分享到



214 29万 文章 阅读

查看TA的文章>

# 学了统计、算法,如何正确应用机器学习?

2016-07-27 18:37

来自: 清华大数据

学习了一段时间的机器学习算法后,再回头看机器学习问题,发现要想利用机器学习去很好 的求解一个问题,其实并不是一件容易办到的事情,尤其是能够对整个模型的解释方面,要 想能够对模型很好的解释,那么难度就会更大。因为利用机器学习处理一个实际的问题就不 仅仅是我们得学会怎么使用机器学习算法,更重要的是如何对整个问题建模。我刚开始是学 习智能计算,当然一个优化问题怎么去建模,建模完成之后就是求解,相对还算比较简单。 但是在机器学习中,问题就变得复杂多了,很多将机器学习的书也都是讲机器学习的算法, 就像我之前的"简单易学的机器学习算法"一样,注重算法的实现,但是机器学习问题中不 仅仅是机器学习算法,还有一些其他的知识需要我们去注意。

所以在这个标题下,我就我看的这篇文章与大家一起分享一下我自己的想法。

# 一、机器学习问题的组成

李航老师在《统计机器学习》中指出:机器学习=模型+策略+算法。其实机器学习可以表示 为: Learning= Representation+Evalution+Optimization。我们就可以将这样的表示和李 航老师的说法对应起来。机器学习主要是由三部分组成,即:表示(模型)、评价(策略)和优 化(算法)。

Representation	Evaluation	Optimization
Instances	Accuracy/Error rate	Combinatorial optimization
K-nearest neighbor	Precision and recall	Greedy search
Support vector machines	Squared error	Beam search
Hyperplanes	Likelihood	Branch-and-bound
Naive Bayes	Posterior probability	Continuous optimization
Logistic regression	Information gain	Unconstrained
Decision trees	K-L divergence	Gradient descent
Sets of rules	Cost/Utility	Conjugate gradient
Propositional rules	Margin	Quasi-Newton methods
Logic programs		Constrained
Neural networks		Linear programming
Graphical models		Quadratic programmin
Bayesian networks		
Conditional random fields		http://blog.csdn.net/google1989

如上图所示,机器学习主要是由上述三个部分组成。对应于每一个部分,都有每一个部分的 基本的方法,这里有的方法在我的以往的博客中都能找到介绍,还有的会在以后陆续的补 充。下面还是具体每一个部分来说说。

#### 1、表示(或者称为:模型): Representation

表示主要做的就是建模,故可以称为模型。模型要完成的主要工作是转换:将实际问题转化 成为计算机可以理解的问题,就是我们平时说的建模。类似于传统的计算机学科中的算法, 数据结构,如何将实际的问题转换成计算机可以表示的方式。这部分可以见"简单易学的机 器学习算法"。

具体对于一个分类问题而言,我们希望能够找到一个映射,这样的映射是将输入空间投影到

#### 24小时热文

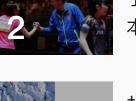


特朗

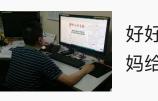
2 登





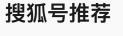




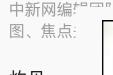












PLAYBO



灼见 讲新知识 点和有愉

澎湃新闻



澎湃新闻"是上海报业集团改革后公布 个成果。澎湃新闻是专注时政与思想



记录大国小民, 呈现人间百态



## 新京报

《新京报》具备鲜明的城市特色。强 意识,浓郁的北京风味。以"负责报道



# 24小时热文



共享单车现首家倒闭企业 扩张中还面临哪些坑?



王楠老公为刘国梁"叫屈" 18个副主席刘国梁负责点



庾澄庆二度当爸 前妻伊能 说了这番话



三星Note 8全力阻击苹果 S8: 你这样我很尴尬!



揭秘网吧通宵玩游戏最尴 件事,看完你一定身有感 输出空间中。而这样的映射的集合我们便把它称为假设空间(hypothesisspace)。我们其实是在这样的假设空间中寻找满足具体问题的一个映射。对于分类问题而言,映射主要分为:

概率模型:主要是条件概率

 $P(y \mid x)$ 

非概率模型: 主要是决策函数

y = f(x)

2、评价(或者称为: 策略): Evalution

评价的目标是判断已建好的模型的优劣。对于第一步中建好的模型,评价是一个指标,用于表示模型的优劣。这里就会是一些评价的指标以及一些评价函数的设计。

3、优化(或者称为: 算法): Optimization

优化的目标是评价的函数,我们是希望能够找到最好的模型,也就是说评价最高的模型。

## 4、总结

机器学习不仅仅是机器学习算法,从上面我们也看到机器学习算法只是属于机器学习的表示,机器学习是由三部分组成,包括表示(模型)、评价(策略)和优化(算法)。构建一个机器学习的应用,我们需要综合考虑这样的三个方面。以上的三个部分也并不是可以任意组合的,主要还是要看具体的问题,采取不同的方式。

# 二、什么是泛化能力

机器学习的目的是利用训练数据集训练出机器学习的模型,然后将该模型应用到测试数据集中。在整个过程中,我们对测试数据集是一无所知的,这是前提。所以该模型对未知数据的预测能力便称为泛化能力(Generalization ability),这是机器学习成功与否的一个很重要的评价指标。同时,预测误差便成为评价学习方法的泛化能力的指标。

机器学习算法中有很多的参数,算法成功往往源自对这些参数的精细调节。尽管我们能够获得的数据越来越多,但是始终我们不能把所有数据都得到,要像调节好这些参数,我们就必须充分利用训练数据。

## 1、cross-validation

交叉验证是充分利用数据最好的方式,其中交叉验证的基本想法就是要重复地使用数据。交叉验证的基本思想是:将数据集划分成训练数据和验证数据,在训练数据上训练整个模型,利用验证数据模拟实际的数据,对整个模型进行调整,最终我们选择在验证数据上表现最好的模型。常用的交叉验证的方法有:

简单交叉验证:与上面类似,简单的划分

S折交叉验证:将数据随机划分成S份,其中的S-1份作为训练,剩下的1份作为验证,重复进行,最终选择平均验证误差最小的模型







留一交叉验证: S折交叉验证的特殊情况

S = N

#### 2、总结

机器学习的最终任务是利用学习好的模型来预测实际的数据,这便是机器学习的泛化能力。 我们是希望能够在优化的过程中直接面对测试数据的优化函数,但是在实际的过程中,我们是无法得到这样的函数的,此时便使用训练过程中的优化函数来代替这个真实的函数。在最优化求解的过程中,有时求出的局部最优解也许比全局最优解更好。

#### 三、为什么仅有数据还不够

因为需要提高的是机器学习的泛化能力,所以在训练数据集上表现优越并不能代表该模型可以在测试集上表现良好,对于我们来说测试集永远是未知的,我们不知道将要面对的测试集是什么样的,对于一个具体的问题,我们也不能得到该问题下的所有的数据集,因为这样的集合根本也没有办法全部获得。所以仅有数据还是不够的。

#### 1、No Free Lunch(NFL,没有免费的午餐)

没有免费的午餐定理最早是在最优化理论中提出的,是由Wolpert和Macerday共同提出。 定理的结论是:由于对所有可能函数的相互补偿,最优化算法的性能是等价的。其含义是说 没有其他任何算法能够比搜索空间的线性列举或者纯随机搜索算法更优。

在机器学习问题中,没有学习器能够比在所有可能的函数中比随机猜测的结果更优。也就是说每个学习器都必须包含一些数据之外的知识或者假设,才能够将数据泛化。

#### 2、关于假设

有一些假设,如平滑(smoothness),相似的样例有相似的类别,有限依赖,或者有限复杂度。其实这些假设在我们的问题中也经常会用到。

#### 3、有关演绎(Deduction)和归纳(Induction)

演绎简单来讲就是由普通的前提推出特殊性结论的推理。而归纳则是通过对特例的分析来引出普遍结论的一种推理形式。由定义可以看出,演绎是由一般到特殊的过程,而归纳则是由特殊到一般的过程。归纳,也正是机器学习所要做的,从一些特殊的例子中得到一般性的规律,对应到机器学习中便是将少量的输入知识转化成为大量的输出知识。

# 四、过拟合问题(Over-fitting)

#### 1、过拟合的含义

过拟合是指学习时选择的模型包含的参数过多,以至于出现这一模型对已知数据预测很好,但是对未知数据预测得很差的现象。

## 2、对过拟合的理解

一种理解过拟合的方式是将泛化误差分解成偏置(bias)和方差(variance):

 $expectedLoss = (bias)^2 + variance + noise$ 

其中,noise为常数。偏置度量了学习器倾向于一直学习相同错误的程度;方差则度量了学习器倾向于忽略真实信号,学习随机事物的程序。其实,在我的理解中,这里很像演化计算中的那些启发式算法的更新公式,如何协调好这两个因素是算法成功的关键。

# 3、过拟合的例子

作者在文中举了一个例子:

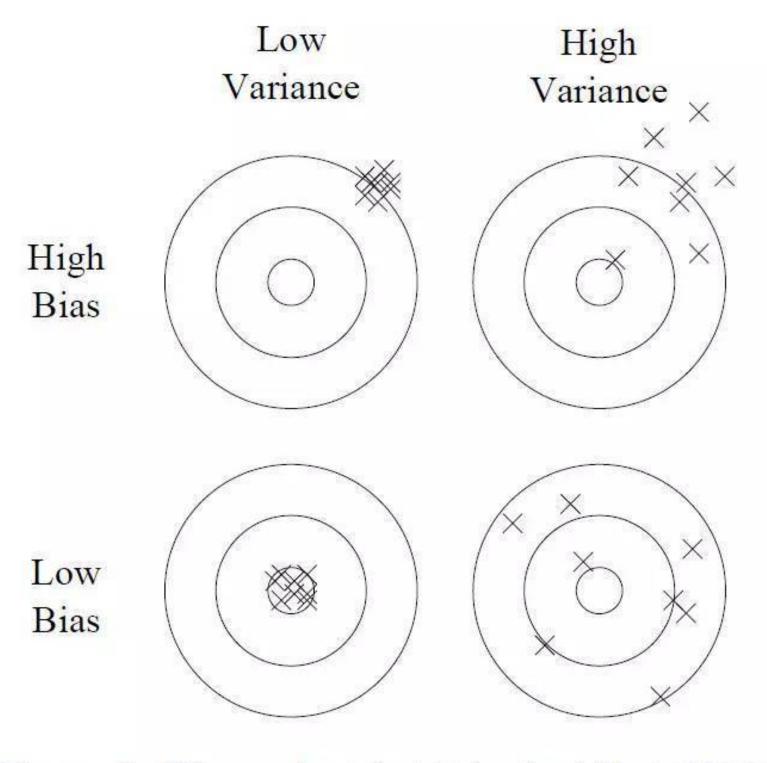


Figure 1: Bias and variance in dart-throwing.

一个线性学习器有高的偏置,因为加入需要分类的问题不是一个线性可分的问题,那么线性学习器是无法学习的;而决策树是有高的方差,因为在不同的训练集上,最终的学习结果是不一样的。

又一个例子说明过拟合的存在对模型的影响:

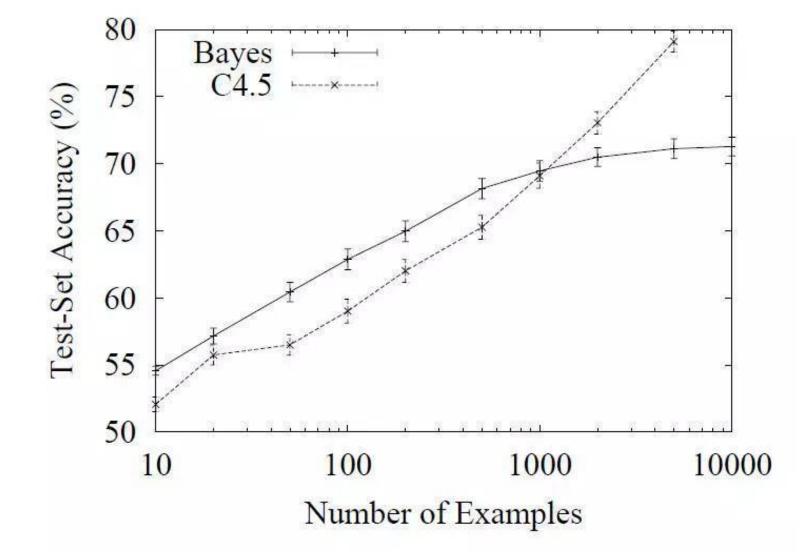


Figure 2: Naive Bayes can outperform a state-of-the-art rule learner (C4.5rules) even when the true classifier is a set of rules. http://blog.csdn.net/google19890102

如上图,在1000个样本以下时,Bayes的方法的准确性要比决策树高,尽管正确的模型是基于规则的,直到数据量超过了1000,决策树的准确性才能高于bayes,这也说明:一个强错误假设比那些弱正确假设更好,因为后者需要更多的数据才能避免过拟合。

#### 4、避免过拟合的方法

避免过拟合的方法主要有两种:

交叉验证(cross-validation)

给评价函数加上正则项

以上的两种方法在数据量比较小的时候还是比较有优势的,其中交叉验证有的时候也不能彻底解决问题,因为加入我们利用交叉验证做了太多的参数选择,那么本身这样的过程就开始有过拟合的可能。加上正则项的目的是惩罚那些包含更多结构的分类器,去选择较少结构的分类器,起到降低过拟合的可能。

## 5、过拟合与数据噪声的关系

有人会认为过拟合是因为数据中的噪声引起的,实际上过拟合并不仅仅是因为数据的噪声引起的,当然,在有噪声的数据中会加剧过拟合的可能性。

## 五、维数灾难(The curse of dimensionality)

#### 1、维数灾难的概念

维数灾难最初是由贝尔曼提出,主要描述的一个现象是:很多的算法在低维空间表现很好,但是在高维空间中就变得无法正确计算。在机器学习中,这样的问题变得更加普遍,随着样本维度(即特征数目)的增长,正确泛化的难度会以指数级增加。这里原因是同等规模的训练集只能覆盖越来越少的输入空间。

#### 2、典型的例子

以K-近邻这样的基于相似度的算法为例,基于相似度的推理在高维空间不再有效。分为两

种情况:

有2个与预测相关的特征,增加98个不相关的特征,这样,这98个特征就会淹没掉2个相关特征,最终导致K-NN是随机预测

这100个都是相关的特征,这时所有的样本表现出来的是都是相似的

3、带来的问题及其解决办法

有人认为在机器学习问题中,特征越多越好,其实这就会引起维数灾难的问题。"blessing of non-uniformity"在一定程度上抵消了维数灾难。

在大多数应用中,样本在空间中并非均匀分布,而是集中在一个低维流形上面或者附近。以手写体识别为例:每一个像素点为一个特征,这样为什么有效?

是因为数字图片的空间要远小于整个可能的空间,学习器可以隐式地充分利用这个有效的更低位空间,也可以显示地进行降维。

六、理论VS实践

1、机器学习中常见的两类理论

能保证泛化所需要样例数目的边界

渐进: 给定无穷数据, 学习器将保证输出正确的分类器

2、理论与实践的关系

机器学习中理论保证的主要作用并不是在实践中作为决策的标准,而是在算法设计中作为理解和驱动的来源。

3、总结

机器学习是理论+实践的过程,里面有很多的理论是边界的条件,在实际的应用中很少会真正碰到这样的约束问题,所以不能把这样的理论作为实践的指导,而是并驾齐驱的过程。

七、特征工程

决定一个机器学习项目成功与否的其中一个关键因素就是特征的选取。

1、特征的难点

应用机器学习的过程是相对较为简单的过程,然而,如果考虑到对数据的收集,整合,清洗和预处理,那么整个过程就显得不会那么简单,因为在整个数据的处理过程中,针对不同的数据,处理的方法是不尽相同的,这里面就会牵涉到很多的领域知识(domain-specific)。对这些领域知识的理解是能否处理好这些数据的关键。

2、如何处理特征工程问题

每一个特征独立地看也许与分类无关,但是组合起来也许就是相关的。如果所要处理的特征很多,那么寻找合适的特征将会成为一个费时费力的事,而且,还很容易导致模型的过拟合。在上面提及到的文章中有一些方法,可以应用到特征的选择中。

More data beats a cleaverer algorithm.

#### 1、提高机器学习准确性的两种方法

当特征已经如上面说的那样处理好,但是依然得不到好的结果,此时有两种方式改进:

设计一个更好的学习算法

获得更多的数据(包括更多的样例和不致造成维数在难度饿更多可能的原始特征)

## 2、更多的数据出现的问题

经验上讲: More data beats a cleaverer algorithm。但是,这样就出现一个问题,更多的数据我们可以学习到更复杂的分类器,但是在实践中,要学习到这样复杂的分类器需要更多的时间,所以在很多情况下,我们依旧选择简单的分类器。这时,就需要一些快速的学习算法。例如一些分布式的计算。

#### 3、机器学习的分类以及工作机制

其实无论简单的学习算法还是复杂的学习算法,他们的工作机制都是类似的: 所有的学习器都是将临界样例归类到同一类别中; 所不同的是每个分类器对这样的"临界"的定义是不一样的。如下图所示: 无论什么样的算法, 都能学出这样的分类边界, 只是在刻画分类边界的时候是复杂的, 还是不复杂的, 这样的情况就有很多。

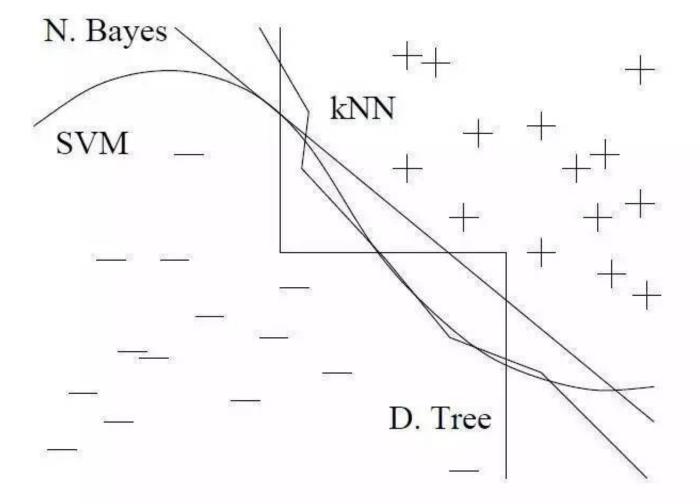


Figure 3: Very different frontiers can yield similar class predictions. (+ and – are training examples of two classes.)

http://blog.csdn.net/google19890102

分类器可以分为两类:

参数学习算法, 如线性分类器

非参数学习算法, 如决策树

首先,在参数学习算法中,其参数是固定的。但是这样的模型存在一个问题就是,当数据超

过一定的数量之后,这样的学习算法就不能再从中获益。

如上图中,朴素贝叶斯在最后的阶段,提升的效果已经很不明显。

第二种是非参数学习算法,如果有足够的数据,理论上是可以学习任何的函数,如上图中的C4.5算法。但是实际中是无法做到的,有这样的一些因素:

一些优化算法会陷入局部最优计算复杂性的限制维数灾难中讲,再多的数据也不会够。

# 4、总结

在选择机器学习算法的过程中,首先还是应当选择简单的学习算法。当然,想要提升学习的准确性,设计不同的学习算法也是相当有必要的,但是前提就是要知道基本的学习算法,从基本的学习算法中得到设计的灵感与方法。

九、多模型的融合。

最初,人们是在一个算法上花很大的功夫,研究各种变种,力图使一个算法变得更通用,但是结果却发现,将算法进行融合,会表现的更好,而且只需花费很少的精力。

1、三种集成模式

Bagging(最简单的一种)

Boosting

Stacking

第一种(Bagging)采用的方法是:通过重采样随机产生若干个不同的子训练集,然后在每个集合上训练一个分类器,最终用投票的方式将结果合并。

第二种(Boosting)采用的方法是:每个训练样例都有一个权重,并且权重会不断变化,每次训练新分类器的时候都集中在那些分类器之前倾向于分错的样例上。

第三种(Stacking)采用的方法是:每个单独的分类器的输出作为更高层分类器的输入,更高层分类器可以判断如何更好地合并这些来自低层的输出。

# 2、Bagging和Boosting方法的比较

Bagging采用重复取样,每个个体分类器所采用的训练样本都是从训练集中按等概率抽取的,因此Bagging的各子训练集能够很好的覆盖训练样本空间,从而有着良好的稳定性。

Boosting注重分类错误的样本,将个体子训练集分类错误的训练样本的权重提高,降低分类正确的样本权重,并依据修改后的样本权重来生成新的训练样本空间并用来训练下一个个体分类器。然而,由于Boosting算法可能会将噪声样本或分类边界样本的权重过分累积,因此Boosting很不稳定,但其在通常情况下,其泛化能力是比较理想的集成算法之一。

# 3、Netflix比赛

随着比赛的进行,很多优秀的算法都是融合了很多不同的算法才能得到如此高的准确性,这也说明了现在的趋势是越来越大型的集成。

4、模型集成(model ensemble)VS 贝叶斯模型平均(Bayesian model averaging, BMA)

模型集成改变了假设空间,而且可以采用多种多样的形式;BMA只是根据某个准则对原始空间的假设赋予不同的权重。

#### 5、总结

主要是在几种模型集成的算法上,为什么会成功,模型集成是一种趋势,其实也是"No Free Lunch"的一个表现。接下来,我也会写点模型集成的实验。

# 十、简单VS准确

简单并不意味着准确。在简单与准确之间并不能完全的画上等号。

# 1、奥卡姆剃刀(Occam's Razor)

奥卡姆剃刀原理是: Entities should not be multiplied unnecessarily(若无必要,勿增实体)。奥卡姆剃刀原理是在说明简单有效的原理,即不要浪费东西去做本身只需较少的东西就能做好的事情。

# 2、机器学习中错误的认识

在机器学习中,有人会认为,对于相同训练误差的两个分类器,比较简单的那个更可能有较低的测试误差。其实这个断言是没有根据的,在很多情况下也是存在问题的。理由主要有以下的三点。

首先是"No Free Lunch"理论。NFL理论中就表明,没有额外的信息是没有办法做到最好的。与上面的表述就是矛盾的。

其次是模型集成。集成模型的泛化误差会随着增加新的分类器而得到不断地改进,即使是在训练误差达到0的情况下。这就说明并不是简单就好。

最后是支持向量机。支持向量机会有很多的参数,这些参数不至于使得支持向量机过拟合。在这里,也就说明模型参数的数量和过拟合之间并没有直接的联系。

#### 3、结论

在机器学习中应当先选择简单假设,因为简单本身就是一个优点,并不是因为所假设的与准确率有什么关系。在很多情况下,我们都会选择简单的模型,吴军老师在《数学之美》上也谈到数学模型的问题时就指出:一个正确的数学模型应当在形式上是简单的。

# 十一、可表示VS可学习

# 1、可表示并不意味着可学习

当一个问题中仅有一个函数是可以被表示的,并不意味着这个函数是可以被学习的。有这样的三个例子:

第一个是标准的决策树。决策树就无法学习出比训练样本更多的叶子节点。

第二个是在连续空间中,我们可以想象一下泰勒展开,对于一个正弦函数,我们要用泰勒展

开也是无穷多项。

第三个是存在局部最优的情况,我们可能就无法找到真正的函数。

所以,对于给定有限数据,时间和内存的情况下,标准学习器智能学到所有可能函数中很有 限的子集,这个子集也会随着表示方法的不同而不同。

#### 2、可表示与可学习之间的关系

对于某些函数来讲,一些表示方法会比其他方法更加精简,从而只需要更少的数据来学习那 些函数。举例来讲,对与一个非线性问题,如果直接处理,可以采用核方法,但是采用多层 神经网络,那就会变成一个线性的问题

## 3、总结

可表示并不意味着可学习,但是表示的恰到好处会大大降低学习的难度与复杂性。

## 十二、相关VS因果

相关性是因果关系的标志,但是相关并不意味着因果。机器学习也只能学习到相关关系。

了解更多,请点击"阅读原文"

# 投稿、合作请发送邮件至contact@dataunion.org

声明:本文由入驻搜狐公众平台的作者撰写,除搜狐官方账号外,观点仅代表作者本人,不代表搜狐立场。

阅读 (25)

※ 不感兴趣 ① 投诉



# 我来说两句

O人参与, O条评论

来说两句吧……

登录并发表

搜狐"我来说两句"用户公约



还没有评论, 快来抢沙发吧!

# 推荐阅读

女子深夜喝酒独行 竟被男司机招呼之后强拽进车

◆ 东方网 · NaN-NaN-NaN

湖北鄂州交通事故致18死 车辆非法改装严重超载(图)





秒杀! 头春金骏眉198/桶

① 广告 ・ 今天 7:14

# 女博士误机打人被拘留,如今被法航永久"拉黑"意味着什么?

SOHU.com

 $\bigcirc$ 

搜狐 SOHU.com 中国打造"超级渔场"交付挪威:一次可养鱼150万条(图)

上游新闻 · NaN-NaN-NaN

 $\bigcirc$ 

搜狐 SOHU.com 港交所董事总经理:要加强海外投资者参与A股上市公司治理 (图)

中国新闻网 · NaN-NaN-NaN

 $\bigcirc$ 

# 石家庄火车站地下公交站台吊顶坍塌:施工人员踩踏所致(组图)

SOHU.com

澎湃新闻网 ・ 昨天 23:33

 $\bigcirc$ 

# 老公,这裤子降价了!秒杀199五件!舒适,透气!值爆了!



(1) 广告 ・ 今天 7:14

# 鹿晗新剧女主疑是关晓彤, 网友却嫌弃, 纷纷喊热巴

🏥 随煮娱 ・今天 00:49

 $\bigcirc$ 

张继科1: 4惨败, 刘国梁急脱衣服没找出败因, 网友却一语 道破?

¼ 篮球幸痒mp · NaN-NaN-NaN

搜狐

迪丽热巴过生日跑男团送祝福 网友: 鹿晗还没送

搜狐娱乐 · NaN-NaN-NaN

 $\bigcirc$ 

搜狐

350余名重庆大学生深入田间地头 让环保种子在农村开花结 果(组图)

 $\bigcirc$ 



男子持刀劫持女主播: 曾刷礼物1万多 现被要求断绝交往



对于哈林二度当爸,前妻伊能静是这么说的...

澎湃新闻网 ・今天 11:11



体育 取消总教练主教练 中国乒乓球队改革推出大动作

人民网 ・ 今天 12:43  娱乐| baby曾公开表决心要戒烟 可如今又开始抽了 SOHU.com 明星粉丝团 ・ 今天 12:14 巴黎香街恐袭案:袭击者合法拥枪 家藏"军火库" 搜狐 SOHU.com 
 中国新闻网 ・ 今天 12:09
 老司机必备,驾驶专用蓝牙偏光镜!偏光减速,蓝牙通话,安全驾驶!2折秒杀



(1) 广告 ・ 今天 23:13

SOHU.com

朝鲜官员: 若美国停止军演 平壤准备暂停核试和试射导弹

参考消息 ・ 今天 11:14

男子持刀劫持女主播:曾刷礼物1万多 现被要求断绝交往

澎湃新闻网 ・今天 11:11

女子公路上坐车顶"自由飞翔"看呆众人 交警部门介入

● 野球网 · 今天 11:07

SOHU.com

股价为何闪崩? 万达复星晚间齐发公告称公司状况良好(组 图)

3条休闲裤+皮带仅199元! 特价包邮货到付款



(1) 广告 ・ 今天 23:13

# 莫雷真狠心,火箭死亡五投解体!抛售3200万组合狂追保罗、格里芬!



# 故宫首次为农民何刚办追思会, 当年捐赠文物却要借钱去北京(组图)

澎湃新闻网 ・ 今天 21:42  $\bigcirc$ 

伊朗与卡塔尔是如何看待沙特更换王储的?





# 紧急 今天教你1个狐臭的绝招 不看后悔「5分钟后删」







气氛诡异, 国乒众将"声援"刘国梁, 张继科退赛

# 对于哈林二度当爸,前妻伊能静是这么说的...





