Accepted for Computers & Mathematics with Applications.

A Multilevel Tabu Search Algorithm for the Feature Selection Problem in Biomedical Data

I. O. Oduntan[†‡], M. Toulouse[†], R. Baumgartner[‡], C. Bowman[‡]
R. Somorjai[‡] and T.G. Crainic[§]

†. Dept. of Computer Science, University of Manitoba
{oduntan,toulouse}@cs.umanitoba.ca

‡. Institute for Biodiagnostics, Winnipeg, Manitoba
{Richard.Baumgartner,Christopher.Bowman,Ray.Somorjai}@nrc-cnrc.gc.ca

§. School of Business Administration and CIRRELT UQAM, Montréal, Canada
theo@crt.umontreal.ca

Abstract

The automated analysis of patients' biomedical data can be used to derive diagnostic and prognostic inferences about the observed patients. Many noninvasive techniques for acquiring biomedical samples generate data that are characterized by a large number of distinct attributes (i.e., features) and a small number of observed patients (i.e., samples). Using these biomedical data to derive reliable inferences, such as classifying a given patient as either cancerous or non-cancerous, requires that the ratio r of the number of samples to the number of features be within the range $5 < r < 10$. To satisfy this requirement, the original set of features in the biomedical data has to be reduced to an '*optimal*' subset of features that most enhances the classification of the observed patients.

In this paper, we propose a new feature selection technique (*multilevel feature selection*) that seeks the '*optimal*' feature subset in biomedical data using a multilevel search algorithm. This algorithm combines a hierarchical search framework with a tabu search method. The framework consists of increasingly coarse forms (i.e., search subspaces) of the original feature space that are strategically and progressively explored by the tabu search method. The result of the search at any given coarse subspace is used to initialize the search at the previous less coarse subspace.

We evaluate the performance of the proposed technique in terms of the solution quality, using experiments that compare the classification inferences derived from the solution found, with those derived from other feature selection techniques such as the sequential forward selection, random feature selection and tabu search feature selection. An equivalent amount of computational resource is allocated to the evaluated techniques to provide a common basis for comparison. The empirical results show that the multilevel feature selection technique finds '*optimal*' subsets that enable more accurate and stable classification than those obtained using the other feature selection techniques.

Keywords: Multilevel search algorithms, feature selection problem, tabu search, biomedical data.

1 Introduction

The advent of noninvasive techniques such as magnetic resonance spectroscopy (**MRS**) [10] and gene microarrays [3] for acquiring biomedical data enables the creation of automated means to identify the presence and monitor the progression of diseases in patients. The biomedical sample that represents each observed patient usually consists of a combination of distinctive characteristics having quantitative measures. Each distinctive characteristic is a feature, and a collection of L ordered features is a feature vector. For instance, the normalized expression level of each gene in gene microarrays corresponds to a feature, and the expression levels of the genes for each observed specimen under a given condition corresponds to a feature vector.

Useful and reliable information and inferences can be derived from biomedical data by applying appropriate pattern recognition techniques (classifiers) [11]. Unfortunately, classifier performance tends to degrade when the sample-to-feature ratio r decreases beyond a certain range. Typical values of r that are necessary for good classifier performance range between 5 and 10; but for biomedical data, r typically ranges between 1/500 and 1/20 [23]. The ratio r can be increased to the required range either by increasing the number of observed patients (i.e., the number of samples) or by reducing the number of features. The former option is usually not practical because of the lack of adequate biomedical samples. The more practical option of reducing the original set to an '*optimal*' subset of features that most enhances classification performance is known as the feature selection problem [14].

Feature selection can be formulated as a combinatorial optimization problem. In application to biomedical data, feature selection appears in two different combinatorial formulations. In one, it consists of finding a subset of features of fixed cardinality m (in the range $5 < r < 10$) that yields the lowest misclassification error rate for a given classifier. In the other formulation, feature selection seeks a subset of features with the smallest cardinality such that the misclassification error rate is below a given threshold. In this paper, we present a configuration of the multilevel feature selection technique for the first formulation. This can be stated as follows:

Given an ordered set of L features $F = (f_1, f_2, \dots, f_L)$, find a subset S of F such that $|S| = m$, and the error rate $c(S)$ of a given classifier is minimized when presented with the feature subset S. That is:

$$\begin{gathered} \min c(S) \\ \text{such that} \quad S \subset F,\ |S| = m, \quad m < L \end{gathered} \tag{1}$$

The problem formulation in (1) can be solved exactly, using an exhaustive enumeration of the $\binom{L}{m}$ different subsets having cardinality m, but this approach is impractical except for very small values of L. To solve this feature selection problem in a practical way, many techniques have been developed using search algorithms that enable the selection of near-optimal feature subsets within practicable computational time. Techniques based on simple heuristics such as greedy sequential search algorithms [1], on evolutionary methods such as the genetic algorithm (**GA**) [20] and on meta-heuristics such as tabu search [26] have been proposed. These techniques have been adapted or possibly enhanced appropriately to suit feature selection problems in particular types of biomedical data. For instance, Nikulin *et al.* [18] proposed a GA-based feature selection technique that is primarily aimed at biomedical spectra, wherein there is evident correlation amongst adjacent features; however, this technique is inappropriate for other types of biomedical data, such as microarrays, where such correlation may not exist [23]. Also, feature selection techniques such as in [2, 6] that focus primarily on microarrays data are usually not flexible enough to exploit the evident correlation that exist amongst features in spectral data. There is a need for a feature selection technique having an underlying search strategy that is flexible enough to adapt effectively to the different types of biomedical data and enhance classification performance.

In this paper, we propose a new technique, that we identify as *multilevel feature selection*, to solve the feature selection problem in biomedical data. The proposed technique is based on the multilevel search algorithm. This algorithm performs a strategic and flexible exploration of the feature space, using the hierarchical approach that is inherent in multilevel algorithms. Starting from the original feature set, the technique creates a hierarchy of feature subspaces with increasing coarseness. The flexibility in the hierarchically coarsening feature subspaces eases the adaptation of the technique to different forms of biomedical data. Starting from the coarsest feature subspace to the least coarse, i.e., the original feature subspace in the hierarchy, a tabu search [8] is used to find a sub-optimal subset of the feature subspace at each level in a progressive manner. The resulting sub-optimal subset at a level is used to initialize the search at the next less coarse feature subspace. The result of the tabu search at the least coarse feature subspace is regarded as the *optimal* feature subset found [4].

The remainder of this paper is organized as follows: Section 2 reviews the existing techniques that address the feature selection problem described in Section 1; Sections 3 and 4 describe respectively the multilevel search paradigm and its application to the feature selection problem; Section 5 provides the results and inferences derived from our experimentations, and the conclusion follows in Section 6.

2 Feature Selection Techniques

Feature selection techniques typically consist of an underlying search or ranking algorithm that explores the feature space and a cost function (e.g., a measure of the classification error rate) that guides the underlying algorithm. Considering the approach for evaluating the cost function of the feature selection techniques, Kohavi and John [15] identify two approaches for designing feature selection techniques: filter and wrapper approaches. The filter-based approach determines the fitness of an examined feature subset without any reference to or feedback from the target classifier. The cost function evaluation is independent of the target classifier that uses the selected subset of features in the subsequent classification of independent datasets. Rather, a generic error estimation function can be used to compute the cost function value that guides the ranking of the individual features or the search for an '*optimal*' subset in the feature search space. On the other hand, the wrapper-based approach determines the fitness of an examined feature subset by referring the subset to the target classifier to get a feedback in the form of an estimation of the classification error rate that will result when the examined subset underlies the design of the target classifier. The wrapper approach usually enables the selection of feature subsets that leads to better classification accuracy than the filter approach. However, the evaluation of each examined subset by the target classifier to determine the fitness is more computationally intensive than for the simple error estimation of filter-based methods. Consequently, since many subsets must be evaluated by the wrapper-based methods, there is an additional computational overhead that results from the evaluations that are based on the target classifier.

Considering the underlying search or ranking algorithm, Guyon and Elisseeff [9] also group feature selection techniques into two broad categories: feature ranking techniques and feature subset selection techniques. Feature ranking techniques order the features according to a relevance criterion such as covariance, and select a subset from the ordered features. Kira and Rendell [13] describe a simple feature ranking technique. The technique scores each feature in the original feature set using a ranking criterion and selects the first m features having the highest scores as the '*optimal*' feature subset, where m is the cardinality of the desired '*optimal*' subset. A primary drawback of feature ranking techniques with respect to classifier design is: a combination of the m highest ranked features is not necessarily the optimal or near-optimal subset of m features for 'best' classifier performance [4].

Given sufficient computation time, feature subset selection techniques such as in [17, 21, 22, 24] implicitly examine all the feature subsets and select the subset having the 'best' cost function evaluation as '*optimal*'. These techniques guarantee finding the optimal feature subset with respect to the estimation of the target classifier performance. However, the computational requirements (time and resources) of the techniques are very intensive and may be impractical when applied to large-scale feature selection problems. Furthermore, the techniques are usually based on assumptions that are not always true in practice. For instance, the branch-and-bound-based feature selection techniques [17, 21, 22] require that the cost function be monotonic on the subset of features; i.e., adding a new feature from the original set to a current subset of features must result in a better value of the cost function.

To solve the feature selection problem in a practical way, some other techniques find an approximate solution that is 'good', hopefully as close as possible to the optimal subset. These techniques intelligently examine some of the possible subsets of features and select as the '*optimal*' subset the 'best' cost function evaluation amongst all the examined subsets. Some of these techniques are discussed in the following. Sequential forward selection (**SFS**) and sequential backward selection (**SBS**) [1] are based on simple greedy deterministic heuristics. **SFS** starts by selecting the empty subset as the current subset and sequentially adds a new feature (from the original set) to the current subset. In each sequence, the added feature satisfies the condition of combining with the current subset to give the 'best' evaluation of the cost function. The selection process stops when a termination criterion is satisfied (e.g., when the addition of a new feature no longer improves the cost function or when the cardinality of the subset equals a set threshold) and the '*optimal*' subset of the selection is the current subset prior to termination. **SBS** is similar to **SFS**, but the selection process is reversed. **SBS** begins with the entire original set as the current subset and sequentially removes a feature from the current subset until a termination criterion is satisfied. **SFS** adds a single feature (and **SBS** removes a single feature) at each search sequence; hence, the discriminatory dependencies that exist amongst some combinations of features are ignored during the search. Stearns [24] proposes the plus-l-take-away-r method to address the shortcoming of possible exclusion. At each sequence, the method adds l features to the current selection using **SFS** and removes r features using **SBS**. The challenging task of this method is: there are presently no theoretical means of choosing a predefined value for l and r that enables finding the '*optimal*' subset. Generalized sequential forward selection (**GSFS**), a generic form of **SFS**, provides a flexible means of finding the '*optimal*' subset by permitting the addition of k features to the current selection at every search sequence. Similarly, there are generic forms for **SBS** and plus-l-take-away-r: generalized sequential backward selection (**GSBS**) and generalized plus-l-take-away-r, respectively.

The aforementioned sequential search techniques do not permit backtracking; that is, a search step cannot be reversed even when subsequent steps reveal the step as impairing to finding the '*optimal*' subset. To resolve the backtracking drawback, Pudil et al. [19] propose the sequential floating forward selection (**SFFS**) and the sequential floating backward selection (**SFBS**). At each search sequence, the **SFFS** method adds to the current selection, using **SFS**, and performs some **SBS** steps as long as the cost function evaluates to a better value. **SFBS** is similar to **SFFS**, but the progressive search sequence is based on **SBS**. Generally, other than the **SFS** and **SBS**, the sequential search techniques are computationally expensive for large-scale feature selection problems.

Siedlecki and Sklansky [20] propose a genetic algorithm (**GA**) approach for feature subset selection. Nikulin *et al.* [18] develop a GA-based technique for selecting the '*optimal*' subset of block of features (regions). The technique does not generate a stable subset of features, however, and may be inadequate for biomedical data not having correlation amongst consecutive features [23].

Zhang and Sun [26] develop a tabu search method for feature subset selection. Tabu search is an iterative search method where, at each iteration, a successor solution s' to a current solution s is selected from a set $\mathcal{N}(s)$ of solutions called the *neighborhood* of s. The successor s' is often obtained by applying a simple rule such the following:

$$s' = \min\{c(s') \mid s' \in \mathcal{D}(s)\}. \tag{2}$$

where $\mathcal{D}(s) \subseteq \mathcal{N}(s)$. To avoid cycling (which happens when $c(s') \geq c(s)\ \forall s' \in \mathcal{N}(s)$), tabu search uses an adaptive memory, known as the *tabu list*, to keep track of solutions (or solution attributes) that have been visited and should be avoided for a number of iterations; the *tabu tenure* determines how long a solution remains in the tabu list. In Zhang and Sun [26], a comparative analysis of the tabu-search-based technique and other feature selection techniques (**SFS**, **GSFS**, **SBS**, **GSBS**, plus-l-take-away-r, **SFFS**, **SFBS** and **GA**) is performed using a synthetic dataset. Although the result of the performance analysis shows tabu search as a promising search heuristic for feature selection problem, the analysis is done using a synthetic dataset and the claims should be verified using real-life datasets.

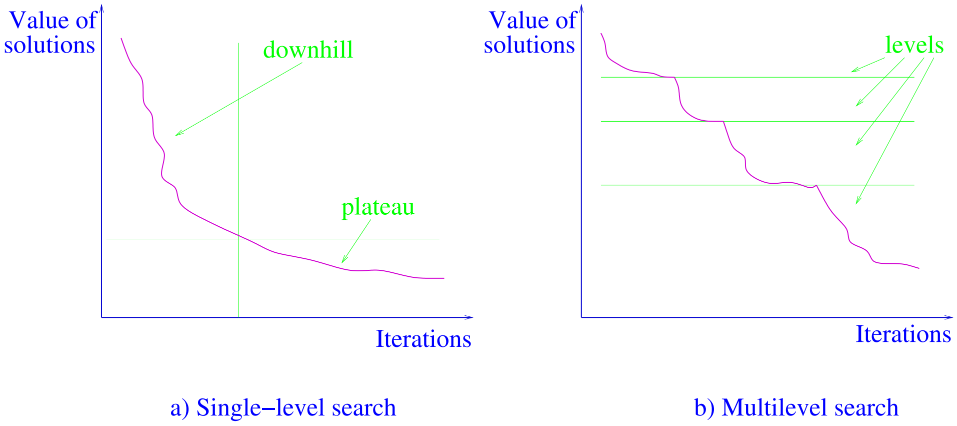
We examine the strength of the basic tabu search feature selection technique using biomedical data and propose a new technique that integrates tabu search and can be adapted to solve the feature selection problem in most forms of biomedical data (e.g., **MR** spectra, microarrays, mass spectra). The proposed technique is based on the multilevel search paradigm [25].

3 The multilevel search paradigm

The basic framework of *multilevel search* algorithms can be described as follows: Starting from an original discrete optimization problem instance P_0 with solution space $\mathcal{S}_0$, a *coarsening phase* projects P_0 into smaller problem instances $P_1, P_2, \ldots, P_l$ by recursively reducing the number of decision variables with respect to P_0. Usually, the projection is such that the search spaces $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_l$ induced respectively from problem instances $P_1, P_2, \ldots, P_l$ satisfy the relation

$$\mathcal{S}_l \subset \mathcal{S}_{l-1} \subset \cdots \subset \mathcal{S}_1 \subset \mathcal{S}_0. \tag{3}$$

During the *initial search phase*, an approximation e_l of the optimal solution for P_l is computed, using some search algorithm. During the *refinement phase*, a solution $s_i \in \mathcal{S}_i$ is derived, by *interpolating* the values of the decision variables in



a) Single−level search

b) Multilevel search

Figure 1: Trends in single-level & multilevel search methods

P_i from the solution e_{i+1}, that is an approximate optimum of the problem instance P_{i+1}. The interpolated solution s_i serves as an initial solution to a search heuristic that optimizes over P_i. The refinement phase interpolates and refines feasible solutions until the values of the decision variables of the original problem P_0 can be interpolated from the best approximation e_1 of coarsened problem instance P_1. This last interpolation is the initial solution for the search heuristic optimizing P_0.

The multilevel exploration of the solution space helps search heuristics like tabu search to cope with challenges arisen from the optimization of large problem instances. One of these challenges is the confinement of the exploration to a particular region of the solution space. As pictured in Figure 1a, the function plotting the value of each current solution of a search heuristic such as tabu search against its respective iteration number is often an exponential. This exponential can be divided into two segments: a downhill (or uphill) segment and a plateau segment. Each segment represents a distinctive phase of the exploration of the solution space [7, 16]. The downhill (or uphill) segment corresponds to a phase where rapid improvements in the value of the solutions attract the exploration in a confined region of the solution space; the plateau segment corresponds to 'sidewalks' of the search in this confined region of the solution space. Note that the restrictions imposed by the tabu tenure in tabu search often resulted in sequences of iterations where the cost of each successor solution s' is greater of equal to the cost of the current solution. Such sequence, where $c(s') \geq c(s)$ for each iteration, allows the exploration to overcome some uphill barriers in the landscape of the cost function. Nonetheless, in the broader context of the exploration, search techniques rarely can overcome large uphill barriers such as those corresponding to downhill segments. Consequently, in a large solution space, search is often confined to a specific region of solution space predetermined by the initial solution of the search.

Several strategies have been proposed to attenuate the impact of confinement, multilevel search can be seen as one of them. Multilevel search re-initializes the search through a sequence of plateau segments, each having the capacity to steer away a search heuristic from the confined region of the solution space associated to the very first initial solution. Figure 1b represents the typical trend in the value of the current solutions against their respective iteration number in a multilevel search. This trend is a sequence of exponential functions, one for each level. The downhill segment of each exponential at a given level i corresponds to rapid improvements made on the value of initial solution s_i, which has been interpolated from e_{i+1}, the best solution of previous level $i + 1$. These rapid improvements stem from assigning values to decision variables in P_i that do not appear in the definition of the problem instance P_{i+1}.

Each downhill segment ends with a plateau. The cost function of coarsened problem instances, particularly at the highest levels, is smoother in comparison to the original problem instance P_0 [25]. The confinement of tabu search to specific regions of a search space is not as definitive because uphills in the landscape of coarsened search spaces can be overcome. As a result, sidewalk explorations cover more broadly the search space of coarsened problem instances, enabling search at level $i - 1$ to 'break away' from the search at level $i + 1$. (Search at level $i - 1$ breaks away from the search at level $i + 1$ if the approximate solutions e_{i+1} and e_i cannot be found in the downhill or plateau segments of a single-level search of $\mathcal{S}_0$.) Through interpolations and sidewalk explorations, multilevel not only re-invents the initial solution of the search at each level, but it does so by taking hints from the optimization process performed at the (immediate) previous level.

Dimensionality reduction of problem instances during the coarsening phase also provides computational advantages to a search heuristic such as tabu search. Solutions in smaller search spaces have fewer neighbors. Consequently, the cost for applying the pivoting rule in each iteration is reduced significantly in coarsened search spaces, where the neighborhoods are substantially smaller. Under certain conditions, it is expected that tabu search, while optimizing the smallest search spaces of the multilevel search structure, can find approximations that are very close to the cost of the optimal solutions of each corresponding search space.

The multilevel approach is naturally suitable to subset problems such as the feature selection problem. It is even more so for feature selection applied to biomedical data, where the cardinality of the desired subset is usually very small in comparison with the total number of features in the problem instances, i.e., $|S| << |F|$. It is likely that the reduction of the number of decision variables during the coarsening phase will not destroy too much relevant information with respect to near optimal subsets in the original feature space. Finally, we suspect that coarsened search spaces will be resistant to overfitting, which will help to improve the generalization capabilities of classifiers.

4 The multilevel feature selection algorithm

As input, we are given a biomedical dataset **V** $= \{V_1, V_2, \ldots, V_k\}$ consisting of k feature vectors (where k is the number of observed samples). Each feature vector V_i is a set $(V_{i_1}, V_{i_2}, \ldots, V_{i_L})$ of real numbers where V_{i_j} is the jth measured feature value for the corresponding sample V_i. Coordinates of the L-dimensional feature space are represented by the feature vector $F = (f_1, f_2, \ldots, f_L)$ of size L.

Our algorithm seeks a feature subset S of F of pre-defined cardinality m, consisting of the features that best enable the classification of the given biomedical dataset. The decision to select a feature in F to be a member of a subset S is expressed using an array x of Boolean variables. Feature $f_j \in F$ is mapped to a *decision variable* $x[j]$. When an entry $x[j]$ is set to '1' during a selection process, the corresponding feature $f_j \in F$ is included in the desired subset; otherwise when $x[j]$ is set to '0', the corresponding feature f_j is excluded. The values of the decision variables are set by a decision process which, in the multilevel feature selection technique, is an underlying search heuristic.

4.1 Coarsening phase

For the feature selection problem, the coarsening phase recursively generates a hierarchy of feature subspaces. Across the hierarchy, a coarse feature subspace consists of features generated from the immediate, less coarse subspace and the dimensionality of the subspaces reduces with increasing coarseness. Given an original feature set $F_0 = F$, the coarsening phase combines a coarsening strategy with parameters such as the *reduction factor* rf and the number of levels l to generate feature sets $F_1, F_2, \ldots, F_l$ such that $F_l \subset F_{l-1} \subset \cdots \subset F_0$, where l is an implicitly or explicitly defined parameter that determines the number of levels in the hierarchy, and rf, the reduction factor, is the ratio of the dimensionality of the given feature subspace to the dimensionality of the immediate coarser feature subspace, i.e., $rf = \frac{|F_i|}{|F_{i+1}|}$. In the present implementation of multilevel search to feature selection, we only consider reduction factors that are constant at each level.

4.1.1 Coarsening strategies

Research on multilevel algorithms suggests two general strategies that can be applied during the coarsening phase to reduce the dimensionality of the feature subspaces: *clustering* of decision variables [12] and fixing the state of some decision variables, i.e., *decision variables pre-setting* [5].

The first approach involves merging a collection of features and then representing the merged features by a single decision variable. This clustering approach can be used to combine features and to generate feature subspaces as follows: for a given level i, the feature subspace F_i is coarsened by aggregating groups of features in F_i such that an approximated form of each group represents a new feature that is an element of a new feature space at the immediate next coarse level F_{i+1} in the multilevel hierarchy. For a given subspace, the groups of features that are approximated to constitute the next coarser subspace can be created using clustering algorithms that identify the possible correlations that may exist amongst the features in the given subspace. For instance, for biomedical datasets wherein correlations exist amongst adjacent features, the groups can be created by selecting consecutive features within predefined window(s) and a statistics (e.g., median, average) of the features within the window can represent an approximation for each group.

Typically, the coarsened feature subspaces generated using the clustering approach are synthetic, i.e., they consist of features that literally may not exist in the original feature space. Furthermore, the characteristics of the features can vary for each subspace in the multilevel hierarchy. Therefore, the task of relating the solutions generated from the synthetic feature subspaces to the desired solution in the original problem instance and to the subsequent interpretation of the desired solution can be quite challenging. This challenge may not be prominent for biomedical datasets such as MRS data, wherein evident correlation typically exists amongst adjacent features, since the resulting clusters in the synthetic subspaces do not necessarily compromise the interpretation of the original features that constitute the clusters. However, for datasets, such as microarray data, wherein such correlation may not exist, the clusters are usually not approximate representations of the constituent original features with respect to interpretation; therefore, the final near-optimal solution selected for such dataset can impair diagnosis and prognosis. A means of addressing the challenge of retaining the originality of features in the feature clustering coarsening approach requires tracking in a subspace the features that are combined to form new features in the coarser subspaces.

The second coarsening approach, decision variables pre-setting, generates the coarse feature subspace F_{i+1} at level $i+1$ from the immediate less coarse subspace F_i by excluding some features from the feature subspace at level i. A feature $f_j \in F_i$ is excluded from level $i + 1$ by fixing the state of the corresponding Boolean variable $x[j]$ to '0' at level $i + 1$. Once a feature is excluded at a given level, it cannot be included in the solutions of the solution space at coarser levels. Fixing the decision variables in this way recursively reduces the dimensionality of the feature subspaces in the multilevel hierarchy and therefore reduces the size of the solution space of the original optimization problem progressively.

In order to create the coarse feature subspaces using the pre-setting approach, there is need for a means of determining which decision variables have their state fixed to '0' at each level. We identify and investigate two strategies for this purpose: *biased selection* and *random selection*. For a given feature subspace, the biased selection strategy determines the feature that belongs to the immediate coarser subspace by examining the discriminatory capability of the features in the given subspace. The discriminatory capability of the features can be determined by applying a feature selection technique to explore the given feature subspace. Any appropriate technique can be used for this purpose; a simple feature ranking technique is used in the present implementation. For a given feature subspace, the ranking technique sorts the features in descending order of discriminatory capability and the first p features are selected, where p is the cardinality of the next coarser subspace. In the random selection strategy, the values of the decision variables that correspond to the features in a given feature subspace are set randomly and recursively. A simple Gaussian random number generator is used to guide the selection.

The above three coarsening strategies differ significantly either in the definition of what is a feature at each level (synthetic versus real features) or in the way features are selected to constitute each level (random versus ranking). We performed experiments to investigate whether those strategies have an impact on the performance of the feature selection technique. In the experiment, we use three instances of the multilevel algorithm that have the same configuration except for the coarsening strategy, to find the near-optimal subset for 10 training dataset instances. The coarsening phase of the multilevel algorithm instances differs in the

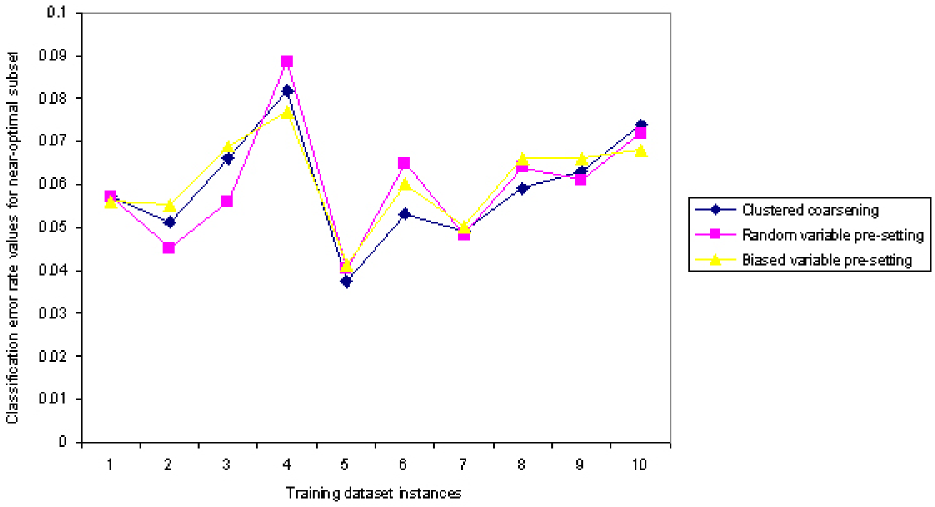


Figure 2: The effect of the coarsening strategies on the multilevel feature selection technique

coarsening strategy that underlies the algorithm; the first instance is based on the clustering strategy, the second and third instances are based on the two versions of the pre-setting strategy (i.e., random and biased selection pre-setting, respectively). For the three multilevel algorithm instances, we compare the estimated values of the classification error rate - CER, derived using leave-one-out cross validation. Figure 2 show the results of the experiment.

As shown in Figure 2, varying the coarsening strategies using the aforementioned strategies does not have obvious influences on the overall performance of the multilevel feature selection algorithm. The clustering and biased pre-setting strategies require more computational resource than the random pre-setting strategy. The clustering strategy requires additional computing resource to track the features that are combined to constitute a coarser subspace at each level. The biased pre-setting strategy requires additional computing resource to determine the features that are selected from a coarse subspace to a coarser subspace. Therefore, the computational cost of the multilevel feature selection technique is higher for both the clustering and the biased pre-setting approach compared to the random pre-setting coarsening approach. We use the random pre-setting coarsening strategy in the present implementations of the multilevel feature selection technique, since this strategy requires the least computation cost and its influence on the multilevel feature selection technique is comparable to the other strategies.

4.1.2 Reduction factor

The reduction factor $rf = \frac{|F_i|}{|F_{i+1}|}$ is a second significant parameter in the coarsening phase. Let E_i denote the best subset of m features discovered by the search procedure optimizing the feature subspace at level i, and let S_i denote the initial solution at level i that has been interpolated from E_{i+1}. Assume the coarsening strategy involves fixing at '0' the state of some decision variables of x_i, the Boolean vector representing the decision variables of problem instance P_i. Finally, we define $x_i[j]$ as a ***free*** decision variable if $x_i[j]$ is not fixed at '0' at level i and the corresponding feature f_j is not in the feature subspace of level $i + 1$.

Coarsening strategies based on small reduction factors produce similar adjacent feature subspaces. The downhill segment of tabu search at level i can be short, since S_i, the initial solution at level i interpolated from E_{i+1}, cannot be substantially improved by assigning states to the free decision variables at level i. Search is likely to yield a prolonged plateau segment at each level. Overall, given that adjacent feature subspaces are similar, breakaways in the search between adjacent levels are less likely to occur. This increases the dependency of the multilevel search on the initial conditions as defined by the search at level l.

On the other hand, coarsening strategies based on large reduction factors produce adjacent feature subspaces which differ substantially. Tabu search at each level i is likely to yield a prolonged downhill segment, refining and improving the value of the initial solution S_i by assigning states to a large set of free decision variables at level i. However, the refinement of S_i by assigning states to free decision variables may create a situation where the best solution E_i at level i is a local optimum of the downhill segment. In other words, the search in the plateau segment is unable to find a solution better than the best solution in the downhill segment. Again, adjacent levels fail to produce breakaways because the search at a given level uniquely refines the best solution of the downhill segment from the previous level.

We perform experiments to determine an appropriate value for the reduction factor for a given problem. In the experiment, we use similar configurations (except for the value of the reduction factor that varies with each instance) of the multilevel feature selection algorithm to find the near-optimal subset for the same biomedical datasets. The dimensionality of the original feature space is 1500 and the number of levels in the multilevel algorithm instances is set to 3. We compare

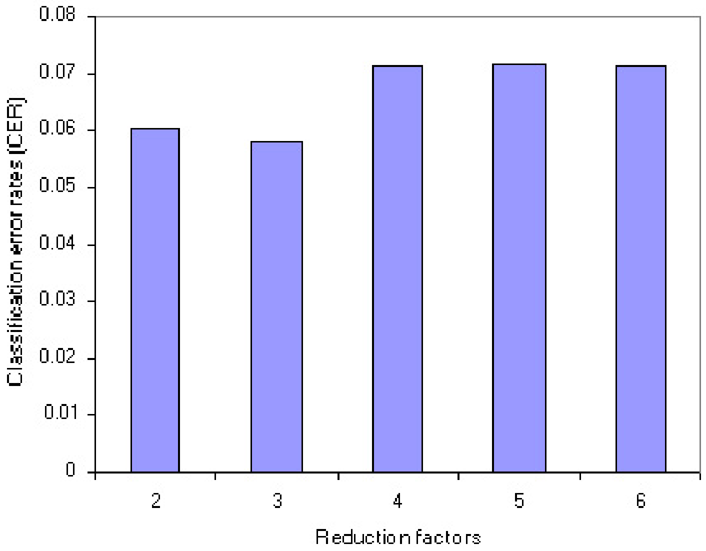


Figure 3: The effect of the reduction factor on the multilevel feature selection technique

the estimated values of the classification error rate CER using leave-one-out cross validation for the multilevel algorithm instances with the reduction factor value set to 2, 3, 4, 5, and 6. Figure 3 show the results of the experiment.

As shown in Figure 3, with reduction factor values of 2 and 3, the multilevel feature selection algorithm consistently finds near-optimal subsets having lower average error rate than for the other reduction factor values (i.e., 4, 5 and 6). Using the empirical results shown above, a reduction factor that coarsen a subspace by 30% to 50% can be recommended for similar problem domain instances. This recommendation agrees with the reduction factor of 2 that is used in most configurations in the literature on multilevel search algorithms. A reduction factor of 3 is used in the present implementations of the multilevel feature selection algorithm.

4.1.3 Number of levels

The *number of levels* parameter of multilevel search algorithms is a function of the size of the original problem instance P_0, the targeted size of the feature subspace at level l, and the reduction factor rf. As with rf, the appropriate value for the number of levels cannot be determined theoretically. Rather, this parameter can

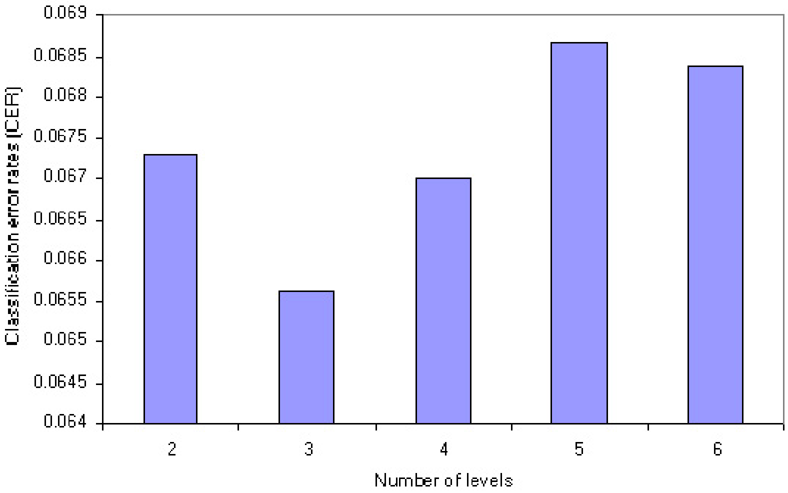


Figure 4: The effect of the number of levels on the multilevel feature selection technique

be explicitly predefined, based on empirical inferences, or implicitly defined as a function of other parameters such as the reduction factor and the dimensionality of the original problem instance. For example, if an exact search algorithm is used to optimize the feature subspace at level l, then the number of features at level l should be small enough to allow a rapid execution of the exact search. In this context, assuming the reduction factor is predefined, the number of levels follows directly from the size of the problem instances P_0 and P_l. There is more flexibility when heuristic search procedures are used at all levels.

We investigate the effect of the number of hierarchical levels on the performance of the multilevel feature selection technique using experiments that compare different instances of the technique having varying number of levels while the other parameters are constant. Similar calibration experiments are performed, using different instances of the multilevel feature selection algorithm, with the number of levels set to 2, 3, 4, 5, and 6, while the other configurations remains constant. Figure 4 show the results of the experiment. This figure shows that setting the number of levels to 3 or 4 is appropriate for the given problem instance since the algorithm maintains a competitive classification error rate.

4.2 Search phase

The search phase involves finding a solution of the smallest problem instance. This solution can be obtained using either an exact or an approximate search method. Exact searches yield optimal solutions for the given feature subspace, but there are strong limitations on the dimensionality of feature subspace that can be solved feasibly. To apply an exact search, the coarsening process has to be performed until a feature subspace with small dimensionality is obtained. In the context of the feature selection problem, there is need to consider the relevance of finding the optimal solution for the coarsest feature subspace to the quality of the desired near-optimal subset for the original feature space. Using a wrapper-based feature selection technique, a solution is usually optimal with respect to a target classifier. That is, the optimal solution can differ for different classifiers. Therefore, using an approximate solution at the coarsest feature subspace may not necessarily impair the quality of the near-optimal subset desired at the original feature space. Heuristic methods provide approximate solutions that are adequate, irrespective of the dimensionality of the original feature space.

4.3 Refinement phase

For the multilevel feature selection technique, the coarsening phase produces a hierarchy of coarse feature subspaces, such that the subspace at level i is (explicitly or implicitly) a subset of the subspace at the next less coarse level $i-1$; the search phase produces the starting solution, i.e., a solution of the coarsest feature subspace; the refinement phase improves upon the starting solution across the feature subspaces in the hierarchy with decreasing coarseness. The refinement phase improves upon the solutions by interpolating them at a coarse level i onto the immediate less coarse level $i-1$, and refining the projected solution in the less coarse subspace.

The interpolation and refining processes depend on how the features in the feature spaces are generated in the coarsening phase. When the coarse feature subspace at level i in the hierarchy consists of synthetic features generated from clusters of features from the next less coarse subspace at level $i-1$, interpolating the solution at level i onto level $i-1$ can involve decomposing each synthetic feature that belongs to the solution at level i into the constituent features at level $i-1$. To refine the solution, the features that result from the projection can be combined to form a subset of features wherein an initial solution can be selected and used to initiate the search heuristics over the subspace at level $i-1$. We refer to the search heuristics that are used for the refinement processes as *refinement heuristics*. When the coarse feature subspaces consist of original features that are selected using the feature pre-setting strategies, the interpolation can simply consist of using the 'best' solution at level i as an initial solution for the refinement heuristics at level $i - 1$. Then, once a solution has been interpolated from level i, it can be improved by the refinement heuristics at level $i - 1$. In the present implementation of the multilevel feature selection technique, the later form of interpolation is used since the coarsening is done using the random pre-setting strategy.

Unlike the search phase, the set of heuristics that can be used in the refinement phase is quite restricted. Search heuristics such as the greedy SFS, SBS, and their variants cannot be used as refinement heuristics since these methods usually create the 'optimal' feature subset from a sequence of addition or elimination of features from a starting subset having a cardinality of 0 or L - the dimensionality of the original feature space. In the present implementation of the multilevel feature selection technique, the tabu search (as implemented in [26]) is used as the refinement heuristic.

To design the refinement phase of the multilevel feature selection technique, we consider the possibility of influencing the behavior of the technique by varying the configurations of the refinement procedure. We investigate the effect of varying the allocation of the refinement resource across the levels on the performance of the technique. The refinement resource in this context refers to the cost of computing the objective function values in order to determine the discriminatory capability of an examined feature subset; the allocation of the resource is based on the number of times the objective function is called during a refinement procedure. Therefore, the allocation of resources directly relates to the number of iterations of the refinement heuristics at each level. We perform experiments to investigate three allocation possibilities: allocating equal amount of resource to refine a solution at each level (i.e., constant resource allocation); increasing the amount of allocated resources with decreasing coarseness of the feature subspaces across the levels; and decreasing the amount of allocated resource with decreasing coarseness of the feature subspaces across the levels. The decrement or increment of the resource allocation across the levels is based on a simple arithmetic progression. The number of iterations for the level having the least amount of allocation is set at a value (i.e., the basic number of iterations) and the subsequent levels are increased progressively by a multiple of the basic number of iterations. In the experiment, we create three instances of the multilevel feature selection algorithm such that each instance is based on one of the allocation possibilities. We compare

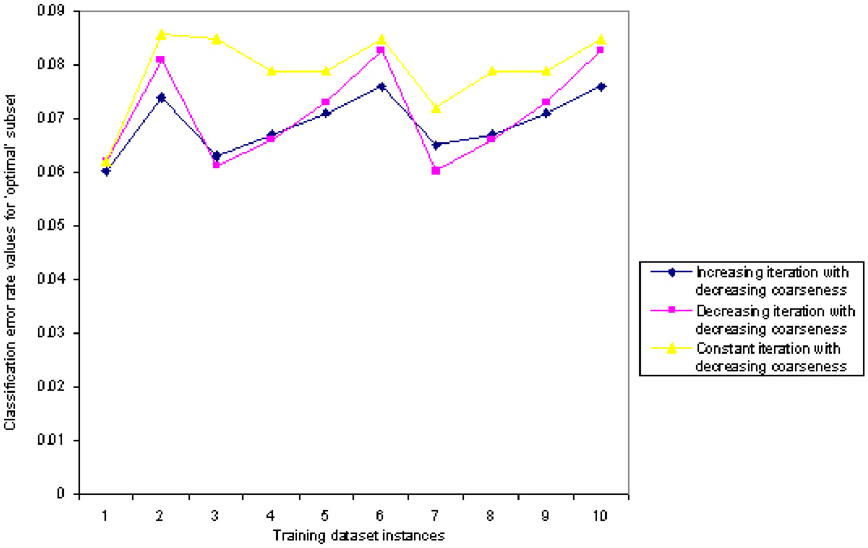


Figure 5: The effect of different refinement possibilities on the performance of the multilevel feature selection technique

the classification error rate derived using the leave-one-out cross validation for the three instances of the multilevel feature selection algorithm. Figure 5 shows the effect of the three refinement scenarios on the performance of the multilevel feature selection technique.

As shown in Figure 5, when the number of iterations across the levels is varied (decreasing or increasing), the multilevel feature selection algorithm finds subsets with consistently smaller classification error than when a constant number of iteration is maintained across the levels. This can be attributed to a more flexible exploration of the hierarchical search framework that is derivable by implicitly allocating more resources to the search method at levels where highly discriminatory subsets can be found. Besides, the configuration of the multilevel feature selection technique wherein the number of iteration increases as the coarseness decreases across the levels enables an intensive search in the subspaces with smaller dimensionality and a more diversified search in subspaces with larger dimensionality. This can explain the relatively stable near-optimal subset generated by this configuration of the multilevel technique. In the present implementation, the allocation of refinement resources increases as the coarseness of the subspaces decreases across the levels.

5 Experimentation

In this section, we present empirical comparisons of the new multilevel feature selection technique with other feature selection approaches. We provide descriptions of the evaluation experiments, the biomedical datasets used, and a discussion of the results.

5.1 Experimental dataset

The dataset used in the experiments is a MRS dataset of biomedical origin from the National Research Councils Institute for Biodiagnostics (NRC-IBD). The dataset consists of 337 labeled samples (175 in class ‘1’ and 129 in class ‘2’) with a feature space dimensionality of 1500 and the cardinality of the desired optimal subset is set at 10. For each complete experimental run, the dataset is randomly partitioned into training and test sets in the ratio 2:1 in a stratified form. That is, the sample size of the training set is 203 (117 from class ‘1’ and 86 from class ‘2’) and the sample size of the test set is 101 (58 from class ‘1’ and 43 from class ‘2’). The *a priori* class labels are used as the basis for the computation of the estimated classification error rates for values, and the classification accuracies. For each training/test set partition the test set is independent of the training set and the test set is used only for external cross-validation.

5.2 Evaluation experiments

Using a synthetic dataset, Zhang and Sun [26] empirically compare the performance of a tabu-search-based feature selection technique with other techniques such as **SFS**, **SBS**, **GSFS**, **GSBS**, plus-l-take-away-r, **SFFS**, **SBFS** and **GA**. The result of the comparison shows that SFS and SBS require the least computational cost to obtain a solution, but these techniques obtain solutions with the worst quality in terms of the estimated error rate. On the other hand, the tabu-search-based feature selection technique obtains better solutions than all the other techniques. These inferences are used to set performance thresholds for our comparison experiments.

Using the described real biomedical dataset, we compare the performance of the new multilevel feature selection technique with the tabu-search-based one, the SFS technique, and a random method. The configuration of the tabu search method underlying the multilevel feature selection technique and the tabu-search-based feature selection technique is based on the recommendations in [26]. For the feature selection problem form by the biomedical dataset, the tabu list size is set to 30, the neighbourhood candidate list size to 100, and the initial solution is randomly selected, based on a Gaussian random number generator. The SFS technique is implemented as a deterministic greedy method. The computational cost of this method is used as the upper limit of the computational requirement for the other techniques. The random feature selection technique simply evaluates the fitness of randomly selected feature subsets and the subset having the best evaluation is considered as the near-optimal subset. This technique is implemented to appraise any claims that the other feature selection techniques select features purely based on probabilistic chances.

To provide a common basis for evaluating the examined feature selection techniques, an equivalent amount of computational cost is assigned to each technique. The computational cost is based on the number of times the objective function value is computed in a complete run of each technique. For instance, the computational cost for SFS can be determined as follows: Given an original feature set F of cardinality L and the cardinality of the desired near-optimal subset is m; the computational cost c_{SFS} of the SFS technique is given as:

$$c_{SFS} \quad = \quad \binom{L}{1} + \binom{L-1}{1} + \cdots + \binom{L-m}{1}. \tag{4}$$

For the feature selection problem instance in context, $L = 1500$ and $m = 10$, therefore, $c_{SFS} = 16445$. A rather lesser amount of computational cost ($c_{SFS} = 15000$) is allocated to the other feature selection techniques. For the multilevel feature selection technique, the computational cost is shared amongst the refinement heuristics according to the implemented refinement option (i.e., increasing computational cost with decreasing coarseness across the multilevel hierarchy). For the 3-level configuration of the multilevel technique, 25 basic iterations are assigned to the coarsest subspace; the next less coarse subspace is assigned 50 basic iterations; and the least coarse subspace is assigned 75 basic iterations. For each coarse subspace, the neighborhood size of the refinement heuristics (tabu search) is set at 100 and this implies that the computational cost per basic iteration is 100. Therefore, the total computational cost c for the multilevel feature selection technique is given as:

$$c = 25 \times 100 + 50 \times 100 + 75 \times 100 = 15000.$$

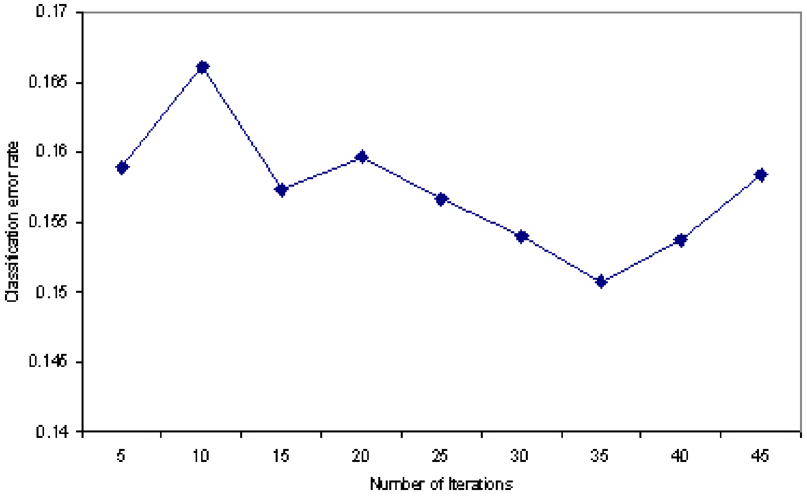


Figure 6: The effect of the number of iterations on the multilevel feature selection technique

The computation cost that is derived by assigning 25 basic iterations to the coarsest subspace is within the range of refinement resource allocation wherein overfitting is less likely to occur using the multilevel feature selection technique.

As shown in Figure 6, the classification error rate on independent test sets begins to increase continually when the number of basic iterations at the coarsest level is set at 35 and beyond. That is, using the near-optimal feature subset selected by the multilevel technique as the underlying feature space in the design of a classifier, the classification performance degrades due to overfitting when the computational cost assigned to the multilevel technique is 21000 (i.e., $c = 35 \times 100 + 70 \times 100 + 105 \times 100 = 21000$) and higher. For the tabu-search-based feature selection technique, the number of basic iterations is set at 150; therefore the total computational cost c is also given as:

$$c = 150 \times 100 = 15000.$$

For the random feature selection technique, the discriminatory capability of 15000 randomly selected subsets is examined and the optimal subset is the subset having the best fitness or evaluation. Therefore, the computational cost for this technique is also 15000.

Some of the examined techniques (multilevel feature selection, tabu-search-based feature selection, and random feature selection) have random components. Examples of the random components are: the random selection of subsets; the random generation of the initial solution and the random selection of candidate sets of solutions from the neighborhood in the tabu search method underlying the tabu-search-based and multilevel feature selection; and the random coarsening of the feature subspaces in the multilevel feature selection. To normalize the randomness in these techniques, the complete run is repeated on the same dataset for a number of times that is predefined by a randomness factor. The average of the evaluation parameters (i.e., the minimization objective function values, the classification accuracies on the training datasets, and the classification accuracies on the independent test datasets) is obtained and analyzed for the evaluated feature selection techniques. The value of the randomness factor is set at 5 for the experiments implemented in this paper. To establish a trend in the comparison of the evaluated feature selection techniques, we perform the experiments on 10 randomly partitioned pairs of training and test sets from the biomedical dataset in order to establish a trend in the evaluation results. The techniques are implemented using Java 2 SDK Standard Edition version 1.4.2 on Microsoft Windows platform and the experiments are executed on a Dell high performance desktop (Pentium 4 CPU 3.00GHz, 1.00GB of RAM) and IBM servers.

5.3 Experimental results and discussion

The results of the experiments are shown in Figures 7 and 8. Figure 7 compares the training and test sets classification accuracies of a simple LDA classifier that is designed using the near-optimal subsets selected by the various techniques. Figure 8 compares the standard deviation on training and test sets. Overall, the comparison of the classification accuracies of the evaluated techniques over independent test set instances shows that multilevel and tabu-search-based feature selection techniques are better than the others feature selection techniques. Moreover, the multilevel technique demonstrates better performance than the tabu-search-based technique in terms of the three evaluation parameters (i.e., classification accuracy on training set, and classification accuracy on independent test set and stability). The performance of the better techniques can be attributed to the exploration strategy of the underlying search method.

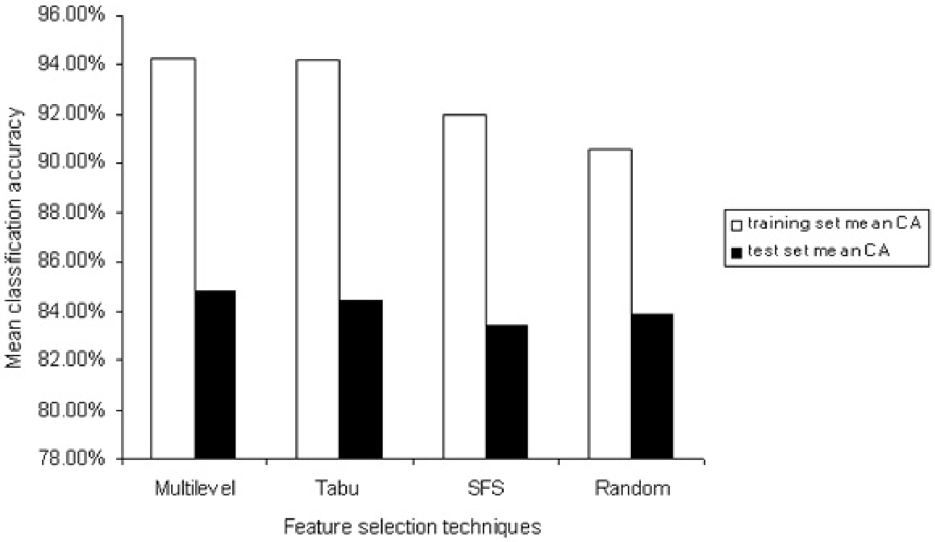


Figure 7: Comparing the classification accuracies

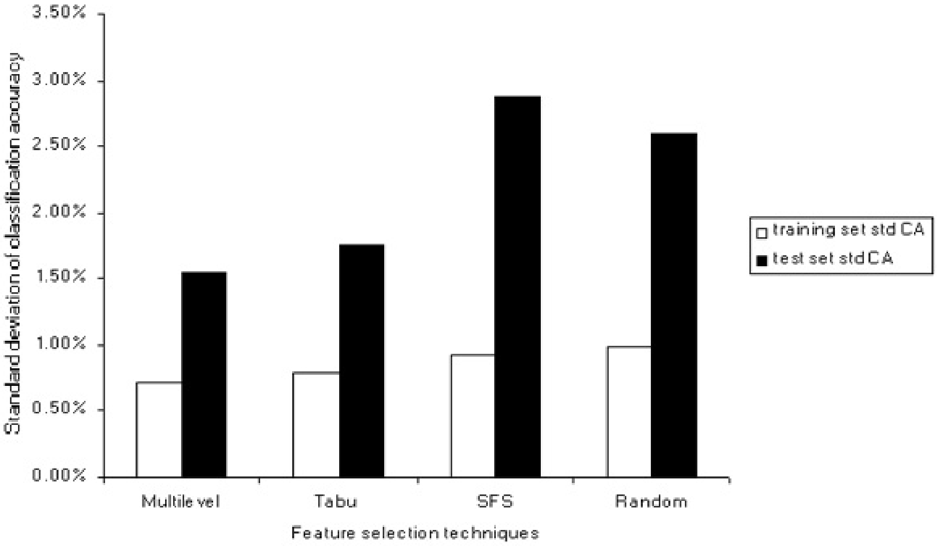


Figure 8: Comparing standard deviations

6 Conclusions & future work

Feature selection techniques for biomedical data are usually based on an underlying search heuristic. We propose a new feature selection technique for biomedical data that is based on a multilevel search algorithm. Using experimental results from the implementation of the multilevel feature selection, a tabu-search-based feature selection, sequential forward selection, and random feature selection, on practical biomedical datasets, we compared the quality of the near-optimal subset found by these techniques, given an equivalent amount of computational cost. The results show tabu search and multilevel search as promising meta-heuristics for finding near-optimal feature subsets in biomedical datasets. These meta-heuristics consistently find near-optimal feature subsets having lower classification error rate estimations (i.e., objective function values) and higher classification accuracies on the training set than those found by **SFS** and the random feature selection. Also, the meta-heuristics consistently produce relatively more stable classification accuracies than the **SFS** and the random feature selection, on independent test datasets, a more relevant consideration.

In our future work, we will design and develop an enhanced version of the multilevel feature selection technique and use it to identify biomarkers in biomedical datasets. We will also investigate the strength of the coarsening strategy in the multilevel technique in adapting to different forms of biomedical datasets.

Acknowledgments

Funding for this project has been provided by the Natural Sciences and Engineering Council of Canada, through its Discovery Grant programs and by the University of Manitoba through the University of Manitoba Graduate Fellowship.

References

[1] D. W. Aha and R.L. Bankert. A comparative evaluation of sequential feature selection algorithms. In *Proc. Fifth International Workshop on Artificial Intelligence and Statistics*, pages 1–7, 1995.

[2] C. Ambroise and G. J. McLachlan. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the National*

Academy of Sciences of the United States of America, 99(10):6562–6566, 2002.

[3] S.A. Bustin and S. Dorudi. The value of microarray techniques for quantitative gene profiling in molecular diagnostics. *Trends in Molecular Medicine*, 8:269–272, 2002.

[4] T. M. Cover. The best two independent measurements are not the two best. *IEEE Trans. Systems, Man, and Cybernetics*, 4:116–117, 1974.

[5] T.G. Crainic, Y. Li, and M. Toulouse. A Simple Cooperative Multilevel Algorithm for the Capacitated Multicommodity Network Design. *Computer & Operations Research*, 33(9):2602–2622, 2006.

[6] C. Furlanello, M. Serafini, S. Merler, and G. Jurman. An accelerated procedure for recursive feature ranking on microarray data. *Neural Networks*, 16:641–648, 2003.

[7] I. P. Gent and T. Walsh. An Empirical Analysis of Search in GSAT. *Journal of Artificial Intelligence Research*, 1:47, 1993.

[8] F. Glover and M. Laguna. *Tabu Search*. Kluwer Academic Publishers, 1997.

[9] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(7-8):1157–1182, 2003.

[10] R.A. Iles, A.N. Stevens, and J. R. Griffiths. NMR Studies of metabolites in living tissue. . *Progress in Nuclear Magnetic Resonance Spectroscopy*, 15(1-2):49–200, 1982.

[11] A. K. Jain, R. P. W. Duin, and J. Mao. Statistical pattern recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):4–37, 2000.

[12] G. Karypis, V. Aggarwal, V. Kumar, and S. Shekhar. Multilevel Hypergraph Partitioning: Application in VLSI Domain. *IEEE Transactions on VLSI Systems*, 07(01):69–79, Mar. 1999.

[13] K. Kira and L.A. Rendell. The feature selection problem: traditional methods and a new algorithm. In *Proc. 10th National Conference on Artificial Intelligence*, pages 129–134, 1992.

[14] J. Kittler. Feature set search algorithms. In C.H. Chen, editor, *Pattern Recognition and Signal Processing*, pages 41–60, 1978.

[15] R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273–324, 1997.

[16] W. Li. Dynamics of Local Search Trajectory in Traveling Salesman Problem. *Journal of Heuristics*, 11(5-6):507–524, 2005.

[17] P. M. Narendra and K. Fukunaga. A branch and bound algorithm for feature subset selection. *IEEE Trans. Computers*, 26(9):917–922, 1977.

[18] A. E. Nikulin, B. Dolenko, T. Bezabeh, and R. L. Somorjai. Near-optimal region selection for feature space reduction: novel preprocessing methods for classifying MR spectra. *NMR Biomedicine*, 11:209–216, 1998.

[19] P. Pudil, J. Novovicova, and J. Kittler. Floating search methods in feature selection. *Pattern Recognition Letters*, 15(11):1119–1125, 1994.

[20] W. Siedlecki and J. Sklansky. A note on genetic algorithm for large-scale feature selection. *Pattern Recognition Letters*, 10(11):335–347, 1989.

[21] P. Somol, P. Pudil, F. J. Ferri, and J. Kittler. Fast branch and bound algorithm in feature selection. *In Proceedings Fourth World Multiconf. Systemics, Cybernetics, and Informatics*, 7(1):646–651, 2000.

[22] P. Somol, P. Pudil, and J. Kittler. Fast branch and bound algorithm in feature selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(7):900–912, 2004.

[23] R. L. Somorjai, B. Dolenko, and R. Baumgartner. Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, and cautions. *Bioinformatics*, 19(12):1484–1491, 2003.

[24] S. Stearns. On selecting features for pattern classifiers. In *Proc. 3rd International Joint Conference on Pattern Recognition*, pages 71–75, 1976.

[25] C. Walshaw. Multilevel refinement for combinatorial optimisation problems. *Annals Oper. Res.*, 131:325–372, 2004.

[26] H. Zhang and G. Sun. Feature selection using tabu search method. *Pattern recognition*, 35(3):701–711, 2002.