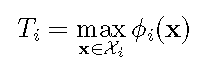
Page10

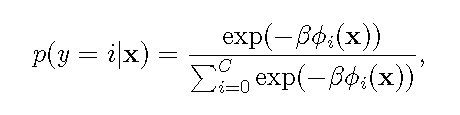
把这些samples放入一个outlier bin 中，而不是像其他已有的分类方法那样，把它们分类为最相似的class。bin中的Samples之后可能会与bin中的其他一些samples分为一组并代表一个new object的出现，或者也可能只是简单地保存在outlier bin中来保证它们不会破坏已经存在的object的模型。我们对每一个class i，基于训练集中样本间可能性分歧的最大值，建立一个该类的可能性分歧的阈值 。

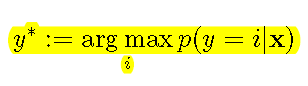
系统对outlier bin中的samples定义y=0。

B Model Trees for Active Learning

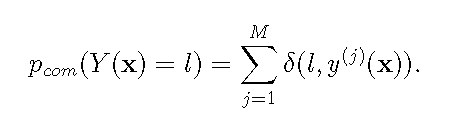
使用已知分类的samples建立models。分类器从自身的决定来学习，如果分类器决定了一个错误的label，就会面临参数可能会变得非常少的问题，性能也会下降。当有操作员与系统交互时，利用人类反馈来保证分类器的自我学习是可靠的。

因此为系统设计了Query-by-Commitee框架，这样每次发出一个query后，只需要简单地移除committee中的成员，系统中并不存在再次训练、再次分类、以及再次部署主动学习的算法的问题。这个框架的难点在于通过对训练集数据特征的学习，生成不同的models的committee。如果committee中现有的可能性的范围没有捕获到每一个sample的不确定性，那么这些不同的模型就不能很好的作为query的基础。所以提出融合了uncertainty Sampling和Query-by-Committee的，通过动态的生成committee来反映samples的不确定性的方法。针对主动学习的在线分类问题调整这个框架，并且改进在主动学习算法中的一个未解决的问题，即何时发出query的问题。

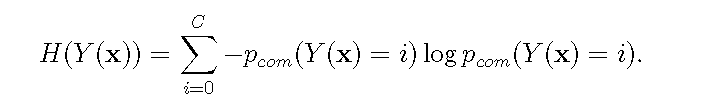
系统中的主动学习算法——model trees，由通过对训练集中已知objects计算建立的每一个class 的可能性 来进行初始化。对一个进入系统中的new sample x，按照算法1来计算特征，转变为Gibbs分布，

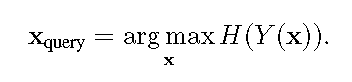
当x对应y\*不为0即不放入outlier bin时，系统使用x来更新其所属的class的可能性参数。然而，如果p(.|x)中最大值和次大值非常相近，系统会考虑不同的选择y\*和y alt，用x同时更新committee中不同的假设和。

通过继续这样的工作，建立一个不同models的tree，其中每一条路径都代表不同的假设。当系统的分类决定带来大量的不确定性时，tree的分枝数目也会大量增加。这样就会指示系统分类器应该通过操作员的反馈来提升性能。这样通过对tree中的分枝数目设置一个限制，系统能够决定何时发出query，这和models间的不确定性的数目相关。选择query的sample要能够最大化的减少相互矛盾的假设的数量。假设间分歧的数量可以通过赋予该sample的labels的entropy来量化。

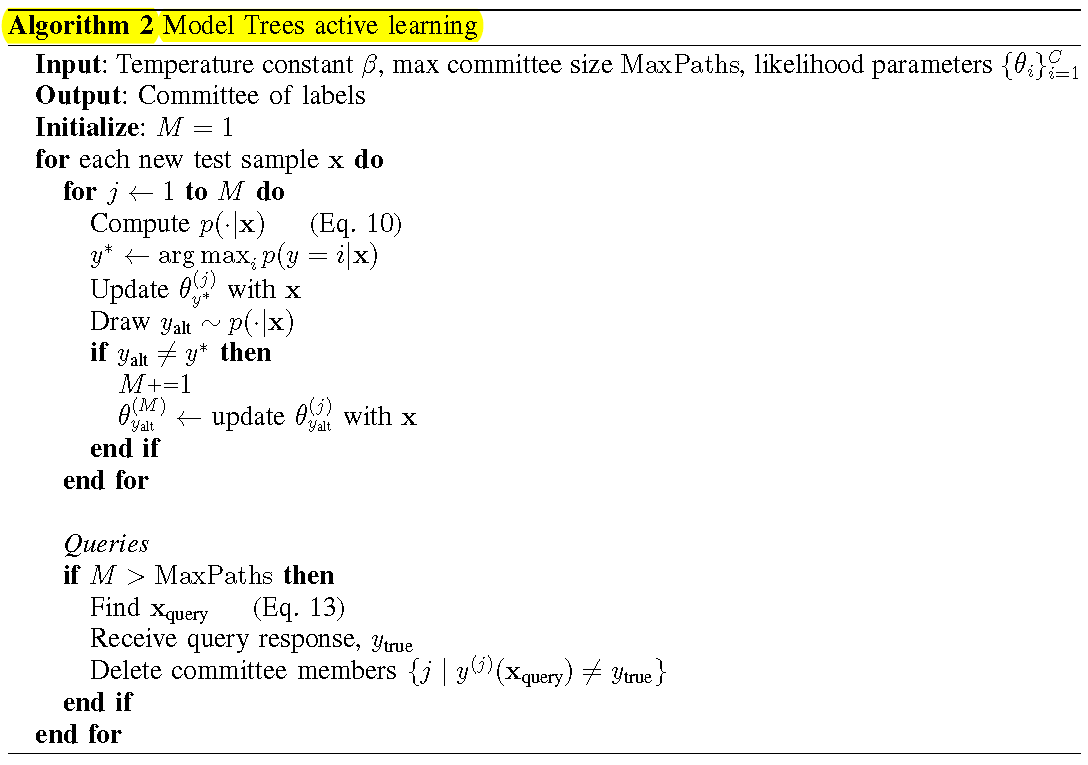
令committee中第个成员赋予x的label为。系统可以建立一个committee投票的直方图，来计算每个sample的entropy。一个大小为M的committee的label分布为for a label outcome l and where is the Kronecker delta.。

Then the Shannon entropy for the label of x is



选择要query的sample是最大label entropy的sample.

系统将要query的sample发送给操作员，人为判断其真实label并发送给系统。系统移除掉model tree中所有对该sample赋予了错误label的分枝，提高模型分类的准确性。如果所有的假设具有相等的可能性，那么最大label entropy query决定能够最大化的减少tree中的分枝。



C.new object的发现

committee中的每一个假设 放置samples到它们自己的outlier bin中。从这些bin中，通过cluster可以识别出组成new object的samples，并对该object建立模型。New objects的数目是未知的吗，但至少从已知的classes中我们有一些估计关于现在一组samples中的变化。由于clustering算法需要一个参数来决定识别clusters的规模，我们可以按照已知classes来调整这个参数。这样能提供一个更有针对性的平衡关于正确和错误的发现。

首先选择一个clustering算法。算法中clusters的数目必须不需要确定，因为outlier bin中可能存在的objects的数目是未知的。通过应用clustering算法给已知类别的训练集数据，系统能够找到最适合已知labels的参数集。Outlier bin在需要的适合执行clustering。Committee中存储的labels随之增加。

