The methods of modeling and structure estimation building for *K*NN classifiers on the basis of small training sets

Vitaliy Tayanov

ipm.lviv.ua, 14/19 Glyboka str., Lviv, 79013, Ukraine, vtayanov@ipm.lviv.ua, www.ipm.lviv.ua

Abstract: *In this paper the fullest conception of the probabilistically combinatorial approach has been presented. This conception is the result of previous long preliminary works. The approach gives the possibility to establish the reasons of algorithms overtraining, to define the possible ways of it reduction and to build the most precise estimates of the recognition probability. The combinatorial approach works with determined data of the recognition process and the probabilistic one determines the probability of these results existence. The most usefulness of the combinatorial approach consists in the possibility to determine the effect of the training set variation on the different algorithms and select the most appropriate one from these algorithms or algorithm composition. The probabilistic part of this approach determines the probability of results, obtained on the basis of combinatorial approach.*

Keywords**:** Overtraining, probabilistically combinatorial approach, training set, classification algorithms, true (false) objects.

1. INTRODUCTION

Among all classification algorithms the most interesting and important for the research are the algorithms, using training. These algorithms are the target of the Machine Learning Theory (MLT) that has been successfully developed during last ten years [1]. This theory considers such important problems as optimal composition of the training set definition, classifiers training and optimal classifiers composition making, the most informative feature selection procedure, etc. The algorithms that can more or less solve these tasks named as Bagging, Boosting and Random Space Method (RSM). The analysis of these algorithms determines one mutual feature of them. All these algorithms oriented on the informativity increasing of the training data (the optimal training set definition and the most informative feature selection) and overindulgence (complexity) of the classification algorithms decreasing. All these approaches have been oriented on classification algorithm overtraining (overfitting) minimization and reaching the minimal rate of the recognition error.

Also the most important task is to determine all these optimal parameters, when the training set is small. For example if the error probability is 10^{-2} and the reliability of determination of this probability is 10^{-3} then according to the Vapnik-Chervonenkis theory the training set have to be of 35539 elements [2]. There is no possibility (in the most of cases) to obtain the training set of such size in practice. That is why it is extremely important to develop these approaches. The idea of such type approaches consists in additional information reception about the objects and their relations. In this paper we propose some of them.

2. SOME IMPORTANT TASKS OF MACHINE LEARNING THEORY

The modern theory of machine learning has two vital problems: to obtain precise upper bound estimates of the overtraining (overfitting) and ways of it overcoming. Now the most precise familiar estimates are still very overrated. So the problem is open for now. It is experimentally determined the main reasons of the overestimation. By the influence reducing they are as follow:

1. The neglect of the stratification effect or the effect of localization of the algorithms composition. The problem is conditioned by the fact that really works not all the composition but only part of it subject to the task. The overestimation coefficient is from several tens to hundreds of thousands.
2. The neglect of the algorithms similarity. The overestimation coefficient for this factor is from several hundreds to tens of thousands. This factor is always essential and less dependent from the task than first one.
3. The exponential approximation of the distribution tail area. In this case the overestimation coefficient can be several tens;
4. The upper bound estimation of the variety profile has been presented by the one scalar variety coefficient. The overestimation coefficient is often can be taken as one but sometimes it can be several tens.

The reason of overtraining effect has been conditioned by the usage of an algorithm with minimal number of errors on the training set. This means that we realize the one-sided algorithms tuning. The more algorithms are going to be used the more overtraining will be. It is true for the algorithms, taken from the distribution randomly and independently. In case of algorithm dependence (as rule in reality they are dependent) it is suggested that the overtraining will be reduced. The overtraining can be in situation if we use only one algorithm from the composition of two algorithms. Stratification of the algorithms by the error number and their similarity increasing reduces the overtraining probability.

Let consider a duplet algorithm-set. Every algorithm can cover a definite number of the objects from the training set. If one uses internal criteria [3] (for example in case of metrical classifiers) there is the possibility to estimate the stability of such coverage. Also we can reduce the number of covered objects according to the stability level. To cover more objects we need more algorithms. These algorithms should be similar and have different error rate.

There is also interesting task of redundant information decrease. For this task it is important to find the average class size, guaranteeing the minimal error rate. The reason in such procedure conditioned also by the class size decrease for the objects, interfering the recognition on the

training phase.

The estimation of the training set reduction influence on the recognition results gives the possibility to define the data structure (the relationship between etalon objects and objects that are the spikes or non-informative ones). Also the less class size the less time needed for the decision making procedure. But the most importance of such approach consists in possibility to learn precisely and to understand much deeper the algorithms overtraining phenomenon.

In this paper we are going to consider the metrical classifiers. Among all metrical classifiers the most applied and simple are the *k*NN classifiers. These classifiers have been used to build practical target recognition systems in different areas of human's activity and the results of such classification can be easily interpreted.

3. PROBABILISTIC APPROACH TO PARAMETRICAL OPTIMIZATION OF THE *K*NN CLASSIFIERS

The most advanced methods for algorithm composition optimization, informative training set selection and feature selection are bagging, boosting and random space method (RSM). These methods try to use the information, containing in the learning sample as possible as they can. Let us consider the metrical classifier optimization in feature space, using different metrics. The most general presentation of the measure between feature vectors **x** and **y** has been realized through Manhatten measure as the simple linear measure with weighted coefficients a_i [4]:

$$d(x,y)=\sum_{i=1}^{n} a_i \mid x_i - y_i \mid, \qquad (1)$$

where $d(x,y)$ is the arbitrary measure between vectors **x** and **y**.

Minkovski measure as the generalized measure in pattern recognition theory can be presented in form of

$$d(x,y)=(\sum_{i=1}^{n} \mid x_i - y_i \mid^p)^{\frac{1}{p}} = (\sum_{i=1}^{n} a_i \mid x_i - y_i \mid)^{\frac{1}{p}} = \\ = C(p)\sum_{i=1}^{n} a_i \mid x_i - y_i \mid, \qquad (2)$$

where parametrical multiplier $C(p)$ have been presented in form of

$$C(p)=(\sum_{i=1}^{n} a_i \mid x_i - y_i \mid)^{\frac{1-p}{p}};\ a_i = \mid x_i - y_i \mid^{p-1};\ p>0. \qquad (3)$$

One can make the following conclusions. An arbitrary measure is the filter in feature space. It determines the weights on features. The weight must be proportional to the increase of one of indexes, when feature has been added to the general feature set that has been used for the class discrimination procedure. Such indexes are: correct recognition probability, average class size, divergence between classes, Fisher discriminant etc. [5]. One can use another indexes, but the way of their usage has to be similar. If one of the features does not provide the index increase (or worsen it) the value of such feature weight should be taken as zero. So by force of supplementary decrease of feature number one can accelerate the recognition process, retaining the qualitative characteristics. The feature optimization problem and the measure selection have been solved uniquely. This procedure has been realized, using weighted features and linear measure with weighted coefficients. Feature selection task at the same time has been solved partially. First the feature subset from general set is determined. Such set has been determined by some algorithm (for example by the number of orthogonal transforms). The algorithm should satisfy the following conditions: class entropy minimization or divergence maximization between different classes. These conditions have been provided by the Principle Component Analysis [4]. The last parameter, using in the model is the decision function or decision rule. Number of decision functions can be divided onto functions, working in feature space and the functions, based on distance functions. For example the Bayes classifier, linear Fisher discriminant, support vector machine etc. work in feature space. The decision making procedure is rather complex in multidimensional feature space, when one uses such decision rules. Such circumstance is especially harmful for continuous recognition process with pattern series that have been recognized. Thus, realizing the recognition system with large databases in practice, one uses classifiers, based on the distance function. The simplest classifier is 1NN. But this classifier has been characterized by the smallest probability indexes. Therefore one should use *k*NN one. So the task consists in selection of *k* value that is optimal for decision making procedure in fiducial interval bound. This interval corresponds to the list of possible candidates. In such case *k* value has upper bound by the class size. In classical approach the nearest neighbor value should be taken rather large, approximating Bayes classifier.

4. PROBABILISTIC APPROACH TO NON-PARAMETRICAL OPTIMIZATION OF THE *K*NN CLASSIFIERS

Let us consider RS with training and self-training. The calculation and analysis of the parameters of such systems is carried out on the basis of learning set. Let there exists the feature distribution in linear multidimensional space or unidimensional distribution of distances. We are going to analyze the type of such distribution. The recognition error probability for the mean $\mu = 0$ could be presented as $\int_{|x|\geq\theta} p(x)dx$, where θ is the threshold. According to the Chebyshev inequality [6] we obtain $\int_{|x|\geq\theta} p(x)dx \leq \frac{\sigma^2}{\theta^2}$.

Let consider the case of mean and variance equality of $p(x)$ distribution. The upper bound for single mode distributions with $\mu = 0$ mode with help of Gauss inequality [7] is equal:

$$P(\mid x-\mu \mid \geq \lambda\tau) \leq \frac{4}{9\lambda^2}, \qquad (4)$$

where $\tau^2 \equiv \sigma^2 + (\mu - \mu_0)^2$.

Let $\mu = \mu_0 = 0$ and $\tau \equiv \sigma$. Then the threshold θ is $\theta = \lambda\tau = \lambda\sigma$ and $\lambda = \dfrac{\theta}{\sigma}$. Thus the Gauss inequality for the threshold θ could be presented in form of:

$$\int_{|x|\geq\theta} p(x)dx \leq \frac{4\sigma^2}{9\theta^2}. \qquad (5)$$

As seen from (5), the Gauss upper bound estimate for the single-mode distribution is better in 2.25 times then for the arbitrary distribution. So the influence of the distribution type on the error probability is significant. The normal distribution has equal values of mode, mean and median. Also this distribution is the most popular in practice. From the other hand the normal distribution has been characterized by the maximum entropy value for the equal values of variance. This means that we obtain the minimal value of classification error probability for the normally distributed classes. For the algorithm optimization one should realize the following steps:

- to calculate the distance vector between objects for the given metric;
- to carry out the non-parametrical estimation of the distance distribution in this vector by the Parzen window method or by the support vector machines;
- to estimate the mean and variance of the distribution;
- on the basis of the estimated values to carry out the standardization of the distribution ($\mu = 0$, $\sigma = 1$);
- to build the distributions both for the theoretical case and estimated one by the non-parametrical methods;
- to calculate the mean square deviation between the distributions;
- to find out the parameter space, when deviation between the distributions less then given δ level.

5. COMBINATORIAL APPROACH

Let present the recognition results for *k*NN classifier in form of binary sequence:

$$\underbrace{\underbrace{\underbrace{11111111110}_{l_1}\underbrace{001}_{m_1}\underbrace{11111111}_{l_2}\underbrace{10000}_{m_2}}_{I_t}\underbrace{11}_{l_3}\underbrace{00}_{m_3}}...$$

Fig.1. The recognition results in form of binary sequence for *k*NN classifier

Using *k*NN classifier, it is important that among *k* nearest neighbours we have the related true objects majority or the absolute one. Let consider more simple case, meaning the related majority. The *k*NN classifier correct work consists in fact that for *k* nearest neighbours it has to be executed the condition

$$\left|\bigcup_i \tilde{l}_i\right| > \left|\bigcup_i \tilde{m}_i\right|, i = 1,2,3..., \qquad (6)$$

where $\tilde{l}_i, \tilde{m}_i$ are the groups that appear after class size decrease. Under the group one understands the homogeneous sequence of elements. In such sequence (see Fig.1) there exist patterns of all classes. In general case there is no the direct conformity between the group number and the class number although.

Let consider the case of non-pair *k* value in *k*NN classifier only. This means that we have the case of synonymous classification. Such univocacy could disappear in case of pair *k* value and votes equality for different classes.

Let estimate the effect of class size reduction in case of *k*NN classifier. Note that reduced class sizes are equal to each other and equal s^*. It is considered the *k*NN classifier correct work condition: $ENT\left(\dfrac{k}{2}\right)+1 \leq s^*$. In contradistinction to 1NN classifier there is no such an importance of the first nearest patterns of the true class. Thus all such sequences one could denote as l_i. Let determine the probabilities that it will be selected s^* patterns from the true class by the combinatorial approach. These probabilities have fiducial sense. This means that for the given part of true objects there will be no selections among the patterns of the false classes by the correspondent combinatorial way. The multiplication of pointed two probabilities determines the probability of *k*NN classifier correct work. Let assign q_j as the recognition error probability for the corresponding m_i groups:

$$q_1 = P\left(\inf\left(\left|\bigcup_i m_i\right|\right) \geq ENT\left(\frac{k}{2}\right)+1\right);$$

$$q_2 = P\left(\inf\left(\left|\bigcup_i m_i\right|\right) + |m_{i+1}| \geq ENT\left(\frac{k}{2}\right)+1\right);$$

$$q_3 = P\left(\begin{array}{l}\inf\left(\left|\bigcup_i m_i\right|\right) + |m_{i+1}| + |m_{i+2}| \geq \\ \geq ENT\left(\dfrac{k}{2}\right)+1\end{array}\right);... \qquad (7)$$

$$q_j = P\left(\begin{array}{l}\inf\left(\left|\bigcup_i m_i\right|\right) + \left|\bigcup_j m_{i+j-1}\right| \geq \\ \geq ENT\left(\dfrac{k}{2}\right)+1\end{array}\right);...$$

The combinatorial expression for q_j probability could be written in form of:

$$q_j = \frac{\sum_{j=ENT\left(\frac{k}{2}\right)+1}^{s^*} C_j^{\left|\bigcup_{i,j} m_{i+j-1}\right|} C_{s-\left|\bigcup_{i,j} m_{i+j-1}\right|}^{s^*-j}}{C_s^{s^*}}, \left|\bigcup_{i,j} m_{i+j-1}\right| \geq ENT\left(\frac{k}{2}\right)+1. \qquad (8)$$

The fiducial probability for arbitrary true pattern sequence is equal:

$$P_{q_j} = \frac{\sum_{j=ENT\left(\frac{k}{2}\right)+1}^{s^*} C_j^{\left|\bigcup_i l_i\right|} C_{s-\left|\bigcup_i l_i\right|}^{s^*-j}}{C_s^{s^*}}. \qquad (9)$$

Thus the correct recognition probability for *k*NN classifier has been determined by probability (9) and addition to probability (8):

$$P_j = P_{q_j}(1-q_j) = \frac{\sum_{j=ENT\left(\frac{k}{2}\right)+1}^{s^*} C_{\left|\bigcup_i l_i\right|}^{j} C_{s-\left|\bigcup_i l_i\right|}^{s^*-j}}{C_s^{s^*}} - \frac{\left(\sum_{j=ENT\left(\frac{k}{2}\right)+1}^{s^*} C_{\left|\bigcup_i l_i\right|}^{j} C_{s-\left|\bigcup_i l_i\right|}^{s^*-j}\right)\left(\sum_{j=ENT\left(\frac{k}{2}\right)+1}^{s^*} C_{\left|\bigcup_{i,j} m_{i+j-1}\right|}^{j} C_{s-\left|\bigcup_{i,j} m_{i+j-1}\right|}^{s^*-j}\right)}{\left(C_s^{s^*}\right)^2}. \quad (10)$$

It is modeled the recognition process with different sequences of patterns of true and false classes for the 1NN and *k*NN classifiers in case of related majority. For the modeling the face recognition system has been taken. The class size (training set) has been taken as 18 according to the database we made. On the Fig.1 the results of modeling of the training set decrease influence on the recognition results for the 1NN classifier have been presented. On the Fig.2 the similar results for the *k*NN classifier under condition $ENT\left(\frac{k}{2}\right)+1=s^*$ have been presented.

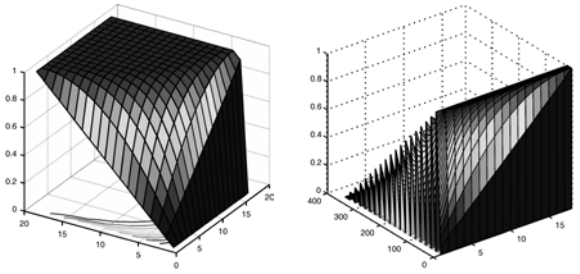

Fig.2. The probability of correct recognition as function of training set (x axis) and number of true/false objects in the target sequence (y axis) for the 1NN classifier

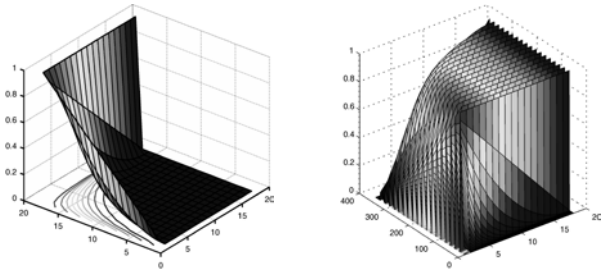

Fig.3. The probability of correct recognition as function of training set (x axis) and number of true/false objects in the target sequence(y axis) for the *k*NN classifier

On the Fig.1,2 x axis means the size of the training set and the y axis means the size of the true pattern sequence (left picture) and sequence of both true and false patterns (right picture). The y axis has been formed by the following way. We organized 2 cycles where we changed the number of true and false patterns. For every combination of these patterns and different class size we calculate the probability of correct recognition.

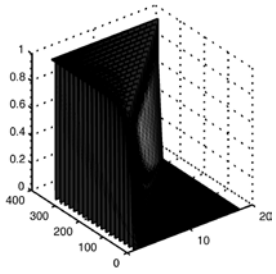

Fig.4. The probability of correct recognition as function of training set (x axis) and $ENT\left(\frac{k}{2}\right)+1$ value (y axis)

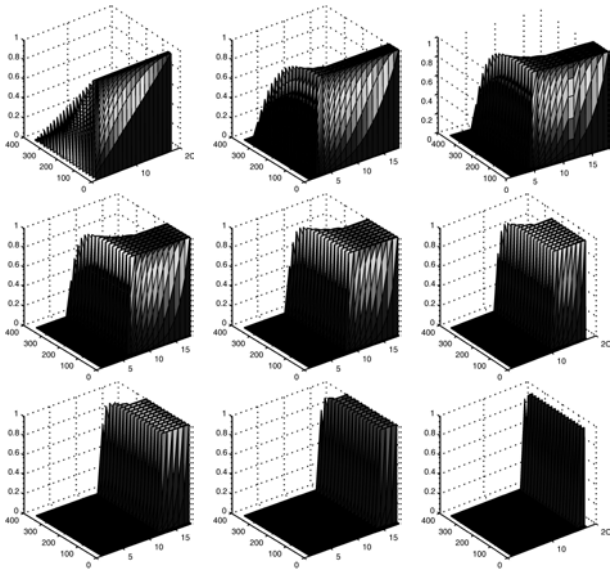

Fig.5. The probability of correct recognition as function of $ENT\left(\frac{k}{2}\right)+1$ (x axis) and number of true/false objects in the target sequence (y axis)

On the Fig.4,5 the results of *k*NN classifier modeling have been presented. Here it has been satisfied the following condition: $ENT\left(\frac{k}{2}\right)+1 \le s^*$. On the Fig.5 the fiducial probability as function of training set size (x axis) and $ENT\left(\frac{k}{2}\right)+1$ value (y axis) has been presented.

6. SOME MOMENTS OF PROBABILISTICALLY COMBINATORIAL APPROACH APPLICATION

The probability part of proposed approach consists in following idea. Despite of combinatorial approach, where the recognition results were determined precisely, we define only the probability of the initial sequence existing. Due to low probability of arbitrary sequence existing (especially for the large sequences) it has been determined the probability of homogeneous sequences existing of the type $\{0\}$ or $\{1\}$. This probability has been determined on the basis of the last object in given sequence as probability of replacing this object (the object from the true class $\{1\}$ by the others objects of the

false classes from the database. This means that the size of homogeneous sequence has been determined by the most "week" object in the homogeneous pattern sequence. The probability of existing of the non-homogeneous sequences is inversely proportional to the $2^{|l+m|}$ value, where $|l+m|$ is the sequence size. This procedure could be realized, using distribution function (fatigue function) of the distances between objects. This approach has been developed for metrical classifiers and classifiers on the basis of distance function in [3,8,9,10]. Thus we need to calculate the probability of sequence with true patterns that has definite size or for the given probability rate we need to calculate the maximal size of the sequence that satisfies this probability. For the binary sequence the sum of the weights of the lower order bits is always less than the next more significant bit. The difference is equal to 1. This means that arbitrary pattern replacement of the true class in the fiducial interval is equivalent to the alternate replacement of the previous ones. The minimal whole order of the scale of notation that has such peculiarity is equal to 2. Thus we need to calculate the weights of the true patterns position and compare them with binary digit. Such representation of the model allows us to simplify the probability calculation of the patterns replacement from the true sequences by the patterns of false classes. From the other side the arbitrary weights can be expressed through the exponent of number 2 that also simplifies the presentation and calculation of these probabilities. So the probability of the homogeneous sequence of the true patterns existence has been calculated on the basis of distance distribution function and is the function of the algorithm parameters. We should select the sequence of the size that has been provided by the corresponding probability. We after apply the combinatorial approach that allows us to calculate the influence effect of the class size decrease on the recognition probability rate. Since the probabilistic part of the given approach has been determined by the recognition algorithm parameters, the integration of both probabilistic and combinatorial parts allows us to define more precisely the influence of the effect of the training set reduction.

Let consider step by step the example of fast computing of the probability of replacement of true pattern from the sequence by the false one, where relation between weights of the objects is whole exponent of number 2. Thus for example the weights can be presented by the following way: $w = \{2^9, 2^6, 2^4, 2^3, 2^2, 2^1, 2^0\}$. As known the probability of replacement of true object from the sequence by the false one, when it is known that replacement is true event is inversely proportional to the weights of these objects. Let define the probability of replacement of the object, having the 2^9 weight comparatively to the object with 2^6 weight. As far as we do not know what object has been replaced the total weight of the fact that there will not be replaced the objects with 2^6 weight and lower is equal: $w = 2^6 + 2^4 + 2^3 + 2^2 + 2^1 + 2^0$. This weight can be expressed trough 2^6 weight accurate within 1 by following way: $2^6(1+0.5) = 1.5 * 2^6$. In case of large sequences this one has week influence on the accuracy. The relation between 2^9 and 2^6 is equal to 8. In case of divisible group of events we obtain that $8\lambda + 1.5\lambda = 1$, where the proportional coefficient λ approximately equal to 0.11. So the probability of non-replacement of the object with 2^9 weight is equal to $8 * 0.11 = 0.88$. The object with 2^6 weight has the corresponding probability equal to $1 - 0.88 = 0.12$. Since we know exactly that replacement is the true event and the last object has weight equal to 1 the accuracy correction that is equal to 1 makes the appropriate correction of probability calculation.

7. CONCLUSION

In this paper the results of both combinatorial and probabilistic approach have been presented. As seen from the figures there was realized the advanced analysis and estimation of the recognition results, when the training set is decreased. So we can make the prognosis of the recognition probability for reduced training sets, using combinatorial approach. The reliability of such method can be provided on the basis of probabilistic approach.

8. REFERENCES

[1] M. Skurichina, L.I. Kuncheva, R.P.W. Duin. Bagging and Boosting for the Nearest Mean Classifier: Effects of Sample Size on Diversity and Accuracy. *Multiple Classifier Systems. Proc. Third International Workshop MCS*, Cagliari, Italy 2002, pp.62-71.

[2] V. Vapnik. *The nature of statistical learning theory*. Springer-Verlag, 2 ed. New York, 2000.

[3] B.E. Kapustii, B.P. Rusyn, V.A. Tayanov. Features in the design of optimal recognition systems, *Automatic Control and Computer Sciences*, 42 (2) (2008). pp. 64-70.

[4] T. K. Moon, W. C. Stirling. *Mathematical methods and algorithms for signal processing*. Prentice-Hall, Inc., N.J., 2000.

[5] A. R. Webb. *Statistical Pattern Recognition*. John Wiley & Sons, Inc. Chichester, West Sussex PO198SQ, England, 2002.

[6] E. W. Weisstein. *Chebyshev Inequality*. From MathWorld – A Wolfram Web Resource., http://mathworld.wolfram.com/ChebyshevInequality.html, 10.12.2008.

[7] E. W.Weisstein. *Gauss's Inequality*. From MathWorld – A Wolfram Web Resource., http://mathworld.wolfram.com/GausssInequality.html, 10.12.2008.

[8] B.E. Kapustii, B.P. Rusyn, V.A. Tayanov. The new approach to definition of correct recognition probability of object sets. *Control systems and machines*, 2 (2005). pp. 8-13.

[9] B.E. Kapustii, B.P. Rusyn, V.A. Tayanov. Mathematical Model of Recognition Systems with Small Databases. *Journal of Automation and Information Sciences*, 39 (2007). pp. 70-80.

[10] B.E. Kapustii, B.P. Rusyn, V.A. Tayanov. Classifier optimization in small sample size condition, *Automatic Control and Computer Sciences*, 40 (5) (2006). pp. 17-22.