贝叶斯信念网络的学习与应用

姓名：王丹

学号：2120151036

班级：2015 级硕 1

一、实验目的

1、剖析贝叶斯信念网络，理解其原理。
2、熟悉 Weka 软件。
3、了解贝叶斯信念网算法编程。

二、实验环境

Weka 3.6.13，IRIS 数据集。

Weka 的全名是怀卡托智能分析环境(Waikato Environment for Knowledge Analysis)，是一款免费的，非商业化的，基于 JAVA 环境下开源的机器学习以及数据挖掘软件。

Iris 也称鸢尾花卉数据集，是一类多重变量分析的数据集。通过花萼长度，花萼宽度，花瓣长度，花瓣宽度 4 个属性预测鸢尾花卉属于（Setosa，Versicolour，Virginica）三个种类中的哪一类。

三、实验内容

用贝叶斯信念网分析样本错分情况。

四、实验原理

1、监督学习

理想情况下，一个 classifier 会从它得到的训练集中进行“学习”，从而具备对未知数据进行分类的能力，这种提供训练数据的过程通常叫做 supervised learning（监督学习）。监督学习的定义是：当样例是输入/输出对给出时，成为监督学习；有关输入/输出函数关系的样例称为训练数据

常见的监督学习方法包括：记忆学习、决策树学习、贝叶斯学习、支持向量机、BP 算法等。

2、贝叶斯学习

从当前观察到的不同类输出数据中归纳出各类别数据所服从的统计规律，进而根据统计学的相应原理，利用这些统计规律对 P(h|d)的绝对大小或相对大小进行推算，这种学习方式主要构建在贝叶斯法则（Bayesian Rule，Bayesian Theorem）上，相应统计学习方法被称为贝叶斯学习。

3、贝叶斯信念网（Bayesian Belief Network，BBN）

（1）函数形式

贝叶斯信念网的函数形式属于隐式表达形式。隐式表达形式是用某种形式隐含地表达出离散或连续函数形式，即采用一种非直观的形式来给出离散或连续函数，通常是以图结构的形式来表达。

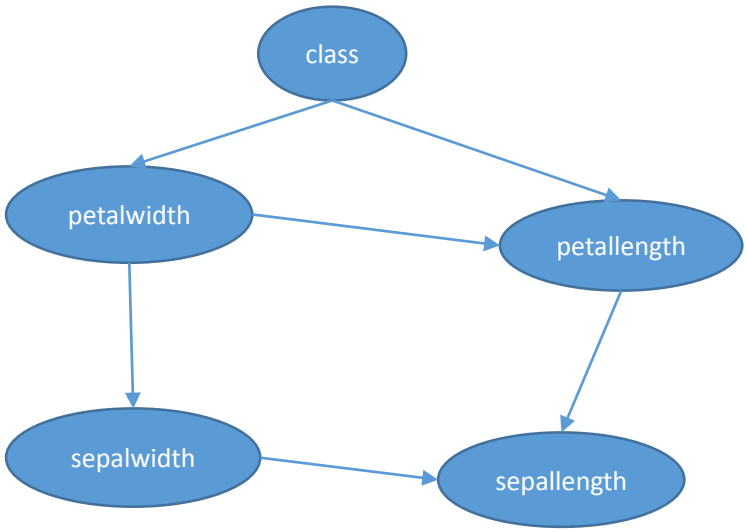
一个贝叶斯信念网络定义包括一个有向无环图（DAG）和一个条件概率表集合。DAG 中每一个节点表示一个随机变量，可以是可直接观测变量或隐藏变量，而有向边表示随机变量间的条件依赖；条件概率表中的每一个元素对应 DAG 中

唯一的节点，存储此节点对于其所有直接前驱节点的联合条件概率。

贝叶斯网络的有向无环图中的节点表示随机变量{X1,X2,X3,…,Xn}，它们可以是可观察到的变量，或隐变量、未知参数等。认为有因果关系（或非条件独立）的变量或命题则用命题或箭头来连接。若两个节点间以一个单箭头连接在一起，表示其中一个节点是“因（parents）”，另一个是“果（children）”，两节点就会产生一个条件概率值。总之，连接两个节点的箭头代表此两个随机变量是具有因果关系，或非条件独立。

贝叶斯信念网络有一条极为重要的性质，就是我们断言每一个节点在其直接前驱节点的值制定后，这个节点条件独立于其所有非直接前驱前辈节点。

下图就是一个贝叶斯信念网的图形式：



（2）优化准则

1）确定优化目标

优化目标设为极大似然估计：

设 D 表示由所有训练数据构成的集合，d 表示其中的任意一个数据，$P_w(D)$ 表示所有训练数据在某概率分布下同时存在的概率，则有

$$P_w(D)=\prod_{d\in D}P_w(d).$$

这里 w 表示由概率分布中所有未知参数构成的向量。这样，我们需要获得最优的 w（设为 w*），使得$P_w(D)$最大化，即

$$w^*=\arg\max_w P_w(D)=\arg\max_w \prod_{d\in D}P_w(d).$$

对上公式取对数，可将计算目标转化为等价但更容易计算的形式：

$$w^*=\arg\max_w \ln P_w(D)=\arg\max_w \sum_{d\in D}\ln P_w(d).$$

在贝叶斯信念网络中，带求解的未知参数 w 是由贝叶斯信念网中所有条件概率值构成的向量。

2）设计优化算法

采用基于函数梯度的最速优化方法。该方法按照函数梯度方向迭代更新函数参数值，指导函数值稳定或超过迭代次数。

此处应采用梯度上升方法求解。

当确定了一个 w 时，就确定了一个贝叶斯信念网，也就确定了相应的$P_w(D)$以及 $\ln P_w(D)$。设$\frac{\partial ln P_w(D)}{\partial w}$表示 $\ln P_w(D)$关于 w 的梯度，则利用梯度上升发求$\ln P_w(D)$的最大值就是不断用以下公式迭代修改 w：

$$w = w + \eta \frac{\partial \ln P_w(D)}{\partial w}$$

直到停止条件满足。停止条件可以是函数值已趋于稳定或者达到最大迭代次数等。η是步长。

为了求解最优 w，需要给出 $\ln P_w(D)$关于 w 中任意一个条件概率变量的偏导数。令w_{ijk}表示 w 中任意一个条件概率变量，即当其所有父节点给定一个确定的值时，节点 i 具有某个确定值的概率。$\ln P_w(D)$关于w_{ijk}的偏导数为：

$$\frac{\partial \ln P_w(D)}{\partial w_{ijk}} = \sum_{d \in D} \frac{P_w(x_{ij}, u_{ik} | d)}{w_{ijk}}.$$

w_{ijk}的迭代更新公式：

$$w_{ijk} = w_{ijk} + \eta \sum_{d \in D} \frac{P_w(x_{ij}, u_{ij} | d)}{w_{ijk}}.$$

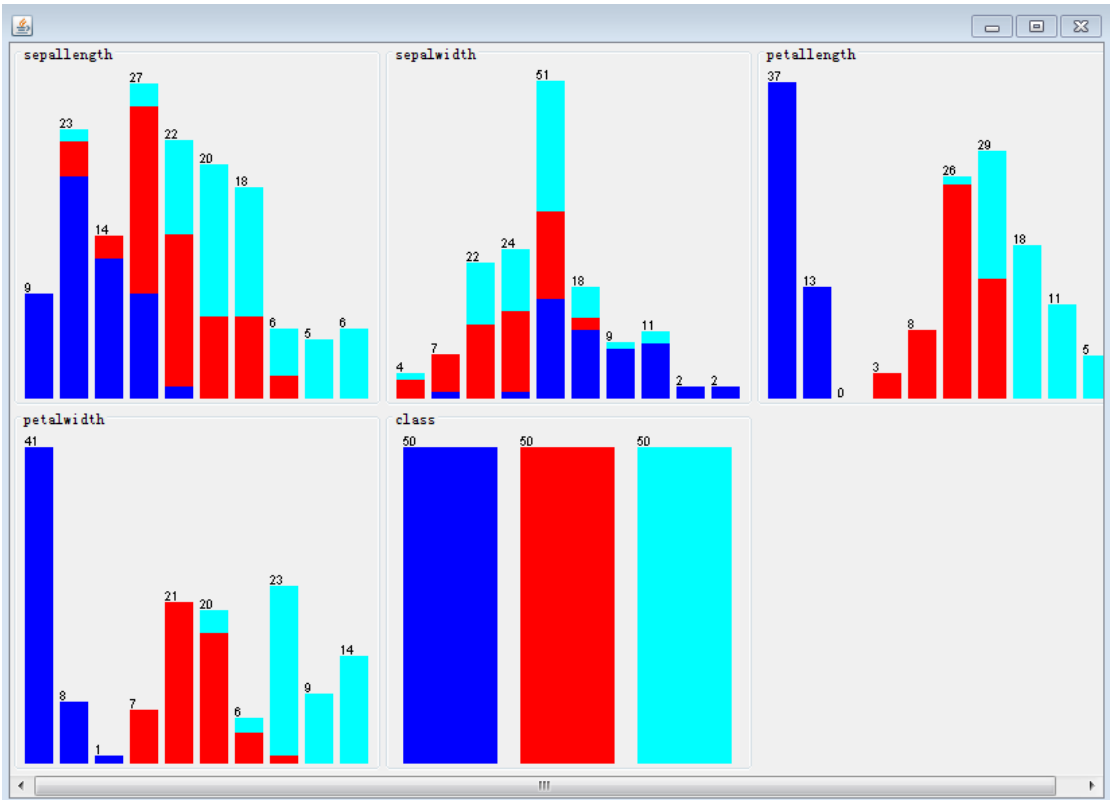
在该公式的基础上，便可根据输入数据，迭代更新贝叶斯信念网中各节点的条件概率表，指导各节点条件概率趋于稳定或达到最大迭代次数为止。

五、实验步骤

1、预处理：

（1）选择 Iris 数据集

（2）选择 Discretize Filter，得到如下结果：



2、分类：

（1）选择 weka/classifiers/bayes/BayesNet;

（2）设置 Percentage Split 为 80%；

六、实验结果及分析

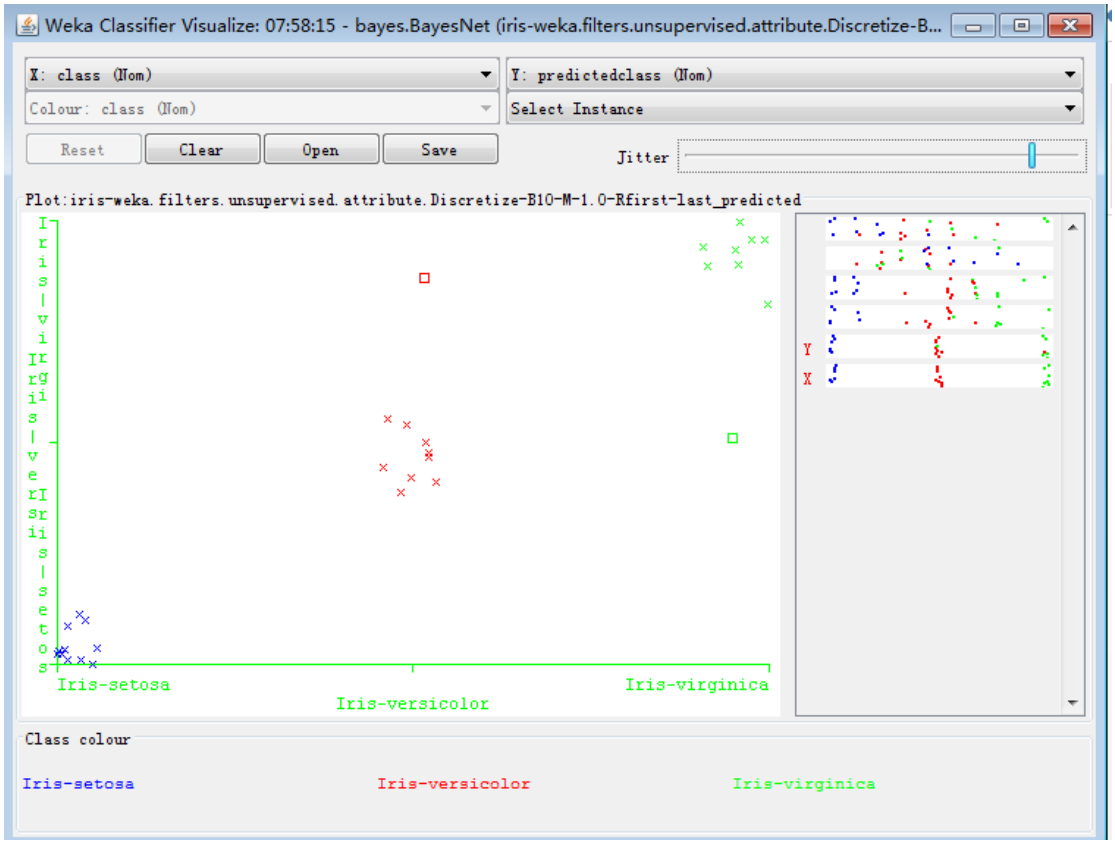
实验结果如下：

```
Time taken to build model: 0 seconds

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances          28               93.3333 %
Incorrectly Classified Instances         2                6.6667 %
Kappa statistic                          0.8997
Mean absolute error                      0.064
Root mean squared error                  0.1951
Relative absolute error                 14.3942 %
Root relative squared error             41.347  %
Total Number of Instances               30

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0          1         1         1          1        Iris-setosa
                 0.9       0.05       0.9       0.9       0.9        0.98     Iris-versicolor
                 0.889     0.048      0.889     0.889     0.889      0.979    Iris-virginica
Weighted Avg.    0.933     0.031      0.933     0.933     0.933      0.987

=== Confusion Matrix ===

  a  b  c   <-- classified as
 11  0  0 |  a = Iris-setosa
  0  9  1 |  b = Iris-versicolor
  0  1  8 |  c = Iris-virginica
```

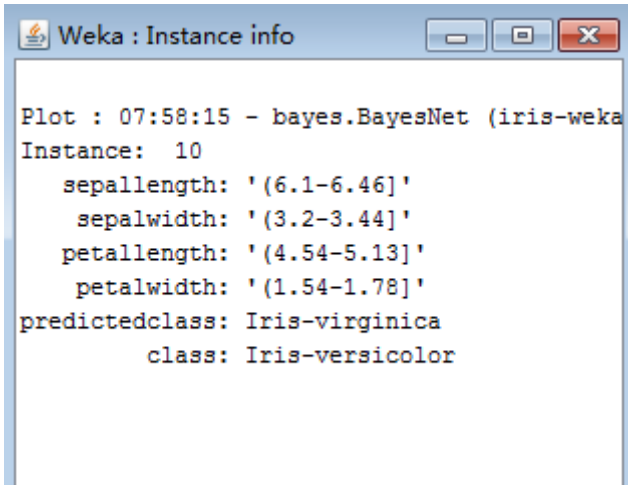
```
=== Run information ===

Scheme:weka.classifiers.bayes.BayesNet -D -Q weka.classifiers.bayes.net.search.local.K2 -- -P 1 -S BAYES -E w
Relation:     iris-weka.filters.unsupervised.attribute.Discretize-B10-M-1.0-Rfirst-last
Instances:    150
Attributes:   5
              sepallength
              sepalwidth
              petallength
              petalwidth
              class
Test mode:split 80.0% train, remainder test

=== Classifier model (full training set) ===

Bayes Network Classifier
not using ADTree
#attributes=5 #classindex=4
Network structure (nodes followed by parents)
sepallength(10): class
sepalwidth(10): class
petallength(10): class
petalwidth(10): class
class(3):
LogScore Bayes: -1137.5169590395597
LogScore BDeu: -1502.2058926390077
LogScore MDL: -1469.2089723239778
LogScore ENTROPY: -1193.624031148684
LogScore AIC: -1303.624031148684
```

从图中可以看出，在将数据集的 80%作为训练集，20%作为测试集的情况下，30 个测试数据中有 28 个被正确分类，占 93.3333%，2 个测试数据被错误分类，占 6.6667%。

分类情况具体如下：



图中方框表示错误分类的样本，以红色方框为例：



该数据实际属于 Iris-versicolor，但被错分到 Iris-virginica。

七、实验小结

通过这次实验，我详细学习了贝叶斯网络相关知识，理解了其原理，并对其应用有了一些了解。其次，还熟悉了 weka 这一工具。这次实验我只是在 weka 上跑了一下数据，希望以后能够自己写程序调用 weka 接口。