Assignment 4: Feature Selection II.

October 20, 2014

John Hancock

Florida Atlantic University

Advanced Data Mining and Machine Learning CAP-6778

jhancoc4@fau.edu

Abstract

For this assignment we are required to examine the performance of the Naive Bayes and 5 Nearest-Neighbors classifiers when we apply feature selection techniques to the input data for these classifiers. We report on the number of features a feature selection technique should retain to get optimal performance in terms of area under the receiver operating characteristic curve. We do an analysis to determine the inluence of classifier performance in terms of classifier, feature selection technique, and feature subset size. We examine the overlap of attributes that various feature selection techniques select.

Introduction

This work investigates the performance of the Naive Bayes (NB) [1] and 5-Nearest Neighbors (5NN) classifiers available in the Weka data mining tool[2]. We configure the Weka tool to

evaluate the performance of these classifiers where we use data that has been treated with a feature selection technique as input to the classifiers. The feature selection techniques we use are all available in the Weka tool. Please see the methodology section for more details on which feature selection techniques we use. We analyze the performance of the classifiers, and look at the results of treating data with feature selection techniques in terms of the overlap of features that the various techniques choose to retain, for feature subset size 6. We also took this assignment as an opportunity to write software that automates using the Weka tool, and stores experiment results in a database.

Methodology

For this work we run two Java programs [1]

All of the source code that produces the results we report on in this work is available at https://github.com/jhancock1975/data-mining-assignment-2/tree/master/cap-6778.

The first program we wrote, Assignment4.java uses the Weka Java API to construct classifiers and then iterate over combinations of classifiers and input data.

The classifiers Assignment4.java creates are: Naive Bayes (NB), 5-Nearest Neighbors (5NN), and J48 (C4.5) Decision Tree (J48). We use default settings for all classifiers, with the exception of setting the number of neighbors to 5 for the nearest neighbors classifier.

The dataset we use for this work is Lymphoma96x4026.arff.

The dataset has two classes, ACL and nonACL.

The Assignment 1 requirements document states, "Type I (False Positive): a nonACL module is classified as ACL." Hence, ACL must be the positive class. Since ACL must be the positive class, nonACL must be the negative class.

[1] the program for generating result data and storing it to a database is at `https://github.com/jhancock1975/data-mining-assignment-2/blob/master/cap-6778/src/main/java/edu/fau/weka/Assignment4.java` and the program for generating the gnuplot graphs is at `https://github.com/jhancock1975/data-mining-assignment-2/blob/master/cap-6778/src/main/java/edu/fau/weka/Assign4Reports.java`

We us the dataset as-is as input for all three classifiers. In addition we apply one of the 6 feature selection techniques to the data to use a subset of the features in the dataset as input to the Naive Bayes and 5-Nearest Neighbors classifiers.

The 6 feature selection techniques we use for this work are: Information Gain (IG), Gain Ratio (GR), Chi-Squared (CS), ReliefF (RF), ReliefF Weighted (RFW), and Symmetric Uncertainty (SU). For an overview of the functioning of these filter-based feature ranking techniques, please see part III. Methodology, section C. Feature Ranking Techniques of [3].

For each feature selection technique we set the number of features retained using the values 5, 6, 7, 8, 9, 10, 20, 50, 100, 200.

As Assignment4.java runs it stores classification results for each combination of dataset and classifier in a database. Due to the requirements in Assignment 4, we store the classifier's false positive rate (FPR), false negative rate (FNR), positive class area under receiver operating characteristic curve (pAUC), negative class AUC (nAUC), and weighted average AUC (wAUC). In addition we store a list of the features each feature selection technique retains for the dataset.

We run a second program, Assign4Reports.java to process the data stored in the database to generate graphics in the results below.

To answer other questions given in Assignment 4, we run queries on the data that Assignment4.java stores.

For all classifications we run, we use 10-fold cross validation.

To summarize the size of the experiments, we run $2 \times 6 \times 10 = 120$ combinations of classifiers, feature selection techniques , and features subset size using 10-fold cross validation. In addition we run 10-fold cross validation each of the NB, 5NN, and J-48 classifiers using the full dataset as input, for a total of 123 sets of classification results.

Results

Patterns

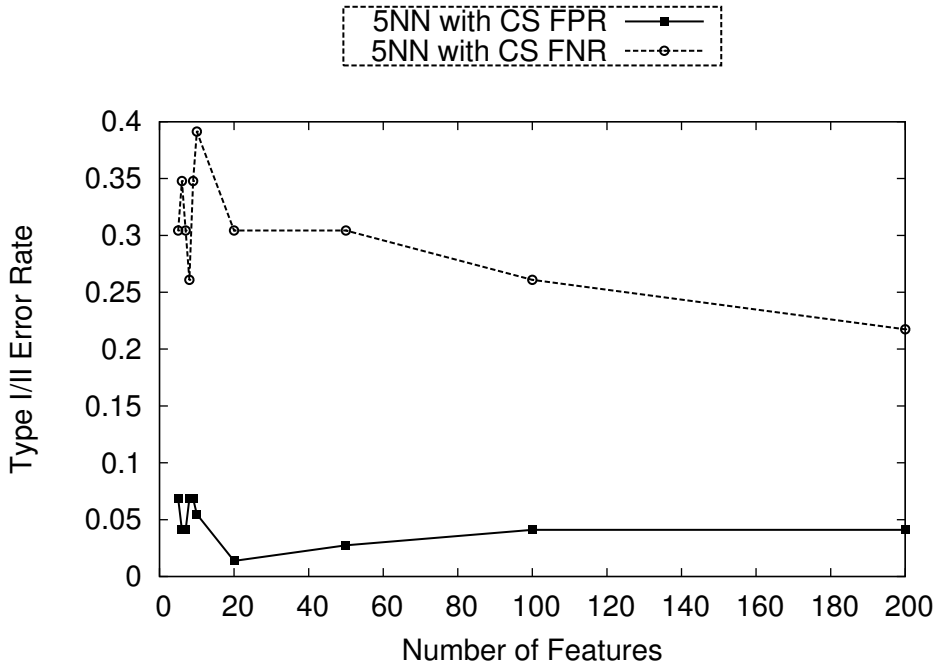
Assignment 4 has the following requirement, "**discover patterns in terms of FPR, FNR, and AUC as the number of features retained changes. Report on these patterns...**" The way we find patterns is to plot FPR, FNR, and AUC versus the number of features changes.

We review the plots (please see Appendix 1 for a full listing of all plots) and find the following patterns:

- For the 5NN classifier, the false negative rate tends to drop steeply when the number of features reaches is about 20, and decrease more slowly for larger numbers of features.
- For the 5NN classifier, the false positive rate rate tends to reach a minimum with around 20 features, and then stay flat or increase slow for larger numbers of features.
- For the 5NN classifier, all three AUC values that Weka collects always overlap. Furthermore, the pattern we see is that AUC values attain a maximum at 20 to 100 features and vary slightly for larger number of features.
- For the NB classifier, we generally see the false negative error rate drop to a minimum at 20 features, and remain at this minimum for higher numbers of features.
- For the NB classifier, the false positive rates dip to minimum around the 20 feature level, and rise steadily after that.
- For the NB classifier, AUC values do not always overlap. However, we see that generally, AUC values attain a local maximum around 20 features. For higher numbers of features, sometimes the AUC values increase slowly after attaining the local maximum, and sometimes they decrease slowly.

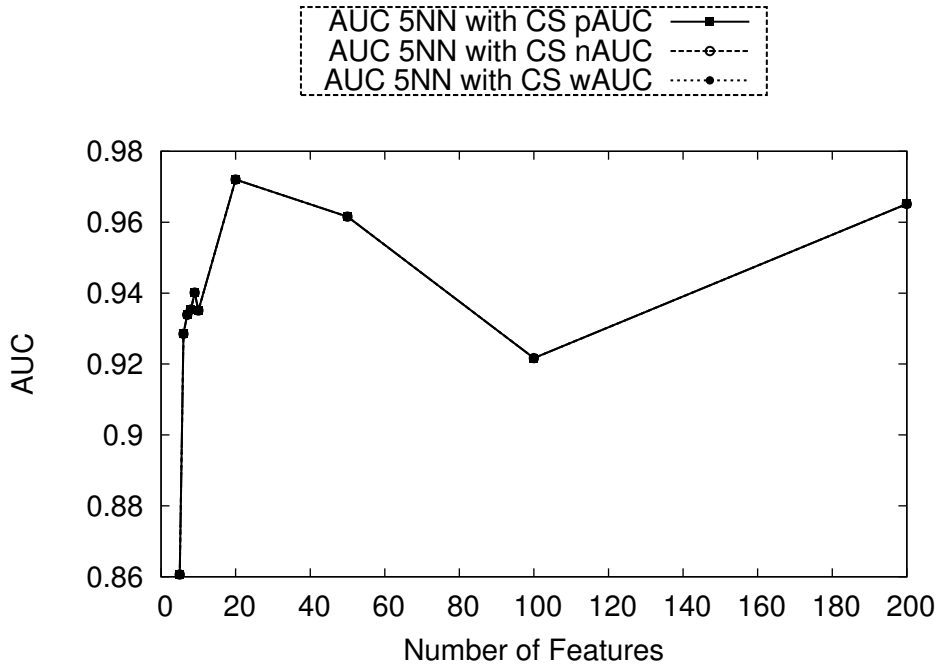
Appendix 1 below contains 24 graphs for all combinations of classifier and feature selection technique. We decided to display these graphs in the Appendix because they are not strictly required in Assignment 4. We show two charts here to give the reader an idea of what is in Appendix 1, in case it is not interesting to examine all of the graphs.

5NN with CS Attribute Selection



Error Rates

Figure 1



AUC

Figure 2

Optimal Number of Features in Terms of AUC

Assignment 4 requires us to report on AUC data in the following manner:

"...including the optimal number of features in terms of AUC, the evidence that led you to conclude this, and the resulting performance (in terms of FPR, FNR, and AUC) when this number is used. Be sure to include the performance of the classifiers on the full set of attributes for comparison."

We must make a note about AUC values Weka collects when it runs a classification. Weka collects 3 AUC values in classification results. One AUC value is associated with the positive class (ACL for the Lymphoma96x4026.arff dataset, we refer to as pAUC), one with the negative class (nonACL, we refer to as nAUC), and a weighted average (we refer to as wAUC). For any of our results involving the 5NN classifier, all 3 AUC values are equal. However, for the NB classifier, results are not always equal. Therefore we may obtain a different optimal number of features depending on which AUC value we are interested in. We present results for all 3 AUC values.

We query our database of results to meet this requirement. For the results in this section, and sections below where we mention that we wrote queries to obtain results, we invite the reader to peruse queries in our source code repository under `https://github.com/jhancock1975/data-mining-assignment-2/tree/master/cap-6778/src/main/resources/sql`

We would also like to emphasize that the table below reports FPR and FNR as required for Assignment 4.

Our first query results give the maximum pAUC value per combination of classifier and ranker.

Classifier	FS Technique	Num. Features	pAUC	FPR	FNR	Figure
5NN	RFW	20	0.977963	0.0136986	0.26087	10
NB	SU	20	0.977367	0.109589	0.0434783	24
NB	CS	20	0.976772	0.0547945	0.0434783	14
NB	GR	20	0.972603	0.0547945	0.0869565	16
5NN	CS	20	0.972007	0.0136986	0.304348	2
NB	IG	20	0.966051	0.0958904	0.0434783	18
5NN	SU	100	0.961584	0.0684932	0.130435	12
5NN	IG	7	0.957117	0.0410959	0.304348	6
5NN	RF	20	0.94461	0	0.521739	8
5NN	GR	200	0.936569	0.0547945	0.173913	4
NB	RFW	200	0.931805	0.123288	0.0869565	22
NB	RF	200	0.928231	0.219178	0.0869565	20
5NN	No feature selection	4026	0.882073	0.109589	0.391304	(N/A)
NB	No feature selection	4026	0.84455	0.178082	0.26087	(N/A)
J48	Embedded feature selection	6	0.776951	0.0958904	0.391304	(N/A)

Table 1: AUC values associated with positive (ACL) class by classifier and feature selec-

tion (FS) technique. The figure column indicates which figure in Appendix 1 of this document the reader may to refer to see the data in the associated row in graphical form. N/A stands for, "not available."

The results are sorted by pAUC in descending order, so the overall highest pAUC value is in the first row, and so on.

We make some points about results in about results in Table 1:

- For the 5NN classifier and IG feature selection, figure 6 shows the maximum AUC value is attained twice, once for 7 features, and once for 40. We report 7 as the optimal number of features as it is our understanding that the data mining community prefers models with smaller feature sets over models with larger feature sets. We decide optimal feature numbers for other tie values similarly.

- For the 5NN classifier with GR feature selection, the maximum pAUC value is associated with 200 features selected, which is the maximum number of features selected in our experiments. However, we also see that using no feature selection technique, the pAUC value is lower. Therefore there must be a point at which adding features to the 5NN classifier GR feature selection technique combination begins to negatively impact classifier results as we see for other classifier and feature selection technique combinations. Also, we would like to point out that the slope of the pAUC curve for 5NN with GR in figure 4 diminishes for number of features equal to 50 and above.

- NB classifier with RF feature selection and NB classifier with RFW feature selection we see that the number of features associated with the highest pAUC value is 200. However the respective plots in Figures 20 and 22 and the pAUC value for NB classifier with no feature selection technique are similar to what we stated in the point above regarding 5NN classifier with GR feature selection.

The table below is similar to Table 1, but we alter our query to search for maximum nAUC values in order to generate it.

Classifier	FS Technique	Num. Features	nAUC	FPR	FNR	Figure
NB	SU	100	0.978559	0.123288	0.0434783	24
5NN	RFW	20	0.977963	0.0136986	0.26087	10
NB	CS	20	0.976772	0.0547945	0.0434783	14
NB	GR	100	0.976176	0.109589	0.0434783	16
NB	IG	100	0.973794	0.109589	0.0434783	18
5NN	CS	20	0.972007	0.0136986	0.304348	2
5NN	SU	100	0.961584	0.0684932	0.130435	12
5NN	IG	7	0.957117	0.0410959	0.304348	6
5NN	RF	20	0.94461	0	0.521739	8
5NN	GR	200	0.936569	0.0547945	0.173913	4
NB	RFW	200	0.926147	0.123288	0.0869565	22
NB	RF	200	0.921977	0.219178	0.0869565	20
5NN	No feature selection	4026	0.882073	0.109589	0.391304	(N/A)
NB	No feature selection	4026	0.82698	0.178082	0.26087	(N/A)
J48	Embedded feature selection	6	0.780524	0.0958904	0.391304	(N/A)

Table 2: AUC values associated with negative (nonACL) class by classifier and feature selection (FS) technique. The figure column indicates which figure in this document the reader may to refer to see the data in the associated row in graphical form. N/A stands for, "not available."

We notice the following for Table 2:

- For the NB classifiers and GR, IG, and SU feature selection techniques, the number of features increases from 20 to 100.
- The positions of NB classifiers and their feature selection techniques in Table 2 differ from their positions in Table 1 since pAUC values can be different from nAUC values.

The next table presents results for the weighted AUC value we recorded for our experiments. It is generated in a manner similar to Tables 1 and 2.

Classifier	FS Technique	Num. Features	wAUC	FPR	FNR	Figure
5NN	RFW	20	0.977963	0.0136986	0.26087	10
NB	SU	20	0.977367	0.109589	0.0434783	24
NB	CS	20	0.976772	0.0547945	0.0434783	14
NB	GR	20	0.972603	0.0547945	0.0869565	16
5NN	CS	20	0.972007	0.0136986	0.304348	2
NB	IG	50	0.966107	0.136986	0.0434783	18
5NN	SU	100	0.961584	0.0684932	0.130435	12
5NN	IG	7	0.957117	0.0410959	0.304348	6
5NN	RF	20	0.94461	0	0.521739	8
5NN	GR	200	0.936569	0.0547945	0.173913	4
NB	RFW	200	0.930449	0.123288	0.0869565	22
NB	RF	200	0.926733	0.219178	0.0869565	20
5NN	No feature selection	4026	0.882073	0.109589	0.391304	(N/A)
NB	No feature selection	4026	0.840341	0.178082	0.26087	(N/A)
J48	Embedded feature selection	6	0.777807	0.0958904	0.391304	(N/A)

Table 3: weighted average AUC values associated by classifier and feature selection (FS) technique. The figure column indicates which figure in this document the reader may to refer to see the data in the associated row in graphical form. N/A stands for, "not available."

We notice that the number of features for optimum wAUC value for NB classifier with IG, SU, or GR feature selection techniques is again different from the optimum number of features for nAUC value in Table 2, changing to 50, 20, and 20, respectively.

Influence of Classifier and Ranker

Assignment 4 gives the requirement, "In addition, discuss how these changes [in number of features for optimum AUC metric] are influenced by the choice of classifier and ranker." We see from the tables above, that for both NB classifier and the 5NN classifier, we get a higher AUC metric when we apply a feature selection technique to our data before feeding the data as input to the classifier. Specifically, one can compare the AUC value in Tables 1, 2, or 3, for rows with, "No feature selection," to rows that list a feature selection technique and see that the AUC value is always higher.

The best performing classifier and ranker combination depends on which AUC metric we are interested in. We see in Tables 1 and 3 that the overall best performing ranker is 5NN with RFW feature selection. However in Table 2 it is NB with SU that has highest nAUC value. However, the relative positions of the classifiers and rankers do not change when we fix the classifier. That is to say, in Tables 1, 2, and 3, the relative order of the NB classifiers and rankers is the same, and the relative order of the 5NN classifiers and rankers is the same. This makes the case that we have a clear best choice for feature selection technique if we are decided on which classifier to use. For the NB classifier, we see the SU feature selection technique as the best choice, and for the 5NN classifier, we see the RFW feature selection technique as the best choice. It is also interesting to note that the best feature selection technique for one classifier is close to the worst feature selection technique for another. For example, RFW is the second to worst feature selection technique for the NB classifier.

In order to formally analyze the influence of classifier and ranker, we used R to do ANOVA on three different linear models where we used one of the three AUC metrics as the response value and classifier, feature selection technique, and number of features retained as factors, as well as all possible interactions. For all three factors and interactions the ANOVA results, in terms of significance, are the same. The table below characterizes the results we find. Source code for queries to generate data for the linear models as well as the R commands to generate the ANOVA results are available in the source code repository for

our project. The ANOVA results all indicate that classifier, number of features selected, and feature selection technique are all significant in terms of AUC. ANOVA results also indicate that the interaction between feature selection technique and number of features selected is significant.

Factor / Interaction	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Classifier	2	0.031912	0.0159561	15.9586	1.039e-06
FS Technique	6	0.127432	0.0212387	21.2420	1.005e-15
Num. Features	1	0.027247	0.0272469	27.2511	1.033e-06
Classifier:FS Technique	6	0.005772	0.0009620	0.9621	0.455128
Classifier:Num. Features	1	0.002552	0.0025519	2.5523	0.113424
FS Technique:Num. Features	5	0.019048	0.0038096	3.8101	0.003427
Classifier:FS Technique:Num. Features	5	0.002818	0.0005637	0.5638	0.727509
Residuals	96	0.095985	0.0009998		

Table 4: Anova results for pAUC metric; results for nAUC, wAUC metrics are similar in that the same set of features and interactions are significant at the P=0.5 level. Interactions between factors are indicated using a ':' to separate factors.

As one's intuition might suggest, ANOVA results imply that the choice of ranker, feature selection technique, the number of features to retain are all significant. Also, the interaction between feature set size and feature selection technique is also significant in terms of AUC.

Best Feature Ranker for 6 Features, and Which Features

For part 2 of Assignment 4, our first requirement is, "For each classifier used in Part 1 of this assignment, and when choosing 6 features, what is the best feature ranker in terms of AUC?"

In order to answer this question we can query the results we have stored in our database. The query to obtain the name of the ranker is included in the sql area of source code we

wrote for this report `https://github.com/jhancock1975/data-mining-assignment-2/tree/master/cap-6778/src/main/resources/sql`. For 5NN classifiers and 6 features, the best feature ranker in terms of AUC is IG. The AUC values we record for pAUC, nAUC, and wAUC are all 0.946.

For NB classifiers and 6 features the the best feature ranker in terms of AUC is also IG. For the combination of NB and IG we get a value of 0.968 for all the variations of AUC metric.

Assignment 4 also requires that we report on which features are selected.

For IG feature selection technique, the features selected are: GENE1610X, GENE3943X, GENE493X, GENE390X, GENE2760X, GENE1537X.

Feature sets with 6 Elements & Compared with J48

Assignment 4 requires a comparison of the features that the attribute selection techniques return when we fix the number of features at 6, and the features that the J48 classifier returns.

Since we store the lists of features each feature selection technique generates in a database when we run the classifiers, it is convenient to write a query to answer this quest. The query is included with source code at the URL in the previous section. The table below gives the result of the query

FS Technique	J48 overlap
RFW	GENE1567X
GR	GENE3941X, GENE1125X
RF	GENE1567X
SU	GENE3941X, GENE1567X
IG	No overlap
CS	No overlap
J48	GENE1125X, GENE1391X, GENE1567X, GENE2996X, GENE3732X, GENE3941X

Table 5, overlap of features selected by various feature selection. The genes J48 selects are listed as well in order to partially justify the overlap we claim, as required in Assignment 4.

We notice that gene GENE1567X occurs most frequently in the table above. This is a strong indication that GENE1567X is an important factor in predicting the class of an instance.

We have a requirement to show our results and justify our conclusion. We have included a listing of all length-6 feature lists that the feature selection techniques generate. The justification for our conclusion is that we checked query results against genes listed below to ensure genes in the table above are indeed in lists in the appendix. We also checked query results against genes in attribute lists in arff files we generate using Weka to do feature selection using the Weka Explorer, and genes listed in classifier output when we run the J48 classifier using the Weka Explorer. In performing these checks we are able to verify that the genes listed in table 4 above are indeed the genes that overlap with what the J48 classifier finds, and that the data in table 4 is exhaustive. This lends confidence in the query results so that we feel comfortable to expand our results to feature subsets of larger size in future work.

To verify that we have no overlapping features between the list J48 generates and the lists IG and CS generates we set these lists down below for the reader's inspection:

FS Technique	FS List
RFW	GENE1567X,GENE1610X,GENE1622X,GENE1637X,GENE1658X,GENE2462X
CS	GENE1537X,GENE1609X,GENE1610X,GENE384X,GENE385X,GENE493X
GR	GENE1125X,GENE3753X,GENE3754X,GENE3755X,GENE3941X,GENE3944X
IG	GENE1537X,GENE1610X,GENE2760X,GENE390X,GENE3943X,GENE493X
RF	GENE1567X,GENE1609X,GENE1610X,GENE1622X,GENE1658X,GENE2462X
SU	GENE1537X,GENE1567X,GENE1609X,GENE385X,GENE3941X,GENE493X
J48	GENE1125X,GENE1391X,GENE1567X,GENE2996X,GENE3732X,GENE3941X

Table 6 listing of genes to justify overlap for for 6 element feature lists generated with the 6 feature selection techniques versus the embedded feature list generated from J48.

Conclusions

The results of the experiments that we are required to conduct for Assignment 4 lead us to conclude the following:

- Not surprisingly, ANOVA shows that choice of classifier, feature selection technique and the number of features to select are all significant in the resulting AUC value for a classification run.
- ANOVA also shows that the interaction between feature selection technique and the number of features to select has an impact on AUC.
- The overlap study we are required to conduct with the genes that the 6 filter based feature selection techniques we use and the genes that the J48 classifier retain shows GENE1567X is most frequently retained for feature subset size 6 for the feature selection techniques we use in this study.

- The choice of classifier and feature selection technique together are important to optimize AUC values as we see in Table 3. The best feature selection technique to use with the NB classifier is SU, whereas the best feature selection technique to use with the 5NN classifier is RFW.

Future Research

The Java programs we wrote for this research are flexible in that the lists of feature selection techniques and classifiers to use are specified independently of the program that runs classifications. Therefore we can easily expand this work to encompass more filter based feature selection techniques, and classifiers. In addition the number of features to use is also easily configurable, so we could collect more data points for the plots we have in Appendix 1.

At present, our program only stores a subset of the data in a Weka API Evaluation object. We would like to use Hibernate tools to generate a schema for, and serialize all data members of an Evaluation object to database. This would more flexibility in enabling researchers to search for more results after running experiments where the questions researchers ask after running the experiments are not necessarily ones they had in mind before conducting their experiments.

The results of these experiments show that using NB and 5NN classifiers with a feature selection technique enhances performance. We are curious to know if we can combine NB and 5NN in a meta-classifier with feature selection techniques to get even better performance.

References

[1] I. Witten and E. Frank, Data Mining (second edition). San Francisco: Elsevier, 2005, ch. 4 pp.89-91

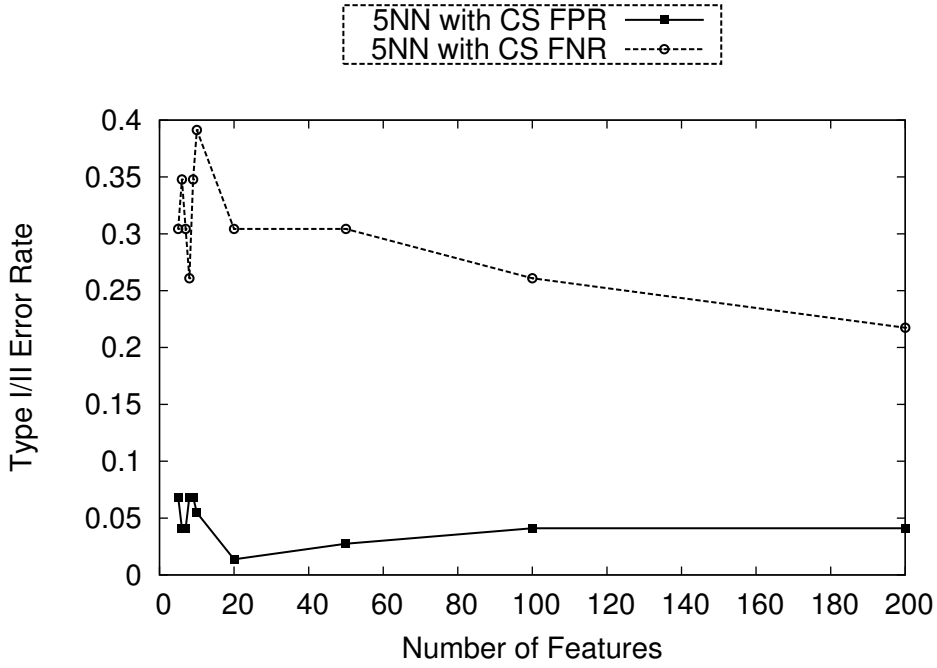
[2] I. Witten and E. Frank, Data Mining (second edition). San Francisco: Elsevier, 2005, ch. 4 pp.128-136

[3] J. R. Wald, T. M. Khoshgoftaar, A. Abu Shanab, "The Effect of Measurement Approach and Noise Level on Gene Selection Stability," in IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Philadelphia, PA, 2012. doi: 10.1109/BIBM.2012.6392713

[4] I. Witten and E. Frank, Data Mining (second edition). San Francisco: Elsevier, 2005, ch. 5 p.169 fig. 5.2.

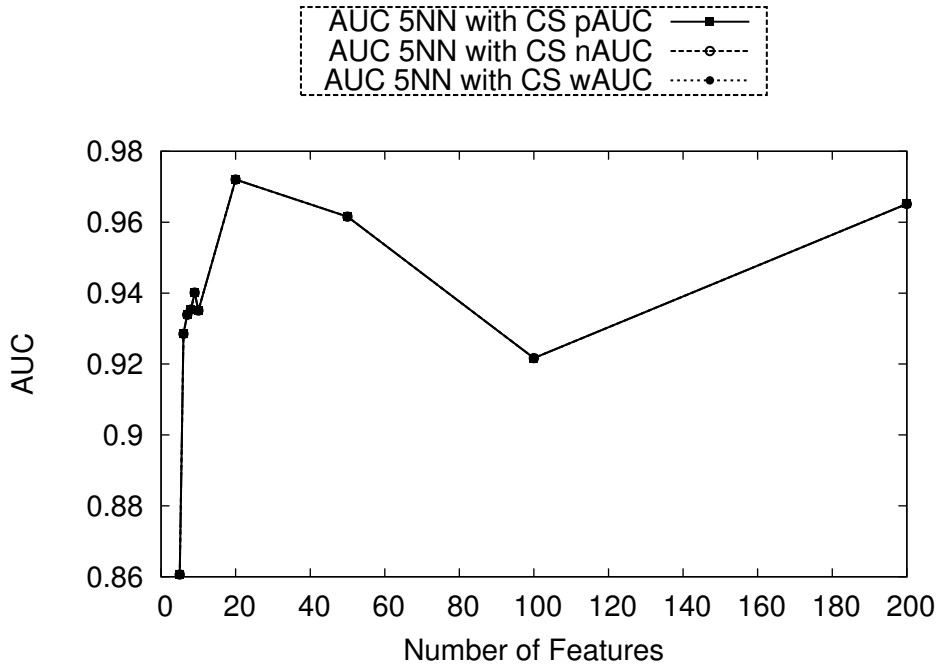
Appendix 1: Complete Listing of All Graphs

5NN with CS Attribute Selection



Error Rates

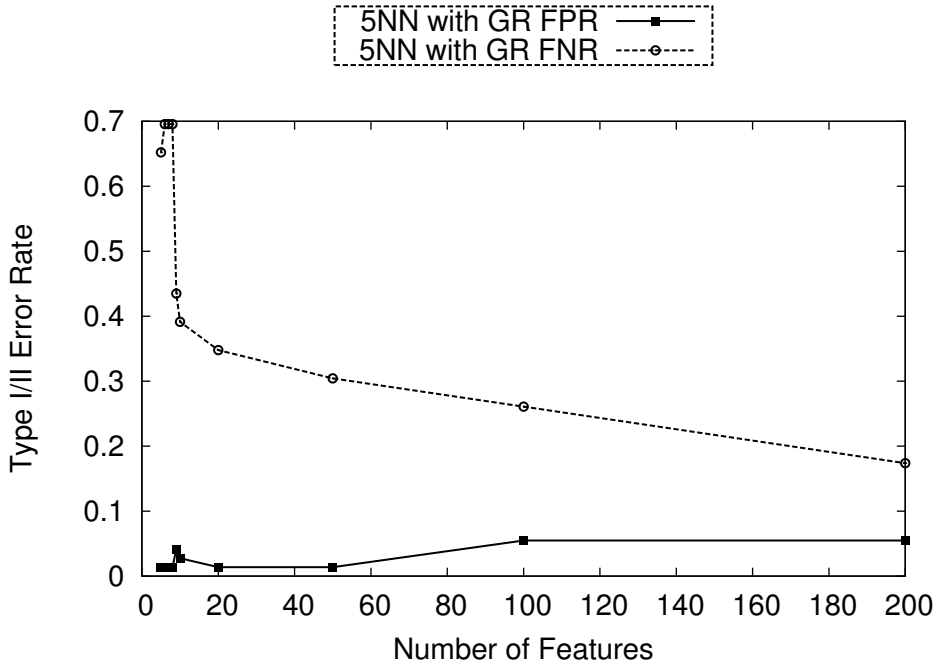
Figure 1



AUC

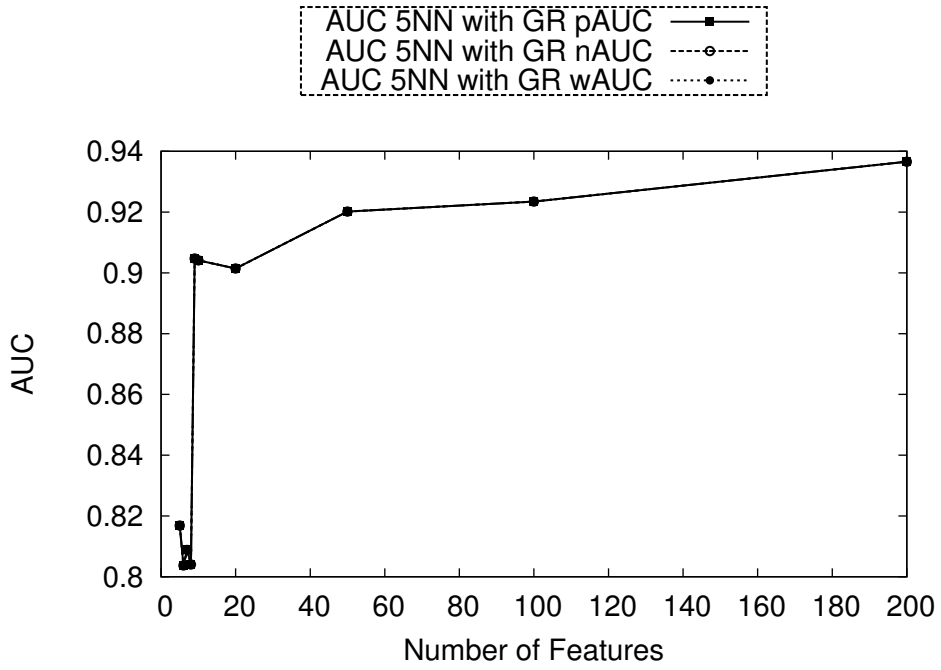
Figure 2

5NN with GR Attribute Selection



Error Rates

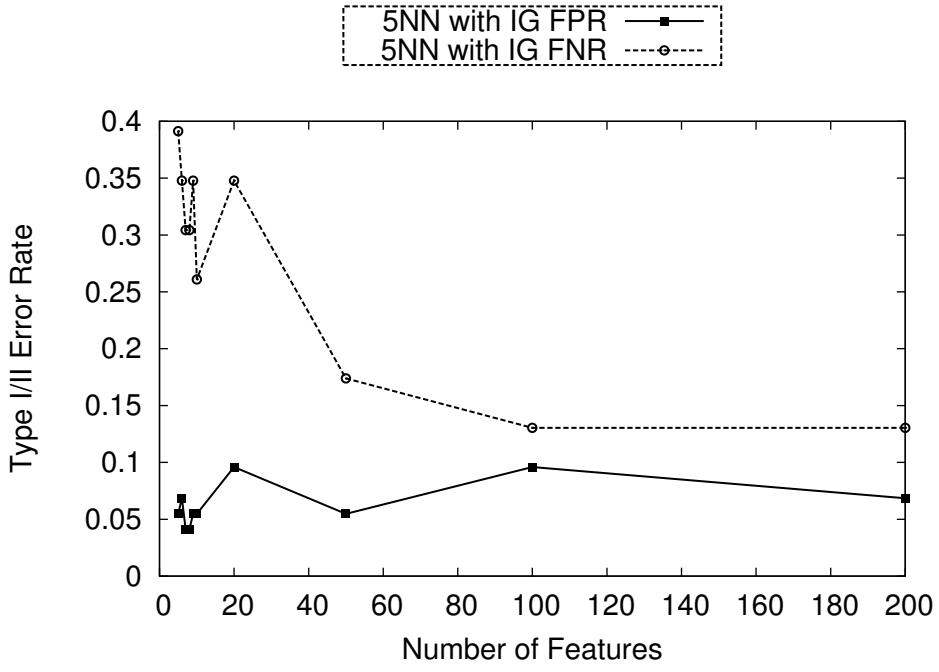
Figure 3



AUC

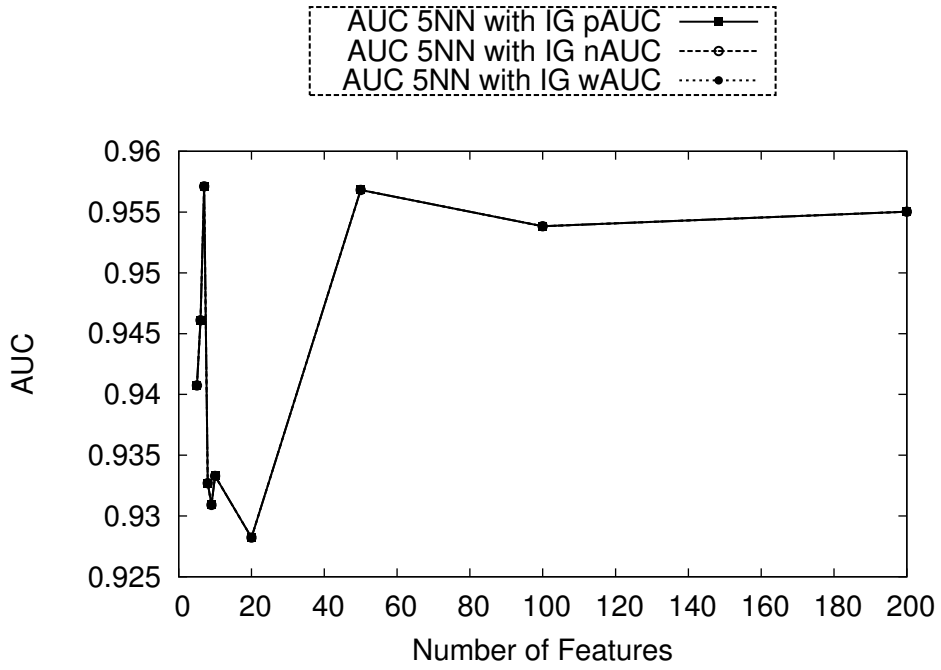
Figure 4

5NN with IG Attribute Selection



Error Rates

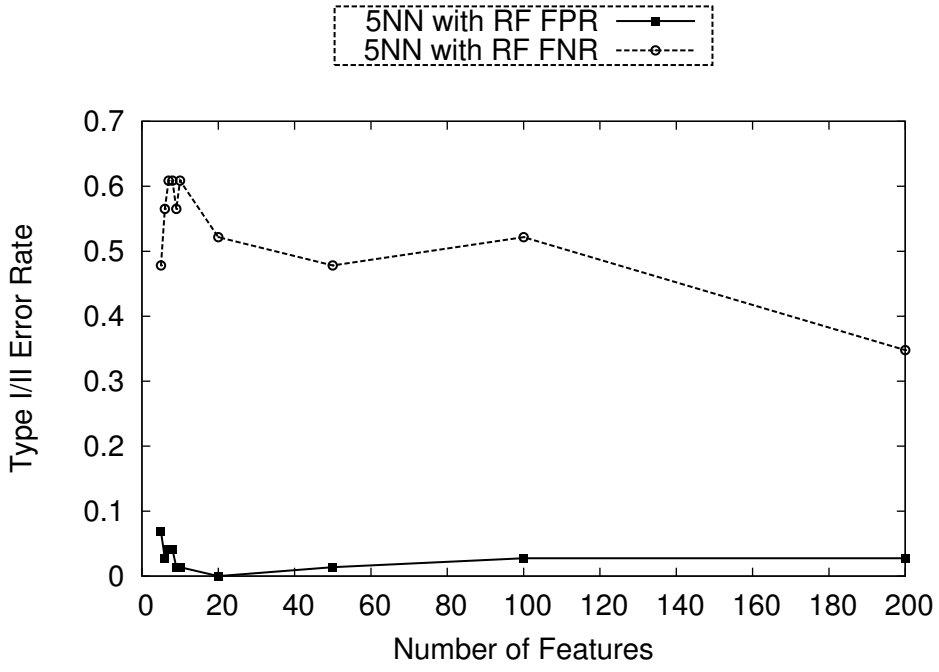
Figure 5



AUC

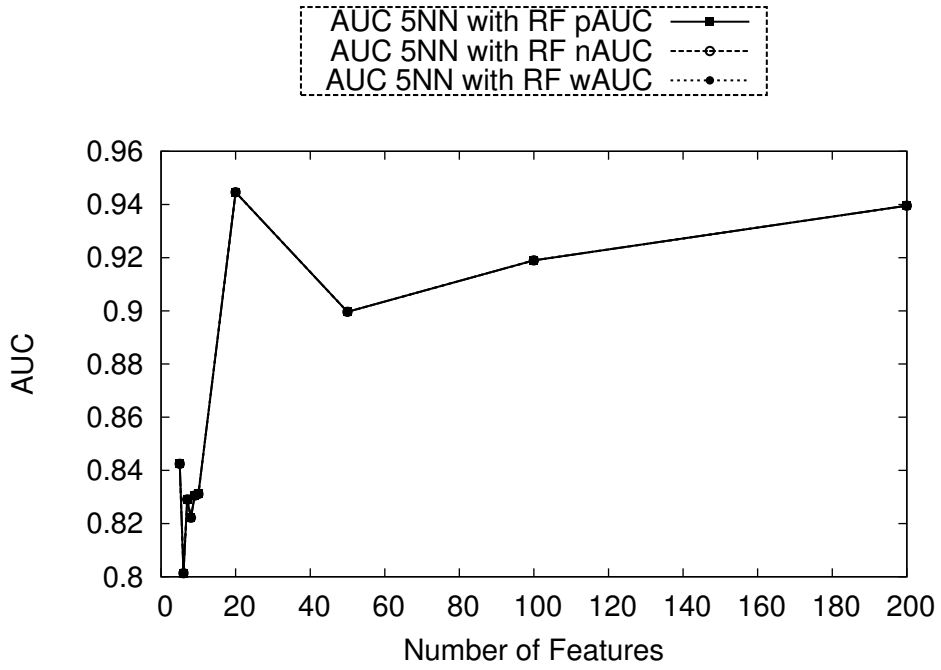
Figure 6

5NN with RF Attribute Selection



Error Rates

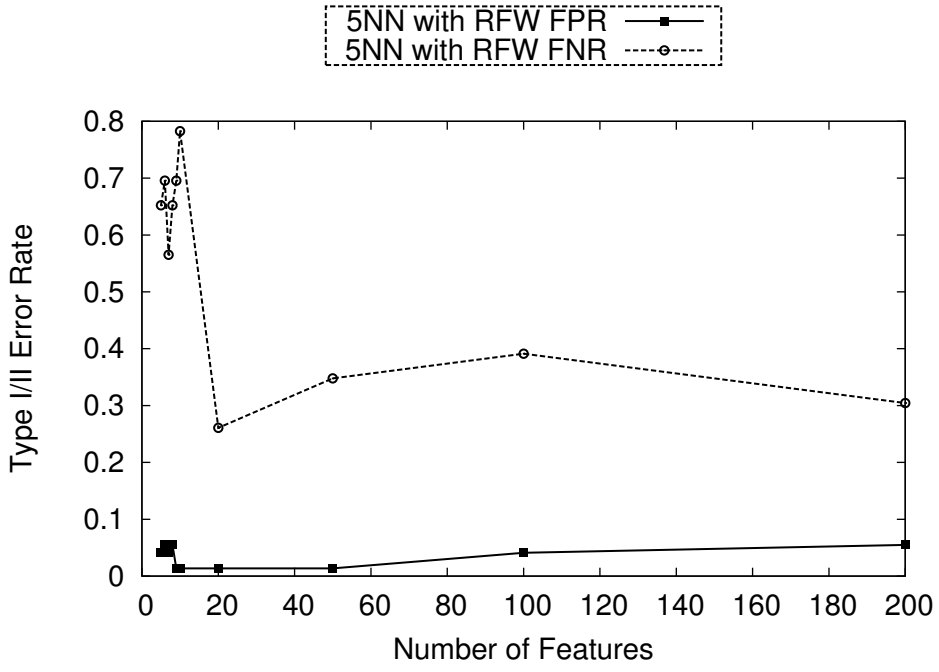
Figure 7



AUC

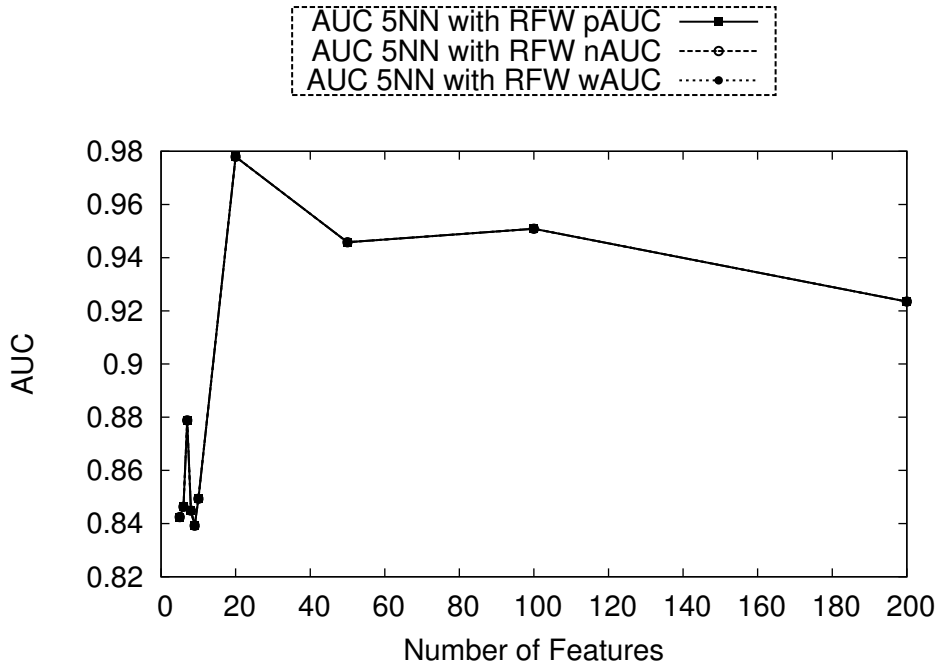
Figure 8

5NN with RFW Attribute Selection



Error Rates

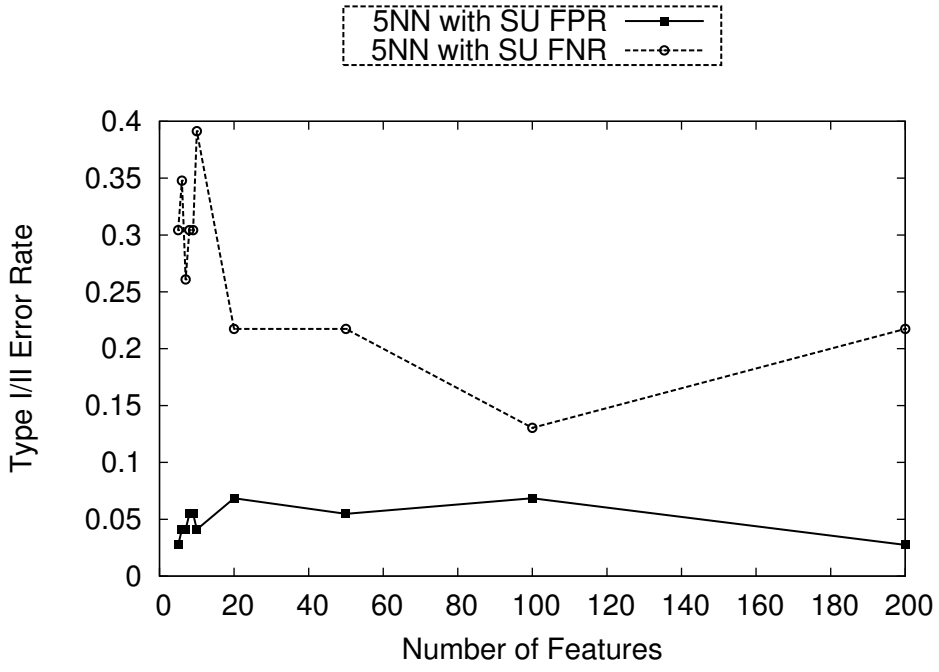
Figure 9



AUC

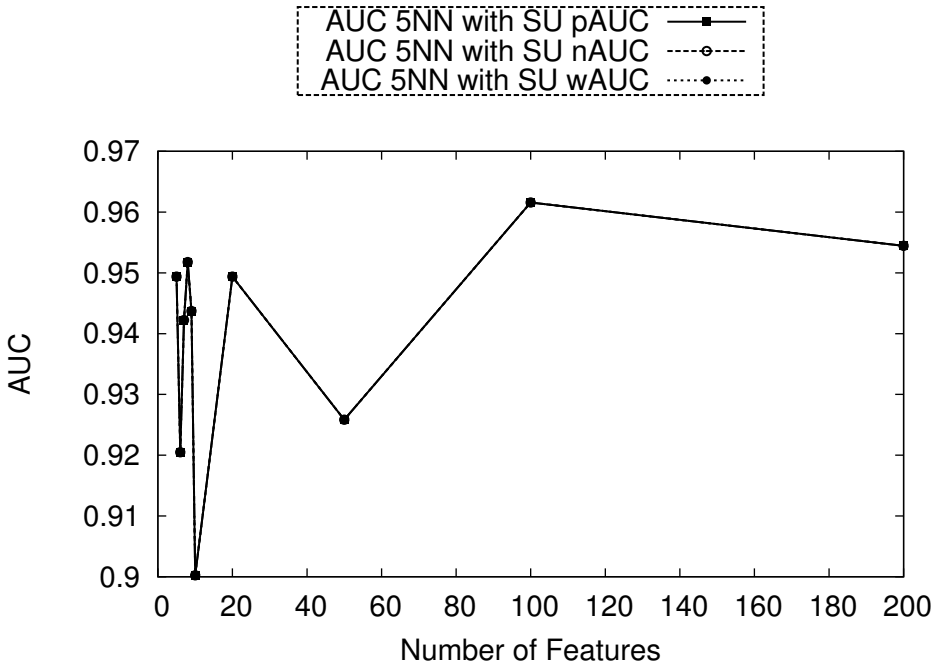
Figure 10

5NN with SU Attribute Selection



Error Rates

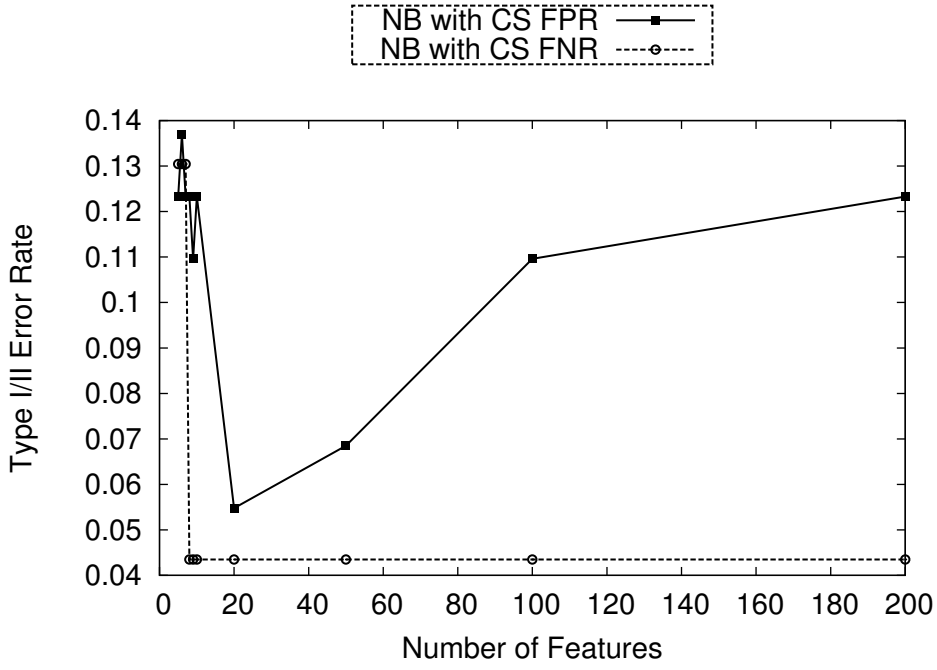
Figure 11



AUC

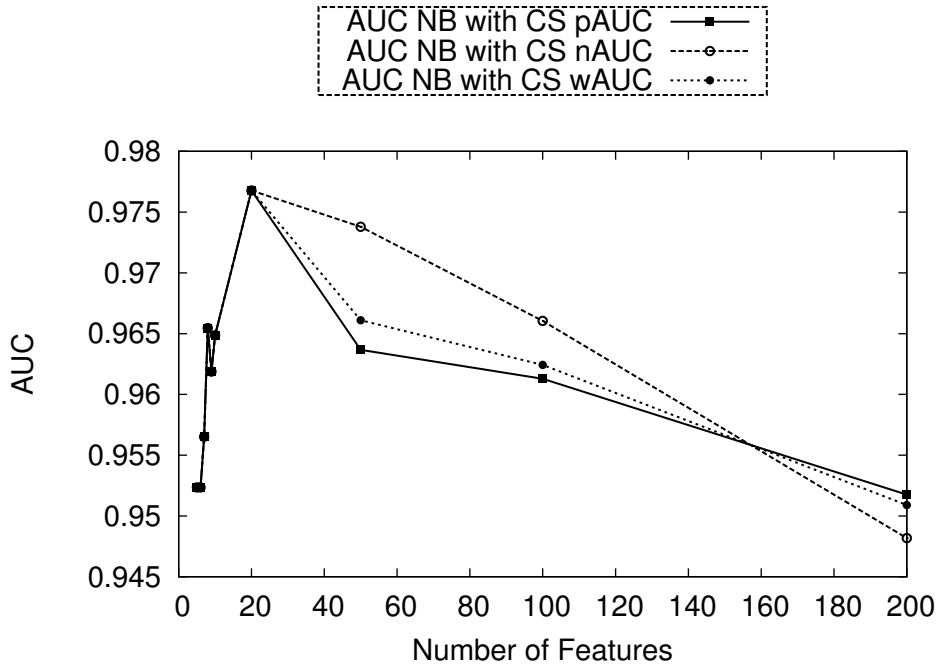
Figure 12

NB with CS Attribute Selection



Error Rates

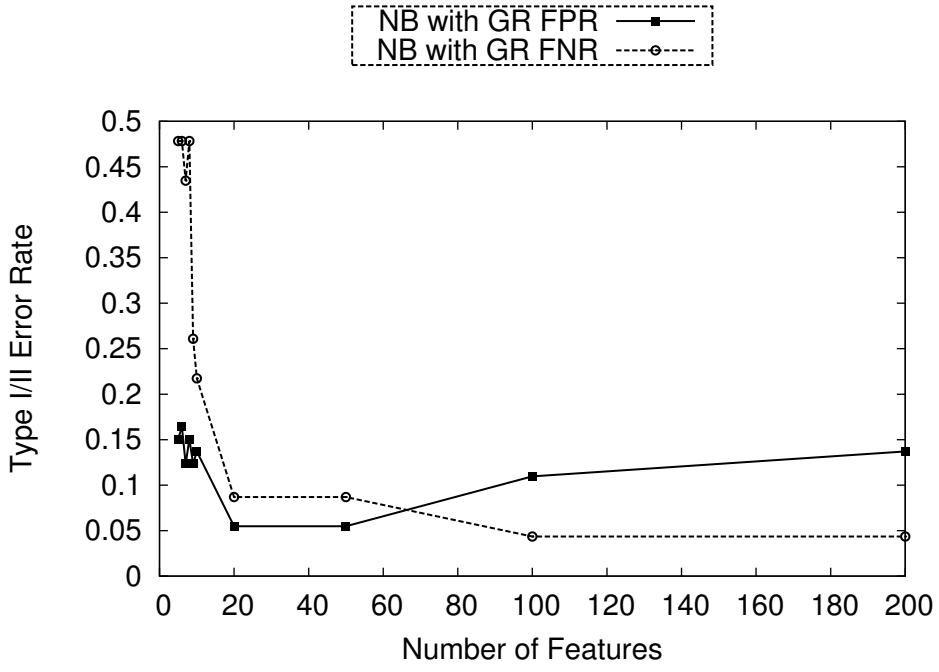
Figure 13



AUC

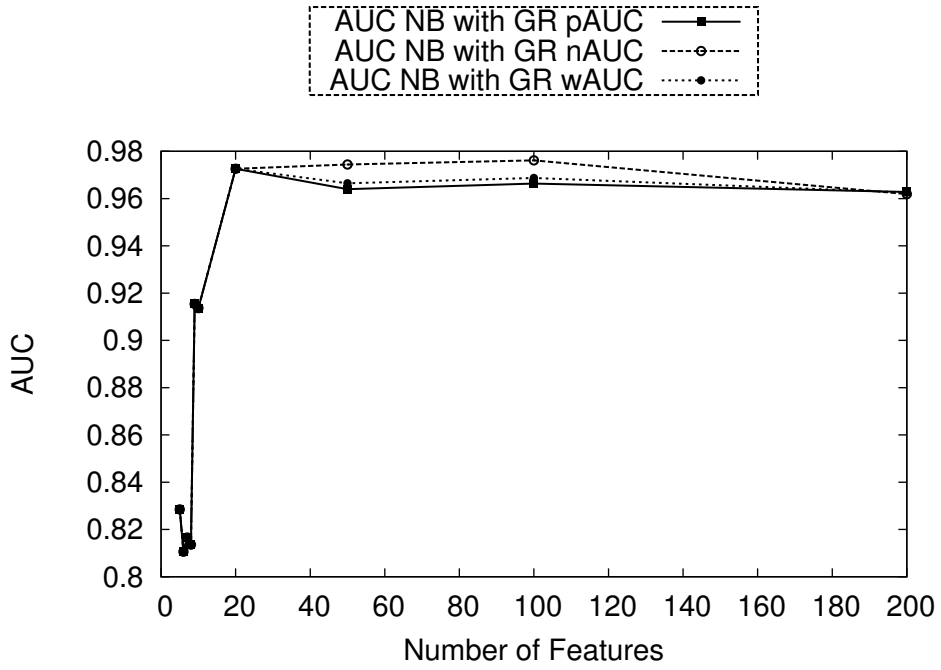
Figure 14

NB with GR Attribute Selection



Error Rates

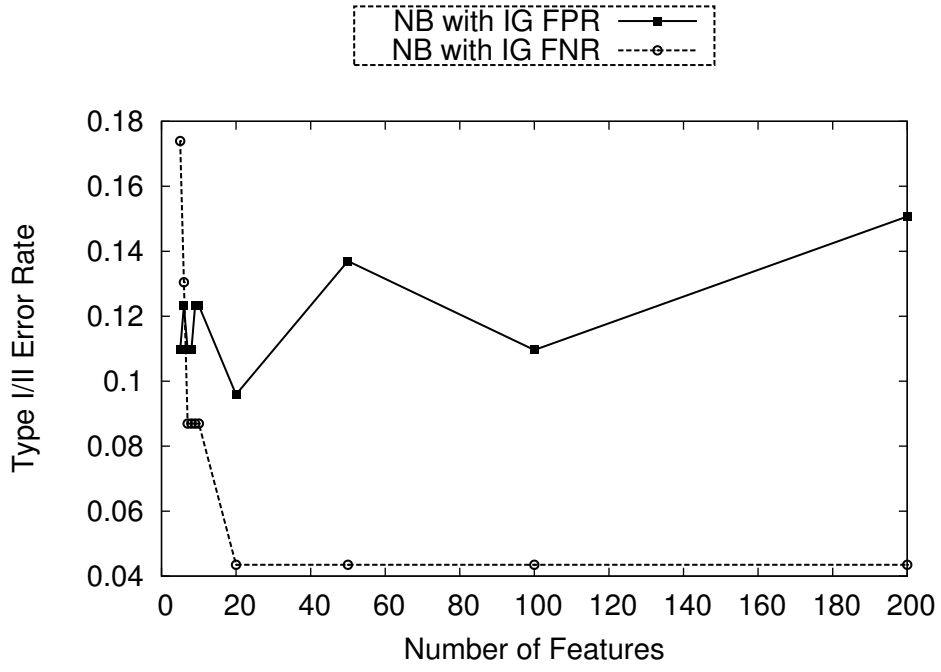
Figure 15



AUC

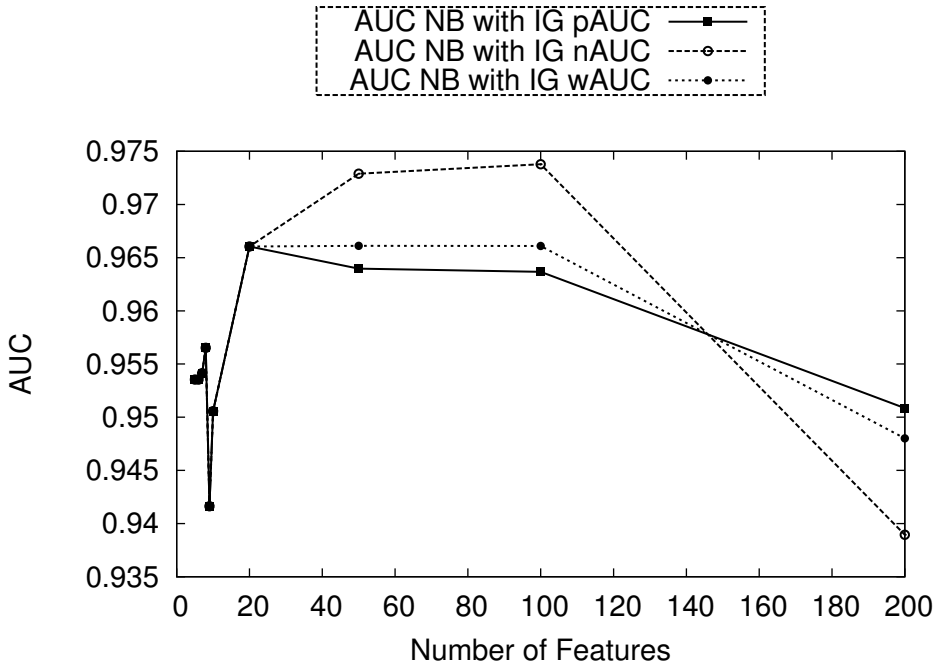
Figure 16

NB with IG Attribute Selection



Error Rates

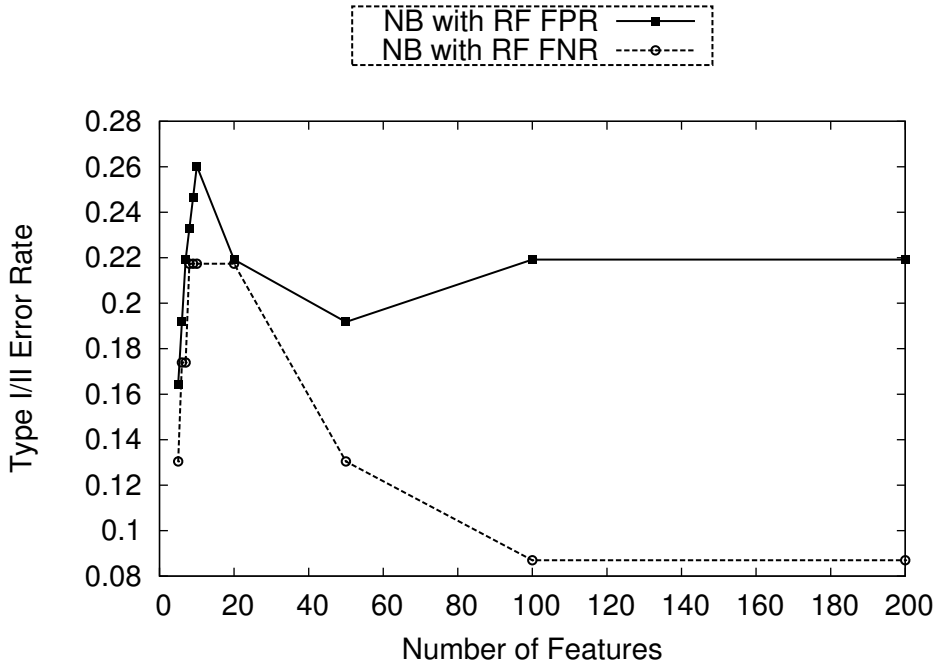
Figure 17



AUC

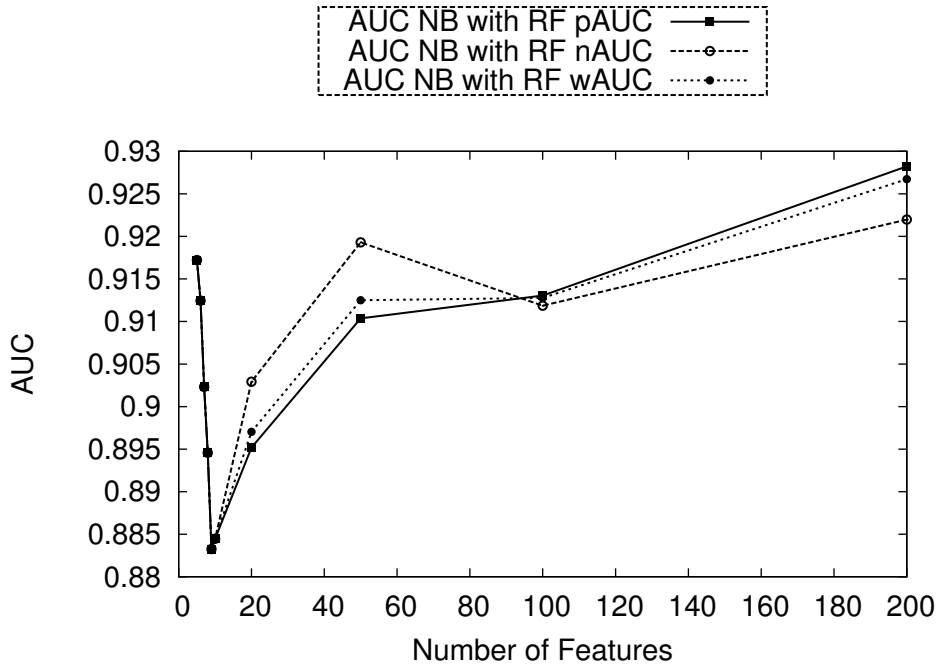
Figure 18

NB with RF Attribute Selection



Error Rates

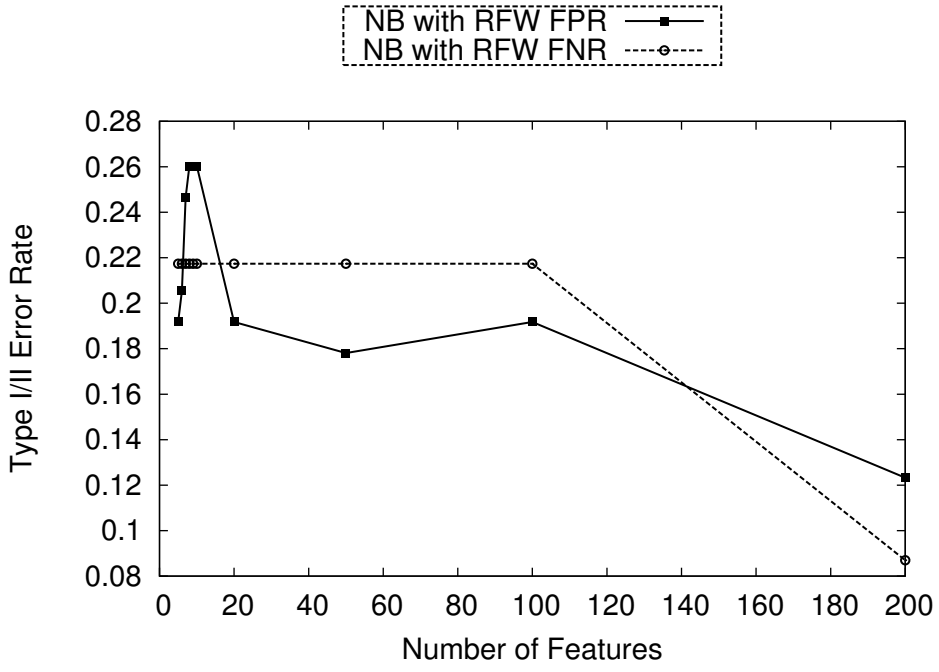
Figure 19



AUC

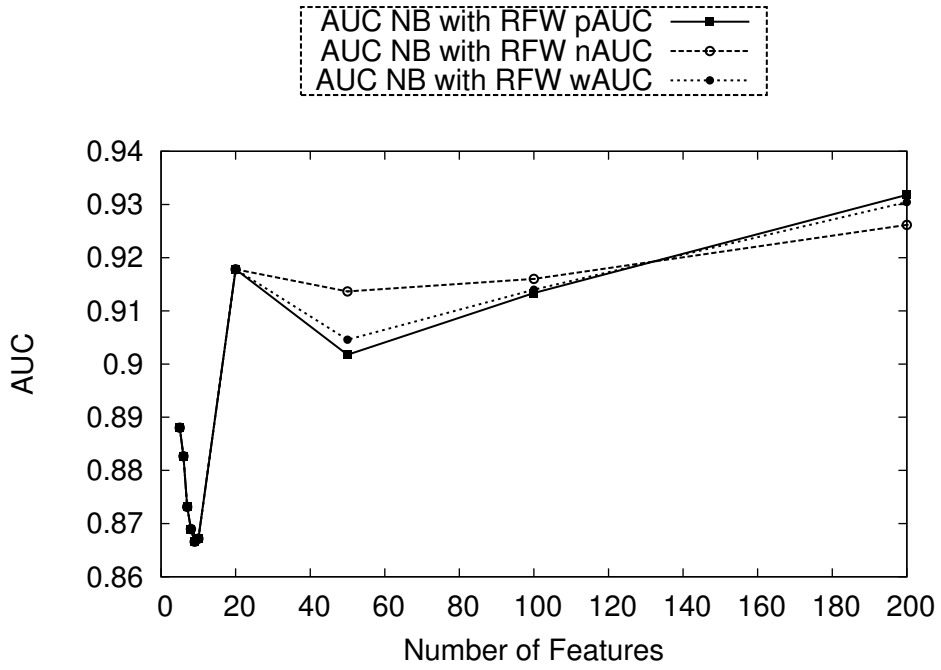
Figure 20

NB with RFW Attribute Selection



Error Rates

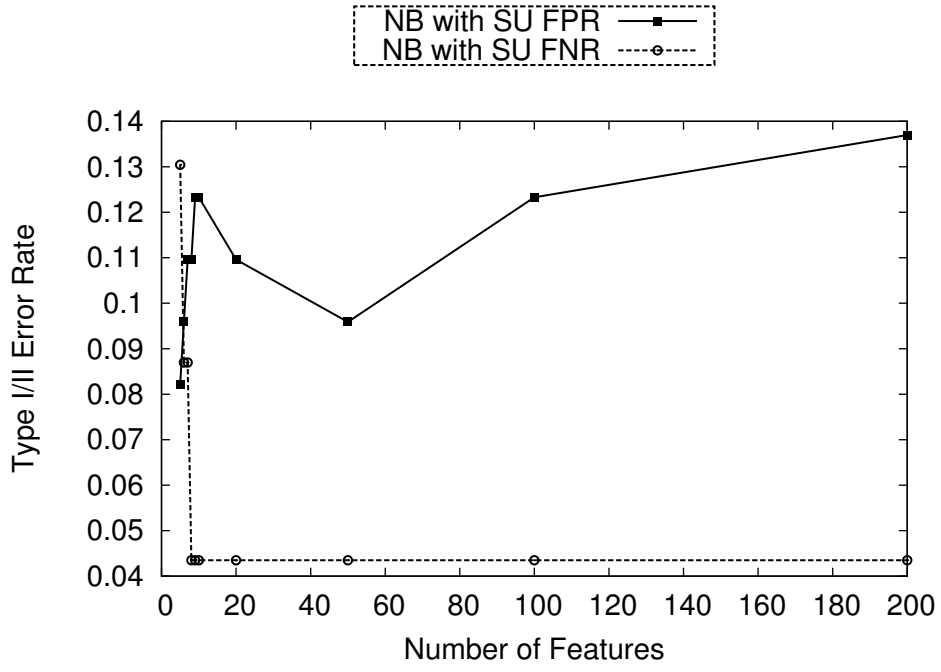
Figure 21



AUC

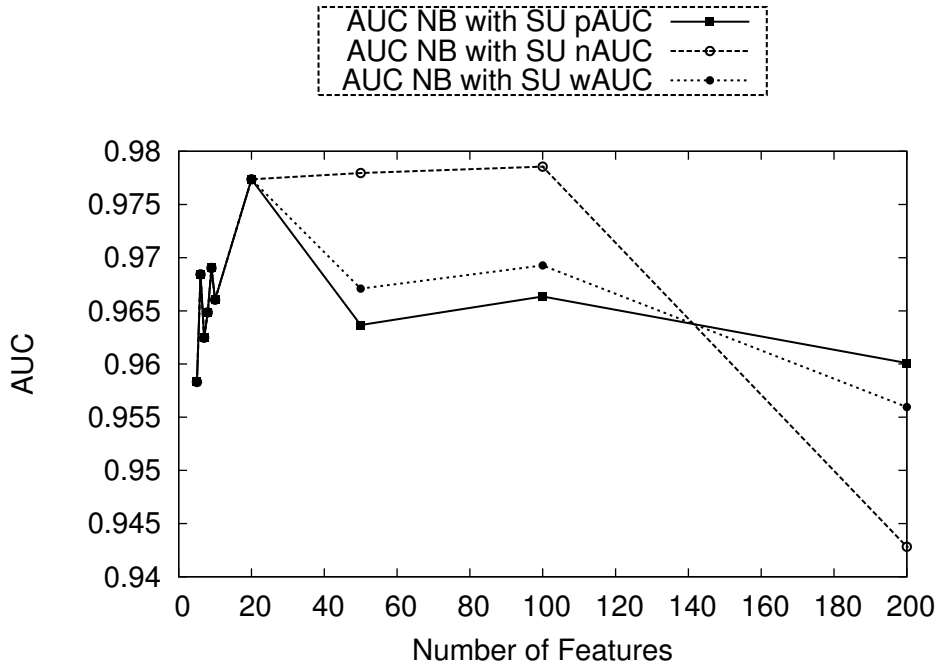
Figure 22

NB with SU Attribute Selection



Error Rates

Figure 23



AUC

Figure 24