A Tabu Algorithm for Homogeneous Partition of Samples

Michel Vasquez, Stefan Janaqi

A Tabu Algorithm for Homogeneous Partition of Samples

Michel Vasquez[*] *Stefan Janaqi*[*]

[*] LGI2P, EMA-EERIE
Parc Scientifique Georges Besse, 30035 Nîmes cedex 01, France
Email: {Michel.Vasquez, Stefan.Janaqi}@site-eerie.ema.fr

1 Introduction

The construction of predicting models by learning algorithms, such as statistical regression or neural networks, needs a *learning set* $\mathcal{A} \subset \mathcal{D}$, where $\mathcal{D}$ is a set of samples $\mathcal{D} = \{(\mathbf{x}_k, y_k) | k = 1, \ldots, n\}$. The output of the learning algorithm is a predicting function $y = L(\mathbf{x})$ that minimizes a given error criteria such as $E(L, \mathcal{A}) = \sum_{\mathbf{x}_k \in \mathcal{A}} (y_k - L(\mathbf{x}_k))^2$. Recent work on learning theory (see [10]) have shown that a "little" learning error $E(L, \mathcal{A})$ does not necessarily imply a "little" error on a new sample $\mathbf{x}$. To illustrate this idea, think of L as a polynomial of a given degree p. When p is near to $|\mathcal{A}|$, the polynomial $L(\mathbf{x})$ fits the data very well, but this high degree will add noisy oscillations even where we do not need them. To get round this difficulty, the statisticians employ a *testing set* $\mathcal{T} \subset \mathcal{D}$ different from $\mathcal{A}$. The testing error is $E(L, \mathcal{T}) = \sum_{\mathbf{x}_k \in \mathcal{T}} (y_k - L(\mathbf{x}_k))^2$. The model L is considered valid when $E(L, \mathcal{T}) \approx E(L, \mathcal{A})$. This supposes that $\mathcal{A}$ and $\mathcal{T}$ realize a *homogeneous partition* of $\mathcal{D}$.

In the following we give a geometric criterion to measure the homogeneity of a partition. We will see that the "best" partition is hard to define and above all, hard to compute. This justifies our choice for a heuristic searching algorithm like *tabu search* which will look for a "good" partition. This tabu search algorithm has a set of important features including an efficient neighborhood and mechanisms for dynamic tabu tenure, intensification and diversification.

Before we go into the details of our approach, we want to emphasize the necessity for a very fast method in order to meet the need to supervise real world processes. The choices and solutions that follow were mostly guided by this *time constraint* as well as the robustness of the *homogeneity criterion*.

2 A geometric criterion of homogeneity

A first measure of the homogeneity of a partition $[\mathcal{A}, \mathcal{T}]$ of $\mathcal{D}$, needs the estimation of the probability densities f_A, f_T and f_D on $\mathcal{A}$,$\mathcal{T}$ and $\mathcal{D}$ respectively. Then, a homogeneous partition is obtained by minimizing

$$K(\mathcal{A}, \mathcal{T}) = d_K(\mathcal{D}, \mathcal{A}) + d_K(\mathcal{D}, \mathcal{T}) \tag{1}$$

where d_K is the Kullback-Leibler distance (see [7, 8]) given by $d_K(\mathcal{D}, \mathcal{A}) = \int (f_D - f_A)^2$. Even when the densities above are given by closed formulas (which is not the case for our instances), the integration is performed by numerical methods which remain rather slow in high dimension. The reason is the number of functional estimations which grows exponentially even for sophisticated integration methods (see [6, 2]). It is clear then, that a searching algorithm that minimizes this measure of homogeneity would hardly satisfy the time constraint.

Thus we use the following geometric criterion. Let E_A, E_T and E_D be the ellipsoids defined by the covariance matrices of $\mathcal{A}$,$\mathcal{T}$ and $\mathcal{D}$ respectively. We can rather easily (see [9]) calculate the volume minimal ellipsoid E_M containing $E_A \cap E_T$. The homogeneity of the partition $[\mathcal{A}, \mathcal{T}]$ is given by:

$$H(\mathcal{A}, \mathcal{T}) = \|g_A - g_T\| + (volume(E_D) - volume(E_M))^2 \tag{2}$$

where, g_A and g_T are the centers of gravity of $\mathcal{A}$ and $\mathcal{T}$. A good feature of the volume is that it is invariant under affine transformation of data. Remember that most of the normalization of the data (as used in statistics) are affine transformations. On the other side, this volume criterion has a serious drawback. It does not work on high dimension. The difficulty comes from the fact that the ratio of the volume of a sphere over the volume of the circumscribed cube tends to 0 when dimension increases. We show latter that one could obtain satisfactory results for dimensions up to 15.

For a given a partition, the subroutine that calculates $H(\mathcal{A}, \mathcal{T})$ is written in `Fortran 90` and makes use of the mathematical functions of IMSL library.

3 A tabu search of a homogeneous partition

Tabu search TS is a heuristic designed for tackling hard combinatorial optimization problems. For a comprehensive presentation of TS, we refer [4]. Contrary to genetic algorithms, where randomness is extensively used, TS visits the *search space* in a more systematic way based on adaptive memory and learning. Given the cardinality of data set $|\mathcal{D}| = n$ and $|\mathcal{A}| = k$, our search space S is composed of all binary vectors $s \in \{0, 1\}^n$ having $k = |\mathcal{A}|$ coordinates at 1 and $n - k = |\mathcal{T}|$ coordinates at 0. A typical instance has $n = 3000, k = \frac{3}{4}n$ giving, by Stirling formula:

$$|S| = \left(\begin{array}{c} n \\ k \end{array} \right) \approx \frac{4^{n+1}}{\sqrt{6\pi n}\, 3^{0.75n}} \approx 2^{2428}$$

Given the correspondence between the partitions of $\mathcal{D}$ and the binary vectors s of S, in the following we will note $H(s)$ for $H(\mathcal{A}, \mathcal{T})$.

3.1 Neighborhood and move

For a given instance (S, H) or our optimization problem, characterized by the search space S and the objective function H (see 2), a neighborhood N is first defined to associate, to each $s \in S$ a nonempty subset $N(s)$ of S. A neighbor s' of s is obtained by choosing two indices i and j such that $s_i = 1$ and $s_j = 0$ and putting $s'_i = 0$ and $s'_j = 1$. We note this movement from s to s' by $move(i, j)$. The choice of the indices i, j is done as follows:

- Take the hyperplane $P(s)$ containing $g = \frac{1}{2}(g_A + g_T)$ and orthogonal to $\mathbf{n} = g_A - g_T$. If, for a given ϵ, $\|g_A - g_T\| < \epsilon$ then $\mathbf{n}$ is the greatest axe of the ellipsoid E_M;
- Choose i, j such that $\mathbf{x}_i \in \mathcal{A}$ is in one side of $P(s)$ and $\mathbf{x}_j \in \mathcal{T}$ is on the other side.

Similar neighborhoods, based on this "adding-dropping" technique has been used in many heuristic algorithms (see [1, 3, 5]). The TS algorithm examines the value of $H(s')$ for each neighbor of s and chooses one that has the minimum value. In order to do this, we keep in a special data structure δ, $move(i, j)$ and the value $H(s') - H(s)$. Each time a move is carried out, the elements of δ affected by the move, are updated accordingly. It is clear that $|N(s)| < k \times (n - k)$. Thus, the initialization and updating of δ requires, in the worst case, time $O(k \times (n - k))$.

3.2 Tabu list management

The role of a *tabu list* is to prevent from short-term cycling. Each time a $move(i,j)$ is carried out, $move(j,i)$ is classified tabu (that is forbidden) for a number of iteration (*tabu tenure*). The tabu tenure $t(j,i)$ is dynamically defined by a function depending on $c(i,j)$ and $freq(i,j)$, where $c(i,j)$ is proportional to the inverse of $d(\mathbf{x}_i, P(s)) + d(\mathbf{x}_j, P(s))$ and $freq(i,j)$ is the number of times the $move(i,j)$ is done. The idea behind the term $c(i,j)$ is to promote the choice of points to exchange far from $P(s)$.

In order to implement the tabu list, a vector T is used (as suggested in [4]) containing the numbers $t(j,i) + iter$, where $iter$ is the current number of iterations. In this way, it is easy to know whether $move(j,i)$ is tabu or not at iteration m: $move(j,i)$ is forbidden if and only if $t(j,i) + iter > m$.

An *aspiration criteria* is however employed to cancel the tabu status of $move(j,i)$ when the partition s' has a strictly better $H(s')$ value than $H(s^*)$, s^* being the best partition found so far.

The tabu search may lead to a state where no move is admissible. In that case, an *intensification phase* can be started based on a heuristic using long-term information. To implement this, the algorithm saves a set KER containing a few partitions s having $H(s)$ very near to $H(s^*)$. Then, starting with a partition s in KER we visit all the neighbors of s.

However, it is possible that KER corresponds to a set of partitions trapped in a local minimum. It is for this reason that the algorithm builds dynamically a *diversification set* DIV containing the indices having an appearing frequency lower than the average. Thus, DIV corresponds to less visited regions of S. During a diversification phase, the algorithm runs from initial configurations, whose coordinates are fixed by using the information in DIV. These exploratory mechanisms have been successfully employed in a VCSP (Value Constrained Satisfaction Problem) problem (see [12, 11])

4 Experimentation and results

Experiments are carried out on a set of 10 simulated instances and 10 realistic ones.

- *Simulated instances:* The simulated instances are samples of size 1500 coming from spaces of dimension 6, ... , 15. For each dimension, the samples are a mixture of a fixed number of normally distributed data. This knowledge on data distribution helps us to better appreciate the quality of the partition procedure. Remark that we limit ourselves to dimension 15. The reason for this is that the ellipsoids are more and more flattened when the dimension grows. Thus, our criterion on ellipsoids' volume is useless for large dimensions. Fortunately, the dimension of the realistic cases we are working on never exceeds 15.

- *Realistic instances:* Five of these instances come from samples gathered by a set of ozone sensors covering the Lyon region (France). These data are of dimension 8. Another set of data come from an oil refinery. The dimension of these data goes from 6 to 10. Their size goes from about 400 to 2500.

For all these instances we will give the size of the instance, the number of iterations, the running time, the best value of H, the first time that value was attained, the mean value of H. For these instances we have calculated in the same time the values of $K(\mathcal{A}, \mathcal{T})$ (1). We then give the correlations between the two criteria H and K.

References

[1] P.C. Chu and J.E. Beasley. *A Genetic Algorithm for the Multidimensional Knapsack Problem*. Journal of Heuristics, 4, 63-86, 1998.

[2] R. Cranley and T.N.L. Patterson. *Randomization of number theoretic methods for multiple integration*. SIAM J. Numer. Anal., 13, 904-914, 1976.

[3] F. Dammeyer and S. Voss. *Dynamic tabu list management using reverse elimination method*. Annals of Operations Research, 41, 31-46, 1993.

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

[5] F. Glover and G.A. Kochenberger. *Critical event tabu search for multidimensional knapsack problem*. In Meta-Heuristics: Theory and Applications, I.H.Osman and L.P.Kelly (eds), Kluwer Academic Publishers, 407-428, 1996.

[6] P. J. Davis and P. Rabinowitz. *Methods of Numerical Integration*. Academic Press, 1984.

[7] S. Kullback. *Information Theory and Statistics*. Wiley, New York 1959.

[8] S. Kullback and R.A. Leibler. *On information and sufficiency*. Ann. Math. Stat., 22, 79-86, 1951.

[9] A. Kurzhanski and I. Vályi. *Ellipsoidal Calculus for Estimation and Control*. IIASA, Birkhaüser, 1997.

[10] V.N. Vapnik. *Statistical Learning Theory*. John Wiley, 1998.

[11] M. Vasquez and J.K. Hao. *A logic-constrained knapsack formulation and a tabu algorithm for the daily photograph scheduling of an earth observation satellite*. To appear in Journal of Computational Optimization and Applications, 20(2): November 2001.

[12] M. Vasquez. *Résolution en variables 0-1 de problèmes combinatoires de grande taille par la méthode tabou*. Thèse de doctorat, LGI2P, École des Mines d'Alès.