文章编号：2095-2163(2019)02-0096-04 中图分类号：TP183 文献标志码：A

数据挖掘在乳腺癌复发预测中的应用研究

程国建，张 晗，魏珺洁
（西安石油大学 计算机学院，西安 710065）

摘 要： 乳腺癌是发生在人体乳腺上的恶性肿瘤，受某些因素的影响，乳腺癌术后会有复发的可能。乳腺癌术后复发不仅会加大乳腺癌的治疗难度，还会对患者的身心健康造成伤害。数据挖掘是知识发现的一个特定步骤，能够利用专门的算法从海量数据中抽取有用的知识。数据挖掘可以完成分类、聚类、预测、关联分析等任务，使用数据挖掘算法预测乳腺癌是否有复发的可能，将会对乳腺癌的治疗提供帮助。文章使用来自南斯拉夫卢布尔雅那大学医疗中心乳腺癌肿瘤研究所、由 Zwitter 和 Soklic 提供的乳腺癌数据，实验利用 C4.5 算法、朴素贝叶斯算法和 SVM 算法并使用十折交叉验证方法对该数据进行分类，进而预测乳腺癌是否有复发的可能。最后，文章对 3 种算法的预测结果进行综合分析，得到各个算法在乳腺癌复发预测中的优势和劣势。

关键词： 数据挖掘；乳腺癌；C4.5 算法；朴素贝叶斯；SVM；十折交叉验证；复发预测

Application of data mining in breast cancer recurrence prediction

CHENG Guojian, ZHANG Han, WEI Junjie
(School of Computer Science, Xi´an Shiyou University, Xi´an 710065, China)

【Abstract】 Data mining is a specific step in knowledge discovery. It can use specialized algorithms to extract useful knowledge from massive data. Breast cancer is a malignant tumor that occurs in the breast. Due to certain factors, breast cancer may have a recurrence after surgery. Postoperative recurrence of breast cancer will not only increase the difficulty of treatment of breast cancer, but also cause damage to the physical and mental health of patients. Data mining can complete tasks such as classification, clustering, prediction, and association analysis. Using data mining algorithms to predict whether breast cancer has recurrence may help breast cancer treatment. The breast cancer data of this article is acquired from the Breast Cancer Research Institute at the University of Ljubljana Medical Center in Yugoslavia, provided by Zwitter and Soklic. The article uses C4.5 algorithm, naive Bayesian and SVM with a ten-fold cross-validation method algorithm to classify the data and predict whether breast cancer has recurrence. Finally, the article comprehensively analyzes the prediction results of the three algorithms, and obtains the advantages and disadvantages of each algorithm in breast cancer recurrence prediction.

【Key words】 data mining; breast cancer; C4.5 algorithm; Naive Bayes; SVM; ten-fold cross-validation; recurrence prediction

0 引 言

数据挖掘的概念是在 1995 年加拿大召开的第一届知识发现和数据挖掘会议中提出的，早期主要研究从数据库中发现知识[1]。数据挖掘通常是指从大量的数据中寻找隐藏的有用信息的过程，主要任务有分类、聚类、关联分析、时序模式、偏差检测和预测。自从数据挖掘被提出以来，就引起了许多专家学者的广泛关注。近年来，随着大数据的兴起，数据挖掘逐渐被应用到各行各业中，例如医疗领域[2]、金融业[3]、电力行业[4]等领域。

乳腺癌是乳腺上皮细胞增生癌变后，形成的一个凹陷肿块。乳腺癌是一种常见的恶性肿瘤，不仅危及女性的生命，也严重影响了患者的身心健康。自从 20 世纪 70 年代末开始，乳腺癌的发病数一直位居女性肿瘤首位，并且每年都有递增的趋势[5]。随着医疗技术的发展和人们对乳腺癌研究的不断深入，现在乳腺癌已经有手术治疗、放射治疗、化学药物治疗、免疫治疗等多种治疗方法。然而，受乳腺癌原发肿块的大小、位置、患者年龄、受侵淋巴数等因素的影响，乳腺癌在手术后两年内有可能发生复发或者转移[6]。因此，利用数据挖掘算法对乳腺癌复发的影响因素进行分析，进而实现对乳腺癌术后是否会复发的预测，可以有效地帮助患者尽早采取措施、积极治疗。

1 数据挖掘

数据挖掘是数据库知识发现的一个重要步骤，

作者简介：程国建(1964-)，男，博士，教授，主要研究方向：机器学习、模式识别、图像处理等；张 晗(1994-)，男，硕士研究生，主要研究方向：数据挖掘、机器学习；魏珺洁(1994-)，女，硕士研究生，主要研究方向：智能计算、可视化技术。

收稿日期：2018-12-12

已有许多经典的算法，例如，常用于分类的决策树算法 C4.5、能够根据属性进行聚类的 K-Means 算法、可以挖掘数据集中的关联规则的 Apriori 算法等。本文主要利用 C4.5 算法、朴素贝叶斯算法和 SVM 算法对乳腺癌复发情况进行分类与预测。对此可探讨分述如下。

1.1　C4.5

C4.5 算法是由澳大利亚悉尼大学 Ross Quinlan 教授提出的，是对 ID3 算法改进后得到的一种决策树分类算法。相比于 ID3 算法，C4.5 算法引入了信息增益率来选择属性，可以对连续属性进行离散化处理[7]。其次，C4.5 还在构造树的过程加入了剪枝，剪枝可以减少模型的复杂度，从而避免过拟合现象[8]，而且 C4.5 算法还能够对不完整数据进行处理。

如图 1 所示，C4.5 算法可以根据数据样本的特征属性构造一棵决策树，树的叶子节点代表经过分类得到的具体类别，树的非叶子节点代表数据的属性，从根节点到叶子节点形成的一条路径就是一条分类规则。该算法的本质就是从训练数据中归纳出一组分类规则，并且这些分类规则是互斥且完备的。C4.5 算法适合在小规模数据集和多属性数据集上使用，并且得到的分类准确率较高。只是在构造决策树的过程中，需要多次扫描和排序，因此该算法的效率较低。

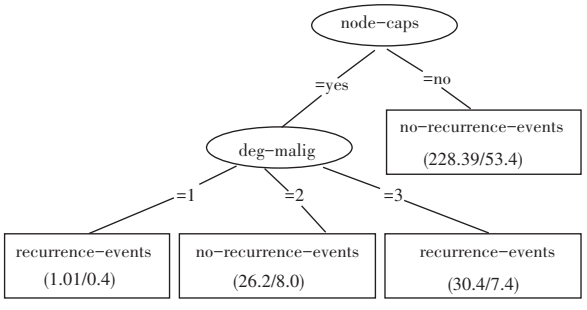

图 1　实现乳腺癌复发预测的 C4.5 决策树

Fig. 1　C4.5 decision tree for breast cancer recurrence prediction

1.2　朴素贝叶斯

朴素贝叶斯算法是 Duda 和 Hart 于 1973 年提出的，是以贝叶斯定理为基础的一种分类方法。之所以称为朴素贝叶斯，是因为该分类器假设数据的每一个属性之间是相互独立的，这些属性都是直接与类属性相关联。朴素贝叶斯算法可以根据数据的一些特征属性，计算各个类别的概率，最终概率最大的类别即为该数据的类。其方法可概述如下：

(1)设 $x=\{a_1,a_2,\cdots,a_m\}$ 为一个待分类项，而每个 a 为 x 的一个特征属性。

(2)有类别集合 $C=\{y_1,y_2,\cdots,y_n\}$。

(3)计算在 X 个属性条件下，所有类别的概率 $P(y_1\mid x),P(y_2\mid x),\cdots,p(y_n\mid x)$。

(4)在所有概率中，选取最大的概率 $P(y_k\mid x)=\max\{P(y_1\mid x),P(y_2\mid x),\cdots,P(y_n\mid x)\}$，则 X 属于概率最大的类别 $x\in y_k$。

朴素贝叶斯算法分类效率稳定、算法简单，尤其对小规模数据的分类效果好，对缺失数据不敏感。由于朴素贝叶斯模型假定属性之间是相互独立的，因此与其它分类方法相比，该算法的误差率可能较低。然而实际上，各个属性之间往往具有一定的相关性。因此当数据集的各个属性实际相关性较小时，朴素贝叶斯分类器分类效果良好，否则，分类效果不好。

1.3　SVM

支持向量机(Support Vector Machine，SVM)是一种监督式学习的方法。支持向量机首先由 Vapnik 和 Corinna Cortes 在 1995 年提出的，通常被广泛地应用在统计学、模式分类和回归分析等方面。SVM 可以在最小化经验误差的同时，最大化几何边缘。因此，SVM 也被称为最大化边缘区分类器。如图 2 所示，限制边缘宽度的向量(点)是支持向量(SV)，2 个异类支持向量到超平面的距离之和称为间隔。支持向量机的基本思想就是将输入数据视为 n 维空间中的 2 组向量。通过在该空间中构建一个分离超平面来对输入数据进行分类，这个超平面使 2 个数据集之间的边界最大化。支持向量机具有以下优点：

(1)通用性：可以在多种函数集中构造函数。

(2)鲁棒性：不需要微调。

(3)有效性：在解决实际问题时是最好的方法之一。

(4)计算简单：方法的实现只需要利用简单的优化技术。

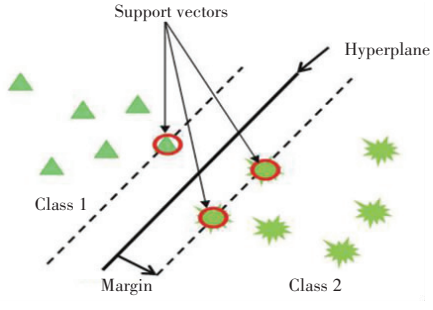

图 2　SVM 原理图

Fig. 2　SVM schematic

2 乳腺癌复发预测

2.1 实验过程

本文使用的是来自南斯拉夫卢布尔雅那大学医疗中心乳腺癌肿瘤研究所,由 Zwitter 和 Soklic 提供的乳腺癌数据[9-12],表 1 展示了该数据集的一部分。该数据集包含 286 个实例和 10 个属性,类属性代表是否会复发,其它 9 个属性分别为 Age(年龄)、Menopause(更年期)、Tumor-size(肿瘤大小)、Inv-nodes(受侵淋巴结数)、Node-caps(有无结节冒)、Deg-malig(恶性肿瘤程度)、Breast(肿块位置)、Breast-quad(肿块所在象限)、Irradiat(是否放疗)[13]。

表 1 乳腺癌数据集

Tab. 1 Breast cancer dataset

No.	Age	Menopause	Tumor-size	Inv-nodes	Node-caps	Deg-malig	Breast	Breast-quad	Irradiat	Class
1	40-49	premeno	15-19	0-2	yes	3	right	left_up	no	no-recurrence-events
2	50-59	ge40	15-19	0-2	no	1	right	central	no	no-recurrence-events
3	50-59	ge40	35-39	0-2	no	2	left	left_low	no	no-recurrence-events
4	40-49	premeno	35-39	0-2	yes	3	right	left_low	yes	no-recurrence-events
5	40-49	premeno	30-34	3-5	yes	2	left	right_up	no	no-recurrence-events
…										
215	30-39	premeno	0-4	0-2	no	2	right	central	no	recurrence-events
216	50-59	ge40	40-44	6-8	yes	3	left	left_low	yes	recurrence-events
217	40-49	premeno	30-34	15-17	yes	3	left	left_low	no	recurrence-events
218	40-49	ge40	20-24	0-2	no	2	right	right_up	no	recurrence-events
219	50-59	premeno	50-54	9-11	yes	2	right	right_up	no	recurrence-events
…										

Weka 是一个拥有可视化界面的数据挖掘平台,在这个平台下,可以简单地完成数据挖掘的整个过程[14]。本文的实验环境使用的是 Weka3.9。过程中,各研发步骤可阐述如下。

(1)启动 Weka,打开 Explorer 面板,在 Preprocess 下点击 Open File 导入乳腺癌数据集(breast-cancer)。

(2)在 Classify 下的 Classifier 中,点击 Choose 选择分类器。本文使用的 3 种分类算法对应的分类器分别为:J48(C4.5 算法)、NaiveBayes(朴素贝叶斯算法)和 LibSVM(SVM 算法)。其中,C4.5 分类器的 confidenceFactor 参数设置为 0.25、numFold 参数值设置为 3、seed 参数设置为 1、reduceErrorPruning 参数设置为 False,即使用 C4.5 剪枝。朴素贝叶斯分类器的 useKernelEstimator 参数和 useSupervisedDiscretization 参数均设置为 False。

C4.5 算法使用 *createNode*( ) 函数为决策树创建新节点;使用 *find_best_split*( ) 函数来选择属性;使用 *Classify*( ) 函数确定叶节点的类别标签;使用 *stopping_cond*( ) 来检查是否要终止决策树的增长。

朴素贝叶斯算法在进行分类时,首先要假设数据的特征属性是相互独立的,并且所有的属性变量都直接与类属性相关联,把类属性作为唯一的父节点。根据此次实验的训练数据,构建朴素贝叶斯分类器模型。

(3)在 Test options 中选择 Cross-validation(交叉验证),由于本文使用的是十折交叉验证,因此 Cross-validation Fold 为 10。

本文采用十折交叉验证的验证方法,相比于其它交叉验证方法,该方法在模型选择中更为有效[15]。在十折交叉验证方法中,数据集被分成 10 份,在进行实验时,轮流将其中的 9 份作为训练数据,1 份作为测试数据。每次实验都会得到一个相应的正确率,对 10 次实验得到的正确率求取平均值,并将该值作为算法最终的正确率。

(4)点击 Start 按钮,开始实验。最后得到如图 3~图 5 所示的运行结果。

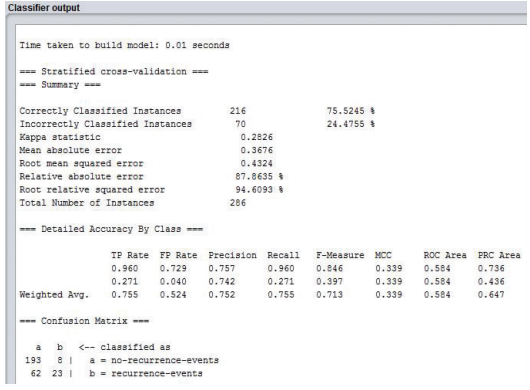

图 3 C4.5 算法预测乳腺癌复发的结果

Fig. 3 C4.5 algorithm predicts breast cancer recurrence

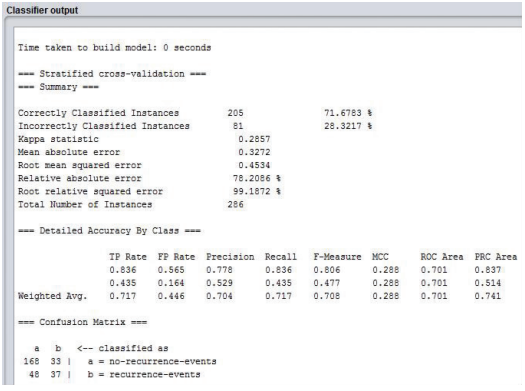

图 4　朴素贝叶斯算法预测乳腺癌复发的结果

Fig. 4　Naive Bayesian algorithm predicts breast cancer recurrence

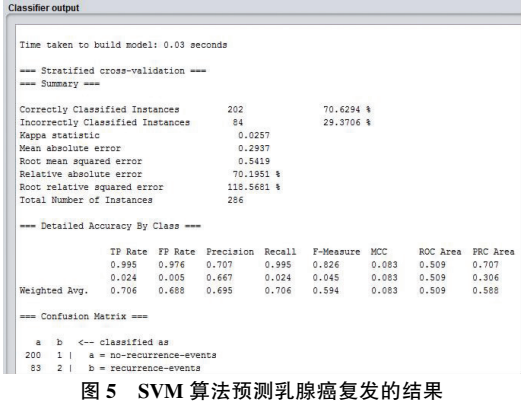

图 5　SVM 算法预测乳腺癌复发的结果

Fig. 5　SVM algorithm predicts breast cancer recurrence

2.2　实验结果分析

基于前文的实验结果，将其归纳整合后详见表 2。本次实验使用的数据集属于小规模、多属性，单从这一点分析，3 种分类算法都易于实现、且性能表现良好。然而从表 2 的数据中可以看出，C4.5 的分类正确率大于朴素贝叶斯分类器和 SVM 分类器的正确率。朴素贝叶斯算法在进行分类时，只考虑了每个属性和类属性之间的关系，而没有考虑到各个属性之间的关系，这就直接影响了算法的分类正确率。而且，根据最后一行可以看到，3 种算法的均方误差也是有差异的，C4.5 算法的均方根误差显然比另外 2 种算法小。综上所述，在乳腺癌数据预测实验中 C4.5 算法效果更好。

表 2　3 种方法实验结果对比

Tab. 2　Comparison of experimental results of three methods

算法	正确率/%	错误率/%	均方根误差
C4.5 算法	75.524 5	24.475 5	0.432 4
朴素贝叶斯	71.678 3	28.321 7	0.453 4
SVM	70.629 4	29.370 6	0.541 9

3　结束语

乳腺癌是一种可能危及女性生命的恶性肿瘤，而且乳腺癌在术后的 2 年内存在复发的风险。文章利用了数据挖掘中的 C4.5 算法、朴素贝叶斯算法和 SVM 算法在 Weka 中对由 Zwitter 和 Soklic 提供的乳腺癌数据进行实验，从而实现对乳腺癌的复发预测。此次实验不仅实现了对乳腺癌的复发预测，还可以对比分析 3 类算法，选出最合适的一种。C4.5 算法的分类正确率较高，且均方根误差明显比另外 2 种算法小，因此，在此次实验中使用 C4.5 算法的效果更佳。

今后主要研究的问题就是在提高 C4.5 算法效率的同时，寻找更优的算法。如今，乳腺癌复发预测是广受关注的一个问题，未来会有越来越多的研究学者提出更好的算法和方案来解决这个问题，并为医学中的乳腺癌治疗提供帮助。

参考文献

[1] 王惠中，彭安群. 数据挖掘研究现状及发展趋势[J]. 工矿自动化，2011，37(2)：29-32.

[2] 苏亚丁. 基于决策树的数据挖掘技术在口腔诊疗中的应用[D]. 石家庄：河北科技大学，2010.

[3] 谢江林，何宜庆，陈涛. 数据挖掘在供应链金融风险控制中的应用[J]. 南昌大学学报(理科版)，2008，32(3)：278-281.

[4] 耿亮，吴燕，孟宪楠. 电力数据挖掘在电网内部及各领域间的应用[J]. 电信科学，2013，29(11)：127-130.

[5] 张忠清，李广灿，叶召. 乳腺癌当前流行趋势分析[J]. 中国肿瘤，2000，9(10)：454-455.

[6] 贾宝洋，李海斌. 乳腺癌复发转移的相关因素分析[J]. 现代预防医学，2009，36(22)：4377-4378.

[7] 高海宾. 基于 Weka 平台的决策树 J48 算法实验研究[J]. 湖南理工学院学报(自然科学版)，2017，30(1)：21-25.

[8] 杨小军，钱鲁锋，别致. 基于 WEKA 平台的决策树算法比较研究[J]. 舰船电子工程，2018，38(10)：34-36，97.

[9] MICHALSKI R S，MOZETIC I，HONG Jiarong，et al. The multi-purpose incremental learning system AQ15 and its testing application to three medical domains[C]//Proceedings of the 5th National Conference on Artificial Intelligence. Philadelphia，PA：AAAI Press，1986：1041-1045.

[10] CLARK P，NIBLETT T. Induction in noisy domains. in progress in machine learning [C]//the Proceedings of the 2nd European Working Session on Learning. Bled，Yugoslavia：Sigma Press，1987：11-30.

[11] TAN M，ESHELMAN L. Using weighted networks to represent classification knowledge in noisy domains[C]//Proceedings of the Fifth International Conference on Machine Learning. Ann Arbor，Michigan：Morgan Kaufmann，1988：121-134.

[12] CESTNIK G，KONONENKO I，BRATKO I. Assistant-86：A knowledge - elicitation tool for sophisticated users [C]// Proceedings of the Second European Working Session on Learning. Bled，Yugoslavia：Sigma Press，1987：31-45.

[13] 周云辉，王娇. 数据挖掘技术在医疗领域中的应用研究[J]. 机械工程与自动化，2013(4)：14-15，18.

[14] 束建华. 基于 WEKA 平台的分类预测模型分析[J]. 蚌埠学院学报，2013，2(2)：26-28.

[15] 范永东. 模型选择中的交叉验证方法综述[D]. 太原：山西大学，2013.