称号及び氏名	博士（工学）Mohd Saberi Bin Mohamad
学位授与の日付	2010 年 3 月 31 日
論　文　名	「Studies on Intelligent Approaches to Select Informative Genes from Gene Expression Data for Cancer Classification」
論文審査委員	主査　大松　繁 副査　市橋　秀友 副査　松本　啓之亮

論文要旨

Gene expression technology namely microarrays, offers the ability to measure the expression levels of thousands of genes simultaneously in biological organisms. Gene expression data produced by the microarrays are expected to be of significant help in the development of efficient cancer diagnoses and classification platforms. Many researchers have analyzed gene expression data to select a small subset of informative genes for cancer classification using various intelligence approaches. As a result, the selection of the small subset has improved classification accuracy. However, due to the small number of samples compared to the huge number of genes (high-dimension), irrelevant genes, and noisy genes, the most approaches face difficulties to select the small subset. Therefore, the ultimate goal of this Ph.D. research is to propose intelligent approaches for selecting a small (near-optimal) subset of informative genes from gene expression data for cancer classification. Support vector machine classifiers (SVMs) were used to measure classification accuracies on the gene subsets that produced by all the proposed approaches. The first six proposed approaches were produced based on genetic algorithms (GAs), whereas the remaining approaches were extensions of particle swarm optimization (PSO).

First, a multi-objective strategy in GAs was proposed to improve the performance of GAs that uses a single-objective approach. It is described in Chapter 2. The strategy has been developed based on classification accuracy maximization and gene subset size minimization. In this strategy, multi-objective problems have been accommodated by using specialized fitness functions in GAs. The ultimate goal of the strategy is to

search and select a nondominated gene subset Pareto front. It was tried on four benchmark gene expression data sets and obtained encouraging results on those data sets as compared with an approach that used a single-objective strategy in GAs.

Second, an approach using two hybrid methods was then introduced. It is discussed in Chapter 3. This approach includes a hybrid of GAs and SVM classifiers (GASVM) and an improved GASVM (GASVM-II). It was developed to overcome the limitations of GASVM and GASVM-II that developed separately. In the first phase, GASVM-II is applied to manually select genes from overall gene expression data to produce a subset of genes. It is used to reduce the dimensionality of the data, and therefore the complexity of the search or solution spaces can also be decreased. In the second phase, GASVM is used to select and optimize a small subset of informative genes from the subset that is produced by the first phase. The approach was assessed and evaluated on four well-known gene expression data sets, showing competitive results.

Third, a cyclic hybrid method based on GASVM-II has been proposed. Chapter 4 describes the detail of the cyclic hybrid method. It differs from other GASVM-based methods in one major part, namely it involves a cyclic approach, whereas the GASVM-based methods did not use any cyclic approach. Basically, the cyclic hybrid method repeats the process of GASVM-II to iteratively reduce the dimensionality of data and produce potential gene subsets. Five real gene expression data sets were used to test the effectiveness of the method. Experimental results show that the performance of the proposed method is superior to other experimental methods and previous related works in terms of classification accuracy and the number of selected genes. In addition, a scatter gene graph and the list of informative genes in the best gene subsets are also presented for biological usage.

Fourth, an iterative approach based on GASVM is then developed. Chapter 5 concerns on the discussion of the iterative approach. Generally, it is completely same with the proposed cyclic hybrid, but it uses GASVM to replace GASVM-II in the process to yield potential gene subsets and reduce the dimensionality of data iteratively. To demonstrate its effectiveness, four gene expression data sets are used. Experimental results show that the approach is efficient in finding genes for classifying cancer classes.

Fifth, a two-stage method was proposed to surmount the drawbacks of

GASVM-based methods in previous related works. It is discussed in Chapter 6. In the first stage, a filter method either gain ratio (GR) or information gain (IG) is applied on overall gene expression data to pre-select genes and finally produces a subset of genes. The dimensionality of data is also can be decreased. The second stage applies GASVM to automatically optimize the gene subset that is produced by the first stage. As a result, it yields a small (near-optimal) subset of informative genes. They were evaluated on four publicly available gene expression data sets. The results show that the proposed method outperforms existing methods and other experimental methods.

Sixth, since a two-stage method does not perform well as expected, a three-stage method that includes frequency analysis in the third stage was proposed. Chapter 7 describes this three-stage method. This frequency analysis is implemented to identify the most frequently selected genes in near-optimal gene subsets. The most frequently selected genes in the near-optimal gene subsets are presumed to be the most relevant for the cancer classification. The three-stage method differs from methods in previous works in one major part. The major difference is that it involves three stages (using a filter method, a hybrid method, and frequency analysis), whereas the previous works usually had only one stage (using a filter method or a hybrid method) or two stages (using a filter method and a hybrid method). The proposed method has been tested and evaluated for gene selection on five gene expression data sets that contain binary classes and multi-classes of tumor samples. Based on the experimental results, the performance of proposed method is better than other methods in previous related works. The list of informative genes in the final gene subset is also presented for biological usage

Seventh, a modification of binary PSO was proposed to overcome the limitations of the conventional version of binary PSO and previous PSO-based methods. It is introduced in Chapter 8. A scalar quantity that called particle's speed and a novel rule for updating particle's positions are introduced in this modified binary PSO. This particle's speed and rule are proposed in order to reduce the probability of genes to be selected for the cancer classification. By performing experiments on 12 different gene expression data sets, including multi-class data sets, the modified binary PSO outperforms other previous related works, including the conventional version of binary PSO in terms of classification accuracy, the number of selected genes, and running times.

Eighth, an enhancement of binary PSO with the constraint of particle's velocities was proposed. It is completely discussed in Chapter 9. The constraint is introduced in the enhanced binary PSO to increase the probability of genes to be unselected for the classification. Experimental results on twelve actual gene expression data sets show that the performance of the proposed approach is superior to other previous related works, as well as to conventional binary PSO tried in this work.

Ninth, a modified sigmoid function and the particle's speed were introduced and implemented in binary PSO. Chapter 10 describes this modified binary PSO. This modified sigmoid function and particle's speed decrease the probability of genes to be selected for the cancer classification. The proposed method was experimentally assessed on twelve well-known gene expression data sets. In this sense, comparisons with the existing of binary PSO and several PSO-based methods show competitive results.

Finally, Chapter 11 gives the conclusion remarks and suggests interesting directions for future researches. As a conclusion, twelve benchmark gene expression data sets have been used to test the effectiveness of the proposed intelligent approaches. Overall, experimental results show that the performances of the proposed approaches are superior to previous related works as well as methods experimented in this work in terms of classification accuracy, the number of selected genes, and running times.

審査結果の要旨

本論文は、遺伝子発現量データから有益な情報を持った遺伝子の小規模集団を選択するために､遺伝的アルゴリズム(GA)と粒子群最適化(PSO)に基づいた知的手法について研究したものであり、以下の成果を得ている。

（１）癌分類を行うために、遺伝子発現量データから有益な情報を持った遺伝子の小規模集団を選択するために遺伝的アルゴリズムに基づく知的手法を提案した。遺伝的アルゴリズムとサポートベクターマシンにおける多目的戦略を提案し、その有効性を検証した。

（２）ハイブリッド手法の組み合わせ、巡回ハイブリッド手法、繰り返し手法、二段階法、三段階法を提案し、次元削減、無関係遺伝子のフィルタ除去、ノイズを含む遺伝子の除去を行い、癌分類のための最適に近い遺伝子集団を抽出した。5 種類の遺伝子表現データに

よる実験により、提案手法の有用性を分類精度と選択された遺伝子数の観点から定量的に検証した。

（３）生物の集団行動規則にヒントを得た二値 PSO を拡張した、改変 PSO、拡張改良 PSO を提案した。ここでは、粒子速度、粒子速度更新の新方式、粒子速度制約、改変シグモイド関数を新たに導入し、12 個のベンチマーク遺伝子表現データセットで評価し、二値 PSO を含む関連する従来研究に比して分類精度と遺伝子選択数の観点から優れた結果を得た。とくに、提案手法は二値 PSO に比べて所要時間の点で優位性があり、実際の遺伝子抽出への適用を可能とした。

以上の成果は、遺伝子発現量データから有益な情報を持った遺伝子の小規模集団を選択するための高精度かつ高速な手法を実現していることから、広範囲な分野の学術的、産業的な発展に貢献するところが大である。また、申請者が自立して研究活動を行うのに必要な能力と学識を有することを証したものである。