

基于属性权重最优化的 k-means 聚类算法

熊 平, 顾 霄

(中南财经政法大学 信息与安全工程学院, 湖北 武汉 430074)

摘 要: 聚类是最常用的数据挖掘算法之一. 为了提高聚类结果的质量, 应用拉格朗日乘数法提出了一种基于属性权重最优化的 k-means 聚类算法. 该算法在计算样本与质心的距离时为各属性赋予相应的权重以表示属性的重要程度, 并在每轮迭代中根据质心向量的变化自动计算最优的属性权重, 使得所有样本与相应质心的距离和最小. 实验结果验证了该方法相对于传统 k-means 算法的优势.

关键词: 聚类算法; 属性权重; 数据挖掘; 目标函数

中图分类号: TP309

文献标识码: A

文章编号: 1000-7180(2014)04-0040-04

A k-means Algorithm Based on Optimization of Attribute Weight

XIONG Ping, GU Xiao

(School of Information and Security Engineering, Zhongnan University of Economics and Law, Wuhan 430073, China)

Abstract: Clustering is one of the most widely used data mining algorithm. In order to improve the quality of clustering results, we present a k-means algorithm based on optimization of attribute weight using Lagrange multiplier method. The algorithm uses the weight value for each attribute to determine the importance of that attribute while computing the distance between an instance and the centroid of a cluster. In each iteration of clustering, it computes the optimal weight of attributes according to the change of centroid vector which minimizes the sum of distance between each instance and the centroid. The experimental results demonstrate that the quality of clustering results can be improved significantly with the algorithm comparing with the traditional k-means algorithm.

Key words: clustering algorithm; attribute weight; data mining; objective function

1 引言

聚类是数据挖掘领域最常用的数据分析方法之一, 用于将数据集中的样本划分到若干个聚类中, 使得一个聚类内的样本具有较高相似度, 而不同聚类内的样本则具有较低相似度^[1]. 常用的相似度指标包括样本间的欧氏距离、马氏距离、切比雪夫距离等. 在众多聚类算法中, k-means 因其简单、易于实现等优点成为最常用的聚类算法之一.

给定数据集 X , 包含 p 个属性, 则 X 中的任一样本 x_i 可表示为向量 $x_i = \langle x_i^1, x_i^2, \dots, x_i^p \rangle$. k-means 算法就是要寻求一种方案 P , 将 X 中的样

本划分为 K 个聚类 $(C_1, C_2, \dots, C_K)$, 并使得

$$W(P) = \sum_{k=1}^K \sum_{x_i \in C_k} d(x_i, g_k) \quad (1)$$

有最小值, 其中 $g_k = \langle g_k^1, g_k^2, \dots, g_k^p \rangle$ 为聚类 C_k 的质心向量, $d(x_i, g_k) = \sum_{j=1}^p (x_i^j - g_k^j)^2$ 为向量 x_i 和 g_k 之间的欧氏距离.

可以看出, 在计算两个样本的相似度时, k-means 算法将所有属性被看作是同等重要的. 然而, 不同的属性对于样本之间的差异所产生的影响往往是不一样的. 基于这一观点, 学者们提出了一些基于属性权重的聚类算法. Tsai 等^[2]提出了一种嵌入到 k-

收稿日期: 2013-05-30; 修回日期: 2013-07-22

基金项目: 国家自然科学基金项目(71173238); 教育部人文社科研究青年基金项目(12YJC630078)

means 算法中的 FWSA 机制,根据聚类内的距离和与聚类间距离和的比值来计算属性权重;Liping 等^[3]针对高维数据集提出了一种基于聚类熵的属性加权 k-means 算法;Yong 等^[4]在 D-优化准则的基础上提出了一种根据向量积矩阵来计算属性权重的方法;Qiang Wang 等^[5]提出了一种基于属性加权的模糊 k-means 算法;武优西等采用改进粒子群优化算法来计算属性权重^[6]. 实验表明,这些算法均能使 k-means 算法的聚类质量得到一定的改善. 但其缺点在于,在确定算法参数时,这些算法都不同程度地依赖于算法执行者的领域知识. 这对于无监督的聚类学习而言,必然会对聚类结果产生间接的影响,使聚类质量无法达到最优.

本文提出了一种基于属性权重最优化的 k-means 算法,根据属性加权 k-means 聚类算法的目标函数,采用拉格朗日乘数法来确定属性权重,并在聚类过程中不断进行自适应的调整,使最终的聚类结果趋于最优. 实验结果验证了该方法的有效性.

2 加权 k-means 算法

2.1 目标函数

传统 k-means 算法就是要寻求一个划分方案 P ,使得式(1)最小,式(1)称为 k-means 算法的目标函数. 对于加权的 k-means 算法,其目标函数需要在式(1)基础上做相应的调整.

给定聚类 C_k ,质心为 $g_k = \langle g_k^1, g_k^2, \dots, g_k^p \rangle$, 设该聚类中各属性的权重分别为 $\lambda_k^1, \lambda_k^2, \dots, \lambda_k^p$. 则聚类 C_k 内样本与质心的距离和为

$$\begin{aligned} \sum_{x_i \in C_k} d(x_i, g_k) &= \sum_{x_i \in C_k} \sum_{j=1}^p (\lambda_k^j (x_i^j - g_k^j)^2) \\ &= \sum_{j=1}^p (\lambda_k^j \sum_{x_i \in C_k} (x_i^j - g_k^j)^2) \end{aligned} \quad (2)$$

所以加权 k-means 算法的目标函数为

$$W(P, \lambda) = \sum_{k=1}^K \sum_{j=1}^p (\lambda_k^j \sum_{x_i \in C_k} (x_i^j - g_k^j)^2) \quad (3)$$

式中, λ 为权重向量集.

因此,加权的 k-means 算法要寻求一个方案 P 和权重向量集 λ ,使得式(3)最小.

2.2 权重计算方法

记 $\varphi_j = \sum_{x_i \in C_k} (x_i^j - g_k^j)^2$, 则式(3)可简化为

$$W(P, \lambda) = \sum_{k=1}^K \sum_{j=1}^p (\lambda_k^j \varphi_j). \text{ 显然,对于所有的 } k=1, 2, \dots, K, \text{ 当 } \sum_{j=1}^p (\lambda_k^j \varphi_j) \text{ 均为最小时, } W(P, \lambda) \text{ 有最小}$$

值. 因此,这就是一个有约束条件的最小优化问题. 按照 Diday 和 Govaert 提出的标准自适应方法^[7], 可以采用拉格朗日乘数法在 $\lambda_k^j > 0$ 和 $\prod_{j=1}^p \lambda_k^j = 1$ 的约束条件下计算权重 $\lambda_k^j (j=1, 2, \dots, p)$, 使得 $\sum_{j=1}^p (\lambda_k^j \varphi_j)$ 最小.

首先构造拉格朗日函数 $L = \sum_{j=1}^p (\lambda_k^j \varphi_j) - \mu \prod_{h=1}^p \lambda_k^h$, 并令 L 对 λ_k^j 的一阶偏导数等于零, 即:

$$\frac{\partial}{\partial \lambda_k^j} (\sum_{j=1}^p (\lambda_k^j \varphi_j) - \mu \prod_{h=1}^p \lambda_k^h) = 0,$$

$$\text{亦即 } \varphi_j - \mu \frac{\prod_{h=1}^p \lambda_k^h}{\lambda_k^j} = 0.$$

因为 $\prod_{h=1}^p \lambda_k^h = 1$, 所以 $\lambda_k^j = \frac{\mu}{\varphi_j}$. 因此约束条件

$\prod_{h=1}^p \lambda_k^h = 1$ 可表述为

$$\begin{aligned} 1 &= \prod_{h=1}^p \frac{\mu}{\varphi_h} = \frac{\mu^p}{\prod_{h=1}^p \varphi_h}, \text{ 所以 } \mu = (\prod_{h=1}^p \varphi_h)^{\frac{1}{p}}. \text{ 所以,} \\ \lambda_k^j &= \frac{\mu}{\varphi_j} = \frac{(\prod_{h=1}^p (\sum_{x_i \in C_k} (x_i^h - g_k^h)^2))^{\frac{1}{p}}}{\sum_{x_i \in C_k} (x_i^j - g_k^j)^2} \end{aligned} \quad (4)$$

2.3 算法流程

本小节提出基于属性权重最优化的 k-means 聚类算法: w_k-means 算法. 与传统 k-means 算法类似, w_k-means 算法首先随机生成初始聚类并得到每个聚类的质心, 然后计算各聚类中每个属性的权重, 在此基础上计算每个样本到各质心的距离并将其分配到距离最近的聚类中. 不断循环这一过程, 直到每个聚类中的样本不再发生改变. 具体算法如下.

w_k-means 算法

输入: 包含 n 个样本的数据集 X , 聚类数量 K

输出: K 个聚类

将 n 个样本随机分配到 K 个聚类中, 生成聚类集 $Y = \{C_1, C_2, \dots, C_K\}$;

stop_flag = 1;

while(stop_flag)

stop_flag = 0;

for $k=1$ to K {

计算 C_k 的质心 $g_k = \langle g_k^1, g_k^2, \dots, g_k^p \rangle$;

for $j=1$ to p 计算 λ_k^j ;}

for $i=1$ to n {

计算 $k' = \arg \min_{k=1, \dots, K} \sum_{j=1}^p \lambda_k^j (x_i^j - g_k^j)^2$;

if $x_i \in C_k$ 且 $k' \neq k$ {

$C_{k'} = C_{k'} \cup x_i$; $C_k = C_k - x_i$;

stop_flag=1;}

return Y;

以上算法中,质心的计算方法与传统 k-means

算法相同,即 $g_k^j = \frac{\sum_{x_i \in C_k} x_i^j}{|C_k|}$, $|C_k|$ 为 C_k 中的样本数量。

3 实验及结果

本文分别用合成数据集和实际数据集对 w_k-means 算法进行测试.采用被普遍接受的 CR 系数 (Corrected Rand Index)^[8]来度量聚类结果的质量。

给定数据集 X , 设 $U = \{u_1, \dots, u_i, \dots, u_R\}$ 和 $V = \{v_1, \dots, v_j, \dots, v_C\}$ 是 X 的两个聚类结果, 分别包含 R 和 C 个聚类, 则两者的 CR 系数为

$$CR = \frac{\sum_{i=1}^R \sum_{j=1}^C \binom{n_{ij}}{2} - \binom{n}{2} \sum_{i=1}^R \binom{n_i}{2} \sum_{j=1}^C \binom{n_j}{2}}{\frac{1}{2} \left[\sum_{i=1}^R \binom{n_i}{2} + \sum_{j=1}^C \binom{n_j}{2} \right] - \binom{n}{2} \sum_{i=1}^R \binom{n_i}{2} \sum_{j=1}^C \binom{n_j}{2}} \quad (5)$$

式中, n_{ij} 表示 u_i 和 v_j 同时包含的样本数量, n_i 表示 u_i 的样本数量, n_j 表示 v_j 的样本数量, n 为总样本数量. CR 系数的取值范围为 $[-1, 1]$, 越接近 1, 表明两种结果的一致程度越高。

实验中分别采用传统 k-means 算法和 w_k-means 算法对数据集进行聚类, 并计算聚类结果与数据集真实分类结果的 CR 系数, 以测试聚类结果与真实分类结果的一致程度. 实验按照蒙特卡洛机制重复 100 次, 取 100 个 CR 系数的平均值作为最终结果. 实验系统为 Windows7, 编程环境为 R Studio.

3.1 合成数据集

为了检验 w_k-means 算法对不同数据集的聚类性能, 我们按照二元正态分布, 生成了两个包含有 350 个样本点的数据集. 数据集包含 3 个类别, 其中数据集 1 中的样本类别区分明显, 而数据集 2 的类别区分不明显. 两个数据集的每个类别样本点的均值和方差如表 1 所示, 其散点图分别如图 1 和图 2 所示。

表 1 构造合成数据集

类别	数据集 1		数据集 2	
class1	$\mu_x = 28$	$\sigma_x^2 = 100$	$\mu_x = 45$	$\sigma_x^2 = 100$
150 个样本	$\mu_y = 22$	$\sigma_y^2 = 9$	$\mu_y = 22$	$\sigma_y^2 = 9$
class2	$\mu_x = 60$	$\sigma_x^2 = 9$	$\mu_x = 60$	$\sigma_x^2 = 9$
150 个样本	$\mu_y = 30$	$\sigma_y^2 = 144$	$\mu_y = 30$	$\sigma_y^2 = 144$
class3	$\mu_x = 45$	$\sigma_x^2 = 9$	$\mu_x = 52$	$\sigma_x^2 = 9$
50 个样本	$\mu_y = 38$	$\sigma_y^2 = 9$	$\mu_y = 38$	$\sigma_y^2 = 9$

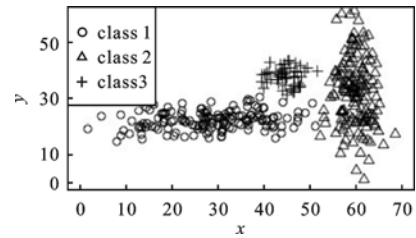


图 1 数据集 1 的散点图

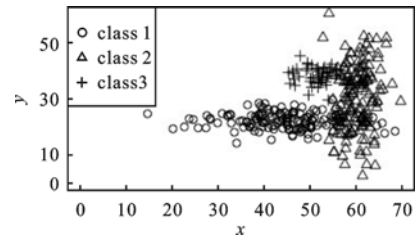


图 2 数据集 2 的散点图

分别应用传统 k-means 算法和 w_k-means 算法对以上两个数据集进行聚类分析, 将聚类结果与样本的真实类别比较, 得到的平均 CR 系数如表 2 所示。

从表 2 的实验结果来看, 本文提出的 w_k-means 算法在聚类准确性上要远优于传统 k-means 算法。

表 2 合成数据集聚类结果的平均 CR 系数

数据集	传统 k-means	w_k-means
数据集 1	0.585 2	0.937 1
数据集 2	0.346 3	0.559 4

3.2 真实数据集

本文采用美国加州大学 UCI Machine Learning 数据库中的三个真实数据集 (Iris、Wine、Haberman's Survival) 来测试 w_k-means 算法的性能. 这些数据集的信息如表 3 所示。

表 3 三个真实数据集

数据集	样本数量	属性数量	类别数量
Iris	150	4	3
Wine	178	13	3
Haberman's Survival	306	3	2

应用传统 k-means 算法和 w_k-means 算法对以上三个数据集进行聚类分析,将聚类结果与样本的真实类别比较,得到的平均 CR 系数如表 4 所示。

表 4 真实数据集聚类结果的平均 CR 系数

数据集	传统 k-means	w_k-means
Iris	0.670 8	0.834 1
Wine	0.365 4	0.768 4
Haberman's Survival	-0.003 1	0.125 7

实验结果表明,对于样本分类较好的数据集(如 Iris, Wine),传统 k-means 能够得到可接受的聚类结果,而 w_k-means 聚类结果的准确性要远好于传统 k-means 算法。但对于样本类别区分不明显的数据集(如 Haberman's Survival),两个算法均无法获得理想的聚类结果。尽管如此,w_k-means 算法在性能上仍优于传统 k-means 算法。

4 结束语

本文将聚类算法中属性权重值的确定作为一个最小优化问题,以聚类算法的原始定义为目标函数,结合属性权重的约束条件,通过拉格朗日乘数法来计算聚类收敛过程中的最优属性权重值。将这一方法嵌入到传统 k-means 算法中,本文提出了基于属性加权的 w_k-means 聚类算法。实验结果验证了该算法相对于传统 k-means 算法的优越性。

参考文献:

- [1] Anil K Jain, Narasimha M Murty, Patrick J Flynn. Data Clustering: A Review[J]. ACM Computing Sur-

veys, 1999,31(3):264-323.

- [2] TsaiChieh-Yuan, Chiu Chuang-Cheng. Developing a feature weight self-adjustment mechanism for a K-means clustering algorithm[J]. Computational Statistics and Data Analysis, 2008,52(10): 4658-4672.
- [3] Liping Jing, Michael K. Ng, Joshua Zhexue Huang. An entropy weighting K-means algorithm for subspace clustering of high-dimensional sparse data[J]. IEEE Trans on Knowledge and Data Engineering, 2007,19(8):1026-1041.
- [4] Yong B. Lim, Yeo Jung Park, Myung-Hoe Huh. D-optimality criterion for weighting variables in K-means clustering[J]. Journal of the Korean Statistical Society, 2009,38(4):391-396.
- [5] Wang Qiang, Ye Yun-ming, Huang J Z. Fuzzy K-means with variable weighting in high dimensional data analysis[C]// Proceedings of the 9th International Conference on Web-Age Information Management. Beijing:Springer,2008.
- [6] 武优西,侯丹丹,李建满,等. 属性权重聚类算法的研究[J]. 小型微型计算机系统,2012,33(3):651-654.
- [7] Diday E, Govaert G. Classification automatique avec distances adaptatives[J]. R. A. I. R. O. Informatique Computer Science, 1977,11(4):329-349.
- [8] Lawrence Hubert, PhippsArabie. Comparing Partitions[J]. Journal of Classification, 1985, 2(1): 193-218.

作者简介:

熊平男,(1974—),博士,副教授.研究方向为数据挖掘、信息安全。

顾霄女,(1990—),硕士研究生.研究方向为数据挖掘。

(上接第 39 页)

- [6] Nguyen K,Cao J. K-Graphs: Selecting top-k data sources for XML keyword Queries [C] //Proc of 22nd Int'l Conf on Database and Expert Systems Applications. Heidelberg: Springer-Verlag, 2011: 425-439.
- [7] 邓松,万常选,刘喜平,等. 基于主题语义的非合作结构化 Top-N 深网数据源选择[J]. 计算机研究与发展, 2012,49(S1):58-64.

作者简介:

邓松男,(1982—),博士研究生,讲师.研究方向为 Web 数据管理、数据挖掘。

万常选男,(1962—),教授,博士生导师.研究方向为 Web 数据管理、XML 信息检索、情感计算和金融数据挖掘。

吁亮男,(1983—),硕士,工程师.研究方向为数据管理、数据分析。

刘德喜男,(1975—),博士,副教授.研究方向为 Web 数据管理、XML 信息检索和自然语言处理。

雷刚男,(1974—),博士研究生,副教授.研究方向为 Web 数据管理、信息检索与挖掘。

王映龙男,(1974—),博士,教授.研究方向为 Web 数据管理、数据挖掘。