

文章编号: 2096-1618(2020)05-0537-05

# 基于奇异值分解的医疗数据信息提取及分类方法

王震, 张海清, 彭莉, 汪杰, 游凤, 李代伟, 唐聃

(成都信息工程大学软件工程学院, 四川 成都 610225)

**摘要:**当医疗数据存在缺失和冗余信息的情况下如何提高预测准确率一直是一个极具挑战的问题。为解决这一挑战,大多数预测模型要么直接删除缺失和冗余的实例,要么使用均值或其他方式对缺失数据进行填补。基于加权 KNN 算法(weighted k-nearest neighbor, WKNN),提出一种改进的医疗数据分类方法,该方法首先利用 KNNI(k-nearest neighbor imputation, KNNI)对包含缺失数据的数据集进行预填补,然后采用奇异值分解(singular value decomposition, 简称 SVD)对填补后完整的数据进行有效信息提取,最后使用修订权重的 WKNN 算法进行分类预测。实验表明,在对数据进行填补和信息提取后,显著提高了分类准确率。在 5 个医疗数据集上,相较于传统的 KNN 算法分类准确率提升 10% 左右。在 8 个医疗数据集上均使用随机森林算法、朴素贝叶斯算法和支持向量机算法进行实验对比,算法分类准确率均取得较好效果。

**关键词:**医疗数据集;缺失值填补;奇异值分解; $k$ 最近邻算法

**中图分类号:**TP312

**文献标志码:**A

**doi:**10.16836/j.cnki.jcui.2020.05.010

## 0 引言

在经济发展和社会进步逐渐加速的今天,人们的健康意识日趋提高。近年来,恶性肿瘤、心脏病、脑血管疾病和呼吸系统疾病等慢性病,已经成为中国城镇居民首要死因之一,其死亡率位于世界较高水平<sup>[1]</sup>,因此对医疗数据的研究非常有意义。应用大数据技术和智能化技术,结合患者生活环境和临床数据,可以辅助医生实现精准的疾病分类和诊断,制定具有个性化的疾病预防和诊疗方案<sup>[2]</sup>,所以有必要提高医疗数据分类的准确率。

研究发现无论是从医院收集的数据,还是从 UCI 上下载的医疗数据均存在数据缺失的情况。若直接使用这些存在缺失的数据集进行建模,会造成如下影响:系统可能会丢失大量的有用信息、系统中表现出的关系可能更加显著、系统中蕴含的确定性成分更难把握并会使挖掘过程陷入混乱,最后导致不可靠的输出<sup>[3]</sup>。因此,需要首先对获取的数据集进行缺失值预处理操作。

现在较为流行的机器学习分类算法有很多,如  $k$ 最近邻( $k$ -nearest neighbor, KNN)<sup>[4]</sup>、人工神经网络(artificial neural network, ANN)<sup>[5]</sup>、朴素贝叶斯<sup>[6]</sup>、支持向量机(support vector machines, SVM)<sup>[7]</sup>、KD-tree<sup>[8]</sup>和随机森林算法(random forest, RF)<sup>[9]</sup>等。其中 KNN 算法是一种最基本的基于实例的学习方法,因其思想

简单,准确率较高,被广泛应用于机器学习与数据挖掘<sup>[4]</sup>。该方法的思路是:在特征空间中,如果一个样本附近的  $k$  个最近(即特征空间中最邻近)样本的大多数属于某一个类别,则该样本也属于这个类别。在医疗数据集中,由于检测项较多,有的条件属性可能对决策属性的影响较小,或者并无直接联系,导致寻找最近邻样本时产生不准确的结果。如果直接使用 KNN 算法分类,可能导致分类准确率较低,可以考虑提取医疗数据集中的主要信息,然后再使用 KNN 算法进行分类,这样可以有效地增加分类精度。在研究过程中还发现,传统的 KNN 算法在进行最近邻计算时,只是单纯地计算被测样本周围每一类决策属性的个数,而忽略了权重的影响,致使分类准确率偏低。

针对以上问题,KNN 对医疗数据集分类上做出改进,首先对存在缺失的数据集进行填补,之后使用奇异值分解对数据集进行特征提取,最后使用 WKNN 算法进行分类,最终在所使用的数据集上取得了一定的成果。

## 1 相关理论知识

### 1.1 缺失值填补

在医疗数据中,缺失值的产生是很常见的,造成这种情况的原因有很多,如病人为了保护自己的隐私、护士的操作失误和病人没有体检完全等。直接使用缺失数据进行分类时,会对准确率造成一定的影响,尤其是

在医疗方面更要小心谨慎,所以在模型训练之前需要对数据进行预处理操作。对缺失数据的预处理操作包括两种:最简单的方法是将存在缺失值的元组进行删除,但是这样可能会造成数据的原始信息丢失<sup>[10]</sup>,导致不可靠的输出。另一种方法是使用有效的科学方法对缺失值进行填补。目前针对缺失值填补的研究主要在两方面,一方面是基于统计学的方法,如众数填补、均值填补和中位数填补等<sup>[11]</sup>;另一方面是基于机器学习的方法,如决策树归纳填补法、人工神经网络填补法、最邻近填补法、SVM 填补法、粗糙集理论填补法<sup>[12]</sup>和随机森林填补法等。

## 1.2 奇异值分解

奇异值分解其实是对矩阵进行分解,与特征值分解不同的是,奇异值分解不要求被分解的矩阵为方阵,可以对任意矩阵进行分解<sup>[13]</sup>。

假设  $M$  是一个  $m \times n$  的矩阵,则使用 SVD 分解可得

$$M_{m \times n} = U_{m \times m} \Sigma_{m \times n} V_{n \times n}^T$$

其中矩阵  $U_{m \times m}$  是矩阵  $M_{m \times n} M_{m \times n}^T$  的所有特征向量  $u_i (i = 1, 2, \dots, m)$  组合而成的;矩阵  $V_{n \times n}$  是矩阵  $M_{n \times m}^T M_{m \times n}$  的所有特征向量  $v_i (i = 1, 2, \dots, n)$  组合而成的。

矩阵  $\Sigma_{m \times n}$  除了对角线上的元素为矩阵  $M$  的奇异值以外,其余元素都为0,并且对角线上的奇异值是从大到小排列的,即  $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_t (t \leq \min(m, n))$ <sup>[14]</sup>,奇异值即为权重,权重越大代表特征信息越大<sup>[15]</sup>。

## 1.3 带权重的最近邻算法(WKNN)

$k$  最近邻算法是采用度量不同样本之间的距离进行分类<sup>[16]</sup>。该算法较为简单实用,但是当样本不平衡时,如果一个类的样本容量很大,而其他类样本容量很小时,有可能导致当输入一个新样本时,该样本的  $k$  个邻居中大容量类的样本占多数,因此可以考虑采取权重的方法来改进,即离被测样本距离越近的样本被赋予的权重越大。

原本的 KNN 算法只是计算  $k$  个最近样本的大多数属于某一个类别,则将该类别赋予未知样本,相当于每一个最近样本的权重都为1。根据上述描述,赋予权重时可以乘以一个不同大小的因子,假设计算前  $k$  个最邻近样本中第  $N$  近样本的权重,每一个因子都使用  $K$  当分母,这样第  $N$  近的样本,只需分母也是第  $N$  大的数,采用  $(K-N+1)$  即可实现。但是实验发现当  $K$  较大时分类效果较差。通过分析发现,当  $K$  较大时,距离未知样本较远位置的权重几乎为0,所以实验采取赋予权重的方法为  $1+(K-N+1)/K$ 。

## 2 基于SVD的信息特征提取及WKNN分类

算法首先对数据集进行预处理操作,将形成的完整矩阵进行奇异值分解,根据经验对矩阵进行特征提取,最后使用带权重的  $k$  最近邻算法进行分类。文中方法对医疗数据集进行分类的具体流程如下:

(1)对数据集进行预处理操作(如果数据集存在缺失值,采用 KNNI 进行填补);

(2)将预处理后的数据的条件属性形成的  $M_{m \times n}$  矩阵进行奇异值分解;

(3)根据经验取前  $r$  个较大的奇异值进行保留;

(4)使用信息提取之后的特征向量进行 WKNN 分类。

①将新来的样本  $B$  使用公式:  $\text{site} = B_{1 \times m}^T \times U_{m \times r} \times \Sigma_r^{-1}$  进行坐标化。

②计算待测样本与每一个样本的欧拉距离:

$$\text{dist} = \sqrt{(x_1 - z_1)^2 + \dots + (x_r - z_r)^2}$$

③计算前  $k$  个距离待测样本最近的样本中,第  $N$  近样本的权重:

$$W_N = 1 + (K - N + 1) / K$$

④计算前  $K$  个最小距离的样本中每个类标号的权重总和。

⑤权重总和最大的类标号作为新样本  $B$  类别的预测值。

## 3 实验

使用  $k$  最近邻算法、随机森林算法、支持向量机算法和朴素贝叶斯算法与 SVD-WKNN 算法进行分类准确率对比。以下所有试验均使用 Windows10 64 位操作系统,Intel i5-5200U CPU @ 2.2GHz,8 GB RAM,Python 3.6 进行编程。

实验所有的分类准确率都是使用十折交叉验证方法求得。

### 3.1 数据集介绍

为验证提出算法的有效性,使用 UCI 的 9 个标准数据集以及 kaggle 的 PimaIndiandibabetes 数据集。医疗数据集描述如表 1 所示,其中标为“% MV”的列表示数据集中所有缺失值的百分比,标为“Ins. with MV”的列表示数据集中至少具有一个缺失值的实例的百分比。

表1 医疗数据集

Data set	样本个数	属性个数	是否缺失	% MV	% Ins. with MV
Breast	699	10	是	0.31	3.15
Statlog( Heart )	270	14	否	0	0
Bupa	345	7	否	0	0
Hepatitis	155	20	是	5.39	48.39
Primary-tumor	339	18	否	0	0
New-thyroid	215	6	否	0	0
Dermatology	366	35	否	0	0
CTG	2126	21	否	0	0
Pima	768	9	是	11.04	56.25
Mammographic	961	6	是	2.81	13.63

3.2 实验参数设置

以上算法所使用的参数设置如表2所示。

表2 实验算法参数设置

算法	参数设置	参数描述
KNN	K 取 1 ~ 40	最近邻个数取从 1 到 40 个
RF	n_estimators = 50 ~400, 步长为 50	随机生成树的个数为: 50, 100, 150, 250, 300, 350 和 400
SVM	Kernel = " linear "	支持向量机的核函数采用线性核
SVD-WKNN	R 取 1 ~ 条件属性 个数, K 取 1 ~ 40	奇异值保留个数从 1 ~ 属性个数, 最近邻个数取 1 ~ 40
KNNI	K = 10	最近邻个数取 10

3.3 实验结果及分析

首先分别使用均值填补方法和 KNNI 方法对缺失值进行填补,之后进行实验,对比实验效果,如图1所示。从图1发现使用 KNNI 填补之后,分类准确率更高,所以本文采用 KNNI 方法进行填补。

同时,结合表1分析可得,当缺失比例越大时,使用 KNNI 算法填补后,分类准确率效果提升越高。

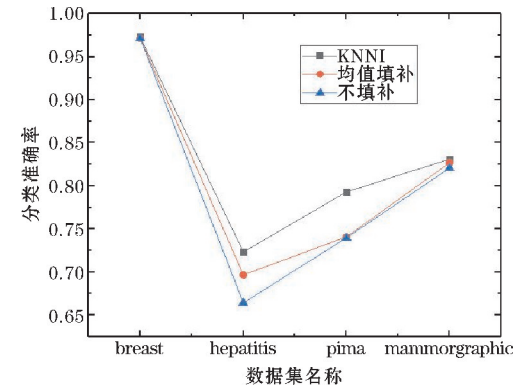


图1 填补算法分类准确率比较

首先对存在缺失数据集使用 KNNI 方法进行缺失值填补处理,然后使用奇异值分解算法对数据集进行特征提取,最后使用带权重的最邻近算法进行分类。

图2 ~ 5 分别是使用 SVD-WKNN 分类算法在 breast、New-thyroid、Dermatology 和 CTG 医疗数据集上

进行的实验结果。图6是使用 KNN 分类算法在所有医疗数据集上进行的实验,其中最近邻个数取 1 ~ 40 个,横轴为最近邻个数,纵轴为分类准确率。图7是使用 RF 分类算法在所有医疗数据集上进行的实验,横轴为随机生成树的个数,纵轴为分类准确率。

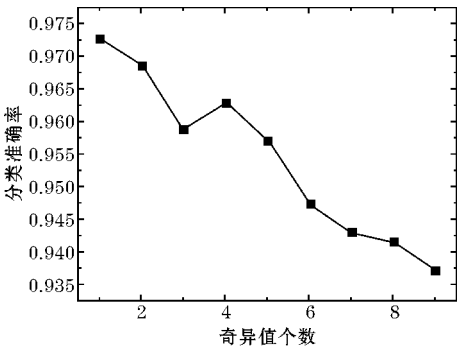


图2 breast数据集实验结果

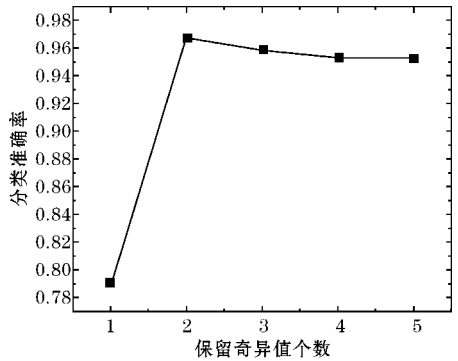


图3 New-thyroid数据集实验结果

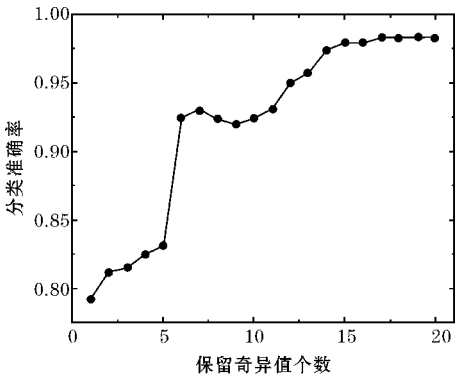


图4 dermatology数据集实验结果

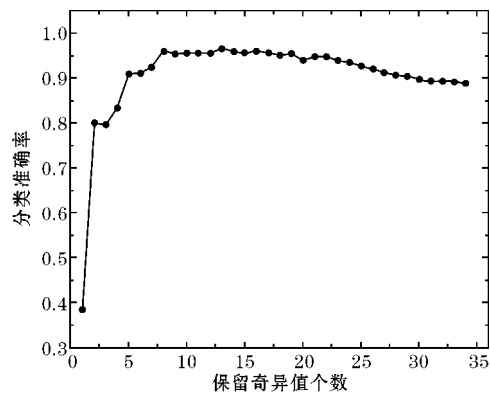


图5 CTG 数据集实验结果

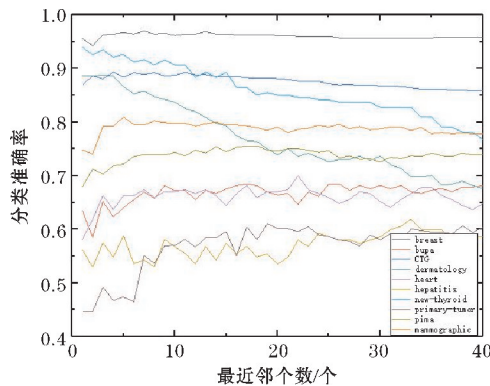


图6 KNN 在所有数据集上的分类准确率

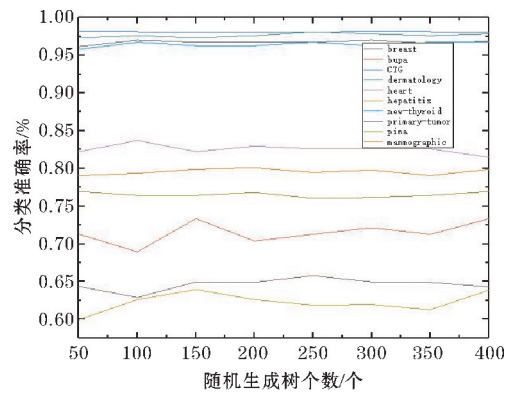


图7 RF 在所有数据集上的分类准确率

根据实验,在所有医疗数据集上算法分类准确率对比结果见表 3。

实验结果表明,在医疗数据集中,使用奇异值分解算法对数据进行特征提取后,再使用最近邻算法进行分类效果优于传统的随机森林算法、支持向量机算法和朴素贝叶斯算法。只有在两个数据集 Bupa 和 Dermatology 中,随机森林算法分类准确率最高,在 CTG 数据集中,SVM 和 NB 算法分类准确率最高,在其他几个数据集中,都是 SVD-WKNN 分类准确率最高。

表 3 医疗数据集算法预测准确率对比

	KNN	RF	SVM	NB	SVD-WKNN
Breast	0.9701	0.9701	0.9672	0.6552	0.9729
Statlog( Heart )	0.70	0.8407	0.8296	0.7852	0.8444
Bupa	0.6845	0.7187	0.6812	0.5684	0.7134
Hepatitis	0.6188	0.6525	0.6192	0.5479	0.7233
Primary-tumor	0.6104	0.6521	0.6994	0.658	0.7261
New-thyroid	0.9398	0.9673	0.9675	0.7349	0.9678
Dermatology	0.8879	0.9785	0.9726	0.9779	0.9666
CTG	0.8928	0.9817	0.984	0.984	0.9834
Pima	0.7553	0.7696	0.7709	0.6445	0.793
Mammographic	0.8094	0.801	0.8166	0.5422	0.8311

注:其中 KNN 算法、RF 算法和 SVD-WKNN 算法在参数范围内均采取分类准确率最高的进行保留

由图 2 ~ 5 和表 3 可以发现,在 Breast 数据集中,随着奇异值保留个数的增加,分类准确率呈下降趋势,当保留奇异值个数为 1 时分类准确率最高为 0.9729,当增加保留奇异值个数时只能增加不相关信息,会对分类效果产生影响;在 Thyroid 数据集中,当保留奇异值个数为 1 时,准确率较低,说明对有效信息的提取不够,当保留奇异值个数为 2 时,分类准确率达到最高为 0.9678,之后再增加奇异值个数,分类准确率会越来越低,说明增加了不相关信息;在 Dermatology 数据集中,奇异值保留个数在 8 之前分类准确率都在大幅提升,当奇异值保留个数为 13 时,分类准确率最高为

0.9666;在 CTG 数据集中,随着奇异值保留个数的增加,分类准确率呈上升趋势,个数为 17 时达到最高为 0.9836,说明想要提高分类准确率就要提取到足够的有效信息才可以。

通过结果也发现,使用奇异值分解算法对数据集进行特征提取之后,再使用带权重的最近邻算法进行分类,比直接使用最近邻算法进行分类,效果要好得多。在 Statlog ( Heart ) 数据集中分类准确率提升了 14.4%,在 Hepatitis 数据集中提高了 11%,在 Dermatology 数据集中提高了 8%,在 CTG 数据集中也提升了将近 10%。



## 4 结束语

通过 SVD-WKNN 算法在医疗数据集上进行分类预测,准确率相较于直接使用 KNN 算法进行分类提升较大,在 5 个医疗数据集中准确率都提升近 10%,相较于 RF、SVM 和 NB 算法相比也具有较高的分类准确率。而且在 4 个医疗数据集中,分类准确率均达到了 96% 以上。可见,SVD-WKNN 算法应用于医疗数据集分类上是可行的。

## 参考文献:

- [1] 周梦丽. 基于广义线性模型的中国主要疾病死亡率统计分析[D]. 成都:西南财经大学,2014.
- [2] 高福,杨宏钧. 推动精准医疗和伴随诊断产业创新发展[J]. 生物产业技术,2018(2).
- [3] 刘星毅,农国才. 几种不同缺失值填充方法的比较[J]. 南宁师范高等专科学校学报,2007,24(3):148-150.
- [4] Guo G, Wang H, Bell D A, et al. An kNN Model-based Approach and Its Application in Text Categorization[C]. Computational Linguistics and Intelligent Text Processing, 5th International Conference, CICLing 2004, Seoul, Korea, 2004:15-21.
- [5] Zhang Zhihua. Multivariate Time Series Analysis in Climate and Environmental Research[M]. Springer 2018.
- [6] 陈曦,张坤. 一种基于树增强朴素贝叶斯的分类器学习方法[J]. 电子与信息学报,2019,41(8).
- [7] Hearst M A, Dumais S T, Osman E, et al. Support vector machines[J]. IEEE Intelligent Systems, 1998,13(4):18-28.
- [8] Wenfeng Hou, Daiwei Li, Haiqing Zhang, et al. An Advanced k Nearest Neighbor Classification Algorithm Based on KD-tree[C]. 2018 IEEE International Conference of Safety Produce Informatization (IICSPI). 2019:902-905.
- [9] Ayyadevara, V Kishore. Pro Machine Learning Algorithms Random Forest[M]. 2018:105-116.
- [10] Wang Fengmei, Hu Lixiao. A missing data filling method based on nearest neighbor rule[J]. Computer engineering, 2012, 38(21):53-55.
- [11] 李璐. 基于 R 语言的缺失值填补方法[J]. 统计与决策, 2012, (17):72-74.
- [12] Zhang Haiqing, Li Daiwei, Wang Tao, 等. Hesitant extension of fuzzy-rough set to address uncertainty in classification[J]. Journal of Intelligent & Fuzzy Systems, 2018, 34(4):2535-2550.
- [13] Golub G H. Singular value decomposition and least squares solutions[J]. Numerische Mathematik, 1970, 14(5):403-420.
- [14] Epps B P, Krivitzky E M. Singular value decomposition of noisy data: noise filtering[J]. Experiments in Fluids, 2019, 60(8).
- [15] 徐锋, 刘云飞. 基于 EMD-SVD 的声发射信号特征提取及分类方法[J]. 应用基础与工程科学学报, 2014(6):1238-1247.
- [16] Leif E. Peterson. K-nearest neighbor[J]. scholarpedia, 2009, 4(2):1883.

## Information Extraction and Classification Method of Medical Data based on Singular Value Decomposition

WANG Zhen, ZHANG Haiqing, PENG Li, WANG Jie, YOU Feng, LI Daiwei, TANG Dan  
(College of Software Engineering, Chengdu University of Information Technology, Chengdu 610225, China)

**Abstract:** Missing values imputation and redundant information reduction have been proved to be significant challenge of improving the prediction accuracy in medical data sets. The traditional prediction models tend to delete the missing instances directly from the data sets or use mean values to fill the missing values, which cannot deeply analyze the internal complex relationships among objects. In order to solve these problems, in this paper, we proposed an improved medical data classification method based on the Weighted k-Nearest Neighbor (WKNN) algorithm. The proposed method firstly pre-filling the incomplete dataset with k-Nearest Neighbor Imputation (KNNI), and then extracting the effective information of the complete data set with Singular Value Decomposition (SVD), finally the revised weighted WKNN algorithm is used to conduct classification prediction. The classification accuracy of 5 medical datasets by this method is higher than that of the traditional KNN algorithm by approximately 10%. The classification accuracy based on experiment performance is higher than the benchmark methods of Random Forest algorithm, Naïve Bayesian algorithm, and Support Vector Machine algorithm (on 8 medical datasets).

**Keywords:** medical data set; missing value imputation; singular value decomposition;  $k$  nearest neighbor algorithm