International Journal of Business, Human and Social Sciences
ISSN: 2517-9411
Vol:4, No:2, 2010

# Information Gain Ratio Based Clustering for Investigation of Environmental Parameters Effects on Human Mental Performance

H. Mehdi, Kh. S. Karimov, and A. A. Kavokin

***Abstract*****—**Methods of clustering which were developed in the data mining theory can be successfully applied to the investigation of different kinds of dependencies between the conditions of environment and human activities. It is known, that environmental parameters such as temperature, relative humidity, atmospheric pressure and illumination have significant effects on the human mental performance. To investigate these parameters effect, data mining technique of clustering using entropy and Information Gain Ratio (IGR) K(Y/X) = (H(X)–H(Y/X))/H(Y) is used, where $H(Y)=-\sum P_i \, ln(P_i)$. This technique allows adjusting the boundaries of clusters. It is shown that the information gain ratio (IGR) grows monotonically and simultaneously with degree of connectivity between two variables. This approach has some preferences if compared, for example, with correlation analysis due to relatively smaller sensitivity to shape of functional dependencies. Variant of an algorithm to implement the proposed method with some analysis of above problem of environmental effects is also presented. It was shown that proposed method converges with finite number of steps.

***Keywords*****—**Clustering, Correlation analysis, Environmental Parameters, Information Gain Ratio, Mental Performance.

## I. INTRODUCTION

ENVIRONMENTAL parameters such as temperature, relative humidity and atmospheric pressure have significant effects on human performance especially the mental performance. Several papers contain reviews and there is substantial evidence of an association between work performance and temperature. Some researches [1,2,3] indicates that most comfort temperature yields optimal work performance. Pepler and Warner [4] investigated the learning performance of university students at six temperatures ranging from $16.7^0$ C to $33.33^0$ C, beside temperature, relative humidity [5], atmospheric pressure [6] and illumination [7] have their own significance at human mental performance. To investigate these parameters effects, data mining technique of clustering based on entropy and information gain ratio (IGR) is used.

H. Mehdi is with the G.I.K. Institute of Engineering Sciences and Technology, Topi, NWFP, 23640 Pakistan (phone: +92-333-4931259; fax: +92-938-271865; e-mail: gcs0804@giki.edu.pk).

Kh. S. Karimov is with the G.I.K Institute of Engineering Sciences and Technology, Topi, NWFP, 23640 Pakistan (e-mail: khasan@giki.edu.pk).

A.A.Kavokin. is with the G.I.K. Institute of Eng. Sciences and Technology, Topi, NWFP, 23640 Pakistan (e-mail: kavokin@giki.edu.pk).

According to [8,9,10] clustering is the classification of objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters), so that the data in each subset (ideally) share some common trait - often proximity according to some defined distance measure. There exists number of algorithms for the clustering of data, which includes hierarchal, partitioning and grid based algorithms. Different implementations of these algorithms also exist and are being used in many application areas. Comprehensive survey of clustering algorithms can be found in [10]. For clustering, relationship between different data items is very important. If we have better understanding of connection between data items then we can easily cluster the data with high accuracy. Different metrics like correlation coefficient have been proposed in the literature in order to find relationships among variables [9, 10].

In this paper, a method for finding optimum bounds of clustering [9,11] is used. Currently this method assumes that first factor (Y) measures according to the fix range scale and the second one (X) according to the uncertain intervals. Let us explain this idea with an application in environmental scenario. Let's investigate the relationship between environmental parameters (X) and mental performance (Y) (e.g. time for solution of a standard task). From the common sense we know the intervals for evaluation of performance (e.g. 30%, 70% and 100%). There is only need to find the optimum bounds of clusters in 2-dimensional space for temperature and performance. For this IGR can be used as distance metric [11].

This paper is organized as follows. Section II provides some key properties of proposed technique (IGR), Section III explains the algorithm to find optimum clustering bounds and summaries the algorithm and also provide some mathematical analysis of the algorithm. Section IV gives an application of this algorithm to investigate the environmental parameters effects on human mental performance and result discussions. Section V gives the concluding remarks.

## II. PROPOSED TECHNIQUE

Let $X=\{x(i)\}$, $Y=\{y(i)\}$ are sets of states $x(i)$ and $y(i)$, that take the investigation factors X and Y into the testing experiments with the corresponding probabilities $\mathbf{P}(.j) > 0$ and $\mathbf{P}(i.) > 0$ , ( $i = 1,...,n; j = 1,...,m$ ). The probability of the compatible appearance of values $y(j)$ and $x(i)$ via $P(i,j)$ is designed. It is supposed that we determined the information gain ratio K(Y/X) as IMC (the informative measure of the connection) for X and Y in the following way [8]:

$$\mathbf{K}(Y/X) = (\mathbf{H}(X) - \mathbf{H}(Y/X)) / \mathbf{H}(Y) \quad (1)$$

Where,

$$\mathbf{H}(Y) = -\sum_{i=1}^{n} \mathbf{P}(i.\} * \mathbf{ln}(\mathbf{P}(i.) \quad (2)$$

is called the entropy of factor Y. The Ratio, **K**(Y/X) has following three properties:

1) **0** <= **K**(Y/X) <= **1**;
2) **K**(Y/X) = **0**,
If X and Y are stochastically independent
3) **K**(Y/X) = **1**, if and only if X and Y are fully functional dependent, that is the each value $x(i) \in X$ corresponds with the unique $y(j) \in Y$. (3)

Following property can obviously be derived from above.

Properties 2) and 3) of the ratio **K**(Y/X) are the basis to use it as IMC.

In order to have same representation about the quantity **K**(*Y/X*) for different degrees of connection between *X* and *Y*, let us reduce results of the numerical experiment for the computer for the line model of the one-factor dispersion analysis with random effects [8], having a view,

y(*i, k*) = $\sqrt{d}$ * *Z*(*i*) + $\sqrt{1-d}$ * *t*(*i, k*),(*i* = 1,...,*r; k* = 1,...,*n*), where *Z*(*i*) and *t*(*i, k* ) are appropriate components of vectors of random numbers with the normal distribution, the null mean and the variance *d* and *(1-d)* respectively. Where *d* is the parameter characterizing (as seen from the determination y(*i,k*)) the closeness of connection.

As one should expect, **K**(*Y/Z*) grows monotonously from ***d*.** **K**(*Y/Z*) = 0 when d = 0 and **K**(*Y/Z*) = 1 when *d* = 1, that it agrees with cases of the full absences of dependence and the presence of the functional dependence. It is for intermediate values **K**(*Y/X*) = 0.2 when d = 0.4 -:-0.5; **K**(*Y/X*) = 0.5 when d = 0.8 -:-0.9. It is to say that values **K**(*Y/X*) = 0.25 – 0.5 intuitively conform the presence of the middle degree of dependence between *X* and *Y*, and **K**(*Y/X*) > 0.5 is the presence of the strong one, that is near to the functional dependence.

It is known from the information theory [13], that:

$$\mathbf{H}(Y/X) = \mathbf{H}(Y,X) - \mathbf{H}(Y), \quad (4)$$

Where **H**(Y/X) is Kullback-Leibler divergence, the conditional entropy of the event under the condition of performance of the X ; **H**(Y,X) is the entropy of the simultaneous appearance of occurrences Y and X. Values on the right side in the formula (4) will be written via probabilities. Considering the probabilities the equation (4) can be written as gives rise to the view

$$\mathbf{H}(Y/X) = -\sum_{j=1}^{m}\sum_{i=1}^{n} P(i,j) * \ln(P(i,j))$$

$$+ \sum_{i=1}^{n} P(i,.) * \left(\ln\left(P(i,j) / P(i,.)\right)\right) \quad (2)$$

$$= -\sum_{j=1}^{m}\sum_{i=1}^{n} P(i,j) * \left(\ln\left(P(i,j) / P(i,.)\right)\right) \quad (5)$$

## III. Algorithm of Adjusting Clustering Bounds Using K(Y/X), [11]

Here we discuss how one can find the optimum bounds of clusters. Let the event $y(j) \in Y$ represent value y for the factor Y in the *i*-th range and $x(i) \in X$ is the multitude of values for the factor X, which locate inside the interval with the length Q(*j*) and boundaries *b*(*j*) and *b*(*j+1*), (*j* <= *m*). It is supposed for the definiteness that values *b*(*j*) were arranged in increasing order, that is $b_n = b(1) < b(2) < ... < b(m+1) = b_k$, where $b_n$ and $b_k$ were the lower and the upper values from the multitude X , that is $b_n = min\{x\}$ ; $b_k = max\{x\}$.

One must define values *b*(*j*) in such manner that the dependence between X and Y would be as closed as possible and be the most near to the functional one. Taking into consideration the definition of **K**(Y/X) in the formula (1) and the constant of **H**(Y), this can be written mathematically in the following form:
One must find out those values *b*(*j*) that:

$$F = \mathbf{H}(X/Y) = -\sum_{j=1}^{m}\sum_{i=1}^{n} \mathbf{P}(i,j) * \mathbf{ln}(\mathbf{P}(i,j)) / \mathbf{P}(i,.))$$

$$\rightarrow \boldsymbol{min} \quad (6)$$

Under next conditions:

$$\sum_{j=1}^{m} Q[j] = b_k - b_n;$$

$$b_n = b(1) < b(2) < ... < b(m+1) = b_k \quad (7)$$

Q(*j*) ≠ **0**, ( i.e exist at least one *x*(*i*) inside of each Q(*j*) ); (8)

Where
$b_n = \min\{x(i)\}$ ; $b_k = \max\{x(i)\}$; (9)

$$\mathbf{P}(i,j) = f(b(j)); \quad (10)$$

$$Q(j) = b(j+1) - b(j) \quad (11)$$

for the each *j = 1,...,m*

The senses of terms (6)-(9),(11) are clear from previous definition and the term (10) indicatives that the probabilities **P**(*i,j*) entering in the minimum function (11), are implicit functions from values of the bounds values of intervals - b(*j*). The solution of the problem (6) - (11) gives us the maximum value of the ratio **K**(Y/X) and this means that it indicates those values of bounds of intervals b(*j*) under which the dependence between X and Y is the most near to the functional one.

International Journal of Business, Human and Social Sciences
ISSN: 2517-9411
Vol:4, No:2, 2010

One can represent computation solution in the form of algorithm that belongs to the category of co-ordinate descent algorithms. Diagram of the proposed algorithm is as in Fig. 1:

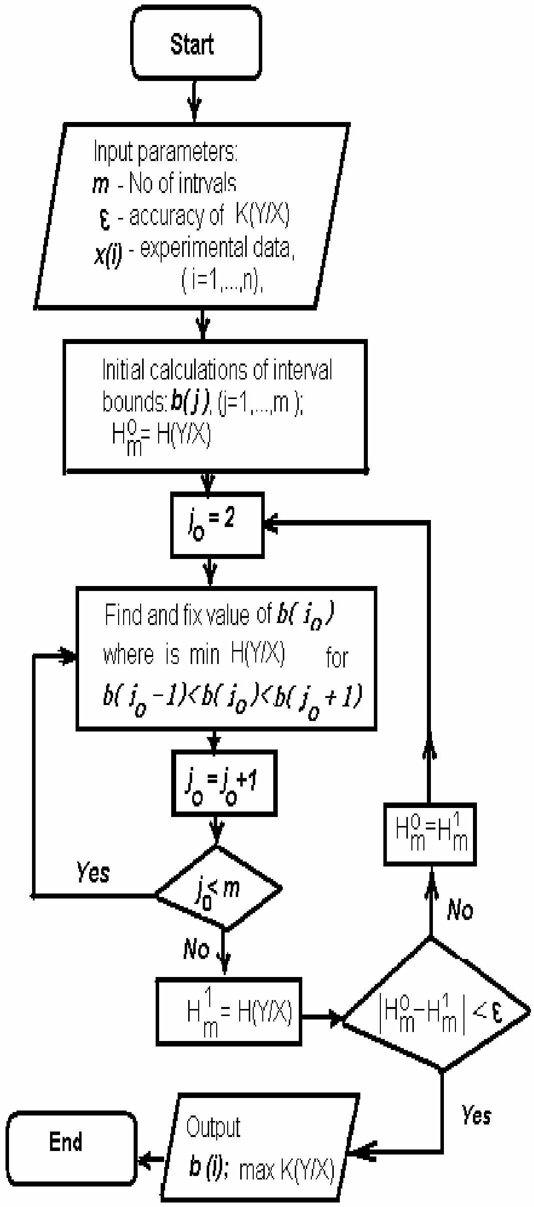


Fig. 1 The algorithm of adjusting clustering bounds using K(Y/X)

Let us see in what case the function F() in (6) will monotonously vanish. For the sake of this, we will discuss property of the ratio **K**(Y/X), that characterizes the closeness of connection between X and Y.

It was shown, [11] that when K(Y/X) →1 the connection between X and Y approaches to functional one. At the same time H(Y/X) will monotonically tends to null.

Therefore calculating values of K(Y/X) will be used for range of difference factors by their entropy. In principle the quest of the global maximum in the whole change region of *b(j)* can be conducted by the method of the sequentially checking of variants [8], however that will practically increased the volume of calculation. For example, if each *b(j)* run for *N* values, it needs $O\left(N^m\right)$ calculations of F(*b(j)*) for the whole searching of the variant, while they are in all $O\left(N*m\right)$ in the proposed method.

The application of the gradient descent method [8] of *min*( F(*b(jo),b(j)*), *b(jo)* is little-effective both by reason of the setting difficulty of the increase step on the co-ordinate and by reason of the existence of local minimum at the function F(*b(j)*).

## IV. Investigation of Environmental Parameters Effects

In order to investigate the effects of environmental parameters on human mental performance, we conducted our experiment in which we involved five persons from our institute. For this made same psychological assessment tests were used and the task was given to every person. All experiments were conducted approximately at the same time of day. During the solution of the task the environmental parameters were measured and we recorded the time of solution as indicator of performance for each person was recorded. In the Table I one can see the *min* and *max* time which participants spent to solve the task throughout the whole period of our experiment.

After end of the experiment *min* and *max* outlier for each person (in average two records of each person) was removed and then data was normalized using (12), where $t_{min}$ and $t_{max}$ are given in Table I. This kind of normalization allowed us to eliminate influence of individual skills of participants (such as specific analytical skills etc.) and investigate only the deviation of performance versus changing of environmental parameters. After normalization first the classical technique of correlation analysis was applied, for investigation of the existence of dependencies between temperature, pressure and time to solve the task. Set of sampled data 104 records about normalized time and temperature in ascending order are represented in Fig. 3 and Fig. 4, where different label shows the results of different participants. Note that measurement of temperature and pressure has been conducted everyday, but mental performance tests were conducted on the availability of participants (see Table II).

$$t = \frac{t_i - t_{\min}}{t_{\max} - t_{\min}} * 100 \tag{12}$$

Fig. 2 shows the relationship between temperatures versus atmospheric pressure. The correlation coefficient between these two parameters is -0.45 and level of significance α<0.01 i.e.; there exists strong dependence between these parameters. Based of this observation it was decided to focus our investigation of dependency only between temperature and time of solution of task.

Correlation analysis shows that there is not exist linear correlation between the temperature and time of solution, because *r*=- *0.01*(correlation coefficient). But on the other hand one can see in Fig. 3 that smallest time of solution i.e.

the best performance is noticeable at middle values of our measured temperature. That was the reason to apply the proposed technique of clustering for investigation of effects. First we split normalized time of solution on three subintervals i.e. 0-31%, 32%-75% and 76%-100%.

Fig. 3 shows the initial clusters of different day's temperatures versus percentage of spent time to solve the task, with K(Y/X)=0.026. When we processed our data according to the algorithm [11], it adjusts cluster boundaries on the basis of entropy and IGR, after the adjustment of boundaries the final clusters will become as shown in Fig. 4. We get the gain ratio coefficient K(Y/X) = 0.43. In mentioned above [11], values of K(Y/X) between 0.25 – 0.5, correspondence with diapason from middle to strong dependencies, hence one can characterize existence dependency between above parameter as middle close to strong. This is in good accordance with results [14], where it was shown that maximal performance (i.e. *min* time of solution) was at temperature 20-23$^0$C.

Analyzing the results of clustering in Fig. 4, one can make conclusion that for investigated group of participants the average increment to the minimal time of solution into intervals of temperatures 19.4 - 24.9$^0$ C was about 47% of the difference between *max* and *min* time of solution (see Table I) and for interval 25.6 - 27.6$^0$ C was about 50% ,whereas in interval of optimal temperature 24.8-25.6$^0$ C average performance was only 18.7% of above difference greater than minimal time of solution for each participants. In other words, average time of solution for each participant' was close to minimal into this interval of temperatures.

TABLE I
PARTICIPANTS' MIN AND MAX SPENT TIME IN SECONDS TO SOLVE THE REQUIRED TASK

| Participants | Min. Time(sec) | Max. Time(sec) |
|---|---|---|
| 1 | 93 | 403 |
| 2 | 66 | 395 |
| 3 | 177 | 361 |
| 4 | 151 | 501 |
| 5 | 92 | 413 |

## V. CONCLUSION

Information Gain Ratio developed in theory of information [15], is more useful in some cases than linear correlation coefficient, because it is less sensible to the shape of correlation curve. This technique can also be applied in combination with other techniques. One possible application of this technique is investigation of dependencies between environmental parameters and performance. In distinguished with linear correlation analysis, describes approach of clustering using information gain ratio found middle to strong degree of dependence between temperature and performance. This result coincides well with results of investigation [14]. Strong linear dependence between indoor temperature and pressure allowed us to eliminate parameter (pressure) and investigate influence of only the parameter (temperature) instead of both (temperature and pressure).

In this work it was investigated the effects of environmental parameters on human being performance. Obtained data was processed by using entropy and information gain ratio based clustering approach, it was found that at the range of temperature 24.8-25.6$^0$ C the investigated group shows the maximum performance. These results can be used for optimization of the work conditions in different areas of human beings activity to facilitate their better performance.

TABLE II
THE EXPERIMENTAL DATA

| S# | Date, DD/MM | Temperature $^0$C | Pressure (kPa) | S# in Fig 3 & 4 |
|---|---|---|---|---|
| 1 | 3/9 | 19.4 | 970 | |
| 2* | 2/9 | 19.5 | 971 | 1 |
| 3* | 15/10 | 23.1 | 977 | 5 |
| 4* | 14/10 | 23.5 | 976 | 6 |
| 5 | 9/10 | 23.6 | 971 | |
| 6* | 4/9 | 24.0 | 968 | 7 |
| 7* | 15/9 | 24.1 | 972 | 8 |
| 8* | 7/10 | 24.1 | 969 | 9 |
| 9* | 7/9 | 24.2 | 965 | 10 |
| 10* | 12/10 | 24.5 | 975 | 11 |
| 11* | 6/10 | 24.7 | 970 | 12 |
| 12 | 28/8 | 24.8 | 970 | |
| 13 | 22/8 | 24.9 | 966 | |
| 14* | 14/9 | 24.9 | 969 | 13 |
| 15 | 27/8 | 25.0 | 968 | |
| 16* | 6/11 | 25.1 | 977 | 14 |
| 17 | 26/8 | 25.2 | 970 | |
| 18 | 21/8 | 25.4 | 965 | |
| 19* | 2/11 | 25.4 | 979 | 15 |
| 20* | 3/11 | 25.4 | 980 | 16 |
| 21 | 12/9 | 25.5 | 971 | |
| 22* | 29/10 | 25.5 | 980 | 17 |
| 23 | 25/8 | 25.6 | 967 | |
| 24* | 4/11 | 25.6 | 979 | 18 |
| 25* | 5/11 | 25.6 | 979 | 19 |
| 26* | 5/10 | 25.8 | 968 | 20 |
| 27* | 27/10 | 25.8 | 978 | 21 |
| 28 | 28/10 | 25.9 | 978 | |
| 29 | 20/8 | 26.0 | 969 | |
| 30* | 31/8 | 26.0 | 970 | 22 |
| 31* | 1/9 | 26.0 | 970 | 23 |
| 32* | 9/9 | 26.1 | 969 | 24 |
| 33 | 11/10 | 26.1 | 974 | |
| 34 | 26/10 | 26.1 | 978 | |
| 35* | 16/10 | 26.3 | 975 | 25 |
| 36 | 10/9 | 26.4 | 968 | |
| 37* | 18/10 | 26.5 | 974 | 26 |
| 38 | 19/8 | 26.6 | 968 | |
| 39 | 5/9 | 26.8 | 967 | |
| 40 | 23/10 | 26.9 | 978 | |
| 41* | 11/9 | 27.3 | 970 | 27 |
| 42* | 6/9 | 27.4 | 966 | 28 |
| 43 | 19/10 | 27.4 | 975 | |
| 44 | 13/8 | 27.5 | 961 | |
| 45* | 29/8 | 27.6 | 968 | 29 |
| 46* | 20/10 | 27.6 | 976 | 30 |
| 47 | 4/10 | 28.1 | 973 | |
| 48 | 18/8 | 28.2 | 966 | |
| 49 | 23/8 | 28.2 | 963 | |
| 50 | 11/8 | 28.3 | 964 | |
| 51 | 12/8 | 28.3 | 964 | |
| 52 | 13/9 | 28.4 | 968 | |
| 53 | 30/8 | 28.5 | 966 | |
| 54 | 10/8 | 28.9 | 963 | |
| 55 | 17/8 | 29.0 | 964 | |
| 56 | 8/8 | 29.9 | 961 | |
| 57 | 14/8 | 30.2 | 962 | |
| 58 | 9/8 | 30.6 | 961 | |

The contents are, Serial no., Temperature in ascending order, atmospheric pressure in kilopascal and serial number in Fig. 3 and Fig. 4, where ( *) represents the day of conducting test.

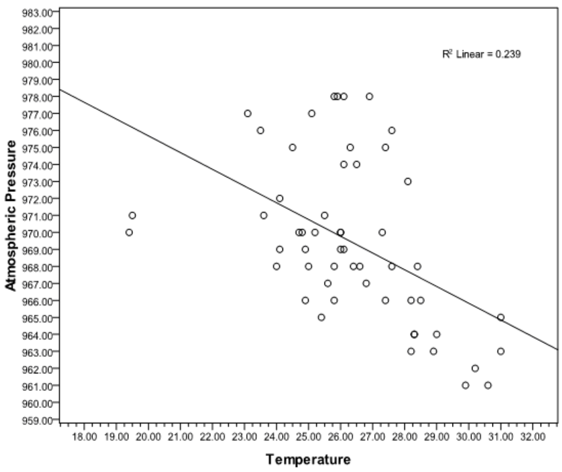


Fig. 2 Indoor temperature of day's vs. indoor atmospheric pressure (see Table II)

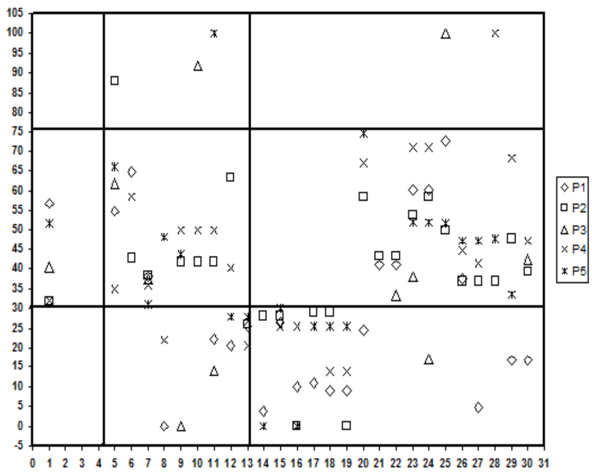


Fig 3 Initial clustering, temperature (horizontal, see Table II) vs percentage of spent time (vertical), with K(Y/X)=0.026

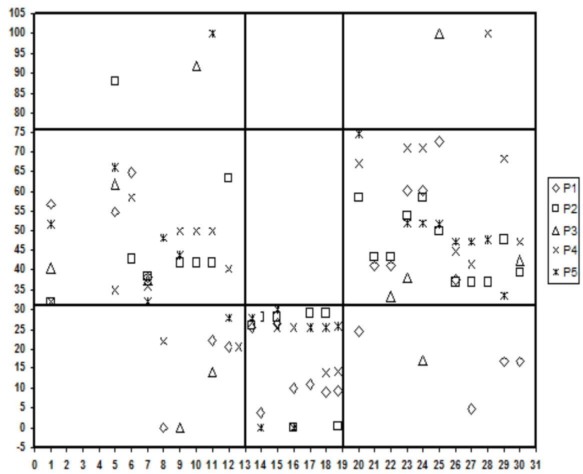


Fig. 4 Final clustering, temperature (axis are same as in Fig. 3) vs percentage of spent time with K(Y/X)=0.43

## Acknowledgment

We are thankful to Higher Education Commission (HEC) of Pakistan for providing financial assistance during the course work of this research. We are also thankful to GIKI administration for support this research and are thankful to all our colleagues in GIKI, who give their valuable time to conduct the psychological assessment tests.

## References

[1] Wyon DP, Andersen IN, and Lundqvis GR, "The e¤ects of moderate heat stress on mental performance", Scandinavian Journal of Work Environment and Health 5: 352-361, 1979.
[2] Wyon DP, "Healthy Buildings and their impact on productivity", Proceedings of Indoor Air, 6th International Conference on Indoor Air Quality and Climate, Helsinki 6:3-13,1993.
[3] Levin, H., "Physical factors in the indoor environment", Occupational Medicine: State of the Art Reviews 10(1):59-94, 1995.
[4] Pepler RD, Warner RE, "Temperature and learning: an experimental study", ASHRAE Transaction 74(II): 211-219, 1968.
[5] Green GH, "The effects of indoor relative humidity on absenteeism and colds in schools", ASHRAE Transaction 8/0:131-141, 1974.
[6] Peltonon, J., Rantamaki, J, Niittymaki, S., Sweins, K., Viitasalo, J. and Rusko, H., "Effects of oxygen fraction in inspired air on rowing performance", Medicine and Science Sports and Exercise, 27: 573-578, 1995.
[7] Smith SW and Rea MS, "Proofreading under different levels of illumination", Journal of Illumination Engineering Society, 8(1): 47-78, 1979.
[8] Eliseeva I.I., Rukavishnikov V.O., "The grouping, the correlativity, the pattern recognition", Moscow, 1977, 273 p. rus.
[9] Jain, Murty and Flynn, "Data Clustering: A Review", ACM Comp. Surv, 1999.
[10] Rui Xu Wunsch, D., II, "Survey of clustering algorithms, Neural Networks", IEEE Transactions, volume: 16, issue: 3, May 2005.
[11] A.A. Kavokin, M.Sarmad Ali, Adeel Mumtaz, Tahir Jameel, Shujaat Ali Rathore, "Adjusting clustering bounds using information gain ratio" International Journal of Software Engineering, Vol. 1. No. 2, pages 17-22.
[12] Duda, R.O., Hart, P.E., Stork, D.G., "Pattern classification" (2nd edition), Wiley, ISBN 0471056693, 2001.
[13] H. A. Taha., "Operations Research: An Introduction", Prentice Hall, 1996.
[14] Olli Seppanen, William J. Fisk, Q. H. Lei, "Effect of Temperature on Task Performance in Office Environment", NTIS, Alexandria, 2006.
[15] Kulback S., "Information Theory and Statistics", Courier Dover Publications, pp. 416.