Computer Science and Application 计算机科学与应用, 2017, 7(3), 199-205
Published Online March 2017 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2017.73025

Hans汉斯

Study on the Classification Scheme of Education Resources Based on Ontology via GA-KNN and Linear Programming

Guan Wang, Xiaoyan Wang

College of Computer Science, Beijing University of Technology, Beijing
Email: 1614616896@qq.com

Received: Mar. 1st, 2017; accepted: Mar. 20th, 2017; published: Mar. 23rd, 2017

Abstract

With the rapid development of information technology, the Internet has accumulated a large number of outstanding teaching resources, and people begin to use the network learning platform to access these excellent resources in order to achieve the purpose of autonomous learning. In order to be able to quickly access to the required resources, we must classify multifarious resources. The traditional artificial classification is difficult to finish this job. Using text automatic classification technology, we can realize the rapid and effective classification of teaching resources. This article will use the GA-KNN and linear programming to achieve the automatic classification of education resources. From the experimental results, it achieves the expected effect, greatly improving the accuracy of classification.

Keywords

GA-KNN, Education Resources, Aid Decision Making, Linear Programming, Neural Networks

借助GA-KNN及线性规划的基于本体的教学资源分类方案的研究

王　冠，王晓燕

北京工业大学计算机学院，北京
Email: 1614616896@qq.com

收稿日期：2017年3月1日；录用日期：2017年3月20日；发布日期：2017年3月23日

文章引用: 王冠, 王晓燕. 借助 GA-KNN 及线性规划的基于本体的教学资源分类方案的研究[J]. 计算机科学与应用, 2017, 7(3): 199-205. https://doi.org/10.12677/csa.2017.73025

摘　要

随着信息技术的高速发展，Internet上积累了大量的优秀的教学资源，人们开始利用网络这个学习平台来访问这些优秀资源，以达到自主学习的目的。为了能够快速访问到所需资源，必须对繁杂的资源进行分类。传统的人工分类难以完成这一工作，利用文本自动化分类技术，则可以实现对教学资源进行快速而有效的分类。本文将利用GA-KNN及线性规划来实现教育资源的自动化分类，从实验结果来看，它基本达到了预期的效果，大大提高了分类的精度。

关键词

GA-KNN，教学资源，辅助决策，线性规划，神经网络

1. 引言

随着网络和网络教育的发展，Internet 上的教学资源也在迅速增长，如何有效管理这类数据对学生、教育科学工作者和家长获取重要的信息具有至关重要的意义。针对教育领域的多资源、多属性、多要素特征及分类机制依赖的资源库必须满足快速检索及时支持用户使用的要求，确立了教学资源分类规范化这一关键技术问题。

本研究以教育部教育资源建设技术规范为基础来对教学资源中的信息进行向量化处理，通过结合客观的资源分类方法和主观的专家经验，研究基于本体的教学资源分类方案。

2. 基于本体的教学资源分类研究现状

本体是一个源于哲学的概念，原意指关于存在及其本质和规律的学说，后来被计算机科学领域引入，特指对共享概念模型所作的明确化、形式化、规范化说明，它强调领域中的本质概念，也强调这些本质概念之间的关联。某个领域的本体能够将该领域中的各种概念及概念之间的关系显性地、形式化地表达出来，从而将概念中包含的语义表达出来[1]。本文中对教育资源分类遵循的标准是 CELTS-41 (教育资源建设技术规范) [2]。教学资源库建设是教育信息化的基础，面对网上的海量信息，人工选择并加以分类整理的做法不但耗费大量的人力、物力，而且还存在分类结果和实际结果一致性不高的问题。文本自动分类技术为解决这个问题提供了一个新思路。国内外在文本自动分类方面的研究取得了可喜的成果。目前，较为常用的分类算法包括支持向量机、K 近邻法、贝叶斯方法、神经网络法、线性最小二乘法等。VSM 在文本分类中被广泛使用，KNN 就是其中的一种，研究表明，KNN 分类算法是向量模型中最好的分类算法之一[3]。

2005 年，钱晓东、正欧在《基于改进 KNN 的文本分类方法》[3]一文针对 VSM (向量空间模型)中 KNN (K 最近邻算法)在文本处理环境下的不足，根据 SOM (自组织映射神经网络)理论、特征选取和模式聚合理论，提出了一种改进的 KNN 文本分类方法，提出应用 SOM 神经网络进行 VSM 模型各维权重的计算。此后，2012 年杜尔斌、李翔、林祥针对我国教学资源的特点在《改进的 KNN 文本分类算法》[4]一文中针对传统 KNN 算法中对特征项的非监督权重分配的不足之处做了改进，采取 x2 统计量方法和信

息增益这两种监督权重分配方法，有效地利用了训练集标签信息，提高了 KNN 算法的精确度。但这两种方法均没有考虑结合主观经验来进行推理，即没有利用教育专家的先验知识参与对机器学习学到参数进行调整来提高分类的精度。

因此，我们提出了一种借助 GA-KNN 及线性规划的基于本体的教学资源分类方案，这种方案可以利用教育专家的先验知识在数据筛选阶段对机器学习学到参数进行调整，从而为训练神经网络提供与待分类资源相关性较高的资源作为训练集，并最终达到提高整个系统的分类精度的目的。

3. 借助 GA-KNN 及线性规划的基于本体的教学资源分类方案

在使用系统输出决策前，需要进行建模。本方案的建模过程分为两个主要阶段：第一阶段，使用 GA-KNN 算法及线性规划选取出与待分类资源最相似的资源的集合。第二阶段，将这些检索出的资源作为训练集，训练一个以资源为输入并且以分类结果为输出的神经网络。

3.1. 筛选相似资源

主要方法是利用 GA-KNN 算法计算出每个量化的属性的权重，然后借助线性规划方法使用专家对机器学习计算出的结果的评价对已学习到的权重进行调整并达到优化的目的。并利用最终得到的权重执行加权的 KNN 算法，检索出与待分类资源最相似的案例集。

3.1.1. GA-KNN 算法

GA-KNN (GA(Genetic algorithm)算法也就是遗传算法)算法采用遗传算法和交叉验证搜索出特征变量的最优权重，将最优权重加入相似度计算以决定最近邻[5]。

1) 初始化阶段

遗传算法采用二维编码来表征属性的权重，染色体对应加权策略。采用随机数生成算法产生一定数量的随机向量，为优化过程提供初始种群。

2) 加权的 KNN 算法

使用加权的 KNN 邻近算法[6]从资源库中搜索出相似资源，该方法强调的是资源的一对一的属性匹配。首先为资源的每一个属性指定一个权值，检索资源的时候可以根据输入资源中各组成成份的权值求出与资源库中各属性的匹配程度，即相似度。具体做法为，给出资源间距离(即相似度)的定义，根据这个定义，计算出当前资源与资源库中所有资源间的距离，然后从中选出距离最小者。其中数值型属性的相似度计算采用基于欧几里德距离公式演化而来的一种计算确定数值型属性的相似度算法。设使用 1 个属性来描述每个资源，并且每个属性具有一个权重，于是权重的集合是 $W=(w_1,w_2,\cdots,w_l)$ 对于资源库 $P=\{X_1,X_2,\cdots X_n\}$ 中的任意的两个资源，$X_i=(x_{i1},x_{i2},\cdots,x_{il})$ 和 $X_j=(x_{j1},x_{j2},\cdots,x_{jl})$。把两个资源的距离记作 d_{ij}，则我们使用下式来计算这个距离

$$d_{ij}=w_k\sum_{k=1}^{l}\mathrm{d}_k \tag{1}$$

$$d_k=\begin{cases}0, & \text{若 } x_{ik}=x_{jk}\\ 1, & \text{若 } x_{ik}\neq x_{jk}\end{cases} \tag{2}$$

对于资源库中的任意两个个体 x_i 和 x_j，定义

$$S_{ij}=\frac{1}{1+d_{ij}} \tag{3}$$

称 S_{ij} 为第 i 个体与第 j 个体之间的相似度。

3) 遗传算法操作过程

为每个染色体计算适应性函数，并进行遗传算法[7]的进化操作。首先，进入初始资源库的构建阶段，通过反复运行传统 GA 获得多个局部最优解，根据局部最优解间的相似程度，将不相似的局部最优解增加到资源库中，保证资源的多样性，为优化过程提供初始种群。在测试数据上进行加权的 KNN 算法的推理，然后计算适应函数的值，选择适应函数比较大的产生下一代，并进行相应的交叉、变异的操作算法使得自带向着适应度函数最大的方向进化。将适应度函数的输出值与算法在测试集上运行得到的结果的正确率成正比，与该结果的错误率成反比，具体可定义为：

$$f\left(W_{\text{iter}}\right)=\frac{r_{\text{iter}}}{N} \tag{4}$$

其中 r_{iter} 表示将该次遗传算法的迭代获得的权重用于加权的 KNN 算法时，得到的正确分类的资源的总数。N 表示测试集中的资源数量。

采用单点交叉[8]方式，再通过轮盘赌的方式选择出的两组权重的二进制码串上随机选择某位，然后以此点为界将其分为左右部分，根据设定的交叉概率大小决定是否将两组权重的左右部分互相交换，最后生成两组新的权重。轮盘赌的方式[9]具体来讲，就是通过各个个体的选择概率计算其累计概率，第 K 个个体的累计概率为 $P_x\left(a_k\right)$ 间然后产生 0 到 1 之间的随机数 e 与 $P_x\left(a_k\right)$ 进行比较来决定选择的个体，若 $a_{k-1}<e<a_k$，则选择第 k 个个体。为了提高遗传算法的全局搜索能力，还需要执行变异操作。从交叉后的每组权重中，对个体编码串以变异概率 P 随机指定某一位或某几位基因进行变异操作。

这个过程循环地执行，直到在所有个体的适应度中的最大值超过预定义的要求，即大于某个实数值，我们将这个数记作 $f_{\max}$。这个过程如图 1 中所示。

3.1.2. 线性规划问题

由于在本文之前的方案并没有利用专家的先验知识对机器学习学到的模型参数(即筛选相似资源的加权 KNN 算法的权重)进行评估，因此，往往得到主观上难以理解的结果，同时，利用先验知识能够有效提高模型的精度。继而，本文提出的方案允许专家参与对模型参数的评价，并能够利用评估的结论进一步优化权重。

GA-KNN 算法返回的结果是与待分类资源最相似的资源的集合。在专家参与参数调整的过程中，令专家得到系统提供的最相似的 n 个资源，然后根据专家经验对这 n 个资源按照与待分类资源的相似度从大到小进行排序，并再次输入系统的线性规划模块。

$$w_1+w_2+\cdots+w_l=1 \tag{5}$$

另外，设专家排序后的资源的序列为 $c_1,c_2,\cdots,c_n$，其中专家认为这些资源与待分类资源的相似度具

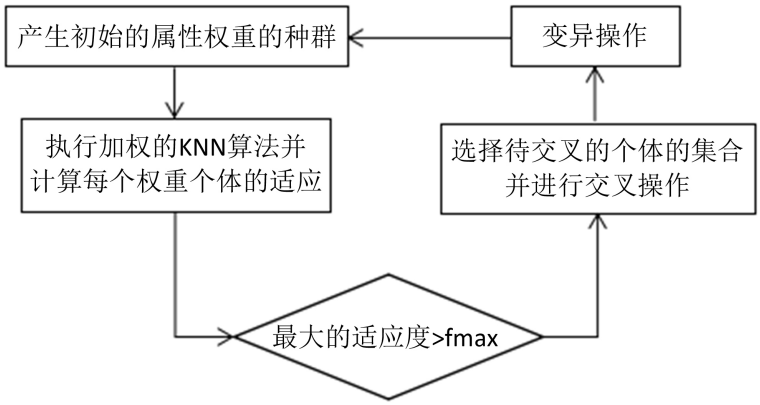

Figure 1. GA-KNN algorithm implementation process
图 1. GA-KNN 算法执行过程

有以下关系：

$$S_{1t} \geq S_{2t} \geq \cdots \geq S_{nt} \tag{6}$$

这些资源到达测试资源 c_t 的距离记作 $d_{1t}, d_{2t}, \cdots, d_{nt}$，其中距离的定义与等式(1)中定义的一致。

由等式(2)和不等式组(5)看出专家给出的这个排序可以看作是按距离的从小到大的排序，因此，我们有关于 $w_1, w_2, \cdots w_l$ 的线性不等式组：

$$\begin{aligned} d_{1t} &\leq d_{2t} \\ d_{2t} &< d_{3t} \\ &\vdots \\ d_{(n-1)t} &< d_{nt} \end{aligned} \tag{7}$$

我们仍希望满足以上等式与不等式的约束的新的权重能够与我们已经求解的权重的差异尽量小，于是最大化两个权重的相关性，即最小化这个相关性的相反数，这个式子仍然是关于 $w_1, w_2, \cdots, w_l$ 是线性的：

$$f(W) = -\left(w_1 \times ga\_w_1 + w_2 \times ga\_w_2 + \cdots + w_l \times ga\_w_l\right) \tag{8}$$

其中 $ga\_W = \left(ga\_w_1, ga\_w_2, \cdots, ga\_w_l\right)$ 是我们之前使用 GA-KNN 算法求解出的最佳权重。

在(5)和(6)的线性约束下最小化一个线性函数(7)，这是典型的线性规划问题。于是，我们将其解出。假设有 m 名专家对最近邻检索的结果进行排序，那么我们就将进行 m 次上面的这种过程。

把我们之前求解的最佳权重与线性规划解出的权重组成一个权重的种群，然后将这个种群作为遗传算法的初始种群，执行 GA-KNN 算法再次执行权重优化。

3.2. 神经网络训练及决策输出阶段

神经网络是对人脑若干基本特性的抽象利模拟。它是以人的大脑工作模式为基础，研究自适应及非程序的信息处理方法。这种工作机制的特点表现为通过网络中大量神经元的作用来体现它自身的处理功能，从模拟人脑的结构和单个神经元功能出发，达到模拟人脑处理信息的目的[10]。它是一种客观的机器学习方法。

为了在利用主观先验知识的同时体现分类结果良好的客观公正性，本文方案将人工参与调解的模型(即得到的加权 KNN 模型)与人工神经元网络进行组合。将资源的各种分类等级输入到神经网络中，利用神经网络的学习和推理能力，得到最终的分类结果，从而为资源的分类提供客观公正的等级参考。

我们将资源分为 16 类，即每一类对应神经网络中的一个输出节点。同时，我们按照提取好的资源特征将教学资源向量化表示，因此，神经网络的输入数据应具有 61 个属性。于是，我们的神经网络应具有 61 个输入节点。本文使用传统的三层神经网络，即该网络具有一个输入层、一个隐藏层和一个输出层。我们使用公式来计算隐藏节点的个数：

$$o = \sqrt{m \times n} \tag{9}$$

其中,m 是可视节点的个数，n 是隐藏节点的个数，o 是通过计算的到的隐藏节点的个数。本文中使用的神经元网络具有 61 个输入节点和 16 个输出节点，因此，我们计算得到的隐藏节点个数为 20 个。另外，在我们的神经网络中的相邻层之间的节点间是全连接的。

在图 2 中，我们给出了具有传统的三层网络结构并且能适用于本文研究的教育资源分类决策的课题的神经网络结构示意图。

4. 结论

本研究的实验数据从 1200 个教学资源文书中获取。以教育部教育资源建设技术规范为基础，将 1200 篇文档中的信息提取为向量形式的数据。我们将其中的1000条数据作为训练集，将另外200条数据作为测试集。

整个过程共有 15 名教育专家参与参数调整的过程。在求解最优权重的过程中，我们进行了 10 轮的 GA-KNN 算法求解过程，每次 GA-KNN 算法求得一个最佳权重以及 20 个与待分类资源最接近的资源，然后由专家从主观经验上对这些相似资源按照与待分类资源的近似程度来进行排序。排序后的结果用于产生新的权重的初始种群，这个初始种群再次经过 GA-KNN 算法来求得最优权重。

以上过程进行 10 轮之后，便得到了最终的最优权重，此权重将用作加权的 KNN 算法的属性权重。

在获取最终的 KNN 算法的最优权重之后，我们就从测试集中随即抽取一个资源作为待分类资源，使用加权的 KNN 算法从 1200 个资源中检索出 100 个与待分类资源最相似的资源作为三层神经网络的训练集。我们将这 100 个最相似资源的属性作为神经网络的输入，将表示其分类结果类型的向量作为神经网络的输出。并使用随机梯度下降法来训练我们的神经网络。这 100 个资源的数据将在训练过程中被重复使用，在本研究中，我们设训练的迭代次数为 100 次到 500 次。

作为对照组实验，我们再次执行由 GA-KNN 算法来求解加权 KNN 算法的最优权重并且使用加权的 KNN 算法检索出来的资源用作三层神经网络的训练数据集过程。这个过程的每个操作的执行次数与上面提出的实验完全相同。这个过程与本文提出的方案的唯一区别在于，这个过程没有专家评审的步骤。

我们提出的方案是一种机器学习结合人类主观经验来进行分类的过程。而对照组的实验的学习过程则完全依赖于机器学习方法来完成。从而，这两组实验所求得的最佳的加权 KNN 算法的属性权重是不同的。理论分析的结果与实验结果符合，如图 3 所示。

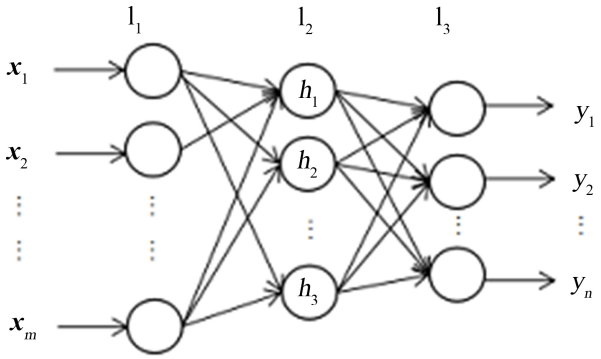

Figure 2. Neural network structure diagram
图 2. 神经网络结构示意图

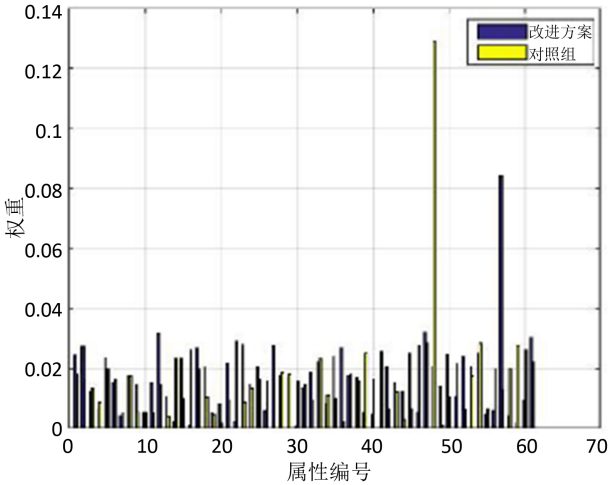

Figure 3. The weight of control experiment contrast
图 3. 对照实验的权重对比

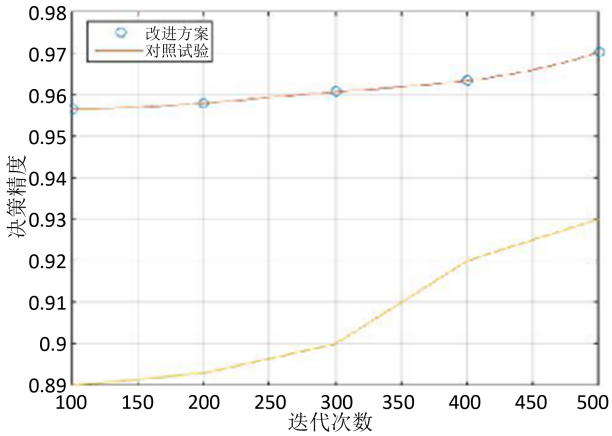

Figure 4. Precision control experiment contrast
图 4. 对照实验精度对比

从图中可以看出，两组实验所求得的权重在某些属性上有显著的差异。从而，尽管我们使用该权重加权的 KNN 算法来检索相似资源并且把这些相似资源作为神经网络的训练数据，同时保证神经网络的结构相同，但是由于使用了不同的权重来检索相似资源从而导致神经网络的训练数据与待分类资源的相似程度不同，因而影响了神经网络的训练效果，进而两组实验获得的决策精度是不同的。神经网络的训练迭代次数是 100 次到 500 次，我们所得到的精度曲线图如图 4 中所示。

因此，本文将 GA-KNN 算法、线性规划与神经网络学习相结合，形成了借助 GA-KNN 及线性规划的基于本体的教学资源分类方案，该方案继承了机器学习客观公正的特点，同时又使用了主观经验来进一步提高系统的分类精度。实验证明，本方案与传统分类方案相比，在分类的精度上有更好的表现。

参考文献 (References)

[1] 杨建林. 基于本体的文本信息检索研究[J]. 情报理论与实践, 2006, 29(5): 598-601.

[2] 余胜泉, 朱凌云. 《教育资源建设技术规范》体系结构与应用模式[J]. 中国电化教育, 2003(3): 51-55.

[3] 钱晓东, 王正欧. 基于改进 KNN 的文本分类方法[J]. 情报科学, 2005, 23(4): 550-554.

[4] 杜尔斌, 李翔, 林祥. 改进的 KNN 文本分类算法[J]. 信息安全与通信保密, 2011(4): 38-39.

[5] Yang, Y. and Liu, X. (1999) A Re-Examination of Text Categorization Methods. *Proceedings of the* 22*nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Berkeley, CA, 15-19 August, 42-49.

[6] 冯璠, 严广乐. 基于优化 KNN 算法的线上拍卖价格预测模型[J]. 信息技术, 2015(3): 40-43.

[7] 周明, 孙树栋. 遗传算法原理及应用[M]. 北京: 国工业出版社, 2001.

[8] 葛继科, 邱玉辉, 吴春明, 等. 遗传算法研究综述[J]. 计算机应用研究, 2008, 25(10): 2911-2916.

[9] 梁宇宏, 张欣. 对遗传算法的轮盘赌选择方式的改进[J]. 信息技术, 2009, 33(12): 127-129.

[10] 许万增. 神经网络的研究及应用[J]. 神经网络的研究及其应用, 1990(1): 23-26.