The k-Feature Set Problem is $W[2]$-complete

Carlos Cotta[1] and Pablo Moscato[2]

[1] Dept. Lenguajes y Ciencias de la Computación, Universidad de Málaga
ETSI Informática (3.2.49), Campus de Teatinos, 29071-Málaga, Spain

[2] Grupo de Engenharia de Computação em Sistemas Complexos,
Departamento de Engenharia de Computação e Automação Industrial,
Universidade Estadual de Campinas,
C.P. 6101, Campinas, SP, CEP 13083-970, Brazil

ccottap@lcc.uma.es

Abstract

We prove the $W[2]$-completeness of the feature subset selection problem when the cardinality of the subset is the parameter. Aside from the many applications the problem has in the data mining literature, the problem is highly relevant in Computational Biology since it arises in differential gene expression analysis using microarray technologies. It is also related to genetic-based prognosis and regulatory interaction discovery using DNA chip technologies.

1 Introduction

Interpreting the functional relationships between genes appears as one of the greatest challenges to be addressed by scientists [7, 12]. This task has posed very interesting optimization problems. Among these, one can cite the conspicuous Feature Set problem. According to Davies and Russell [3], the decision version of this problem is defined as follows:

- **Input**: A set X of examples (which are composed of a binary value specifying the value of the *target feature* and a vector of n binary values specifying the values of the other features) and an integer $k > 0$.
- **Question:** Does there exist a set S of k *non-target* features ($S \subseteq \{1, \cdots, n\}$) such that no two examples in X that have identical values for all the features in S have different values for the target feature ?

The Feature Set problem emerges as a crucial component in areas such as gene discovery, disease diagnosis, drug discovery or pharmacogenomics, toxicogenomics [10], cancer research [8], and predictive genomic medicine just to mention a few. For example, diagnostic classifiers for preventing diseases, such as cancer, can be built on the basis of *labeled* data sets, obtained by measuring the expression levels of a number of genes in two kinds of tissue, one with a tumor and the other of a non-tumor. Determining the relevant genes for prognosis purposes can be clearly reduced to Feature Set. A similar scenario can be found in the inference of Boolean Networks for modeling gene regulation mechanisms and/or genetic networks [1, 6]. In this case, each gene and each biological *stimulus* (i.e., any chemical or physical factor that influences the genetic network and is itself neither a gene nor a gene product) is represented by a node in a directed graph. A Boolean function is assigned to each of these nodes in order to determine the expression level (*expressed* or *non-expressed*) of the corresponding gene product according to the expression levels of other gene products (incoming arcs in the graph). Obviously, these incoming arcs must allow constructing a truth table matching experimental data. Again, selecting appropriate inputs for each node can be casted as a Feature Set instance.

It is interesting to note that Feature Set can be shown to be NP-complete by a reduction from Vertex Cover [3]. This work is concerned with its parameterized complexity though. In this sense, it is natural to consider the parameterized version of Feature Set in which the cardinality of the set is the parameter. Consider that a large number of expression levels can be measured in the same experiment using the *DNA microarray* [2] technology, but only a few observations can be done (typically the ratio of these two is on the order of 1/100). Hence, large feature sets are prone to overfitting. Additionally, small feature sets play a central role in many supervised learning problems for several reasons, e.g., generalization performance, running time requirements, and problem-dependent interpretational issues among others. This is precisely the case of learning approaches based on *support vector machines* [11], one of the most popular pattern recognition tools currently used for genomic information interpretation.

The main result of this work is showing that this parameterized version of Feature Set is $W[2]$-complete. This is done in the next section.

2 Parameterized Complexity of Feature Set

In order to analyze the parameterized complexity of the Feature Set problem, let us conveniently reformulate the problem as follows:

- **Instance**: A set of m examples $X = \{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$, such that for all i, $x^{(i)} = \{x_1^{(i)}, x_2^{(i)}, \ldots, x_n^{(i)}, t^{(i)}\} \in \{0,1\}^{n+1}$, and an integer $k > 0$.
- **Question:** Does there exist a *feature set* S, $S \subseteq \{1, \cdots, n\}$, with $|S| = k$ and such that for all pairs of examples $i \neq j$, if $t^{(i)} \neq t^{(j)}$ then exists $l \in S$ such that $x_l^{(i)} \neq x_l^{(j)}$?

Now, the main result of this section can be stated in the following proposition:

Proposition 1: *The parameterized version of* FEATURE SET *in which the number of features is taken as a parameter is* $W[2]$*-complete.*

Proof: To prove the $W[2]$-completeness of FEATURE SET it is necessary to show that (a) the problem is $W[2]$-hard, and (b) the problem belongs to $W[2]$. A proof of the first item is given in Subsection 2.1. As to the second item, a proof is provided in Subsection 2.2. □

2.1 FEATURE SET is $W[2]$-hard

To prove the $W[2]$-hardness of FEATURE SET a parameterized reduction from the DOMINATING SET problem (a $W[2]$-complete problem [4]) will be shown. The source problem is defined as follows:

- **Instance:** An undirected graph $G(V, E)$, with $|V| = n$, an integer $k > 0$.
- **Question:** Does there exist a set D of k nodes ($D \subseteq V$), such that for every $w \in V \backslash D$, an edge $(v, w) \in E$ exists, with $v \in D$?

The size of the set D will be the parameter. Now, a *parametric transformation* from a parameterized language L to a parameterized language L' is an algorithm that transforms a pair $\langle x, k\rangle$ into a pair $\langle x', k'\rangle$ such that $k' = g(k)$, and the computation is accomplished in $\mathbf{TIME}(f(k)|x|^c)$, where $|x|$ is the size of x, c is a constant, and f, g are arbitrary functions.

Let L be the DOMINATING SET problem, and let L' be the FEATURE SET problem. Let the Boolean matrix X represent an instance of L'. The reduction is done as follows: The first n rows of X encode the *solid neighborhood* of every node $v \in V$, i.e., $x_j^{(i)} = 1 \Leftrightarrow [(i = j) \vee (v_i, v_j) \in E]$. Also, $t^{(i)} = 1,\ 1 \leq j \leq n$. The $(n + 1)$-th row of X represents a *null* example, i.e., $x_j^{(n+1)} = 0,\ 1 \leq j \leq n$, and $t^{(n+1)} = 0$. Clearly, this reduction can be done in $\mathbf{TIME}\ (n^2)$. Now, it will be shown that $D \subseteq V$ is a valid solution of the DOMINATING SET problem having $\langle G(V, E), k\rangle$ as input if, and only if, D is a valid solution of the corresponding reduced instance $\langle X, k\rangle$ of the FEATURE SET problem.

First, assume that D is a solution of L. Then, for every $w \notin D$, there exists $v \in D$ such that $(v, w) \in E$. This means that for every $v_i \notin D$, $x_j^{(i)} = 1$, with $v_j \in D$. It is easy to see that the vertex indexes in D define a set S which is a solution of L', since the only way for not being a feature set is having row r ($r \leq n$) with 0s at all the positions indicated by S.

Conversely, assume that S is a solution of L'. This means that for all $i \neq j$ if $t^{(i)} \neq t^{(j)}$ then there exists an $l \in S$ such that $x_l^{(i)} \neq x_l^{(j)}$. By the construction process, there exists only one example (the null example) which has a different output. Then if S is a solution we have that for all $i \neq n + 1$ exists at least one l (which we can denote as $l(i)$) such that

$x^{(i)}_{l(i)} = 1$. This means that all vertices in V are connected by an edge to a vertex defined by the set S. Hence S induces a dominating set in V. Since $k' = k$ and the instance can be constructed in **TIME** $(|x|)$, the proof is completed.

2.2 FEATURE SET belongs to W[2]

The proof of membership to $W[2]$ is done via Boolean circuits. It is shown that a circuit allowed by the definition of $W^*[2]$ can be constructed such that the existence of satisfying weight-k input (the weight of a Boolean vector is the number of 1s in it) implies the existence of a weight-k solution of the FEATURE SET problem instance (notice that different circuits can be defined for different values of k). Since $W^*[2] = W[2]$ (see [5]), this completes the proof.

Similarly to [4], let $V = \{v[b,i] : b = 1, 2, \cdots, k \text{ and } i = 1, 2, \cdots, n\}$ be the inputs of the circuit. V can be viewed as a set of k "choice blocks" of size n. The circuit will be defined so as to require that any satisfying truth assignment make exactly one variable in each of these blocks `true`, thus indicating a set of k variables. Let us now define the following expressions:

$$E_1^1 = \bigwedge_{1 \le b < b' \le k} \bigwedge_{1 \le i \le n} \left(\neg v[b,i] \vee \neg v[b',i]\right) \tag{1}$$

$$E_1^2 = \bigwedge_{1 \le b \le k} \left(\bigvee_{1 \le i \le n} v[b,i] \right) \tag{2}$$

$$E_1 = E_1^1 \wedge E_1^2 \tag{3}$$

It is easy to see that E_1 is `true` if, and only if, exactly k different variables are selected as input (one in each block). Now, consider the expressions below:

$$E_2 = \bigwedge_{t^{(i)} \neq t^{(j)}} E(i,j) \tag{4}$$

$$E(i,j) = \bigvee_{1 \le r \le k} E^r(i,j) \tag{5}$$

$$E^r(i,j) = \bigvee_{x_s^{(i)} \neq x_s^{(j)}} v[r,s] \tag{6}$$

Clearly, $E^r(i,j)$ is satisfied if, and only if, a variable taking different values in examples i and j is selected in block r. Thus, $E(i,j)$ is satisfied if, and only if, a variable taking different values in examples i and j is selected in any block. Finally, E_2 is satisfied if, and

only if, for each pair of examples with different target values, a variable taking different values in these examples has been selected. Notice that if small gates are allowed fan-in bounded by $f(k) = k$, the weft number of E_2 (i.e., the maximum number of gates with fan-in not bounded by k in an input-to-output path) is two, i.e., the big and in E_2 and the big or in $E^r(i, j)$.

Hence the expression $E = E_1 \wedge E_2$ has weft-two (E_1 has weft-one, but it is not in the same path of E_2), and is satisfied if, and only if, a weight-k input is given, and the selected variables allow discriminating between any two examples with different target values. Furthermore, the depth of the circuit is constant, so it verifies all conditions for being in $W^*[2] = W[2]$.

As mentioned above, having found a reduction from DOMINATING SET (a $W[2]$-complete problem) to FEATURE SET, and having established the membership of the latter to $W[2]$ implies the $W[2]$-completeness of FEATURE SET. Notice that the current measure of intractability according to parameterized complexity is $W[1]$-hardness [4]. Thus, this $W[2]$-completeness result poses a reasonably strong argument against the availability of efficient algorithms for solving general instances of this problem.

3 Conclusions

This work has studied the parameterized complexity of FEATURE SET, a crucial problem for gene expression data mining. It has been shown that the problem is $W[2]$-complete. This highlights the intrinsic difficulty of approaching arbitrary instances of this problem via exact algorithms, thus suggesting two main lines for future developments. On one hand, the identification of tractable subclasses of the problem can be approached. In this sense, we have some preliminary results regarding the number of active bits per example. On the other hand, the use of heuristic techniques (e.g., memetic algorithms [9]) for solving this problem is an appealing option. This raises interesting theoretical issues, mainly with respect to the innards of *recombination* procedures for this problem. Work is in progress in this area as well.

Acknowledgments

The authors thank M. Fellows and V. Raman for their comments on an earlier version of this manuscript. P.M. also wants to thank M. Fellows and F. Rosamond for their hospitality and useful discussions while at University of Victoria as well as support by CNPq, Brazil, grant Proc. 52.1100/01-1. C.C. is partially supported by Spanish CICYT under grant TIC1999-0754-C03.

References

[1] T. Akutsu, S. Miyano, and S. Kuhara. Algorithms for identifying boolean networks and related biological networks based on matrix multiplication and fingerprint function. *Journal of Computational Biology*, 7:331–343, 2000.

[2] P.O. Brown and D. Botstein. Exploring the new world of the genome with DNA microarrays. *Nature Genetics*, 21:33–37, 1999.

[3] S. Davies and S. Russell. NP-completeness of searches for smallest possible feature sets. In R. Greiner and D. Subramanian, editors, *AAAI Symposium on Intelligent Relevance*, pages 41–43, New Orleans, 1994. AAAI Press.

[4] R. Downey and M. Fellows. Fixed parameter tractability and completeness I: Basic theory. *SIAM Journal of Computing*, 24:873–921, 1995.

[5] R. Downey and M. Fellows. *Parameterized Complexity*. Springer-Verlag, 1998.

[6] T.E. Ideker, V. Thorsson, and R.M. Karp. Discovery of regulatory interactions through perturbation: Inference and experimental design. *Pacific Symposium on Biocomputing*, 5:302–313, 2000.

[7] E.V. Koonin. The emerging paradigm and open problems in comparative genomics. *Bioinformatics*, 15:265–266, 1999.

[8] J. Marx. DNA arrays reveal cancer in its many forms. *Science*, 289:1670–1672, 2000.

[9] P. Moscato and C. Cotta. A gentle introduction to memetic algorithms. In F. Glover and G. Kochenberger, editors, *Handbook of Metaheuristics*, pages 105–144. Kluwer Academic Publishers, Boston MA, 2003.

[10] E.F. Nuwaysir, M. Bittner, J. Trent, J.C. Barrett, and C.A. Afshari. Microarray and toxicology: The advent of toxicogenomics. *Molecular Carcinogenesis*, 24:153–159, 1999.

[11] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik. Feature selection for SVMs. In S.A. Solla, T.K. Leen, and K.-R. Muller, editors, *Advances in Neural Information Processing Systems 13*. MIT Press, 2001.

[12] J.C. Wooley. Trends in computational biology: a summary based on a RECOMB plenary lecture. *Journal of Computational Biology*, 6:459–474, 1999.