

An Extensive Study on Data Thinning in Single and Ensemble Classifiers

¹Indulekha E Thambi, ²R. Anju and ³Ranjidha Rajan

¹Department of Computer Science & IT,
School of Arts and Sciences, Amrita University, Kochi.

indulekhaethambi@gmail.com

²Department of Computer Science & IT,
School of Arts and Sciences, Amrita University, Kochi.

anjurevathy15@gmail.com

³Department of Computer Science & IT,
School of Arts and Sciences, Amrita University, Kochi.

ranjidharajan@gmail.com

Abstract

Nowadays feature selection techniques have become a significant part in many data mining and analytics applications. Most of the data lakes are flooded with relevant and non-relevant data. For most of the machine learning methods need filtering of relevant data into the model, which implies the importance of Data Thinning or feature selection. Feature selection perform an important role in improving the performance of machine learning algorithms in terms of reducing the time to build the learning model and increasing the accuracy in the learning process. Therefore, for improving the performance of machine learning algorithms researchers pay more attention for feature selection. Classification is a data mining task which is an effective method to classify the data in the process of knowledge data discovery. The classification accuracy and execution time are two important parameters in the selection of classification algorithms. In our experiments, 5 different ensemble algorithms, and 5 single classifiers are compared according to their accuracies and test time over 4 datasets. The results show that ensemble classifiers when combined with correlation based feature selections has the highest accuracy.

Key Words: Classification, data thinning, ensemble, feature selection, random forest.

1. Introduction

Feature selection or data thinning is actually a technique of data reduction. It provides a reduced representation of data. Through feature selection we can reduce the number of features in a data set and also speed up the learning process thus making the dataset easier to understand. It may be performed by using several trial and error of all possible subsets of the original dataset. Feature selection is performed on original dataset by several trial and error methods. And there are several systematic methods to manage this process. The potential difficulties in feature selection are small sample space and confusion while choosing objective function. Number of features is directly proportional to the feature subset that is, in the case of an increase in features, there may be an increase in feature subset also.

2. Necessity of Feature Selection

Internet has been widely used by all the people and it has become very popular in this last decades. So there exist large amount of data. So it is very difficult to select relevant features from the large amount of data. Many researchers have conducted study on these to extract data. One of the solutions to this problem is feature selection.

When we organize data in high dimensional spaces we require high memory space and we need high computational cost. Usually a dataset is represented by a matrix where the rows are the instances and columns represent the features. So to find the problem of dimensionality we can make matrices of dataset narrower. The process of reducing the number of random variables is known as dimensionality reduction.

One of the dimensionality reduction techniques is feature extraction. It consists of two steps: construction and selection. Feature construction means converting the data into a set of useful features. During feature construction phase some construction methods do not lose information. Reducing the number of initial features is the main goal of feature selection methods. So to deal with large amount of data and features in our society in all areas there is need for a solution, which is feature selection. Next section briefly describes feature selection techniques.

3. Feature Selection Technique

The main goals of feature selection are:

- 1) To improve model performance
- 2) To avoid over fitting
- 3) Gaining a deeper insight into the underlying processes that generated the data.

Feature selection technique is of three categories: filter methods, wrapper

methods and embedded methods.

A. Filter Method

Filter methods select only the best feature from a subset of features. It is one of the feature ranking techniques. Filter methods assess the relevance of features using measures like distance, information, correlation and consistency. Filter methods are very fast in computation and easy to interpret, and they are scalable for high-dimensional datasets like micro array data.

Table I: Advantages and disadvantages of filter method

Feature selection method	advantage	disadvantage	examples
Filter : univariate	Fast. Scalable. Independent of the classifier.	Ignores feature dependencies. Ignore interaction with the classifier.	Chi-square test. Euclidean distance. I – test. Information gain, Gain ratio.
Filter: Multivariate	Models feature dependencies Independent of the classifier Better computational complexity than wrapper methods	Slower than univariate techniques Less scalable than univariate techniques Ignores interaction with the classifier	Correlation-based feature selection (CFS). Markov blanket filter (MBF) Fast correlation-based feature selection (FCBF).

B. Wrapper Method

The wrapper method is slower and expensive than filter method. The main advantage of wrapper method is interaction between the feature subsets and maintaining the dependencies between the features. They are better in defining optimal features rather than simply relevant features. To remove insignificant features from the subset wrapper method uses the backward elimination. To identify relevant feature wrapper method, need some predefined learning algorithm. Some of the wrapper selection methods are

- 1) SFS-sequential forward selection
- 2) Genetic algorithms
- 3) SBE-sequential backward elimination
- 4) Randomized hill climbing

Table II: Advantages and disadvantages of wrapper method

Feature selection method	Advantage	disadvantage	example
Wrapper: deterministic	Simple Interacts with the classifier. Model feature dependencies. less computationally intensive than randomized methods	Risk of over fitting More prone than randomized algorithms to getting stuck in a local optimum (greedy search) Classifier dependent selection	Sequential forward selection (SFS). Sequential backward elimination. (Beam search).
Wrapper: randomized	Less prone to local optima	Computationally intensive Models feature dependencies Classifier dependent selection Higher risk of overfitting than deterministic algorithms.	Genetic algorithm. Estimation of distribution algorithm.

C. Embedded Method

Embedded method uses the independent criteria to decide the optimal subset to know cardinality. It captures feature dependencies and it not only considers the relationship between one input feature and the output feature but also searches locally for feature that allow better local discrimination. The main approach is to incorporate the feature selection as part of the training process. This approach has the advantage of including the interaction with the classification model, while at the same time being far less computationally intensive than wrapper methods. Compared to wrapper approach, embedded methods reduce the cost of computation.

Table III: Advantages and disadvantages of embedded method

Feature selection method	Advantage	disadvantage	example
embedded	Interacts with the classifier better computational complexity than wrapper method models feature selection dependencies	Classifier dependent selection.	Decision tree Weighted naive Bayes Feature selection using the weight vector of SVM.

4. Methods for Selecting Feature

Here we done a comparative study in single and ensemble classification algorithms. And also describes all the single and ensemble classification algorithms that we used in this experiment. In our work we used 5 single and 5 ensemble classification algorithms.

A. Single Algorithms

Below table shows the set of single algorithms and their short form.

Table IV: Single algorithms

Algorithm name	Abbreviation
ZeroR	ZR
Naive Bayes	NB
Support Vector Machine	SMO
C4.5 Decision Tree	J48
Random Tree	RT

B. Ensemble Algorithms

Below table shows the set of ensemble algorithms and their short form.

Table V: Ensemble algorithms

Algorithm name	Abbreviation
Random Forest	RNDF
Bagging	BG
Vote	Vote
AdaBoost	ADB
Stacking	Stacking

C. Experiment Dataset

All the datasets are taken from UCI repository. In our experiment we have taken 4 datasets and we evaluated these datasets by using above specified classifiers.

Table VI: Dataset

Datasets	No. of Attributes	No. of Instances
Anneal	39	898
Glass	10	214
Mushroom	23	8124
Zoo	18	101

D. Correlation based Feature Selection

For selecting features from dataset we use correlation based feature selection. Here we use attribute evaluator as CorrelationAttributeEval and search technique is ranker method. It is efficient feature selection technique that will filters good features from the set of features.

Table VII: Features selected using correlation based F.S

DATASET	FEATURE SELECTION TECHNIQUE	SEARCH TECHNIQUE	TOTAL ATTRIBUTES	SELECTED ATTRIBUTES
Anneal	Correlation Attribute Evaluator	Ranker	39	38
Glass	Correlation Attribute Evaluator	Ranker	10	9
Mushroom	Correlation Attribute Evaluator	Ranker	23	22
Zoo	Correlation Attribute Evaluator	Ranker	18	17

E. Information gain based Feature Selection

In information gain based feature selection the attribute evaluator is InfoGainAttributeEval and the search method is ranker.it is a commonly used feature selection method. But is less efficient with compared to other existing feature selection methods.

Table VIII: Features selected using information gain based F.S

DATASET	FEATURE SELECTION TECHNIQUE	SEARCH TECHNIQUE	TOTAL ATTRIBUTES	SELECTED ATTRIBUTES
Anneal	Information Attribute Evaluator	Ranker	39	38
Glass	Information Attribute Evaluator	Ranker	10	9
Mushroom	Information Attribute Evaluator	Ranker	23	22
Zoo	Information Attribute Evaluator	Ranker	18	17

F. Learner based Feature Selection Method

Learner based feature selection method selects best feature among other features. It filters only less number of attributes compared to all other feature selection methods. Here the attribute evaluator is wrapper subset evaluator and

the search method is best first.

Table IX: Features selected using learner based F.S

DATASET	FEATURE SELECTION TECHNIQUE	SEARCH TECHNIQUE	TOTAL ATTRIBUTES	SELECTED ATTRIBUTES
Anneal	Wrapper subset Evaluator	Best First	39	9
Glass	Wrapper subset Evaluator	Best First	10	5
Mushroom	Wrapper subset Evaluator	Best First	23	5
Zoo	Wrapper subset Evaluator	Best First	18	5

From the above tables it is clear that learner based feature selection is best as it selects more accurate features.

5. Experimental Results

These experiments are conducted using Weka tool. Cross validation is performed on each dataset to identify which algorithm has highest accuracy. We performed different feature selection methods on different datasets. After that we applied these results on different classifiers.

Table X: FS method-correlation based feature selection

	Anneal			glass		Mushroom		Zoo	
Type of classifier	Name of Algorithm	Accuracy (%)	Time (Sec)	Accuracy (%)	Time (sec)	Accuracy (%)	Time (Sec)	Accuracy (%)	Time (sec)
Single	ZeroR	98.99	0.01	35.51	0	38.74	0	40.59	0
Single	NB	88.86	0.03	47.66	0.01	64.69	0.03	95.04	0.01
Single	SMO	98.99	0.28	56.07	0.2	61.22	39.46	93.06	0.18
Single	RT	98.21	0.03	69.15	0	47.03	0.06	95.04	0.03
Single	J48	98.99	0.07	68.69	0.01	61.90	0.1	92.07	0.03
Ensemble	RF	98.77	0.2	79.90	0.08	47.30	5.54	97.02	0.04
Ensemble	Vote	98.99	0.01	35.51	0	38.74	0.01	40.59	0
Ensemble	Bagging	98.99	0.25	70.56	0.01	52.08	1.82	92.07	0.02
Ensemble	AdaBoost	98.99	0.1	44.85	0	55.04	0.17	60.39	0.01
Ensemble	Stacking	98.99	0.01	35.51	0	38.74	0.03	40.59	0.02

Table XI: FS method-information gain based feature selection

	Anneal			glass		Mushroom		Zoo	
Type of classifier	Name of Algorithm	Accuracy (%)	Time (Sec)	Accuracy (%)	Time (sec)	Accuracy (%)	Time (Sec)	Accuracy (%)	Time (sec)
Single	ZeroR	98.99	0	35.51	0	38.74	0	40.59	0
Single	NB	88.86	0.02	47.66	0	64.69	0.03	95.04	0
Single	SMO	98.99	0.06	56.07	0.14	61.22	44.58	96.03	0.18
Single	RT	98.21	0	69.15	0	47.03	0.01	66.33	0
Single	J48	98.99	0.02	68.69	0.02	61.90	0.07	95.04	0
Ensemble	RF	98.77	0.07	79.90	0.07	47.30	4.84	97.02	0.02
Ensemble	Vote	98.99	0	35.51	0	38.74	0	42.57	0
Ensemble	Bagging	98.99	0.04	70.56	0.01	52.08	1.43	92.07	0.03
Ensemble	AdaBoost	98.99	0.06	44.85	0	55.04	0.04	60.39	0
Ensemble	Stacking	98.99	0	35.51	0	38.74	0.02	40.59	0

Table XII: FS method-learner based feature selection

Type of classifier	Anneal			glass		Mushroom		Zoo	
	Name of Algorithm	Accuracy (%)	Time (Sec)	Accuracy (%)	Time (sec)	Accuracy (%)	Time (Sec)	Accuracy (%)	Time (sec)
Single	ZeroR	76.16	0	35.51	0	51.79	0	40.59	0
Single	NB	84.07	0.01	49.53	0	98.38	0	81.18	0
Single	SMO	97.32	0.43	52.80	0.08	98.84	2.6	83.16	0.19
Single	RT	98.77	0.01	68.69	0	99.24	0	81.18	0
Single	J48	98.10	0	70.56	0	99.24	0.01	82.17	0
Ensemble	RF	98.99	0.2	76.16	0.06	99.27	0.2	83.16	0.01
Ensemble	Vote	76.16	0	35.15	0	51.79	0	40.59	0
Ensemble	Bagging	97.77	0.1	73.36	0.01	99.24	0.1	82.17	0
Ensemble	AdaBoost	83.63	0.03	44.85	0	98.52	0.17	60.39	0
Ensemble	Stacking	76.16	0.02	35.51	0	51.79	0.02	40.59	0

From the above experiments of classifier algorithms, it is clear that the ensemble classifier is the best. It shows higher accuracy than single classifier. Even though for small datasets the time taken is less in zeroRule considering accuracy, random forest shows higher accuracy.

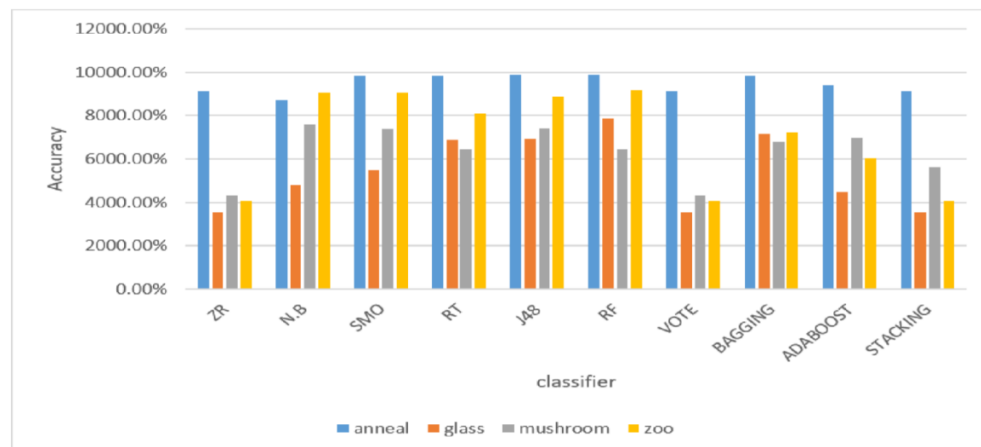


Fig. 1: Accuracy of both single and ensemble classifiers

6. Conclusion

Feature selection has been widely used as a pre-processing step that reduces dimensions of problem and improves classification accuracy. This paper analysed need for thinning of data by selecting significant features and briefly reviewed most feature selection methods and comparison of classifiers and also compared single and ensemble classifiers on the basis of accuracy and execution time. In this study, 5 single classifiers and 5 ensemble classifiers were compared over 4 datasets according to classification accuracy and execution time.

Finally, following conclusions were reached. Best classifier is ensemble classifier. Among that Random forest classifier when combined with correlation based feature selection found to be the best. Here accuracy of ensemble classifier is higher than single classifier among that Random forest classifier found to be the best when compared with others in terms of accuracy and bigger

data. It is clear that even new information technologies arise in the field of feature selection there will be datasets with millions of features. We are planning to extend our studies in Deep learning and to compare the results with ensemble methods along with mutual induction feature selections.

References

- [1] Donoho D.L., High-dimensional data analysis: the curses and blessings of dimensionality, AMS Math. Challenges Lect. (2000), 1–32.
- [2] Lavanya D., Rani D.K.U., Analysis of feature selection with classification: Breast cancer datasets, Indian Journal of Computer Science and Engineering (IJCSE) 2(5) (2011), 756-763.
- [3] Aboul Ella Hassaneian, Classification and feature selection of breast cancer data based on decision tree algorithm, Studies and Informatics Control 12(1) (2003).
- [4] Asha GowdaKaregowda, M.A.Jayaram, A.S. Manjunath, Feature Subset Selection Problem using Wrapper Approach in Supervised Learning, International Journal of Computer Applications 1(7) (2010), 13–17.
- [5] Asir Antony Gnana D., Jebamalar Leavline E., Literature Review on Feature Selection Methods for High Dimensional Data, International Journal of Computer Applications 136(1) (2016).
- [6] Bharesh Patankar, Dr.Vijaychavda, Improving classifier performance using Feature selection with Ensemble learning, International Journal of scientific Research in computer science, Engineering and Information Technology 1(1) (2016).
- [7] Kucheva L.I., Whitaker C.J., Measures of Diversity in Classifier Ensemble and their Relationship with the Ensemble Accuracy, Machine Learning 51(2) (2003).
- [8] Shweta Srivastava, Nikita Joshi, Madhvi Gaur, A Review Paper on Feature Selection Methodologies and Their Applications, International Journal of Engineering Research and Development, 7(6) (2013).
- [9] Ladha L., Deepa T., Feature Selection Methods and Algorithms, International Journal on Computer Science and Engineering (IJCSE) 3(5) (2011), 1787-1797.

