Data Mining Assignment 2
Cost sensitive classifiers

08305006: Prashanth K
08305023: Prashant Borole
08305028: Sriram Kashyap
08305045: Anup Kulkarni

October 10, 2008

1 Introduction

1.1 Basic Info

- Classifiers Implemented:
  - Cost sensitive Decision Tree
  - Cost sensitive Naive Bayes Classifier
- Files Created/Modified:
  - weka/classifiers/trees/CostTree.java
  - weka/classifiers/bayes/CostSensitiveNaiveBayes.java
  - weka/gui/GenericObjectEditor.props
  - costs.cfg

1.2 Instructions

Instructions to run:

- Run weka.jar from the command-line: java -jar weka.jar
- Edit the costs.cfg file so that the first line is the file path of the attribute cost file and the second line is the file path of the cost matrix file (The remaining lines are ignored).

- Load the dataset, and run one of these classifiers:
  - trees.CostTree
  - bayes.CostSensitiveNaiveBayes

Note: The classifiers work for data sets with numeric/nominal/missing attributes, and nominal classes (numeric classes are currently not supported).

2 Cost Sensitive Naive Bayes Classifier

2.1 Formulation

Fiven (D, C, T), where:

- D is a training dataset consisting of N samples $(x_1, x_2, ..., x_N)$ from P classes $(c_1, c_2, ..., c_P)$ Each sample x_i is described by M attributes $(A_1, x_2, ..., A_M)$ among whom there can be missing values.
- C is a misclassifcation cost matrix. Each entry $C_{ij} = C(i, j)$ specifies the cost of classifying a sample from class c_i as belonging to class $c_j (1 \leq i, j \leq P)$. Usually, $C_{ii} = 0$.
- T is a test-cost vector. Each entry $T_k = T(k)$ specifies the cost of taking a test on attribute $A_k (1 \leq k \leq M)$;

Build a test-cost sensitive naive Bayes classi?er *csNB* and for every test case, a test strategy with the aim to minimize the sum of the misclassification cost C_{mc} and test cost C_{test}.

2.1.1 Strategy

A sequential test strategy is as follows. During the process of classification, based on the results of previous tests, decisions are made sequentially on whether a further test on an unknown attribute should be performed, and if so, which attribute to select.

Suppose that $x = (a_1, a_2, ..., a_M)$ is a test example. Each attribute a_i can be either known or unknown. Let $\tilde{A}$ denote the set of known attributes among all the attributes A and $\bar{A}$ the unknown attributes. The expected misclassification cost of classifying x as class c_j based on $\tilde{A}$ is:

$$R(c_j|x) = R(c_j|\tilde{A}) = \sum C_{ij} \times P(c_i|\tilde{A}, 1 \leq j \leq P \tag{1}$$

c_{j*} with the minimum expected cost is predicted as the class label.

To decide which attribute $\tilde{A} \in \bar{A}$ to select, we define the *utility* of testing an unknown attribute $\bar{A}_i$ as follows:

$$Util(\bar{A}_i) = Gain(\tilde{A}, \bar{A}_i) - T_i \tag{2}$$

$$Gain(\tilde{A}, \bar{A}_i) = C_{mc}(\tilde{A}) - C_{mc}(\tilde{A} \cup \bar{A}_i) \tag{3}$$

$$C_{mc}(\tilde{A}) = min_j R(c_j|\tilde{A}) \tag{4}$$

$$C_{mc}(\tilde{A} \cup \bar{A}_i) = \sum_{k=1}^{|\bar{A}_i|} P(\bar{A}_i = x_k|\tilde{A}) \times min_j R(c_j|\tilde{A}, = A_i \qquad = x_k) \tag{5}$$

Overall, an attribute $\bar{A}_i$ is worth testing if testing it offers more gain than the cost it brings. We select the attribute A_i with the maximum utility, read its value, update $\tilde{A}$ and $\bar{A}$ with A_i, and continue this process until none of the remaining attributes are worth testing.

3 CostTree

The cost sensitive decision tree is based on a modification of the algorithm suggested by Ling et al. in "Decision Trees with Minimal Costs". It has been built using weka's REP-Tree as a base, and modifiying the gain functions that it uses.

3.1 Formulation

The decision tree is made cost-sensitive by selecting those attributes that have highest gain, at each stage of the tree building process. The gain is defined as:

$$Gain = priorCost - cCost - attribCost \times N \tag{6}$$

Here, *priorCost* is the cost of misclassification before the split, and *cCost* is the cost of misclassification after the split. *attribCost* is the cost of evaluating the attribute over which the split is taking place. This cost is multiplied by the number of instances (N) for which this attribute should be evaluated.

$$currentCost = \sum_{i=0}^{n} \sum_{j=0}^{d} (N * dist_j) * C_{jk} \tag{7}$$

where: n is the number of values that the attribute can take,
N is the number of instances,
d is the number of attributes,
$dist_j$ is the probability of class value j,
C_{jk} is the cost of misclassifying an instance of class j as that of class k, where k is the dominating class of the split.

Given a distribution for c classes, the dominating class i for that node is calculated as follows:

$$\arg\min_i \quad cost = \sum_{j=0}^{c} dist_j * C_{ji} \tag{8}$$

It has been shown that if a tree is built using these rules, the best attributes to evaluate are at the top of the tree, and the most expensive attributes move to the bottom of the tree. This means that evaluating the class of a new instance using this tree is just the matter of traversing the tree, till the leaf node is reached. Any attributes which are un-economical to expand are not included in the tree building process.

The class of an instance is calculated from the distributions by weighing each class distribution by the misclassification costs of that class, as shown above.

4 Results

Pima Indians

	CS Naive Bayes			SimpleNaiveBayes		
cs dataset 1						
	a	b	0.68	a	b	0.42
a	151	349		a	422	78
b	9	259		b	104	164
cs dataset 2						
	a	b	0.59	Confusion Matrix		
a	500	0		RMS Error		
b	268	0		cost matrix		
cs dataset 3						
	a	b	0.59			
a	486	14				
b	249	19				
cs dataset 4						
	a	b	0.81			
a	0	500				
b	0	268				
cs dataset						
	a	b	0.51			
a	426	74				
b	126	142				

Cancer dataset

	CS Naive Bayes			SimpleNaiveBayes		
cs dataset 1						
	a	b	0.19	a	b	0.21
a	330	14		a	324	20
b	5	174		b	4	175
cs dataset 2						
	a	b	0.59	Confusion Matrix		
a	344	0		RMS Error		
b	179	0		cost matrix		
cs dataset 3						
	a	b	0.26			
a	332	12				
b	22	157				
cs dataset 4						
	a	b	0.81			
a	0	344				
b	0	179				
cs dataset						
	a	b	0.2			
a	331	13				
b	7	172				

Pima Indians

	CostTree			REPTree	
cs dataset 1					
	a	b	0.65	a	b	0.43
a	0	10	a	0	500	a	422	78
b	70	0	b	0	268	b	113	155
cs dataset 2					
	a	b	0.48	Confusion Matrix
a	0	1	a	500	0	RMS Error
b	1	0	b	268	0	cost matrix
cs dataset 3				
	a	b	0.56	
a	0	70	a	500	0	
b	10	0	b	268	0	
cs dataset 4				
	a	b	0.81	
a	0	0	a	0	500	
b	70	0	b	0	268	
cs dataset				
	a	b	0.47	
a	0	70	a	441	59	
b	70	0	b	197	71	

Cancer dataset

	Cost Tree			REP Tree	
cs dataset 1					
	a	b	0.28	a	b	0.22
a	0	100	a	315	29	a	332	12
b	400	0	b	4	175	b	19	160
cs dataset 2					
	a	b	0.47	Confusion Matrix
a	0	1	a	344	0	RMS Error
b	1	0	b	179	0	cost matrix
cs dataset 3				
	a	b	0.34	
a	0	400	a	338	6	
b	100	0	b	57	122	
cs dataset 4				
	a	b	0.81	
a	0	0	a	0	344	
b	300	0	b	0	179	
cs dataset				
	a	b	0.25	
a	0	400	a	327	17	
b	400	0	b	19	160	