熵在数据科学中的应用

王湘峰 PB19030861

摘　要

熵的概念由德国物理学家克劳修斯于 1865 年提出，最初用于度量热力学中系统的无序程度。1948 年香农提出信息熵的概念，用以度量样本的信息量，此后熵[1]便在信息学和数据科学领域不断发挥着重要的作用。

本文将从熵的概念以及诞生背景入手，逐步深入地介绍熵在数据分析、数据挖掘以及机器学习[2]领域的应用，包括其在数据压缩、自然语言处理、决策树等方面的应用，以及展望未来熵在数据科学领域应用的前景。

关键字：信息熵 数据挖掘 数据分析 机器学习

The Application of Entropy in Data Science

Xiangfeng Wang

Abstract

The concept of entropy was first proposed by the German physicist Clausius in 1865 as a measure of the disorder of systems in thermodynamics. In 1948, Shannon proposed the concept of information entropy to measure the amount of information in sample. Since then, the Shannon entropy has been playing an important role in the field of informatics and data science.

Starting from the concept of entropy and the background of its birth, this paper will introduce the application of entropy in data analysis, data mining and machine learning step by step, including its application in data compression, natural language processing, clustering, decision tree and other aspects, and look into the future application of entropy in the field of data science.

Keywords： Information Entropy　Data Mining　Data Analysis　Machine Learning

1　熵的一般定义与含义

• 熵的定义

[1] 下文的熵都代指信息熵，与热力学熵相区分

[2] 事实上，数据挖掘和机器学习有很多重合的地方，本文将综合考虑

根据 Shannon 在论文《通信的数学原理》[1]中的定义，随机变量 X 的熵值 **H** 为：

$$\mathbf{H}(X) = \mathrm{E}[\mathrm{I}(X)] = \mathrm{E}[-\ln(\mathrm{P}(X))]$$

其中 P 是 X 的概率分布函数，I(X)表示 X 的自信息

$$\mathrm{I}(X) = \log_b \mathrm{P}(\mathrm{X})$$

因此熵可以表示为

$$\mathrm{H(X)} = \sum_i P(x_i) \cdot I(x_i) = -\sum_i P(x_i) \cdot log_b P(x_i)$$

X 为离散随机变量，或

$$\mathrm{H}(X) = -\int f(x) \cdot log_b f(x)\, dx$$

X 为连续随机变量。

值得注意的是，对于不同的 b 的取值，计算出的熵的单位也不同。一般来说 b=2 时单位为 bit；b=e 的时候单位为 nat；b=10 的时候单位为 Hart。

• 熵的含义

不难发现，一条信息的信息量与它的不确定性有很大的关联。一个事件如果不确定度越大，那么我们越要花费更多的额信息去了解它；反之事件愈是确定，则它能够带来的信息愈少。那么如何定量的描述信息量呢？

举个例子，假如下一次诺贝尔奖有 16 位候选人，并且最终只能选出一位颁奖。倘若我错过了颁奖典礼，之后我去问一位观众“谁是诺贝尔奖得主”？他不愿意直接告诉我而只愿意回答“是”或“不是”，并且每次回答都要收取一元小费，那么不难想到，我至少要花费 4 元才能知道谁是诺贝尔奖得主，因此这条信息的“价值”是 4 元钱。

类似的，在信息的世界中，香农用比特(bit)而不是钱来衡量信息量。沿用刚才的例子的话，“谁是诺贝尔奖得主”这条信息价值 4 比特。进一步的，我们发现，有些情况下甚至不需要四次就能知晓答案：假如爱因斯坦、牛顿仍然活着并且参与候选，那么我们可以将最可能获奖的分为一组，其他的再分若干组，这样可能二到三次就可以得出结果。此时这条信息的信息量应该小于 4 比特。香农认为，这条信息的严格的信息熵应为

$$\mathrm{H}= -(p_1 \cdot \log p_1 + p_2 \cdot \log p_2 + \cdots p_{16} \cdot \log p_{16})$$

其中 p_i 代表每个候选人获奖的概率，而这这恰好是熵的定义的具体化。

总而言之，熵代表着信息的不确定性或信息量，二者在信息的世界中是“等价”[3]的。

2 熵在数据分析中的作用

• 熵与数据压缩

在日常生活中，我们常常会用到压缩软件，原因无他，就是可以降低文件的存储空间，在有限的空间内存储更多的信息。可是有没有想过，如果把一个已经被压缩的文件反复压缩，那么文件会一直缩小吗？如果不是，那么是否存在一个压缩上限是不可被超越的？幸运的是，有了熵的概念之后，我们可以在数学上严格的证明文件压缩的上限。

首先我们要了解压缩的原理，一个简单而朴素的思想是：把重复的内容用更短的符号代替。例如“AAAAAAAAA”可以压缩为“9A”；相应的，如果信息内容杂乱无章甚至不同信息均匀分布，那么它将几乎无法压缩。

考虑一般情况，在分布较为均匀的情况下，如果一个元素（字符）在文件中出现的概率为 p，那么在该位置最多可以出现 1/p 种情况。对于一个由 n 个部分组成的文件，每个部分的内容在文件中的出现概率分别为 p1、p2、...p_n。那么至少应需要的二进制字符数为

$$log_2 p_1 + log_2 p_2 + \cdots log_2 p_n = \sum_i log_2 p_i$$

考虑到每个部分出现的概率，上式应该改为

$$E[log_2 p] = \sum_i p_i log_2 p_i$$

这恰好与上文信息熵的计算公式不谋而合。因此信息用二进制字符压缩的极限恰好在数值上等于它的信息熵大小[4]。而这个思想，也是由数学家香农首先提出的。

[3] 指度量标准和计算规则相同，具体意义应当根据使用场景具体判断

[4] 这里的信息熵的单位是 bit

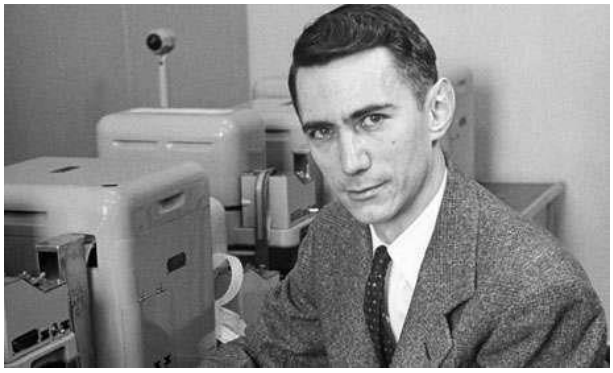

图 1 香农（Shannon，1916-2001）

因此，熵可以告诉我们信息的压缩极限[5]，我们可以不用像前人那样一味地追求压缩率，熵在数据的压缩中具有指导性的意义。

• 熵与信息相关性

由前文可知，信息与不确定度有着很大的关联，通过信息我们可以消除不确定度。在日常生活中我们发现，有些时候多个信息之间会产生关联，从而额外地消除了一些不确定度：例如在网页上搜索“长城”时，搜索引擎并不知道我想要的是万里长城还是长城汽车，可是如果在后面加上“股票”，那么我们往往认为是指长城汽车。那么为什么“相关的”信息会带来不确定度的消除（即熵的减少），以及如何衡量这种减少呢？

为此我们引出**条件熵**(Conditional Entropy)的概念：

假设 X 和 Y 是是两个随机变量 ，已知 X 的概率分布$P(x)$以及 X 和 Y 的联合概率分布$P(x,y)$，那么 X 的熵为

$$\mathrm{H(X)} = -\sum_{x\in X} P(x)\cdot logP(x)$$

给定 Y 取值的条件下的条件熵为

$$H(X|Y) = -\sum_{x\in X, y\in Y} P(x,y)\cdot logP(x|y)$$

有许多学者已经证明$H(X) \geq H(X|Y)$[6]，也就是说在已知 Y 的信息的情况下，X 的不确定度至少不会上升。事实上，在统计语言模型中，如果把 Y 看作是 X 的

[5] 更严谨的说法是香农第一定理

[6] 这里不再赘述

前一个字，那么 X 的不确定度会比单独存在时更小。这说明二元模型比一元模型更好。类似的三元模型一般比二元模型更好，即存在如下关系

$$H(X|Y,Z) = -\sum_{x\in X,y\in Y,z\in Z} P(x,y,z)\cdot logP(x|y,z) \le H(X|Y)$$

理论上模型的变元越多，每个变元的条件熵就会越小，可是考虑到计算的复杂度随变量个数的指数级增长，无限制扩充条件熵仅存在理论意义。另外值得一提的是，上式等号成立的条件为，条件中的变量与我们关心的变量 X 毫无关联。例如知晓今天的天气对预测下一次掷骰子的点数毫无帮助。

很自然的，如果我们要量化随机变量 Y 与 X 的相关性，或者说了解 Y 对预测 X 的帮助，我们可以用 H(X)-H(X|Y)来计算。这即是**互信息**[7](Mutual Information)的概念

$$I(X;Y) = \sum_{x\in X,y\in Y} P(x,y)\cdot log\frac{P(x,y)}{P(x)P(y)}$$

容易验证，$I(X;Y) = H(X) - H(X|Y)$

现在我们知道，所谓两个事件的相关性，就是在已知一个事件的前提下，对消除另一个事件的不确定度的影响。当 X 与 Y 完全相关时，I(X;Y)=H(X)；当完全无关时，I(X;Y)=0.

- **熵与机器翻译**

在自然语言处理中，计算语言特征的互信息并不是一件难事，在拥有充足的样本的前提下，$P(X,Y)$, $P(X)$以及 $P(Y)$可以被很好的估计出来，进而计算出它们之间的互信息。互信息被广泛地应用于度量语言现象的相关性。

让互信息大显身手的一个领域是机器翻译。机器翻译遇到的最大问题是词汇的歧义性(Ambiguation)。例如 Carpenter 既可以作为木匠，也可以指歌手 Karen Carpenter[8].那么如何才能让机器正确的翻译呢？仅仅考虑语法的话，Carpenter 不论是木匠还是歌手都是名词且都是人，都是符合语法规则的。如果企图通过添加详细的规则如“流行音乐家做主语时，翻译为卡朋特，否则翻译为木匠”来优化的话，那工作量将是难以承受的，而且不具备泛化能力。

[7] 互信息同样是由香农率先提出的

[8] Karen Carpenter,1950-1983,美国歌手，代表作《yesterday once more》（昨日重现）

图 2 卡伦卡朋特（左）

解决这个问题的是一个简单却实用的方法，那就是：首先从大量文本中找出和木匠互信息最大的词汇，如锯子、木质家具、雕刻等 ；再找到与歌手卡朋特一起出现的词汇，如流行音乐、《昨日重现》、康涅狄格州等。在翻译 Carpenter 时对比上下文哪种词汇多就好了。这个方法是由 William Gale, Kenneth Church 和 David Yarowsky 于 1992 年在论文《A Method for Disambiguating Word Senses in a Large Corpus》[2]中提出的。

we will sort contexts c by

$$score(c) = \prod_{token \text{ in } c} \frac{Pr(token|sense_1)}{Pr(token|sense_2)}$$

where $Pr(token|sense)$ is an estimate of the probability that *token* appears in the context of $sense_1$ or $sense_2$.

图 3 论文中以翻译单词 token 举例

• 熵与信息检索

讲完了条件熵与互信息，接下来要引出的是信息论中的另一个重要的概念—**相对熵**[9](Relative Entropy,或 KL 散度)。相对熵首先是由 Kullback 和 Leibler 提出的，因此又叫 KullbackLeibler Divergence.相对熵也是用来衡量变量的相关性的，不过与互信息不同的是，相对熵用来衡量随机变量分布的相似性。在统计模型推断中，相对熵又称为讯息增益(information gain)或讯息散度(information

[9] 有的文献称之为交叉熵，具体情况下文有解释

divergence)[10].定义如下

$$H(f(x)||g(x)) = KL(f(x)||g(x)) = \sum_{x\in X} f(x)\cdot log\frac{f(x)}{g(x)}$$

注意到

$$\sum_{x\in X} f(x)\cdot log\frac{f(x)}{g(x)} = \sum_{x\in X} f(x)\cdot \log f(x) - f(x)\cdot \log g(x)$$

$$= -\sum_{x\in X} f(x)\cdot log\ g(x) - H(f)$$

此时第一项即为**交叉熵**，可以看出当*f(x)*分布不变时，交叉熵和相对熵只差一个常数。

对于相对熵，我们可以直观的看出：

1. 两个完全相同的分布的相对熵为 0

2. 相对熵是分布差异的量化，相对熵越大则差异越大，反之越小

由于交叉熵可以衡量分布之间的差异，因此 KL 散度（即交叉熵）又被称为 KL 距离，但应当注意的是，这里的“距离”并非是一个良好的比喻，因为交叉熵是不对称的，即

$$KL(f(x)||g(x)) \neq KL(g(x)||f(x))$$

为了解决这个不便之处，詹森和香农将它们进行线性组合以消除不对称性[3]

$$JS(f(x)||g(x)) = \frac{1}{2}[KL(f(x)||g(x)) + KL(g(x)||f(x))$$

上式被称为詹森-香农散度(Jesen-Shannon Divergence)有了相对熵，我们就可以引出一个在信息检索中的主要概念：词频率-逆向文档频率(TF-IDF).这个概念将会在下一章节详细的讲解。

3 熵在数据挖掘、机器学习中的应用

• 交叉熵与 TF-IDF

首先介绍词频率(TF)的概念。设一篇文章中有 m 个词汇，其中单词 w_i 的出现次数为 n_i，则该单词的词频率为

$$TF(w_i) = \frac{n_i}{m}$$

[10] 引用自 wiki 百科

若假设资料库中共有 D 篇文章，其中出现单词 w_i 的文档数为 $D(w_i)$，那么该单词的逆文档频率(IDF)定义为

$$IDF(w_i) = log_2 \frac{D}{D(w_i)}$$

现在考虑“如何提取文章的关键词”这个问题，不难想到，如果一个词汇在一篇文章中出现的概率远高于这个词在整个数据库中的概率，那么这个词更有可能成为这篇文章的关键词[11]。在开始计算之前，我们需要做一个理想化假设：

1.每篇文档大小基本相同，大约为 m 个词。

2.以频率估计概率

那么词 w_i 在全部文档中出现的概率 $q(w_i)$ 为

$$q(w_i) = \frac{n_i D(w_i)}{mD}$$

词 w_i 在文档 D 中出现的概率 $p(w_i)$ 为

$$p(w_i) = \frac{n_i}{m}$$

考虑它们的 KL 距离：

$$D_{KL}(p||q) = p(w_i) log_2 \frac{p(w_i)}{q(w_i)} = \frac{n_i}{m} log_2 \frac{n_i/m}{n_i D(w_i)/mD}$$
$$= \frac{n_i}{m} log_2 \frac{D}{D(w_i)} = TF(w_i) \cdot IDF(w_i)$$

根据前文的分析我们知道，交叉熵越大，则二者的分布差异就越大。在文本中，这说明词汇w_i的指向性越强，越能描述文档 D，因此它越能胜任关键词的地位。

• 熵与决策树

决策树(Decision Tree)是在已知各种情况发生概率的基础上，通过构成决策树来求取净现值的期望值大于等于零的概率，评价项目风险，判断其可行性的决策分析方法，是直观运用概率分析的一种图解法。由于这种决策分支画成图形很像一棵树的枝干，故称决策树。在机器学习中，决策树是一个预测模型，他代表的是对象属性与对象值之间的一种映射关系。[12]

[11] 这个想法最早由 Karen Sparck Jones 提出，详见参考文献[4]

[12] 引用自百度百科

在构建决策树的过程中，最大的难题是如何划分分支，或者说以什么标准来划分。以文献**错误!未找到引用源。**中的例子来看，如果我们要对“这是好瓜吗?”这样的问题进行决策时，通常会进行一系列的判断或“子决策”我们先看“它是什么颜色?”，如果是“青绿色”，则我们再看“它的根蒂是什么形态?”，如果是“蜷缩”，我们再判断“它敲起来是什么声音?”最后，我们得出最终决策:这是好瓜。

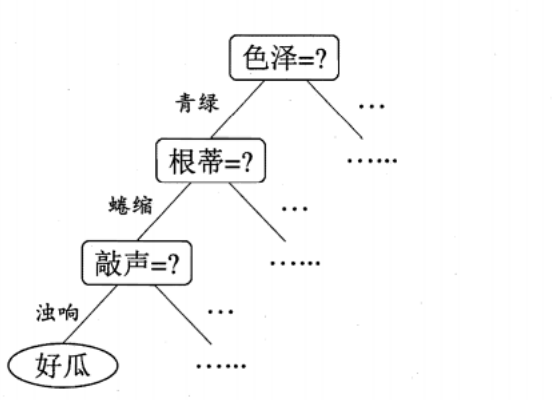

图 4 关于西瓜的决策树

不难发现，决策树学习的关键在于找到最优划分属性。一般而言，随着划分过程不断进行，我们希望决策树的分支结点所包含的样本尽可能属于同一类别，即结点的“纯度” (purity) 越来越高。

这个时候我们可以用信息熵作为一个良好的评估标准。即对于某个属性的划分，分别计算划分前后的总的信息纯度，然后挑选对纯度增益最大的属性即可。这里引入**信息增益**(information gain)的概念

假设一个样本集合 D 中共含有 m 类样本，经过属性 a 的划分后产生了 n 个子集，对于每个子集跟它们包含的元素数量赋予它们权重$|D^v|/|D|$.

$$H(D) = \sum_{k=1}^{m} p_k log_2 p_k$$

$$Gain(D, a) = H(D) - \sum_{v=1}^{n} \frac{|D^v|}{|D|} H(D^v)$$

一般来说，信息增益越大，划分后的分支越“纯净”，或者说确定度越高，在做决策时的置信度越高，决策树性能越好。这正是著名的 ID3 决策树学习算法[5]

- **最大熵原理与最大熵模型**

在金融投资领域，人们常说“不要将鸡蛋放在同一个篮子里”以降低风险，在数学上，这种思想被称为**最大熵原理**(The Maximum Enropy Principle). 最大熵原理在日常生活中也被广泛地应用，例如学生在复习备考时，如果仅知道考试范围的话，往往会将所有的知识点和内容进行复习，并且投入的时间大致相同，此时模型的熵达到最大值。一个直观的理解是，最大熵原理保留了最大的不确定度以降低风险。进一步地，如果这位同学知道要考试的三章中第一章的分值占 50%，那么一般来说这位同学会花费 50%的时间复习第一章，而分别花费 25%的时间复习第二三章，即使他不知道第二三章的占比究竟如何。

朴素地来讲，当我们遇到不确定的问题时，应当保留最大的可能性。这个想法同样适合机器学习，通过构建最大熵模型来处理实际问题。那么这样的模型是否一定存在呢？香农奖得主 I.Csiszar 证明了，对于任意一组不矛盾的信息，最大熵模型存在且唯一[6].

例如对于一个根据上下文和主题来预测下一个词的模型，它具有如下的指数形式

$$P(w_3|w_1,w_2,s)=\frac{1}{Z(w_1,w_2,s)}e^{\lambda_1(w_1,w_2,w_3)+\lambda_2(s,w_3)}$$

其中 Z 是归一化因子。此后我们要做的就是训练参数λ和 Z.训练的方法极为复杂，最初的迭代算法是通用迭代算法(Generalized Iterative Scaling)，但是由于收敛速度过慢以及空间复杂度极高，慢慢的就被更为先进的改进迭代算法(Improved Iterative Scaling)淘汰掉了，具体内容可以参考文献[7][8].

总之，最大熵模型可以将多种已知信息整合到统一的模型中，同时具有较为良好的性质：形式上简单而优美，符合奥卡姆原理(Ockham's Razor)；性能上，模型可以满足多种约束条件，且具有平滑性(Smooth)，应用十分广泛（包括但不限于金融[11]、计算机图形学[12]、气象学[13]等领域）。但是缺点是计算代价过于高昂，如何降低最大熵模型中的计算复杂度如今仍然是一个充满活力的课题。

4 总结与分析

• 课题总结

数据科学(Data Science)包括数据科学主要以统计学、机器学习、数据可视

化以及某一领域知识为理论基础，其主要研究内容包括数据科学基础理论、数据预处理、数据计算和数据管理等。同时它也为自然科学和社会科学研究提供一种新方法，称为科学研究的数据方法[13]。如今数据科学在经营决策、市场分析、金融预测以及自然科学研究等领域提供了有力的支持，成为时下较为流行的科学工具。

本课题着重介绍了信息论中的重要概念——信息熵在数据科学中的应用，并且通过前人的探索经验和诸多应用实例来揭示熵的重要作用，同时为其他领域的科研工作者提供一种数据科学的思路，为信息熵和其他领域的交叉提供一些参考。

- **主要工作**

本文主要调研了信息熵在数据科学领域的一些应用案例，如信息压缩、决策分类、自然语言处理、风险管控等。本文首先引出信息熵的概念以及诞生背景，再通过文本压缩的案例来解释信息熵的含义，让读者对熵有着更直观的认识；其次介绍了熵在数据分析领域的应用，其中又分别介绍了相对熵、条件熵、交叉熵以及它们在机器翻译、信息检索、相关性分析等方面的应用；最后总结了 KL 散度、JS 散度以及最大熵模型在数据挖掘和机器学习中的应用，包括其在构建决策树、文本关键字检索等模型中的应用，并且指出最大熵模型的良好性质以及它在不同领域可能的应用。

- **展望**

尽管熵的概念很早就提出来了，可是它应用在信息学和数据科学领域却是最近 100 年的事情，而且关于熵的研究仍在继续。在此感谢为之努力的科学家们，比如香农，他提出的信息熵和香农三大定理奠定了现代通信技术的根基。尽管现实中的数据是非常巨大的，但是其中蕴含的信息熵是可知的。我们可以利用数据科学的方法，例如数据挖掘等技术，将其中真正有价值的信息挖掘出来，探索和发现其中的奥秘。在今天，信息熵仍然焕发着活力：基于信息熵理论的信息压缩技术在如今的大数据时代显得尤为重要，节约了大量的人力物力财力；交叉熵理论被应用在音频搜索[15]或者手写识别[14]中；最大熵模型被数据

[13] 引用自百度百科

工程师应用在空间特征分类中[16]等。随着大数据时代的到来以及不断被提出的各种新问题，熵在未来的数据科学中还有很大的发展空间。

参考文献

[1] Shannon C E. A mathematical theory of communication[J]. The Bell system technical journal, 1948, 27(3): 379-423.

[2] Gale W A, Church K W, Yarowsky D. A method for disambiguating word senses in a large corpus[J]. Computers and the Humanities, 1992, 26(5): 415-439

[3] B. Fuglede and F. Topsoe, "Jensen-Shannon divergence and Hilbert space embedding," *International Symposium onInformation Theory, 2004. ISIT 2004. Proceedings.*, 2004, pp. 31-, doi: 10.1109/ISIT.2004.1365067.

[4] Jones K S. A statistical interpretation of term specificity and its application in retrieval[J]. Journal of documentation, 1972.

[5] Quinlan J R. Induction of decision trees[J]. Machine learning, 1986, 1(1): 81-106.

[6] Csiszár I. I-divergence geometry of probability distributions and minimization problems[J]. The annals of probability, 1975: 146-158

[7] Csiszar I. A geometric interpretation of Darroch and Ratcliff's generalized iterative scaling[J]. The Annals of Statistics, 1989: 1409-1413.

[8] Della Pietra S, Della Pietra V, Lafferty J. Inducing features of random fields[J]. IEEE transactions on pattern analysis and machine intelligence, 1997, 19(4): 380-393.

[9] 周志华，机器学习，清华大学出版社，2016,1：73-95

[10] 吴军，数学之美，人民邮电出版社，2019,28: 179-185

[11] 张阚, 丰雪. 最大熵——均值方差保费原则[J]. 数学理论与应用, 2006, 026(002):32-34.

[12] 陈果, 左洪福. 图像分割的二维最大熵遗传算法[J]. 计算机辅助设计与图形学学报, 2002(06):530-534.

[13] 曹鸿兴, 罗乔林. 气象历史序列的最大熵谱分析[J]. 科学通报, 1979(08):351-355.

[14] 金忠, 胡钟山, 杨静宇,等. 手写体数字有效鉴别特征的抽取与识别[J]. 计算机研究与发展, 1999, 36(12):1484-1489.

[15] 欧智坚, 林晖. 基于交叉熵的音频指纹快速搜索方法:, CN101853262A[P]. 2010.

[16] 宋国杰, 唐世渭, 杨冬青,等. 基于最大熵原理的空间特征选择方法[J]. 软件学报, 2003(09):45-51.