Optoelectronics 光电子, 2011, 1, 33-38
http://dx.doi.org/10.12677/oe.2011.12007 Published Online December 2011 (http://www.hanspub.org/journal/oe)

The Research of Handwritten Digit Cluster Based on Artificial Bees Colony Algorithm*

Guangbiao Wang[1], Shuying Yang[1], Fan Feng[1], Bokai Wang[1], Zijuan Jia[1], Guang Zhu[2]
[1]School of Computer and Communications Engineering, Tian Jin University of Technology, Tianjin
[2]Radio Film and TV Administration of Shan Dong Province, Shandong
Email: guangbiaowang@126.com
Received: Sep. 23rd, 2011; revised: Oct. 11th, 2011; accepted: Nov. 3rd, 2011.

Abstract: Handwritten digit cluster is an important study of pattern recognition, because of the application of traditional evolutionary algorithm for clustering analysis of handwritten digit has much more problems, such as slow convergence, easily fall into local optimization and so on. To overcome those problems, we present a novel approach to solve the problem of digital clustering by using artificial bees colony algorithm, and propose 3 kinds of operators for bees' location of updating, and establish a dynamic update of 3 operators in formulas. Finally we elaborate the concrete steps of handwriting digital cluster by using this approach. We do the simulation experiments with some typical handwritten digital instances. Experiments show that our approach can make a good implementation of handwriting digital cluster, overcome the phenomenon of premature convergence, and accelerate the convergence rate in a way.

Keywords: Artificial Bees Colony Algorithm; Handwritten Digit Cluster; Combinatorial Optimization

基于人工蜂群算法的手写数字聚类研究*

王光彪[1]，杨淑莹[1]，冯　帆[1]，王博凯[1]，贾紫娟[1]，朱　光[2]
[1]天津理工大学，计算机与通信工程学院，天津
[2]山东省广播电影电视局，山东
Email: guangbiaowang@126.com
收稿日期：2011 年 9 月 23 日；修回日期：2011 年 10 月 11 日；录用日期：2011 年 11 月 3 日

摘　要：手写数字聚类是模式识别研究中的一个重要研究方向，但应用传统的进化算法对手写数字进行聚类分析往往存在着收敛速度慢，易陷入局部最优等问题，本文提出了用蜂群算法求解数字聚类问题，并且提出了 3 种蜜蜂的位置更新算子，建立了 3 种算子的动态更新公式，最后阐述了利用该算法对手写数字聚类的具体步骤。通过典型的手写数字实例进行了仿真实验，实验表明：该算法能够很好的实现手写数字聚类，并且克服了过早收敛的现象，而且能够加快收敛速度。

关键词：人工蜂群算法；手写数字聚类；组合优化

1. 引言

近年来，应用群体智能来解决组合优化问题的研究成为了人们关注的焦点，多个简单个体组成的群体，具有通过相互之间的简单协作完成问题求解的能力就被称为群体智能[1]。一些学者根据昆虫群体进行研究，提出了很多用于组合优化问题的求解算法理论，如蜂群算法[2]，蚁群算法[3]等,手写数字聚类是组合优化的一个具体应用，即在错误概率最小的情况下，将特征相同或者相近的数字归为一类，实现聚类划分。

蜂群算法是建立蜜蜂的自组织模型和群体智能基础上的一种非数值优化计算方法。Seeley[4]于 1995 年最先提出了蜂群的自组织模拟模型，在该模型中，虽然各社会阶层的蜜蜂只完成了一种任务，但是蜜蜂以“摆尾舞”、气味等多种方式在群中进行信息的交

*基金项目：国家 863 计划项目(2007AA01Z188)，国家自然基金项目(60773073)，天津市高等学校科技发展基金(20071308)。

流，使得整个群体可以完成诸如喂养、采蜜、筑巢等多种工作。Karaboga[5]于 2005 年将蜂群算法成功的应用于函数的数值优化问题，系统的提出了人工蜂群算法(ABC)(Artificial Bee Algorithm)，2006 年在文献[6]提出了蜜蜂算法，这些算法均有其特色，主要应用于多峰值函数寻优等优化问题，作者成功的将蜂群算法应用于手写数字聚类的求解问题。

2. 蜂群算法基本原理

蜜蜂是一种群居昆虫，单个蜜蜂的行为极其简单，但是由单个简单的个体组成的群体却表现出极其复杂的行为。现实生活中的蜜蜂能够在任何环境下，以极高的效率获得食物；同时它们能够适应环境的改变。

蜂群算法中，主要包括这四个基本元素：引领蜂、跟随蜂、侦查蜂、蜜源，我们将算法的寻优过程转变为蜜蜂寻找蜜源的过程，某一蜜蜂及其所对应的蜜源都是值问题的一个可行解。蜜源的丰富程度(即蜜源的收益度)即代表着可行解的质量。蜂群算法开始时，蜜蜂都是以侦查蜂的身份在蜂巢附近寻找蜜源，即初始化产生 N 个可行解；然后根据蜜源的收益度将收益度排名前 50%的解作为蜜源位置，即前 50%为引领蜂，后 50%为跟随蜂。蜜源的个数不会随着迭代过程的进行而改变。之后根据贪婪原则，引领蜂先对对应的蜜源做一次邻域搜索，如果搜索到的新蜜源的收益度比原来蜜源好，则忘记原来蜜源，开采新的蜜源；否则，继续开采原来蜜源。所有的引领蜂完成搜索后，在舞蹈区将蜜源的信息与跟随蜂共享。跟随蜂依照概率公式来选择去哪个蜜源采蜜。蜜源收益度越高则被选择的概率越大。之后跟随蜂根据选择的蜜源，对记忆蜜源做一次邻域搜索，同引领蜂一样，比较新搜索到的蜜源的收益度，如果新蜜源收益度比原来蜜源好，则开采新的蜜源，否则，继续开采原来的蜜源。当同一个蜜源被开采的次数超过了限定的 limit 次之后，此时采集该蜜源的蜜蜂变成侦查蜂，并且由侦查蜂随机在解空间中产生一个新的蜜源来代替原来的蜜源。

蜂群算法与一般所求问题的对应关系如表 1 所示。

表 1. 蜂群算法与最优问题的对应关系
Table 1. The correspondence of ABC and optimal problem

蜂群采蜜行为	具体问题
蜜源的位置	具体问题的可行解
蜜源的大小收益度	可行解的质量
寻找和采集蜜源的速度	求解的速度
最大收益度的蜜源	问题最优解

3. 蜜蜂更新算子

3.1. 引领蜂算子

在侦查蜂搜索到蜜源，在蜂巢内卸下蜂蜜之后，根据蜜蜂所搜索到的蜜源的收益度，根据算法之前设定的引领蜂的比例来确定将搜索到高收益度蜜源的那部分蜜蜂变成引领蜂，收益度低的部分变成跟随蜂。蜂群算法中，引领蜂与其所采的蜜源相对应，引领蜂能够记忆自己所采蜜源的相关信息，并且将这些信息在蜂巢的舞蹈区中与跟随蜂分享，从而招募更多的蜜蜂去采蜜。

蜂群算法中引领蜂的位置更新主要是通过引领蜂算子实现的，类似于遗传算法中的染色体的变异操作，就是在蜂群范围内随机搜索，通过改变个体上的部分或者全部基因来产生新的个体，其位置更新如公式(1)所示。

$$X_i(t+1) = X_i(t) + \mathrm{r} \times \left(X_i(t) - X_k(t) \right) \tag{1}$$

式中，$X_i(t+1)$ 代表了新产生的第 i 个蜜源位置；$X_i(t)$ 代表了原来的第 i 个蜜源位置；r 为在[-1,1]范围内的随机数；$\mathrm{k} \in \{1,2,\cdots,\mathrm{N}\}$，是随机产生的下标 $\mathrm{k} \neq i$，其中 M 代表个体的基因长度；由公式(1)可看出,引领蜂在其可视范围内随机地向某一个蜜蜂前进，这个蜜蜂可能是引领蜂也可能是跟随蜂，随着参数 X_i 与参数 X_k 之差的变小，对位置 X_i 的扰动也越来越小。因此,随着对最优解的逼近，步长会自适应的缩减。

3.2. 跟随蜂算子

引领蜂在蜂巢内的舞蹈区，将有关蜜源的信息(即蜜源的收益度)告诉给跟随蜂，然后跟随蜂就会根据蜜源的信息来选择一个蜜源跟随着引领蜂去采蜜。其选择的标准也是根据蜜源的收益度来判别的，跟随蜂更倾向于到收益度高的蜜源那里采蜜。一般在算法中都是根据采用轮盘赌的选择方法决定跟随蜂到哪个蜜源去采蜜，每个蜜源被选择的概率是根据其收益度来决定的，即蜜源被选择的概率是该蜜源的收益度在整个

备选蜜源总收益度中所占的比例，如公式(2)所示。蜜源的收益度越高，被跟随蜂选择的概率越大。

$$P_i = \frac{\text{fitness}_i}{\sum_{j=1}^{N} \text{fitness}_j} \tag{2}$$

式中，P_i 代表第 i 个蜜源被选择的概率；fitness_i 为基于第 i 个蜜源的收益度；N 为蜜源的总个数。

在确定到哪个蜜源采蜜之后，跟随蜂将会在此蜜源的附近搜寻一个新的蜜源。设第 i 个引领蜂对应的蜜源被选择，$X_i(t)$ 表示第 i 个引领蜂。跟随蜂 $G_k(t)$ 表示第 k 个跟随蜂，r 为在[–1,1]范围内的随机数。跟随蜂是根据引领蜂传递过来的信息，在此基础上产生一个变化的位置，代表在引领蜂附近采蜜，计算公式为(3)所示。

$$\mathrm{G}_K(t+1) = X_i(t) + r \tag{3}$$

计算跟随蜂搜索到的新蜜源的收益度，并比较新蜜源是否好于原蜜源，如果新蜜源的收益度比原蜜源的收益度高，则忘掉旧的蜜源开采新的蜜源；否则，继续开采原来的蜜源。

3.3. 侦查蜂算子

随着采蜜工作的进行，蜜源的丰富程度将会降低，很有可能出现蜜源枯竭的现象，即蜜源的收益度过低的情况，这样的情况下使得解的质量变坏。为了防止这种现象的发生，需要对蜜源的采集次数进行限制，算法中规定蜜源的最大采集次数为 limit。

当同一蜜源被采集了 limit 次之后，其蜜源的丰富程度必然降到很低了，则放弃这个蜜源，此时蜜蜂的身份变为侦查蜂，并由侦查蜂在整个解空间中随机的产生一个新的蜜源。这样增加了解的多样性，有利于求解到全局最优解。

侦查蜂随机产生新解的公式如(4)所示。

$$X_{\mathrm{i}}(t+1) = r \tag{4}$$

式中，$X_{\mathrm{i}}(t+1)$ 代表了新产生的第 i 个蜜源位置；r 为在[–1,1]范围内的随机数。

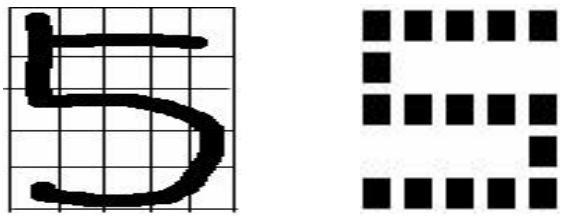
4. 蜂群算法手写数字聚类

4.1. 手写数字特征提取

要实现手写数字有效的的聚类划分，首先我们必须对相应的手写数字进行特征提取。在本文中，为每个数字图形定义一个 $N \times N$ 的模板，将样品的长度和宽度进行 N 等分，然后将每一份内的黑像素个数在本份中所占的比例作为样品的特征初值。如图 1(a)所示，对样品的长度和宽度 5 等分，构成一个 5×5 均匀区域，在每一区域中统计黑像素个数，然后利用黑像素的个数除以该小区域的面积总数，即得特征值，如图 1(b)所示，我们可以根据需要进行修改 N 值，N 值越大，特征越多，区分物体的能力也就越强，但同时会相应的增加计算量，延长运行等候的时间，所需要的样本库也成倍增加，样本库的个数一般为特征数的 5～10 倍，这里特征总数为 5 × 5 = 25，每一种数字就至少需要 75 个标准样本，10 个数字则需要 750 个标准样本。如果 N 值过小，则不容易进行不同物体的区分。

这种特征提取方法在一定程度上解决了同一类样品在不同大小情况下的区分效果。对于定义的 N × N 的模板，要求物体的长宽至少大于 N 个像素，否则无法正确分类。

一幅图像中包含多个手写数字，在对数字进行聚类分析时需要对不同的数字作分割标识，如图 2 所示，通过画图工具手写了(7、8、4、7、4、4、7、8、3)共 8 个待分类样品，需要通过计算机将它们自动的分成 4 类。



(a) Sample area
(a) 对样品分成 5 × 5 的区域

(b) Templates feature value
(b) 5 × 5 模板特征值示意图

Figure 1. 5 × 5 Template feature extraction method
图 1. 5 × 5 模板提取特征法

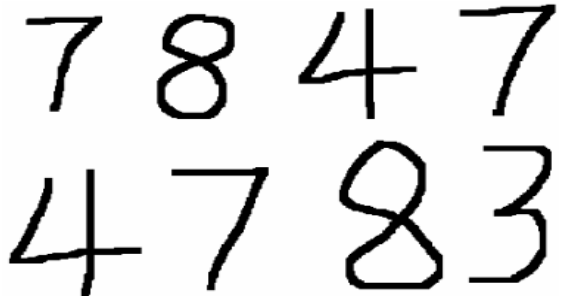


Figure 2. Sample number
图 2. 样品数字

4.2. 解的编码形式

对图 2 所示的 8 个物体进行编号，样品编号如图 3 所示，在每个样品的右上角，不同的样品编号不同，而且编号始终固定。

采用符号编码，位串长度 L 取 8 位，分类号代表样品所属的类号(1~4)，样品编号是固定的，也就是说某个样品在每个解中的位置是固定的，而每个样品所属的类别随时在变化。如果编号为 n，则其对应第 n 个样品，而第 n 个位所指向的值代表第 n 个样品的归属类号。

每个解包含一种分类方案。设初始解的编码为：(2，3，4，1，2，1，3，4)，这是处于假设分类情况，并不是最优解，其含义为：第 1 、5 个样品被分到第 2 类；第 2、7 个样品被分到第 3 类；第 3、8 个样品被分到第 4 类；第 4、6 个样品被分到第 1 类。如表 2 所示。

经过蜂群算法找到的最优解如图 3 所示。蜂群算法找到的最优解编码如表 3 所示。过比对可以发现相同的数据被归为一类，分到相同的类号，而且全部正确。

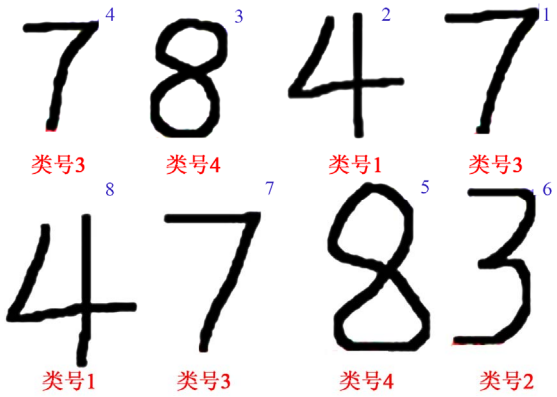


Figure 3. Number of the test sample
图 3 待测样品的编号

Table 2. Initial solution
表 2. 初始解

样品值	(7)	(4)	(8)	(7)	(8)	(3)	(7)	(4)
分类号	2	3	4	1	2	1	3	4
样品编号	1	2	3	4	5	6	7	8

Table 3. Optimal solution
表 3. 最优解

样品值	(7)	(4)	(8)	(7)	(8)	(3)	(7)	(4)
分类号	3	1	4	3	4	2	3	1
样品编号	1	2	3	4	5	6	7	8

4.3. 聚类准则函数

系统初始化了 N 个蜜蜂，根据收益度选取前 50% 的蜜蜂作为引领蜂，并且蜜源的个数=引领蜂的个数。蜜源与引领蜂相同，对应着所求问题的解。

设模式样品集为 $X=\{X_i,i=1,2,\cdots,n\}$ 其中 X_i 为 D 维模式向量，根据解的含义不同，通常可以分为两种方法。一种是以聚类结果为解，一种是以聚类中心集合为解。本文中我们采用的是基于聚类中心集合作为蜜蜂的对应解，也就是每个蜜蜂的位置是由 k 个聚类中心组成，样品向量维数为 D 聚类问题中，每个蜜蜂 i 由四部分组成，蜜蜂结构 i 表示为：

$$\begin{aligned}\mathrm{Bee}(i)=\{&\mathrm{location}[],\\&\mathrm{String},\\&\mathrm{oBas},\\&\mathrm{fitness}\}\end{aligned}\tag{5}$$

其中 location 代表蜜蜂的位置编码结构表示为：

$$\mathrm{Bee}(i)\cdot\mathrm{location}[\ ]=[C_1,C_2,.....,C_k]\tag{6}$$

其中，C_j 表示第 j 类的聚类中心，是一个 D 维矢量。

蜜源被开采度 oBas 表示蜜源被采集的次数，主要是为了防止蜜源枯竭，导致解的质量下降。

String 为一整数，表示了样品的分类号。

蜜源收益度值 fitness 为一实数，表示蜜源的收益度。可以采用以下方法计算其收益度。

按照最近邻法式(7)，确定该蜜蜂的聚类划分。

当聚类中心确定时，聚类的划分可由最近邻法则决定。即对样品 X_i，若第 j 类的聚类中心 C_j 满足式(7)，则 X_i 属于类 j。

$$\mathrm{d}(X_i,C_j)=\min_{l=1,2,\ldots,k}\mathrm{d}(X_i,C_l)\tag{7}$$

根据聚类划分，重新计算聚类中心，按照式(8)计算总的类内离散度 J_c。

$$J_c=\sum_{j=1}^{k}\sum_{X_i\in C_j}\mathrm{d}(X_i,C_j)\tag{8}$$

其中 C_j 为第 j 个聚类的中心，$d(X_i,C_j)$ 为样品到对应聚类中心距离，聚类准则函数 J_c 即为各类样品到对应聚类中心距离的总和。

蜜源的收益度可表示为式(9)。

$$\mathrm{Bee}\times\mathrm{fitness}=k/J_c\tag{9}$$

其中 J_c 是总的类内离散度和，k 为常数，根据具体情况而定。即蜜源所代表的聚类划分的总类间离散度越小，蜜源的收益度越大。

4.4. 蜂群算法实现聚类分析具体步骤

① 开始蜜蜂都是以侦查蜂的身份在蜂巢附近搜索蜜源，产生初始解，初始化各个参数如聚类书目、蜂群总数、蜜源被采集次数及同一蜜源的开采极限等。

② 根据收益度函数计算各个蜜源的收益度，并记录并选取收益度较高的 50%的蜜蜂作为引领蜂。

③ 开始进入迭代过程，判断同一蜜源被采集的次数是否大于开采极限，若是，则此时该蜜蜂的身份变为侦查蜂，并在解空间中随机产生一个解，并令新蜜源的开采度值零；否则，由引领蜂在蜂巢的舞蹈区招募跟随蜂到蜜源附近采蜜。

④ 计算新蜜源的收益度值，并判断是否好于原来蜜源，如果比原来蜜源好，则开采新的蜜源，并且其开采度置零，否则，继续开采原来蜜源，并且其开采度加 1。

⑤ 计算各个蜜源的收益度，记录并保留当前的最优解。

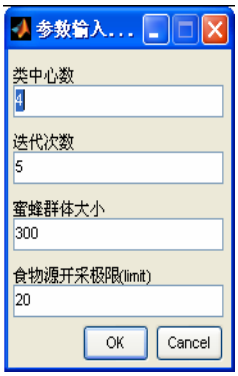
⑥ 判断是否满则终止条件，若没有满足，则转到步骤③继续执行，否则，输出最优解，结束程序。

5. 典型手写数字实例仿真及对比试验

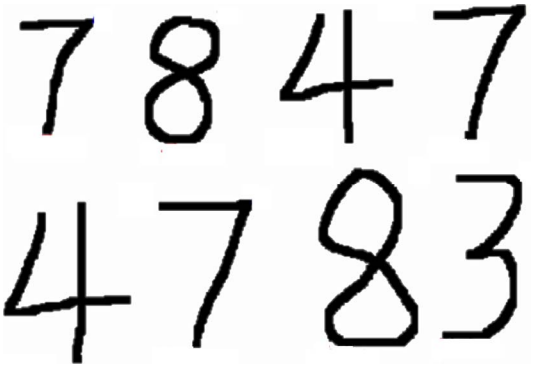
下面是一个基于手写数字聚类效果图，如图 4 所示。从结果上可以看出，蜂群算法能够很好的实现手写数字的聚类分析。

在 MATLAT 环境下，软件实现了基于蜂群算法的手写数字的聚类，在 Inter(R) Core(TM) 2 Duo T7200 处理器的计算机上进行了相应的聚类实验，如图 5 所示，本文利用 50 组待聚类样品通过蜂群算法与遗传算法和粒子群算法的比较表明采用基于蜂群算法的手写数字聚类不但能够很好的实现聚类分析，而且在收敛速度和算法性能上有较大的提高。

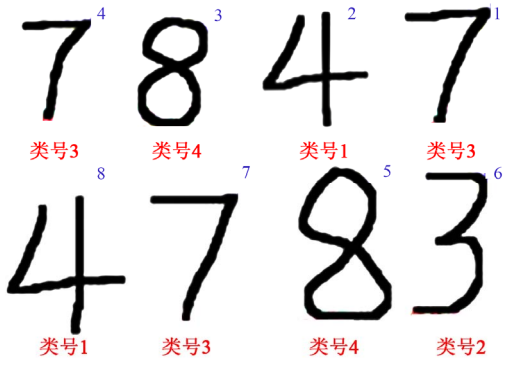
(a) Dialog of distance selection
(a) 距离选择对话框



(b) Parameter input dialog
(b) 参数输入对话框



(c) Clustering sample
(c) 待聚类的样品



(d) The result of clustering
(d) 输出聚类结果

Figure 4. The ABC applied in handwritten digit clustering
图 4. 蜂群算法应用于数字聚类分析

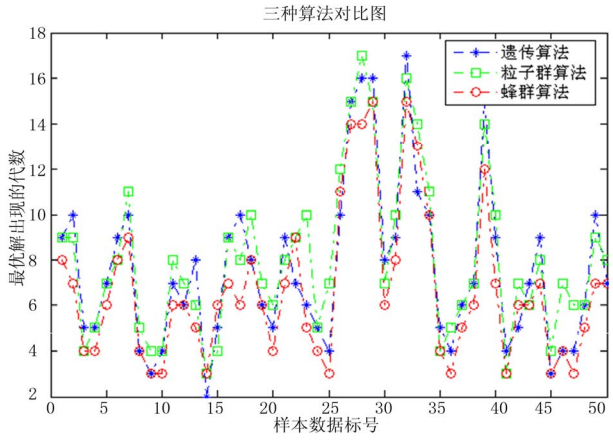


Figure 5. The result of algorithm comparison
图 5. 蜂群算法与遗传算法和粒子群算法的比较

6. 总结

本文主要基于人工蜂群算法，提出了蜂群算法的三种位置更新算子，并基于这些算子提出了一种有效的手写数字聚类方法，该方法通过蜂群的群体智能能够很好的实现手写数字聚类，并且通过实验表明该方法较传统的进化算法具有更好的收敛性能，提高的算法的效率，从整体上减少了手写数字所花费的时间，但该方法也存在一定的问题，在侦查蜂搜索蜜源的过程中只是一个随机搜索，没有参考之前所采蜜源，其解的质量可有会有一定退化。这将是接下来重点研究的工作之一。

参考文献 (References)

[1] G. Theraulaz, E. Bonabeau and J. L. Deneubourg. Response threshold reinforcement and division of labour in insect societies. Proceedings of the Royal Society of London, London, January 1998: 327-332.

[2] G. Theraulaz, S. Goss and J. Gervet. Task differentiation in polistes wasp colonies models for self-organizing groups of robots. Pairs Proceedings of the First International Conference on Simulation of Adaptive Behavior on From Animals to Animals, MIT Press, Paris, 1991: 346-355.

[3] M. Dorigo, V. Maniezzo and A. Colomi. The ant system: Optimization by a colony of cooperation agents. IEEE Transactions on Systems, Man and Cybemetics Party B, 1996, 26(1): 1.

[4] T. D. Seeley. The wisdom of the hive the social physiology of honey bee colonies. Harvard University Press, Cambridge, 1995.

[5] D. Karaboga. An idea based on bees swarm for numerical optimization. Erciyes University, Kayseri, 2005.

[6] D. T. Pham, A. Ghanbarzadeh and E. Koc. The bees algorithm—a novel tool for complex optimization problems. In Proceedings of the Abstracts of 10th EWGT Meeting, Poznan, 13-16 September 2005: 13-16.

