Hans Journal of Data Mining 数据挖掘, 2020, 10(2), 136-142
Published Online April 2020 in Hans. http://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2020.102014

Hans汉斯

Fault Diagnosis Methods Based on Support Vector Machine and Cosine Similarity

Zengsheng Kuang[1], Zhe Zhang[2], Xiuli Shao[1], Bingbing Chang[3]

[1]College of Computer Science, Nankai University, Tianjin
[2]Nankai University Binhai College, Tianjin
[3]Tianjin Feiyu Curtain Wall Decoration Engineering Co., Ltd, Tianjin
Email: kuangzs@mail.nankai.edu.cn

Received: Mar. 9th, 2020; accepted: Apr. 9th, 2020; published: Apr. 16th, 2020

Abstract

Fault diagnosis is a kind of engineering technology widely used in enterprises. Effective fault diagnosis can save a lot of expenses in manpower and material resources for the enterprise. Traditional text fault diagnosis mostly uses the cosine similarity algorithm. When the matching is wrong, the data falls behind, and the amount of data is large, it often fails to meet the real-time needs of customers. Therefore, this paper uses the support vector machine algorithm to coarsely divide the fault description text sentences input by the user to screen out the large categories with similar characteristics. Based on the rough classification results, this paper further uses the cosine similarity algorithm to perform accurate matching, so as to select the cause of the fault with the highest matching similarity and preventive measures to feedback customers. Experimental results show that the fault diagnosis algorithm proposed in this paper can effectively perform fault diagnosis and bring considerable economic benefits to the enterprise.

Keywords

SVM, Cosine Similarity, Fault Diagnosis Methods

基于支持向量机和余弦相似度的故障诊断方法

匡增晟[1]，张　喆[2]，邵秀丽[1]，常兵兵[3]

[1]南开大学计算机学院，天津
[2]南开大学滨海学院，天津
[3]天津飞宇幕墙装饰工程有限公司，天津
Email: kuangzs@mail.nankai.edu.cn

收稿日期：2020年3月9日；录用日期：2020年4月9日；发布日期：2020年4月16日

文章引用: 匡增晟, 张喆, 邵秀丽, 常兵兵. 基于支持向量机和余弦相似度的故障诊断方法[J]. 数据挖掘, 2020, 10(2): 136-142. DOI: 10.12677/hjdm.2020.102014

摘　要

故障诊断是一种广泛应用于企业的工程技术，有效的故障诊断可以为企业节省大量的人力和物力的开销。传统的文本故障诊断大多采用余弦相似度算法，当匹配出错、数据靠后以及数据量较大时，往往无法满足客户的实时需求。因此，本文采用支持向量机算法对用户输入的故障描述文本语句进行粗划分，筛选出具有相似特征的大类。在此基础上，依据粗分类结果，进一步使用余弦相似度算法进行精确匹配，从而选取匹配相似度最高的故障产生原因和防治措施以反馈客户。实验结果表明，本文所提的故障诊断算法可以有效地进行故障诊断，为企业带来可观的经济效益。

关键词

支持向量机，余弦相似度，故障诊断方法

1. 引言

近年来，由于计算机技术的快速发展，故障诊断技术广泛的应用于各大企业，但还未取得显著性的进展。有效的故障诊断技术可以提升企业的竞争力，为企业带来可观的经济效益。因此，研究故障诊断技术是非常有必要的。

现在大多数故障诊断的研究都是基于图像的，基于文本的故障诊断研究还是相对较少，本文主要借鉴的是智能医疗诊断的方法。传统的智能故障诊断方法的思想是以疾病的数值表示与专家的推理相结合，包括贝叶斯公式、模糊数学等方法[1] [2]。接着机器学习的方法被成功地应用到智能医疗诊断领域[3] [4] [5]。2007 年，何凯[6]将支持向量机方法运行在智能医疗诊断系统中的应用与研究中，并取得了良好的效果。然而，当类别很多时，单纯地使用支持向量机算法需要更多的训练样本。2015 年，林予松等[7]基于 VSM 权重改进算法，以及徐奕枫等[8]在 2017 年做了 TF-IDF 权重改进算法研究，实现了智能导医系统。当匹配的数据条目比较靠后，该算法将十分耗时。最近，深度学习在文本领域取得很大的进步[9]。2019 年，陈实[10]将神经网络应用在中医诊断中，很大程度地提升了准确性和降低了时间成本。

在实际场景中，故障诊断通常具有较强的领域相关性，且故障诊断的文本数据收集较为困难，成本较高。本文使用的数据集主要来源于企业网站报修单上的实际用户故障描述这种类型的文本数据，以期在企业的实际应用中减少故障诊断的时间、人工成本以及提升企业竞争力。由于收集的故障数据集相对较小，且故障数据文本一般都是短文本，在 100 字以内。所以本文暂未选用深度学习算法。综合考虑数据集大小以及故障诊断的需求和现状，本文主要以余弦相似度算法为主要框架[7] [8]。考虑到当匹配出错、数据靠后以及数据量较大时，单一使用余弦相似度算法往往无法满足客户的实时需求。本文提出首先采用 SVM 算法对用户输入的故障描述文本语句进行粗划分，筛选出具有相似特征的大类[11] [12] [13]。在此基础上，依据粗分类结果，进一步使用余弦相似度算法进行精确匹配[8] [14]，从而选取出匹配相似度最高的故障产生原因和防治措施以反馈客户，帮助客户自助诊断常见的故障问题[15] [16]。

本文的结构介绍如下，第 2 部分给出了文本数据预处理以及向量化工作。在第 3 部分介绍基于 SVM 的故障粗分类算法的具体实现。在第 4 部分介绍基于余弦相似度的故障精确匹配。在第 5 部分将介绍算

法的对比实验。最后，在第 6 部分总结本文方法所取得的效果以及对未来的展望。

2. 文本数据预处理以及向量化

在进行 SVM 模型处理以及余弦相似度匹配之前，需要对数据集进行预处理工作。数据集由 SVM 模型训练需要的故障文本数据集以及余弦相似度匹配算法所需要的故障通病现象数据集两部分数据组成(数据集将在第 5 部分的故障数据介绍部分进行详细介绍)。

2.1. 数据预处理

在进行去停用词和分词两个部分的数据预处理工作时，本文考虑到一般常用百度停用词列表、哈工大停用词表和四川大学机器智能实验室停用词库。因此，使用整理去重后的三者合集作为本文使用的停用词表去停用词。由于 Jieba 分词准确，速度快。本文使用 Jieba 分词工具对去停用词后的故障样本集以及故障通病现象数据集进行分词处理，分别用列表格式存储。

2.2. 数据向量化

分别对故障样本集以及故障通病现象数据集的列表格式数据进行向量化处理，本文考虑到故障文本数据集以及故障通病现象数据集的每个文本之间相差较大，因此使用 TF-IDF 算法[17] [18] [19]进行文本向量化处理(算法公式如公式(1)。

$$TFIDF_{ij} = \frac{tf_{ij}}{tf_{\bullet j}} \log \frac{df}{df_i},\ i = 1, 2, \cdots, m;\ j = 1, 2, \cdots, n \tag{1}$$

式中 tf_{ij} 是单词 w_i 出现在故障文本 d_j 中的频数，$tf_{\bullet j}$ 是文本 d_j 中出现的所有单词的频数之和，df_i 是含有 w_i 的文本数，df 是故障文本数据集的全文本数。这样每个故障文本 d_j 就可以用向量矩阵的第 j 列向量表示。

通过 Genism 模型中的 Dictionary 统计式中 tf_{ij} 是单词 w_i 出现在故障文本 d_j 中的频数以及 df_i 是含有 w_i 的文本数。然后利用 Genism 中的 TF-IDF 模型根据 Dictionary 统计的 tf_{ij} 和 df_i，给每个词赋予相应的权重。因为每个句子中所包含的词的个数是不一样，所以在这里用到了 Genism 模型中的 Sparse Matrix Similarity 将句子的词向量表示成稀疏矩阵的形式，记作 X(在矩阵中，x_{ij} 表示在故障文本 d_j 中权值，$x_{\bullet j}$ 表示文本 d_j 的向量化形式)。

$$x = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \tag{2}$$

3. 基于 SVM 的故障粗分类

支持向量机(Support Vector Machine, SVM)是 Vapnik 在统计学理论基础上提出的一种机器学习算法，在解决小样本、非线性和高维空间的实际问题中有较大优势。它通过构造最优的超平面，使得各样本到超平面的误差总和最小。对于线性可分的样本，它通过间隔最大化函数确定最优超平面。而对于线性不可分的样本，SVM 使用核函数将低维空间映射到高维，使样本线性可分。最优超平面由少数的支持向量决定，故 SVM 在小样本中也能取得不错的效果。

故障诊断的文本数据集属于线性不可分的问题，本文选择 SVM 算法对故障文本数据集进行分类[20] [21] [22]。

下面给出算法的具体实现。

本文使用 Python 语言 Sklearn 包中的 svm.SVC()函数构建故障文本分类器。由于故障数据集一共有八大类故障，因此将 svm.SVC()函数中的 Decision_Function_Shap 参数设置为“OVR”(一对多分类法，One-Versus-Rest)。

故障分类训练输入：将 2.2 模块得到的故障文本数据集的稀疏矩阵形式，以及用 0，1，2，...，7 分别表示八大类故障类别组成训练集 T(这里只以第 0 类分类为例，其他类分类以此类推，稀疏矩阵中的每一个列向量为 x_i，将第 0 类的 y 值设为 1，其他类设置为−1)。$T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$，$x_i\in\chi=R^n, y_i\in\gamma=\{-1,+1\}$ 选择一个适合的惩罚函数 $C>0$，构造并求解故障分类的最优化问题：

$$\min_{\alpha}\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}a_i a_j y_i y_j\left(x_i\cdot x_j\right)-\sum_{i=1}^{N}a_i \tag{3}$$

使得：

$$\sum_{i=1}^{N}a_i y_i=0, 0\le a_i\le C, i=1,2,\cdots,N \tag{4}$$

对其求解后，我们可以得到最优方案 $a^*=\left(a_1^*,a_2^*,\cdots,a_N^*\right)$。

计算 $w^*=\sum_{i=1}^{N}a_i^* y_i x_i$，选择 a^* 的一个小于 C 的正分量，并根据此计算：

$$b^*=y_j-\sum_{i=1}^{N}y_i a_i^*\left(x_i\cdot x_j\right) \tag{5}$$

之后构建一个超平面 $w^*\cdot x+b^*=0$，来获取决策函数：

$$f(x)=sign\left(w^*\cdot x+b^*\right) \tag{6}$$

最终通过实验发现，当取 Kernel = “Linear”，C = 1.5，Decision_Function_Shape = “OVR”时，故障文本分类效果最好(其中 Linear 表示线性核，OVR 代表分类类别为一个类别与其他类别的划分)。SVM 分类效果，本文在实验部分给出。

4. 基于余弦相似度的故障精确匹配

余弦相似度算法适合于短文本，而不适合长文本[20]。因为故障诊断文本数据是短文本数据，故此本文使用余弦相似度进行细化分，做文本的相似度匹配[21] [22]。余弦相似度用向量空间中两个向量夹角的余弦值作为衡量两个个体间差异的大小。如用户输入故障问题文本数据与故障的通病现象文本数据构成的向量 a，b。当余弦值越接近 1，就表明夹角越接近 0 度，也就是两个向量越相似，即故障文本 a，b 相似，这就叫“余弦相似性”。

在基于余弦相似度的故障相似度匹配算法中，首先利用 SVM 算法八大类故障分类结果，确定是哪一大类的故障通病现象数据作为余弦相似度匹配数据集。然后执行前面提到的数据预处理以及数据向量化操作将确定大类的故障通病数据集转化为稀疏矩阵格式。再将在 SVM 算法中已经向量化过的用户输入的故障描述文本向量 a 与故障通病现象数据集中的稀疏矩阵格式中的每一个文本向量 b 用余弦相似度算法计算其相似度(算法公式如公式 7)。最后，选取相似度最大的故障通病现象文本的产生原因以及防治措施反馈给用户。

$$\cos(\theta)=\frac{a\cdot b}{\|a\|\times\|b\|}=\frac{\sum_{i=1}^{n}(x_i\times y_i)}{\sqrt{\sum_{i=1}^{n}(x_i)^2}\times\sqrt{\sum_{i=1}^{n}(y_i)^2}} \tag{7}$$

5. 实验结果与分析

5.1. 故障诊断数据介绍

本文实验的数据主要来源于某家幕墙公司提供的近三年的1000份用户报修单中的有效数据。报修单上详细记录了用户对故障问题的描述，以及维修人员所提供的报修类型、产生原因和故障防治措施所组成的数据。这些数据分为两个部分：第一个部分是本文分类所需要的粗划分数据，由用户对故障问题的描述评论以及所属故障大类两个字段组成，如表1所示(表1中类别用数字表示)。

Table 1. The first part of the data samples
表 1. 第一部分数据样例

故障描述	所属大类
预埋件，后补埋件与结构接触不紧密	0
横梁的安装不合格，使得横梁不水平，有点翘	1
...	……
胶中硅油渗出，污染板面	7

所有的故障包括：“预埋件故障”，“龙骨故障”，“金属幕墙故障”，“玻璃幕墙故障”，“玻璃幕墙窗故障”，“幕墙窗、门、开启扇故障”，“保温、防火、避雷故障”，“打胶(耐候密封胶)故障”等八大类故障。这里选取了其中一个大类中词频最高的前50词别表示如图1所示。

Figure 1. Word cloud for failure data set
图 1. 故障数据集的词云

第二部分数据是余弦相似度匹配算法所需的细化分数据集。由于每一大类故障里，根据损坏的方式，严重程度以及部位的不同，对大类里的每一种故障情况其防治措施也截然不同。因此，第二部分数据有以下四个字段，分别为：“故障大类类别”，“通常现象”，“原因”，“防治措施”。这一部分数据是为用户对故障问题的描述评论做相似度匹配，数据样例如表2所示。

Table 2. The second part of the data samples
表 2. 第二部分数据样例

所属大类	通病现象	产生原因	防治措施
0	在预埋件中，螺母未拧紧	工人遗漏，检查不到位	交底中要明确规定，安排专人进行检查
...	...	...	...

5.2. 实验结果及分析

实验选择随机森林、朴素贝叶斯、逻辑回归和SVM四种算法进行故障分类，并采用准确率、召回率

和 F1 三个指标作为模型性能的评估指标。四种算法在三个指标上的分类结果如表 3 所示。

Table 3. Classification results of four models on three indicators
表 3. 四种模型在三个指标上的分类结果

模型	准确率(%)	召回率(%)	F1(%)
随机森林	70%	71.%	70.5%
NB	86%	85%	85.5%
逻辑回归	86%	86%	86.0%
SVM	93%	92%	92.5%

从表 3 可以看出，SVM 模型在所有评价指标上均获得了最好的分类效果，逻辑回归和 NB 模型次之，随机森林模型最差。实验证明了 SVM 能够较好地完成基于小数据集线性不可分的多分类任务。因此，本文选择 SVM 算法进行幕墙的故障诊断分类。

最后将传统的只使用相似度匹配故障诊断算法与先用 SVM 进行分类然后进行相似度匹配的故障诊断算法进行时间和准确度上的对比，对比结果如表 4 所示。由表 4 可以看出，使用 SVM 和余弦相似度的组合算法比只使用余弦相似度算法的速度提升了将近两倍，而且准确度也提高了三个百分点。

Table 4. Comparison of the running time and accuracy of the two methods
表 4. 两种方法的运行时间与准确度对比

算法类别	余弦相似度	基于 SVM 的余弦相似度
运行时间	40.89 ms	25.93 ms
准确度	88%	91%

6. 结语

本文针对传统故障诊断方法存在的问题，结合智能医疗诊断方法的基础上，设计实现了基于 SVM 分类和余弦相似度的故障文本相似度匹配的智能故障诊断方式。通过实验证明，该方法取得了良好的效果，并且已经在企业网站上应用，解决了传统的故障诊断方式所存在的问题，为企业节省大量的人力、财力，提升了企业的竞争力。但是，本论文尚有不足之处。以往未有基于文本的故障诊断这方面的研究且故障数据量以及数据种类有限，若以后能有更多的数据集，可以使故障诊断方法更具有说服力。另一方面是由于故障数据量的匮乏，也没有使用深度学习算法进行对比实验。随着以后数据集数量的增加，可以进一步在深度算法上研究。

基金项目

天津市智能制造专项资金项目(201810602，201907206，201907210，20191009)；天津市互联网先进制造专项资金项目 18ZXRHGX00110。

参考文献

[1] 郑鹏, 刘海青, 等. 模糊聚合算子在医疗诊断中的应用[J]. 计算机工程与应用, 2001(24): 170-171.

[2] Ding, W.-P., *et al.* (2006) Application of Fuzzy Logic Reasoning in Intelligent Assistant Diagnosis System of Electronic Patient Record. *Journal of Nantong University* (*Natural Science*), **4**, 77-81.

[3] 胡寿松, 王源. 基于支持向量机的非线性系统故障诊断[J]. 控制与决策, 2001, 16(5): 617-620.

[4] 陈钦界. 基于机器学习的智能医疗诊断辅助方法研究[D]: [硕士学位论文]. 北京: 国防科学技术大学, 2017.

[5] 梁耀波. 智能医疗诊断系统的研究与实现[J]. 北京理工大学, 2016, 32(9): 81-83.

[6] 何凯. 支持向量机方法在智能医疗诊断系统中的应用与研究[D]: [硕士学位论文]. 杭州: 浙江大学, 2007.

[7] 林予松, 梁璐, 崔勇, 等. 基于 VSM 权重改进算法的智能导医系统[J]. 计算机应用与软件, 2015, 32(9): 81-83.

[8] 徐奕枫, 刘利军, 黄青松, 等. 智能导医系统中 TF-IDF 权重改进算法研究[J]. 计算机工程与应用, 2017, 53(4): 238-243.

[9] Li, X.Z., *et al*. (2019) Intelligent Diagnosis with Chinese Electronic Medical Records Based on Convolutional Neural Networks. *BMC Bioinformatics*, **20**, 62. https://doi.org/10.1186/s12859-019-2617-8

[10] 陈实. 神经网络及在中医智能诊断中的应用[J]. 电子技术与软件工程, 2019(20): 155-156.

[11] Mitra, V., Wang, C.-J. and Banerjee, S. (2006) Text Classification: A Least Square Support Vector Machine Approach. *Applied Soft Computing Journal*, **7**, 908-914. https://doi.org/10.1016/j.asoc.2006.04.002

[12] 朱远平, 戴汝为. 基于 svm 决策树的文本分类器[J]. 模式识别与人工智能, 2005, 18(4): 412-416.

[13] 李毅. 基于 SVM 的文本分类应用研究[D]: [硕士学位论文]. 成都: 成都电子科技大学, 2014.

[14] 郭庆琳, 李艳梅, 唐琦. 基于 VSM 的文本相似度计算的研究[J]. 计算机应用与研究, 2008, 25(18): 3256-3258.

[15] 王茜. 基于电子病历的医疗诊断辅助系统设计与实现[D]: [硕士学位论文]. 郑州: 郑州大学, 2016.

[16] 宋安. 基于电子病历的医疗诊断模型的研究与应用[D]: [硕士学位论文]. 杭州: 浙江理工大学, 2017.

[17] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. and Dean, J. (2013) Distributed Representations of Words and Phrases and Their Compositionality. *Advances in Neural Information Processing Systems*, **26**, 3111-3119.

[18] 黄承惠, 印鉴, 侯昉. 一种结合词项语义信息和 TF-IDF 方法的文本相似度量方法[J]. 计算机学报, 2011, 34(5): 856-863.

[19] 付德宇, 代成琴. 一个面向文本分类的中文特征词自动抽取方法[J]. 计算机工程与应用, 2006, 42(15): 165.

[20] 刘江华, 程君实, 等. 支持向量机训练算法综述[J]. 信息与控制, 2002, 31(1): 45-50.

[21] Chang, C.-C. and Lin, C.-J. (2011) LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology*, **2**, 21-27. https://doi.org/10.1145/1961189.1961199

[22] 张高祥. 基于 SVM 的文本信息过滤算法研究[D]: [硕士学位论文]. 长春: 吉林大学, 2016.