百家论坛

# 基于不确定性的大数据学习模型

王熙照[1]，朱 红[2] / 1. 深圳大学 2. 澳门科技大学

近年来，基于不确定性的机器学习模型研究得到了飞速的发展[1-7]。不确定性处理（包括其描述、度量、建模、处理等）对整个数据分析和模型学习过程有着非常重要的影响。如果对不确定性进行了不恰当的处理，学习方法的性能会被大大地降低。

## 不确定性的定义

目前，尚不存在不确定性对于所有情况都适用的通用定义。我们通常在某个特定的背景下，对不确定性进行讨论。这里列出了五种对于不确定性的度量标准，即香农熵 (SE)[4]、分类熵 (CE)[6]、模糊度[1-2]、非特异性[5]和粗糙度[7]。不确定性通常指某个概念不能被清晰准确地描述。在数学层面上还没有对于不确定性的一般定义，但是在不同的知识背景下，会得到不确定性在对应背景下的特定定义。表 1 是对于几种从数学角度进行阐释的不确定性的简介。

表 1 不同类型的不确定性

| 不确定性 | 研究对象 | 不确定性的来源 |
|---|---|---|
| 香农熵 | 概率 distribution | 由随机现象引起的不确定性 |
| 分类 | 精确集 | 集合中元素分布的杂乱性程度 |
| 模糊性 | 模糊集 | 由界限不清晰引起的不确定性 |
| 不明指向性 | 模糊集 | 处理一对多关系时产生的不确定性 |
| 粗糙度 | 粗糙集 | 上 / 下近似 |

下面讨论一种典型的不确定性——模糊集的模糊性。模糊性被用来描述两个语义之间的不明确性程度，比如热和冷。模糊性最早是由 Zadeh 在 1968 年提出的，他也是模糊集理论[8]的提出者。Zadeh 模糊集理论的基本思想是，隶属度的函数值从原来的只为 0 或 1 扩展到了区间 [0,1]。由于主观上有对于语义理解的不确定性，所以隶属度的函数值范围被扩展了。在模糊集理论的基础上，Luca 和 Termini 在 1972 年提出模糊性是一种由模糊集描述的不确定性，而且他们用类似于香农信息熵的非概率熵定义了模糊性的度量标准[9]。他们还提出模糊性应该满足三条性质，由这些性质可以得出，如果所有元素关于某个集合的隶属度都相等，则该集合的模糊度达到最大值；如果所有元素关于某个集合的隶属度为 0 或 1，则该集合的模糊度达到最小值。此外，Luca 和 Termini 将熵的定义扩展到了模糊集领域[10]。这一扩展得到的定义不仅可以是一个数量值，也可以是一个列矩阵或向量。

## 大数据不确定性学习的研究

一个建立在常规数据集上的学习模型和算法一般是不能拓展到大数据的，原因有多个。基于不确定性的学习模型自然也是如此。不确定性的处理对大数据学习更为重要，有些与不确定性有关的问题只有在大数据集上才有，在常规数据集上原本不是问题。我们在此简要介绍两种基于大数据学习的不确定性的研究，一种是基于模糊性的半监督学习；另一种是基于不可指定性的处理混合条件属性的模型树。其中，第一项研究工作，基本满足如图 1 所示的基于不确定性的大数据学习的一般框架[2]。

图 1 中，分类器 A 的训练精度与分类器 B 的训练精度相同，但是 A 的不确定性小于 B 的不确定性

（例如模糊性或不明确性）。我们称对于某些类型的大数据（并非所有类型），分类器 A 比 B 有更强的泛化能力。与传统的模式识别观点相比，A 的这一优势为学习算法的设计提供了一个截然不同的思路。

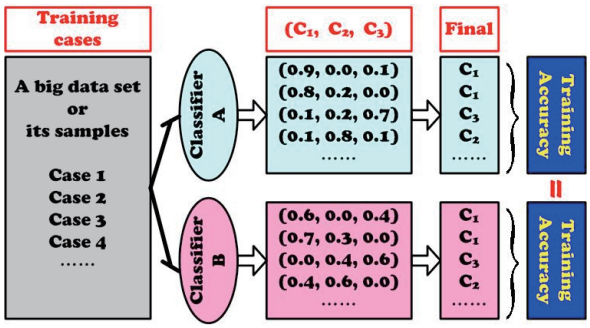


图 1 基于不确定性的大数据学习的一般框架

### 1. 基于模糊性的半监督学习

假设 A 是一个大数据集，并且 A 中的大部分样例没有类标；B 是 A 中一小部分样例组成的集合，并且 B 中的每个样例都有类标。利用数据集 B 我们可以训练得到分类器，但我们不能保证这样的分类器对 A-B 中的样例有较好的预测结果。基于对数据集 A-B 中的每个样例的预测结果，我们想从 A-B 中挑选出一些样例（连同对这些样例的预测结果）加入到数据集 B 中。再次利用 B 进行训练得到的分类器对于 A-B 中样例的预测精度将会有所提升。此刻需要明确的关键问题是训练得到的分类器应该满足哪些条件和我们应该怎样从 A-B 中挑选样例。理论上讲，训练得到的分类器必须满足训练精度大于 0.5。在以下描述的算法 1 中我们将从不确定性的角度讨论选择样例的策略。

算法 1：基于模糊性的样例选择

步骤 1：将数据集 A 随机划分为训练集 B 和测试集 A-B;
步骤 2：基于集合 B 训练得到一个基本的分类器；
步骤 3：对于每个既在训练集又在测试集中的样例，得到基于上述基本分类器的模糊向量输出；
步骤 4：计算每个输出结果的模糊度；
步骤 5：分别基于训练集中的模糊度和测试集中的模糊度对样例进行 排序；
步骤 6：基于步骤 5 中的排序结果，将训练集和测试集分别划分成三组，即高模糊度组 G1，中模糊度组 G2 和低模糊度组 G3；
步骤 7：G1 组和 G3 组连同它们的预测类标将会被添加到集合 B 中用以进行下一轮训练。

需要特别注意的是为了提高学习性能，我们通常只使用 G3 组，然而在此学习算法中 G3 组和 G1 组都被用到了。

我们采集了一个关于中国象棋游戏局面分类（CCGSC）的大数据集，作为示例来说明分类器的训练过程。该数据集所占计算机的存储空间为 1.86 GB，包含了 107 条象棋游戏记录，多于 109 条棋局记录。这是一个典型的基于非结构化数据的半监督学习，其中大量的棋局没有类标。为了得到复杂棋局的预测结果，我们需要请教象棋大师，这是一项耗费相当巨大的工作。传统的预测方法是根据棋局预测函数计算出一个数值，然后根据这个数值得到对棋局结果的预测，但是用该方法得到的精度很低。基于 CCGS 分类数据的实验结果表明，基于模糊性的半监督学习算法可以得到很高的预测精度。对不确定性的适当处理能够十分显著地提升分类系统的性能，这一事实进一步证明了我们的陈述。

### 2. 基于不可指定性的处理混合条件属性的模型树

模型树是处理混合条件属性（大数据多模态的一个特例）分类问题的一种有效方法，其中混合条件属性是指在信息决策表中部分条件属性的取值是符号型的，而另一部分条件属性的取值是数值型的。从全局来看，模型树是一种树结构，但在每一个叶子节点都有一个特定模型被构建。在基于不明确性的模型树 (AMT) 中，决策树的构建原则是尽量减少父节点划分产生子节点过程中的歧义。模型树的叶子节点是一个由极速学习机 (ELM) 算法 [11-13] 训练得到的三层前馈神经网络。在 AMT 中，我们分别用决策树和 ELM 来处理离散型属性和连续型属性。以下列出的算法 2 对基于不可指定性的模型树的生成过程进行了简要地描述。近年来深度学习 [14] 一

直是一个非常热门的课题，通过与深度学习的结合，AMT 可以被扩展到属性是图像和文本的问题中。深度学习本质上是一个自动特征选择策略，最初开发深度学习的目的是对图像进行特征提取和分类。对于属性为图像的大数据的分类问题而言，结合深度学习的模型树将是一个非常有效的方法。最近的一些研究 [15-16] 表明，在性能方面 ELM 自动编码器要优于多种不同技术水平的深度学习算法。

算法 2：基于不明确性的模型树 (AMT)

输入：混合属性的大数据集 S
输出：基于不明确性的模型树
步骤 1：选出具有最小不明确性的条件属性 作为模型树的根节点；
步骤 2：根据离散型条件属性的取值将当前父节点划分为个子节点；
步骤 3：对于每个子节点，选择出不明确性小于划分属性的离散型条件属性；
步骤 4：重复步骤 2 和 3，直到各个子节点不明确度的最大值小于给定的阈值；
步骤 5：将不再被划分的子节点作为叶子节点，在该叶子节点上对连续型条件属性的样例进行训练得到一个 ELM。

几个大数据集（样例个数超过两百万）的实验结果表明，我们所提方法的并行化算法有良好的性能。并行 AMT 算法的训练时间随着计算机数量的增多而减少，这表明并行算法是可以减少计算时间的；实验结果还表明，我们所提的 AMT 算法有很好的泛化能力。在基于 15 个数据集的对比实验中我们可以看到在大多数数据集上 AMT 算法的测试精度要高于功能树 [17]、朴素贝叶斯树 [18] 和逻辑模型树 [19-20] 的测试精度。

## 结束语

到目前为止，大数据还没有一个数学定义，但它可以被一些特性描述，比如它的 5v 特性。本文主要关注第四个特性，即不确定性，试图说明：① 一些关于不确定性处理的问题，如数据集中每个样例都有 80% 以上的数据缺失问题，该问题只在大数据环境中出现；② 处理嵌入到数据分析整个过程中的不确定性对于大数据的学习性能有重大的影响。在图 2 中我们对大部分处理大数据计算的方法进行了总结，并且突出了数据的规模从大到小变化的效果。

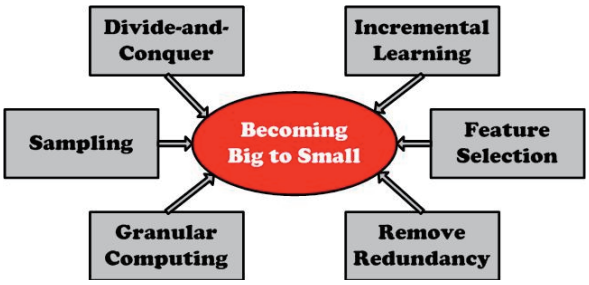


图 2 大数据分析的主要内容是将数据的规模由大变到小。不确定性模型的处理方法对于变化效果起着关键作用

参考文献

[1] X. Z. Wang, R. A. R. Ashfaq, and A. M. Fu. Fuzziness Based Sample Categorization for Classifier Performance Improvement. Journal of Intelligent & Fuzzy Systems, 2015, 29(3): 1185-1196.
[2] X. Z. Wang, H. J. Xing, Y. Li, Q. Hua, C. R. Dong, and W. Pedrycz. A Study on Relationship between Generalization Abilities and Fuzziness of Base Classifiers in Ensemble Learning. IEEE Transactions on Fuzzy Systems, 2015, 23(5): 1638-1654.
[3] X. Z. Wang, R. Wang, H. M. Feng, and H. C. Wang. A New Approach to Classifier Fusion Based on Upper Integral. IEEE Transactions on Cybernetics, 2014, 44(5): 620-635.
[4] X. Z. Wang, Y. L. He, and D. D. Wang. Non-Naive Bayesian Classifiers for Classification Problems with Continuous Attributes. IEEE Transactions on Cybernetics, 2014, 44(1): 21-39.

[5] X. Z. Wang, L. C. Dong, and J. H. Yan. Maximum Ambiguity Based Sample Selection in Fuzzy Decision Tree Induction. IEEE Transactions on Knowledge and Data Engineering, 2012, 24(8): 1491-1505.
[6] X. Z. Wang and C. R. Dong. Improving Generalization of Fuzzy If-Then Rules by Maximizing Fuzzy Entropy. IEEE Transactions on Fuzzy Systems, 2009, 17(3): 556-567.
[7] Z.B Xu, J. Y. Liang, C.Y. Dang, and K.S. Chin. Inclusion Degree: A Perspective on Measures For Rough Set Data Analysis. Information Sciences, 2002, 141(3-4): 227-236.
[8] L.A. Zadeh. Probability measures of fuzzy events. Journal of Mathematical Analysis and Applications 23 (1968), 421-427.
[9] A. De Luca and S. Termini. A definition of a nonprobabilistic entropy in the setting of fuzzy sets theory. Information and Control 20 (1972), 301-312.
[10] A. De Luca and S. Termini. Entropy of L-fuzzy sets. Information and Control 24 (1974), 55-73.
[11] G. B. Huang, D. H. Wang, and Y. Lan. Extreme Learning Machines: A Survey. International Journal of Machine Learning and Cybernetics, 2011, 2(2): 107-122.
[12] G. B. Huang, H. Zhou, X. Ding, and R. Zhang. Extreme Learning Machine for Regression and Multiclass Classification. IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics, 2012, 42(2): 513-529.
[13] A. M. Fu, C. R. Dong, and L. S. Wang. An Experimental Study on Stability and Generalization of Extreme Learning Machines. International Journal of Machine Learning and Cybernetics, 2015, 6(1): 129-135
[14] G. Hinton and R. Salakhutdinov. Reducing the Dimensionality of Data with Neural Networks. Science, 2006, 313: 504-507.
[15] L. L. C. Kasun, H. Zhou, G. B. Huang, and C. M. Vong. Representational Learning with Extreme Learning Machine for Big Data. IEEE Intelligent Systems, 2013, 28(6): 31-34.
[16] J. Zhang, S. F Ding, N. Zhang, and Z. Z. Shi. Incremental Extreme Learning Machine Based on Deep Feature Embedded. International Journal of Machine Learning and Cybernetics, 2015, DOI: 10.1007/s13042-015-0419-5.
[17] J.Gama. Functional Trees. Machine Learning, 2004, 55(3): 219-250.
[18] R. Kohavi. Scaling Up the Accuracy of Naïve Bayes Classifiers: A Decision-Tree Hybrid. In Proceedings of KDD'96, 1996, pp. 202-207.
[19] N. Landwehr, M. Hall, and E. Frank. Logistic Model Trees. Machine Learning, 2005, 59(1): 161-205.
[20] M. Sumner, E. Frank, M. Hall. Speeding Up Logistic Model Tree Induction. In Proceedings of PKDD'05, Lecture Notes in Computer Science, 2005, 3721: 675-683.

**王熙照**

博士，深圳大学计算机与软件学院教授、博士生导师，大数据研究所副所长。IEEE Fellow，Springer杂志 Machine Learning and Cybernetics 主编。主要研究方向为机器学习与不确定性信息处理。

**朱 红**

澳门科技大学资讯科技学院在读博士研究生。主要研究方向为决策树、神经网络。