加权闵可夫斯基 K-Means 的指数选取策略①

王法云, 何振峰

(福州大学 数学与计算机科学学院, 福州 350108)

摘 要: 与传统 K-Means 相比, 加权闵可夫斯基 K-Means(MWK-Means)需要自适应获取特征权重并选择合适的闵可夫斯基指数. 无监督选取指数策略是计算每个指数的三种尺度值, 根据三种尺度的选取标准得到各自最好的指数, 然后选取较接近的两个指数求均值. 在这种策略的启发下, 提出了基于排名的闵可夫斯基指数选取策略, 将三种尺度的值分别进行排名, 每个指数通过选取两个较接近的排名相加得到综合排名来确定指数. 用这两种指数选取策略分别对 UCI 数据集进行实验, 结果表明, 基于排名的选取策略较优.

关键词: 聚类; 闵可夫斯基指数; 无监督; 排名; MWK-Means

Selection of the Minkowski Exponent for MWK-Means

WANG Fa-Yun, HE Zhen-Feng

(School of Mathematics and Computer Science, Fuzhou University, Fuzhou 350108, China)

Abstract: Compared to the traditional K-Means, the MWK-Means needs to obtain feature weights adaptively and select suitable exponent. Choosing the Minkowski exponent in an unsupervised setting is a way to calculate three-scale values of each exponent. It gets the best of each exponent based on the selection criteria of three scales, and then gets the mean of two closer exponents. According to this strategy, we put forward a new strategy of selecting the Minkowski exponent based on ranking, ranked the values of three scales each. Then, we added the two closer rankings of each exponent as comprehensive rank and used them to determine the final exponent. This paper used the above two strategies of selecting Minkowski exponent to test UCI dataset. The result shows that the new strategy is better.

Key words: clustering; Minkowski exponent; unsupervised; ranking; MWK-Means

聚类分析是数据挖掘领域的一个重要分支. 聚类就是将数据集划分为若干类或簇, 使得同一个簇内的数据样本具有较高的相似度, 而不同簇中的数据样本则是不相似的[1,2]. 现有的聚类方法主要分为以 K-Means、Fuzzy K-Means 为代表的划分型聚类, 以 Cure、Birch 为代表的层次聚类和以 Dbscan、Optics 为代表基于密度的聚类等[3].

K-Means 作为划分型聚类的代表, 基本思想是: 随机地选择 K 个样本作为初始中心点, 对剩余的每个样本根据其与各个类中心的距离, 将它分配给最近的类; 然后重新计算每个类中样本的平均值作为聚类中心点, 不断重复该过程, 直到每个聚类不再发生变化为止. K-Means 存在的缺点有: 聚类数 K 必须事先给定, 聚类的结果依赖于初始的中心点, 距离度量的选取问题[4]等.

到目前为止, 对于前两个问题的解决方法已经很多, 确定聚类数 K 的方法主要是通过引入某种聚类质量评估标准来确定, 如结合高斯混合模型(GMM)的最小消息长度(MML)准则(Figueiredo and Jain, 2002), 贝叶斯信息准则(BIC)等; 初始聚类中心的选择主要有最大最小距离法和通过得到 K 个异常聚类求初始聚类中心[4-6]. 但对于距离度量选取的研究还相对较少, K-Means 采用的是欧氏距离, 它能很好的发现凸球面形状的簇, 但不适于发现非凸球面形状的或者大小差别很大的簇, 为了解决这一问题, Amorim 等提出了 MWK-Means[7,8], 因闵可夫斯基指数对于聚类的结果

① 收稿时间:2014-05-14;收到修改稿时间:2014-06-04

影响较大, 于是指数选择就成为一个重要研究课题.

1 MWK-Means (Minkowski Weighted K-Means) 算法

MWK-Means 算法包含特征权重的自适应调整, 闵可夫斯基距离度量以及不同闵可夫斯基指数的聚类中心的求取. 每一次迭代需要根据当前样本的划分状态, 分别动态的确定各个类对应的特征权重; 聚类中心的求取采用闵可夫斯基中心算法[6], 它是对聚类的各个维度的中心进行独立的求取.

加权闵可夫斯基距离度量函数为:

$$d_p(y_i, c_k) = \sum_{v=1}^{V} {w_{kv}}^p \mid y_{iv} - c_{kv} \mid^p \quad (1)$$

其中, V代表特征数, w_{kv}是第k类第v个特征的权重, p是闵可夫斯基指数, y_{iv}是第i个样本的第v个特征值, c_{kv}是聚类中心第k类的第v个特征值.

MWK-Means的目标函数为:

$$W_p(S, C, w) = \sum_{k=1}^{K} \sum_{i \in S_k} \sum_{v=1}^{V} {w_{kv}}^p \mid y_{iv} - c_{kv} \mid^p \quad (2)$$

对于特征权重的计算, 思路源自于Chan[9]的观点, 即在一个特定的类中, 一个样本的特征和类内的其他样本的特征的距离之和越小, 其对应的特征权重就越大. 这一思路对应自适应求特征权重的算法[10], 特征权重的计算公式如下:

$$w_{kv} = \frac{1}{\sum_{u \in V} [D_{kv} / D_{ku}]^{1/(p-1)}} \quad (3)$$

其中 D_{kv} 的计算公式为:

$$D_{kv} = \sum_{i \in S_k} \mid y_{iv} - c_{kv} \mid^p \quad (4)$$

通常指数p按1:0.1:5进行取值, 但当更新权重时, p等于1是无法用公式(3)来求取的, 采取的方法是将同一个聚类中对应的D_{kv}最小的特征权重设为1, 其余的为0[11].

MWK-Means算法的基本步骤:

1) 随机选取K个中心点, 并将初始的特征权重设为属性个数的倒数, 即w_{kv}=1/V.

2) 用公式(1)计算每个样本到K个聚类中心点的加权闵可夫斯基距离(传统K-Means采用欧氏距离, 即p=2), 将样本分配到距离中心点最近的类中.

3) 更新每个类的闵可夫斯基中心(传统K-Means是聚类样本的均值). 如果前后两次的聚类中心不再发生变化就停止, 否则继续.

4) 用公式(3)重新计算每个类对应的特征权重(传统 K-Means 的特征权重始终等同于 1/V), 转到 2).

2 闵可夫斯基指数选取

目前闵可夫斯基指数的选取主要有两类方法: 一类是半监督学习, 一类是无监督学习. 半监督选取指数的前提是数据集有部分数据是被正确标记的, 无监督选取指数是没有任何隐藏的划分信息, 通常采用评分函数来自适应选取.

2.1 半监督选取策略

半监督选取指数[12]是选取标记数据样本中的部分或者全部作为实验数据, 对应准确率最高的指数作为最佳指数.

半监督选取指数的步骤:

1) 按照 p=1.1:0.1:5 对选取的数据样本运行 MWK-Means 算法;

2) 选取使得标记数据中准确率最高的闵可夫斯基指数作为数据样本的最优指数 p*;

半监督算法的前提是必须有足够数量的标记样本, 而通常标记过程比较复杂, 导致代价较高; 相比之下, 无监督算法只需选择合适的评价函数得来到较好的效果.

2.2 无监督选取策略

现有的无监督策略系 Amorim 等提出的. 它是在比较指数的聚类质量基础上进行的, 评价划分的质量, Amorim 等采用了三种尺度, 即闵可夫斯基聚类系数(MCI, Minkowski Clustering Index), 基于闵可夫斯基距离的 Silhouette 系数(MSI, Minkowski Silhouette Index)和基于欧式距离的 Silhouette 系数(ESI, Euclidean Silhouette Index)[12]. 三种尺度选取指数的标准分别是选取 MCI 最小值, MSI 最大值和 ESI 最大值对应的指数, 然后选择三个指数中较接近的两个指数求均值(MES, Mean Exponent Selection).

其提出的无监督选取指数算法的基本步骤:

1) 对数据集进行归一化处理, 按 p=1.1:0.1:5 运行 MWK-Means, 对每一个 p 计算 MCI, MSI 和 ESI.

2) 分别选取对应 MCI 最小的指数 p_1, MSI 最大的 p_2 和 ESI 最大的 p_3.

3) 用 MES 策略获取最佳的指数.

闵可夫斯基聚类系数(MCI)的定义是:

$$MCI = \frac{W_p(S, C, w)}{\sum_{k=1}^{K}\sum_{i\in S_k}\sum_{v=1}^{V} | w_{kv}y_{iv} |^p} \quad (5)$$

对于给定划分, 一个样本的Silhouette Index为:

$$S(y_i) = \frac{b(y_i) - a(y_i)}{\max\{a(y_i), b(y_i)\}} \quad (6)$$

其中$a(y_i)$是样本y_i与类内其他样本相异程度的平均值, $b(y_i)$是y_i与其他类中样本的
相异程度平均值的最小值.

在求取Silhouette Index时, 样本y_i和样本y_j距离度量采用闵可夫斯基距离和欧式距离:

闵可夫斯基距离:

$$d_p{}^p(y_i, y_j) = \sum_{i\in S_k}^{V} w_{kv}{}^p | y_{iv} - y_{jv} |^p \quad (7)$$

欧式距离:

$$d(y_i, y_j) = \sum_{i\in S_k}^{V} | y_{iv} - y_{jv} |^2 \quad (8)$$

对于Silhouette Index, 因为$S(y_i)$越大, 说明y_i更有可能属于这一类, 所以对MSI和ESI分别选择对应最大值的指数.

3 基于排名的闵可夫斯基指数的自适应选取算法

MES 策略得到指数往往不一定是好的, 就如同两个山峰的中间往往是山谷. 图 1 是 Iris 数据集实验结果, 横坐标是闵可夫斯基指数, 纵坐标是对应的排序, 其中 MCI 按降序排序, MSI, ESI 以及实际的聚类准确率按照升序进行排序. 由图中可看出 MSI 对应最大值的指数是 4.5, ESI 对应最大值的指数是 1.1, MCI 对应最小值的指数是 3, 基于这三个参数, MES 策略得到的指数是 3.8, 在图中已经标记出来, 可以看出这一策略得到的指数离实际最高的准确率差距很大, 同时其对应三种尺度排名也不理想.

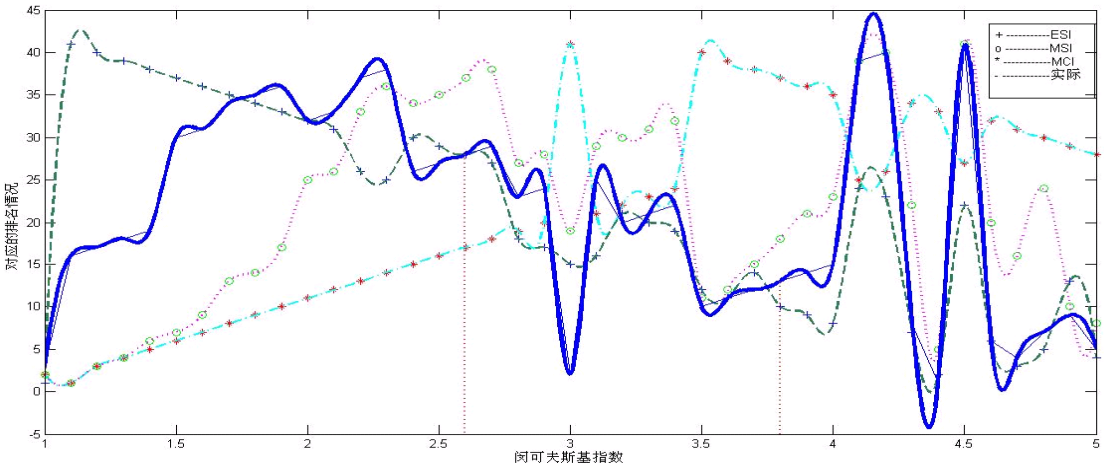

图 1 Iris 数据集的结果分析图

针对 MES 策略存在的不合理之处, 而排名可以综合考虑到这三种尺度对于指数的综合影响, 于是我们提出了基于排名的闵可夫斯基指数的自适应选取算法, 对每一个指数对应的三种尺度进行排名, 然后选择其中较接近的两个排名进行相加, 最终每个指数对应一个综合排名, 排名最靠前对应的指数即为所求(RES Ranking based Exponent Selection). 采用 RES 策略得到的指数为 2.6, 从图 1 中可以看出, 其对应的聚类准确率更接近实际的最高准确率.

基于排名的闵可夫斯基指数自适应选取算法基本步骤:

1) 随机选取 K 个初始的聚类中心点, 并设置初始权重为所有特征数的倒数.

2) 取 p=1:0.1:5 进行实验, 用 MWK-Means 对每一个 p 进行 100 次的随机实验, 对每一次的聚类结果求取对应的 MCI, MSI 和 ESI.

3) 选取每一个 p 对应的 100 次实验中目标函数即公式(2)最小的情况下所对应的三种尺度的值.

4) 用 RES 策略选择指数.

4 实验

为了验证闵可夫斯基指数选取方法的有效性和可行性, 本文选用 4 个 UCI 数据集(Iris, Wine, Glass 和 Seed)进行实验, 以此来比较 RES(Ranking based Exponent Selection)策略和 MES(Mean Exponent Selection)策略的优劣. 表 1 给出了选取 UCI 数据集的信息.

表 1 选取的 UCI 数据集

数据集	对象数	特征数	聚类数
Iris	150	4	3
Wine	178	13	3
Glass	214	9	6
Seed	210	7	3

运行提出的基于排名的闵可夫斯基指数自适应选取算法, 按照三种尺度的选取标准得到相应的指数以及 MES 策略和 RES 策略所对应的指数, 结果如表 2 所示.

表 2 三种尺度及两种指数选择策略的结果

数据集	ESI	MSI	MCI	MES	RES
	准确率(p值)	准确率(p值)	准确率(p值)	准确率(p值)	准确率(p值)
Iris	87.37(1.1)	94.95(4.5)	73.38(3)	86.23(3.8)	93.41(2.6)
Wine	72.95(1.1)	59.28(5)	59.28(5)	59.28(5)	59.28(5)
Glass	70.62(1.3)	64.45(4.8)	70.08(1.1)	69.51(1.2)	70.08(1.1)
Seed	86.66(1.1)	59.56(5)	59.80(3.4)	59.56(4.2)	59.80(3.5)

由表 2 可知, 采用 MES 策略得到的四个数据集的指数 p 分别是 3.8, 5, 1.2 和 4.2, 而 RES 策略得到的 p 值分别 2.6, 5, 1.1 和 3.5. 对比聚类的准确率可以看出 RES 策略较 MES 策略更优一些.

5 结语

加权闵可夫斯基作为距离度量受到权重和指数的双重影响, 对于自适应权重已经得到了比较好的解决, 本文主要是解决指数的选取问题, 先分析了 MES 策略中存在的问题, 进而提出了 RES 策略. 由四个数据集的实验结果可以看出本文提出的 RES 策略较 MES 策略得到的聚类效果好, 从而验证了这种指数选取策略的正确性.

参考文献:

1 Rui X, Donald WII. Survey of custering algorithms. IEEE Trans. on Neural Networks, 2005, 16(3): 645–678.

2 Han JW, Kamber M.范明,孟小峰译.数据挖掘:概念与技术 第 2 版.北京:机械工业出版社,2007.

3 周杨,苗夺谦,岳晓冬.基于自适应权重的粗糙K均值聚类算法.计算机科学,2011,38(6):237–241.

4 Anil KJ. Data clustering: 50 years beyond K-means. Pattern Recognition Letters, 2010, 31(8): 651–666.

5 王留正,何振峰.基于全局性分裂算子的进化K-Means算法. 计算机应用,2012,32(11):3005–3008.

6 Renato CDA, Boris M. Minkowski metric, feature weighting and anomalous cluster initializing in K-Means clustering. Pattern Recognition, 2012, 45(3): 1061–1075.

7 Renato CDA. Constrained clustering with Minkowski weighted K-Means. 13th IEEE International Symposium on Computational Intelligence and Informatics. 2012. 13–17.

8 Renato CDA, Peter K. On initializations for the Minkowski weighted K-Means. Advances in intelligent Data Analysis XI Lecture Notes in Computer Science, 2012, 7619: 45–55.

9 Chan EY, et al. An optimization algorithm for clustering using weighted dissimilarity measures. Pattern Recognition, 2004, 37(5): 943–952.

10 Huang JZ, Ng MK, Rong H, Li Z. Automated variable weighting in k-means type clustering. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2005, 27(5): 657–668.

11 Renato CDA, Boris M. Selecting the Minkowski exponent for intelligent K-Means with feature weighting. Accepted for Publication as a Chapter of Clusters, Orders, Trees: Methods and Applications, 2013.