Chapter 7

Evaluation of Dataset Selection for One-Versus-All Aggregate Modeling

It was stated in chapter 6 that the proposed methods of training dataset selection were aimed at supporting the creation of aggregate models for multi-class prediction tasks. For such models, the two proposed methods for creating the base models are One-Versus-All (OVA) and positive-Versus-negative (pVn) classification. The experiments to study OVA base model design, training dataset selection for OVA base models and the performance of OVA base models and aggregate models are presented in this chapter. Questions 1 and 2 below were posed in chapter 6. The studies reported in this chapter are aimed at providing answers to these questions, in the context of OVA modeling.

1. How should training datasets be designed in order to create base models that are syntactically diverse and highly expert at prediction for aggregate models?

2. How should training datasets for the base models be designed in order to achieve high accuracy for the aggregate model?

This chapter is organised as follows: Section 7.1 provides a discussion of OVA modeling. Experiments to study 5NN OVA model performance and See5 OVA model performance are respectively discussed in sections 7.2 and 7.3. Sections 7.4 and 7.5 respectively provide a discussion and conclusions for the chapter.

7.1 OVA modeling

This section provides the motivation for OVA modeling. The methods for creating OVA aggregate models are also presented. The motivation for OVA modeling is discussed in section 7.1.1. The design of training samples for OVA base models is presented in section 7.1.2. The experimental procedure for this chapter is presented in section 7.1.3.

7.1.1 Motivation for OVA modeling

It was stated in chapter 6 that OVA classification was selected as one of the problem decomposition methods to be studied, for several reasons. Firstly, by definition OVA classification enables the creation of base models where each base model is an expert on classification for one specific class. Secondly, since each OVA classifier solves a 2-class problem, the training sample size required to achieve a high level of accuracy is reduced. This is an implication of the Probably Approximately Correct (PAC) learning theory as discussed in section 6.1 of chapter 6.

A third reason for selecting OVA modeling is as follows: It was stated in section 2.8 that increasing the amount of training data for the modeling process results in the reduction of the variance component of the prediction error. However, an excessively large training dataset results in overfitting and modeling of phantom structure. If several moderately sized training datasets are used for the modeling process, then the amount of training data is increased while at the same time overfitting is avoided. The use of OVA base models enables the above approach.

7.1.2 Sample composition for OVA base model training datasets

The methods for training sample selection for base models was discussed in section 6.3.3 and illustrated in figure 6.2. It should be highlighted here that each base model was created with a different training set from the other base models. Two options for sample composition for a dataset with *k* classes were studied for OVA base models design. The options cover the use of un-boosted and boosted OVA base models. The boosted OVA base models were designed based on information obtained from a confusion matrix for a single *k*-class model. The confusion matrix was discussed in section 4.7 of chapter 4. The two options that were studied are as follows:

Option 1

Use 50% of instances from the class c_i that the OVA classifier specialises in and for each of the other classes use *50/(k-1)%* instances. This option results in the creation of un-boosted OVA base models. This option was used to test whether the increase in the quantity of training data through OVA modeling provides increased predictive performance.

Option 2

Use 50% of instances from the class c_i that the OVA classifier specialises in and for each of the *j (j < k)* classes which are predominantly confused with class c_i, based on the values in the confusion matrix, use *50 / (j-1)%* instances. Recall that the confusion matrix was discussed in chapter 4. Option 2 results in the creation of boosted OVA base models. This option was used to test whether the use of boosting in addition to increasing the quantity of training data through OVA modeling provides additional increases in predictive performance.

7.1.3 Experiment design for the study of OVA modeling

Three categories of experiments on OVA aggregate modeling were conducted as follows:

(1) To compare the performance of *un-boosted* OVA models with single *k*-class models for both 5NN and See5 classification.
(2) To compare the performance of *boosted* OVA models with single *k*-class models for both 5NN and See5 classification.
(3) To compare the performance of *un-boosted* OVA models with *boosted* models for both 5NN and See5 classification.

The methods for OVA base model design and implementation, dataset partitioning and sampling, training dataset selection, model aggregation, and analysis of model performance were presented in chapter 6. These methods were used for the experiment categories listed above. The forest cover type and KDD Cup 1999 datasets were used for the experiments. The 5NN and See5 algorithms were used for the creation of the base models.

7.2 Experiments to study OVA models for 5NN classification

The empirical studies of 5NN OVA classification based on the experiment design of section 7.1.3 are discussed in this section. Section 7.2.1 reports the experiments to compare the predictive performance of single models with un-boosted 5NN OVA models. The process that was followed for the design of boosted OVA models is

discussed in section 7.2.2. Section 7.2.3 presents experimental results to compare the predictive performance of single, un-boosted and boosted 5NN OVA models.

7.2.1 Predictive performance of un-boosted 5NN OVA models

The predictive performance of un-boosted OVA base models and aggregate models is presented in this section. The training sets for the un-boosted OVA base models were designed using option 1 of section 7.1.3. A training sample size of 4000 was used for OVANORMAL, OVADOS, OVAPROBE, and OVAR2L for the KDD Cup 1999 base models. A training set size of 1000 was used for the OVAU2R model in order to limit the amount of bootstrap sampling for the U2R class. Table 7.1 gives the experimental results for the predictive performance of 5NN un-boosted base models for the forest cover type and KDD Cup 1999 datasets. Columns 3 and 4 respectively show the mean and 95% confidence interval for the TPRATE and TNRATE measures as percentages.

Table 7.1: Predictive performance of 5NN OVA un-boosted base models

Dataset, Training sample size, Test set size	Base model name	Mean performance for base models	
		Mean TPRATE%	Mean TNRATE%
Forest cover type (12000) (350 x 10)	OVA1-unboosted	91.8 ± 2.5	85.0 ± 0.8
	OVA2-unboosted	83.8 ± 2.6	80.5 ± 1.1
	OVA3-unboosted	90.4 ± 1.1	85.3 ± 0.9
	OVA4-unboosted	95.6 ± 1.5	94.3 ± 0.6
	OVA5-unboosted	99.6 ± 0.5	90.8 ± 0.8
	OVA 6-unboosted	98.4 ± 1.0	84.6 ± 0.8
	OVA7-unboosted	99.2 ± 0.6	93.7 ± 0.5
KDD Cup 1999 (4000) (350 x 10)	OVANORMAL-unboosted	99.3 ± 0.6	73.0 ± 1.5
	OVADOS-unboosted	69.1 ± 4.5	97.8 ± 0.7
	OVAPROBE-unboosted	95.9 ± 1.2	88.5 ± 3.4
	OVAR2L-unboosted	76.7 ± 2.8	82.0 ± 1.6
	OVAU2R-unboosted	54.3 ± 0.0	97.7 ± 0.6

The results of table 7.1 indicate that the forest cover type base models have very high TPRATE and TNRATE values and are therefore highly competent at predicting the classes they are designed to predict. It remains to be seen if combining these highly competent base models into an aggregate model provides performance gains. The OVANORMAL and OVAPROBE base models for the KDD Cup 1999 dataset have very high TPRATE and TNRATE values while the OVADOS, OVAR2L and OVAU2R have much lower values.

The 5NN OVA base models were combined into aggregate models. The predictions of the individual 5NN OVA models on each test instance were combined into a single prediction using the algorithm of figure 6.4 presented in section 6.4.3. Recall that the algorithm in figure 6.4 uses the probabilistic scores assigned by the base models to determine the best prediction. When there is more than one prediction with the highest probabilistic score (tied scores) the distances to the nearest neighbours are used to break the tie. Single *k*-class models were created and also tested on the same instances as the aggregate models. The single 7-class model for forest cover type was created from a training dataset of 12000 instances with an equal class distribution for all the classes.

The KDD Cup 1999 single 5-class model was created from a training dataset of 4000 instances. The training dataset for the KDD Cup 1999 single model was composed of 500 instances for the class U2R and 3500 instances for the remaining four classes in equal proportions. Table 7.2 shows the results for the 5NN single and un-boosted OVA aggregate models for the forest cover type and KDD Cup 1999 datasets. The details for predictive accuracy and TPRATE values for the single and aggregate models for the forest cover type dataset are given in the appendix tables F.1 and F.2. The details for predictive accuracy and TPRATE values for the single and aggregate models for the KDD Cup 1999 dataset are given in the appendix tables F.9 and F.10.

Table 7.2: Predictive performance of 5NN single and un-boosted OVA aggregate models

Dataset, (training set size), (test set size)	Class	Mean predictive performance of models	
		Single model	un-boosted OVA aggregate model
		Mean TPRATE%	Mean TPRATE%
Forest cover type (12000) (350 x 10)	ALL(accuracy)	74.7 ± 1.0	80.5 ± 0.9
	1	62.8 ± 3.4	70.0 ± 4.3
	2	48.8 ± 2.8	58.4 ± 2.7
	3	56.8 ± 4.1	71.8 ± 1.9
	4	92.4 ± 1.8	89.8 ± 1.9
	5	91.2 ± 2.0	95.8 ± 3.1
	6	75.0 ± 2.1	80.8 ± 4.5
	7	96.0 ± 1.3	96.6 ± 0.6
KDD Cup 1999 (4000) (350 x 10)	ALL (accuracy)	68.5 ± 1.4	72.4 ±1.1
	NORMAL	84.4 ± 3.1	92.7 ±2.8
	DOS	66.3 ± 5.0	66.0 ±4.4
	PROBE	95.7 ± 1.2	95.2 ±1.0
	R2L	64.7 ± 3.6	65.4 ±3.6
	U2R	31.6 ± 0.3	42.6 ± 0.4

Student's paired samples t-test and the *Diff(A,S)* and *Ratio(A,S)* measures discussed in section 6.4 were used to compare the performance of the single models with that

of the aggregate models. Tables 7.3 and 7.4 respectively give the results of the statistical tests for the forest cover type and KDD Cup 1999 datasets. The paired t-test results of table 7.3 indicate that for the forest cover type dataset, the un-boosted aggregate model provides statistically significant increases in accuracy and the TPRATE values for five out of seven classes. The Diff(A,S) measure indicates an increase in accuracy of 5.8%. The increases in the class TPRATE values range between 4.6% and 15%. The Ratio(A,S) measure indicates a relative improvement of 0.2 for the accuracy and relative improvements that range between 0.2 and 0.5. Recall that the maximum improvement as measured by Ratio(A,S) is 1.0.

Table 7.3: Statistical tests to compare the performance of 5NN single and un-boosted OVA aggregate models for forest cover type

Group names and mean accuracy / TPRATE% for 10 test sets		Student's paired t-test (9 df)			Performance improvement measures	
Group A un-boosted aggregate model	Group S single model	95% CI of mean difference	P value (2 tail)	Group A better than Group S?	Diff(A,S)%	Ratio(A,S)
All classes-A (80.5 ± 0.9)	All classes-S (74.7 ± 1.0)	[4.1, 7.5]	0.000	yes	5.8	0.2
Class1-A (70.0 ± 4.3)	Class1-S (62.8 ± 3.4)	[0.9, 13.5]	0.029	yes	7.2	0.2
Class2-A (58.4 ± 2.7)	Class2-S (48.8 ± 2.8)	[6.3, 12.9]	0.000	yes	9.6	0.2
Class3-A (71.8 ± 1.9)	Class3-S (56.8 ± 4.1)	[11.8, 18.3]	0.000	yes	15.0	0.3
Class4-A (89.8 ± 1.9)	Class4-S (92.4 ± 1.8)	[-4.6, -0.6]	0.018	no	-2.6	-0.3
Class5-A (95.8 ± 3.1)	Class5-S (91.2 ± 2.0)	[0.8, 8.4]	0.022	yes	4.6	0.5
Class6-A (80.8 ± 4.5)	Class6-S (75.0 ± 2.1)	[0.5, 11.1]	0.036	yes	5.8	0.2
Class7-A (96.6 ± 0.6)	Class7-S (96.0 ± 1.3)	[-1.2, 2.4]	0.468	no	0.6	0.1

The paired t-test results of table 7.4 indicate for the KDD Cup 1999 dataset that the un-boosted aggregate model provides statistically significant increases in accuracy and the TPRATE values for two out of five classes. The Diff(A,S) measure indicates an increase in accuracy of 3.9%. The increases in the class TPRATE values are 8.3% for class NORMAL and 11% for class U2R. The Ratio(A,S) measure indicates a relative improvement of 0.1 for the accuracy and relative improvements of 0.2 for the class U2R and 0.5 for the class NORMAL. Overall, the use of OVA base models based on option 1 of section 7.1.2 for training dataset selection, provides significant improvements in predictive performance. Recall that the training set for each un-boosted OVA base model is composed of 50% for the class that the base model predicts and 50% for all the other classes combined.

Table 7.4: Statistical tests to compare the performance of 5NN single and un-boosted OVA aggregate models for KDD Cup 1999

Group names and mean accuracy / TPRATE% for 10 test sets		Student's paired t-test (9 df)			Performance improvement measures	
Group A Aggregate model	Group S Single model	95% CI of mean difference	P value (2 tail)	Group A better than Group S?	*Diff(A,S)%*	*Ratio(A,S)*
All classes-A (72.4±1.1)	All classes-S (68.5±1.4)	[2.6, 5.0]	0.000	yes	3.9	0.1
NORMAL-A (92.7±2.8)	NORMAL-S (84.4±3.1)	[5.0, 11.6]	0.000	yes	8.3	0.5
DOS-A (66.0±4.4)	DOS-S (66.3±5.0)	[-2.6, 2.0]	0.790	no	-0.3	0.0
PROBE-A (95.2±1.0)	PROBE-S (95.7±1.2)	[-1.4, 0.3]	0.164	no	-0.5	-0.1
R2L-A (65.4±3.6)	R2L-S (64.7±3.6)	[-1.9, 3.3]	0.560	no	0.7	0.0
U2R-A (42.6±0.4)	U2R-S (31.6±0.3)	[10.3, 11.8]	0.000	yes	11.0	0.2

7.2.2 Design of boosted 5NN OVA base models

The results of section 7.2.1 have demonstrated that un-boosted OVA base models result in improvements in predictive performance. Boosting was discussed in sections 2.8.2 and 2.10.2 as a method of bias error reduction. Option 2 of section 7.1.2 involves the use of boosted base models. The author hypothesised that boosting of OVA base models should lead to further improvements in predictive performance compared to un-boosted models. Recall from chapter 2 that boosting is a statistical technique for directing the greatest effort towards those areas of the instance space where prediction is most difficult. It was further hypothesised that examination of the confusion matrix for the single *k*-class model should provide information about those areas of the instance space where correct prediction is most difficult to achieve. Confusion matrices were discussed in section 4.7.

It was further hypothesized that a predictive model makes incorrect decisions in those regions which are class boundary regions in the instance space. The term *confusion regions*, was used by the author to refer to these regions. *Confusion regions* were discussed in section 2.7. Table 7.5 shows the confusion matrix for the single k-class model for the forest cover type dataset based on 5 test sets. For simplicity of presentation only the counts for the off-diagonal cells are shown in the confusion matrix.

Table 7.5: Confusion matrix for the 5NN single model for the forest cover type dataset

Single model confusion matrix, training size =12000, test set size = 250 per class								Total confusion	
Actual class	Predicted class								
	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7	SUMS	PCNT
Class 1		38	2		10	1	49	**100**	40
Class 2	38		16	1	53	12	7	**127**	50.8
Class 3		3		41	6	65		**115**	46
Class 4			8			14		**22**	8.8
Class 5		2	8			6		**16**	6.4
Class 6		4	37	25	10			**76**	30.4
Class 7	4	3						**7**	2.8

A confusion matrix can be used to identify the existence or possible absence of a confusion region between the decision regions of any two classes. Identification of the confusion regions was done based on the confusion matrices, using the following simple deductive logic:

(1) Given two classes c_i and c_j, if the entry (c_i, c_j) in the confusion matrix is zero, then c_i and c_j do not have a common boundary in the instance space, and so, do not share a confusion region

(2) If the entry (c_i, c_j) for two classes c_i and c_j is non-zero, then the two classes share a common decision boundary in the instance space, and the value in the cell (c_i, c_j) indicates the intensity of the class confusion between the two classes.

The confusion matrix of table 7.5 indicates for the 5NN single model of the forest cover type dataset that class 1 gets predominantly confused with classes 2 and 7. On the other hand class 2, is never confused with classes 3, 4 or 6. Class 7 is predominantly confused with classes 1 and 2, but is never confused with classes 3, 4, 5 or 6. The following strategy could be used to reduce the confusion between class 1 and class 2: Select the training set sample for OVA1 with class 1 as the positive instances and classes 2 and 7 as negative instances. This should provide higher instance space coverage of the confusion regions between classes 1 and 2, and classes 1 and 7. In other words, the training sample for the OVA1 base model is boosted so that there are more instances for the classes that are difficult to separate.

Table 7.6 shows the sample composition that was used for the OVA base models for the forest cover type training datasets for purposes of reducing class confusion. The entries in the second column have the following interpretation. If the counts of

confusion matrix cells (c_i, c_j) and (c_j, c_i) are high then c_i is predominantly confused with c_j. The rationale behind the training sample composition for base model OVA_i was to ensure that training instances of the classes appearing in column 2 are included in the training dataset as negative instances. The number of positive and negative instances should be equal as was done for the un-boosted models.

Table 7.6: 5NN training sample composition to reduce class confusion for forest cover type

Class	Predominantly Confused with	Training sample composition for OVA base model	
		Percentage of positive instances	Percentage of negative instances
C1:	C2, C7	C1: 50	C2: 25, C7: 25
C2	C1,C3,C5	C2: 49	C1:17, C3:17, C5: 17
C3	C2,C4,C6	C3: 49	C2:17, C4:17, C6: 17
C4	C3,C6	C4: 50	C3: 25, C6: 25
C5	C2	C5: 50	C2: 50
C6	C3,C4,C5	C6: 49	C3:17, C4: 17, C5: 17
C7	C1,C2	C7: 50	C1: 25, C7: 25

Table 7.7 shows the confusion matrix for the KDD Cup 1999 dataset for the single model with a training set size of 4000 instances. Based on the information in the confusion matrix, training set samples for OVA base models were designed to provide a higher coverage of the confusion regions. The training set sample design is shown in table 7.8. It should be noted that the sample composition for the OVANORMAL base model is the same as for the un-boosted base model.

Table 7.7: Confusion matrix for the 5NN single model for the KDD Cup 1999 dataset

Single model, training size = 4000, test set size = 350 instances per class							
	Predicted class					Total confusion	
Actual class	NORMAL	DOS	PROBE	R2L	U2R	SUM	PCNT
NORMAL		13	**29**	4	2	48	13.7
DOS	13		**91**	14	3	121	**34.6**
PROBE	11				4	15	4.3
R2L	**107**	1			6	114	32.6
U2R	**170**			**69**		239	**68.3**

Table 7.8: Training sample composition to reduce class confusion for 5NN models for KDD Cup 1999

Class	Predominantly Confused for:	Training sample composition for OVA base models		
		Percentage of positive instances	Percentage of negative instances	Training sample size
NORMAL	R2L,U2R, DOS,PROBE	NORMAL: 50	R2L:12.5, U2R:12.5, DOS:12.5, PROBE:12.5	4000
DOS	NORMAL,PROBE, R2L	DOS: 49	NORMAL:17, PROBE: 17, R2L:17	4000
PROBE	NORMAL, DOS, U2R	PROBE: 49	NORMAL:17, DOS:17, U2R:17	4000
R2L	NORMAL	R2L: 50	NORMAL:50	4000
U2R	NORMAL, R2L	U2R: 50	NORMAL:25, R2L:25	1000

7.2.3 Predictive performance of boosted 5NN OVA models

Boosted 5NN base models were created based on the sample designs shown in tables 7.6 and 7.9 for the forest cover type and KDD Cup 1999 datasets. Implementation of the aggregate models based on the boosted base models as shown in table 7.6 and 7.8 did not result in performance improvements over the single models. However, the approach of using a combination of boosted and un-boosted base models resulted in performance improvements for the forest cover type aggregate model. The base models used for the boosted version of the OVA aggregate models for the forest cover type and KDD Cup 1999 datasets are given in table 7.9. The rationale for choosing boosted base models was as follows: If a boosted base model had a higher TPRATE value than the un-boosted version, the boosted version was selected. This was the case, for example, for the OVA4 forest cover type base model. If a boosted base model had a TPRATE comparable (equal) to that of the un-boosted version then the boosted base model was included in the aggregate model. If a performance improvement was realized, then the boosted base model was retained, otherwise it was replaced with the un-boosted version.

Table 7.10 shows the predictive performance results for the single, un-boosted and boosted OVA aggregate models based on the boosted OVA base models for the forest cover type and KDD Cup 1999 datasets. The details for predictive accuracy and TPRATE measure for the boosted OVA aggregate models for the forest cover type and KDD Cup 1999 datasets are respectively given in appendix tables F.3 and F.11

Table 7.9: Predictive performance of 5NN OVA boosted base models

Dataset, Training sample size, Test set size	Base model name	Mean performance for base models	
		TPRATE%	TNRATE%
Forest cover type (12000) (350 x 10)	OVA1-unboosted	91.8 ± 2.5	85.0 ± 0.8
	OVA2-unboosted	83.8 ± 2.6	80.5 ± 1.1
	OVA3-unboosted	90.4 ± 1.1	85.3 ± 0.9
	OVA 4-boosted	100.0 ± 0.0	96.3 ± 0.6
	OVA 5-boosted	99.6 ± 0.5	89.0 ± 0.9
	OVA 6-boosted	94.2 ± 0.9	87.3 ± 1.3
	OVA 7-unboosted	99.2 ± 0.6	93.7 ± 0.5
KDD Cup 1999 (4000) (350 x 10)	OVANORMAL-unboosted	99.3 ± 0.6	73.0 ± 1.5
	OVADOS-boosted	68.3 ± 4.8	97.3 ± 0.8
	OVAPROBE-unboosted	95.9 ± 1.2	88.5 ± 3.4
	OVAR2L-boosted	68.2 ± 3.3	82.0 ± .2
	OVAU2R-unboosted	54.3 ±0.0	97.7 ±0.6

Table 7.10: Predictive performance of 5NN single, un-boosted and boosted OVA aggregate models

Dataset, (training set size), (test set size)	Class	Mean predictive performance of models		
		single model	un-boosted OVA aggregate model	boosted OVA aggregate model
		Mean TPRATE%	Mean TPRATE%	Mean TPRATE%
Forest cover type (12000) (350 x 10)	ALL(accuracy)	74.7 ± 1.0	80.5 ± 0.9	82.0 ± 0.6
	1	62.8 ± 3.4	70.0 ± 4.3	70.0 ± 4.3
	2	48.8 ± 2.8	58.4 ± 2.7	62.0 ± 3.4
	3	56.8 ± 4.1	71.8 ± 1.9	71.0 ± 1.3
	4	92.4 ± 1.8	89.8 ± 1.9	100.0 ± 0.0
	5	91.2 ± 2.0	95.8 ± 3.1	97.0 ± 0.9
	6	75.0 ± 2.1	80.8 ± 4.5	77.6 ± 2.0
	7	96.0 ± 1.3	96.6 ± 0.6	96.6 ± 0.6
KDD Cup 1999 (4000) (350 x 10)	ALL (accuracy)	68.5 ± 1.4	72.4 ±1.1	71.0 ±1.2
	NORMAL	84.4 ± 3.1	92.7 ±2.8	92.4 ±3.0
	DOS	66.3 ± 5.0	66.0 ±4.4	66.0 ±5.1
	PROBE	95.7 ± 1.2	95.2 ±1.0	95.4 ±1.2
	R2L	64.7 ± 3.6	65.4 ±3.6	60.9 ±3.8
	U2R	31.6 ± 0.3	42.6 ± 0.4	40.5 ±1.4

Table 7.11 shows the results of the statistical tests to compare the predictive performance of the single, un-boosted and boosted OVA aggregate models for the forest cover type dataset. The paired t-test results of table 7.11 compare the boosted OVA aggregate model with the single model. The results indicate that the boosted aggregate model provides statistically significant increases in accuracy for the forest cover type dataset. The boosted model also provides increased TPRATE values for six out of seven classes. The Diff(A,S) measure indicates an increase in accuracy of 7.3%. The increases in the class TPRATE values range between 2.6% and 14.2%. The Ratio(A,S) measure indicates a relative improvement of 0.3 for the accuracy and

relative improvements that range between 0.1 and 1.0. Recall that a Ratio(A,S) value of 1.0 indicates maximum improvement.

Table 7.11: Statistical tests to compare the 5NN single, un-boosted and boosted OVA aggregate models for forest cover type

Group names and mean accuracy / TPRATE for 10 test sets		Student's paired t-test (9 df)			Performance improvement measures	
Group A aggregate model	Group B single model	95% CI of mean difference	P value (2 tail)	Group A better than Group B?	Diff(A,B)%	Ratio(A,B)
boosted All classes-A (82.0 ± 0.6)	single All classes-S (74.7 ± 1.0)	[5.8, 8.8]	0.000	yes	7.3	0.3
boosted Class1-A (70.0 ± 4.3)	single Class1-S (62.8 ± 3.4)	[0.9, 13.5]	0.029	yes	7.2	0.2
boosted Class2-A (62.0 ± 3.4)	single Class2-S (48.8 ± 2.8)	[9.8, 16.6]	0.000	yes	13.2	0.3
boosted Class3-A (71.0 ± 1.3)	single Class3-S (56.8 ± 4.1)	[9.8, 18.6]	0.000	yes	14.2	0.3
boosted Class4-A (100.0 ± 0.0)	single Class4-S (92.4 ± 1.8)	[5.5, 9.7]	0.000	yes	7.6	1.0
boosted Class5-A (97.0 ± 0.9)	single Class5-S (91.2 ± 2.0)	[3.8, 7.8]	0.000	yes	5.8	0.7
boosted Class6-A (77.6 ± 2.0)	single Class6-S (75.0 ± 2.1)	[1.2, 4.0]	0.002	yes	2.6	0.1
boosted Class7-A (96.6 ± 0.6)	single Class7-S (96.0 ± 1.3)	[-1.2,2.4]	0.468	no	0.6	0.1
boosted All classes-A (82.0 ± 0.6)	un-boosted All classes-A (80.5 ± 0.9)	[0.5,2.7]	0.009	yes	1.5	0.1
boosted Class1-A (70.0 ± 4.3)	un-boosted Class1-A (70.0 ± 4.3)	no variance	no variance	no	0.0	0.0
boosted Class2-A (62.0 ± 3.4)	un-boosted Class2-A (58.4 ± 2.7)	[1.8,5.4]	0.001	yes	3.6	0.1
boosted Class3-A (71.0 ± 1.3)	un-boosted Class3-A (71.8 ± 1.9)	[-3.1,1.5]	0.443	no	-0.8	0.0
boosted Class4-A (100.0 ± 0.0)	Class4-A (89.8 ± 1.9)	[8.0,12.4]	0.000	yes	10.2	1.0
boosted Class5-A (97.0 ± 0.9)	un-boosted Class5-A (95.8 ± 3.1)	[-2.9,5.3]	0.520	no	1.2	0.3
boosted Class6-A (77.6 ± 2.0)	un-boosted Class6-A (80.8 ± 4.5)	[-7.7,1.2]	0.137	no	-3.2	-0.2
boosted Class7-A (96.6 ± 0.6)	un-boosted Class7-A (96.6 ± 0.6)	no variance	no variance	no	0.0	0.0

The paired t-tests to compare the boosted and un-boosted aggregate models indicate for the forest cover type dataset that the boosted aggregate model provides statistically significant increases in accuracy. The boosted model also provides increased TPRATE values for two out of seven classes. The Diff(A,S) measure indicates an additional increase in accuracy of 1.5%, due to boosting. The increases in the class TPRATE values are 3.6% for class 2 and 10.2% for class 4. The Ratio(A,S) measure indicates a relative improvement of 0.1 for the accuracy and relative improvements of 0.1 for class 1 and 1.0 for class 4.

Table 7.12 shows the results of the statistical tests to compare the predictive performance of the boosted and un-boosted OVA aggregate models for the KDD Cup 1999 dataset. A comparison of the test results of tables 7.4 and 7.12 indicates that the use of un-boosted 5NN OVA base models results in performance improvements over the single model for the KDD Cup 1999 dataset. However, there are no performance gains released due to boosting of 5NN OVA base models for the KDD Cup 1999 dataset.

Table 7.12: Statistical tests to compare the 5NN single, un-boosted and boosted OVA aggregate models for KDD Cup 1999

Group names and mean accuracy / TPRATE for 10 test sets		Student's paired t-test (9 df)			Performance improvement measures	
Group A aggregate model	Group B single model	95% CI of mean difference	P value (2 tail)	Group A better than Group B?	*Diff(A,B)%*	*Ratio(A,B)*
boosted All classes-A (71.0±1.2)	un-boosted All classes-A (72.4±1.1)	[-2.1, -0.6]	0.002	no	-1.3	0.0
boosted NORMAL-A (92.4±3.0)	un-boosted NORMAL-A (92.7±2.8)	[-0.9, 0.4]	0.343	no	-0.3	0.0
boosted DOS-A (66.0±5.1)	un-boosted DOS-A (66.0±4.4)	[-1.5, 1.5]	0.988	no	0.0	0.0
boosted PROBE-A (95.4±1.2)	un-boosted PROBE-A (95.2±1.0)	[-0.2, 0.7]	0.168	no	0.3	0.1
boosted R2L-A (60.9±3.8)	un-boosted R2L-A (65.4±3.6)	[-7.4, -1.7]	0.005	no	-4.6	-0.1
boosted U2R-A (40.5±1.4)	un-boosted U2R-A (42.6±0.4)	[-4.1, -0.2]	0.031	no	-2.2	0.0

7.3 Experiments to study OVA models for See5 classification

The empirical studies of See5 OVA classification based on the experiment design presented in section 7.1.3 are discussed in this section. Section 7.3.1 reports the experiments to compare predictive performance of single models with un-boosted See5 OVA models. The design of boosted OVA models is discussed in section 7.3.2. Section 7.3.3 presents experimental results to compare predictive performance of single, un-boosted and boosted See5 OVA models.

7.3.1 Predictive performance of un-boosted See5 OVA models

The training datasets that were used for the un-boosted 5NN OVA base models were also used for the experiments to compare See5 single and un-boosted OVA aggregate models. Table 7.13 gives the experimental results for the predictive performance of See5 OVA un-boosted base models. Columns 3 and 4 respectively show the mean and 95% confidence interval for the TPRATE and TNRATE measures as percentages.

Table 7.13: Predictive performance of See5 OVA un-boosted base models

Dataset, Training sample size, Test set size	Base model name	Mean performance for base models	
		Mean TPRATE%	Mean TNRATE%
Forest cover type (12000) (350 x 10)	OVA1-unboosted	92.6 ± 2.3	82.7 ± 0.7
	OVA2-unboosted	85.6 ± 1.7	82.0 ± 0.7
	OVA 3-unboosted	93.2 ± 1.7	86.8 ± 0.5
	OVA 4-unboosted	99.0 ± 0.9	95.9 ± 0.6
	OVA 5-unboosted	98.6 ± 1.3	93.7 ± 0.7
	OVA 6-unboosted	92.2 ± 1.9	88.0 ± 0.5
	OVA 7-unboosted	99.6 ± 0.5	96.1 ± 0.5
KDD Cup 1999 (4000) (350 x 10)	OVANORMAL-unboosted	98.4 ± 0.7	82.7 ± 0.9
	OVADOS-unboosted	53.2 ± 4.6	99.6 ± 0.1
	OVAPROBE-unboosted	88.6 ± 1.3	90.3 ± 1.0
	OVAR2L-unboosted	37.4 ± 3.6	88.9 ± 0.8
	OVAU2R-unboosted	65.7 ± 0.0	96.8 ± 0.8

The results of table 7.13 indicate that the forest cover type base models have very high TPRATE and TNRATE values and are therefore highly competent at predicting the classes they are designed to predict. It remains to be seen if combining these highly competent base models into an aggregate model provides performance gains.

The OVANORMAL and OVAPROBE base models for KDD Cup 1999 have high TPRATE and TNRATE values. While OVADOS, OVAR2L and OVAU2R have high TNRATE values, the TPRATE values for these base models are low.

The See5 OVA base models were combined into aggregate models. The predictions of the individual See5 OVA base models on each test instance were combined into a single prediction using the combination algorithm of figure 6.3 that was presented in section 6.4.3. Recall that the algorithm in figure 6.3 uses the probabilistic scores assigned by the base models to determine the best prediction. Single *k*-class models were created and also tested on the same instances as the aggregate models. Table 7.14 shows the results for the single and aggregate models for the forest cover type and KDD Cup 1999 datasets. The details for predictive accuracy and TPRATE measure for the forest cover type single and aggregate models are respectively given in appendix tables F.5 and F.6. The details for predictive accuracy and TPRATE measure for the KDD Cup 1999 single and aggregate models are respectively given in appendix tables F.13 and F.14.

Table 7.14: Predictive performance of See5 single and un-boosted OVA aggregate models

Dataset, (training set size), (test set size)	Class	Mean predictive performance of models	
		Single model	un-boosted OVA aggregate model
		Mean TPRATE%	Mean TPRATE%
Forest cover type (12000) (350 x 10)	ALL(accuracy)	76.9 ± 1.0	75.3 ± 0.7
	1	57.4 ± 3.4	60.6 ± 2.6
	2	63.8 ± 3.0	49.8 ± 3.6
	3	60.8 ± 3.3	64.0 ± 1.8
	4	96.8 ± 1.0	86.6 ± 1.7
	5	86.2 ± 2.4	94.4 ± 1.8
	6	77.8 ± 3.3	79.2 ± 2.0
	7	95.6 ± 1.6	92.8 ± 2.5
KDD Cup 1999 (4000) (350 x 10)	ALL (accuracy)	63.8 ± 1.3	63.3 ± 1.2
	NORMAL	86.0 ± 3.1	98.3 ± 0.7
	DOS	82.0 ± 3.8	50.1 ± 4.4
	PROBE	36.8 ± 2.4	88.0 ± 1.3
	R2L	37.7 ± 3.3	34.3 ± 3.3
	U2R	77.1 ± 0.0	45.7 ± 0.0

Student's paired samples t-test and the *Diff(A,S)* and *Ratio(A,S)* measures were used to compare the performance of the single models with that of the aggregate models. Tables 7.15 and 7.16 respectively give the results of the statistical tests for the forest cover type and KDD Cup 1999 datasets. The results of the paired samples t-tests for the forest cover type models indicate that there is a general degradation in performance due to the use of the un-boosted aggregate model. The accuracy and TPRATE values for 6 out of 7 classes are lower for the un-boosted OVA aggregate

model compared to the single model. The statistical tests of table 7.16 indicate that there is no overall improvement in accuracy due the un-boosted OVA aggregate model. However, there is a significant improvement in the TPRATE values for the NORMAL and PROBE classes.

Table 7.15: Statistical tests to compare the performance of See5 single and un-boosted OVA aggregate models for forest cover type

Group names and mean accuracy / TPRATE for 10 test sets		Student's paired t-test (9 df)			Performance improvement measures	
Group A un-boosted aggregate model	Group S single model	95% CI of mean difference	P value (2 tail)	Group A better than Group S?	Diff(A,S) %	Ratio(A,S)
All classes-A (75.3±0.7)	All classes-S (76.9 ± 1.0)	[-2.5, -0.8]	0.002	no	-1.6	-0.1
Class1-A (60.6±2.6)	Class1-S (57.4 ± 3.4)	[-1.1, 7.5]	0.125	no	3.2	0.1
Class2-A (49.8±3.6)	Class2-S (63.8 ± 3.0)	[-17.6, -10.4]	0.000	no	-14	-0.4
Class3-A (64.0±1.8)	Class3-S (60.8 ± 3.3)	[-1.3, 7.7]	0.141	no	3.2	0.1
Class4-A (86.6±1.7)	Class4-S (96.8 ± 1.0)	[-12.5, -7.0]	0.000	no	-10.2	-3.2
Class5-A (94.4±1.8)	Class5-S (86.2 ± 2.4)	[5.8, 10.6]	0.000	yes	8.2	0.6
Class6-A (79.2±2.0)	Class6-S (77.8 ± 3.3)	[-2.1,4.9]	0.390	no	1.4	0.1
Class7-A (92.8±2.5)	Class7-S (95.6 ± 1.6)	[-5.3, -0.4]	0.029	no	-2.8	-0.6

Table 7.16: Statistical tests to compare the performance of See5 single and un-boosted OVA aggregate models for KDD Cup 1999

Group names and mean accuracy / TPRATE% for 10 test sets		Student's paired t-test (9 df)			Performance improvement measures	
Group A un-boosted aggregate model	Group S single model	95% CI of mean difference	P value (2 tail)	Group A better than Group S?	*Diff(A,S)%*	*Ratio(A,S)*
All classes-A (63.3 ± 1.2)	All classes-S (63.8 ± 1.3)	[-2.0, 0.9]	0.430	no	-0.5	0.01
NORMAL-A (98.3 ± 0.7)	NORMAL-S (86.0 ± 3.1)	[9.0, 15.6]	0.000	yes	12.3	0.9
DOS-A (50.1 ± 4.4)	DOS-S (82.0 ± 3.8)	[-38.2, -25.5]	0.000	no	-31.9	-1.8
PROBE-A (88.0 ± 1.3)	PROBE-S (36.4 ± 2.4)	[48.2, 54.9]	0.000	yes	52.6	0.8
R2L-A (34.3 ± 3.3)	R2L-S (37.7 ± 3.3)	[-7.5, 0.5]	0.082	no	-3.4	-0.1
U2R-A (45.7 ± 0.0)	U2R-S (77.1 ± 0.0)	no variance	no variance	no	-31.4	-1.4

7.3.2 Design of See5 boosted OVA base models

Table 7.17 and 7.18 show the confusion matrices for the single models created from the training samples with an equal class distribution for the forest cover type and KDD Cup 1999 datasets. For simplicity of presentation, only the off-diagonal counts are given. A comparison of the confusion matrices for forest cover type for the 5NN and See5 models reveals that the nature of the class confusion is fairly similar for both models. However, there a significant change in the level of confusion between the PROBE and U2R classes of the KDD Cup 1999 dataset. The 5NN OVA training sample designs given in table 7.6 for forest cover type were also used for the implementation of the See5 OVA base models. The sample design for KDD Cup 1999 See5 OVA base models is shown in table 7.19. It should be noted that the sample composition for the OVANORMAL, OVAPROBE and OVAR2L base models is the same as that for the un-boosted base models.

Table 7.17: Confusion matrix for See5 classification tree single 7-class model for forest cover type

See5 single model, training set size = 12000, test set = 250 per class									
Actual class	**Predicted class**							**Total confusion**	
	Class 1	**Class 2**	**Class 3**	**Class 4**	**Class 5**	**Class 6**	**Class 7**	**SUMS**	**PCNT**
Class 1		**60**			3	2	**38**	**103**	41.2
Class 2	**43**		5		**32**	8	8	**96**	38.4
Class 3				**26**	**11**	**50**		**87**	34.8
Class 4			6					**6**	2.4
Class 5		**17**	6			6		**29**	11.6
Class 6		4	**30**	**23**	3			**60**	24
Class 7	**16**							**16**	6.4

Table 7.18: Confusion matrix for See5 classification tree single 5-class model for KDD Cup 1999

See5 single model, training set = 4000, test set size = 350 instances per class							
	Predicted class					Total confusion	
Actual class	NORMAL	DOS	PROBE	R2L	U2R	SUM	PCNT
NORMAL		1	**30**	11	1	43	12.3
DOS	**32**		15	10		57	16.3
PROBE	4	17			**198**	219	62.6
R2L	**185**		8		20	213	60.9
U2R	**70**	10				80	22.9

Table 7.19: See 5 Training sample composition to reduce class confusion for KDD Cup 1999

Class	Predominantly Confused for:	Training sample composition for OVA base models		
		Percentage of positive instances	Percentage of negative instances	Training sample size
NORMAL	R2L,U2R, DOS,PROBE	NORMAL: 50	R2L:12.5, U2R:12.5, DOS:12.5, PROBE:12.5	4000
DOS	NORMAL,PROBE, R2L	DOS: 49	NORMAL:17. PROBE: 17, R2L:17	4000
PROBE	NORMAL, DOS, R2L,U2R	PROBE: 50	NORMAL:12.5, DOS:12.5, R2L:12.5, U2R:12.5	4000
R2L	NORMAL, DOS,U2R	R2L: 49	NORMAL:17, DOS:17, U2R:17	4000
U2R	NORMAL, DOS, PROBE, R2L	U2R: 50	NORMAL:12.5, DOS:12.5, PROBE: 12.5, R2L:12.5	1000

The performance of the boosted base models and aggregate models is discussed in the next section.

7.3.3 Predictive performance of boosted See5 OVA models

Boosted See5 base models were created based on the training set designs of table 7.6 for forest cover type and table 7.19 for the KDD Cup 1999 dataset. The TPRATE and TNRATE values for the base models are given in table 7.20. A comparison of the un-boosted base models of table 7.9 and the boosted base models of table 7.20 reveals that the boosted base models generally have lower mean TPRATE values.

Table 7.20: Predictive performance of See5 OVA boosted base models

Dataset, Training sample size, Test set size	Base model name	Mean performance for base models	
		Mean TPRATE%	Mean TNRATE%
Forest cover type (12000) (350 x 10)	OVA1-boosted	75.0 ± 2.9	92.7 ± 0.7
	OVA2-boosted	81.4 ± 1.8	83.3 ± 0.9
	OVA3-boosted	85.8 ± 2.4	91.8 ± 0.7
	OVA4-boosted	99.0 ± 0.7	97.5 ± 0.4
	OVA5-boosted	96.4 ± 1.4	90.4 ± 0.7
	OVA6-boosted	93.2 ± 1.2	91.3 ± 0.8
	OVA7-boosted	97.6 ± 1.1	98.3 ± 0.3
KDD Cup 1999 (4000) (350 x 10)	OVANORMAL-unboosted	99.3 ± 0.6	73.0 ± 1.5
	OVADOS-boosted	56.3 ± 4.3	88.5 ± 0.2
	OVAPROBE -unboosted	95.9 ± 1.2	88.5 ± 3.4
	OVAR2L-boosted	51.0 ± 4.4	88.2 ± 1.4
	OVAU2R-unboosted	54.3 ± 0.0	97.7 ± 0.6

Boosted aggregate models were created using the base models of table 7.20. Table 7.21 shows the predictive performance results for the See5 single, un-boosted and boosted OVA aggregate models for the forest cover type and KDD Cup 1999 datasets. The details for predictive accuracy and TPRATE measures for the forest cover type boosted aggregate model are given in appendix tables table F.7. The

details for predictive accuracy and TPRATE measures for the KDD Cup 1999 boosted aggregate model are given in appendix table F.15.Table 7.22 shows the results of the statistical tests to compare the predictive performance of the forest cover type single, un-boosted and boosted aggregate models.

Comparison of the test results of tables 7.15 and 7.22 indicates that there is degradation in performance when un-boosted OVA base models are combined into an aggregate model. However, comparison of the forest cover type single and boosted OVA aggregate models indicates that there are significant performance improvements in the accuracy and TPRATE values for 3 out of 7 classes. The Diff(A,S) measure indicates an increase of 2.5% in accuracy and increases of TPRATE values of 2.2% (class 7), 6.0% (class 2), and 7.6% (class 1).

Table 7.21: Predictive performance of See5 single, un-boosted and boosted OVA aggregate models

Dataset, (training set size), (test set size)	Class	Mean predictive performance of models		
		single model	un-boosted OVA aggregate model	boosted OVA aggregate model
		Mean TPRATE%	Mean TPRATE%	Mean TPRATE%
Forest cover type (12000) (350 x 10)	ALL(accuracy)	76.9 ± 1.0	75.3 ± 0.7	79.4 ± 0.6
	1	57.4 ± 3.4	60.6 ± 2.6	65.0 ± 2.9
	2	63.8 ± 3.0	49.8 ± 3.6	69.8 ± 2.4
	3	60.8 ± 3.3	64.0 ± 1.8	63.2 ± 3.3
	4	96.8 ± 1.0	86.6 ± 1.7	95.4 ± 1.3
	5	86.2 ± 2.4	94.4 ± 1.8	88.4 ± 2.3
	6	77.8 ± 3.3	79.2 ± 2.0	76.0 ± 1.9
	7	95.6 ± 1.6	92.8 ± 2.5	97.8 ± 1.1
KDD Cup 1999 (4000) (350 x 10)	ALL (accuracy)	63.8 ± 1.3	63.3 ± 1.2	61.7 ± 0.9
	NORMAL	86.0 ± 3.1	98.3 ± 0.7	99.2 ± 0.6
	DOS	82.0 ± 3.8	50.1 ± 4.4	56.3 ± 4.3
	PROBE	36.8 ± 2.4	88.0 ± 1.3	89.3 ± 1.4
	R2L	37.7 ± 3.3	34.3 ± 3.3	23.6 ± 3.4
	U2R	77.1 ± 0.0	45.7 ± 0.0	40.0 ± 0.0

Table 7.23 shows the results of the statistical tests to compare the predictive performance of the single and boosted aggregate models for KDD Cup 1999. The results show that the predictive accuracy of the aggregate model on all the classes combined is not better than that of the single model. Secondly, the TPRATE values of the aggregate model on the classes DOS, PROBE and R2L are lower than the TPRATE values of the single model on the same classes. However, the aggregate model provides significant improvements on the TPRATE values for the classes NORMAL and U2R. Overall, both the Student's paired t-test results and the *Diff(A,S)* and *Ratio(A,S)* measures demonstrate that there are no impressive gains to be

realized by using the aggregate model. This is in contrast to the forest cover type dataset where the aggregate model provides significant gains over the single model.

Table 7.22: Statistical tests to compare the See5 single, un-boosted and boosted OVA aggregate models for forest cover type

Group names and mean accuracy / TPRATE for 10 test sets		Student's paired t-test (9 df)			Performance improvement measures	
Group A model	Group B model	95% CI of mean difference	P value (2 tail)	Group A better than Group B?	Diff(A,B)%	Ratio(A,B)
boosted All classes-A (79.4 ± 0.6)	single All classes-S (76.9 ± 1.0)	[1.6, 3.4]	0.000	yes	2.5	0.1
boosted Class1-A (65.0 ± 2.9)	single Class1-S (57.4 ± 3.4)	[3.1, 12.1]	0.004	yes	7.6	0.2
boosted Class2-A (69.8 ± 2.4)	single Class2-S (63.8 ± 3.0)	[2.4, 9.6]	0.004	yes	6.0	0.2
boosted Class3-A (63.2 ± 3.3)	single Class3-S (60.8 ± 3.3)	[-0.9, 5.7]	0.132	no	2.4	0.1
boosted Class4-A (95.4 ± 1.3)	single Class4-S (96.8 ± 1.0)	[-3.1, 0.3]	0.088	no	-1.4	-0.4
boosted Class5-A (88.4 ± 2.3)	single Class5-S (86.2 ± 2.4)	[-1.9, 6.3]	0.258	no	2.2	0.2
boosted Class6-A (76.0 ± 1.9	single Class6-S (77.8 ± 3.3)	[-4.0, 0.4]	0.096	no	-1.8	-0.1
boosted Class7-A (97.8 ± 1.1)	single Class7-S (95.6 ± 1.6)	[0.6, 3.8]	0.012	yes	2.2	0.5
boosted All classes-A (79.4 ± 0.6)	un-boosted All classes-A (75.3±0.7)	[3.7, 4.5]	0.000	yes	4.1	0.2
boosted Class1-A (65.0 ± 2.9)	un-boosted Class1-A (60.6±2.6)	[1.7, 7.1]	0.005	yes	4.4	0.1
boosted Class2-A (69.8 ± 2.4)	un-boosted Class2-A (49.8±3.6)	[16.5, 23.5]	0.000	yes	20	0.4
boosted Class3-A (63.2 ± 3.3)	un-boosted Class3-A (64.0±1.8)	[-4.3, 2.7]	0.619	no	-0.8	0.0
boosted Class4-A (95.4 ± 1.3)	un-boosted Class4-A (86.6±1.7)	[7.4. 10.2]	0.000	yes	8.8	0.7
boosted Class5-A (88.4 ± 2.3)	un-boosted Class5-A (94.4±1.8)	[-9.0, -3.0]	0.001	no	-6.0	-1.1
boosted Class6-A (76.0 ± 1.9	un-boosted Class6-A (79.2±2.0)	[-5.7, -0.8]	0.016	no	-3.2	-0.2
boosted Class7-A (97.8 ± 1.1)	un-boosted Class7-A (92.8±2.5)	[2.3, 7.7]	0.002	yes	5.0	0.7

It was stated in sections 2.2.4 and 6.2.3 that syntactic diversity and high competence of base models should lead to performance improvements for an aggregate model. The statistical test results of table 7.16 indicate that the See5 un-boosted OVA aggregate models for the KDD Cup 1999 dataset did not provide a statistically significant increase in predictive accuracy. The statistical test results of table 7.23 indicate that the See5 boosted OVA aggregate model resulted in a statistically significant reduction in predictive accuracy. Two problems were observed for the See5 OVA aggregate models for the KDD Cup 1999 dataset. The first problem was that only two base models (OVANORMAL and OVAPROBE) had a high level of competence, based on the results of tables 7.13 and 7.20.

The second problem was that the prevalence of 'no prediction' was high for both the un-boosted and boosted aggregate models. Recall from section 6.4.3 that it is possible for all OVA base models to predict the class '**other**'. When this happens, then the aggregate model prediction is '**none**' to indicate that there is no valid prediction. The prevalence of '**none**' predictions for the un-boosted OVA aggregate model ranged between 11.4% and 13.4% on the ten test samples. Boosting had the desirable effect of reducing the '**none**' prediction to between 5.4% and 7.7%. However, the rate of incorrect predictions also increased in the boosted version of the model.

Both the See5 un-boosted and boosted base models for the forest cover type dataset had a high level of competence, based on the results of tables 7.13 and 7.20. The occurrence of '**none**' predictions was very low for the forest cover type aggregate models, varying from 0.3% to 1.4% for the un-boosted model and 0.6% to 1.7% for the boosted model. The reduction in predictive performance for the See5 un-boosted OVA aggregate model is due to the occurrence of '**none**' predictions and tied predictions which could not be resolved. The problem of unresolved tied predictions is further discussed in section 7.4.

Table 7.23: Statistical tests to compare the See5 single and boosted OVA aggregate models for KDD Cup 1999

Group names and mean accuracy / TPRATE for 10 test sets		Student's paired t-test (9 df)			Performance improvement measures	
Group A boosted aggregate model	Group S single model	95% CI of mean difference	P value (2 tail)	Group A better than Group S?	*Diff(A,S)%*	*Ratio(A,S)*
All classes-A (61.7 ± 0.9)	All classes-S (63.8 ± 1.3)	[-3.6,0.8]	0.008	no	-2.1	-0.01
NORMAL-A (99.2 ± 0.6)	NORMAL-S (86.0 ± 3.1)	[9.9,16.4]	0.000	yes	13.2	0.9
DOS-A (56.3 ± 4.3)	DOS-S (82.0 ± 3.8)	[-32.6,-18.6]	0.000	no	-25.7	-1.4
PROBE-A (89.3 ± 1.4)	PROBE-S (36.4 ± 2.4)	[49.5,56.3]	0.000	yes	52.9	0.8
R2L-A (23.6 ± 3.4)	R2L-S (37.7 ± 3.3)	-18.3,-10.0]	0.000	no	-14.1	-0.2
U2R-A (40.0 ± 0.0)	U2R-S (77.1 ± 0.0)	no variance	no variance	no	-37.1	-1.6

7.4 Discussion

OVA modeling was studied as a method of problem decomposition with a potential to reduce the bias variance components of the prediction error. It has been demonstrated through the experimental results of this chapter that highly competent and syntactically diverse base models can be obtained through OVA modeling. Recall from chapter 2 and section 6.2 that several researchers (e.g. Sun & Li, 2008; Ho, 1998; Ali & Pazzani, 1996; Krogh & Vedelsby, 1995; Kwok & Carter; 1990; Hansen & Salamon, 1990) have argued that high competence and syntactic diversity of base models lead to aggregate models with improved predictive performance. The experiments reported in this chapter were conducted in order to establish:

(1) Whether the use of OVA base models, each with a different training set, results in increased performance for an aggregate model.
(2) Whether the use of boosting in addition to OVA base models results in additional increased performance for the aggregate model.

Table 7.24 provides a summary of the conclusions from the OVA modeling experiments. The use of OVA modeling alone resulted in increased performance for the 5NN algorithm. The use of OVA modeling alone did not result in increased performance for the See5 algorithm. However, for the forest cover type dataset, the use of boosting in addition to OVA modeling resulted in increased performance for the See5 algorithm.

Table 7.24: Summary of the conclusions from the OVA modeling experiments

Dataset	Algorithm	Is performance improvement realized for the aggregate model when the base models used are:	
		un-boosted OVA?	boosted OVA?
Forest cover type	5NN	yes	yes
	See5	no	yes
KDD Cup 1999	5NN	yes	no
	See5	no	no

Recall that the combination algorithm for the 5NN aggregate models uses probabilistic scores as well as distances to the nearest neighbour in order to resolve tied predictions. On the other hand, the combination algorithm for See5 does not have a second measure available for resolving tied predictions, except to break ties randomly. It was observed by the author that even though the occurrence of tied predictions is rare for the See5 aggregate models, ties do occur. A sample of the output of the See5 combination algorithm is given in table 7.25 for the forest cover type un-boosted OVA aggregate model. Recall that an OVA base model predicts the one class it is designed to predict or it predicts the value 10 to represent 'other'. The instances in the first two rows are correctly predicted since there are no tied predictions with the highest score values. The third instance is incorrectly predicted as the tie between the class 1 and class 2 predictions cannot be correctly resolved.

Table 7.25: Sample of the output for the See5 combination algorithm

OVA1	score1	OVA2	score2	OVA3	score3	OVA4	score4	OVA5	score5	OVA6	score6	OVA7	score7	predicted	actual	score
1	0.91	2	0.85	10	0.99	10	1	10	1	10	1	10	0.91	1	**1**	0.91
1	0.91	10	0.85	10	0.99	10	1	10	1	10	1	10	1	1	**1**	0.91
1	0.91	2	0.91	10	0.99	10	1	10	1	10	1	10	1	1	**2**	0.91

7.5 Conclusions

The first question posed in this chapter was: *How should training datasets be designed in order to create base models that are syntactically diverse and highly expert at prediction for aggregate models?* The experimental results have demonstrated that the use of OVA modeling results in base models that are highly expert in predicting instances in specific regions of the instance space. However, the experimental results also demonstrated that expertise of base models, as measured

in terms of the predictive accuracy of the individual models, is not always enough to guarantee high predictive performance of the aggregate model.

The second question was: *How should training datasets for the base models be designed in order to achieve high accuracy for the aggregate model?* The experimental results demonstrated that one limiting factor for the predictive performance of aggregate models, created through parallel combination of the base model predictions, is the level of conflicting predictions for the base models. The experimental results for the 5NN algorithm demonstrated that the use of un-boosted OVA aggregate models results in performance improvements. Recall that the algorithm that was used for the combination of predictions for the 5NN base models has the ability to resolve conflicting predictions which are tied on the scores.

The experiments also demonstrated that when training datasets for base models are selected with the objective of minimising conflicting predictions, a high level of predictive performance may be realised. This was the case for the forest cover type 5NN and See5 boosted OVA aggregate models. For the experiments reported in this chapter, the minimisation of class confusion was realised through boosting which was achieved through the selection of training datasets that provide a high coverage of the confusion regions for the classes. It was demonstrated that boosting can result in improvements to predictive performance when OVA base models have conflicting predictions.

Further studies of the proposed training dataset selection method are reported in the context of pVn modeling in the next chapter.

Chapter 8

Evaluation of Dataset Selection for Positive-Versus-Negative Aggregate Modeling

It was stated in chapter 6 that the proposed methods of training dataset selection were aimed at supporting the creation of aggregate models for multi-class prediction tasks. The last chapter presented an evaluation of OVA modeling. This chapter presents the experiments to study training dataset selection for positive-Versus-negative (pVn) models, a discussion of pVn model performance, and a comparison of predictive performance of single, OVA and pVn aggregate models. Recall that each pVn base model specialises in the prediction of a subset of the classes (the p-classes). Also recall that the following two questions were posed in chapter 6, and answers to these questions were provided in chapter 7 for OVA classification:

1. How should training datasets be designed in order to create base models that are syntactically diverse and highly expert at prediction for aggregate models?

2. How should training datasets for the base models be designed in order to achieve high accuracy for the aggregate model?

This chapter presents further studies for the purpose of answering the above questions in the context of pVn modeling. Section 8.1 provides a discussion of pVn modeling. Experiments to study 5NN pVn model performance and See5 pVn model performance are respectively discussed in sections 8.2 and 8.3. Section 8.4 provides a discussion of the statistical tests used to compare the predictive coherence of single, OVA and pVn models. Sections 8.5 and 8.6 respectively provide discussions and conclusions for the chapter.

8.1 pVn modeling

The motivation for pVn modeling is presented in this section. The methods used for the design of pVn base models, and the creation and testing of pVn base models and pVn aggregate models are also discussed. Section 8.1.1 provides a discussion of the motivation for pVn modeling. The methods used to design the base models are discussed in section 8.1.2. The experiment design for the study of pVn modeling is presented in section 8.1.3.

8.1.1 Motivation for pVn modeling

pVn classification is a proposed modification of OVA classification. The initial motivation for using pVn base classifiers was given in chapter 6. Briefly stated, pVn modeling results in a reduction of the number of base models in comparison to OVA modeling. A further motivation for pVn modeling is as follows: The experimental results of chapter 7 demonstrated that there are datasets for which OVA base models do not result in aggregate models that provide a higher level of predictive accuracy. This is the case, for example, for the KDD Cup 1999 dataset where only the un-boosted 5NN model showed a small improvement in performance. It is useful to compare aggregate models based on OVA classification and with aggregate models based on pVn classification in order to establish whether pVn base models can result in predictive performance which is better than that of a single model which can predict any one of k ($k > 2$) classes.

8.1.2 Design of pVn base models

The following three questions need to be answered for pVn classification:

(1) What pVn models should be created?
(2) Which classes should be the positive classes, and which classes should be the negative classes for each pVn base model?
(3) What should be the class distribution for the positive and negative classes for the training dataset of each pVn base model?

An algorithm was designed by the author to provide answers to questions 1 and 2 above. The algorithm uses the information in the confusion matrix for the single *k*-class model to determine the number of models and the class composition of each pVn base model. This algorithm is presented in the next section. The methods used to answer question 3 above are also discussed in the next section. After the decisions have been made on the composition of the pVn base models, the training datasets must be selected. The selection process that was presented in chapter 6, and depicted in figure 6.2, was followed. The feature subset used for all pVn base models was the same as that for the single model, for both the 5NN and classification tree models.

8.1.3 Experiment design for the study of pVn modeling

Experiments were conducted to study the effectiveness of the proposed pVn base model design. The forest cover type and KDD Cup 1999 datasets were used for the experiments. The 5NN and See5 algorithms were used for the creation of the base models. The base models were combined into aggregate models using the combination algorithm in figure 6.3 (for See5 base models) and figure 6.4 (for 5NN base models). The analysis of pVn model performance was conducted as follows:

(1) To compare the predictive performance of the single and pVn aggregate models for both 5NN and See5 classification.

(2) To compare the predictive coherence of single, OVA, and pVn aggregate models for both 5NN and See5 classification.

Models were compared on predictive performance using the accuracy and class TPRATE measures as discussed in section 6.4.5. Student's paired t-test and the *Diff(A,S)* and *Ratio(A,S)* measures discussed in section 6.4.5 were used to establish whether the aggregate models provide significant improvements in predictive performance.

8.2 Experiments to study pVn models for 5NN classification

This section provides a discussion of the experiments on pVn classification for the 5NN algorithm. Section 8.2.1 presents the methods for base model design and training dataset selection. Sections 8.2.2 and 8.2.3 respectively provide a discussion of the experimental results for base model and aggregate model performance.

8.2.1 Design of training datasets for 5NN pVn base models

Several interesting observations arose out of the experiments on OVA modeling. The following observations can be made for the forest cover type 5NN OVA aggregate models, based on table 7.6. A training sample of 50% class 1, 25% class 2 and 25% class 7 was used for the boosted OVA1 base model. A training sample of 25% class 1, 25% class 2, and 50% class 7 was used for the boosted OVA7 base model. For both models, the main reason behind this decision was due to the fact that there is significant class confusion between classes 1, 2 and 7. A question that comes to mind is: *Would the performance of one base model, based on a sample with an equal class distribution for classes 1, 2, and 7 provided better performance than that of the two OVA base models, OVA1 and OVA7?* In fact, the other OVA base models could be similarly combined based on the observations made from the confusion matrix of table 7.5.

A structure that was referred to as a *confusion graph* was designed by the author for purposes of graphically representing the information in a confusion matrix. Figures 8.1 and 8.2 respectively show the confusion graphs for the forest cover type and KDD Cup 1999 5NN single *k*-class models that were presented in section 7.2. The nodes in a confusion graph represent the classes for the prediction task. The arc (c_i, c_j) means that class c_i is predicted as class c_j. That is, classes c_i and c_j share a confusion region. The number in brackets in a node indicates the number of arcs connected to the node. The value labelling an arc represents the level of confusion between classes c_i and c_j. This value comes from cell (c_i, c_j) of the confusion matrix. For simplicity of presentation, the arcs of the confusion graphs with values of 5 or less are shown as dashed lines and are not labeled.

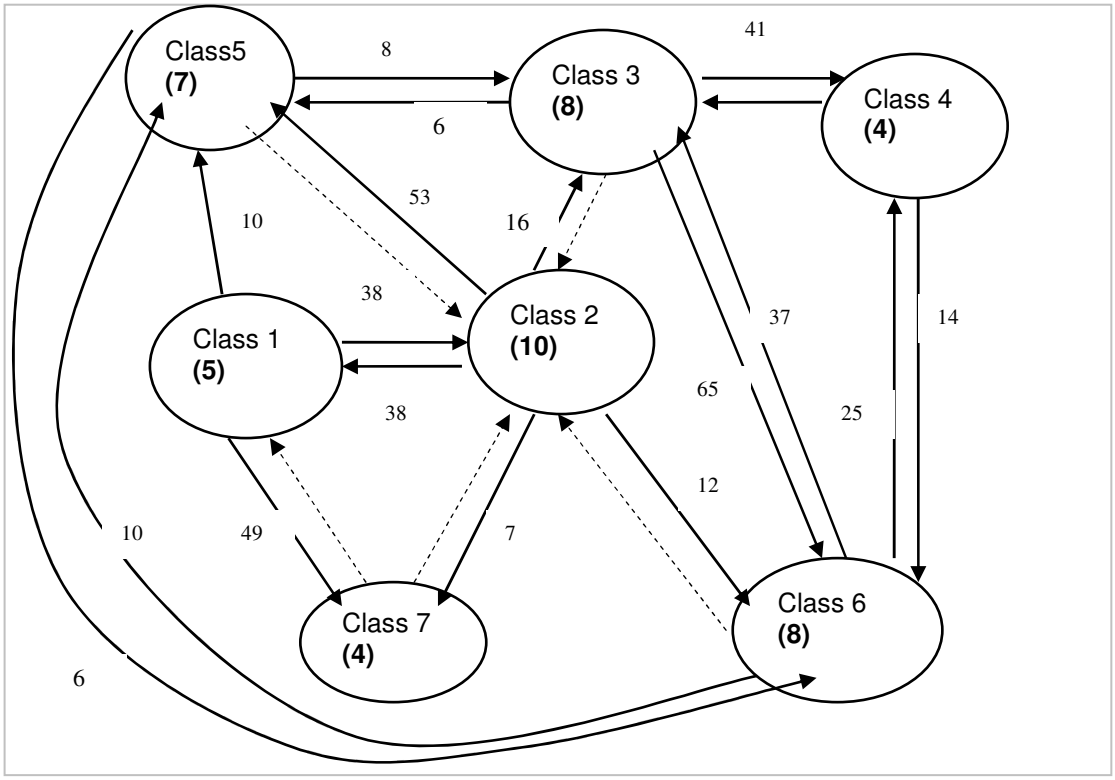

Figure 8.1: Confusion graph for the 5NN single 7-class model for Forest cover type for training set size of 12000 instances

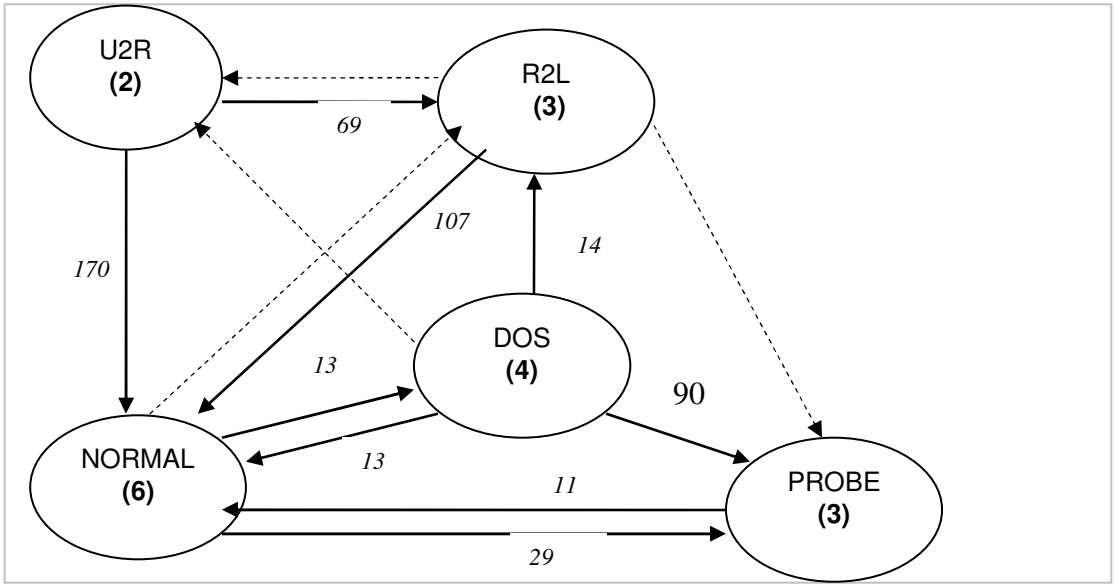

Figure 8.2: Confusion graph for the 5NN single 5-class model for KDD Cup 1999 for training set size of 4000 instances

The algorithm shown in figure 8.3 was designed by the author for selecting classes to include in each of the pVn base models. The objectives of the algorithm are as follows: When selecting the positive (p) classes for each base model, include those classes that share confusion regions. Exclude those classes that do not share

confusion regions with all the selected classes. The motivation here is to identify groups of classes which should be modelled together. Each model based on a subset of classes needs negative instances. The negative instances should be drawn from those classes that have a confusion region with at least one of the positive classes included in the model.

Table 8.1 provides a demonstration of the execution of the algorithm on the confusion graph of figure 8.1 for the forest cover type dataset. The last row of table 8.1 indicates that four pVn base models are identified by the algorithm. These models are: M346 for the positive classes 3, 4 and 6, M127 for the positive classes 1, 2 and 7, M125 for the positive classes 1, 2 and 5, and M2356 for the positive classes 2, 3, 5 and 6. The algorithm was also applied to the confusion graph for the KDD Cup 1999 single model shown in figure 8.2. The pVn base models that were identified are MNRU for the positive classes NORMAL, R2L and U2R, MNDR for the positive classes NORMAL, DOS and R2L, and MNDP for the positive classes NORMAL, DOS and PROBE.

A. Make a copy of the confusion graph

B. Repeat

Step 1: Select the node with the lowest connectivity and call it $c_{i,\min}$ (break ties randomly)

Step 2: Create a set whose members are $c_{i,\min}$ plus all the nodes connected to the node $c_{i,\min}$

Step 3: Delete all the arcs connected to $c_{i,\min}$ and delete $c_{i,\min}$ from the graph

Until there are no arcs left in the graph

C. For each set created in step 2 of B above, if the set is a subset of another set, delete the set (since its nodes are already included in another set)

D. Assign each of the remaining sets as the positive classes for one model

E. For each model, determine the negative classes as follows: For each positive class in the model, add all classes that have an arc to the class node, but are not one of the positive classes of the model.

Figure 8.3: Algorithm for class selection for the pVn base models

Table 8.2 shows the training set composition that was used to study the predictive performance of the pVn base models identified by the algorithm in figure 8.3 for the forest cover type and KDD Cup 1999 datasets. Each base model was composed of

instances from the indicated classes and sample percentages for each class. The rationale behind the samples composition was to ensure that each of the positive classes has nearly the same number of instances as the other positive classes, and nearly the same number of instances as all the negative classes combined. The training sample size for the MNRU model was reduced to 1900 instances to avoid excessive bootstrapping of the U2R instances.

Table 8.1: Trace of the class selection algorithm for the5NN forest cover type graph

Iterations for steps B1, B2 and B3			
Iteration	B1: selected node	B2: created set	B3: deleted arcs and node
1	$c_{i,\min} = 4$	class set: { 3, 4, 6 }	arcs: { 3→4, 4→3, 6→4, 4→6 } node: 4
2	$c_{i,\min} = 7$	class set: {1, 2, 7 }	arcs: { 1→7, 7→1, 2→7, 7→2 } node: 7
3	$c_{i,\min} = 1$	class set: {1, 2, 5 }	arcs: { 1→5, 1→2, 2→1} node: 1
4	$c_{i,\min} = 5$	class set: { 2, 3, 5, 6 }	arcs: { 2→5, 5→2, 3→5, 5→3, 6→5, 5→6 } node:5
5	$c_{i,\min} = 6$	class set: { 2,3,6 }	arcs: { 2→6, 6→2, 3→6, 6→3 } node: 6
6	$c_{i,\min} = 3$	class set: { 2, 3 }	arcs: { 2→3, 3→2 } node: 3
Final results of iterations of steps B1, B2, B3:		{ { 3, 4, 6 }, {1, 2, 7 }, {1, 2, 5 }, { 2, 3, 5, 6 }, { 2,3,6 }, { 2, 3 } }	
Steps C, D and E			
Step	Action	Results	
C	Delete subsets of other sets	deleted sets: {2,3} and {2,3,6} remaining sets: { { 3, 4, 6 }, {1, 2, 7 }, {1, 2, 5 }, { 2, 3, 5, 6 } }	
D	Assign positive classes	M346: positive classes = { 3, 4, 6 } M127: positive classes = { 1, 2, 7 } M125: positive classes = { 1, 2, 5 } M2356: positive classes = { 2, 3, 5, 6 }	
E	Determine negative classes	M346: negative classes = { 2, 5 } 2 borders with 3, 5 borders with 3 and 6	
		M127: negative classes: = { 3, 5, 6 } 3 borders with 2, 5 borders with 1 and 2, 6 borders with 2	
		M125: negative classes = {3, 6, 7} 3 borders with 2, 6 borders with 5, 7 borders with 1 and 2	
		M2356: negative classes = {1, 4, 7} 1 borders with 2, 4 borders with 3 and 6, 7 borders with 2 (but confusion level is very low)	
Algorithm output	Model definitions	M346: positive classes = { 3, 4, 6 }; negative classes = { 2, 5 } M127: positive classes = {1, 2, 7 }; negative classes= { 3, 5, 6 } M125: positive classes = {1, 2, 5 }; negative classes = {3, 6, 7 } M2356: positive classes = { 2, 3, 5, 6 }; negative classes= { 1, 4} class 7 ignored	

Table 8.2: 5NN training set composition for the pVn base models for forest cover type and KDD Cup 1999

Dataset	Model ID	p (positive) classes		n (negative classes)		Training sample size
		Classes used	sample percentage	classes used	sample percentage	
Forest cover type	M125	C1,C2,C5	80: (27,27,26)	C3,C6,C7	20: (7,7,6)	12000
	M127	C1,C2,C7	80: (27,27,26)	C3,C5,C6	20: (7,7,6)	
	M2356	C2,C3,C5,C6	80: (20,20,20,20)	C1,C4	20: (10,10)	
	M346	C2,C3,C6	80: (27,27,26)	C2,C5	20: (1,10)	
KDD Cup 1999	MNRU	NORMAL, R2L,U2R	80: (27,27,26)	DOS, PROBE	20: (10,10)	1900
	MNDR	NORMAL,DOS, R2L	80: (27,27,26)	PROBE, U2R	20: (10,10)	4000
	MNDP	NORMAL,DOS, PROBE	80: (27,27,26)	R2L, U2R	20: (10,10)	4000

8.2.2 Predictive performance of the 5NN pVn base models

The performance of the 5NN pVn base models for the forest cover type and KDD Cup 1999 dataset is shown in table 8.3. Columns 3 and 4 of table 8.3 show the mean TPRATE and mean TNRATE values for the base models. The TPRATE in this context is the predictive accuracy on the test instances for the p-classes while the TNRATE is the predictive accuracy on the test instances for the n-classes.

Table 8.3: Predictive performance of 5NN pVn base models

Dataset (Training size) (test size)	Base model ID	Base model performance		single model performance
		Mean TPRATE% (p instances)	Mean TNRATE% (n instances)	Mean TPRATE% for single model on p instances
Forest cover type (12000) (350 x 10)	M125	75.3 ± 2.3	85.1 ± 1.2	67.3 ± 7.3
	M127	74.4 ± 1.4	91.6 ± 0.7	66.9 ± 7.2
	M2356	57.9 ± 0.5	70.8 ± 3.5	67.1 ± 6.7
	M346	81.3 ± 1.5	94.1 ± 0.7	72.3 ± 5.6
KDD Cup 1999 (4000) (350 x 10)	MNRU	76.3 ±0.8	97.4 ±1.0	60.2 ± 8.1
	MNDR	88.8 ±1.7	71.1 ±1.1	71.8 ± 3.9
	MNDP	74.4 ±5.5	68.9 ±7.4	82.1 ± 4.9

Column 5 of table 8.3 shows the mean TPRATE values for the single 7-class model for forest cover type and the single 5-class model for KDD Cup 1999. The results of table 8.3 indicate that three out of four pVn models for forest cover type have a higher TPRATE value on the p-classes than for the single model. Two out of three pVn models for the KDD Cup 1999 dataset have a higher TPRATE value than the single 5-class model. It remains to be seen whether the aggregate model based on

these base models provide higher predictive performance compared to the single models.

8.2.3 Predictive performance of the 5NN pVn aggregate models

The pVn base models for forest cover type and KDD Cup 1999 were combined into aggregate models using the combination algorithm that was given in figure 6.4 of section 6.4.3. The experimental procedure that was used for aggregation was presented in section 6.4. Table 8.4 shows the results of the predictive performance of the 5NN aggregate model for the forest cover type pVn models. The details of predictive performance are given in table F.4. Table 8.4 also shows the results of the predictive performance of the 5NN single 7-class and OVA aggregate models of chapter 7 for the forest cover type dataset. Table 8.5 shows the results of the statistical tests used to compare the performance of the single 7-class model and the pVn aggregate model.

Table 8.4: Mean Predictive performance of the 5NN single, OVA and pVn aggregate models for forest cover type

Class name	5NN Mean accuracy / TPRATE% (10 test sets of size 350)			
	Single model	un-boosted OVA aggregate model	boosted OVA aggregate model	pVn aggregate model
All classes	74.7 ± 1.0	80.5 ± 0.9	82.0 ± 0.6	78.6 ± 1.2
1	62.8 ± 3.4	70.0 ± 4.3	70.0 ± 4.3	67.8 ± 5.1
2	48.8 ± 2.8	58.4 ± 2.7	62.0 ± 3.4	57.8 ± 2.1
3	56.8 ± 4.1	71.8 ± 1.9	71.0 ± 1.3	65.0 ± 2.3
4	92.4 ± 1.8	89.8 ± 1.9	100.0 ± 0.0	97.0 ± 1.2
5	91.2 ± 2.0	95.8 ± 3.1	97.0 ± 0.9	94.2 ± 2.1
6	75.0 ± 2.1	80.8 ± 4.5	77.6 ± 2.0	75.0 ± 2.9
7	96.0 ± 1.3	96.6 ± 0.6	96.6 ± 0.6	93.2 ± 2.4

The results of Student's paired t-test and the *Diff(A,S)* and *Ratio(A,S)* performance improvement measures provide the following evidence: The pVn aggregate model has a higher level of performance compared to the single model. The aggregate model results in an accuracy increase of 3.9% for all classes combined. The *Diff(A,S)* measure indicates that the pVn aggregate model provides significant increases of 3.0% to 9 % on the TPRATE for four out of seven classes, namely classes 2, 3, 4 and 5. The *Ratio(A,S)* measure indicates increases between 0.2 and 0.6 for classes 2, 3, 4 and 5. However, for classes 1, 6 and 7 there are no statistically significant improvements in the TPRATE. The best 5NN OVA aggregate model for forest cover

type reported in chapter 7 provided a mean accuracy of 82.0±0.6 as shown in table 8.4. The mean accuracy of the pVn aggregate model was 78.6±1.2. This leads to the conclusion that both the OVA and pVn aggregate models can provide improvements in predictive performance for the forest cover type dataset.

Table 8.5: Statistical tests to compare the performance for 5NN single and pVn aggregate models for forest cover type

Group names and mean accuracy / TPRATE% for 10 test samples		Student's paired t-test (9 df)			Performance improvement measures	
Group A Aggregate model	Group S Single model	95% CI of mean difference	P value (2 tail)	Group A better than Group S?	*Diff(A,S)%*	*Ratio(A,S)*
All classes-A (78.6±1.2)	All classes-S (74.7 ± 1.0)	[2.5, 5.2]	0.000	yes	3.9	0.2
Class1-A (67.8±5.1)	Class1-S (62.8 ± 3.4)	[-2.1, 12.1]	0.146	no	5.0	0.1
Class2-A (57.8±2.1)	Class2-S (48.8 ± 2.8)	[4.9, 13.1]	0.001	yes	9.0	0.2
Class3-A (65.0±2.3)	Class3-S (56.8 ± 4.1)	[3.7, 12.8]	0.003	yes	8.2	0.2
Class4-A (97.0±1.2)	Class4-S (92.4 ± 1.8)	[3.1, 6.1]	0.000	yes	4.6	0.6
Class5-A (94.2±2.1)	Class5-S (91.2 ± 2.0)	[1.5, 4.6]	0.002	yes	3.0	0.3
Class6-A (75.0±2.9)	Class6-S (75.0 ± 2.1)	[-1.8, 1.8]	1.000	no	0.0	0.0
Class7-A (93.2±2.4)	Class7-S (96.0 ± 1.3)	[-5.3, -0.4]	0.029	no	-2.8	-0.7

Table 8.6 shows the results of the Predictive performance of the 5NN pVn aggregate model for the KDD Cup 1999 dataset. The detailed results are given in the appendix table F.12. The results for the single 5-class and OVA aggregate models of chapter 7 are also shown in table 8.6. Table 8.7 shows the results of the statistical tests used to compare the predictive performance of the single 5-class model and the pVn aggregate model. The results of Student's paired samples t-test clearly indicate that the pVn aggregate model performance is much higher than that of the single 5-class model. The pVn model provided an increase of 11.8% in the mean accuracy for all the classes. The *Ratio(A,S)* measure indicates an increase of 0.4. The *Diff(A,S)* measure indicates an increase in the TPRATE ranging between 2.7% and 31% for four out of five classes. The *Ratio(A,S)* measure indicates high increases of between 0.5 and 0.9.

Table 8.6: Mean Predictive performance of single, OVA and pVn aggregate 5NN models for KDD Cup 1999

Class name	5NN Mean accuracy / TPRATE% for 10 test sets of size 350			
	Single model	un-boosted OVA aggregate model	boosted OVA aggregate model	pVn aggregate model
All classes	68.5 ± 1.4	72.4 ± 1.1	71.0 ± 1.2	80.3 ± 1.1
NORMAL	84.4 ± 3.1	92.7 ± 2.8	92.4 ± 3.0	98.7 ± 0.9
DOS	66.3 ± 5.0	66.0 ± 4.4	66.0 ± 5.1	97.3 ± 1.7
PROBE	95.7 ± 1.2	95.2 ± 1.0	95.4 ± 1.2	98.4 ± 0.9
R2L	64.7 ± 3.6	65.4 ± 3.6	60.9 ± 3.8	81.4 ± 4.1
U2R	31.6 ± 0.3	42.6 ± 0.4	40.5 ± 1.4	25.7 ± 2.2

Table 8.7 Statistical tests to compare the 5NN single and pVn aggregate models for KDD Cup 1999

Group name and mean accuracy / TPRATE% for 10 test sets		Student's paired t-test (9 df)			Performance improvement measures	
Group A Aggregate model	Group S Single model	95% CI of mean difference	p value (2 tail)	Group A better than Group S?	*Diff(A,S)%*	*Ratio(A,S)*
All classes-A (80.3 ±1.1)	All classes-S (68.5 ±1.4)	[10.1,13.4]	0.000	yes	11.8	0.4
NORMAL-A (98.7 ±0.9)	NORMAL-S (84.4 ±3.1)	[10.9,17.8]	0.000	yes	14.3	0.9
DOS-A (97.3 ± 1.7)	DOS-S (66.3 ±5.0)	[25.5,36.6]	0.000	yes	31.0	0.9
PROBE-A (98.4 ±0.9)	PROBE-S (95.7 ±1.2)	[1.2,4.2]	0.002	yes	2.7	0.6
R2L-A (81.4 ±4.1)	R2L-S (64.7 ±3.6)	[13.1,20.3]	0.000	yes	16.7	0.5
U2R-A (25.7 ±2.2)	U2R-S (31.6 ±0.3)	[-8.4,-3.3]	0.001	no	-5.9	-0.1

In comparison to the KDD Cup 1999 OVA aggregate models of chapter 7, the best OVA aggregate model had a mean predictive accuracy of 72.4±1.1 as shown in table 8.6, while the pVn aggregate model has a mean predictive accuracy of 80.3±1.1. This comparison indicates that the pVn aggregate model has a much higher level of predictive performance. The foregoing observations provide evidence that pVn aggregate modeling can provide much higher performance improvements than OVA modeling.

8.3 Experiments to study pVn models for See5 classification

pVn aggregate modeling was also tested using the See5 classification tree algorithm. A discussion of the experiments and the predictive performance of the See5 base models and aggregate models for the forest cover type and KDD Cup 19999

datasets are provided in this section. The training dataset design for the base models is presented in section 8.3.1. The predictive performance results for the base models and aggregate models are respectively presented in sections 8.3.2 and 8.3.3.

8.3.1 Design of training datasets for pVn base models

The confusion graphs for the forest cover type and KDD Cup 1999 See5 single models are shown in figures 8.4 and 8.5 respectively. The algorithm in figure 8.3 was used to determine the class composition of the pVn classification tree models for both the forest cover type and the KDD Cup 1999 datasets. It became evident that the algorithm in figure 8.3 is not suitable for determining the class composition for the KDD Cup 1999 dataset because the confusion graph for the KDD Cup 1999 See5 single model is a maximally connected (fully interconnected) graph. When a maximally connected confusion graph is used as input to the algorithm of figure 8.3, the first iteration of step B will create a set of nodes which includes all the nodes in the graph. When step C is executed, all the sets of nodes created after the first iteration of step B will be deleted, since they will be subsets of the first set of nodes. A modification of the algorithm in figure 8.3 is given in figure 8.4. The motivation for the modification was to reduce the level of connectivity in the graph while at the same time retaining all the information about the regions with the highest levels of class confusion. The rationale behind step I of the algorithm in figure 8.4 is to ignore those regions that have low levels of confusion and favour those regions which have higher levels of class confusion.

The application of step I of the algorithm in figure 8.4 to the confusion graph for the KDD Cup 1999 dataset resulted in the confusion graph of figure 8.5. The algorithm in figure 8.3 was applied to the confusion graph for the forest cover type dataset. The modified algorithm in figure 8.6 was applied to the confusion graph for the KDD Cup 1999 dataset. The resulting pVn base model designs are shown in table 8.8. Column 2 of table 8.8 shows the names of the pVn models. Each model is identified by the positive classes it is designed to predict. The training sample composition for each pVn base model is also shown in table 8.8. The training sample sizes for the MNRU and MNPU base models were reduced to 1900 instances to avoid excessive bootstrapping of the U2R instances.

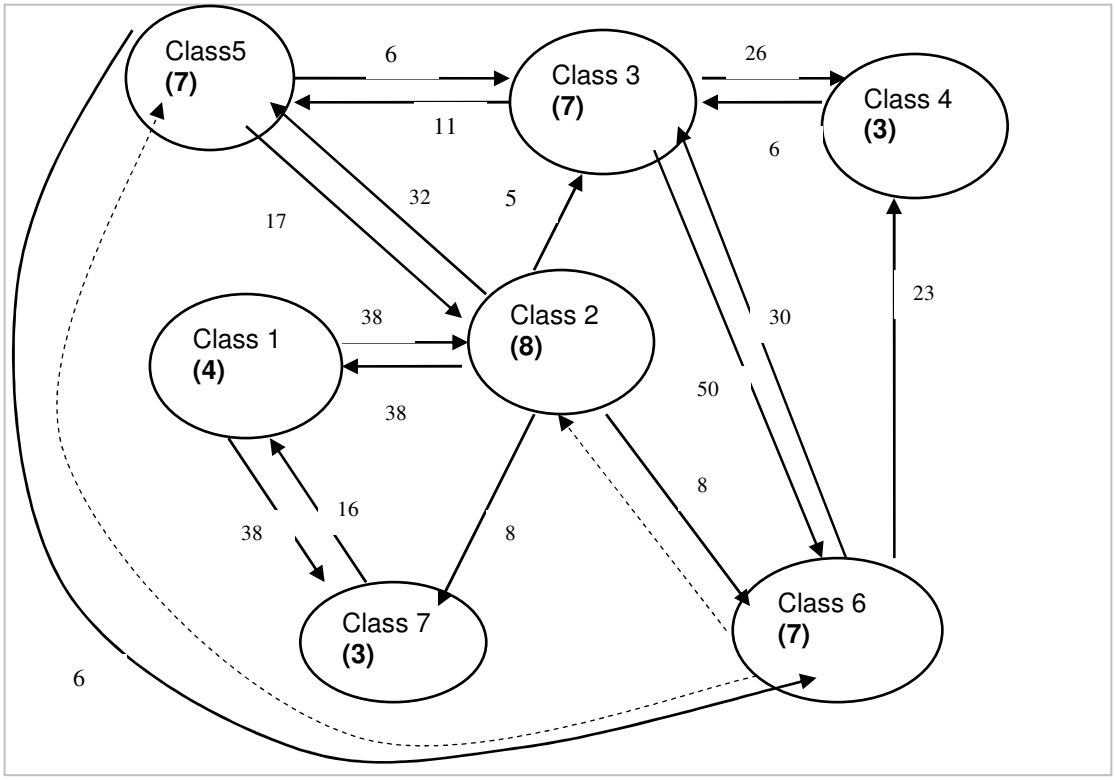

Figure 8.4 Confusion graph for theSee5 single 7-class model for forest cover type for training set size of 12000 instances

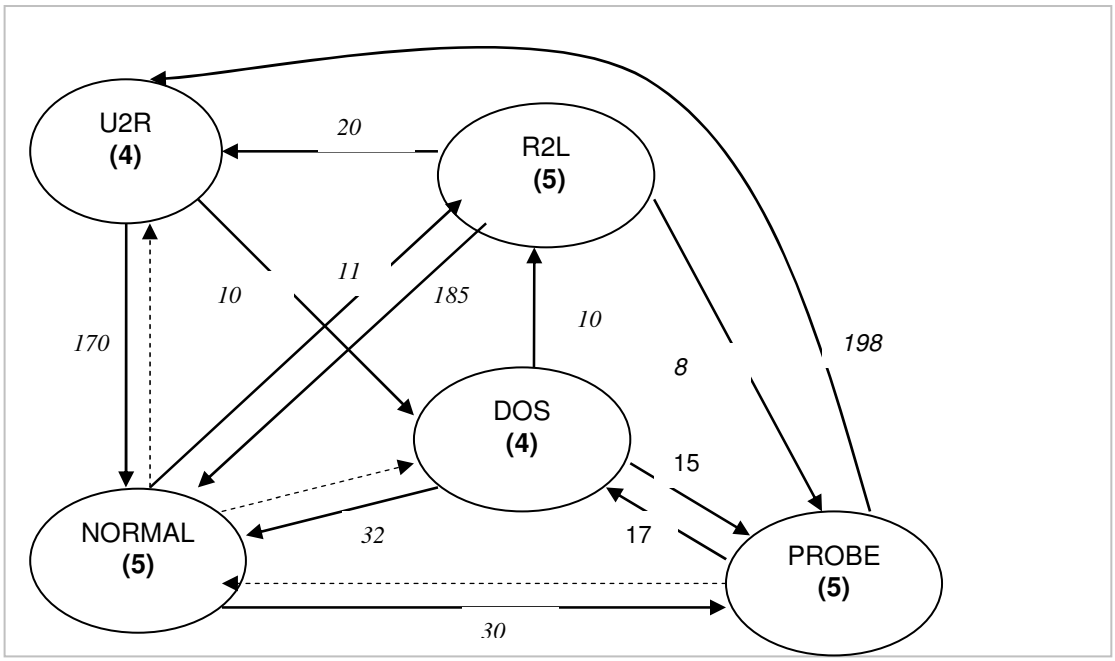

Figure 8.5: Confusion graph for the See5 single 5-class model for KDD Cup 1999 for training set size of 4000 instances

I. Analyse the confusion graph as follows:

If each node is fully connected to all the other nodes then

for each node

delete the weakest outgoing link (the outgoing arc with the smallest weight)

end-for

II. Process the confusion graph as follows:

A. Make a copy of the confusion graph

B. Repeat

Step 1: Select node with the lowest connectivity and call it $c_{i,\min}$ (break ties randomly)

Step 2: Create a set whose member are $c_{i,\min}$ plus all the nodes connected to the node $c_{i,\min}$

Step 3: Delete all the arcs connected to $c_{i,\min}$ and delete $c_{i,\min}$ from the graph

Until there are no arcs left in the graph

C. For each set of nodes created in step 2 of B above, if the set is a proper subset of another set, delete the set.

D. Assign each of the remaining sets as the positive classes for one model.

E. For each model, determine the negative classes. For each positive class in the model, add all classes that have an arc to the class node, but are not one of the positive classes for the model.

Figure 8.6: Modified algorithm for class selection for the pVn base models

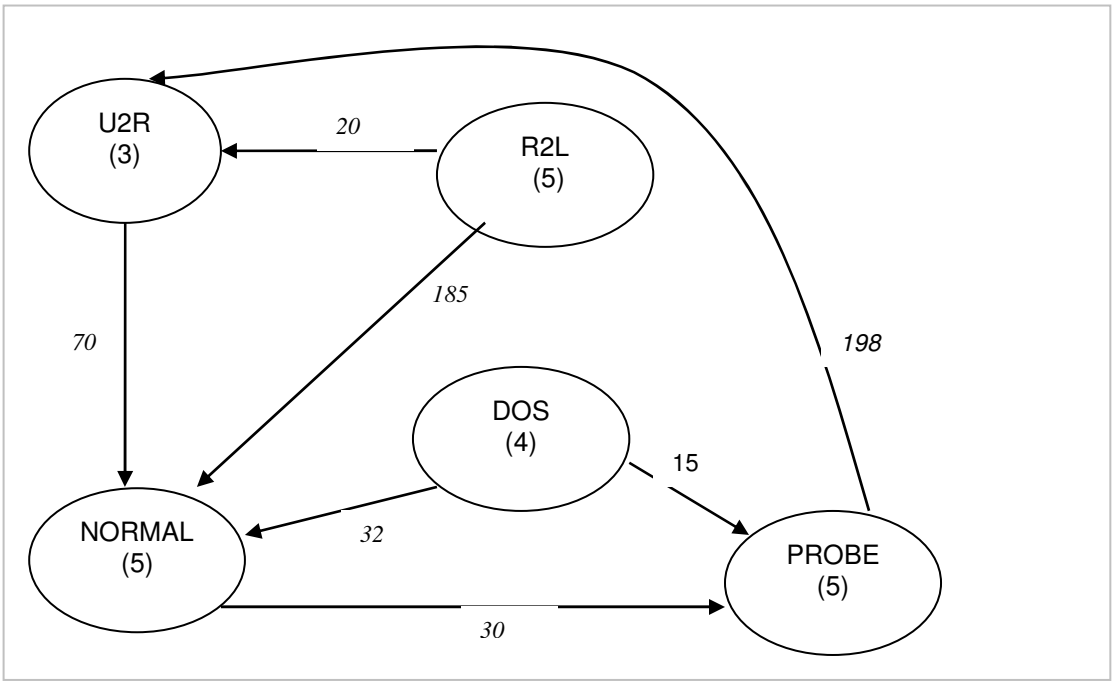

Figure 8.7: Simplified confusion graph for the See5 single 5-class model for KDD Cup 1999

Table 8.8: Training set composition for theSee5 pVn base models

Dataset	Model ID	p (positive) classes		n (negative classed)		Training sample size
		Classes used	sample percentage	classes used	sample percentage	
Forest cover type	M127	C1,C2,C7	80: (27,27,26)	C3,C5,C6	20: (7,7,6)	12000
	M2356	C2,C3,C5,C6	80: (20,20,20,20)	C1,C4	20: (10,10)	
	M346	C2,C3,C6	80: (27,27,26)	C2,C5	20: (7,7,6)	
KDD Cup 1999	MNRU	NORMAL, R2L,U2R	80: (27,27,26)	DOS, PROBE	20: (10,10)	1900
	MNDP	NORMAL, DOS, PROBE	80: (27,27,26)	R2L, U2R	20: (10,10)	4000
	MNPU	NORMAL, PROBE, U2R	80: (27,27,26)	DOS,R2L	20: (10,10)	1900

8.3.2 Predictive performance of the See5 pVn base models

The performance of the See5 pVn base models for the forest cover type and KDD Cup 1999 dataset are shown in table 8.9. Columns 3 and 4 of table 8.9 show the mean TPRATE and mean TNRATE values for the base models. The TPRATE in this context is the predictive accuracy on the p-classes while the TNRATE is the predictive accuracy on the n-classes.

Table 8.9: Predictive performance of See5 pVn base models

Dataset (Training sample size)	Base model ID	Base model performance		single model performance
		Mean TPRATE% (p instances)	Mean TNRATE% (n instances)	Mean TPRATE% (p instances)
Forest cover type (12000)	M127	76.7 ± 1.5	89.9 ± 0.9	72.3 ± 1.4
	M2356	76.8 ± 1.3	81.5 ± 2.0	72.2 ± 1.7
	M346	82.3 ± 0.9	96.9 ± 0.6	78.5 ± 1.8
KDD Cup 1999 (4000)	MNRU	77.4 ± 2.6	84.7 ± 3.2	67.0 ± 1.6
	MNDP	91.1 ± 1.9	63.9 ± 1.3	68.1 ± 1.7
	MNPU	74.8 ± 0.4	77.3 ± 1.4	66.5 ± 1.3

Column 5 of table 8.9 shows the mean TPRATE values for the single 7-class model on the p-classes for forest cover type, and the single 5-class model for KDD Cup 1999. The results in table 8.9 indicate that the pVn base models M127, M2356 and M346 for forest cover type each have higher TPRATE values on their p-classes compared to the single 7-class model on the same classes. The pVn models MNRU, MNDP and MNPU for the KDD Cup 1999 dataset also have significantly higher TPRATE values on their p-classes compared to the single 5-class model.

8.3.3 Predictive performance of the See5 pVn aggregate models

The pVn base models for the forest cover type and KDD Cup 1999 datasets were combined into aggregate models using the algorithm in figure 6.3. Table 8.10 shows the results of the predictive performance of the See5 pVn aggregate model for the forest cover type dataset. The detailed performance results are given in the appendix table F.8. Table 8.10 also gives the performance results for the single 7-class and OVA aggregate models of chapter 7. Table 8.11 shows the results of the statistical tests used to compare the performance of the single 7-class model and the pVn aggregate model.

Table 8.10: Predictive performance of the See5 single, OVA and pVn models for forest cover type

Class name	See5 Mean accuracy / TPRATE%			
	Single model	un-boosted OVA aggregate model	boosted OVA aggregate model	pVn aggregate model
All classes	76.9 ± 1.0	75.3 ± 0.7	79.4 ± 0.6	79.9 ± 1.0
1	57.4 ± 3.4	60.6 ± 2.6	65.0 ± 2.9	64.6 ± 2.9
2	63.8 ± 3.0	49.8 ± 3.6	69.8 ± 2.4	65.5 ± 4.2
3	60.8 ± 3.3	64.0 ± 1.8	63.2 ± 3.3	71.8 ± 3.3
4	96.8 ± 1.0	86.6 ± 1.7	95.4 ± 1.3	94.6 ± 1.7
5	86.2 ± 2.4	94.4 ± 1.8	88.4 ± 2.3	88.6 ± 1.8
6	77.8 ± 3.3	79.2 ± 2.0	76.0 ± 1.9	82.2 ± 2.6
7	95.6 ± 1.6	92.8 ± 2.5	97.8 ± 1.1	92.0 ± 2.8

The results of Student's paired sample t-test and the *Diff(A,S)* and *Ratio(A,S)* performance improvement measures provide the following evidence: The pVn aggregate model has a significantly higher level of performance compared to the single model. The aggregate model results in an increase of 3% in accuracy for all classes combined. For the TPRATE values of the individual classes, the aggregate model provides a significantly higher level of performance with an increase in the TPRATE of 11% on class 3 and 7.2% on class 1. The aggregate model provided a performance improvement of 4.4% in theTPRATE for class 6. However, there is no statistically significant improvement in the TPRATE values for the remaining four classes. In fact, the single model provided higher TPRATE values on two of these classes. The best See5 OVA aggregate model for the forest cover type dataset that was reported in chapter 7 provided a mean accuracy of 79.4 ± 0.6 as shown in table 8.10. The mean accuracy of the pVn aggregate model was 79.9 ± 1.0. This leads to the conclusion that both the OVA and pVn aggregate models can provide a comparable improvement in Predictive performance for the forest cover type dataset.

Table 8.11: Statistical tests to compare the performance for See5 classification tree single and pVn aggregate models for forest cover type

Group mean accuracy / TPRATE% for 10 test sets		Student's paired t-test (9 df)			Performance improvement measures	
Group A Aggregate model	Group S Single model	95% CI of mean difference	P value (2 tail)	Group A better than Group S?	*Diff(A,S) %*	*Ratio(A,S)*
All classes-A (79.9 ± 1.0)	All classes-S (76.9 ± 1.0)	[1.9, 4.0]	0.000	yes	3.0	0.1
Class1-A (64.6 ± 2.9)	Class1-S (57.4 ± 3.4)	[2.9,11.5]	0.004	yes	7.2	0.2
Class2-A (65.2 ± 4.2)	Class2-S (63.8 ± 3.0)	[-1.7, 4.5]	0.334	no	1.4	0.0
Class3-A (71.8 ± 3.3)	Class3-S (60.8 ± 3.3)	[6.8, 15.2]	0.000	yes	11.0	0.3
Class4-A (94.6 ± 1.7)	Class4-S (96.8 ± 1.0)	[-4.6, 0.2]	0.066	no	-2.2	-0.7
Class5-A (88.6 ± 1.8)	Class5-S (86.2 ± 2.4)	[-1.4, 6.2]	0.188	no	2.4	0.2
Class6-A (82.2 ± 2.6)	Class6-S (77.8 ± 3.3)	[1.2, 7.6]	0.014	yes	4.4	0.2
Class7-A (92.0 ± 2.8)	Class7-S (95.6 ± 1.6)	[-5.5, -1.7]	0.002	no	-3.6	-0.8

Table 8.12 shows the results of the Predictive performance of the pVn aggregate model for the KDD Cup 1999 dataset. The performance details are given in the appendix table F.16. The results for the single 5-class and OVA aggregate models are also shown in table 8.12. Table 8.13 shows the results of the statistical tests to compare the predictive performance of the single 5-class model and the pVn aggregate model. The results of Student's paired t-test clearly indicate that the aggregate model performance is much higher than that of the single 5-class model. The *Diff(A,S)* and *Ratio(A,S)* measures indicate that the increase in the TPRATE for three of the classes is between 12.1% and 60.6%. The TPRATE increase for the PROBE class is 60.6%, which is remarkably high. Overall, the accuracy increase over all the classes is 15.2%.

In comparison to the KDD Cup 1999 OVA aggregate model of chapter 7, the best See5 OVA aggregate model had a mean predictive accuracy of 61.7±0.9 as shown in table 8.12, while the pVn aggregate model has a mean predictive accuracy of 79.0±2.1. This comparison clearly indicates that the pVn aggregate model has a much higher level of predictive performance. Again, this provides evidence that pVn aggregate modeling can provide much higher performance gains compared to OVA modeling.

Table 8.12: Predictive performance of See5 single, OVA and pVn aggregate models for KDD Cup 1999

Class name	See5 Mean accuracy / TPRATE (10 test sets of size 350)			
	Single model	un-boosted OVA aggregate model	boosted OVA aggregate model	pVn aggregate model
All classes	63.8 ± 1.3	63.3 ± 1.2	61.7 ± 0.9	79.0 ± 2.1
NORMAL	86.0 ± 3.1	98.3 ± 0.7	99.2 ± 0.6	98.1 ± 0.6
DOS	82.0 ± 3.8	50.1 ± 4.4	56.3 ± 4.3	68.4 ± 6.5
PROBE	36.8 ± 2.4	88.0 ± 1.3	89.3 ± 1.4	97.0 ± 1.0
R2L	37.7 ± 3.3	34.3 ± 3.3	23.6 ± 3.4	54.1 ± 6.9
U2R	77.1 ± 0.0	45.7 ± 0.0	40.0 ± 0.0	77.1 ± 0.0

Table 8.13 Statistical tests to compareSee5 single and pVn aggregate models for KDD Cup 1999

Group name and mean TPRAE% for 10 test samples		Student's paired t-test (9 df)			Performance improvement measures	
Group A Aggregate model	Group S Single model	95% CI of mean difference	p value (2 tail)	Group A better than Group S?	*Diff(A,S)%*	*Ratio(A,S)*
All classes-A (79.0 ± 2.1)	All classes-S (63.8 ± 1.3)	[12.8,17.5]	0.000	yes	15.2	0.4
NORMAL-A (98.1 ± 0.6)	NORMAL-S (86.0 ± 3.1)	[8.7,15.6]	0.000	yes	12.1	0.9
DOS-A (68.4 ± 6.5)	DOS-S (82.0 ± 3.8)	[-18.6,8.6]	0.000	no	-13.6	-0.8
PROBE-A (97.0 ± 1.0)	PROBE-S (36.4 ± 2.4)	[60.0,63.5]	0.000	yes	60.6	0.97
R2L-A (54.1 ± 6.9)	R2L-S 37.7 ± 3.3)	[9.5,23.2]	0.000	yes	16.4	0.3
U2R-A (77.1 ± 0.0)	U2R-S (77.1 ± 0.0)	no variance	no variance	same	0.0	0.0

8.4 Comparison of performance variability for single and aggregate models

Given two systems or methods, the system or method with more predictable behaviour should be preferred (Cohen, 1995:pg 205). In the context of predictive modeling, the method with predictive performance which has lower variability should be preferred to one which exhibits erratic performance. A model with low performance variability has more predictable behaviour. The F-test for variances which was discussed in chapter 4, was used to test the null hypothesis that the variance of predictive accuracy for a single *k*-class model is the same as that for the OVA or pVn aggregate model. There are two available rules for the rejection of the null hypothesis for the 2-tail F-test. The first rule states that the null hypothesis should be rejected if the p-value for the test is less than the critical p-value. The second rule states that the null hypothesis should be rejected if the value of the F-statistic is greater or equal to the critical value of the F-statistic. The second rule was

used for the F-test inference given in table 8.14. The results of the F-tests indicate that, in general, there is no significant difference in performance variability between the single *k*-class models and OVA aggregate models, and between single *k*-class models and pVn aggregate models. This leads to the conclusion that both the single and aggregate models exhibit equal predictive coherence.

Table 8.14: F- tests for comparison of performance variability for single and aggregate models

Dataset	Algorithm	Variance of predictive accuracy		F-test for variance of accuracy on 10 test sets (9 x 9 df, $F_{critical} = 3.18$)		
		Single model (S)	Aggregate model (A)	F value = *Max{VarA ,VarS} / Min{VarA,VarS}*	p-value (F ≤ f) 1-tail	A has same coherence as S?
forest cover type	5NN	single (2.9)	un-boosted OVA (2.3)	1.26	0.37	yes
			boosted OVA (0.9)	3.13	0.05	yes
			pVn (3.8)	1.34	0.33	yes
	See5	single (2.5)	un-boosted OVA (1.1)	2.17	0.13	yes
			boosted OVA (0.8)	2.95	0.06	yes
			pVn (2.4)	1.01	0.49	yes
KDD Cup 1999	5NN	single (4.9)	un-boosted OVA (3.2)	1.53	0.27	yes
			boosted OVA (3.9)	1.28	0.36	yes
			pVn (3.4)	1.47	0.29	yes
	See5	single (4.7)	un-boosted OVA (3.8)	1.24	0.38	yes
			boosted OVA (2.0)	2.42	0.10	yes
			pVn (11.9)	2.52	0.09	yes

The following general conclusions can be made from the statistical tests of chapter 7 for the comparison means and the statistical tests of this chapter for the comparison of means and comparison of variances: Both OVA and pVn aggregate models provided a higher level of predictive performance compared to a single 7-class model for the forest cover type dataset. The single and aggregate models exhibited similar levels of predictive coherence, so that overall the aggregate models should be preferred to the single 7-class model. The pVn aggregate model provided a higher level of predictive performance compared to a single 5-class model for the KDD Cup 1999 dataset. The level of predictive coherence is similar for the single and pVn aggregate model, so that the aggregate model should be preferred.

It should be emphasized that the variance shown in table 8.14 is not the same as the variance component of the prediction error. Recall from section 2.8 that variance error is defined as variability in prediction of an instance ***x*** from one training sample to the next. For a given algorithm and modeling method, the measurement of variance error requires the creation of many models each based on a different training sample. The variance error is then estimated using the same test set for the different models (Kohavi & Wolpert, 1996).

8.5 Discussion

The benefits of pVn modeling are summarised in this section. The performance of OVA and pVn models is compared to the performance of single models. The limitations of the proposed methods for training dataset selection are discussed. Section 8.5.1 provides a summary of the benefits of pVn modeling. Section 8.5.2 presents a comparison of OVA and pVn modeling. Section 8.5.3 discusses the limitations of the proposed dataset selection methods.

8.5.1 Dataset selection for pVn modeling

pVn modeling was proposed as a method of problem decomposition with a potential to reduce the bias (errors in the model estimation process) and variance (sensitivity to the training sample) components of the prediction error. Secondly, the initial motivation for proposing pVn modeling was to reduce the number of base models as required for OVA modeling. The experimental results demonstrated that pVn modeling enables the creation of syntactically diverse and highly competent base models. The pVn models were designed based on the lessons learned from OVA modeling. Confusion graphs derived from confusion matrices were used as input to the proposed algorithm for determining the class composition for the pVn base models. The experimental results reported in this chapter have demonstrated that the design of the base models based on the proposed algorithms results in pVn base models that provide a high level of predictive performance when combined into an aggregate model. The pVn aggregate models provided a much higher level of predictive performance compared to a single *k*-class model for the two datasets and two algorithms used for the experiments.

8.5.2 Comparison of OVA and pVn modeling

Table 8.15 provides a summary of the predictive performance of the OVA and pVn models for the datasets and algorithms used in the experiments. One small dataset, namely Wine quality (white) (Cortez et al, 2009) was also used to test performance of OVA and pVn dataset selection and modeling. The experimental results for the forest cover type and KDD Cup 1999 datasets were discussed in detail in chapter 7 and in this chapter. The details of the test results for the wine quality dataset are provided in appendix tables F.17 through F.26.

Table 8.15: Summary of performance improvements for OVA and pVn models

Dataset (size)	Algorithm	Is there a performance improvement compared to single model for the:		
		un-boosted OVA aggregate model?	boosted OVA aggregate model?	pVn aggregate model?
Forest cover type (large)	5NN	yes	yes	yes
	See5	no	yes	yes
KDD Cup 1999 (large)	5NN	yes	no	yes
	See5	no	no	yes
Wine quality - white (small)	5NN	no	no	yes
	See5	no	no	yes

OVA modeling provided performance gains for the forest cover type dataset for both the 5NN and the See5 algorithms. The un-boosted version of OVA modeling provided a small performance improvement for KDD Cup 1999 for the 5NN algorithm. The boosted version of OVA modeling did not provide any performance gains for the KDD Cup 1999 dataset for the 5NN and See5 algorithms. OVA modelling did not provide any performance gains for the wine quality dataset. pVn modeling provided performance gains for the forest cover type, KDD Cup 1999 and wine quality datasets for both algorithms. The performance improvements for the pVn aggregate models were far more impressive for the KDD Cup 1999 dataset compared to the forest cover type and wine quality datasets. An examination of the confusion graphs of figures 8.1, 8.2, 8.4, 8.5 and 8.7 reveals that one main difference between the prediction tasks for forest cover type and KDD Cup 1999 is that there is one class (NORMAL) for the KDD Cup 1999 whose node is connected to all the other nodes (classes) in the graph. This is not the case for the forest cover type confusion graphs. This observation could help to explain why, for a dataset such as KDD Cup 1999, OVA modeling as proposed in chapter 7 does not provide significant performance

gains, while pVn modeling provides significant gains. Further studies are required before firm conclusions can be made.

The F-tests for variance indicated that, in general, both OVA and pVn aggregate models exhibit the same level of predictive coherence. This leads to the conclusion that the OVA or pVn aggregate model should be preferred if such a model provides a higher level of mean predictive performance compared to a single *k*-class model.

It was observed from the experiments on OVA and pVn modeling, that OVA and pVn modeling can be used to reduce the problems associated with creating predictive models from datasets with skewed class distributions, especially when one or more classes are severely under-represented in the dataset. This is the case, for example, for the U2R class in the KDD Cup 1999 dataset. For the 52 instance of the U2R class, a combination of bootstrap sampling, training sample design to include only the necessary classes in the OVA and pVn models, and reduction of the training sample size were implemented for the OVAU2R, MNPU and MNRU base models. This scheme resulted in performance improvements on the TPRATE for the U2R class for the OVA aggregate models using the 5NN algorithm. The U2R TPRATE for the single 5-class model was 31.6±0.3, for the un-boosted OVA aggregate model the TPRATE was 42.6±0.4, and for the boosted OVA aggregate model the TPRATE was 40.5±1.4. However, for the See5 algorithm, the OVA and pVn aggregate models did not provide an increase in the TPRATE for the U2R class.

8.5.3 Classification problems where proposed boosting methods are not appropriate

Two-class problems are very common in data mining especially in business applications (Giudici & Figini, 2009; Witten & Frank, 2005; Giudici, 2003; Berry & Linoff, 2000). It was stated in chapter 2 that the OVA and pVn base model design and dataset selection methods proposed in this thesis are not appropriate for 2-class problems, but rather to *k*-class problems where $k > 2$. However if each of the classes for a 2-class problem is located in more than one contiguous region of the instance space, then it should be possible to apply the proposed methods to that dataset. For example, suppose that a 2-class dataset has classes c_1 and c_2 with the instances of class c_1 located in regions g_1 and g_2 while the instances of class c_2 are located in

regions g_3 and g_4. Classes c_1 and c_2 can be re-labelled as $c_1g_1, c_1g_2, c_2g_3, c_2g_4$ so that the classification task becomes a 4-class prediction problem to which the proposed methods can be applied. Liu and Motoda (1998) have observed that cluster analysis is commonly used as a pre-processing step in data mining. Samoilenko and Osei-Bryson (2008), and Osei-Bryson (2010) have observed that clustering is commonly used as a step prior to predictive modeling for purposes of improving the performance of predictive models. The author of this thesis hypothesised that identification of 1-class contiguous regions in the instance space of a 2-class problem can be achieved through cluster analysis. Experiments to test this hypothesis are left for future work.

The datasets used for the empirical studies on boosting have the desirable property that their confusion matrices have off-diagonal entries $CM(c_i, c_j)$ with $i = 1,.., k,\ j = 1,.., k$ and $i \neq j$ which do not have an equal (or nearly equal) distribution of instances. In fact, some of the entries in the off-diagonal confusion matrix cells are zero. The proposed OVA and pVn base model design and training dataset selection for boosted OVA and pVn base models were based on this property. The training samples for each OVA_i boosted base model or pVn_i base model were designed as follows: Each training sample included only instances of the classes where the off-diagonal entries $CM(c_i, c_j)$ and $CM(c_j, c_i)$ for $i \neq j$ in the matrix cells have large values, and to exclude instances of the classes with small or zero counts. There are *k*-class datasets for which the above property does not hold as shown in tables 8.16 and 8.17.

Table 8.16: See5 single 3-class model confusion matrix for abalone3C

Single model confusion matrix, training size = 3000, 10-fold cross validation			
Actual class	Predicted class		
	young	middle	old
young		206	51
middle	183		316
old	74	272	

For such datasets the (off-diagonal) entries in the class confusion cells all have nearly the same instance counts. The 3-class abalone3C dataset is a case in point. The 3-class waveform dataset (Blake & Merz, 1998; Breiman et al, 1984) was also identified as fitting this category. The confusion matrices for these two datasets for the See5 classification algorithm are given in tables 8.16 and 8.17.

Table 8.17: See5 single 3-class model confusion matrix for waveform

Single model confusion matrix, training size = 5000, 10-fold cross validation			
Actual class	Predicted class		
	Class 0	Class 1	Class 2
Class 0		269	217
Class 1	160		151
Class 2	179	140	

The foregoing observations led the author to formulate the following property for *k*-class confusion matrices:

Sparse confusion matrix property:
A *k x k* confusion matrix with exactly one off-diagonal cell having a zero count is minimally sparse. A *k x k* confusion matrix with all *k(k-1)* off-diagonal cells having zero counts is maximally sparse. A *k x k* confusion matrix with *j* off-diagonal cells, $1 \leq j \leq k(k-1)$ having zero counts is a sparse confusion matrix.

The implication of the above property is that there are classes in the dataset that do not share a common region of class confusion. The two large datasets that were used in the OVA and pVn studies for boosting training datasets both have the sparse confusion matrix property for the single *k*-class models. For this reason, it was possible to design boosted training datasets for OVA and pVn base models which resulted in increased predictive performance. It should be noted that it is possible that a non-sparse confusion matrix has off-diagonal cells with counts that are much smaller than the counts of all the other off-diagonal cells. Such a matrix can be converted into a sparse confusion matrix by setting the off-diagonal cell counts with small values to zero.

8.6 Conclusions

The first question that was posed for the studies on aggregate modeling and training dataset selection was: *How should training datasets be designed in order to create base models that are syntactically diverse and highly expert at prediction for aggregate models?* The experimental results reported in this chapter have demonstrated that the design of pVn models based on the information in the confusion matrix and confusion graph for a single k-class model and the new pVn model design algorithm presented in this chapter, results in the design of pVn base

models that are syntactically diverse and highly expert at prediction. The discussion of section 8.5.3 has however made it clear that the pVn and boosted OVA base model designs that are proposed are only applicable to datasets for which the single k-class predictive model has a sparse confusion matrix.

The second question that was posed was: *How should training datasets for the base models be designed in order to achieve high accuracy for the aggregate model?* The experimental results reported in this chapter have demonstrated that when pVn base models are designed as described above, the aggregation of such base models results in increased predictive performance. This was shown to be the case for the datasets and the algorithms that were used in the experiments. The experimental results also demonstrated that the predictive performance increases achieved through the proposed OVA and pVn aggregate modeling methods do not come at the cost of reduced coherence in the predictions.

The models discussed in chapter 7 and this chapter were assessed for performance using mean values for accuracy and TPRATE values as well as the variance in accuracy. Evaluation of model performance using Receiver Operating Characteristic (ROC) analysis is presented in the next chapter.