

Abstract

This thesis compares three feature selection methods: through Correlation Based Feature selection (CFS), Relief, and Wrapper methods. Three machine learning algorithms were used: J48 (a decision tree learner), naive Bayes (Bayesian Network), And Multilayer Perceptron (MLP) (Artificial Neural Networks). The purpose of comparison is to extract best set of features that leads enhance performance of classifiers. As the method is study_case_based SEER data is selected for this purpose.

The study showed that classification accuracy using the reduced feature set is equal and in some cases outperform the complete data set.

Moreover, as expected the performance of J48 decreases with the reduced data set.

CFS selected five features, WRAPPER returned eight features and RELIEF returned list of ranked features.

By comparing selected classifier methods Naïve Bayes is showed better results in this study. It produced a significant increase in accuracy with CFS, RELIEF, and WRAPPER methods.

المستخلص

يقدم هذا البحث دراسة مقارنة ثلاث خوارزميات Feature selection methods:

Correlation Based Feature selection (CFS), Relief and Wrapper methods.

تم استخدام ثلاث خوارزميات من Machine Learning:

J48 (a decision tree learner), naive Bayes (Bayesian Network), And Multilayer Perceptron (MLP) (Artificial Neural Networks) .

الغرض من هذه المقارنة هو تقليل حجم مجموعة البيانات التي تحتوي على كمية كبيرة من البيانات وذلك بإختيار أفضل مجموعة من العدد الكلي لل Features والتي تؤدي الى تحسين أداء ال classifiers. وقد تم إختيار بيانات SEER لهذا الغرض.

قامت خوارزمية CFS بتحديد خمسة Features، أما ال WRAPPER قامت بإختيار ثمانية Features، وال RELIEF قامت بإرجاع قائمة ال Features مرتبة بعد تقييمها حسب الأفضلية.

أيضاً أثبتت التجارب أن خوارزمية Naive Bayes حصلت على أفضل النتائج في هذه الدراسة وحقت زيادة واضحة في الدقة مع الثلاث خوارزميات CFS, Relief and Wrapper methods.

Acknowledgements

First and foremost I would like to acknowledge my supervisor, Dr. Mohamed Elhafiz Mustafa Musa. Also I thank my best friends Neama Hussein Eman Abraheem thanks for their technical assistance and helpful comments. Special thanks must also go to my family. They have provided unconditional support and encouragement of my time in graduate university.

Table of Contents:

Contents	Page No.
Abstract.....	I
Acknowledgements.....	III
List of Figures.....	IV
List of Tables.....	VI
Chapter one: Introduction	
1.1 Introduction.....	1
1.2 Motivation.....	1
1.3 Thesis statement.....	2
1.4 Research Objectives	2
1.5 Outline.....	2
Chapter Two: Data Mining and Feature Selection	
2.1 Introduction.....	3
2.2 Knowledge Data Discovery.....	3
2.3 Data Preprocessing.....	4
2.3.1 Feature Selection.....	4
2.4 Supervised Machine Learning	5
2.5 Classification.....	5
2.5.1 Classification Techniques.....	6
2.6 Artificial Neural Network.....	6
2.6.1 Introduction.....	6
2.6.2 Artificial Neural Network Definition.....	6
2.6.3 Training of Artificial Neural Networks.....	6
2.6.4 Network Architectures.....	6
2.7 Types of Neural Networks	8
2.7.1 Multilayer perceptron.....	8
2.8 The Backpropagation Algorithm.....	9
2.9 Decision Tree.....	10
2.9.1 J48 algorithm.....	11

2.10 Bayesian Networks.....	11
2.10.1 Naive Bayes.....	12
2.11 Data reduction	13
2.12 Dimensionality Reduction	14
2.12.1 Feature Selection	14
2.13 Feature Selection Heuristic Methods	15
2.13.1 Forward Selection.....	16
2.13.2 Backward Elimination.....	16
2.13.3 Combination of Forward and Backward.....	16
2.13.4 Decision Tree Induction.....	16
2.14 Characteristics of Feature Selection Algorithms.....	17
2.15 Feature Selection Methods	19
2.15.1 Correlation-based Feature Selection (CFS)	19
2.14.2 Wrapper Evaluator Method.....	20
2.14.3 RELIEF Evaluator Method.....	21

Chapter Three: Experiment and Results

3.1 Introduction.....	22
3.2 Dataset Description.....	22
3.3 Weka Tool.....	23
3.4 Experiments Methodology.....	24
3.5 Results	24
3.5.1 Feature Selection Results.....	24
3.5.2 Classification Results.....	29
3.5.3 J48 Results.....	29
3.5.3.1 CFS	29
3.5.3.2 RELIEF	29
3.5.3.3 Wrapper.....	30
3.5.4 Naïve Bayes Results.....	30
3.5.4.1 CFS.....	30
3.5.4.2 RELIEF.....	30

3.5.4.3 Wrapper.....	31
3.5.5 Multilayer Perceptron Results.....	31
3.5.5.1 CFS.....	31
3.5.5.2 RELIEF.....	31
3.5.5.3 Wrapper.....	32
3.6 Discussion.....	33
Chapter Four: Conclusion and Future Work	
5.1 Conclusion	34
5.2 Future Work.....	34

List of Figures

Figure No.	Page No.
2.1 knowledge discovery process.....	4
2.2 Neural Network in Human Body.....	6
2.3 Human Neurons.....	7
2.4 Artificial Neuron.....	7
2.5 Network Architecture.....	9
2.6 Multilayer Perceptron Network (MLP).....	11
2.7 Data Reduction techniques.....	14
2.8 Feature Selection Methods.....	17
2.9 Steps for feature selection.....	18
2.10 Correlation-based Feature Selection Method (CFS).....	20
2.11 wrapper feature selectors.....	21
3.1 this figure shows selected useful feature by J48 decision tree when using as classifier.....	27
3.2 This figure shows the result of J48 when using different number of feature ranked RELIEF.....	27
3.3 result of Naïve Bayes when using different number of feature ranked RELIEF.....	28
3.4 result of Multi Layer Perceptron when using different number of feature ranked RELIEF.....	28

List of Tables

Table No.	Page No.
3.1 Description of seer breast cancer dataset.....	23
3.2 selected features by CFS, Wrapper and Relief Methods.....	25
3.3 selected ranked features by RELIEF Method.....	26
3.4 Accuracy Results before Feature Selection.....	29
3.5 Accuracy with J48.....	30
3.6 Accuracy with Naïve Bayes.....	31
3.7 Accuracy with Multilayer Perceptron.....	32