文章编号：1000-8608(2021)06-0645-07

基于 M-distance 算法思想的优化加权 KNN 算法

程 勖，高雍政，郭 芳*

（大连工业大学 管理学院，辽宁 大连 116032）

摘要：为快速对数据进行特征选择以实现精确分类，采用 M-distance 算法思想进行数据集簇聚类，对样本数据进行预处理；设计加权 K 近邻算法缩减样本间距并构建样本分类模型；采用模拟简谐振动的方法遍历样本数据，求解最优加权特征向量，实现样本分类. 实验结果表明：设计的算法是正确的，分类模型是合理的. 在样本数据特征中，分离出的消费者最为关心的前 10 个样本特征符合消费者的行为选择，说明算法设计有一定实用性.

关键词：样本近邻；加权特征向量；谐振子；样本分类

中图分类号：TP181 **文献标识码：**A **doi：**10.7511/dllgxb202106012

0 引 言

数据特征选择已成为推荐系统重点研究领域，目的是快速实现最优分类，且要精准识别特征信息. 数据特征选择主要通过两种途径实现：其一，通过用户商品数据库进行过滤，采用矩阵分解[1]、Slope One[2]；其二，学习用户偏好的描述性模型，采用评分机制，如贝叶斯网络[3]、神经网络分类[4]等. 常见的分类方法有决策树、KNN 算法、SVM 算法、贝叶斯网络、神经网络等. 传统 KNN 算法因其简单高效最为常用. 它是一种惰性分类算法，特点在于样本数据不需要训练，使用便捷，但是时间复杂度较高，且将单一变量（距离）作为相似度衡量标准，导致推荐精度低且体验度不够完善，如 Jaccard 相似性[5]、Manhattan 距离[6]等. 针对 KNN 算法的不足，主要解决方案集中在裁剪[7-9]与降维[10-12]两个方面. 裁剪方法虽然可快速去除噪声数据，但以损失分类精度为代价；降维方法虽然提高了运算速度，但以牺牲特征数据为代价. 如何在保证分类精度的前提下，降低时间复杂度，提升运算速度备受关注.

本文通过 M-distance 算法思想进行簇聚类，对样本数据进行预处理. 相对于 k-means 方法，它的时间复杂度比较小，聚类效果比较显著[13]. 然后对数据进行加权处理，以便缩小数据间的距离，并兼顾数据之间的相关性，更加准确地对数据进行分类. 最后，通过简谐振动原理，计算数据相似度距离，优化遍历数据过程，提高搜索速度.

1 加权 K 近邻算法的特征选择方法

1.1 K 近邻算法简介

在模式识别中，K 近邻算法的优势在于简单且对异常值不敏感，泛化能力强. 其核心思想内容[14-15]：任意样本在数据集中的 K 个最相似样本中，如果大部分样本归并为某一类别，那么此样本也属于这个类别. 其意义在于通过待测样本周边的已知数据类别来预测此样本的归属问题，极大弱化了数据集的高维度、高耦合给数据特征分析带来的困难. 优点：对数据噪声有较强的忍耐能力，非常适合零售业复杂多变的特征选择情景；分析数据特征时，只关注最相邻的 K 个样本即可，极大地降低了数据集的维度. 缺点：K 近邻算法属于惰性算法，不能主动学习；K 值的选择对数据分析的结果有影响.

Ayyad 等[16]提出一种加权策略（MKNN）算法，在基因序列分类中，已知类别中心样本到其他样本加权距离，使其待测样本的辨别程度不同，但

收稿日期：2021-03-19； **修回日期：**2021-09-16.
基金项目：国家自然科学基金资助项目(71802035)；辽宁省纺织之光教学改革研究资助项目(BKJGLX333).
作者简介：程 勖(1980-)，男，博士，副教授，E-mail：chengxu_666@163.com；郭 芳*(1974-)，女，副教授，E-mail：80118596@qq.com.

此方法对权值的选取并没有考虑特征的重要程度. Li 等[17]通过随机选择特征子集,使用 KNN 算法评价其分类准确率,并计算其平均值作为置信度的衡量标准. 该方法的泛化能力提高了,但是对权值贡献度的衡量有待评估,且没有说明权值是如何影响分类效果的. 文献[18]提出的计算近邻算法,通过设置安全边际点来完善 KNN 模型,并未考虑样本之间距离关系对模型的影响. 文献[19]提出的移动 K 近邻查询算法,采用双目标函数通过调节两者之间的作用权重,来求解最优目标函数值. 但是对于例如零售业样本集维度高、参数调节困难等情况,必将直接导致计算能力失衡. 文献[20]提出的自然邻居算法,主要将自然邻居特征值作为所有样本的近邻,从而避免了 K 值的人为设定,但是其对所有样本都采用同样的 K 值,会导致在稀疏矩阵的样本中,由于分布密度过高而导致计算速度大幅下降. 为将样本集按照某种条件进行排序,在测试集参与计算之前,将离群点或弱相关样本直接删除,再将经过处理的数据集按照升序或降序进行有规律的计算,可能会更加快速地得到合理的 K 值.

综上分析,在特征分析过程中,不仅要考虑每个待测样本之间的距离对类别的贡献,更要把每个特征对样本分类的重要程度融入权值计算当中.

1.2 基于 M-distance 思想的加权 K 近邻数据特征选择

本文将样本特征的重要性作为参数融入求解权值当中,简称为 IFKNN(importance feature integrate into K-nearest neighbors)算法. 首先获取待测样本数据,将数据中的个体作为初始数据特征,使用 IFKNN 算法预测样本类别,并设计谐振子摆动的目标函数,最后通过经典谐振动的接受准则搜索最优数据特征.

1.2.1 相关定义 本文针对新零售业作相关数据分析,定义相关数据集. 数据集 $X=\{x_i\}_1^M$ 表示 M 个样本构成的集合,$x_i=(x_{i1},x_{i2},\cdots,x_{iN})$ 是第 i 个样本的观测值;样本特征集合为 $F=\{f_1,f_2,\cdots,f_N\}$;数据集 $Y=\{y_i\}_1^N$ 是目标样本构成的集合. 每个样本的评价数量 $B=\{b_i\}_1^p$,每个样本的月成交量 $C=\{c_i\}_1^q$.

定义样本数据贡献度,作为样本最主要的数据簇. 贡献度簇集表达样本在整个观测数据集中作用大小. 贡献度越大,样本特征越显著,即商品成交概率越大. 也就是说贡献度对商品成交影响是正向的,那么影响贡献度的主要因素,就可以作为重要的营销特征.

每个样本在成交时,它的评价数量与月成交量的商作为此样本对成交的贡献度,记为 A. $A=\{\alpha_1,\alpha_2,\cdots,\alpha_n\}(0.005\leqslant\alpha_i\leqslant1)$,$\alpha_i=b_i/c_i$.

通过贡献度阈值的设定,可以更加精确地筛选观测样本数据,也就是说,围绕在中心点周围的样本基本都是线性相关的,几乎没有异常值. 如果 $\alpha_i<0.005$(参照相关系数设定)就认为此样本属于弱相关或无关样本,直接删除[21].

在初步过滤样本数据时,将样本观测数据集按照样本贡献度的大小降序排列. 这样在遍历样本的过程中,不仅减少了求样本中心点的迭代次数,同时,也减小了近邻样本点之间的信息增益,提高了运算速度.

计算相关参数:

(1)对贡献度数据集 A 求均值,$\overline{A}=\dfrac{1}{M}\sum\limits_{i=1}^{M}\dfrac{b_i}{c_i}$.

(2)对贡献度数据集 A 求均方差 δ_A,$\delta_A=\sum\limits_{i=1}^{M}(\alpha_i-\overline{A})$.

(3)将贡献度进行归一化处理得到 A':$A'=\left\{\dfrac{\alpha_i}{\delta_A}\right\}_{i=1}^{M}$.

(4)定义特征权重向量 $\boldsymbol{w}_f=(w_{f_1}\quad w_{f_2}\quad\cdots\quad w_{f_N})$,将每个样本的贡献度融入特征权重中,即 $w_{f_a}=w_{f_a}\cdot\alpha_i'$,设样本数据 x_i 与 x_j 之间的距离为 d_{ij},融入贡献度的特征权重的欧几里得距离公式为

$$d_{ij}=\sqrt{\sum_{a=1}^{N}w_{f_a}^2\cdot(x_{ia}-x_{ja})^2}\tag{1}$$

(5)通过距离公式可以分析出,距离越小,属于同一类别的概率越大. 运用距离倒数加权方法对权值进行加权作用. 距离加权函数为

$$W(d)=\mathrm{e}^{-d}\tag{2}$$

样本特征进行加权后的样本预测数据集为

$$P_i(\boldsymbol{w}_f)=\frac{1}{\sum\limits_{i\neq j}W(d_{ij}(\boldsymbol{w}_f))}\sum_{i\neq j}W(d_{ij}(\boldsymbol{w}_f))\cdot y_i\tag{3}$$

定义损失函数 $L(y_i,P_i)$,使用预测数据集 P 与目标数据集 Y 的差(曼哈顿距离表示),即:

$$L(y_i,P_i)=|y_i-P_i(\boldsymbol{w}_f)| \tag{4}$$

求损失函数 $L(y_i,P_i)$ 的均值函数 $\overline{L}$，$\overline{L}$ 越小，预测误差就越小，说明加权特征求解方法越准确. 公式表示为

$$\overline{L}=\frac{1}{M}\sum_{i=1}^{M}L(y_i,P_i(\boldsymbol{w}_f)) \tag{5}$$

1.2.2　*模拟谐振子运动*　简谐振动系统中谐振子的振动轨迹总是从势能最大(离平衡点距离最远)的一端运动到平衡点，再运动到另一端势能最大处，这样持续作往返运动.

(1)初始状态[22]

基态步长 $L_d=2$，设定 $\overline{A}$ 点为初始振动中心点. 由于贡献度数据集 A 与样本数据集 X 是一一对应的，遍历贡献度数据集 A，可以实现遍历样本数据集 X 的过程，这样可以加快计算速度. 设定目标函数：

$$f(d)=\overline{L}=\frac{1}{M}\sum_{i=1}^{M}L(y_i,P_i(\boldsymbol{w}_f)) \tag{6}$$

(2)设定振幅 S_l

$$S_l=\max(f(d_{il}))-\min(f(d_{jl})) \tag{7}$$

(3)求解振动能量级

取谐振子作简谐振动，计算能级差 U_d：离平衡点越近，能级差越小；反之，能级差越大. 由能级变化式(8)可知，简谐振动的能级以 $\frac{x^2-(x-1)^2}{S^2}$ 递减方式进行[23].

$$U_{d_1}=\frac{1^2}{S^2},\ U_{d_2}=\frac{2^2-1^2}{S^2},\ \cdots,$$
$$U_{d_x}=\frac{x^2-(x-1)^2}{S^2},\ \cdots,$$
$$U_{d_S}=\frac{S^2-(S-1)^2}{S^2} \tag{8}$$

(4)描述问题转换思想

假定将在求的目标函数所有过程解(计算每个过程的最优解)均映射到简谐振动的能量级中. 其中，所有区间的过程解都杂乱无规律地分布其中. 设最大解 $f_{\max}$ 与最小解 $f_{\min}$ 之差等于最大势能 U_{d_S}. 随机计算两个过程解 f_i、f_j，则 $\Delta f=|f_i-f_j|$ 表示两个样本点的过程解之间的相对距离. 为了可视化方便，将其作归一化处理，即：

$$\frac{\Delta f}{f}=\frac{\Delta f}{f_{\max}-f_{\min}} \tag{9}$$

假定 f_i 就是最优解，那么在不断计算的过程中，也就是逐渐靠近最优解的过程. 所以 $\Delta f/f$ 就是新解与最优解的相对距离. 确定了新解的能量区间，就可以确定新解的选择策略(特征值选择方法). 经典简谐振动能级选择策略为将 $f_{\text{end}}-f_i$ 近似看作简谐振动的最大势能.

$$Q=\begin{cases}1; & \Delta f<0\\ \left(\frac{L_i}{L_0}\right)^2-\frac{\Delta f}{f_{\text{end}}-f_i}; & \Delta f\geqslant 0,L\in[L_0,L_i]\\ 0; & \Delta f\geqslant 0,L\in[1,L_0]\end{cases} \tag{10}$$

其中 L_0 为初始步长. 综上分析，搜索最优数据特征采用经典谐振子算法来遍历样本特征空间，就是求解 $\min(\Delta f)$ 最优的过程，最终确定最优特征权值，从而得到最优样本数据特征.

1.2.3　*数据特征选择*　将样本数据集 X 按照贡献度数据集 A 的索引顺序，排列既有前驱也有后继. 即每当遍历一个样本数据时，它能接受前驱的样本特征信息，并将该信息处理完毕直接传递给后继样本，这样求最优解(最优特征)可表达成一个序列：$\varphi_1\rightarrow\varphi_2\rightarrow\cdots\rightarrow\varphi_n$. 即整个解空间 D 的规模变为 $(n-1)!/2$. 同时，这样构建数据结构的好处是，在求解最优特征权值的过程中，保存的最优解向量的索引会一起传递，最终形成最优特征序列，也就是所要挖掘的消费者所关注的消费理念与行为数据.

使用 Python 语言在 PyCharm 集成开发环境下，对样本数据集 X 采用 5 种分类方法进行分类计算. 实验初始化数据集 X 中包含 5 000 个样本数据，每个样本设置一个权重向量并初始化为 $\mathbf{0}$. 对当前数据集 X 的每个样本用 IFKNN 算法计算出预测值 P 以及损失函数 $L(y_i,P_i)$ 的均值函数 $\overline{L}$，并将其作为模拟简谐振动的目标函数 $f(d)$. 为了找到谐振子整个能量级最优解，将求解 $\min(\Delta f)$ 每个过程中的局部最优解及所对应的样本信息存入结果级 Results 中，Results 数据集中的索引也是所关注的样本特征序列的索引. 由于样本数据集的数据结构呈链状，并且数据集 X 根据对应贡献度数据集 A 作降序处理，所以得到的最优解同样也是降序储存. 这样所有能量级中 $\min(\Delta f)$ 所对应的 $\boldsymbol{w}_f$ 就是全局最优解.

IFKNN 算法特征选择方法具体流程为

Input：具有 M 个样本和 N 个特征的数据集 X

Output：最优特征

(1)按照前文描述，初始化数据集，求解贡献

度数据集 A. 样本数据集 X 按照样本的贡献度数据集 A 中的索引顺序从大到小依次排列，并且首尾相连，构建成链状数据结构，样本都拥有前驱与后继.

(2)计算贡献度数据集 A 的均值与方差.

(3)进行迭代操作：设迭代次数为 T

For($i=0;i<T;i++$)

For：对每个样本进行数据处理

计算数据集 X 中的预测值 P

计算损失函数，将损失函数作为选择最优特征权重的目标函数 f，进行迭代操作，使得目标函数的值最小，也就是最接近最优解，即在遍历数据集时，求解目标函数最小值.

计算 Δf，按照经典简谐振动的选择策略进行筛选. 获得最优解.

2 实验与性能分析

2.1 数据集

以 2020 年“双 11 购物节”部分重要日常生活用品(卫生纸)数据为主要研究对象，以 2020-11-30 为终点抓取消费者购买卫生纸的相关数据. 卫生纸品牌选择具有代表性的两个品牌，作为集合 $Y=${维达，心相印}，对集合 Y 作 Map 处理：$Y=$ {维达：0，心相印：1}. 每种售出商品向前抓取 2 500 个用户留言，即共收集 5 000 个样本进行分析. 在分析用户留言时，利用分词方法，随机抓取频繁出现的关键词 120 个记录到特征数据集 F 中. 每个关键词出现的次数记为 n，按照降序排列，如果次数相同，按照先后顺序，依次存入特征数据集 F 中. 将每个顾客留言出现特征数据的次数除以 5 000 的商，按照特征数据集 F 中样本特征出现的顺序依次记录到数据集 X 中，数据集 X 共存储 5 000 条样本信息.

2.2 实验结果及性能分析

本文通过特征权值的设定重点关注分类的正确性，所以采用准确率(accuracy，A_c)、精准率(precision，P_r)、召回率(recall rate，R)以及综合正确率和召回率的指标 F-Measure. 其中 T_p 为正例被正确预测为正例的数目，T_n 为反例被正确预测为反例的数目，F_p 为反例被错误预测为正例的数目，F_n 为正例被错误预测为反例的数目. 准确率的计算公式为 $A_c=\dfrac{T_p+T_n}{T_p+T_n+F_p+F_n}$. 精准率是针对预测结果而言的，即在所有被预测为正例的样本当中，实际为正例样本的概率. 公式定义为 $P_r=\dfrac{T_p}{T_p+F_p}$. 召回率是针对原样本数据而言的，即在实际为正例的样本中被预测为正例样本的概率. 公式定义为 $R=\dfrac{T_p}{T_p+F_n}$. 但是，精准率与召回率是一对此消彼长的参数，即 P_r 越高，R 就越低. F-Measure 指标是兼顾 P_r 和 R 的加权调和平均. 公式定义为 $F_1=\dfrac{2\times P_rR}{P_r+R}$，$F_1\in[0,1]$. 1 代表模型输出结果最好，0 代表模型输出结果最差.

从图 1 可以直观地看出 IFKNN 算法准确率比较高. 图 2 表明每个邻居样本数据都被特征的选择数量所影响，当特征数量小于 30 的时候，预测出的数据不太符合真实的情景，当样本数据大于 100 的时候，又会发生过拟合，同样也不太适合. 所以，选择样本特征数量对于能否成功得到分类结果是非常关键的. 图 3 为 5 种算法的召回率分布情况. 样本特征数量从 30 到 70，召回率下降比较快而且数值比较高，然后召回率缓慢下降最

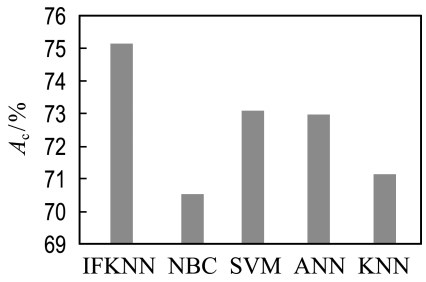

图 1 5 种算法分类的准确率

Fig. 1 Classification accuracy of five algorithms

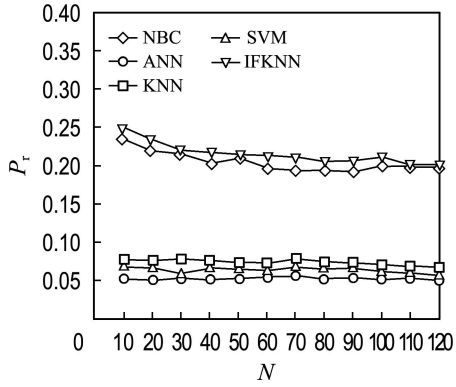

图 2 不同样本特征数量的 5 种算法的精准率

Fig. 2 Precision of five algorithms with different sample feature numbers

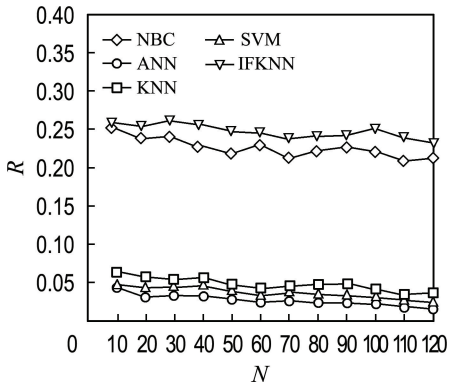

图 3　不同样本特征数量的 5 种算法的召回率

Fig. 3　Recall rates of five algorithms with different sample feature numbers

后趋于一致.

表 1 分别列出 5 种算法获得的最高分类性能(*F*-Measure)和特征数量(N),可以看出,IFKNN 算法的分类性能比其他几种算法好,不仅达到的指标最高,而且使用的特征数量也最少. 由图 4 可以直观地分析出,IFKNN 算法分类性能最高,而且在很多求解过程中都能最快获取最高分类性能. 当选取前 20 个特征时 IFKNN 算法相对其他 4 种算法能明显快速达到最好分类性能. 在样本特征数量为 30～50 时,步调基本趋于一致,这可能是由于在模拟简谐振动的某部分能量级中陷入了局部最优. 当样本特征数量在 40～60 时,IFKNN算法的效果又显著增强. 这也许是由于跳出了局部最优(其他 4 个算法有明显先下降之后显著上升的过程),最后 5 种算法结果趋于一致. 由图 5 可以分析出:在其他参数都是最优值,选择的邻居数量 K 小于 20 的情况下,分类效果是逐步增强的,尤其是当训练数据集比较少的情况下,变动幅度比较大. 当邻居数量 K 大于 20 时,分类效果开始逐步降低,最后趋于稳定. 实验表明,样本邻居数量 K 选择在 20 左右作分类预测时,分类结果比较好. 由图 6 可以看出,消费者最关注的特征是物流服务,而且关注度非常显著. 售后服务质量与商品质量关注度相差不多,而对大家极为看好的红包特征并没有明显关注. 因此,拟订营销策略时,首先要完善物流服务,其次是重视售后服务与商品质量,也就是企业在网商平台上要介绍商品信息,做好售后服务,而并非把注意力放在红包上.

表 1　5 种算法指标对比

Tab. 1　Index comparison of five algorithms

算法	*F*-Measure	N
IFKNN	0.732 7	48
KNN	0.682 4	73
SVM	0.652 8	68
NBC	0.692 7	64
ANN	0.674 6	74

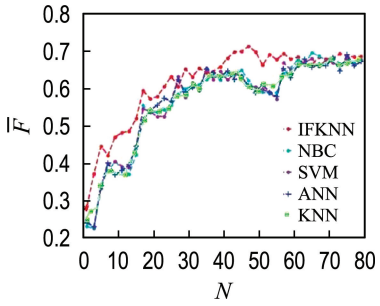

图 4　不同样本特征数量的 *F*-Measure 指标分布

Fig. 4　*F*-Measure index distribution of different sample feature numbers

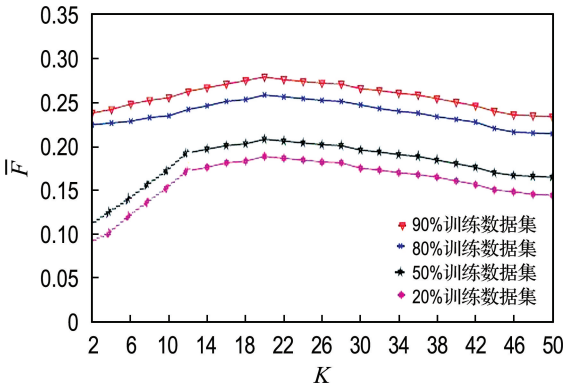

图 5　样本邻居数量选择对指标 *F*-Measure 的影响

Fig. 5　Influence of sample neighbors number selection on index *F*-Measure

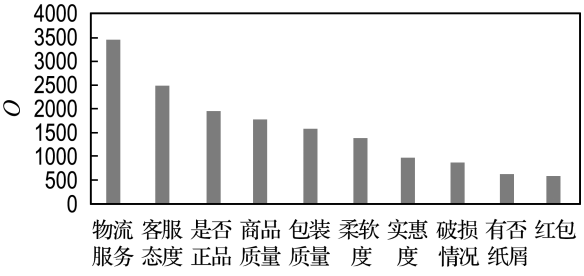

图 6　消费者关注前 10 特征序列分布

Fig. 6　Distribution of top 10 feature sequences of consumer concern

3 结 语

本文提出基于M-distance思想的加权K近邻特征选择算法,在模拟简谐振动算法基础上搜索最优特征权重向量进行精确分类.实验表明:IFKNN算法分类性能最好、速度最快、使用特征向量最少,并且邻居数量K值在20左右效果比较理想.构建的样品特征向量集有助于分析消费者日常关注的消费理念与行为,辅助构建消费策略.在未来的工作中,可以使用其他的计算距离公式提高普适性,利用其他的遍历矩阵方法来搜索最优特征向量,提高分类精度及速度.

参考文献:

[1] ZHANG Fuzhi, LU Yuanli, CHEN Jianmin, *et al*. Robust collaborative filtering based on non-negative matrix factorization and R1-norm [J]. **Knowledge-Based Systems**, 2017, **118**(23): 177-190.

[2] 李桃迎, 李 墨, 李鹏辉. 基于加权Slope One的协同过滤个性化推荐算法 [J]. 计算机应用研究, 2017, **34**(8): 2264-2268.
LI Taoying, LI Mo, LI Penghui. Personalized collaborative filtering recommendation algorithm based on weighted Slope One [J]. **Application Research of Computers**, 2017, **34**(8): 2264-2268. (in Chinese)

[3] PARK H S, YOO J O, CHO S B. A context-aware music recommendation system using fuzzy Bayesian networks with utility theory [J]. **Lecture Notes in Computer Science**, 2006, **4223**: 970-979.

[4] WU S, REN W, YU C, *et al*. Personal recommendation using deep recurrent neural networks in NetEas [C] // **Proceedings of the International Conference on Data Engineering**. India: IJSDA, 2016.

[5] LI Bin, ZHANG Bo, LIU Xuejun, *et al*. A collaborative filtering recommendation algorithm based on Jaccard similarity and locational behaviors [J]. **Computer Science**, 2016, **43**(12): 200-205.

[6] PHUEAKNUMPOL V, BUDSABONG Z, KERDPRASOP K, *et al*. Product recommendation system by approximate search based on Manhattan distance measurement [C] // **Proceedings of the 11th International Conference on Computational Intelligence**. Singapore: Springer, 2012.

[7] TAHIR M A, KITTLER J, MIKOLAJCZYK K, *et al*. A multiple expert approach to the class imbalance problem using inverse random under sampling [J]. **Lecture Notes in Computer Science**, 2009, **5519**: 82-91.

[8] 王艳飞, 郝卫杰, 范支菊, 等. 基于聚类和密度裁剪的改进KNN算法 [J]. 青岛大学学报(自然科学版), 2017, **30**(2): 62-68.
WANG Yanfei, HAO Weijie, FAN Zhiju, *et al*. An improved KNN method for reducing the amount of training samples based on clustering and density [J]. **Journal of Qingdao University (Natural Science Edition)**, 2017, **30**(2): 62-68. (in Chinese)

[9] 肖绍武, 王子牛, 高建瓴. 基于中心抽样的KNN算法在文本分类中的应用 [J]. 贵州大学学报(自然科学版), 2018, **35**(1): 78-81.
XIAO Shaowu, WANG Ziniu, GAO Jianling. The application of KNN algorithm based on central sampling in text categorization [J]. **Journal of Guizhou University (Natural Science)**, 2018, **35**(1): 78-81. (in Chinese)

[10] 殷亚博, 杨文忠, 杨慧婷, 等. 基于搜索改进的KNN文本分类算法 [J]. 计算机工程与设计, 2018, **39**(9): 2923-2928.
YIN Yabo, YANG Wenzhong, YANG Huiting, *et al*. KNN text classification algorithm based on search improvement [J]. **Computer Engineering and Design**, 2018, **39**(9): 2923-2928. (in Chinese)

[11] 张玉芳, 万斌侯, 熊忠阳. 文本分类中的特征降维方法研究 [J]. 计算机应用研究, 2012, **29**(7): 2541-2543.
ZHANG Yufang, WAN Binhou, XIONG Zhongyang. Research on feature dimension reduction in text classification [J]. **Application Research of Computers**, 2012, **29**(7): 2541-2543. (in Chinese)

[12] 刘楠楠. 文本分类中特征降维算法的研究与应用 [D]. 成都: 电子科技大学, 2018: 70-88.
LIU Nannan. Research and application of feature dimension reduction algorithm in text classification [D]. Chengdu: University of Electronic Science and Technology, 2018: 70-88. (in Chinese)

[13] BACHE K, LICHMAN M. UCI Machine Learning Repository [EB/OL]. [2020-10-20]. http://archive.ics.uci.edu/ml.

[14] MEHRABIAN A, RUSSELL J A. **An Approach to**

Environmental Psychology [M]. Cambridge: The MIT Press, 1974.

[15] CHEN C C, LIN Y C. What drives Live-Stream usage intention the perspectives of flow, entertainment, social interaction and endorsement [J]. **Telematics and Informatics**, 2018, **35**(1): 293-303.

[16] AYYAD S M, SALEHL A I, LABIB M. Gene expression cancer classification using modified *K*-nearest neighbors technique [J]. **Biosystems**, 2019, **176**: 41-51.

[17] LI Chuanwen, GU Yu, QI Jianzhong, *et al*. Processing moving KNN queries using influential neighbor sets [J]. **Proceedings of the VLDB Endowment**, 2014, **8**(2): 113-124.

[18] GU Yu, LIU Guanli, QI Jianzhong, *et al*. The moving *k* diversified nearest neighbor query [J]. **IEEE Transactions on Knowledge and Data Engineering**, 2016, **28**(10): 2778-2792.

[19] NUTANONG S, ZHANG R, TANIN E. Analysis and evaluation of V-diagram: An efficient algorithm for moving KNN queries [J]. **VLDB Journal**, 2010, **19**(3): 307-332.

[20] ZHU Qingsheng, FENG Ji, HUANG J. Natural neighbor. A self-adaptive neighborhood method without parameter [J]. **Pattern Recognition Letters**, 2016, **80**(1): 30-36.

[21] 张初兵，李东进，吴　波，等. 购物网站氛围线索与感知互动性的关系 [J]. 管理评论，2017, **29**(8): 91-100.
ZHANG Chubing, LI Dongjin, WU Bo, *et al*. The relationship between atmospheric cues and perceived interactivity on the online shopping websites [J]. **Management Review**, 2017, **29**(8): 91-100. (in Chinese)

[22] 程　勖，李文辉，刘裕斌. 基于模拟谐振子算法的服务调度技术 [J]. 大连海事大学学报，2013, **39**(2): 78-81.
CHENG Xu, LI Wenhui, LIU Yubin. Service scheduling technique based on simulated harmonic oscillator algorithm [J]. **Journal of Dalian Maritime University**, 2013, **39**(2): 78-81. (in Chinese)

[23] 程　勖，李文辉，张明会. 基于路径优化的服务调度方法 [J]. 北京工业大学学报，2015, **4**(10): 1537-1542.
CHENG Xu, LI Wenhui, ZHANG Minghui. Service scheduling method research based on route optimization [J]. **Journal of Beijing University of Technology**, 2015, **4**(10): 1537-1542. (in Chinese)

Optimized weighted KNN algorithm based on M-distance algorithm idea

CHENG Xu, GAO Yongzheng, GUO Fang*

(School of Management, Dalian Polytechnic University, Dalian 116032, China)

Abstract: To select features quickly and classify the data accurately, the idea of M-distance algorithm is used to cluster the data set, and the sample data is preprocessed. The weighted *K*-nearest neighbor algorithm is designed to reduce the sample spacing and construct the sample classification model. The method of simulating harmonic vibration is used to traverse the sample data, solve the optimal weighted eigenvector and realize the sample classification. The experimental results show that the designed algorithm is correct and the classification model is reasonable. In the sample data features, the top 10 sample features that consumers are most concerned about are separated, which are in line with consumers′ behavior choice, indicating that the algorithm design has a certain practicality.

Key words: sample nearest neighbor; weighted eigenvector; harmonic oscillator; sample classification