American Journal of Applied Sciences 11 (1): 1-7, 2014
ISSN: 1546-9239
©2014 Science Publication
doi:10.3844/ajassp.2014.1.7 Published Online 11 (1) 2014 (http://www.thescipub.com/ajas.toc)

DISTINGUISHABILITY BASED WEIGHTED FEATURE SELECTION USING COLUMN WISE K NEIGHBORHOOD FOR THE CLASSIFICATION OF GENE MICROARRAY DATASET

[1]Jeyachidra and [2]Punithavalli

[1]Department of Computer Science and Applications,
Periyar Maniammai University, Vallam-613 403, Thanjavur, Tamilnadu, India
[2]Department of Computer Science, Sri Ramakrishna Engineering College, Coimbatore, Tamilnadu, India

Received 2013-09-06, Revised 2013-10-17; Accepted 2013-11-13

ABSTRACT

In data mining, much research is being carried out to discover the previously unknown, valid, novel, useful and understandable patterns in large databases. The patterns must be actionable so that they might be used for decision making to a variety of applications in healthcare. In this study, feature subset selection is an important area, where many approaches have been proposed. Hence, the authors chosen three existing feature selection algorithms analyzed their performance using the publicly available standard colon tumor dataset. The performance of the existing three methods evaluated and compared each method with DWFS-CKN under study.

Keywords: Feature Selection, Microarray Data, Classification, C4.5, Bayes

1. INTRODUCTION

Microarrays provided lot of information that have significance in various medical domain. In recent years there had been an explosion in the rate of acquisition of biomedical data. Different types of microarray used different technologies for measuring mRNA expression levels.

Machine learning and statistical techniques applied to gene expression data had been used to address the questions of distinguishing tumor morphology. Analysis of microarray presented a number of unique challenges for data mining. The main types of data analysis needed for biomedical applications including gene selection, classification and clustering. One of the major goals of microarray data analysis was discovery of biological knowledge. In this, the importance of feature selection in machine learning came from its ability of improving learning performance. Several feature selection techniques developed and discussed for many years. However, the problem of finding the optimal feature selection still remains to be a very necessary, so far difficult problem. In order to solve this problem and find a solution of the problem, the authors made a selection of three feature selection algorithms which were compared and discussed with the proposed DWFS-CKN in this study.

Feature selection was a topic that concerns selecting a subset of features among the full features that shows the best performance in classification accuracy . The process of feature selection consists of 4 steps. Starting point, Search strategy, Subset Evaluation and Stopping criteria. The starting point, the search for feature subsets started with no features or with all features , the search strategy -theoretically, the best subset of features could be found by evaluating all the possible subsets, the third, point is the subset evaluation-after generated subsets of features, the authors needed to evaluate them. To Evaluate the subset features, there two methods namely filter approach and wrapper approach used Kira and Rendell (1992) and to stop the criteria-finally, the researchers decided the criteria for halting the search. In this study

Corresponding Author: Jeyachidra, Department of Computer Science and Applications, Periyar Maniammai University, Vallam-613 403, Thanjavur, Tamilnadu, India

the authors proposed a simple and efficient feature selection algorithm called “Distinguishability Based Weighted Feature Selection Using Column Wise K Neighborhood (DWFS-CKN)”. The performance of the proposed algorithm has been compared with three algorithms Gini Index, MRMR and Relief-F since these three algorithms performed well in our previous evaluation and also the accuracy tested with two popular classification algorithms Bayes and C4.5 and validated by k-fold validation and Leave-one-out cross validation by considering accuracy as metrics. The obtained results proved that the proposed DWFS-CKN algorithm performed better accuracy as well as speed.

1.1. Objectives and Scope

Microarray experiments were expected to contribute considerably to progress in cancer treatment by enabling a precise early diagnosis, eventhough it is difficult. The objectives of the research were:

- To eliminate the redundant, irrelevant or noisy data
- To get better the data quality furthermore minimize the feature space
- To develop a new algorithm for feature selection to maximize classification accuracy

The aim of the present study was to verify whether the data selection dependent on the algorithm or not. The scope of the present study was restricted to the adoption of three algorithms for analyzing the already available data.

1.2. Previous Works

Many successful feature selection algorithms had been devised. Gheyas and Smith (2010) were involved in the study of goodness of a feature subset. Huang *et al*. (2005) suggested the well organized choice of discriminative genes from microarray gene expression data for cancer diagnosis. Dai *et al*. (2006) demonstrated the Dimension Reduction for Classification with Gene Expression Microarray Data. Wang and Palade (2007) recognized a comprehensive fuzzy based framework for cancer microarray data gene expression analysis. This method used three microarray cancer datasets namely Leukemia, colon cancer and Lymphoma cancer. A novel fuzzy based system was used for both gene selection and classification by applying the microarray gene expression data. The performance achieved by that method was more viable. Yeh *et al*. (2007) followed the data mining techniques for cancer classification using Gene data. Feature Selection from microarray dataset carried out using t-statistics (t-GA) based algorithm. The decision based classifier was used on the top datasets. Wang *et al*. (2007) proposed the approach for cancer classification using an expression of very few genes. There were two types involved in that method. The first type was of an important gene selection that was done by the use of the gene ranking scheme. The second type was of the classification accuracy of gene combination carried out by using a fine classifier. Hang and Wu (2009) described a new approach called “Sparse Representation” using Microarray gene expression profiles for cancer diagnosis. Nine human tumor types were used as data set in their research. Rejani and Selvi (2009) projected a tumor discovery as of mammogram, extracting features which categorized tumors. Microarray data analysis was conducted by Osareh and Shadgar (2010) for cancer classification. An automated system was developed for consistent cancer analysis based on gene microarray expression data. The researchers used the microarray datasets which included both binary and multi-class cancer problems.

1.3. The Proposed Distinguishability Based Weighted Feature Selection Using Column Wise K-Neighborhood

In this section the authors present a algorithm called “Distinguishability based Weighted Feature Selection using Column wise k Neighborhood (DWFS-CKN)”.

In the proposed algorithm, feature weights were calculated based on the classifiable/distinguishable nature of the corresponding member points of that features using a column wise k-neighborhood method. It meant that for a particular column of a feature, most of the points were definitely belonging to any one of the class and distinguishable from the other classes based on k-neighborhood of each value, then the feature weight of that particular column was high. So, a feature which had highest feature weight was the most important attribute of the data and a feature which had lowest feature weight was the least important attribute of the data. So, for classification tasks, the authors selected a small set of first few features which were high feature weights. The following algorithm explained the proposed Data Distinguishability based Weighted Feature Selection using Column wise k Neighborhood (DWFS-CKN).

Algorithm_ DWFS-CKN

Let

D be the set of Microarray Data of m rows of n features

```
        T be the corresponding class id's of m records of D.
        The dataset D can be grouped in to c number of sub groups based on the class membership as follows
        $D=\{ g_1, g_2, .. g_c, \}$
        Where
        $g_1, g_2, .. g_c$, are the c number of sub sets of data belonging to c classes.
        $\bar{g}_1, \bar{g}_1,.....\bar{g}_c$ are the colum-wise average of $g_1, g_2, .. g_c$,
        W-array of size of 1×n to hold the feature weights
        Dist- array of size of 1×n to hold the minimum distance.
for i = 1 to n //for every feature in the data do this
{
    for j = 1:m //for every row in the data do this
    {
        //k-neighbor Detection
        for k = 1: m//again for every row in the data do this {
        //calculate the distance between
        //the selected attribute point
        //and other points
         $d(k) = |D(j,i)^2-D (k)^2 |^{1/2}$
         }
         //we will have the set of distances of size m×1
         $d = \{ d_1, d_2,….. d_m\}$
         //sort the distances in ascending order
         idx = sort(d)
         //Now we will find top k neighbors
         Neighbors = T(idx(1: kn ))

         //find the index of neighbors which are in the same class T(j)
         Idx = find(Neighbors == T(j))

         /If there are at least k/2 neighbors belong to the class T(j)
         //then that data point is a classifiable one-increase weight
         If size(idx) >k/2 {
                W(i) = W(i)+1;
            }
         }
    }
   Features=sort(W,'descend' );
```

Now, the first n features can be used as the primary features.

1.4. The Feature Selection Algorithms

1.4.1. Gini Index

The Gini coefficient or Index was measure of inequality developed by the Italian statistician Corrado Gini and published in his 1912 paper "Variabilità e mutabilità". The Gini coefficient was often calculated by:

$$G = \left| 1 - \sum_{k=1}^{n} (X_k - X_{k-1})(Y_k + Y_{k-1}) \right|$$

1.5. MRMR

Maximum Relevance-Minimum Redundancy (MRMR) was the scheme in feature selection was to select the features that correlate the strongest with a classification variable Peng *et al*. (2005).

1.6. Relief F

Relief-F was a feature selection strategy that chosen instances randomly and changed the weights of the feature relevance based on the nearest neighbor.

1.7. Metrics Used for Performance Evaluation-Classifiers, Accuracy and Validation Methods

The most popular two classifiers namely Bayes Classifier and C4.5 Classifier were used and it was proposed by Quinlan (1993). C4.5 was the most popular and the most efficient algorithm in Decision tree-based approach these two classification algorithms were more frequently used by the previous researchers. The metrics calculated using the following formulas:

Accuracy = (TP+TN) / (TP + FP + TN + FN)

In this study the authors have used k-fold cross validation as well as leave-one-out cross validation for evaluating the performance.

2. MATERIALS AND METHODS

2.1. About the Implementation

The researchers used the feature selection tool box called 'fspackage' provided by Arizona State University. The authors implemented the proposed DWFS_CKN algorithm under MATLAB and compared their performance with three algorithms Gini Index, MRMR and Relief-F.

2.2. The Colon Tumor Microarray Dataset:

Dataset	Number of genes	classes	Training data	Test data	References
Colon tumor	2000	Normal Cancer	22 40	---	http://www.molbio.princeton.edu/colondata

The authors strong-willed to use the colon tumor dataset for this study. Because, some of the previous researchers used and highlighted the complication of this dataset. This dataset contains 62 samples collected from Colon Tumor patients and it is a publicly available standard dataset. Among them, 40 tumor biopsies were from tumors (labeled as “negative”) and 22 normal (labeled as “positive”). Each sample was represented by 2000 genes. So, the data set contains 62×2000 continuous variables and 2000 class ids.

3. RESULTS AND DISCUSSION

The **Table 1** shows the accuracy and error rate of classification by Bayes and J48 (C4.5) with respect to first 50 features selected by different feature selection algorithms. The metrics were calculated by doing Leave-One-Out (LOO) cross validation Jeyachidra and Punithavalli (2013).

The **Fig. 1** shows the accuracy of classification by Bayes and J48 (C4.5) while using the first 50 features selected by four different feature selection algorithms. The performance of the proposed DWFS_CKN was better than compared to the other three algorithms.

The **Fig. 1**, the set of bars at the right most of the chart belongs to the proposed DWFS_CKN method.

The **Table 2** shows the average accuracy, average error, maximum accuracy and minimum error achieved by Bayes classifier and J48 classifier. It was calculated by with respect to repeating the 10 fold cross validation for 25 times (each time, the data was kept in a random order).

The **Fig. 2** shows the average error of the 25 iterations of 10 fold cross validation and the performance of the proposed DWFS_CKN was better than compared to three other algorithms with respect to average error of 10 fold validation.

The **Fig. 3** shows the average accuracy of the 25 iterations of 10 fold cross validation and the performance of the proposed DWFS was better than compared to the three algorithms with respect to average accuracy of 10 fold validation.

The **Table 3** shows the time taken by the three different algorithms. In the case of MRMR, the time taken for selecting the primary features would increase with increase in the number of features, MRMR consumed more time and the performance of the MRMR was poorer than that of the other compared algorithms Jeyachidra and Punithavalli (2012).

The **Fig. 4** shows performance of the feature selection algorithms in terms of run time and in this case the performance of the proposed DWFS- CKN was better than the three algorithms except Relief-F.

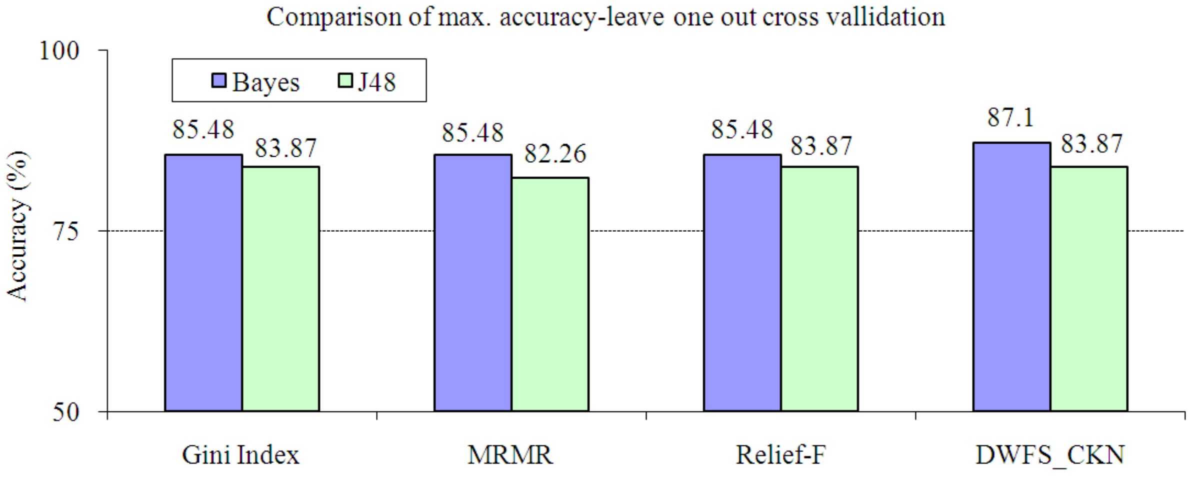

Fig. 1. The accuracy found through leave one out cross validation with respect to 50 features

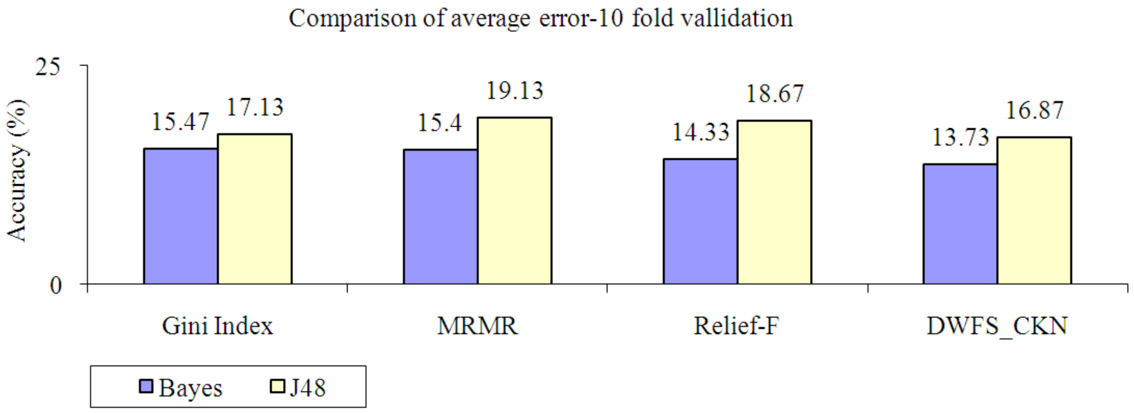

Fig. 2. Average Error of 25 Iterations of 10 Fold cross validation

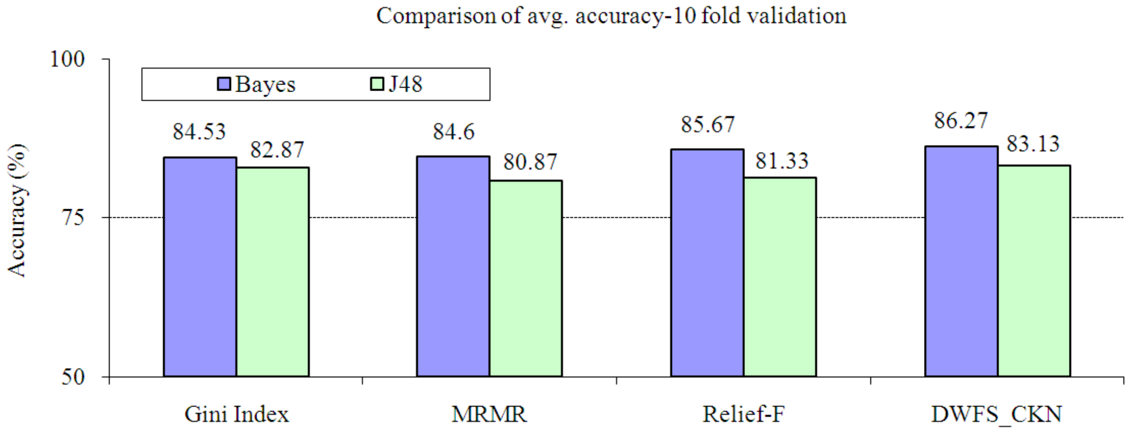

Fig. 3. The average accuracy found through the average of 25 runs of k fold cross validation (k = 10)

Table 1. Comparison between Bayes classifier and J48 classifier with respect to 50 features using LOOCV

	Bayes (%)		J48 (%)	
Feature selection methods	Accuracy	Error	Accuracy	Error
Gini Index	85.48	14.52	83.87	16.13
MRMR	85.48	14.52	82.26	17.74
Relief-F	85.48	14.52	83.87	16.13
DWFS-CKN under study	87.10	12.90	83.87	16.13

Table 2. 10-Fold cross validation using 50 Features-the average, maximum and minimum of 25 iterations

	Bayes (%)				J48 (%)			
Feature Selection Methods	Average Accuracy	Average Error	Maximum Accuracy	Minimum Error	Average Accuracy	Average Error	Maximum Accuracy	Minimum Error
Gini Index	84.53	15.47	86.67	13.33	82.87	17.13	86.67	13.33
MRMR	84.60	15.40	86.67	13.33	80.87	19.13	88.33	11.67
Relief-F	85.67	14.33	86.67	13.33	81.33	18.67	90.00	10.00
DWFS-CKN under study	86.27	13.73	88.33	11.67	83.13	16.87	87.13	12.87

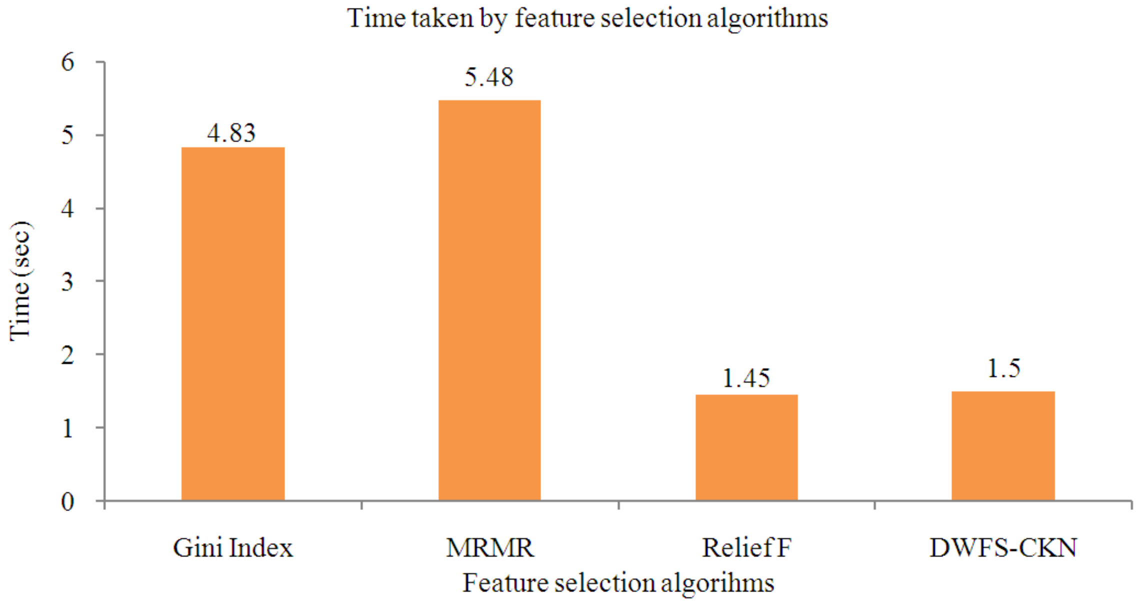

Fig. 4. The time taken for feature selection

Table 3. The top 10 primary features according to different algorithms

Feature selection method	Time taken (sec)	Index of the first 10 selected features
Gini Index	4.83	1671, 249, 493, 765, 1423, 513, 1771, 245, 267, 1772
MRMR	5.48	1671, 249, 493, 765, 1772, 625, 1042, 1423, 513, 1771
Relief-F	1.45	267, 245, 249, 1423, 822, 765, 1892, 66, 493, 897
Proposed DWFS_CKN	1.50	249, 1671, 1423, 513, 765, 245, 267, 493, 1892, 415

4. CONCLUSION

In this study the authors addressed simple, fast, effective and an efficient feature selection algorithm called DWFS_CKN under study and compared its performance with three other classical feature selection algorithms using a complex microarray dataset. The performance of the proposed algorithms have shown improved performance in terms of accuracy of the feature size, consumed less time and the classification accuracy of the DWFS-CKN was better than the three existing algorithms.

4.1. Future Work

Based on the study, the performance, characteristics and the accuracy of the feature selection algorithms, still there are possibilities to advance the performance of the proposed DWFS_CKN algorithm by using appropriate distance calculation procedure to find more and more noticeable features. This study results are in the hands of the future researchers.

5. REFERENCES

Dai, J.J., L. Lieu and D. Rocke, 2006. Dimension reduction for classification with gene expression microarray data. Stat. Applic. Genet. Mol. Biol. PMID: 16646870

Gheyas, I.A. and L.S. Smith, 2010. Feature subset selection in large dimensionality domains. Patt. Recogn., 3: 5-13. DOI: 10.1016/j.patcog.2009.06.009

Hang, X. and F.X. Wu, 2009. Sparse representation for classification of tumors using gene expression data. J. Biomed. Biotechnol., 2009: 403689-403694. DOI: 10.1155/2009/403689

Huang, D., T.W.S. Chow, E.W.M. Ma and J. Li, 2005. Efficient selection of discriminative genes from microarray gene expression data for cancer diagnosis. IEEE Trans. Circ. Syst., 52: 1909-1918. DOI: 10.1109/TCSI.2005.852013

Jeyachidra, J. and M. Punithavalli, 2012. An evaluation of the performance and characteristics of feature selection algorithms using gene microarray dataset. Eur. J. Scient. Res., 93: 214-225.

Jeyachidra, J. and M. Punithavalli, 2013. An investigation into the impact of the feature subset selection methods for classification of gene expression profiles of microarray dataset. Int. J. Scient. Eng. Res., 4: 2395-2402.

Kira, K. and L.A. Rendell, 1992. A practical approach to feature selection. Proceedings of the 9th International Workshop on Machine Learning, Jul. 1-3, Morgan Kaufmann Publishers Inc. San Francisco, CA, USA., pp: 249-256.

Osareh, A. and B. Shadgar, 2010. Microarray data analysis for cancer classification. Proceedings of the 5th International Symposium on Health Informatics and Bioinformatics, Apr. 20-22, IEEE Xplore Press, pp: 125-132. DOI: 10.1109/HIBIT.2010.5478893

Peng, H., F. Long and C. Dong, 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance and min-redundancy. IEEE Trans. Patt. Anal. Mach. Intell., 27: 1226-1238. DOI: 10.1109/TPAMI.2005.159

Quinlan, J.R., 1993. C4.5: Programs for Machine Learning. 1st Edn., Morgan Kaufmann, San Mateo, ISBN-10: 1558602380, pp: 302.

Rejani, Y.I.A. and S.T. Selvi, 2009. Early detection of breast cancer using SVM classifier technique. Int. J. Comput. Sci. Eng., 1: 127-130.

Wang, L., F. Chu and W. Xie, 2007. Accurate cancer classification using expressions of very few genes. IEEE/ACM Trans. Comput. Biol. Bioinform., 4: 40-52. DOI: 10.1109/TCBB.2007.1006

Wang, Z. and V. Palade, 2007. A comprehensive fuzzy-based framework for cancer microarray data gene expression analysis. Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering, IEEE Xplore Press, pp: 1003-1010. DOI: 10.1109/BIBE.2007.4375680

Yeh, J.Y., T.S. Wu, M.C. Wu and D.M. Chang, 2007. Applying data mining techniques for cancer classification from gene expression data. Proceedings of the International Conference on Convergence Information Technology, Nov. 21-23, IEEE Xplore Press, Gyeongju, pp: 703-708. DOI: 10.1109/ICCIT.2007.153