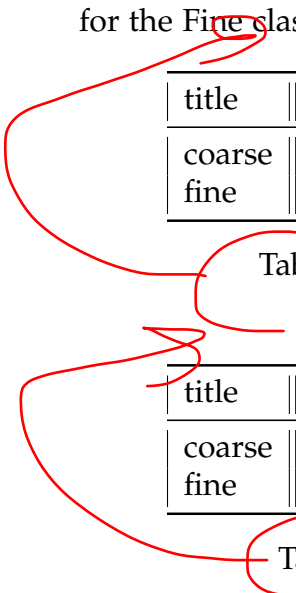


Chapter 4

Results and Analysis

4.1 SVM and LogReg Classifier Performance

Both the SVM and the LogReg classifiers show a slight advantage for the Fine classifier over the Coarse classifier in terms of the PR-AUC metric. The ROC-AUC metric is close to identical between fine and coarse for both classifiers, a slight advantage of 0.002 exists for the Fine classifier in the SVM classifier.



title	pr	roc	acc	f1	conf (tn/fn)	conf (fp/tp)
coarse	0.870	0.871	0.787	0.268	(1503.2 / 17.8)	(410.4 / 78.3)
fine	0.875	0.871	0.913	0.403	(1776.5 / 37.3)	(137.1 / 58.8)

Do you refer to these tables in the main text?

Table 4.1: LogReg entire dataset results after parameter tuning

title	pr	roc	acc	f1	conf (tn/fn)	conf (fp/tp)
coarse	0.892	0.880	0.866	0.347	(1669.5 / 24.8)	(244.1 / 71.3)
fine	0.898	0.882	0.942	0.485	(1839.0 / 41.5)	(74.6 / 54.6)

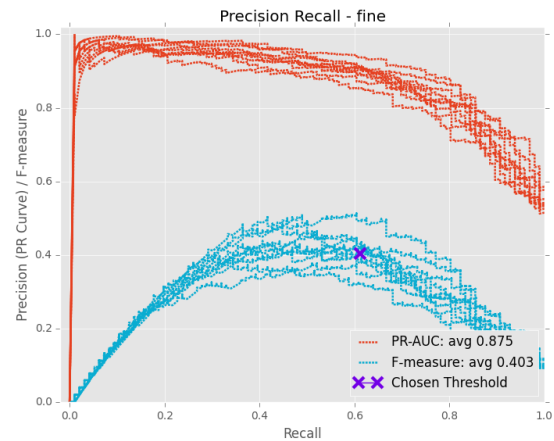
Table 4.2: SVM entire dataset results after parameter tuning

The coarse classifier in both the LogReg and SVM classifier has a greater amount of

false positives at the default threshold. A further examination of these values is shown in Figures 4.2-4.1 and Figures 4.4-4.3. The figures plot the PR and the ROC curves for each of the 10 folds. Each point on the PR and ROC curve has a corresponding F-measure or Accuracy value, these values are plotted on the graphs as a blue line. The graphs demonstrate that the coarse and fine classifiers have close to equivalent average AUC, on the order of 0.007 max difference between Fine and Coarse. At the default threshold the Fine appears to outperform Coarse for Accuracy and F-measure metrics, but the inspection of the plots shows that a coarse threshold can be chosen to match the fine output for both Accuracy and F-measure. The PR-AUC does show a slight advantage for Fine, which warrants application of the HAL algorithm and Active over-labeling approach on this dataset.

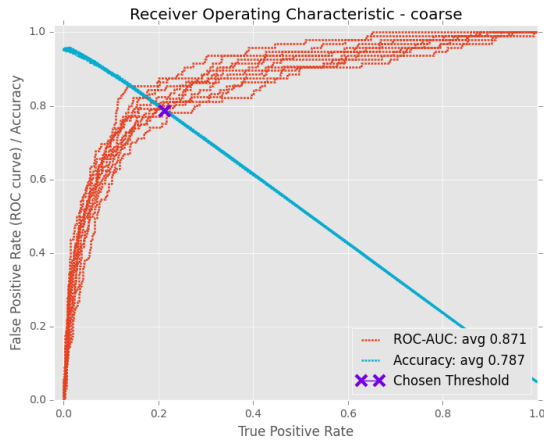


(a) Log Reg Pr Curves - Coarse

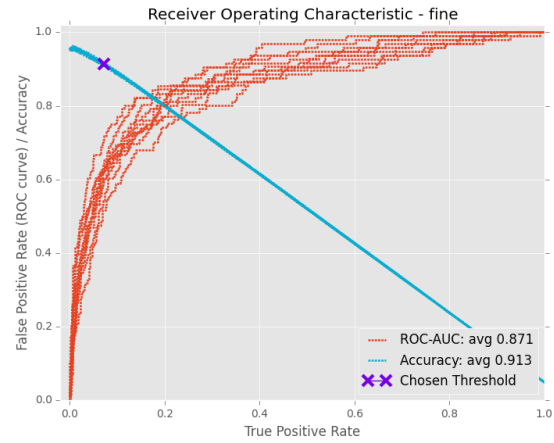


(b) Log Reg Pr Curves - Fine

Figure 4.1: The Fine default threshold occurs at a point on the PR curve associated with a higher F-measure score compared to the Coarse curves.

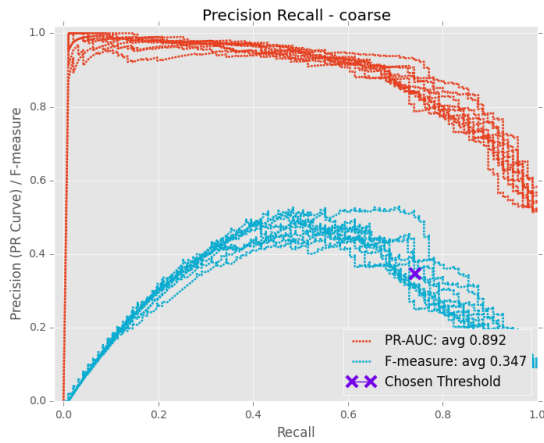


(a) Log Reg ROC Curves - coarse

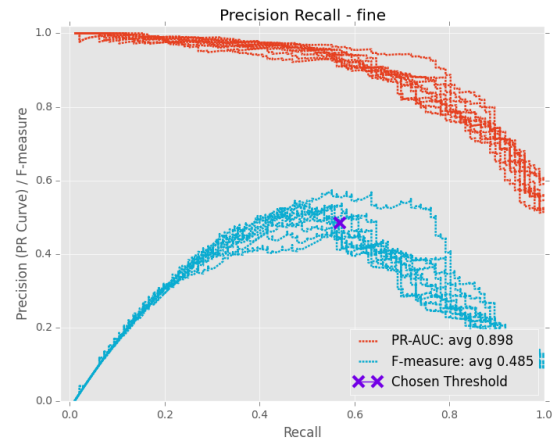


(b) Log Reg ROC Curves - fine

Figure 4.2: Fine has a higher accuracy than coarse at the default threshold for the LogReg classifier.

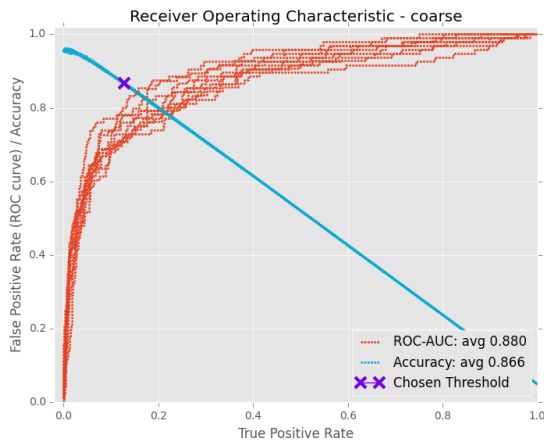


(a) SVM Pr Curves - Coarse

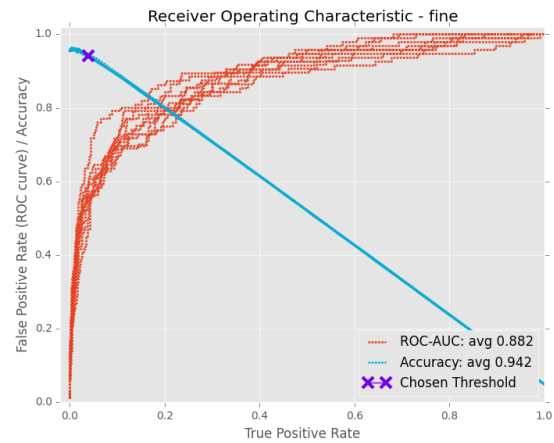


(b) SVM Pr Curves - Fine

Figure 4.3: SVM results for PR curves and F-measure have Coarse and Fine picking different parts of the curves for their respective thresholds. This results in a slight advantage for Fine at the default threshold, similar to the results for the LogReg classifier.



(a) SVM ROC Curves - Coarse



(b) SVM ROC Curves - Fine

Figure 4.4: SVM Accuracy results are similar between Coarse and Fine.

4.2 Active vs Passive curves

The plots in Figures 4.5--XXX

The following plots were obtained with a round batch size of 100 and a starter set of 1040 instances out of the total 20098 instances. The plots are the average of 10 folds, for each fold a test set of 2010 instances is used. The test set remains constant throughout the rounds and contains a representative proportion of each of the classes. The starter set is chosen out of the remaining 18088 and it also contains representatives from each class in proportion to that class's prominence in the dataset. The 17048 non-test set, non-starter set instances are added to the training set in batches of 100. This results in total of 171 rounds, 170 batch selecting rounds and 1 starter set round. The Passive approach selects 100 random instances and adds them to the train set. The Active approach runs the classifier on the eligible instances, orders them by their uncertainty and adds the 100 most uncertain instances to the train set. Coarse and fine classifiers share the same starter set. During each round, Coarse and Fine classifiers are trained on their corresponding sets, which are independent of one another, metrics are outputted on the held out test set which is the same for both Coarse and Fine.

4.2.1 Plots for Logistic Regression Active vs Passive curves

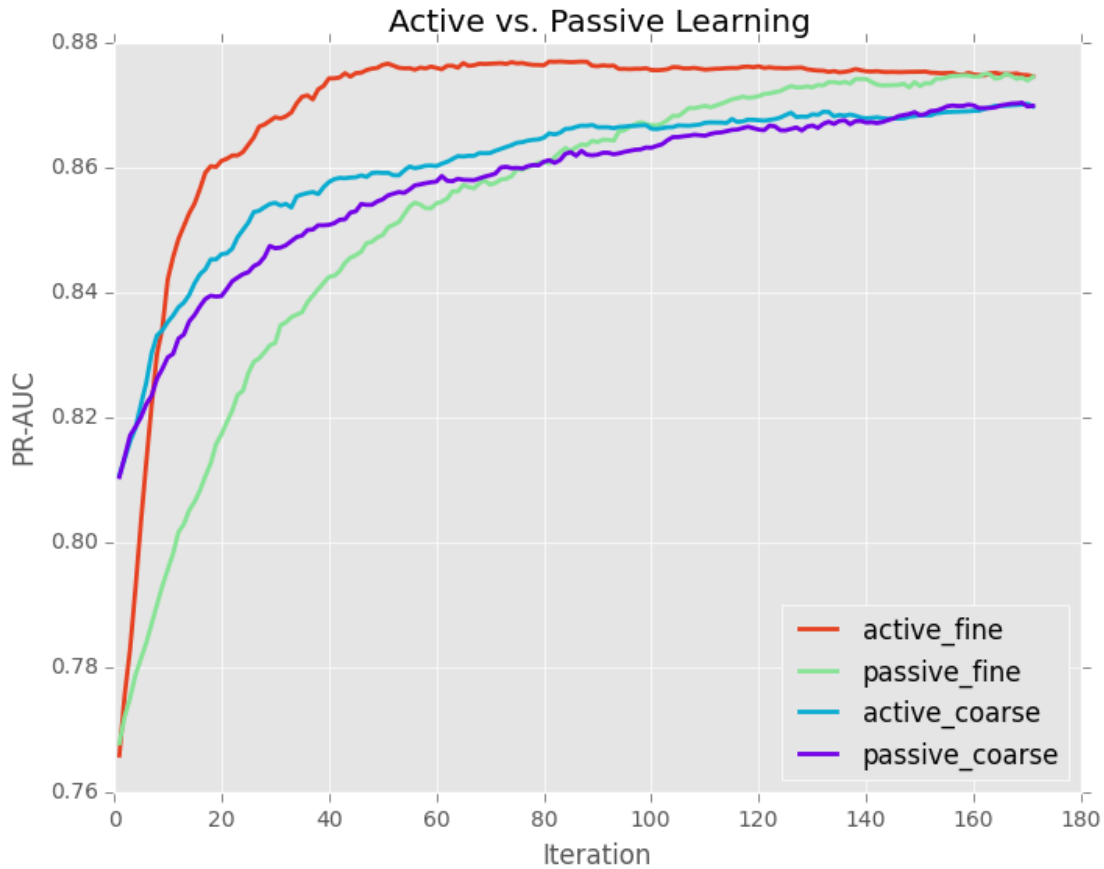


Figure 4.5: The PR-AUC curves for rounds with the Logistic Regression classifier conforms to expectations, with Active Fine having the highest performance, and Active outperforming Passive for both Coarse and Fine classifier types.

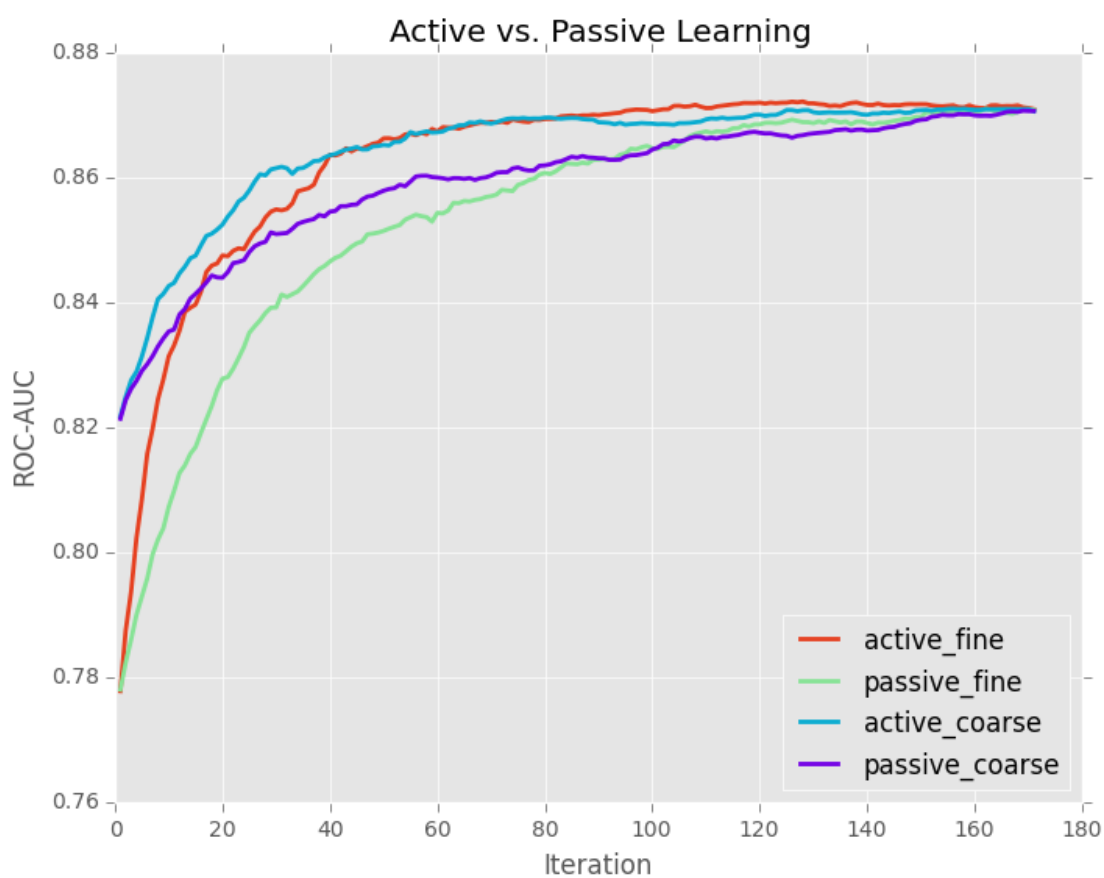
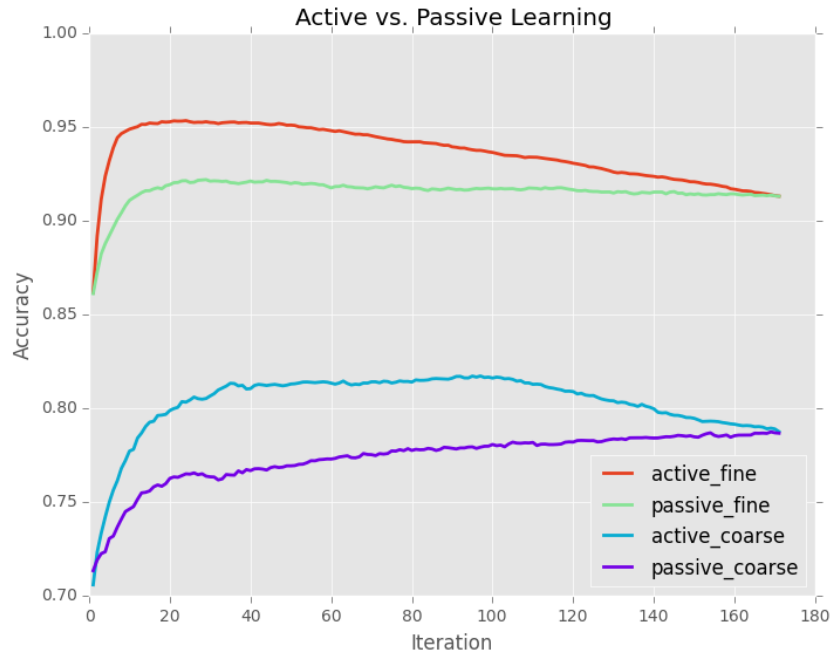
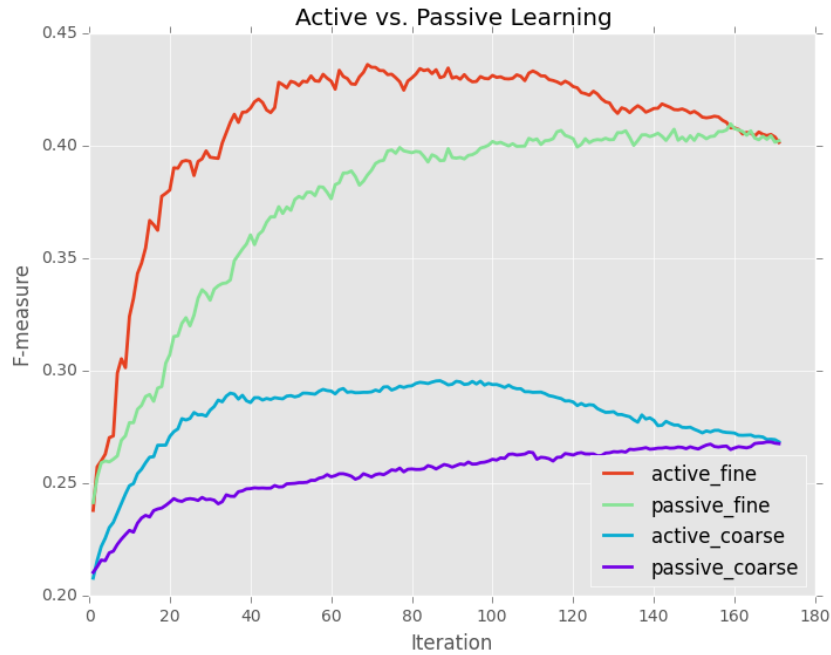


Figure 4.6: The ROC-AUC curves for rounds with the Logistic Regression classifier. The active curves beat out the passive curves for both Coarse and Fine. Note that Active Fine ROC curve doesn't converge to the Active Coarse ROC curve until round 40. This is contrasted to a dominance of the Active Fine PR curve after round 10.



(a) LogReg Accuracy



(b) LogReg F-measure

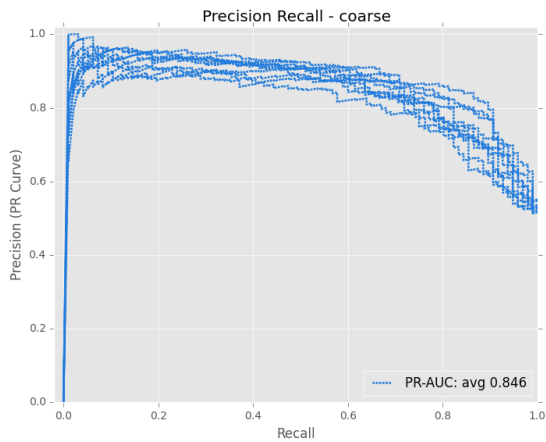
Figure 4.7: The Accuracy of the classifiers stays at roughly the same rate throughout the rounds, this is due to an effective weighting scheme. Both curves show a dominance of Fine over Coarse and Active over Passive.

Note that the Active Fine PR-AUC curve surpasses Active Coarse after round 10 while the Active Fine ROC-AUC curve is still well below the Active Coarse at that round. These curves are shown in *Figures 4.8- 4.9*. This is counter-intuitive, ^{because} according to a proof in Davis [3], “For a fixed number of positive and negative examples, one curve dominates a second curve in ROC space if and only if the first dominates the second in Precision-Recall space”. The theorem uses the following definition of dominance: that every value in the first curve is above the corollary value in second curve. The correlation between PR and ROC curves is that Recall in the PR curve is equivalent to the True Positive Rate in the ROC curve. The average PR-AUC concept is different than that of a plot of PR curves for a round, but if all of the PR curves for Fine dominate the curves for Coarse then we would expect all of the ROC curves for Fine to dominate the ROC curves for coarse and both the ROC-AUC and PR-AUC averages for Fine to be greater than that for Coarse. However, it is shown in *Figure 4.8* that the PR curves for Fine do not completely dominate the PR curves for coarse, and similarly for the ROC curves in *Figure 4.9*. Active Fine PR-AUC curve does not satisfy the theorem’s definition of dominance, since each individual ROC and PR curve contains intersection points between Coarse and Fine. Thus, given that the average PR-AUC for Fine is great at round 20 than average Coarse PR-AUC, this relationship is not expected to hold between the average ROC-AUC curves.

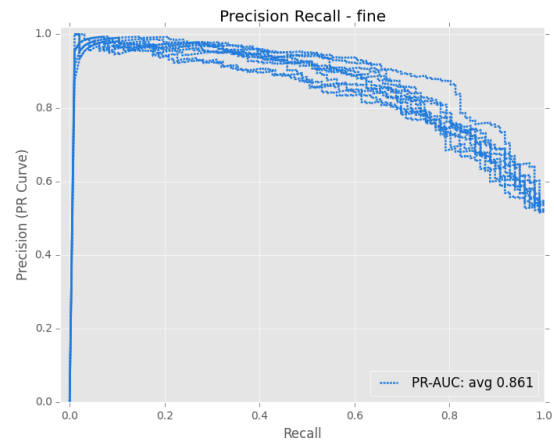
According to Davis [3], a large change in the number of false positives can still yield only a small change in the number of false positives and thus not affect ROC curve performance. However, Davis states, “Precision, on the other hand, by comparing false positives to true positives rather than true negatives, captures the effect of the large number (incorrectly classified) of negative **exaples** on the algorithm’s performance” [3]. Since our dataset demonstrates a heavy class imbalance with a roughly 1:20 ratio of Positive to Negative instances, the algorithm’s ability to classify negative instances should be taken into account when considering overall classifier performance. The PR

So, if you're saying that performance on negatives is very important, we should be emphasizing ROC over PR? Isn't that the opposite of what we discussed, or am I missing something?

curve's ability to capture and have a greater sensitivity to the increased number of False Positives, reveals the advantage that the Fine classifier has over the Coarse classifier. This justifies purchasing Fine-grained labels over Coarse-grained labels to improve classifier performance.

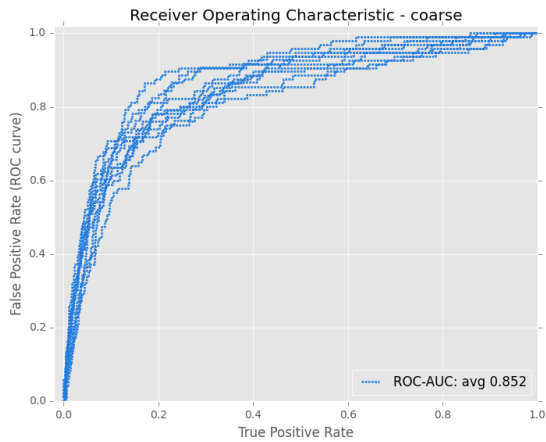


(a) Coarse PR curves at Round 20

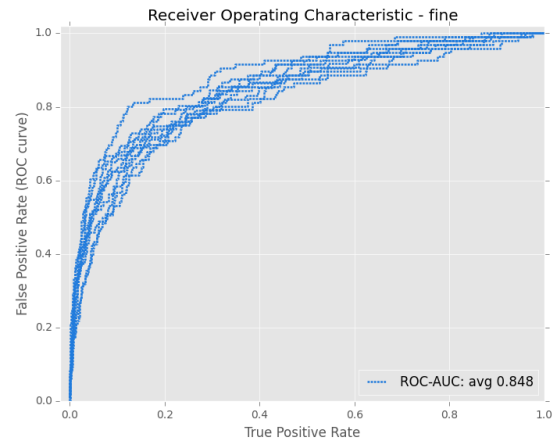


(b) Fine PR curves at Round 20

Figure 4.8: PR curves for each fold at Round 20



(a) Coarse ROC curves at Round 20



(b) Fine ROC curves at Round 20

Figure 4.9: ROC curves for each fold at Round 20

4.2.2 Plots for SVM Active vs Passive curves

The SVM Active vs Passive experiment is performed with the same methodology as the previous section detailed with the exception that a SVM classifier is substituted for the LogReg classifier. Due to the greater advantage of average PR-AUC in the LogReg classifier, the SVM is not used in the Fixed Fine ratio experiments in the [section that follows](#).

Which section?

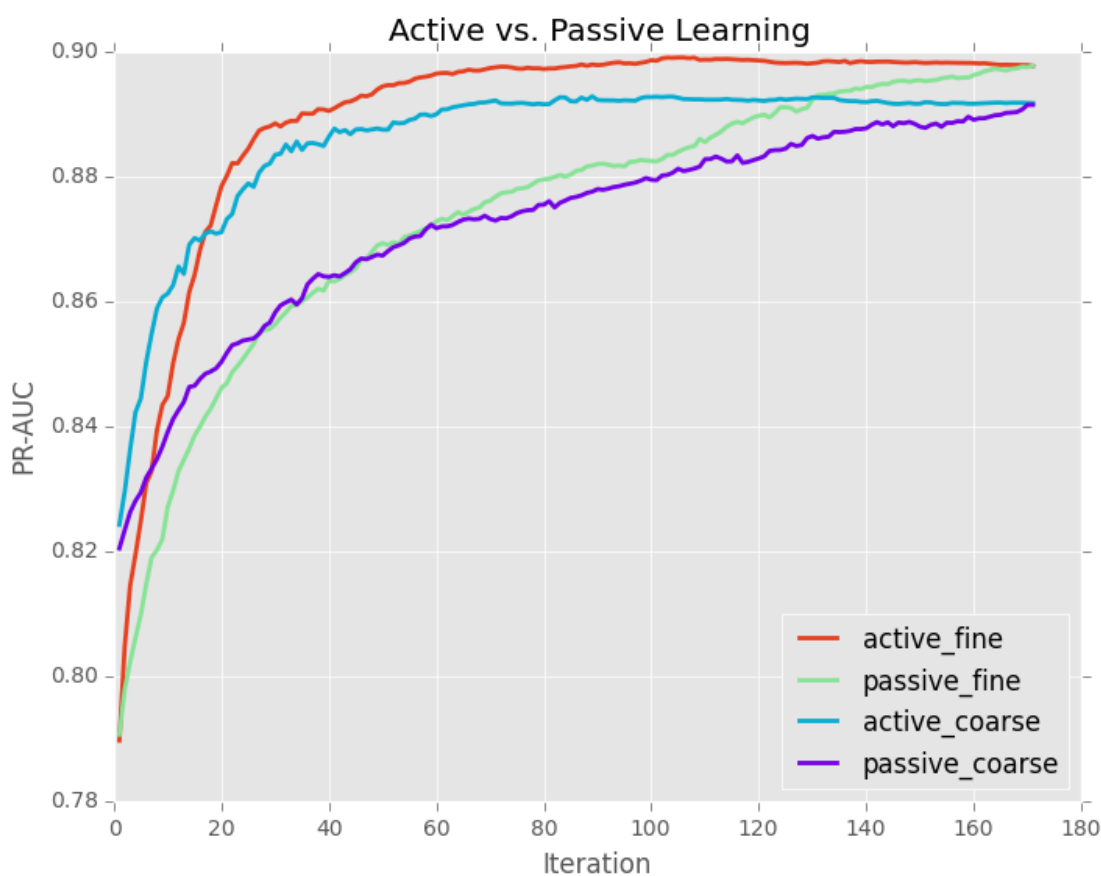


Figure 4.10: The PR AUC curves for SVM show a slight advantage for Active Fine, similar to the LogReg results.

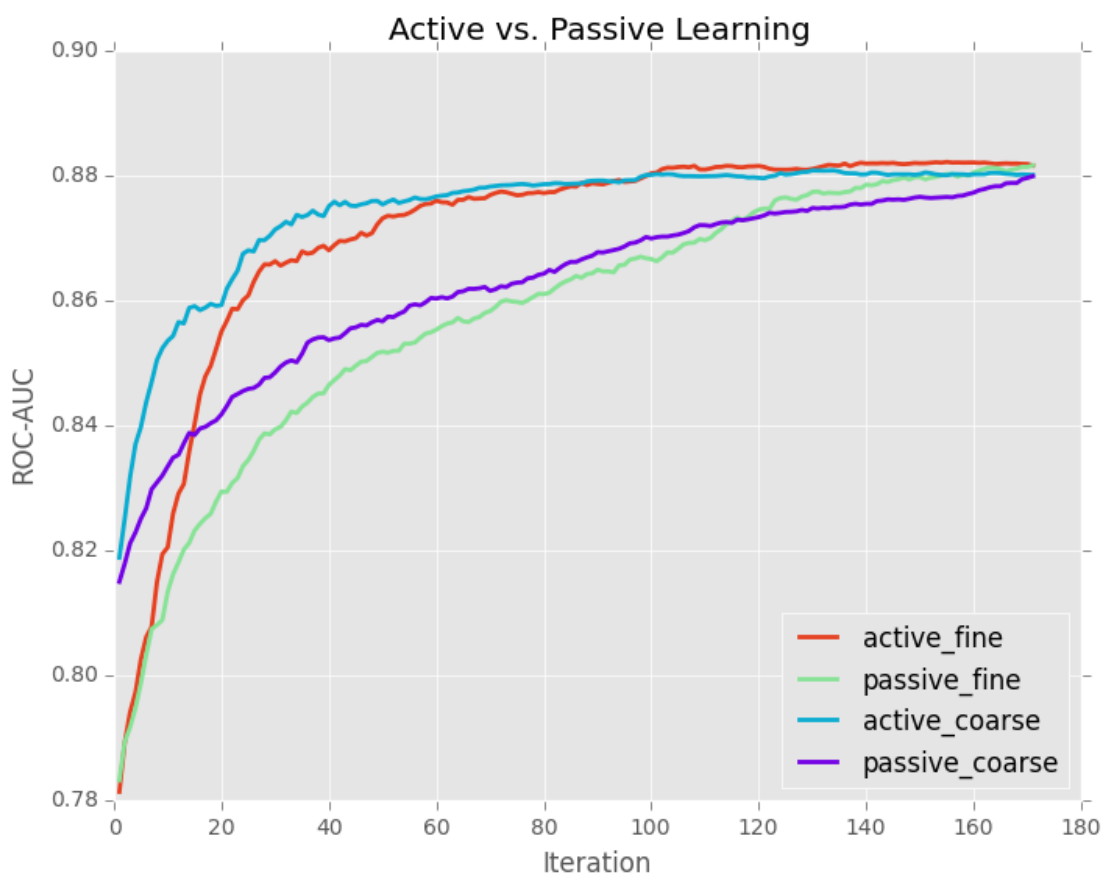


Figure 4.11: The ROC AUC curves for SVM match the LogReg results, the convergence of Active Fine to Active Coarse takes slightly longer, round 60 compared to round 40.

4.3 Plots for Fine Fixed Ratio experiments

The strategy is changed from purchasing a set number of instances per round to having a set budget per round and spending a portion of that budget on fine and coarse-grained labels. The Fine Fixed Ratio (FFR) ranges from 0.0 to 1.0 in increments of 0.1. Note that the FFR 0.0 should roughly correlate to the Active Coarse curve shown in *Figure 4.5*. Likewise the Active Fine curve should roughly correlate to the FFR 1.0 curve. However, the correlation is not exact since the FFR experiments use a combination classifier, it trains Fine and Coarse classifiers on a starter set of the same size and proportion as used in the LogReg Active vs Passive experiment, then uses the confidence of both of those classifiers and the end prediction is the max of the two classifiers. Thus even for the FFR 0.0 and FFR 1.0 the starter set trained Fine or Coarse classifier still contributes to the PR-AUC curve even at the final round 180. The results are an average of 10 folds.

To determine the number of instances to purchase each round, the FFR is multiplied by the round budget of 100, then the round budget for the coarse labels are purchased at a cost of 1.0. The cost of the fine labels will vary and if a decimal occurs it is resolved by randomly purchasing an extra fine label with the probability of the decimal value. For example, if the fine cost is 16 and FFR is 0.5, 50 instances are bought for coarse and 3.125 instances are bought for fine. The remainder 0.125 is then turned into a 0.125 chance for any round to purchase an extra fine label. The round size for the FFR 1.0 curve is very small, with only 7 labels purchased per iteration.

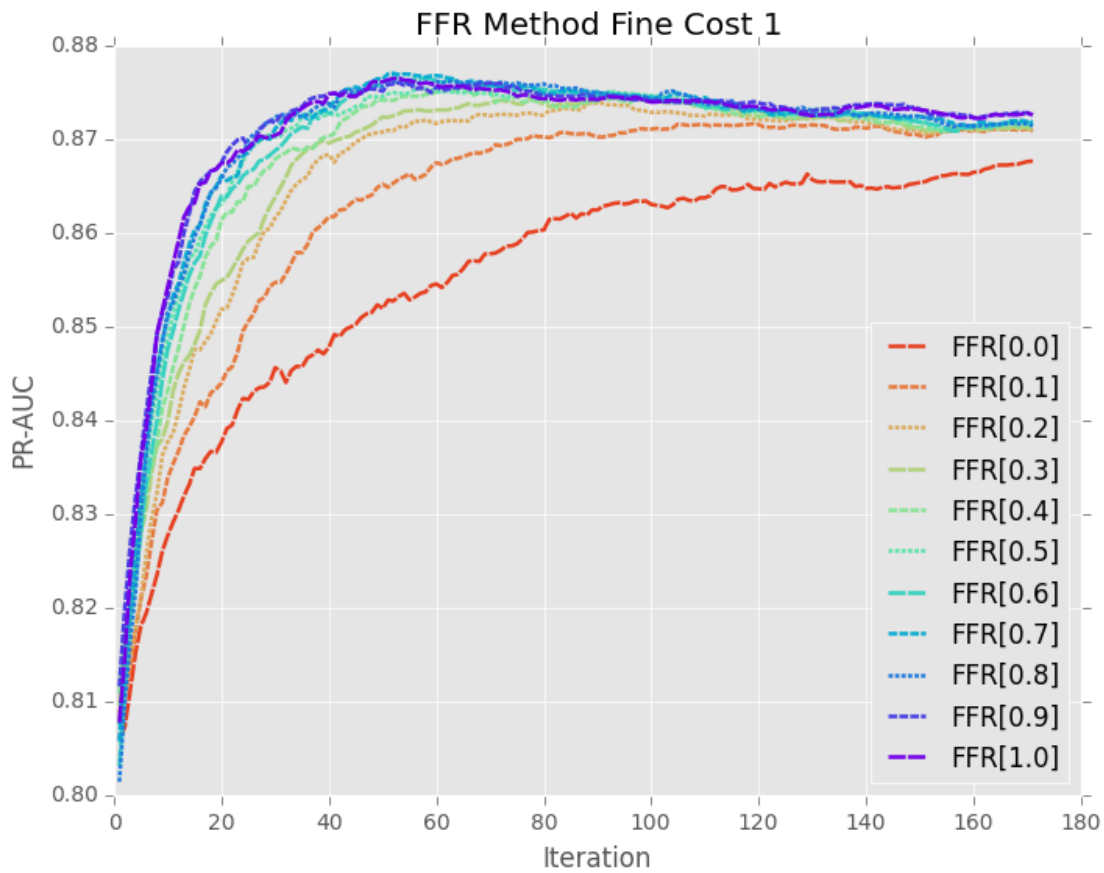


Figure 4.12: For this curve the fine and coarse grain labels both have a cost of 1. The purple 1.0 curve shows that if only fine grained labels are purchased, the highest performing PR-AUC can be obtained. All FFR ratios end at the same round since the cost of the Fine and Coarse instances is the same the budget .

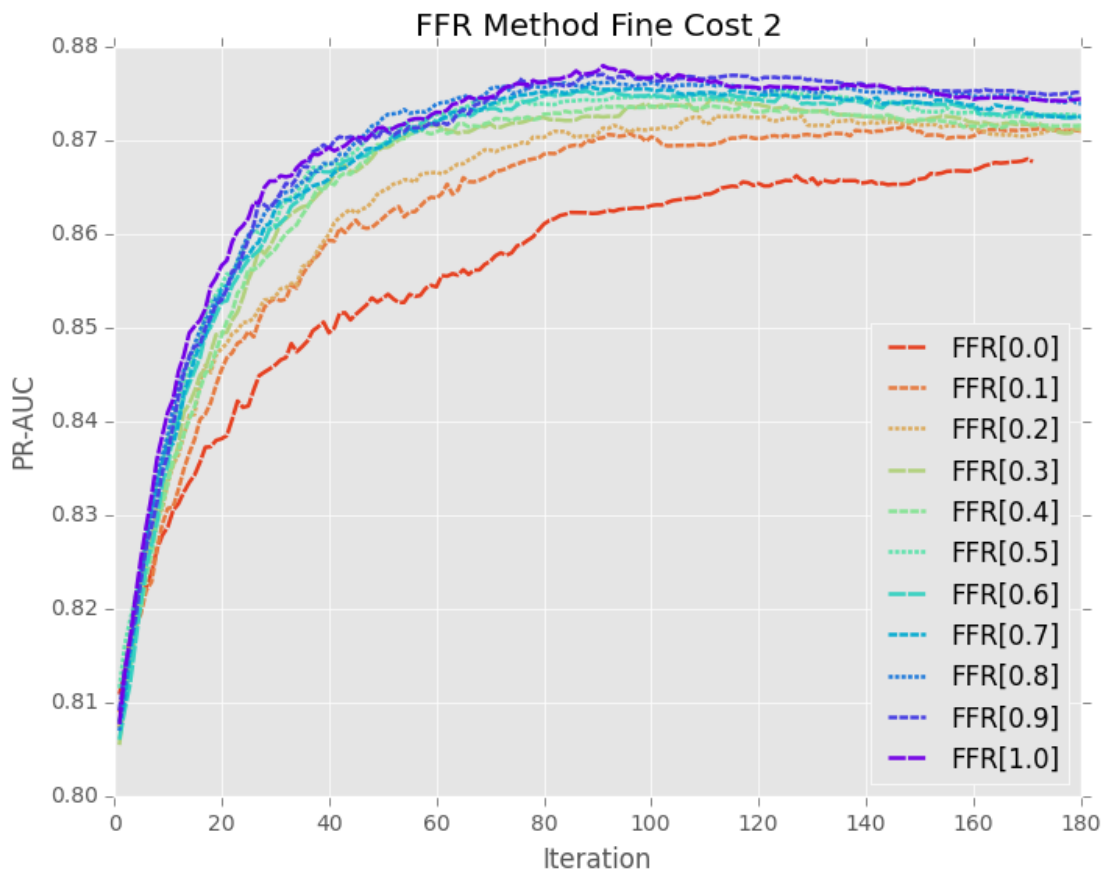


Figure 4.13: At fine cost 2, advantage of the higher FFR values decreases but the ordering of the curves remains unchanged.

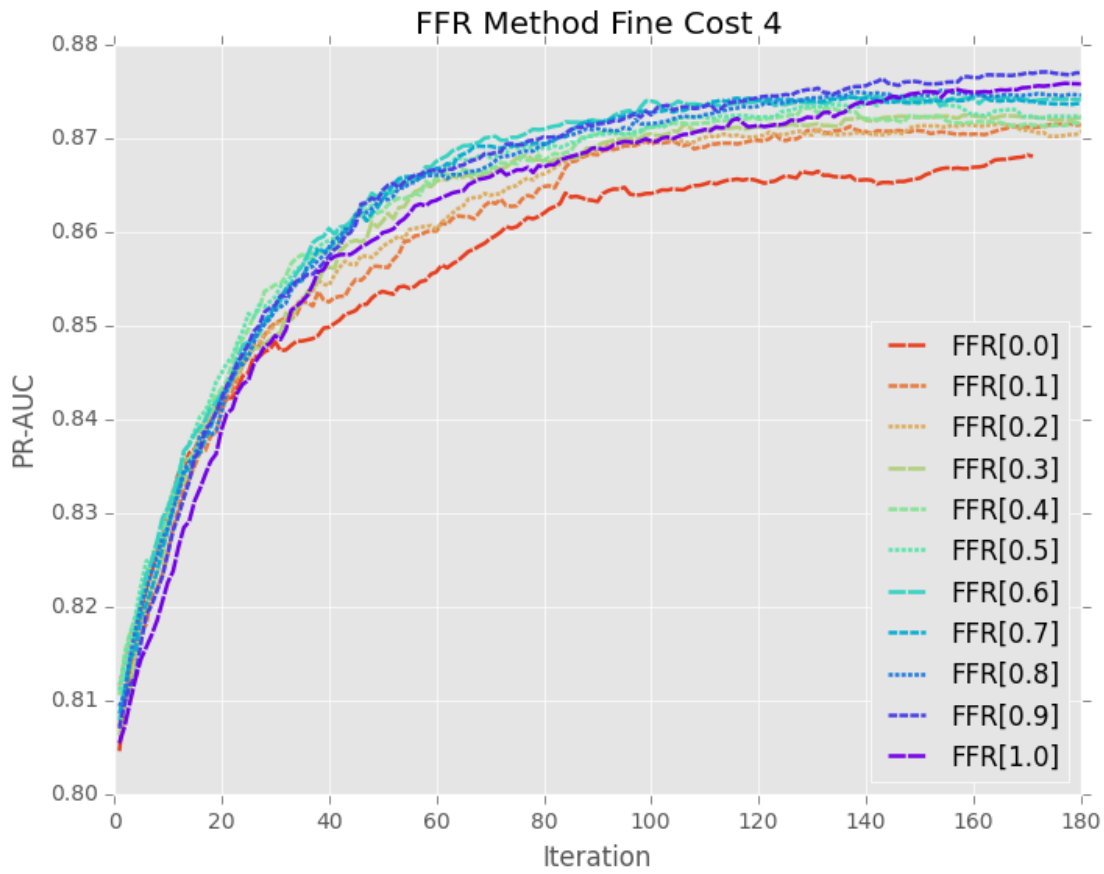


Figure 4.14: At fine cost 4, the highest FFR 1.0 is no longer preferred, the cost is too high for Fine instances PR-AUC utility to overcome the PR-AUC increase gained by purchasing more Coarse instances.

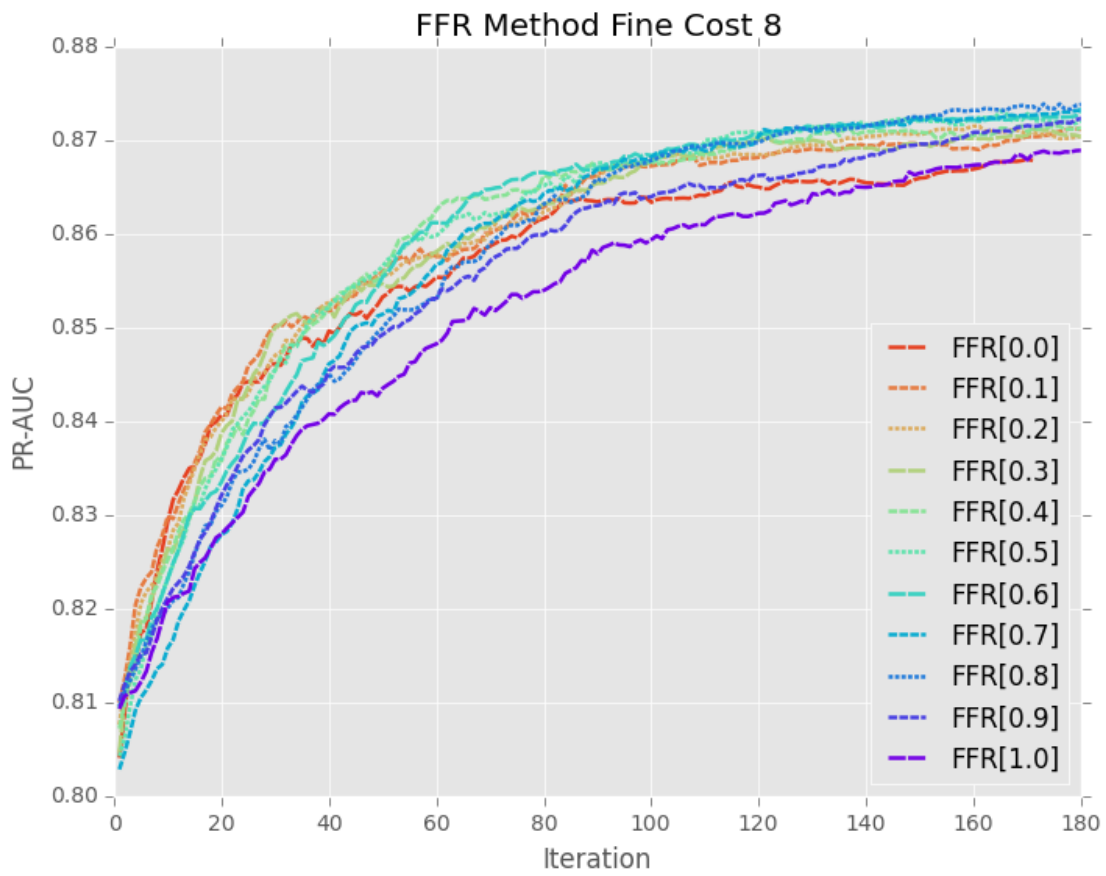


Figure 4.15: At fine cost 8 the middle FFR values outperform the extreme values for rounds 0 to 180.

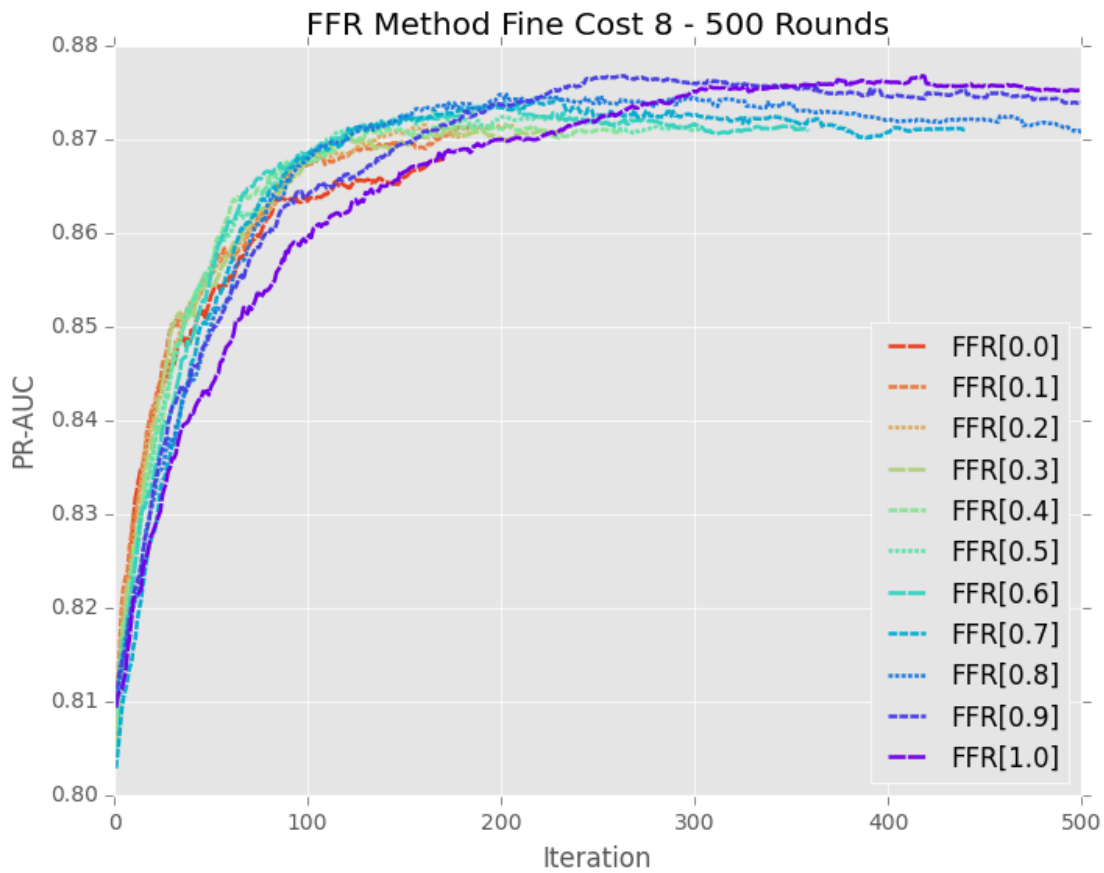


Figure 4.16: This shows the iterations continuing through round 500, the curves with the higher fine rates eventually settle to the same end point that the curves with the high rates of coarse labels purchased achieved at previous iterations.

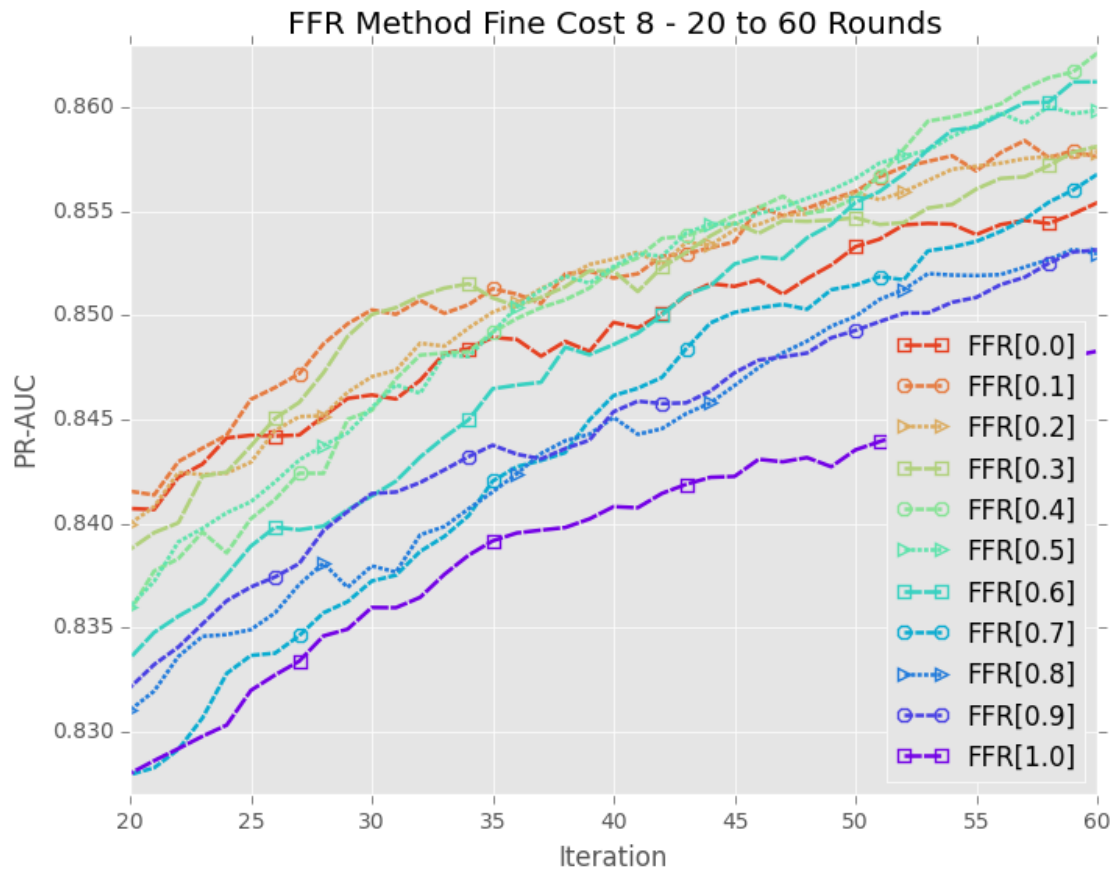


Figure 4.17: The fine cost 8 curves shown expanding the rounds 20-60. If a round budget of 40 occurs than the recommended Fine Fixed Ratio would be 0.2 .

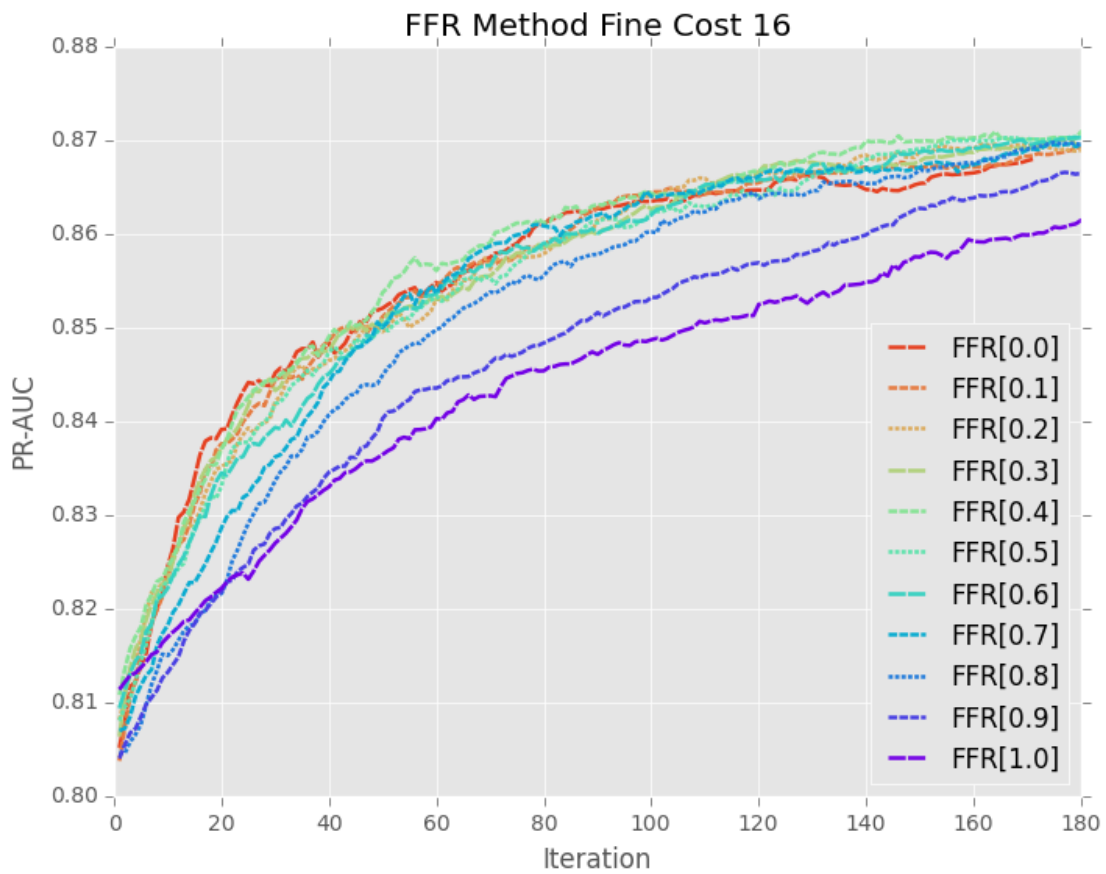


Figure 4.18: The fine cost is increased to 16. The cost is too high for the fine label advantage to offset the decreased number of instances purchased.