第 29 卷第 2 期　　四川理工学院学报(自然科学版)　　Vol. 29　No. 2
2016 年 4 月　　Journal of Sichuan University of Science & Engineering(Natural Science Edition)　　Apr. 2016
文章编号:1673-1549(2016)02-0025-05　　DOI:10.11863/j.suse.2016.02.06

基于预测强度的变量自动加权 K－Means 算法的研究与应用

盛靖友, 张洪伟

(成都信息工程大学计算机学院, 成都 610225)

摘　要:为了克服传统 K－Means 算法 k 值不能确定问题和不具备变量自动选择能力,将预测强度和变量自动加权 K－Means 算法相结合,提出基于预测强度的变量自动加权 K－Means 算法。预测强度表示聚类模型对未知数据的预测能力,预测能力越强,则聚类结果越佳,主要用于 k 值的确定;变量自动加权 K－Means 算法具有在聚类过程中自动调整变量权重的能力,对于噪声变量和冗余变量削弱其对距离的贡献,使聚类结果反映最真实的聚类结构。实验表明,算法具有较强的分类能力和预测能力。

关键词:K－Means;预测强度;变量自动加权

中图分类号:TP301.6　　**文献标志码**:A

引　言

聚类分析是数据挖掘的一项基本任务[1],K－Means 算法是一种非常经典的聚类算法,因为其简单、高效和快速收敛的特点而得到广泛应用[2]。K－Means 算法在聚类过程中将数据对象中的每个变量都同等对待,但在大多数情况下真实的聚类结构仅限于变量集合的一个子集上,噪声变量和冗余变量通常会掩盖真实的聚类结构;同时,K－Means 算法在执行前,需要事先指定 k 值,而在大数据聚类时几乎是不可能提前预知 k 值的。

目前,针对 k 值问题研究主要有两个方向:(1)基于数据的背景资料,对聚类数给出建议。但是其主观性较强,缺乏必要的客观数据支撑,得到的经验不具有一般实用性;(2)建立检验聚类有效性的函数指标,通过有效性指标寻找较优的 k [3]。第二种方式是目前研究的热点。文献[4]通过理论证明了广泛使用的 k 值取值范围的经验规则 $k \leqslant \sqrt{n}$,并提出了一种距离代价函数作为 k 值选取的有效性指标;文献[5]根据类间相似度最大和类内相似度最小原则对文献[4]中的有效性指标进行了改进,并取得了很好的效果;文献[6]采用二分 K－Means算法和类间相似性度量指标反复对数据簇进行分离合并,最终得出聚类结果;文献[7]通过类间相异度和类内相异度决定聚类初始中心和 k 值;文献[8]采用势函数法确定初始聚类中心的同时也确定了 k 值;文献[9]采用基于密度的思想构建聚类数准则函数;文献[3]总结了许多聚类有效性函数指标,并提出以多个有效性指标相结合的方式来决定 k。

聚类分析是无监督学习[10],学习到的模型对未知数据的预测能力称为泛化能力,如果一味追求对已知训练数据准确率,而忽略对未知数据的预测能力,则会出现过拟合的现象[11]。预测强度作为衡量训练模型对未知数据预测能力强弱的指标,很早就被提出来了,预测强度的思想是一个好的聚类结果应该能对未知的样本进行预测,且预测的结果与该样本自身的聚类结果做到很好的吻合[12]。变量自动加权 K－Means 算法是 K－Means 算法的一种变体,通过动态调整各变量的权值,真实地反映各变量对距离的贡献,从而得到更真实的聚类结

收稿日期:2016-02-23

作者简介:盛靖友(1989-),男,四川巴中人,硕士生,主要从事计算智能方面的研究,(E-mail)583155350@qq.com

构[13]。本文通过变量自动加权的 K – Means 算法和预测强度相结合,提出基于预测强度的变量自动加权 K – Means算法,并通过实验证明了它的有效性。

1 算法体系结构

设数据集 $D\{x_1,x_2,\ldots x_n\}$,其中 $x_i(x_{i1},x_{i2},\ldots x_{im},y_i)$ 是集合 D 中的数据对象,$x_{i1},x_{i2},\ldots x_{im}$ 为数据对象的属性,本文中也称变量,y_i 为数据对象 x_i 的目标属性。

1.1 K – Means 算法

首先,指定将数据集 D 聚成 k 个簇,$2 \leqslant k < n$。K – Means算法执行步骤如下[14]:

Step1:从集合 D 中随机选取 k 个数据点作为初始簇中心 $C(c_1,c_2,\ldots c_3)$;

Step2:将数据集 D 中的元素划分到距离最近的簇。距离计算公式如下:

$$d(x_i,c_k) = \sqrt{\sum_{j=1}^{m}(x_{ij}-c_{kj})^2} \tag{1}$$

Step3:根据划分结果重新计算所有数据簇的每个维度的算术平均值,作为新的簇中心;

Step4:比较新旧簇中心是否变化。如果变化转到步骤 Step2;否则,算法结束,输出结果。

使用误差平方和[15]作为度量聚类质量的目标函数:

$$P = \sum_{l=1}^{k}\sum_{i=1}^{|C_i|}\sum_{j=1}^{m} d(x_{ij},c_{lj})^2 \tag{2}$$

1.2 变量自动加权 K – Means 算法

变量自动加权 K – Means 算法中,变量权重的分配原则是给那些重要的变量赋予较大的权重,冗余变量和噪声变量分配较小甚至 0 权重。在每次 K – Means 迭代过程中,变量权重自动调整。

在开始算法前指定各变量的初始权重为 $w\{w_1,w_2,\ldots w_m\}$。其中 w_j 表示第 j 维变量权重,值为:

$$w_j = \frac{1}{m}$$

即初始指定相同的权重。指定类簇数为 k。变量自动加权 K – Means 算法执行步骤如下:

Step1:从 D 中随机选择 k 个初始点作为类簇初中心 $C(c_1,c_2,\ldots c_k)$;

Step2:将数据集 D 中的所有数据对象按照加权距离最近原则划分到与之距离最近的簇。其中关于距离的计算如下:

$$d(x_i,c_k) = \sum_{j=1}^{m} w_j^{\beta}(x_{ij}-c_{kj})^2 \tag{3}$$

Step3:根据划分结果重新计算所有簇中各自维度的算术平均值,作为新的簇中心;

Step4:如果计算前后簇中心未改变,结束算法,输出结果。否则,按照当前聚类结果调整变量权值,然后转到 Step2。变量权值调整方式如公式(4)所示。理论证明见文献[13];

$$w_j = \begin{cases} 0, if(D_j = 0) \\ \dfrac{1}{\sum_{t=1}^{h}(\frac{D_j}{D_t})^{\frac{1}{\beta-1}}} if(D_j \neq 0) \end{cases} \tag{4}$$

其中 h 表示 $D_j \neq 0$ 的变量的个数。D_j 计算方式如下:

$$D_j = \sum_{t=1}^{k}\sum_{i=1}^{|c_t|} d(x\,ij,c_{ij}) \tag{5}$$

度量聚类质量的目标函数为:

$$P = \sum_{l=1}^{k}\sum_{i=1}^{|C_i|}\sum_{j=1}^{m} w_j^{\beta} d(x_{ij},clj) \tag{6}$$

其中 β 作为参数,取值范围为 $\beta < 0$ 和 $\beta > 1$。

1.3 预测强度

预测强度的计算方式如下:

Step1:随机划分 D 为训练集和测试集 $D_{tr} \cup D_{te}$;

Step2:调用聚类算法对 D_{tr} 和 D_{te} 分别聚类,得聚类结果分别为 $\{C_{tr}^l\}_{l=1}^k$ 和 $\{C_{te}^p\}_{p=1}^k$;

Step3:使用 $\{C_{tr}^l\}_{l=1}^k$ 对 D_{te} 中的所有数据对象按相似度最大原则进行划分,得划分结果为 $\{\widetilde{C}_{te}^l\}_{l=1}^k$;

Step4:计算预测强度:

$$ps(k) = \frac{1}{k}\sum_{p=1}^{k}\left(\frac{\sum_{x_i,x_j \in C_w^p}\delta_{ij}}{(|C_{te}^p|(|C_{te}^p|-1)}\right) \tag{7}$$

其中,$|C_{te}^p|$ 表示该簇数据对象个数;δ 为 0 – 1 矩阵,其计算公示式为:

$$\delta_{ij} = \begin{cases} 1, if(i \neq jandx_i,x_j \in C_{te}^p) \\ then\ \exists l(p) \in x_i,x_j \in \widetilde{C}_{te}^l \\ 0, others \end{cases} \tag{8}$$

1.4 算法总结

本文中将预测强度和变量自动加权 K – Means 算法相结合,预测强度作为最终 k 值选取的标准,采用穷举法搜索较优 k 值,k 值搜索范围为 $[2,\sqrt{n}]$ [4]。使用得到的 k 值对数据集进行聚类,聚类结果作为训练模型,通过训练模型对未知数据样本进行分类,得到最终的分类结果。算法步骤如下:

Step1:设置初始值 k 为 2,随机划分 $D = D_{tr} \cup D_{te} \cup$

D_{tv} 为训练集、测试集和验证集;

Step2:调用自动加权 K－Means 算法对 D_{tr} 和 D_{te} 分别聚类,得聚类结果分别为 $\{C_{tr}^{l}\}_{l=1}^{k}$ 和 $\{C_{te}^{p}\}_{p=1}^{k}$;

Step3:使用模型 $\{C_{tr}^{l}\}_{l=1}^{k}$ 对 D_{te} 划分,得分类结果 $\{\widetilde{C}_{te}^{l}\}_{l=1}^{k}$。利用 $\{C_{te}^{p}\}_{p=1}^{k}$ 和 $\{\widetilde{C}_{te}^{l}\}_{l=1}^{k}$ 根据式(7)计算预测强度 $ps(k)$;

Step4: $if(k \leqslant \sqrt{n})$, 则 $k++$, 转到 Step2;否则,转到 Step5;

Step5:取 $\max[ps(k)]$ 为最终 k 值,合并 $D_{tr} \cup D_{te}$ 为 D_{tr2}。调用自动加权 K－Means 算法对 D_{tr2} 进行聚类,得聚类结果为 $\{C_{tr2}^{l}\}_{l=1}^{k}$, 使用多数投票法[16]标记各类簇目标属性,计算聚类准确率;

Step6:使用模型 $\{C_{tr2}^{l}\}_{l=1}^{k}$, 按照加权距离最近原则,即式(3),对 D_{tv} 进行划分,得最终划分结果 $\{C_{tv2}^{l}\}_{l=1}^{k}$, 即是所求。利用多数投票法标记 $\{C_{tv2}^{l}\}_{l=1}^{k}$ 中各类簇目标属性并计算分类准确率,从而得分类结果。

2　实验与结果分析

本文选用 UCI Machine Learning Repository 中的数据集 Iris Dataset 和 Wine Dataset 进行实验。实验过程中对数据采用归一化处理,归一化为 [0,1] 的数,以便降低各维度数据之间的差距,使用的归一化方法为:

$$x'_{ij} = \frac{(x_{ij} - \min(x_{1j}, x_{2j} \ldots x_{nj})) \times 1000}{\max(x_{1j}, x_{2j} \ldots x_{nj}) - \min(x_{1j}, x_{2j} \ldots x_{nj})} \quad (9)$$

实验结果性能评价标准采用训练集聚类正确数(TC)、训练集聚类正确率(TP)、验证集分类正确数(VC)、验证集分类准确率(VP)。

2.1　Iris Dataset 实验结果分析

Iris 数据集中包含 150 个样本数据,4 个连续属性和一个离散的目标属性。其中按照目标属性分类可分为三类,分别为 setosa、versicolor、virginica,每种类型有 50 个样本。

实验前,按照经验规则,将 k 值取值范围设定为 $2 \leqslant k \leqslant \sqrt{150}$。实验中,随机将 Iris 样本划分为训练集、测试集和验证集。其中训练集为 105 个,每种类型各 35 个;测试集 30 个,每种类型 10 个;验证集 15 个,每种类型 5 个。

试验中,由于 Iris 样本数据偏少,首先使用训练集和测试集寻找 k, 然后将训练集和测试集合并,作为新的训练集,此时训练集变为 135 个,使用已知 k 值构建训练模型。对于指定 k 值聚类时,由于变量自动加权 K－Means 算法也具有局部收敛的特点,所以多次调用聚类算法,选式(6)取值最小的聚类结果为本次 k 值的较优聚类结果。通过 10 次调用算法,其最终结果见表 1。

表 1　Iris 10 次实验结果数据

Num	K	TC	TP	VC	VP
1	3	120	88.89%	14	93.33%
2	3	121	89.63%	13	86.66%
3	3	119	88.15%	15	100.00%
4	3	119	88.15%	15	100.00%
5	3	119	88.15%	14	93.33%
6	3	120	88.89%	14	93.33%
7	3	121	89.63%	13	86.67%
8	3	120	88.89%	14	93.33%
9	4	120	88.89%	13	86.67%
10	3	119	88.15%	15	100.00%

对训练集训练结束后,各变量权值见表 2。

表 2　变量权值

Num	Weight1	Weight2	Weight3	Weight4	Sum
1	0.142783	0.256861	0.147675	0.452682	1.0
2	0.152763	0.258431	0.145419	0.443388	1.0
3	0.155535	0.261948	0.146158	0.436359	1.0
4	0.145826	0.263302	0.155047	0.435826	1.0
5	0.134517	0.242064	0.155165	0.468254	1.0
6	0.156408	0.390813	0.0985356	0.354244	1.0
7	0.145058	0.257229	0.148772	0.448942	1.0
8	0.130488	0.241134	0.143581	0.484798	1.0
9	0.146387	0.24838	0.156004	0.44923	1.0
10	0.165894	0.42193	0.0908914	0.321285	1.0

通过表 1 数据分析,10 次调用该算法, 其中 9 次得出最终 k 值为 3,1 次为 4,与数据真实分类情况基本相符。使用得出的 k 值,生成训练模型,然后对验证集样本进行分类。分类结果见表 1 中 VC 和 VP,其中最多时,只有 2 个样本分类错误,由此可得结论:该训练模型有较强的分类能力。通过表 2 中可以看出 Weight2 和 Weight4 明显大于其它权重,说明属性 2 和属性 4 相比与其它属性的重要性更高,对距离贡献更大,同时削弱了属性 1 和属性 3 的作用。由此可得结论:该算法能自动调整各变量的权值,这在高维聚类中可以有效的识别和噪声变量和冗余变量,并同时降低他们对聚类结果的影响。

2.2　Wine Dataset 实验结果分析

Wine Dataset 是对意大利某地 3 种不同品种的葡萄酒进行化学分析后得到的数据,共有 13 个属性和一个目标属性,目标属性分为 1、2、3 类,依次有 59、71、48 个

样本。

试验中随机划分 Wine Dataset 为训练集、测试集和验证集,其中训练集 108 个样本,测试集 36 个样本,验证集 34 个样本。首先,使用训练集和测试集得出较优 k 值;然后将训练集和测试集合并成新的训练集,使用求得的 k 值构建训练模型;最后使用该训练模型,对保留的验证集进行分类。10 次实验结果见表 3。

表 3 Wine Dataset 10 次实验结果

Num	K	TC	TP	VC	VP
1	4	140	97.22%	33	97.06%
2	3	138	95.83%	31	91.18%
3	5	139	96.53%	32	94.11%
4	3	136	94.44%	33	97.06%
5	4	139	96.53%	33	97.06%
6	4	140	97.22%	32	94.11%
7	3	139	96.53%	30	88.24%
8	4	140	97.22%	32	94.12%
9	4	139	96.53%	33	97.06%
10	4	137	95.12%	33	97.06%

对表 3 数据进行分析,10 次算法结果中,k 值 6 次取 4,3 次取 3,一次取 5,k 取值与真实数据样本类别划分本相符。10 次对 144 个训练样本的训练中,最高有 140 个样本分类正确,最少也有 138 个分类正确。对验证集的分类准确率结果也比较高,验证集样本总数 34 个,其中最高时能达到 33 个。由此可得结论:算法能得出较合理的 k,且聚类准确率和预测准确率均较强。

3 结束语

针对 K – Means 算法 k 值不确定问题和不能进行变量选择问题,提出将变量自动加权 K – Mean 算法和预测强度相结合,通过预测强度寻找较优的 k,在聚类过程中为变量增加自适应权值,能有效削弱、甚至剔除冗余变量和噪声变量对聚类结构的影响,这在高维聚类中非常有用。实验表明,该算法能有效得出与真实数据类别数目比较接近的 k 值,且算法有较强的分类能力和预测能力。

参 考 文 献:

[1] 陈苏蓉,朱晓辉.基于模糊逻辑的 K-Means 算法研究[J].计算机工程与科学,2012,34(12):155-159.

[2] RAJESWARI k,ACHARYA O,SHARMA M,et al. Improvement in K-Means Clustering Algorithm for Data Clustering[C]//Proceeding of 2015 International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, February 26-27, 2015:367-369.

[3] MAEDER A,MATAWIEK,MEHAR AM.Determining an Optimal Value of K in K-means Clustering[C]// Proceeding of 2013 IEEE International Conference on Bioinformatics and Biomedicine(BIBM 2013),Shanghai, December 18-21,2013:51-55.

[4] 杨善林,李永森,胡笑旋,等.K-means 算法中的 K 值优化问题研[J].系统工程理论与实践.2006,26(2):97-101.

[5] 韩凌波.一种新的 K-means 最佳聚类数确定方法[J].现代计算机:专业版,2013(30):12-15.

[6] LIN Y J,LUO T,YAO S,et al.An improved clustering method based on K-Means[C]//Proceeding of 2012 9th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD2012),Chongqing,China,May 29-31,2012:734-737.

[7] 张杰,卓灵,朱韵攸.一种 K-Means 聚类算法的改进与应用[J].电子技术应用,2015,41(1):125-128.

[8] 叶于林,夏秀渝,莫建华,等.对 K-Means 及势函数聚类算法的研究与改进[J].计算机系统应用,2015,24(4):209-213.

[9] 张琳,陈燕,汲业,等.一种基于密度的 K-Means 算法研究[J].计算机应用研究,2011,28(11):4071-4073.

[10] TAN L. A Clustering K-means Algorithm based on improved PSO Algorithm[C]//Proceeding of 2015 Fifth International Conference on Communication Systems and Network Technologies, Gwalior, India, April 4-6, 2015:940-944.

[11] 李航.统计学习方法[M].北京:清华大学出版社,2012:15-19.

[12] 王星.大数据分析:方法与应用[M].北京:清华大学出版社,2013:111-115.

[13] HUANG J Z,NG M K,RONG H Q,et al.Automated variable weighting in K-Means type clustering[J].IEEE Transactions on Pattern Analysis and Machine Intelli-

gence,2005,27(5):657-668.

[14] LAN X,LI Q,ZHENG Y. Density K-Means: A new algorithm for centers initialization for K-Means [C]// Proceeding of 2015 6th IEEE International Conference on Software Engineering and Service Science (ICSESS),Beijing, September 23-25,2015:958-961.

[15] 许晴,李凡长,邹鹏.基于 Finsler 几何的 k-means 算法[J].中国科学技术大学学报,2014(7):570-575.

[16] 徐太征,徐中宇.Fisher 理论和多数投票法相结合的数据融合算法[J].科技信息,2009(27):96.

Research and Application of Automatically Variable Weighting K-Means Algorithm Based on Forecasting Intensity

SHENG Jingyou, *ZHANG Hongwei*

(College of Computes, CUIT, Chengdu 610225, China)

Abstract: In order to overcome the problems of unknown *k* value before clustering and less properties automatically choosing ability of conventional K-Means algorithm, automatically variable weighting K-Means algorithm based on forecasting Intensity is proposed. Forecasting Intensity indicates the forecasting capability of training mode to those extra unknown data sample. Better the forecasting is, better the clustering results are. It was mainly used to get the *k* value of clustering algorithm. Automatically variable weighting K-Means algorithm could adjust the weights of each properties. This can reduce the distance which noise properties and redundant properties contributed in the processing of clustering and reflect the real cluster structure. This article combines forecasting intensity, automatically variable weighting K-Means algorithm and empirical results show its classification ability and forecasting ability.

Key words: K-Means; forecasting intensity; automatically variable weighting