Hans Journal of Data Mining 数据挖掘, 2013, 3, 41-49
http://dx.doi.org/10.12677/hjdm.2013.34008 Published Online October 2013 (http://www.hanspub.org/journal/hjdm.html)

Support Vector and Multi-Exemplar: Two Main Approaches for Nonlinear Clustering

Changdong Wang[1], Jianhuang Lai[2]

[1]School of Mobile Information Engineering, Sun Yat-sen University, Guangzhou
[2]School of Information Science and Technology, Sun Yat-sen University, Guangzhou
Email: wangchd3@mail.sysu.edu.cn, stsljh@mail.sysu.edu.cn

Received: Aug. 12th, 2013; revised: Aug. 20th, 2013; accepted: Aug. 29th, 2013

Abstract: As a fundamental method for data mining, data clustering has been widely used in various fields such as computer science, medical science, social science and economics. According to the data distribution of clusters, the data clustering problem can be categorized into linearly separable clustering and nonlinearly separable clustering. Due to the complex manifold of the real-world data, nonlinearly separable clustering is one of the most popular and widely studied clustering problems. In this paper, we will first make a brief survey on the recent research works in nonlinear clustering, from four perspectives, namely, kernel-based clustering, multi-exemplar clustering, graph-based clustering and support vector-based clustering. Then, we will particularly introduce our two research works in nonlinear clustering, namely, position regularized support vector clustering (PSVC) and multi-exemplar affinity propagation (MEAP). We will analyze their merits and limitations and point out the future research directions.

Keywords: Nonlinear Clustering; Support Vector Clustering; Multi-Exemplar Clustering; PSVC; MEAP

支持向量和多中心点：非线性聚类的两大方法

王昌栋[1]，赖剑煌[2]

[1]中山大学移动信息工程学院，广州
[2]中山大学信息科学与技术学院，广州
Email: wangchd3@mail.sysu.edu.cn, stsljh@mail.sysu.edu.cn

收稿日期：2013 年 8 月 12 日；修回日期：2013 年 8 月 20 日；录用日期：2013 年 8 月 29 日

摘　要：作为数据挖掘的基础方法之一，数据聚类被广泛应用各个不同领域，例如计算机科学、医学、社会科学和经济学等。根据类的样本点的分布，数据聚类问题通常可以划分成线性可分聚类和非线性可分聚类。由于现实世界的数据分布流形的复杂性，非线性聚类是最流行和最被广泛研究的聚类问题之一。本文首先从四个角度对非线性聚类的近期工作做一个简要的综述，包括基于核的聚类算法、多中心点聚类算法、基于图的聚类算法以及基于支持向量的聚类算法。接着，我们将特别地介绍我们在非线性聚类研究方面的两个主要工作，分别是位置正则化的支持向量聚类(PSVC)以及多中心点近邻传播算法(MEAP)。我们将介绍这些方法的优势与局限性，同时指出未来的研究方向。

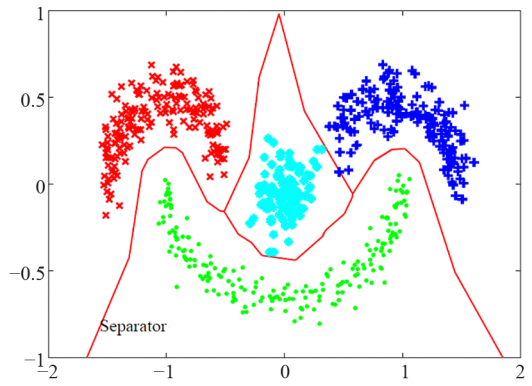
关键词：非线性聚类；核聚类；多中心点聚类；PSVC；MEAP

1. 引言

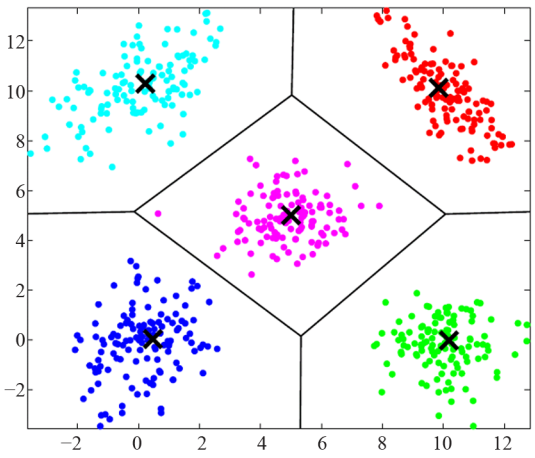
我们处于信息大爆炸的时代，每天的社会生产和生活都产生了大量数据，如国家经济数据、股票行情、银行数据、商业数据、企事业报表、人口数据、交通

数据、天气资料、多媒体信息(视频、图像和语音)等等。如何有效地分析和利用这些数据是一个一直困扰人们的难题。将所获得的数据分类是一种有效简化数据处理和分析的方法，也是人类最基本的处理事物的手段。作为数据挖掘中最基本的方法之一，数据聚类在各科学领域的数据分析中扮演着重要的角色，如计算机科学、医学、社会科学和经济学等[1,2]。给定一个由样本点组成的数据集，聚类的目标是将样本点划分成若干类，使得属于同一类的样本点非常相似，而属于不同类的样本点不相似。根据类的分布形状，聚类问题可以分成线性可分聚类问题和非线性可分聚类问题。

对一个数据集，若至少包含一个非凸形状边界的类，则该数据集称为非线性可分的。图1(a)展示了一个非线性可分的数据集，作为对比，图1(b)展示了一个线性可分的数据集。从类的建模的角度，非线性可分聚类与线性可分聚类的本质区别是：线性可分的聚类可以通过一个中心点表达一个类，但是非线性可分的聚类无法通过一个中心点表达一个类。也就是，线性可分的聚类之间的划分边界可以通过由类中心点形成的Voronoi图来刻画。如图1(b)所示，这五个线性可分的类的划分边界通过由五个类中心点形成的Voronoi图来刻画。对非线性可分的数据集进行有效划分的聚类算法称为非线性聚类算法。



(a)



(b)

Figure 1. Nonlinearly separable dataset and linearly separable dataset (a) Nonlinearly separable dataset (b) Linearly separable dataset

图1. 非线性可分数据集和线性可分数据集 (a) 非线性可分数据集 (b) 线性可分数据集

最近十几年，研究者从不同的角度展开了非线性聚类研究，提出了若干非线性聚类技术，代表性的有基于核的聚类(kernel-based clustering)[3-5]，多中心点方法(multi-exemplar method)[6,7]，基于图的方法(graph-based method)[8,9]，以及基于支持向量的聚类(support vector-based method)[10,11]。由于现实数据的复杂流形分布，非线性可分聚类是最流行且最被广泛研究的聚类问题之一。

本文的第二节首先从四个角度对非线性聚类的近期工作做一个简要的综述，包括，基于核的聚类算法、多中心点聚类算法、基于图的聚类算法以及基于支持向量的聚类算法。接着，我们将特别地介绍我们在非线性聚类研究方面的两个主要工作，分别是第三节：位置正则化的支持向量聚类(PSVC)[11]，以及第四节：多中心点近邻传播算法(MEAP)[7]。我们将介绍这些方法的优势与局限性，同时在两个算法的章节后面指出未来的研究方向。

2. 非线性聚类综述

在本节，我们从四个角度对非线性聚类的近期工作做一个简要的综述，包括，基于核的聚类算法、多中心点聚类算法、基于图的聚类算法以及基于支持向量的聚类算法。

2.1. 核聚类

核聚类(kernel-based clustering)算法是最广泛地用于解决非线性聚类的算法之一[12]。核聚类算法的基本思想是，首先通过一个非线性核映射-，将样本点集从原始空间映射到核空间，使得样本点集在核空间中是线性可分的，再在核空间中找一个合适的样本集的

类分配函数，使得类内样本点的相似性很高，同时类间样本点的相似性很低。在实际应用中，非线性核映射 ϕ 一般是不可知的，并且核空间的维数非常高。所以，一般地，核空间都是通过核函数 κ 以及对应的核矩阵 K 来描述。核空间中两个样本点 $\phi(x)$ 和 $\phi(y)$ 的内积 $\langle\phi(x),\phi(y)\rangle$ 通过核矩阵中的对应元素来描述，这种技术称为核技巧(kernel trick)[12]。然而，在核空间中穷举最佳的类分配函数从计算复杂度的角度是不可能的[12]。因为，所有可能的聚类划分的个数按数据集的大小呈指数级增长。所以，亟需一个高效的搜索算法以找到足够好的局部最优解。

经典的k-均值(k-means)正是这样一个高效的搜索算法[13]，它能够在核空间中寻找合适的类分配函数。这就是核k-均值聚类算法(kernel k-means)[3]。尽管如此，k-均值聚类算法存在着一个严重的缺陷，也就是在病态初始化(ill-initialization)的情况下，它的聚类性能将急剧退化。例如，在随机初始化的过程中，若某个类只随机分配了少数几乎独立的样本点，使得这个类的中心点距离任何样本点都比较远，则在接下来的迭代中，这个中心点将无法分配任何样本点，导致一个空的类。虽然存在若干策略解决这类病态初始化导致的退化问题，如全局搜索策略(global search)[14]和进化算法策略(evolutionary mechanism)[15]，但是这些策略要么无法适应于核空间，要么具有非常高的计算复杂度。

我们的研究工作提出一个新的核聚类算法，称为频数敏感在线学习(conscience on-line learning, COLL)算法[5]，来解决核聚类算法中的最优化问题。这是一个基于在线学习框架的算法。在COLL算法中，我们首先采用与k-均值算法类似的随机初始化方法得到一个初始的聚类划分。但是，在接下来的迭代过程中，与k-均值算法不同的是，对每一个随机挑选的样本点，COLL方法通过频数敏感机制选择优胜者中心点，然后通过在线学习法则更新该优胜者中心点。这种在线学习过程每次只需要将优胜者中心点朝着新样本点移动很短的距离，而不是重新计算每个类的均值，从而取得了更快的收敛率，同时能够方便地将其它竞争机制如频数敏感机制整合到学习过程中。

2.2. 多中心点聚类

若数据集中的某些类含有多个子类，则称该数据集具有多子类结构。含有多子类的类显然无法通过一个中心点来表达。从而，由非线性可分的定义，具有多子类结构的数据集也是非线性可分的。例如，在自然场景图像分类应用中，一个场景类经常包含多个主题[16]，比如，街道场景可能包含如马路、车、行人、建筑等主题，每个主题对应一个子类。类似地，在人脸分类应用中，同一个人在不同表情下拍摄的照片将对应此类中的不同子类[17]。而在光学字符识别和手写数字分类问题中，一个字符类或者数字类可能包含若干个子类，每个子类对应一种书写风格[18]。

刻画多子类结构的一个有效方法就是多中心点模型：一个中心点表达一个对应的子类，再通过合并多个中心点来表达一个类。基于有监督分类的多中心点模型被广泛应用于对多子类结构的数据集进行模。但是在无监督分类(聚类)的情况下，很难采用多中心点模型对多子类结构进行建模。这是因为，缺少训练样本很难估计类的个数以及每个类的子类个数。并且，多中心点模型需要比单中心点模型调整更多的参数。例如，文献[6]中的多中心点聚类(multi-prototype clustering, MPC)算法需要启发式地调整若干敏感的参数以实现多中心点聚类，并且这些参数直接影响多子类结构的建模以及聚类个数的估计。为了采用多中心点模型对具有多子类结构的数据集进行聚类，需要设计多中心点模型的优化方法，使得无需调整参数，同时得到与初始化无关的解。

从多中心点模型(multi-exemplar model)和子类分析(sub-class analysis)的角度，我们的研究工作提出了多中心点近邻传播算法，将单中心点模型扩展到多中心点模型[7]。在多中心点模型里，每个类都由自动确定数量的中心点和一个超级中心点(super-exemplar)来建模。每个样本点被分配给最合适的中心点，同时每个中心点被分配给最合适的超级中心点。超级中心点就是能够最佳表达对应的类的代表性中心点。我们的模型的目标是最大化：样本点与相应中心点之间的相似性之和加上中心点与相应超级中心点之间的链接度之和。直接解决这个目标函数是NP难的。为此，我们采用最大和置信传播算法[19]来产生不依赖于初始化并收敛到近邻最优的聚类结果[20]。

2.3. 基于图的聚类

基于图的算法是另一种广泛应用于解决非线性

可分聚类问题的算法之一。其基本思想是，首先从数据集构建一个图，数据集的样本点作为图的顶点而样本点之间的近邻相似性则用于构建图的边；然后再利用图的有效边连接或者相似性矩阵的特征值/特征向量生成具有任意形状的聚类。例如，在共享近邻聚类(shared nearest neighbor clustering)算法[8]中，首先基于欧氏距离，为每个样本点找k个最近的样本点，然后再定义两个样本点之间的相似性为它们共享的近邻样本点数目，从而构建数据集的相似性矩阵。基于相似性矩阵，定义内核点集(core-point set)，再基于内核点集生成任意形状的聚类。但是，直接基于内核点生成聚类划分并未考虑到非内核点对聚类边界的影响，这将导致类边界样本点的错误划分。另一个基于图的聚类算法是谱聚类算法，如标准化切割(normalized cut, Ncut)[21]。谱聚类的基本思想是，将数据集的聚类问题转化成图的边切割问题。也即是，由数据集构建一个图，再基于某个目标函数对图进行切割，切割结果的每个子图对应一个类。最有名的图切割目标函数是标准化切割(normalized cut, Ncut)[21]，它考虑了子图之间的平衡关系。图切割的最优化问题转化成计算相似性矩阵的特征值/特征向量问题。从而，谱聚类算法具有较高的时间复杂度。

结合图方法、多中心点建模表达及竞争学习算法的优势，我们的研究工作提出了一个新的非线性聚类算法，称为基于图的多中心点竞争学习聚类算法(graph-based multi-prototype competitive learning, GMPCL)[9]。该算法首先通过图方法产生一个初始的粗聚类；接着再提出一个多中心点竞争学习机制来改进该粗聚类，从而产生任意形状的聚类。

2.4. 基于支持向量的聚类

支持向量聚类(support vector clustering, SVC)[10]是由支持向量域描述(support vector domain description, SVDD)[22]发展而来的一种有效划分非线性可分数据集的聚类算法。SVC算法的基本思想是：首先通过核函数将数据集从原始数据空间映射到紧凑的核空间，然后在核空间中找一个包围大部分样本点的超球体，再通过核函数的逆将超球体的球面逆映射到原始数据空间形成原始空间里的若干封闭轮廓线，这些封闭轮廓线将原始数据空间划分成若干连通分块，每个连通分块对应数据集的一个类。与其它聚类算法相比，SVC算法的最大优势在于能够划分任意形状的类，同时可以处理奇异点。但是，SVDD算法所产生的数据域描述非常依赖于一个权衡(trade-off)参数。该权衡参数的选取直接决定了超球体的大小，从而影响了超球体表面样本点的分布，也即SVC算法的聚类结果。并且，对所有样本点采用相同的权衡参数将可能导致错误的SVC聚类结果。所以，提出有效的机制消除该权衡参数成为另一个研究热点。

文献[23]提出一个局部约束的支持向量聚类(locally constrained support vector clustering, LSVC)。通过采用因子分析混合(mixture of factor analyzers，MFA)方法对每个样本点学习一个权重，以代替消除权衡参数，该方法比原始的SVC算法性能有较大的提高。但是，该方法的权重计算依赖于MFA的结果，不同个数的分析子将导致不同的权重，同时估计分析子的个数本身是一个难题。

我们的研究工作提出了一个位置正则化的支持向量聚类(position regularized support vector clustering, PSVC)算法[11]。通过对每个样本点赋予基于位置的权重，而不是所有样本点采用同一个权衡参数，PSVC算法能够自适应地得到一个合适的超球体，从而产生精确的聚类结果。

3. 位置正则化的支持向量聚类

3.1. 支持向量聚类

给定一个包含 N 个样本点的数据集 $X=\left\{x_i \in R^d \middle| i=1,\cdots,N\right\}$ 及一个从原始数据空间投影到高斯核空间的非线性映射 ϕ，我们需要学习一个核空间中围住大部分映射样本点的最小超球体。采用球体中心点 μ 以及球体半径 R 来表示超球体，则需要在满足约束

$$\left\|\phi\left(x_i\right)-\mu\right\|^2 \le R^2+\xi_i,\ \forall i=1,\cdots,N \tag{1}$$

的前提下最小化如下目标函数

$$F\left(R,\mu,\xi_i\right)=R^2+C\sum_{i=1}^{N}\xi_i \tag{2}$$

其中权衡参数 C 权衡了超球体的体积与数据域描述的精度之间的关系，$\xi_i \ge 0$ 是使得允许软边界存在的

松弛变量(slack variables)。

由拉格朗日方法，目标函数可以转换为乌尔夫形式

$$\max_{\beta_i} W = \sum_{i=1}^{N} \beta_i K(x_i, x_i) - \sum_{i,j=1}^{N} \beta_i \beta_j K(x_i, x_j)$$
$$\text{满足 } \sum_{i=1}^{N} \beta_i = 1,\ 0 \le \beta_i \le C,\ \forall i = 1, \cdots, N \tag{3}$$

根据拉格朗日乘子 $\beta_i, i = 1, \cdots, N$ 的值，数据集的样本点可以分成三种类型：

- 内点(inner point, IP)：$\beta_i = 0$，内点位于超球体内部。
- 支持向量(support vector, SV)：$0 < \beta_i < C$，支持向量位于超球体表面。
- 边界支持向量(bounded support vector, BSV)：$\beta_i = C$，边界支持向量位于超球体外面。

核空间样本点 $\phi(x)$ 的核半径函数定义为 $\phi(x)$ 离超球体中心 μ 的欧氏距离，即

$$R(x) = \|\phi(x) - \mu\| = \sqrt{1 - 2\sum_{i=1}^{N} \beta_i K(x_i, x) + \sum_{i,j=1}^{N} \beta_i \beta_j K(x_i, x_j)} \tag{4}$$

从而，超球体半径定义为

$$R = \max\left\{ R(x_i) \middle| x_i \text{是支持向量，即} 0 < \beta_i < C \right\}$$

在原始数据空间，包含数据集大部分样本点的轮廓线定义为 $\{x | R(x) = R\}$。这些轮廓线就是数据集的数据域描述。

通过将数据集从核空间逆映射到原始数据空间，超球体表面对应原始数据空间中的轮廓线，这些轮廓线将数据集划分成若干连通分量，每个连通分量覆盖一个类的样本点。图2展示了支持向量聚类。与其它聚类算法相比，SVC算法的最大优势在于能够划分任意形状的聚类，同时可以处理奇异点。

支持向量聚类算法的一个缺陷是，由于超球体的构建对权衡参数 C 的选择非常敏感，从而得到的聚类结果也依赖于 C。特别地，通过直接控制目标函数(2)中的惩罚项 $C\sum_{i=1}^{N} \xi_i$，超球体的体积以及数据域描述(从而聚类结果)直接依赖于参数 C 的选择。

3.2. 位置正则化的支持向量聚类

在支持向量聚类中出现对参数 C 敏感的主要原因是，对所有样本点使用相同的权衡参数C将使得每个样本点成为奇异点的概率相同。因为样本点的数据密度分布或者样本点相互关系的变化，假设所有样本点具有相同的概率成为奇异点在实际情况下是不成立的。为了改进这个问题，我们提出了位置正则化的支持向量聚类(position regularized support vector clustering, PSVC)算法[11]。通过对每个样本点赋予基于位置的权重，而不是所有样本点采用同一个权衡参数，PSVC算法能够自适应地得到一个合适的超球体，从而产生精确的聚类结果。

为了计算基于位置的权重 $W_i, \forall i = 1, \cdots, N$，我们首先计算核距离矩阵 $D^{\phi} = \left[D_l^{\phi} \middle| l = 1, \cdots, N \right]$ 如下

$$D_l^{\phi} = \left\| \phi(x_l) - \frac{1}{N} \sum_{j=1}^{N} \phi(x_j) \right\|^2 = K(x_l, x_l) + \frac{1}{N^2} \sum_{i,j=1}^{N} K(x_i, x_j) - \frac{2}{N} \sum_{j=1}^{N} K(x_l, x_j)$$

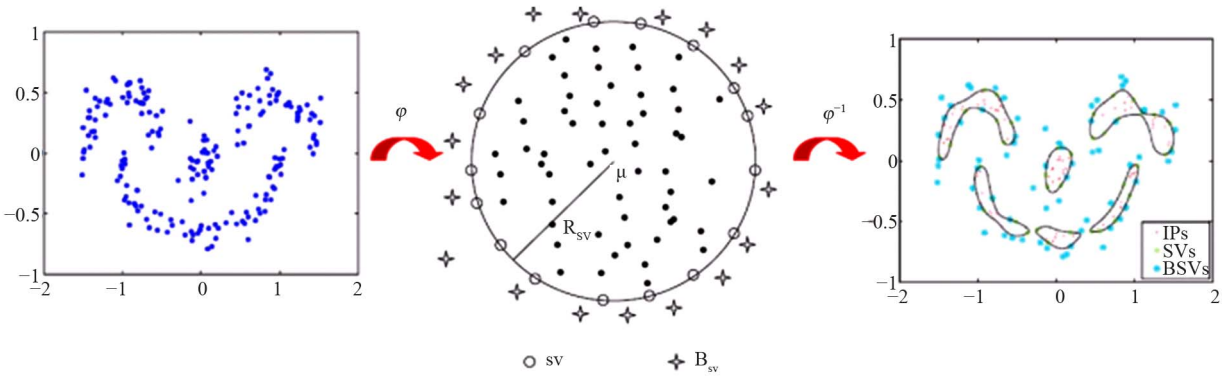


Figure 2. Demonstration of SVC
图2. 支持向量聚类图示

从而，基于位置的权重 $W_i, \forall i=1,\cdots,N$ 可以计算为

$$W_i = \max_{l=1,\cdots,N}\left\{D_l^{\phi}\right\} - D_i^{\phi},\ W_i = \frac{W_i}{\max\limits_{l=1,\cdots,N} W_l} \tag{5}$$

图3展示了权重的概念。粗型星号表示所有核空间投影点的均值，而每个样本点的权重大小由点的大小及颜色来表示。由图可知，样本点的核空间投影点距均值越远，则对应的权重 W_i 越小。权重 $W_i, \forall i=1,\cdots,N$ 则代替权衡参数 C 用来正则化超球体的体积。也就是，在满足约束

$$\left\|\phi(x_i)-\mu\right\|^2 \le R^2+\xi_i,\ \forall i=1,\cdots,N \tag{6}$$

的前提下，我们最小化超球体半径

$$F(R,\mu,\xi_i) = R^2 + \sum_{i=1}^{N} W_i\xi_i \tag{7}$$

与原来的支持向量聚类不同的是，在公式(7)中，每个权重 W_i 分别正则化对应样本点 x_i 的奇异点可能性。权重 W_i 越小，则松弛变量 ξ_i 越大。而松弛变量 ξ_i 则直接用于产生超球体软边界和BSV。从而，这种基于位置的加权机制能自适应地正则化每个样本点是否成为奇异点。

位置正则化对应的乌尔夫形式如下

$$\begin{aligned}&\max_{\beta_i} W = \sum_{i=1}^{N}\beta_i K(x_i,x_i) - \sum_{i,j=1}^{N}\beta_i\beta_j K(x_i,x_j)\\ &\text{满足}\ \sum_{i=1}^{N}\beta_i = 1,\ 0\le\beta_i\le W_i,\ \forall i=1,\cdots,N\end{aligned} \tag{8}$$

注意到，拉格朗日乘子 $\beta_i, i=1,\cdots,N$ 的上界不再相等。相反，每个 β_i 分别由对应的权重 W_i 来控制。

根据拉格朗日乘子 $\beta_i, i=1,\cdots,N$ 和对应的权重 $W_i, i=1,\cdots,N$ 的值，数据集的样本点可以分成三种类型：

- 内点(inner point, IP)：$\beta_i = 0$，内点位于超球体内部。
- 支持向量(support vector, SV)：$0<\beta_i<W_i$，支持向量位于超球体表面。
- 边界支持向量(bounded support vector, BSV)：$\beta_i = W_i$，边界支持向量位于超球体外面。

核空间样本点 $\phi(x)$ 的核半径、超球体半径以及原始数据空间轮廓线均采用与支持向量聚类一样的定义。

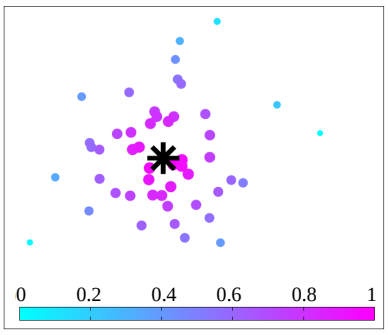


Figure 3. Concept of position-based weighting in kernel space
图3. 核空间中基于位置的权重概念

3.3. 基于支持向量的核聚类的优势及未来研究方向

相对比其它非线性聚类算法，基于支持向量的核聚类不仅仅能够划分任意形状的聚类，同时它还能处理含有奇异点的数据集，更重要的是，它不需要进行聚类初始化，从而无需估计聚类个数。其聚类个数通过将核空间中的超球体表面样本点逆映射回原始空间得到的轮廓线自动刻画。但是，基于支持向量的核聚类仍然存在着亟需解决的问题。例如，虽然位置正则化的支持向量聚类解决了对权衡参数 C 的敏感问题，但是两种基于支持向量的聚类都存在着需要选择能够有效地刻画超球体结构的核参数的问题。虽然，大量实验验证，可以通过基于下采样的稳定性理论来找到合适的高斯核，但是却缺乏理论上的解释。从而，选择一个合适的核映射仍然是基于支持向量的核聚类最困难的待解决问题。

4. 多中心点近邻传播聚类

4.1. 多中心点近邻传播算法

假设有相似性矩阵 $\left[s_{ij}\right]_{n\times n}$ 和链接矩阵分别存储样本点之间的相似性和存储中心点 i 与潜在的超级中心点 j 之间的链接度。多中心点模型需要找两个映射：$\varphi_1:\{1,\cdots,N\}\to\{1,\cdots,N\}$ 将样本点 i 分配给中心点 $\varphi_1(i)$ 和 $\varphi_2:\{\varphi_1(1),\cdots,\varphi_1(N)\}\to\{\varphi_1(1),\cdots,\varphi_1(N)\}$ 将中心点 $\varphi_1(i)$ 分配给超级中心点 $\varphi_2(\varphi_1(i))$。该模型的目标是最大化：样本点与相应中心点之间的相似性之和 S_1 加上中心点与相应超级中心点之间的链接度之和 S_2。

令 $C=\left[c_{ij}\right]_{N\times N}$ 表示类分配矩阵，定义如下

$$c_{ij}=\begin{cases}1 & \text{若}j\text{是}i\text{的中心点} \\ 0 & \text{否则}\end{cases} \quad \forall i\neq j$$
$$c_{ii}=\begin{cases}k\in\{1,\cdots,N\} & \text{若}k\text{是}i\text{的超级中心点} \\ 0 & \text{否则}\end{cases} \tag{9}$$

那么样本点与相应中心点之间的相似性之和 S_1 及中心点与相应超级中心点之间的链接度之和 S_2 可以分别表达成

$$S_1=\sum_{i=1}^{N}\sum_{j=1}^{N}s_{ij}\cdot\left[c_{ij}\neq 0\right],$$
$$S_2=\sum_{i=1}^{N}\sum_{j=1}^{N}l_{ic_{ii}}\cdot\left[c_{ii}\neq 0\right]$$

我们定义一个矩阵 $\left[S_{ij}\left(c_{ij}\right)\right]_{N\times N}$，其中非对角线元素表达了样本点 i 和潜在的中心点 j 之间的相似性 s_{ij}，而对角线元素表达了中心点优先权 s_{ii} 加上中心点 i 与超级中心点 c_{ii} 之间的链接度 $l_{ic_{ii}}$，定义如下：

$$S_{ij}\left(c_{ij}\right)=\begin{cases}s_{ij} & \text{若}i\neq j\text{且}c_{ij}\neq 0 \\ s_{ii}+l_{ic_{ii}} & \text{若}i=j\text{且}c_{ii}\neq 0 \\ 0 & \text{否则}\end{cases} \tag{10}$$

从而，我们有 $S_1+S_2=\sum_{i=1}^{N}\sum_{j=1}^{N}S_{ij}\left(c_{ij}\right)$。一个有效的类分配矩阵 C 必须满足如下三个约束条件：

1) 中心点的N选1约束：每个样本点 i 必须分配给恰好一个中心点，

$$I_i\left(c_{i1},\cdots,c_{iN}\right)=\begin{cases}-\infty & \text{若}\sum_{j=1}^{N}\left[c_{ij}\neq 0\right]\neq 1 \\ 0 & \text{否则}\end{cases}$$

2) 中心点一致性约束：若存在一个样本点 i 选择样本点 j 作为其中心点，则样本点 j 本身必须是中心点，

$$E_j\left(c_{1j},\cdots,c_{Nj}\right)=\begin{cases}-\infty & \text{若}c_{jj}=0\text{但}\exists i:c_{ij}=1 \\ 0 & \text{否则}\end{cases}$$

3) 超级中心点一致性约束：若某个中心点 i 选择中心点 k 作为其超级中心点，即 $c_{ii}=k$，则中心点 k 本身必须是超级中心点，

$$F_k\left(c_{11},\cdots,c_{NN}\right)=\begin{cases}-\infty & \text{若}c_{kk}\neq k\text{但}\exists i:c_{ii}=k \\ 0 & \text{否则}\end{cases}$$

多中心点模型的目标是最大化下面的目标函数

$$J(C)=\sum_{i=1}^{N}\sum_{j=1}^{N}S_{ij}\left(c_{ij}\right)+\sum_{i=1}^{N}I_i+\sum_{j=1}^{N}E_j+\sum_{k=1}^{N}F_k \tag{11}$$

图4展示了多中心点模型。通过两个映射 φ_1 和 φ_2，样本点、中心点和超级中心点形成一个双层结构。映射 φ_1 对下面的一层进行建模：样本点与对应的中心点之间的相似性之和，即 S_1，反映了子类内部的紧凑性(within-subcluster compactness)；而映射 φ_2 对上面的一层进行建模：中心点和对应的超级中心点之间的链接度之和，即 S_2，反映了类内部的紧凑性(within-cluster compactness)。根据中心点表达模型的理论，一个最优的多中心点模型应该同时最大化子类内部的紧凑性和类内部的紧凑性。也就是说，在满足生成有效聚类划分的前提下(即满足约束 I,E,F)，最大化 S_1+S_2 能够有效地刻画具有多子类结构的类。

直接搜索最大化目标函数(11)的类分配矩阵 C 是NP难的。我们采用了最大和置信传播算法来解多中心点模型的最优化问题。多中心点模型的因子图如图5所示；对应地，变量与函数之间的7种消息如图6所示。通过采用数学技巧将这7种消息进行一系列简化，我们得到下面的消息传递

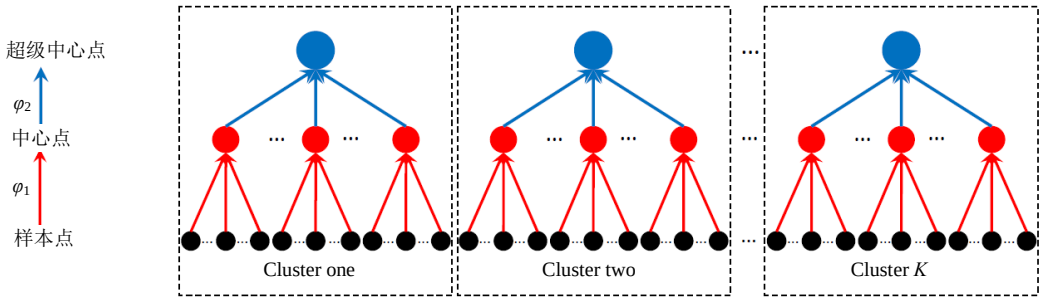


Figure 4. Two-layer structure of multi-exemplar model
图4. 多中心点模型双层结构

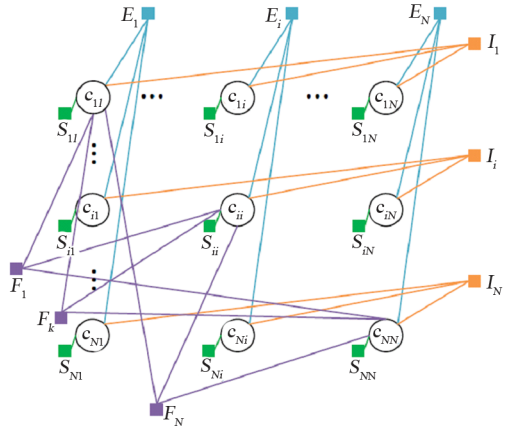


Figure 5. Factor graph of multi-exemplar model
图5. 多中心点模型的因子图

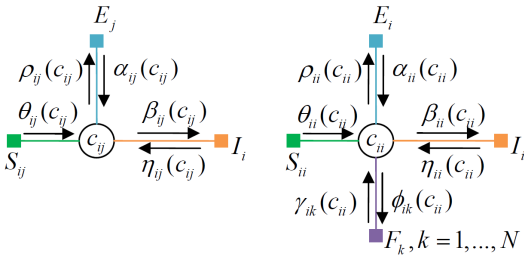


Figure 6. Messages of multi-exemplar model
图6. 多中心点模型的消息

$$i \neq j$$

$$\tilde{\rho}_{ij} \leftarrow s_{ij} - \max\left[\begin{array}{l}\max\limits_{j' \notin \{j,i\}}\left[s_{ij'} + \tilde{\alpha}_{ij'}\right], \\ \max\limits_{m\{1,\cdots,N\}}\left[l_{im} + \tilde{\gamma}_{im}\right] + s_{ii} + \tilde{\alpha}_{ii}\end{array}\right]$$

$$\tilde{\alpha}_{ij} \leftarrow \min\left[0, \max_{m\{1,\cdots,N\}} \tilde{\rho}_j^m + \sum_{i' \notin \{i,j\}} \max\left[0, \tilde{\rho}_{i'j}\right]\right]$$

$$\forall i = 1,\cdots,N, \quad k = 1,\cdots,N$$

$$\tilde{\rho}_i^k \leftarrow s_{ii} + l_{ik} - \max_{i' \neq i}\left[s_{ii'} + \tilde{\alpha}_{ii'}\right] + \tilde{\gamma}_{ik}$$

$$\tilde{\alpha}_{ii} \leftarrow s\sum_{i' \neq i} \max\left[0, \tilde{\rho}_{i'i}\right]$$

$$\tilde{\phi}_{ik} \leftarrow \min\left[\begin{array}{l} l_{ik} - \max\limits_{m \neq k}\left[l_{im} + \tilde{\gamma}_{im}\right], \\ \tilde{\alpha}_{ii} + \tilde{\rho}_i^k - \tilde{\gamma}_{ik}\end{array}\right]$$

$$\tilde{\gamma}_{kk} \leftarrow \sum_{i' \neq i} \max\left[0, \tilde{\phi}_{i'k}\right]$$

$$\tilde{\gamma}_{ik} \leftarrow \min\left[0, \tilde{\phi}_{kk} + \sum_{i' \notin \{i,k\}} \max\left[0, \tilde{\phi}_{i'k}\right]\right], k \neq i$$

这些消息初始化为0，接着上述消息传递公式不断进行更新，直至这些消息值在收敛为止。

为了估计类分配矩阵 C 的元素 c_{ij} 的值，我们将所有输入到 c_{ij} 的消息值累加起来，最优的 c_{ij} 就是使得这个消息累加值最大化的值。简化之后，也就是

$$\hat{c}_{ij} = \begin{cases} 1 & \text{若} \tilde{\alpha}_{ij} + \tilde{\rho}_{ij} \geq 0 \\ 0 & \text{否则} \end{cases} \quad \forall i \neq j$$

$$\hat{c}_{ii} = \begin{cases} \arg\max\limits_k \tilde{\rho}_i^k & \text{若} \tilde{\alpha}_{ij} + \max_k \tilde{\rho}_i^k \geq 0 \\ 0 & \text{否则} \end{cases}$$

根据类分配矩阵，我们便可以得到两个映射，从而得到最终的类分配结果。

4.2. 多中心点近邻传播算法的局限性以及未来研究方向

虽然多中心点近邻传播算法能够有效地划分具有多子类结构的非线性可分数据集，但是该算法却需要调整一个存储了中心点与潜在的超级中心点之间的链接度矩阵。与相似性矩阵可以很容易通过数据的特征计算得到不同的是，虽然，我们可以通过相似性矩阵的比例得到链接度矩阵，但是却没有一个理论上的法则去计算链接度矩阵。从而，研究一个理论法则去计算链接度矩阵是一个颇具理论意义的未来研究方向。

5. 结论

非线性聚类是聚类研究中最热门的研究问题之一。本文首先从四个角度对非线性聚类的近期工作做了一个简要的综述，包括，基于核的聚类算法、多中心点聚类算法、基于图的聚类算法以及基于支持向量的聚类算法。接着，我们特别地介绍了我们在非线性聚类研究方面的两个主要工作，分别是位置正则化的支持向量聚类(PSVC)以及多中心点近邻传播算法(MEAP)。我们介绍了这些方法的优势与局限性，同时指出了未来的研究方向。

参考文献 (References)

[1] R. Xu, D. Wunsch II. Survey of clustering algorithms. IEEE Transactions on Neural Networks, 2005, 16(3): 645-678.

[2] 孙吉贵，刘杰，赵连宇. 聚类算法研究[J]. 软件学报，2008, 19(1): 48-61.

[3] B. Scholkopf, A. Smola and K. R. Muller. Nonlinear Component analysis as a kernel eigenvalue problem. Neural Computation, 1998, 10(5): 1299-1319.

[4] 张莉，周伟达，焦李成. 核聚类算法[J]. 计算机学报，2002,

25(6): 587-590.
[5] C. D. Wang, J. H. Lai and J. Y. Zhu. A conscience on-line learning approach for kernel-based clustering. IEEE 10th International Conference on Data Mining, Sydney, 13-17 December 2010, 531-540.
[6] M. Liu, X. Jiang and A. C. Kot. A multi-prototype clustering algorithm. Pattern Recognition, 2009, 42(5): 689-698.
[7] C. D. Wang, J. H. Lai, J. Y. Zhu and C. Y. Suen. Multi-exemplar affinity propagation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(9): 2223-2237.
[8] L. Ertoz, M. Steinbach and V. Kumar. Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data. Proceedings of the Third SIAM International Conference on Data Mining, 2003, 112: 47-58.
[9] C. D. Wang, J. H. Lai and J. Y. Zhu. Graph-based multiprototype competitive learning and its applications. IEEE Transactions on Systems, Man, and Cybernetics—Part C, 2012, 42(6): 934-946.
[10] A. Ben-Hur, D. Horn, H. T. Siegelmann, et al. Support vector clustering. Journal of Machine Learning Research, 2001, 2: 125-137.
[11] C. D. Wang, J. H. Lai. Position regularized support vector domain description. Pattern Recognition, 2013, 46(3): 875-884.
[12] J. Shawe-Taylor, N. Cristianini. Kernel methods for pattern analysis. Cambridge: Cambridge University Press, 2004.
[13] J. MacQueen. Some methods for classification and analysis of multivariate observations. Proceedings of the 15th Berkeley Symposium on Mathematical Statistics and Probability, 1967, 1: 281-297.
[14] A. Likas, N. Vlassis and J. J. Verbeek. The global k-means clustering algorithm. Pattern Recognition, 2003, 36(2): 451-461.
[15] B. Abolhassani, J. E. Salt and D. E. Dodds. A two-phase genetic k-means algorithm for placement of radioports in cellular networks. IEEE Transactions on Systems, Man, and Cybernetics—Part B, 2004, 34(1): 533-538.
[16] L. Fei-Fei, P. Perona. A bayesian hierarchical model for learning natural scene categories. Proceedings of CVPR, 2005. 524-531.
[17] M. Zhu, A. M. Martinez. Subclass Discriminant analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(8): 1274-1286.
[18] H. I. Avi-Itzhak, J. A. V. Mieghem and L. Rub. Multiple subclass pattern recognition: A maximin correlation approach. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1995, 17(4): 418-431.
[19] F. R. Kschischang, B. J Frey and H. A. Loeliger. Factor graphs and the sum-product algorithm. IEEE Transactions on Information Theory, 2001, 47(2): 498-519.
[20] Y. Weiss, W. T. Freeman. On the optimality of solutions of the max-product belief-propagation algorithm in arbitrary graphs. IEEE Transactions on Information Theory, 2001, 47(2): 736-744.
[21] J. Shi, J. Malik. Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(8): 888-905.
[22] D. Yankov, E. Keogh and K. F. Kan. Locally constrained support vector clustering. Proceedings of the 7th International Conference on Data Mining, Omaha, 28-31 October 2007, 715-720.
[23] D. M. Tax, R. P. Duin. Support vector domain description. Pattern Recognition Letters, 1999, 20(11-13): 1191-1199.

