First International Conference
Modelling and Development of Intelligent Systems
Sibiu - Romania, 22-25 October, 2009

Guided Maximum Entropy Method

Milan Tuba

Abstract

The maximum entropy method (MEM) has been successfully applied in many different areas for solving under-determined systems. It favors uniform distribution and tends to make variables as equal as possible, satisfying constraints. The problem is that for the initial adjustment of some problem for the MEM applications variables have to be selected in such a way that the solution is feasible, but may be far from desirable. This paper presents an improvement of the MEM model by introduction of new variables, constraints and weight factors that shift solution from feasibility to optimality. This modified method exploits the property of the MEM that it can smoothly move from cases where constraints can be satisfied to cases where constraints become desirable goals that are satisfied as much as possible. A software system was developed which includes all the mentioned features.

1 Introduction

The maximum entropy method was recently used with great success in many different areas where under-determined systems are involved. It is most frequently used in chemistry ([1]), but also in many other very diverse areas: computer network design ([2]), character recognition ([3]), data analysis ([4]), image processing ([5], [6]), economy ([7]). Theoretical developments also continue ([8]).

The basic idea is to get a unique solution from the under-determined system by introducing the additional constraint that the entropy function should be maximized. The other methods that were used for solving under-determined systems use the same technique: they introduce additional, artificial constraints that make the number of constraints equal to the number of unknowns. The difference is that the maximum entropy method introduces the most natural additional constraint: one that does not introduce any new, arbitrary and unwarranted information. It uses only the information that is given and makes no assumptions about missing information.

Before going to the formal definition of the maximum entropy principle, it is interesting to mention that, besides very pragmatic uses, there are very extensive and still open philosophical discussions about the real meaning of this principle. The predecessor of the maximum entropy principle is the principle of insufficient reason (James Bernoulli: *Ars Conjectandi,* 1713). It states that in the absence of any information (knowledge), all outcomes should be considered equally possible. This principle was involved in the discussions about prior probabilities (probabilities of one event, state of the knowledge) and relative frequencies. Relative frequencies become predominant and some useful works from Laplace and Bayes were criticized. Shannon's works on information theory opened a new opportunity for revitalization of the principle of insufficient reason, this time as a more sophisticated maximum entropy principle which was introduced by Jaynes. Philosophical discussions about the real meaning of the maximum entropy method are interesting, but since method was successfully applied in many areas, for any new area the most important criterion is not how well the relation between the MEM and that area can be explained, but how useful are the results of the application of the method.

2 The MEM Formalism

General model of the MEM calls for random variables and probabilities, but for most problems more suitable description is:

A system of k equations with n variables v_i, $k < n$ represents constraints. Maximum entropy solution which respects constraints and makes variables as equal as possible is looked for. It is interesting to notice that the same goal can be attained by using some other function that has maximum when all variables are equal. One very simple such function is the product of all variables. The product function expression seems simpler then the entropy function expression which involves logarithms, but the fact that partial derivatives are needed points out that entropy function is better since it separates variables.

$$\begin{aligned} x_{1,1}v_1 + x_{1,2}v_2 + ... + x_{1,n}v_n &= l_1 \\ x_{2,1}v_1 + x_{2,2}v_2 + ... + x_{2,n}v_n &= l_2 \\ &\vdots \\ x_{k,1}v_1 + x_{k,2}v_2 + ... + x_{k,n}v_n &= l_k \end{aligned} \tag{1}$$

Variables v_i are converted to probabilities by normalization: $p_i = v_i / \sum_{j=1}^{n} v_j$ and $m_i = l_i / \sum_{j=1}^{n} v_j$. The system (1) then becomes

$$\sum_{i=1}^{n} x_{r,i} p_i = m_r \ , \qquad r = 1, 2, ..., k \tag{2}$$

This is equivalent to the classical definition of the MEM where it is assumed that for a discrete random variable X the values $x_1,\ x_2, ...,\ x_n$ that it can take are known, but the corresponding probabilities $p_1,\ p_2, ...,\ p_n$ are not known. The expected values for $k < n-1$ functions of X (for example, the first k moments) are also known and represent constraints:

$$E[\ f_r(X)] = m_r \qquad r = 1, 2, ..., k. \tag{3}$$

Equation (2) (or (3)) gives (together with $\sum p_i = 1$) $k+1 < n$ constraints for n unknown variables $p_1,\ p_2,\ ...,\ p_n$. This system is under-determined and has an infinite number of solutions. The unique solution that maximizes the entropy of the system is looked for. That is the best solution in the sense that it uses only the information given. It is neutral to the missing information (it does not introduce any hidden assumptions). This additional constraint can be expressed as:

Maximize the entropy function

$$H(p_1,\ p_2,\ ...,\ p_n) = -K \sum_{i=1}^{n} p_i \ ln(p_i). \tag{4}$$

For $K = 1$, entropy will be expressed in natural units (rather than in bits).

2.1 Solution

The method of Lagrange multipliers is used. This will not guarantee that probabilities are non-negative. The substitution $p_i = e^{-q_i}$ is introduced, but this gives a stronger constraint than the one required: all probabilities are now positive definite (none of them can be zero). The problem now is to maximize

$$H(q_1,\ q_2, ...,\ q_n) \ = \ \sum_{i=1}^{n} q_i\ e^{-q_i} \tag{5}$$

under the conditions

$$\sum_{i=1}^{n} e^{-q_i} \ = \ 1 \tag{6}$$

$$\sum_{i=1}^{n} e^{-q_i} f_r(x_i) \ = \ m_r\ , \qquad r = 1, 2, ..., k \tag{7}$$

Lagrange multipliers $\lambda,\ \mu_1,\ \mu_2, ...,\ \mu_k$ are introduced with the function:

$$F(q_1,\ q_2, ...,\ q_n) \ = \ \sum_{i=1}^{n} q_i\ e^{-q_i} \ + \ \lambda \sum_{i=1}^{n} e^{-q_i} \ + \ \sum_{r=1}^{k} \mu_r \sum_{i=1}^{n} e^{-q_i}\ f_r(x_i) \tag{8}$$

All partial derivatives should be zero:

$$\frac{\delta F}{\delta q_i} \ = \ e^{-q_i}\ [1 - q_i - \lambda - \sum_{r=1}^{k} \mu_r\ f_r(x_i)] \ = \ 0\ , \qquad i = 1, 2, ..., n \tag{9}$$

Since e^{-q_i} is never zero

$$q_i \ = \ 1 - \lambda - \sum_{r=1}^{k} \mu_r\ f_r(x_i)\ , \qquad i = 1, 2, ..., n \tag{10}$$

The problem is now solved: Equations (6), (7), and (10) give $n+k+1$ equations for $n+k+1$ unknown variables $p_1,\ p_2, ...,\ p_n,\ \mu_1,\ \mu_2, ...,\ \mu_k,\ \lambda$. The system should have unique solution, but it is not linear and some numerical method has to be used.

To make the calculations easier, the partition function is introduced:

$$Z(\mu_1,\ \mu_2, ...,\ \mu_k) \ = \ \sum_{i=1}^{n} p_i\ e^{-\lambda} \ = \ \sum_{i=1}^{n} e^{-\lambda - q_i}$$

$$Z(\mu_1,\ \mu_2, ...,\ \mu_k) \ = \ \frac{1}{e} \sum_{i=1}^{n} e^{\sum_{r=1}^{k}\ \mu_r\ f_r(x_i)} \tag{11}$$

It is easy to see that

$$\lambda \ = \ - \ln Z(\mu_1,\ \mu_2, ...,\ \mu_k) \tag{12}$$

$$m_r \ = \ \frac{\delta}{\delta \mu_r} \ln Z(\mu_1,\ \mu_2, ...,\ \mu_k) \tag{13}$$

or

$$m_r \ = \ \sum_{i=1}^{n} [m_r - f_r(x_i)] e^{\sum_{j=1}^{k}\ \mu_j\ f_j(x_i)} = 0\ , \qquad r = 1, 2, ..., k \tag{14}$$

Equation (14) represents k equations for k unknown variables $\mu_1,\ \mu_2, ...,\ \mu_k$. When it is solved, from Equation (12) λ is calculated, and then from Equation (10) $q_1,\ q_2, ...,\ q_n$ are determined , and finally, from $p_i = e^{-q_i}$ the probabilities $p_1,\ p_2, ...,\ p_n$ are calculated.

Substitution $t_j = e^{\mu_j},\ \ j = 1, 2, ...,\ k$ can be introduced. Then Equations (12) and (14) become:

$$\lambda \ = \ 1 - ln[\sum_{i=1}^{n} \Pi_{j=1}^{k} t_j^{f_j(x_i)}] \tag{15}$$

$$\sum_{i=1}^{n} [m_r - f_r(x_i)] \Pi_{j=1}^{k} t_j^{f_j(x_i)} \ = \ 0, \qquad r = 1, 2, ..., k \tag{16}$$

There is an algorithm to solve this system. However, the function that is to be minimized is not convex even in the simplest case when there is only one constraint: expected value. The standard Newton-Rapson procedure will not work. But the Jacobian matrix for this system is symmetric and positive definite. This gives a scalar potential function which is strictly convex and whose minimum is easy to find. The use of the second order Taylor expansion is recommended. However, after much experience with the algorithm, our impression is that it is not even worth trying to find the exact value for α that determines how far to go along a certain direction, let alone inverting the Jacobian matrix every time. For our software system we developed a heuristic that performs well.

2.2 Selection Principle

The previous model has constraints $p_i \ > \ 0,\ i = 1, 2, ..., n$. This may be too strong since the probabilities need only to be nonnegative. To make $p_i \ \geq \ 0$, $p_i \ = \ q_i^2$ can be introduced instead of $p_i \ = \ e^{-q_i}$, which was used before. In that case, the problem becomes to maximize

$$H(q_1,\ q_2, ...,\ q_n) \ = \ -2 \sum_{i=1}^{n} q_i^2 ln(q_i) \tag{17}$$

under the conditions

$$\sum_{i=1}^{n} q_i^2 \ = \ 1 \tag{18}$$

$$\sum_{i=1}^{n} q_i^2 \ f_r(x_i) \ = \ m_r, \qquad r = 1, 2, ..., k \tag{19}$$

Lagrange multipliers are introduced:

$$F(q_1,\ q_2, ...,\ q_n) \ = \ -2 \sum_{i=1}^{n} q_i^2 \ ln(q_i) \ + \ \lambda \sum_{i=1}^{n} q_i^2 \ + \ \sum_{r=1}^{k} \mu_r \sum_{i=1}^{n} q_i^2 \ f_r(x_i) \tag{20}$$

Partial derivatives should be zero:

$$\frac{\delta F}{\delta q_i} \ = \ -2q_i[2ln(q_i) + 1 - \lambda - \sum_{r=1}^{k} \mu_r f_r(x_i)] \ = \ 0, \qquad i = 1, 2, ..., n \tag{21}$$

Now, the selection has to be made: any q_i can be zero.

$$q_i = 0 \quad or \quad q_i \ = \ e^{(\ -1+\lambda+\sum_{r=1}^{k} \ \mu_r \ f_r(x_i)\ )^{0.5}}, \qquad i = 1, 2, ..., n \tag{22}$$

When it is decided which q_i are to be zero, the remaining equations will give as many equations as there are unknown variables. The partition function is equal as in the previous model, and the whole discussion repeats. The only difference is that summations are not carried for all $i = 1$ to n, but only for those i for which $q_i \neq 0$.

This new model is used only to show how the case p_i=0 for some i can be included. In practice, we have to decide which p_i will be zero. We can do it in advance and consider a model that has only $n-m$ probabilities (if m probabilities are selected to be zero). If we select too many probabilities to be zero, the system may become over-determined.

3 The Guided MEM

For many problems initial adjustment for the MEM application requires that variables of the system be determined in such a way that a feasible solution is obtained. This may not be a desirable solution for the optimization, but constraints have to be satisfied first.

It is possible to modify the MEM model and include a mechanism to guide the process of optimization. Once the necessary constrains are satisfied, artificial variables can be introduced that will guide the optimization process in the desirable direction.

MEM guidance will be demonstrated on an example, similar to Brandeis Dice Problem.

A die, possibly irregular, is considered. The number of spots that shows up when the die is tossed defines a random variable with possible outcomes and corresponding probabilities.

$$X = [1, 2, 3, 4, 5, 6]$$

$$P_{(6)} = [p_1, p_2, p_3, p_4, p_5, p_6]$$

The constraint that the sum of the probabilities is 1 is always present and in usual terminology not counted as an additional constraint. Without any (additional) constraints the expected value $E(X)$ is 3.5 and the solution for the probabilities is an uniform distribution: $p_i = 0.167, \quad i = 1, 2, ...6$.

For a single constraint EX=4.4 there is one (additional) constraint:

$$1p_1 + 2p_2 + 3p_3 + 4p_4 + 5p_5 + 6p_6 = 4.4$$

and the MEM solution is:

$$P_{(6)} = [0.063, 0.087, 0.121, 0.169, 0.234, 0.325]$$

As expected, the probabilities density is shifted towards larger outcomes since expected value shifted in that direction.

If the elementary probabilities were not the goal of equalization but some coarser variables, additional constraint can be included. If, for example, the goal is to to make $p_x = p_1+p_2+p_3$ equal to $p_y = p_4+p_5+p_6$, a system of two constraints can be used:

$$1p_1 + 2p_2 + 3p_3 + 4p_4 + 5p_5 + 6p_6 = 4.4$$

$$1p_1 + 1p_2 + 1p_3 - 1p_4 - 1p_5 - 1p_6 = 0$$

In this case it is possible to have a solution that will satisfy both constraints:

$$P_{(6)} = [0.004, 0.042, 0.454, 0.004, 0.042, 0.454] \tag{23}$$

The problem with this approach is that it limited to cases when the guidance goal (in this case the total equalization of p_x and p_y) is possible. However, the main advantage of the MEM method is its ability to push towards the guidance goal even when exact goal satisfaction is not possible.

This can be illustrated on the previous example, but with changed requirement that $E(X) = 4.6$. It is easy to see that the constraint

$$p_1 + p_2 + p_3 = p_4 + p_5 + p_6$$

can not be satisfied. The maximum value for $E(X)$ is reached when probabilities density is pushed toward higher values:

$$P_{(6)} = [0, 0, 0.5, 0, 0, 0.5]$$

The value for $E(X)$ is in that case equal to 4.5. For any higher value of $E(X)$ exact equalization (which is the second constraint) is not possible.

To make the sums $p_1 + p_2 + p_3$ and $p_4 + p_5 + p_6$ as equal as possible, new variables are introduced: $p_6 = p_x = p_1 + p_2 + p_3$ and $p_7 = p_y = p_4 + p_5 + p_6$. Two new constraints that define these new probabilities are added. The fact that new variables are mentioned as constraints will make them participate in the equalization process.

Care mast be taken about normalization. New probabilities (p_7 and p_8) are not independent from the old ones and the sum of all probabilities becomes 2. Considering that the sum of all probabilities has to be 1 and that the sum of old probabilities (only old probabilities participate in the first constraint) is only 0.5, the first constraint has to be redefined.

Three constraints now become:

$$1p_1 + 2p_2 + 3p_3 + 4p_4 + 5p_5 + 6p_6 + 0p_7 + 0p_8 = 2.3$$

$$1p_1 + 1p_2 + 1p_3 + 0p_4 + 0p_5 + 0p_6 - 1p_7 + 0p_8 = 0$$

$$0p_1 + 0p_2 + 0p_3 + 1p_4 + 1p_5 + 1p_6 + 0p_7 - 1p_8 = 0$$

and the corresponding MEM solution is:

$$P_{(8)} = [0.020, 0.039, 0.076, 0.055, 0.106, 0.205, 0.135, 0.365]$$

or, when only $P_{(6)}$ is denormalized:

$$P_{(6)} = [0.040, 0.078, 0.152, 0.109, 0.211, 0.409]$$

This solution represents smooth extrapolation of the previous case. All constraints are satisfied. Expected value is 4.6. However, p_7 and p_8 are not equal since that was not the requirement any more. These variables were mentioned in the system of constraints so they participate in the process of equalization, but only to some extent. In this case (after denormalization), $p_7 = 0.270$ and $p_8 = 0.730$. This is far from being equal, the ratio p_8/p_7 is 2.7. We can make them closer to being equal by forcing them to contribute more significantly in the optimization process. This can be accomplished by redefining them in such a way that the larger mass of the probability is concentrated in them. If the constraints $p_6 = p_1 + p_2 + p_3$ and $p_7 = p_4 + p_5 + p_6$ are replaced with $p_6 = 9p_1 + 9p_2 + 9p_3$ and $p_7 = 9p_4 + 9p_5 + 9p_6$ only the 10% of the probability mass will remain in the old probabilities and 90% will be concentrated in the new probabilities. This will make new probabilities more significant in the equalization process, but the first constraint has to be redefined to reflect the fact that old probabilities, that define it, now contribute 10 times less. The new set of constraint is:

$$1p_1 + 2p_2 + 3p_3 + 4p_4 + 5p_5 + 6p_6 + 0p_7 + 0p_8 = 0.46$$

$$9p_1 + 9p_2 + 9p_3 + 0p_4 + 0p_5 + 0p_6 - 1p_7 + 0p_8 = 0$$

$$0p_1 + 0p_2 + 0p_3 + 9p_4 + 9p_5 + 9p_6 + 0p_7 - 1p_8 = 0$$

The corresponding MEM solution is:

$$P_{(8)} = [0.001, 0.007, 0.031, 0.002, 0.010, 0.049, 0.348, 0.552]$$

or, when only $P_{(6)}$ is denormalized:

$$P_{(6)} = [0.014, 0.066, 0.307, 0.022, 0.104, 0.487]$$

New probabilities p_7 and p_8 are now closer to being equal since ratio p_8/p_7 is 1.6.

We can push this process further in that direction by making old probabilities contain only 2% of the probability mass, which is equivalent of making new probabiities 50 times more important.

The new set of constraints is now:

$$1p_1 + 2p_2 + 3p_3 + 4p_4 + 5p_5 + 6p_6 + 0p_7 + 0p_8 = 0.092$$

$$49p_1 + 49p_2 + 49p_3 + 0p_4 + 0p_5 + 0p_6 - 1p_7 + 0p_8 = 0$$

$$0p_1 + 0p_2 + 0p_3 + 49p_4 + 49p_5 + 49p_6 + 0p_7 - 1p_8 = 0$$

The corresponding MEM solution is:

$$P_{(8)} = [0.000, 0.000, 0.009, 0.000, 0.000, 0.010, 0.444, 0.536]$$

or, when only $P_{(6)}$ is denormalized:

$$P_{(6)} = [0.001, 0.019, 0.434, 0.001, 0.023, 0.523]$$

New probabilities p_7 and p_8 are now even closer to being equal since ratio p_8/p_7 improved to 1.2.

For significance of new probabilities equal to 100, the corresponding probabilities are $P_{(8)} = [0.000000, 0.000038, 0.004603, 0.000000, 0.000044, 0.005314, 0.459509, 0.530491]$, $P_{(6)} = [0.0000, 0.0038, 0.4603, 0.0000, 0.0044, 0.5314]$ and ratio $p_8/p_7 = 1.15$.

The process that is described shows that it is possible to adjust MEM for some constrained optimization problem and then guide it in the desired direction, but there is no universal way how to do it, each problem has to be investigated separately.

4 Conclusion

The maximum entropy method can successfully be used for optimization with constraints that are represented by under-determined systems. A software system is developed that includes standard MEM solution with some improvements which include a heuristics for speeding up the calculations. For each particular application a problem has to be transferred into the form usable for the MEM. This often leads to MEM solution that is only feasible. Introduction of artificial variables and appropriate coefficients allows to guide optimization process in the desired direction. The system was tested on the computer network design problem where good quality initial topology and routing were obtained. This software system represents a tool that is universal for all MEM applications, each particular problem, however, requires very careful adjustments and that part can not be automatized.

Acknowledgment: This paper is founded from the research Project No. 144007, Ministry of Science, Republic of Serbia.

References

[1] Ding YS, Zhang TL, Gu Q, et al., Using Maximum Entropy Model to Predict Protein Secondary Structure with Single Sequence *Protein and Peptide Letters* Volume 16, Issue 5, pp. 552-560, 2009

[2] Tuba, Milan: Maximum Entropy Method and Underdetermined Systems Applied to Computer Network Topology and Routing, Plenary Lecture, 9th WSEAS International Conference on APPLIED INFORMATICS and COMMUNICATIONS (AIC '09) Moscow, Russia, August 20 - 22, 2009, chapter in Recent Advances in Applied Informatics and Communications, A Series of Reference Books and Textbooks, WSEAS Press 2009, pp. 18 and 127-132.

[3] Xuan Wang, Lu Li, Lin Yao, Anwar, W., A Maximum Entropy Approach to Chinese Pin Yin-To-Character Conversion. *2006 IEEE International Conference on Systems, Man, and Cybernetics*, 2006, Taipei, Taiwan

[4] Teh, Chee Siong Lim, Chee Peng, A probabilistic SOM-KMER model for intelligent data analysis, *WSEAS Transactions on Systems* Vol. 5, no. 4, pp. 825-832. Apr. 2006

[5] Zhengmao Ye, Habib Mohamadian1, Yongmao Ye, Practical Approaches on Enhancement and Segmentation of Trimulus Color Image with Information Theory Based Quantitative Measuring, *WSEAS Transactions on Signal Processing*, Issue 1, Volume 4, January 2008, pp. 12-20

[6] Heric, Dusan; Zazula, Damjan, Reconstruction of Object Contours Using Directional Wavelet Transform, *WSEAS Transactions on Computers*. Vol. 4, no. 10, pp. 1305-1312. Oct. 2005

[7] Ciavolino E, Dahlgaard JJ, Simultaneous Equation Model based on the generalized maximum entropy for studying the effect of management factors on enterprise performance, *Journal of Applied Statistics* Volume 36, Issue 7, pp. 801-815, 2009

[8] Aladdin Shamilov, Generalized entropy optimization problems and the existence of their solutions, *Physica A: Statistical Mechanics and its Applications*, Volume 382, Issue 2, August 2007, pp. 465-472

[9] Tuba, Milan: Cost Function for Communication Links in Computer Networks, *Bulletins for Applied Mathematics (BAM)*, LXXIII-1028/94, pp. 115-122, Budapest, 1994

[10] Tuba, Milan: A Mathematical Model for Routing Comparison in Computer Networks, *Bulletins for Applied Mathematics (BAM)*, LXXXVI-A 1565/98, Arad, July 1998, pp. 493-503

[11] Tuba, Milan: Parameters for the Internet Optimization on the Local Level, *Applied & Computing Mathematics*, Vol. II, pp. 139-142, Kosice, 1997

[12] Tuba, Milan: Relation between Static and Dynamic Optimization in Computer Network Routing, *Recent Advances in Artificial Intelligence, Knowledge Engineering and Data Bases*, WSEAS Press 2009, pp. 484-489

[13] Tuba, Milan: Computer Network Routing Based on Imprecise Routing Tables, *WSEAS Transactions on Communications*, Issue 4, Volume 8, April 2009, pp. 384-393

[14] Abd-El-Barr M: Topological network design: A survey, *Journal of Network and Computer Applications*, Volume 32, Issue 3, pp. 501-509, 2009

[15] Karavetsios, P., Economides, A.: Performance Comparison of Distributed Routing Algorithms in Ad Hoc Mobile Networks, *WSEAS Transactions on Communications*, Vol. 3, Issue 1, 2004, pp. 317-321

[16] Sokullu, R., Karaca, O.: Comparative Performance Study of ADMR and ODMPR in the Context of Mobile Ad Hoc Networks and Wireless Sensor Networks, *International Journal of Communications*, Issue 1, Volume 2, 2008, pp. 45-53

[17] Kumar, D., Bhuvaneswaran, R.: ALRP: Scalability Study of Ant Based Local Repair Routing Protocol for Mobile Ad Hoc Networks, *WSEAS Transactions on Computer Research*, Vol. 3, Issue 4, Apr 2008, pp. 224-233

Milan Tuba
Megatrend University Belgrade
Faculty of Computer Science
Bulevar umetnosti 29
SERBIA
E-mail: *tubamilan@ptt.rs*