Towards a Better Feature Subset Selection Approach

Omar A. A. Shiba

Department of Computer Science
Faculty of Science
University of Sebha
Sebha, LIBYA
Abumoad99@hotmail.com

ABSTRACT

The selection of the optimal features subset and the classification has become an important issue in the data mining field. We propose a feature selection scheme based on slicing technique which was originally proposed for programming languages. The proposed approach called Case Slicing Technique (CST). Slicing means that we are interested in automatically obtaining that portion 'features' of the case responsible for specific parts of the solution of the case at hand. We show that our goal should be to eliminate the number of features by removing irrelevant once. Choosing a subset of the features may increase accuracy and reduce complexity of the acquired knowledge. Our experimental results indicate that the performance of CST as a method of feature subset selection is better than the performance of the other approaches which are RELIEF with Base Learning Algorithm (C4.5), RELIEF with K-Nearest Neighbour (K-NN), RELIEF with Induction of Decision Tree *Algorithm (ID3) and RELIEF with Naïve Bayes (NB), which are mostly used in the feature selection task.*

Keywords
Feature selection, classification accuracy, slicing, irrelevant features.

1.0 INTRODUCTION

Selecting an optimal set of features for a given task is a problem which plays an important role in a wide variety of contexts including pattern recognition, adaptive control, machine learning, and classification algorithms.

The problem of Feature Selection can be defined as the task of selection of subset features that describe the hypothesis at least as well as the original set (John, H.George & Kohavi, Ron & Pfleger Karl, 1994) There have been several techniques that have been advocated for feature preprocessing such as Feature Extraction (Huan & Hiroshi, 1998), Feature Selection (Kohavi & John, 1997) , Feature Weighting (Aha, 1998), Feature Construction and Feature Transformation (Zupan, Blaz & Bohanec, Marko & Demsar, Janez & Bratko, Ivan, 1998). The performance of most practical classifiers improved when correlated or irrelevant features of case are removed (Ming & Ravi, 2003 ; Omar A. A. Shiba, Md. Nasir Sulaiman, Ali Mamat & Fatimah Ahmad, 2006). Based on this fact, and the previous classification accuracy results obtained by other researchers, which are not good enough, it is interesting to investigate the optimal way to improve the classification accuracy. The result of the investigation produces a new approach reduce the number of features that will improve the classification accuracy.

2.0 RELATED WORK

In this section, background knowledge and previous research work related to our problem area are briefly described. The most common sequential search algorithms for feature selection are variants of forward sequential selection (FSS) and backward sequential selection (BSS). FSS begins with zero attributes, evaluates all feature subsets with exactly one feature, and selects the one with the best performance. It then adds to this subset the feature that yields the best performance for subsets of the next larger size. This cycle repeats until no improvement is obtained by extending the current subset. BSS instead begins with all features and repeatedly removes the feature that, so removed, the maximal performances increase results. (Aha, 1994).

2.1 Feature Selection

The objective of feature selection is to reduce the number of features used to characterize a dataset so as to improve an algorithm's performance on a given task.

Definition 1: The process of selecting the best subset of features that describes the hypothesis (at least as well as the original set).

$$\mathbf{F`} \subset \mathbf{F} \qquad (1)$$

Where F in equation (Eq. 1) is the set of original 'n' features and F` is the output by a feature selector with 'm' features.

2.1.1 Irrelevant Attributes

An attribute is irrelevant if it contributes nothing to the target hypothesis, i.e. it makes no meaningful contribution towards the classification task. Nearest Neighbour algorithms are especially susceptible to the inclusion of irrelevant attributes in the dataset. The performance of most practical classifiers improved when correlated or irrelevant features of case are removed (Ming & Ravi, 2003; Takao & Hidehiko, 1994).

There are two approaches to feature selection namely,

- The Wrapper Approach
- The Filter Approach.

2.1.2 Wrapper Approach

The Wrapper approach to feature selection conducts a feature space search for evaluating features. The wrappers include the learning algorithm as a part of their evaluation function. Wrappers usually provide better accuracy but are computationally more expensive than the Filter schemes (Baranidharan & Thomas, 2002). Wrapper algorithms typically use forward selection, i.e. they start from an empty list of features and add relevant features as they are discovered (Ming & Ravi, 2003).

The following sequence of steps, adopted from (Kohavi & John, 1997), illustrates a typical wrapper approach to subset selection based on hill climbing;

(1) Let v ← empty set of features.
(2) Expand v. typically, this generates new states by adding or deleting a single feature from v.
For example, if n = 3 and v = (0 0 0), then expansion of v might lead to the following states: (1 0 0), (0 1 0), and (0 0 1).
(3) Use the classifier and an error estimation procedure (such as bootstrapping) to find the fitness of each subset that resulted from the expansion of v.
(4) Let v' be the subset with the highest fitness.
(5) If fitness of v' is greater than that of v, v ← v' and go to step (2). Else terminate with v as the final subset.

2.1.3 Filter Approaches

Filter approaches for feature subset selection attempts to assess the features and their merits using the data available.(Huan & Hiroshia,1998), FOCUS (Kohavi & John, 1997) , RELIEF (Aha, 1998) and its variants (Zupan et al.,1998) are some of the widely known Filter algorithms. The Filter algorithms consider the features independent of the classifiers that use them. Statistical and Information theoretic measures like Information gain, Cross-entropy, etc are used to weigh the feature (John et al., 1994). These measures capture the relationship of the feature with the target feature assuming conditional independence with all other features (Baranidharan & Thomas, 2002).

3.0 THE PROPOSED APPROACH AND RELATED TERMS

3.1 Definitions and Notation

A slice provides the answer to the question "What case features potentially affect the similarity computation;
Sim (New_case, Old_case) at case C?"
This section provides some basic definitions related to our slicing approach;

Definition 2: A case slicing: is a process for automatically obtaining subparts (features) of a case with a collective meaning.

Definition 3: A slicing criterion: denotes the conditions of the slice computation, with respect to which and for which case a slice is required.

Definition 4: Sliced case contains all features that could have direct relations with the features of interest at new case.

3.2 The Main Idea of The New Approach

Conceptually, our model is a variation of the nearest neighbour algorithms. The first step is assigning weights to new cases and also to the training cases in the data file. The second step is slicing the cases with respect to selected features. Slicing cases is removing such features that are irrelevant to the case at hand and also to cases in training set. Slicing cases means that we are interested in automatically obtaining that portion 'features' of the case responsible for specific parts of the solution of the case at hand. By slicing the case with respect to important features we can obtain new case with a small number of features or with only important features. The process of slicing approach is shown in (Fig. 1).

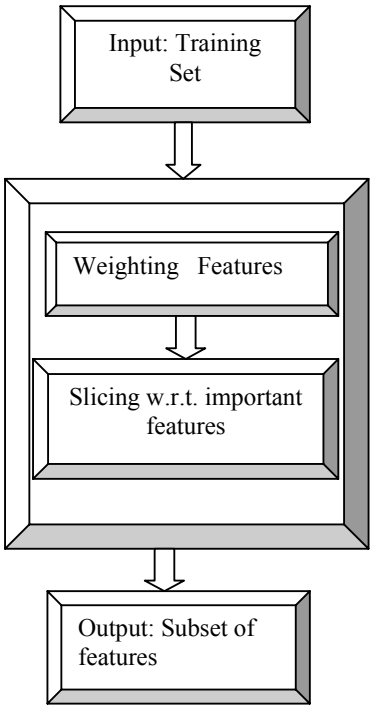

Figure .1: The proposed approach process

As shown in Fig. 1, the brief description of the process of the proposed approach for feature selection is as follows:

- *Firstly, the case is inserted into <weighting features stage> for assigning weights to each feature in the case using a statistical approach called "conditional probability" which will assign high correlation weights to the most relevant features among others.*
- *Secondly, after the weights have been assigned, the slicing technique will take place to slice the case according to its features weights by removing only the features that are irrelevant.*
- Lastly, the outcome of the second step is a subset of features, which are needed by the classification algorithm.

Our model consists of a database, one calculation module to identify the importance of each attribute and one slicing technique to select the important features or attributes for a decision outcome.

3. 2. 1 Database Representation

In a typical supervised machine learning task, data is represented as a table of examples or instances. Each instance is described by a fixed number of measurements, or features, along with a label that denotes its class. Features (attributes) are typically one of two types; nominal (values are members of an unordered set), or numeric (values are real numbers).

Our model requires a set of past cases as the input. This set of cases is represented as a relational data file. Each case is a record in this data file, and consists of two parts. The first part is used as the predictors for the value of the second part which is the goal variable. The structure of the data file is shown in (Table. 1).

Table 1: Structure of the data file

Features	**Feat. 1**	**Feat. 2**	**„**	**Feat. N**	**Classes**
Case 1	„ .. .	..	..	…	… …
[← ……..……→]					[↔]
Predictors					Goal

4.0 EXPERIMENTAL RESULTS

In this section, we compare the proposed CST as a feature selection approach against four selected feature selection approaches which are RELIEF with Base Learning Algorithm (C4.5), RELIEF with K-Nearest Neighbour (K-NN), RELIEF with Induction of *Decision Tree* Al gorithm (ID3) and RELIEF with Naïve Bayes (NB), which are mostly used in the feature selection task on four datasets. Here in Table 2. On completion of all experiments on the four selected datasets from different domains, we can see that all the techniques give good classification accuracy. From this result we can see that our approach is better than other feature selection approaches, because feature selection in our approach is based on features weights and class label, where the feature selection in the other approaches is based on rules extraction, which sometimes produce number of features which are irrelevant to the case and sometimes becomes very weak and not supported by any case. Fig. 2 shows the difference in classification accuracy of our approach against selected approaches with feature selection using RELIEF.

Table 2: Classification accuracy of our approach against selected approaches with feature selection using RELIEF

Methods / Datasets	C4.5+ RELIEF	k-NN+ RELIEF	ID3+ RELIEF	NB+ RELIEF	New approach
BCO	74.375	72.375	71.75	73.12	99.30
GERM	86.42	79.57	86.143	79.42	98.00
DNA	82.5	71.5	75.75	72.56	96.10
VOTING	95.125	93.5	94.625	90.93	97.30

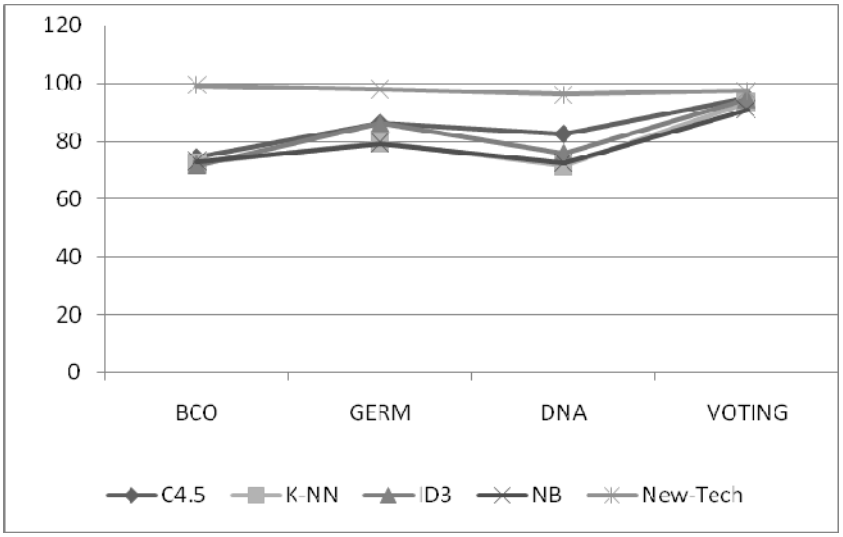

Figure 2. Difference in classification accuracy of the selected algorithms

5.0 CONCLUSION

This paper has briefly described some related work to the purpose of features subset selection in solving the classification problems. The paper has also presented and discussed a new technique for the same purpose. The new technique was supported with experiments on four datasets. The experiments show that the new technique improves the classification accuracy and it gave very high percentage of classification accuracy, because it based on slicing the features with respect to relevant features only. However the technique is not viewed as a replacement technology, but rather as a complementary technology to the approaches for the purpose of feature subset selection.

REFERENCES

Aha, D. W. (1998). Feature Weighting for Lazy Learning Algorithms. In Feature Extraction, Construction, and Selection, A Data Mining Perspective, Huan, Liu and Hiroshi, Motoda (eds.) *Kluwer Academic Publisher*, 13-32.

Aha, D W. (1994). *Feature Selection for Case-Based Classification*. AAAI Technical Report WS-94-01

Raman, B., & Thomas R.Ioerger, (2002). Instance Based Filter for Feature Selection. *Journal of Machine Learning Research, 1*, 1-23.

Liu, H., & Motoda, H. (1998). Feature Extraction, Construction, and Selection, A Data Mining Perspective, *Kluwer Academic Publisher.*

John, H.George, Kohavi, R., & Karl, P. (1994). Irrelevant Features and the Subset Selection Problem. *In International Conference on Machine Learning,* pp. 121-129.

Kohavi, Ron & John, George.H. (1997). Wrappers for Features Subset Selection. *In Artificial Intelligence, 1*(2), 273-324.

Ming Dong, & Ravi Kothari. (2003). *Feature Subset Selection Using a New Definition of Classifiability Computer Science Department*, Wayne State University.

Omar A. A. Shiba, Md. Nasir Sulaiman, Ali Mamat & Fatimah Ahmad, (2006). An Efficient and Effective Case Classification Method Based On Slicing. *International Journal of The Computer, the Internet and Management, 14*(2), 15-23.

Mohri, T., & Tanaka, H. (1994). An Optimal Weighting Criterion of Case Indexing for Both Numeric and Symbolic Attributes. *Workshop, Technical Report ws-94-01. Menlo Park, CA. AIII press,* 123-127.

Zupan, Blaz and Bohanec, Marko and Demsar, Janez and Bratko, Ivan, (1998). Feature Transformation by Function Decomposition, *In IEEE Intelligent Systems, 13*(2), 38-43.