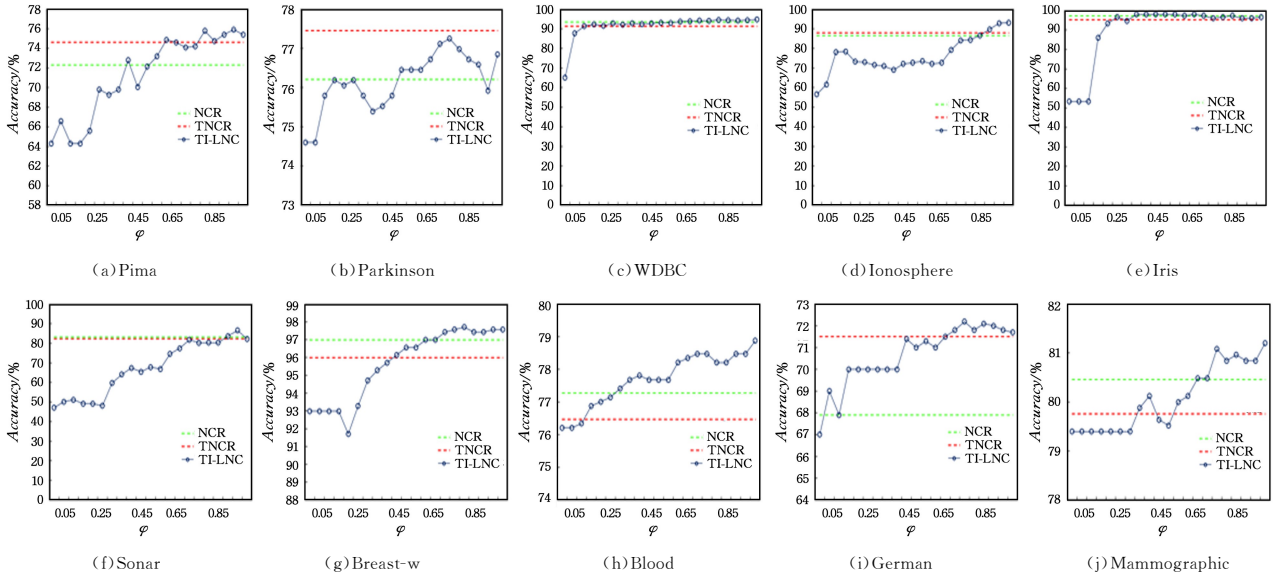

图 3 使用不同参数值 φ 在 10 个数据集上的分类精度

Fig. 3 Classification accuracy with different φ in 10 data sets

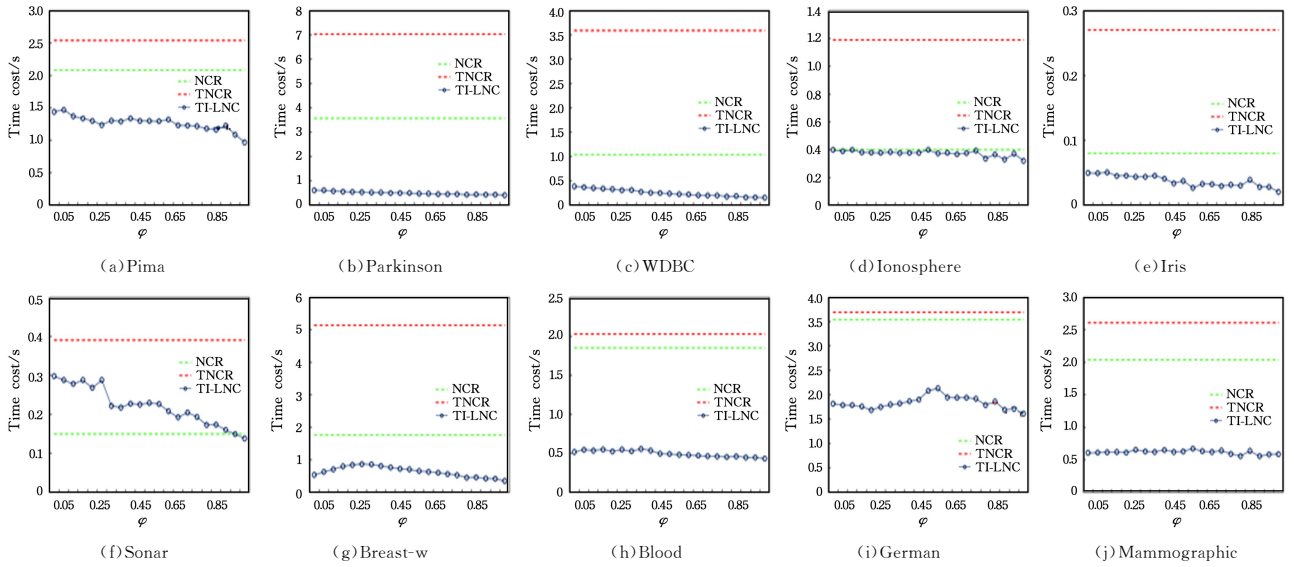

图 4　使用不同参数值 φ 在 10 个数据集上的运行时间

Fig. 4　Time cost with different φ in 10 data sets

4.3　分类效果分析

表 2 所列为分类精度指标下，7 个模型在 10 个数据集上的实验结果，其中包括模型在单一数据集上的分类精度对比以及在全部 10 个数据集上的平均分类精度的对比。

表 2　分类精度对比

Table 2　Comparison of classification accuracy

（单位：%）

数据集	LSVM	CART	10-NN	NB	NCR	TNCR	TI-LNC
Pima	**77.09**	73.17	74.23	75.65	72.29	74.62	75.40
Parkinson	79.79	79.89	**83.60**	78.44	76.21	77.45	76.85
WDBC	95.25	93.31	**96.30**	93.49	93.49	91.38	94.72
Ionosphere	87.75	89.73	81.77	88.02	86.49	87.89	**93.16**
Iris	**97.33**	95.33	95.33	95.33	**97.33**	95.33	96.67
Sonar	79.35	67.36	79.83	68.29	**83.17**	82.20	82.20
Breast-w	96.14	93.42	96.85	95.99	96.99	95.99	**97.57**
Blood	76.74	77.15	77.55	75.40	77.27	76.47	**78.88**
German	**75.50**	72.90	73.30	71.69	67.90	71.50	71.70
Mammographic	81.33	**82.28**	80.36	80.96	80.46	79.76	81.20
Avg	84.62	82.45	83.91	82.33	83.16	83.25	**84.84**

从表 2 中可以看出，TI-LNC 在分类精度指标下，在 10 个数据集中的 3 个数据集上获得了最优分类精度，并且 TI-LNC 的平均分类精度最优。相比 NCR，TI-LNC 在 10 个数据集中的 8 个数据集上分类效果都得到了提升，平均分类精度提升 1.68%。与 TNCR 相比，TI-LNC 也拥有更好的分类效果。在 10 个数据集中，TI-LNC 在其中 8 个数据集上的类效果得到了提升，平均分类精度提升 1.59%。与经典机器学习算法 LSVM，CART，NB 以及 KNN 相比，TI-LNC 也拥有更好的分类效果。在实验的 10 个数据集上，相比 LSVM，TI-LNC 的平均分类精度提高了 0.22%。相比 CART，TI-LNC 的平均分类精度提高了 2.39%。对于 KNN，本文中用于对比的是 10-NN 算法。相比 10-NN，TI-LNC 的平均分类精度提高了 0.93%。相比 NB，TI-LNC 的平均分类精度提高了 2.51%。综上所述，TI-LNC 确保了较好的分类精度。

4.4　效率分析

表 3 所列为构建邻域覆盖的运行时间的对比实验结果。与 NCR 和 TNCR 相比，TI-LNC 完成模型训练所使用的时间更短。在 Pima，Parkinson，WDBC，Iris，Breast-w，Blood，German 以及 Mammographic 数据集上，相比 NCR，TI-LNC 在运行效率上取得了显著的提升，运行时间分别缩短了 52.88%，88.24%，84.62，75%，80.11%，76.22%，54.24%，71.08%。即在所涉及的 10 个数据集中 8 个数据集的运行效率上取得了显著的提升。在另外两个数据集 Ionosphere 以及 Sonar 上，TI-LNC 也使用更短的时间完成了训练，但训练时间降低幅度较小，分别减少了 20%和 6.7%。相比 NCR，TI-LNC 的平均运行时间减少了 69.70%。相比 TNCR，TI-LNC 在所有数据集上的运行效率均取得了较为显著的提升，在 Pima，Parkinson，WDBC，Ionosphere，Iris，Sonar，Breast-w，Blood，German 以及 Mammographic 上的运行时间分别减少了 61.42%，94.03%，95.54%，73.10%，92.59%，64.1%，93.19%，78.33%，56.1%，77.39%。相比 TNCR，TI-LNC 的平均运行时间减少了 82.46%。由实验结果可知，TI-LNC 在运行效率上有较为明显的优势。

表 3　运行时间对比

Table 3　Comparison of time cost

（单位：s）

数据集	NCR	TNCR	TI-LNC
Pima	2.08	2.54	**0.98**
Parkinson	3.57	7.03	**0.42**
WDBC	1.04	3.59	**0.16**
Ionosphere	0.4	1.19	**0.32**
Iris	0.08	0.27	**0.02**
Sonar	0.15	0.39	**0.14**
Breast-w	1.76	5.14	**0.35**
Blood	1.85	2.03	**0.44**
German	3.54	3.69	**1.62**
Mammographic	2.04	2.61	**0.59**
Avg	1.65	2.85	**0.50**

结束语　传统的邻域覆盖模型计算复杂度较高，在模型训练阶段需要消耗较多时间。针对此问题，本文提出了一种

高效的邻域覆盖模型。该模型引入了距离间的三角不等式关系，定义了局部邻域覆盖，分别在邻域构建以及邻域覆盖构建阶段提升了模型的运行效率。此外，本文基于所提出的邻域覆盖模型设计了新的分类算法，提升了分类精度。在 UCI 数据集上的实验结果表明，所提模型具有一定的有效性和合理性，能够在提升模型运行效率的同时保证分类效果。特征选择对于分类任务来说至关重要，故基于邻域覆盖模型设计高效的特征选择算法是未来的一个研究重点。

参 考 文 献

[1] PAWLAK Z. Rough Set[J]. International Journal of Computer and Information Sciences, 1982, 11(5): 341-356.

[2] HU Q H, YU D R, XIE Z X. Neighborhood Classifiers [J]. Expert Systems with Applications, 2008, 34(2): 866-876.

[3] LIN T Y. Granular Computing on Binary Relations I : Data mining and Neighborhood Systems[J]. Rough Sets in Knowledge Discovery, 1998, 1: 107-121.

[4] WANG Q, QIAN Y H, LIANG X Y, et al. Local Neighborhood Rough Set[J]. Knowledge-Based Systems, 2018, 153: 53-64.

[5] HU M, TSANG E C C, GUO Y T, et al. A Novel Approach to Attribute Reduction Based on Weighted Neighborhood Rough Sets[J]. Knowledge-Based Systems, 2021, 220(5): 106908.

[6] CHEN Y M, XUE Y, MA Y, et al. Measures of Uncertainty for Neighborhood Rough Sets[J]. Knowledge-Based Systems, 2017, 120: 226-235.

[7] XU S P, YANG X B, YU H L, et al. Neighborhood Collaborative Representation Based Classification Method [J]. Computer Science, 2017, 44(9): 234-238.

[8] HU Q H, YU D R, LIU J F, et al. Neighborhood Rough Set Based Heterogeneous Feature Subset Selection[J]. Information Sciences, 2008, 178(18): 3577-3594.

[9] HU Q H, PEDRYCZ W, YU D R, et al. Selecting Discrete and Continuous Features Based on Neighborhood Decision Error Minimization[J]. IEEE Transactions on Systems, Man, and Cybernetics, 2009, 40(1): 137-150.

[10] YAO P, LU Y H. Neighborhood Rough Set and SVM Based Hybrid Credit Scoring Classifier[J]. Expert Systems with Applications, 2011, 38(9): 11300-11304.

[11] CHEN H M, LI T R, CAI Y, et al. Parallel Attribute Reduction in Dominance-based Neighborhood Rough Set[J]. Information Sciences, 2016, 373: 351-368.

[12] XIA S Y, ZHANG H, LI W H, et al. GBNRS: A Novel Rough Set Algorithm for Fast Adaptive Attribute Reduction in Classification[J/OL]. IEEE Transactions on Knowledge and Data Engineering. https://ieeexplore.ieee.org/document/9099413.

[13] JIANG Z H, WANG Y B, XU G, et al. Multi-scale Based Accelerator for Attribute Reduction[J]. Computer Science, 2019, 46(12): 250-256.

[14] DU Y, HU Q H, ZHU P F, et al. Rule Learning for Classification based on Neighborhood Covering Reduction[J]. Information Sciences, 2011, 181(24): 5457-5467.

[15] ZHU P F, HU Q H, YU D R. Ensemble Learning Based on Randomized Attribute Selection and Neighborhood Covering Reduction[J]. Acta Electronica Sinica, 2012, 40(2): 273-279.

[16] ZHANG B W, MIN F, CIUCCI D. Representative-based Classification Through Covering-based Neighborhood Rough Sets[J]. Applied Intelligence, 2015, 43(4): 840-854.

[17] YUE X D, CHEN Y F, MIAO D Q, et al. Tri-partition Neighborhood Covering Reduction for Robust Classification[J]. International Journal of Approximate Reasoning, 2017, 83: 371-384.

[18] YUE X D, CHEN Y F, MIAO D Q, et al. Fuzzy Neighborhood Covering for Three-way Classification[J]. Information Sciences, 2020, 507: 795-808.

[19] YUE X D, ZHOU J, YAO Y Y, et al. Shadowed Neighborhoods Based on Fuzzy Rough Transformation for Three-way Classification[J]. IEEE Transactions on Fuzzy Systems, 2020, 28(5): 978-991.

[20] PAN Y W, PAN Z B, WANG Y K, et al. A New Fast Search Algorithm for Exact K-nearest Neighbors Based on Optimal Triangle-inequality-based Check Strategy[J]. Knowledge-Based Systems, 2019, 189: 105088.

[21] WANG X Y. A Fast Exact K-nearest Neighbors Algorithm for High Dimensional Search Using K-means Clustering and Triangle Inequality[C]// The 2011 International Joint Conference on Neural Networks. IEEE, 2011: 1293-1299.

[22] CHANG J Y, HE C X. K-means Algorithm Based on Triangle Inequality[J]. Computer Engineering and Design, 2007, 28(21): 5094-5096.

[23] WILSON D R, MARTINEZ T R. Improved Heterogeneous Distance Functions[J]. Journal of Artificial Intelligence Research, 1997, 6: 1-34.

[24] ZHANG Z L, CAO Z Y, LI Y T. Research Based on Euclid Distance with Weights of K_means Algorithm [J]. Journal of Zhengzhou University(Engineering Science), 2010, 31(1): 89-92.

CHEN Yu-si, born in 1994, postgraduate. His main research interests include rough sets, machine learning and uncertain information processing.

（责任编辑：柯颖）