doi:10.16104/j.issn.1673-1891.2023.03.007

基于Seeded-Kmeans和SVM的分类算法

陈婉茹

(西昌学院理学院,四川 西昌 615013)

摘　要:支持向量机(support vector machines, SVM)在人像识别、文本分类等模式识别问题中有广泛的应用,可以有效地解决一些实际生活中的分类问题。针对半监督两分类问题,提出了基于Seeded-Kmeans和SVM的分类算法(SK-SVM)。用Seeded-Kmeans算法对无标签点进行处理,使其获得初始标签,再选取有效的标签点加入已有带标签点中,构成新的带标签训练集,最后结合SVM进行分类。选取UCI中的8个数据集进行数值实验,基于Seeded-Kmeans和SVM的分类算法的有效性得到了验证。

关键词:k-means算法;seeded-kmeans;支持向量机(SVM);半监督支持向量机(S3VM)

中图分类号:TP181;TP301.6　　**文献标志码:**A　　**文章编号:**1673-1891(2023)03-0040-04

A Classification Algorithm Based on Seeded-Kmeans and SVM

CHEN Wanru

(School of Science, Xichang University, Xichang, Sichuan 615013, China)

Abstract: Support vector machine is widely used in pattern recognition problems such as the portrait recognition and the text classification recognition. It can effectively solve some classification problems in real life. In this paper, a classification algorithm based on Seeded-Kmeans and SVM (SK-SVM) is proposed for the semi-supervised two classification problem. The Seeded-Kmeans algorithm is used to process the unlabeled points to obtain initial labels. Then, effective label points are selected and added to the existing labeled points to form a new labeled training set. Finally, SVM is combined to classify the unlabeled points.

Keywords: k-means algorithm; seeded-kmeans; support vector machines (SVM); semi-supervised support vector machines (S3VM)

0 引言

机器学习是人工智能中发展最快速的分支之一,致力于研究如何通过计算的手段,利用经验来改善自身的性能[1]。监督学习[2]、半监督学习[3]和无监督学习[4]是机器学习的3种方式。1995年,Vapnik等[5]首次提出了支持向量机,文本分类[6]、基因序列[7]和图像识别[8]等是机器学习在生产生活中的应用,但由于实际情况样本数据标签的收集较难。因此,半监督学习具有重要的研究意义。半监督两分类问题可描述如下:给定2类分类问题的训练集$T=\{(x_1,y_1),\cdots,(x_m,y_m)\}\cup\{x_{m+1},\cdots,x_{m+q}\}$。其中,$x_i\in R^n,i=1,\cdots,m+q,y_i\in Y=\{1,-1\},i=1,\cdots,m$。

由已有训练点寻找无标签点对应的标签y_i的值以及决策函数$f(x)=\mathrm{sgn}(g(x)),g(x)$是$R^n$上的一个实值函数。

结合实际情况,额外的一些监督信息还是能够获取到,考虑到不浪费获取到的标签信息,提出了半监督聚类算法。Seeded-Kmeans算法和Constrianed-Kmeans算法是由Sugato等[9]提出的2个算法,均能有效地使用带标签点信息,并确定带标签点为初始聚类中心,聚类效果得到了提高。陈婉如[10]、唐晓亮[11]、任江涛等[12]将Seeded-Kmeans算法应用于各个领域。

Seeded-Kmeans算法步骤如下:

输入:数据集$\chi=\{x_1,\cdots,x_N\},x_i\in R^d(i=1,2,\cdots,N)$,聚为$K$类,令$S=\bigcup_{l=1}^{k}S_l$为种子集。

输出:将χ分为K个部分$\{\chi_l\}_{l=1}^{K}$。

步骤:

收稿日期:2023-04-19

作者简介:陈婉茹(1995—),女,四川西昌人,助教,硕士,主要研究方向:机器学习,e-mail:2889340991@qq.com。

a. 初始聚类中心：$\mu_h^{(0)} \leftarrow \frac{1}{|S_h|}\sum_{x\in S_h} x, h = 1, \cdots, K$；$t \leftarrow 0$。

b. 重复(i)(ii)(iii)步骤直至收敛：

(i)根据式子：$h^* = \arg\min_h \|x - \mu_h^{(t)}\|^2$，确定 x 属于的类别 h^*；

(ii)重新计算聚类中心：$\mu_h^{(t+1)} \leftarrow \frac{1}{|\chi_h^{(t+1)}|}\sum_{x\in\chi_h^{(t+1)}}^{h} x$；

(iii)$t \leftarrow (t+1)$。

Constrianed-Kmeans算法步骤如下：

输入：数据集 $\chi = \{x_1, \cdots, x_N\}, x_i \in R^d (i = 1, 2, \cdots, N)$，聚为 K 类，令 $S = \bigcup_{l=1}^{K} S_l$ 为种子集。

输出：将 χ 分为 K 个部分 $\{\chi_l\}_{l=1}^K$。

步骤：

a. 初始聚类中心：$\mu_h^{(0)} \leftarrow \frac{1}{|S_h|}\sum_{x\in S_h} x, h = 1, \cdots, K$；$t \leftarrow 0$.

b. 重复(i′)(ii′)(iii′)直至收敛：

(i′)当 $x \in S$ 时，若 $x \in S_h$，即 x 属于第 h 类，则令 $\chi_h^t = \chi_h^{(t+1)}$；

当 $x \notin S$ 时，根据式子：$h^* = \arg\min_h \|x - \mu_h^{(t)}\|^2$，确定 x 属于的类别 h^*，则令 $\chi_{h^*}^t = \chi_{h^*}^{(t+1)}$；

(ii′)重新计算聚类中心：$\mu_h^{(t+1)} \leftarrow \frac{1}{|\chi_h^{(t+1)}|}\sum_{x\in\chi_h^{(t+1)}}^{h} x$；

(iii′)$t \leftarrow (t+1)$。

然而，半监督聚类算法只能得到粗糙的聚类结果，对于半监督两分类问题无法得到决策函数，并且分类的准确率不高。支持向量机分类准确率高，分类效果很好，但是需要大量带标签的训练点，而标签点获取需要大量的人力和物力。

Seeded-Kmeans 算法与 SVM 结合，可以实现 Seeded-Kmeans算法与SVM算法的优势互补。首先使用半监督聚类算法中的Seeded-Kmeans算法对样本数据进行聚类，聚类结束后，所有样本点有了初始标签，这为SVM的有效实现做了准备工作。考虑到样本点离聚类中心越近，其与聚类中心的相似度越高，这类离聚类中心近的样本点聚类后的标签可信度也就越高，因此选取靠近聚类中心的样本点构成SVM分类器的训练集。最后得到决策函数，并将所有数据进行分类。

1 基于Seeded-Kmeans与SVM的分类算法

1.1 支持向量机(SVM)

对于两分类问题，给定训练集：$T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_l, y_l)\} \in (R^n \times Y)^l$。其中 $x_i \in R^n, y_i \in Y = \{1, -1\}, i = 1, 2, \cdots, l$。据此寻找 R^n 空间中的一个实值函数 $g(x)$，使得推断任意输入一个 x 都有对应的 y 输出。

在文献[5]中，通过求解下列最优化问题得到：

$$\min_{\omega, b, \xi} \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{l}\xi_i \text{ s.t. } y_i((\omega \cdot x_i) + b) \geqslant 1 - \xi_i, i = 1, 2, \cdots, l, \xi_i \geqslant 0, i = 1, 2, \cdots, l。$$

式中：ξ_i 是松弛变量；C 为惩罚参数，$C > 0$。

引入 Lagrange 函数，求解上述问题的对偶问题：

$$\min_{\alpha} \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j (x_i \cdot x_j)\alpha_i\alpha_j - \sum_{j=1}^{l}\alpha_j \text{ s.t. } \sum_{i=1}^{l} y_i\alpha_i = 0, \quad 0 \leqslant \alpha_i \leqslant C, i = 1, 2, \cdots, l。$$

式中：α_i 是拉格朗日乘子；C 为惩罚参数，$C > 0$。

对于线性不可分的情况，可引入核函数，最常用的核函数有：多项式核函数和 Gauss 径向基核函数，本文使用的核函数是 Gauss 径向基核函数，其表达式如下：

$$K(x, x') = \exp\left(-\frac{\|x - x'\|^2}{\sigma^2}\right)$$

式中：σ 为参数。对于线性不可分问题，引入 Gauss 径向基核函数后，具体的算法如下：

a. 对训练集

$T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_l, y_l)\} \in (R^n \times Y)^l$，

式中：$x_i \in R^n, y_i \in Y = \{1, -1\}, i = 1, 2, \cdots, l$。

b. 选取适当的核函数 $K(x, x')$ 以及惩罚参数 $C > 0$；

c. 求解凸二次规划问题：

$$\min_{\alpha} \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j K(x_i, x_j)\alpha_i\alpha_j - \sum_{j=1}^{l}\alpha_j \text{ s.t. } \sum_{i=1}^{l} y_i\alpha_i = 0, \quad 0 \leqslant \alpha_i \leqslant C, i = 1, 2, \cdots, l。$$

得解 $\alpha^* = (\alpha_1^*, \alpha_2^*, \cdots, \alpha_l^*)^T$；

d. 计算 b^*：选取位于开区间 $(0, C)$ 中 α^* 的分量 α_j^*，据此计算

$$b^* = y_j - \sum_{i=1}^{l}\alpha_i^* y_i K(x_i, x_j)$$

e. 构造决策函数 $f(x) = \text{sgn}(g(x))$，其中 $g(x) = \sum_{i=1}^{l} y_i\alpha_i^* K(x_i, x) + b^*$。

1.2 基于Seeded-Kmeans与SVM的分类算法

SVM需要大量有标签点进行训练,在半监督聚类后的无标签样本点中选取可靠的数据点及标签加入带标签点中训练SVM,考虑到样本点离聚类中心越近,其与聚类中心的相似度越高,这类离聚类中心近的样本点聚类后的标签可信度也就越高。因此选取靠近聚类中心的样本点以及已有的带标签样本点构成SVM分类器的训练集,最后将所有数据点进行分类,得到决策函数。

SK-SVM步骤如下:

a.输入数据集

$T = \{(x_1, y_1), \cdots, (x_k, y_k)\} \cup \{(x_{k+1}, y_{k+1}), \cdots, (x_{k+m}, y_{k+m})\}$,其中前$k$个为有标签点,$x_i \in R^n, i = 1, \cdots, k+m, y_i \in Y = \{1, -1\}, i = 1, \cdots, k$.即种子集$S = \{(x_1, y_1), \cdots, (x_k, y_k)\}$。

b.Seeded-Kmeans算法进行聚类,初始聚类中心由种子集产生,具体步骤为:

(i′′)初始聚类中心:$\mu_h^{(0)} \leftarrow \frac{1}{|S_h|}\sum_{x \in S_h} x, h = 1, 2$; $t \leftarrow 0$;

(ii′′)根据式子:$h^* = \arg\min_h \|x - \mu_h^{(t)}\|^2$,确定$x$属于的类别$h^*$;

(iii′′)计算新的聚类中心:$\mu_h^{(t+1)} \leftarrow \frac{1}{|\chi_h^{(t+1)}|}\sum_{x \in \chi_h^{(t+1)}}^{h} x$, $t \leftarrow (t+1)$;

(iv′′)若收敛,则算法结束;否则回到(i)。

c.选择到与带标签点距离近的无标签点构成训练集T'。

d.用步骤c中的训练集T'训练SVM分类器,将所有数据点放入分类器得到标签,SVM再分类。

SK-SVM算法的流程图如图1所示:

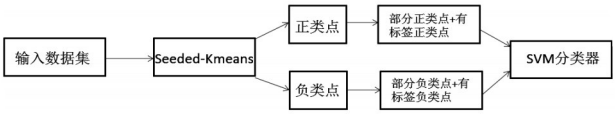

图1 SK-SVM算法的流程

2 数值结果

为了证明该方法的有效性,我们先用Seeded-Kmeans算法对半监督数据集进行聚类,再分别用聚类后的部分有标签点和聚类后部分点加有标签点2种方法进行训练,计算了8个数据集。选取UCI中的WDBC数据集,对数据集进行整理[16],该训练集的所有点构成$m \times n$维矩阵A,$m = 569, n = 30$。对该数据集训2次,第一次带标签点分别为良性中的第201~205行以及恶性中第4~8行,剩下的作为无标签点;第二次带标签点分别为良性患者第1~5行和恶性患者第353~357行,剩下的作为无标签点,其余6个数据集处理方式类似。采用十折交叉验证法得到分类准确率,以这个准确率作为评价分类算法优劣的标准。数值结果如表1所示。

表1 8个数据集的分类准确率

数据集(正类点个数+负类点个数)	无标签点/个	有标签点/行	SLA准确率/%	基于绝对值不等式准确率/%	未加有标签点的准确率/%	KM-SVM的准确率/%
WDBC (357+212)	559	201~205, 4~8	75.40	76.79	88.59	88.59
WDBC (357+212)	559	1~5, 208~212	81.72	85.71	86.39	86.92
Heart-Statlog (150+120)	260	146~150, 116~120	72.00	64.00	85.19	88.89
Australian (307+383)	680	303~307, 379-383	75.00	67.65	76.47	70.11
Wine (59+71)	120	55~59, 67~71	71.43	85.71	85.00	90.00
CMC (333+511)	834	329~333, 507~511	58.33	61.05	57.94	64.69
German (631+271)	892	1~5, 1~5	63.27	69.00	85.30	85.30
PD-speech (564+192)	746	559~564, 188~192	64.00	66.67	91.08	91.08

由表1的数值结果可知,对于数据集WDBC,数值结果表明,仅仅只进行聚类,其准确率分别为88.59%和86.39%,利用收集到的有标签点后,准确率由88.59%和86.39%提升到了88.59%和86.92%。Heart-Statlog准确率由85.19%提升为88.89%。数据集Wine也由85.00%提升为90.00%,提高了

5.00%。而数据集German和PD-speech考虑到聚类后选取的训练点中已经包含有标签点,所有加入有标签点与否准确率均为85.30%与91.08%。

3 结束语

对于半监督两分类问题,本文对无标签点先采用Seeded-Kmeans算法进行处理,再使用SVM进行分类。较好地利用了少量有标签点的带标签信息,将较少的带标签点转化为带标签点,实现了半监督问题转化为监督问题来解决,再使用SVM进行分类。信息得到了较好的利用,并且准确率有了一定的提高。

参考文献:

[1] 周志华.机器学习[M].北京:清华大学出版社,2016:197-214.

[2] [2]KUPPILI V.Advances in pattern classification using improved ann, spiking neural network and decision tree[J].Lecture Notes in Computer Science,2014,5664(41):242-253.

[3] CHAPELLE O,ZIEN A. Semi-Supervised learning (adaptive computation and machine learning)[J]. The MIT Press, 2006, 51:34-42.

[4] FARRAR C R, WORDEN K.Unsupervised learning-novelty detection[M].New York:John Wiley & Sons Ltd,2012:198-225.

[5] 邓乃扬, 田英杰. 支持向量机:理论、算法与拓展[M]. 北京:科学出版社, 2009.

[6] NGUYEN T P,HO T B.Detecting disease genes based on semi-supervised learning and protein-protein interaction networks[J]. Artificial Intelligence in Medicine, 2012, 54(1):63-71.

[7] HASSANI MS, GREEN JR. A semi-supervised machine learning framework for microRNA classification[J]. Human Genomics, 2019,13: 43.

[8] 杜阳, 姜震, 冯路捷. 结合支持向量机与半监督-means的新型学习算法[J]. 计算机应用, 2019, 39(12):5.

[9] QIN Y, DING SF, WANG LJ, et al. Research progress on semi-supervised clustering[J]. Cognitive Computation, 2019, 11: 599-612.

[10] 陈婉茹.求解半监督分类问题的支持向量机[D]. 重庆:重庆师范大学, 2020.

[11] 唐晓亮.基于神经网络的半监督学习方法研究[D]. 大连:大连理工大学, 2009.

[12] 任江涛, 吴海建, 吴向军,等. 一种基于遗传算法的分裂式层次化聚类算法[J]. 计算机应用, 2005, 25(11):2618-2620.

(上接第33页)

播规律,但模拟的网络还是过于理想,接下来将考虑收集真实的网络数据进行隐性知识传播的模拟验证。

参考文献:

[1] 张目,吕知远.基于文本信息的科技型中小企业信用风险识别信号博弈模型[J].软科学,2022,36(5):131-136.

[2] TEECE D J. Strategies for managing knowledge assets: the role of firm structure and industrial context[J]. Long Range Planning, 2000, 33(1): 35-54.

[3] 胡绪华,蒋苏月,王为东.集群知识传播绩效的实现机制研究——基于政府干预与企业知识活动的视角[J].技术经济与管理研究,2015(8):8-12.

[4] LIAO S G, YI S P. Modeling and analyzing knowledge transmission process considering free-riding behavior of knowledge acquisition: a waterborne disease approach[J]. Physica A: Statistical Mechanics and its Applications, 2021, 569: 125769.

[5] 杨湘浩,段哲哲,王筱莉.考虑遗忘机制的企业隐性知识传播SIR模型研究[J].中国管理科学,2019,27(7):195-202.

[6] MIN L, NAN L. Scale-free network provides an optimal pattern for knowledgetransfer[J]. Physica A, 2010(389): 473-480.

[7] KRISTINA BOGNER. Knowledge diffusion in formal networks: the roles ofdegree distribution and cognitive distance[J]. Int. J. Computational Economics and Econometrics, 2018(8): 388-404.

[8] 朱宏淼,闫辛,靳祯,等.耦合网络视角下企业社交媒体对知识传播影响[J].系统工程学报,2020,35(2):153-162+221.

[9] LIAO S G, YI S P. Modeling and analysis knowledge transmission process in complex networks by considering internalization mechanism[J]. Chaos, Solitons & Fractals, 2021, 143: 110593.

[10] 王志平,王佳.基于超网络的舆论演化动态模型[J].复杂系统与复杂性科学,2021,18(2):29-38.

[11] 谢能刚,代亚运,王萌,等.考虑情感的三策略囚徒困境博弈模型与合作演化[J].运筹与管理,2022,31(3):93-99.

[12] 朱宏淼,靳祯,齐佳音,等.线上线下双层耦合网络上的知识传播动力学研究[J].系统工程理论与实践,2020,40(2):403-414.