International Journal of Information Systems and Social Change

January-March 2012, Vol. 3, No. 1

Table of Contents

Special Issue on IT Enabled Business Management

Guest Editorial Preface

Research Articles

Breast Cancer Data Prediction by Dimensionality Reduction Using PCA and Adaptive Neuro Evolution

R. R. Janghel, Indian Institute of Information Technology and Management, Gwalior, India

Ritu Tiwari, Indian Institute of Information Technology and Management, Gwalior, India

Rahul Kala, Indian Institute of Information Technology and Management, Gwalior, India

Anupam Shukla, Indian Institute of Information Technology and Management, Gwalior, India

ABSTRACT

In this paper a new approach for the prediction of breast cancer has been made by reducing the features of the data set using PCA (principal component analysis) technique and prediction results by simulating different models namely SANE (Symbiotic, Adaptive Neuro-evolution), Modular neural network, Fixed architecture evolutionary neural network (F-ENN), and Variable Architecture evolutionary neural network (V-ENN). The dimensionality reduction of the inputs achieved by PCA technique to an extent of 33% and further different models of the soft computing technique simulated and tested based on efficiency to find the optimum model. The SANE model includes maximum number of connections per neuron as 24, evolutionary population size of 1000, maximum neurons in hidden layer as 12, SANE elite value of 200, mutation rate of 0.2, and number of generations as 100. The simulated results reflect that this is the best model for the prediction of the breast cancer disease among the other models considered in the experiment and it can effectively assist the doctors for taking the diagnosis results as its efficiency found to be 98.52% accuracy which is highest.

Keywords: *Breast Cancer, Data Prediction, Fixed Architecture Evolutionary Neural Network (F-ENN), Principal Component Analysis (PCA), Symbiotic, Adaptive Neuro-Evolution (SANE)*

INTRODUCTION

The use of computer technology in medical decision support is now widespread and pervasive across a wide range of medical area, such as cancer research, gastroenterology, hart diseases, brain tumors etc. (Sengur, 2007). Fully automatic malignant and benign breast cancer classification is of great importance for research and clinical studies.

Since the neural networks are sensitive to the number of inputs, as large number of inputs may lead to under-training, immensely

DOI: 10.4018/jissc.2012010101

large training time, loss of generality, inability to model ideal functional surfaces, etc.. In such a context, the solution lies in making effective dimensionality reduction techniques for reducing the number of input attributes, whereby having the least loss of information.

The PCA performs the task of dimensionality reduction. It takes as its input a database that consists of a large number of attributes, and mines out the most interesting attributes or combination of attributes. The resultant attributes may be better suited for solving the concerned problem, and have a smaller dimensionality (Sengur, 2007).

The different models of soft computing are available in the literature and comparisons there comparison has been done. The Hybrid approaches form a very exciting field of work and research for large and complex data sets. In these systems we combine Neural Networks, Evolutionary Algorithms, Fuzzy Logic and heuristics in numerous ways to make a much more efficient system (Janghel et al., 2010).

Evolutionary Computing techniques are search algorithms based on the mechanisms of natural selection and genetics. That is, they apply Darwin's principle of the survival of the fittest (Darwin, 1859) among computational structures with the stochastic processes of gene mutation, recombination, etc. Central to all evolutionary computing techniques is the idea of searching a problem space by evolving an initially random population of solutions such that better - "fitter" – solutions obtained.

The evolutionary algorithm searches for the most productive decision strategies using only the infrequent rewards returned by the underlying system. Together evolutionary algorithms and neural networks offer a promising approach for learning and applying effective decision strategies in many different situations.

Co-evolutionary algorithms are great problem solving tools in such scenarios, where the complete problem of optimization of a highly dimensional fitness landscape may be broken down into multiple sub-problems that together constitute the main problem. The sub-problems are simpler to solve or optimize, and hence aid in providing effective components for the solution of the main problem. It is however important for the different sub-problems to co-operate with each other, in order to enable effective evolution that ultimately results in locating global minima.

Novel neuro-evolution mechanisms called SANE (Symbiotic Adaptive Neuro-evolution) are efficient for sequential decision learning. Unlike most approaches, which operate on a population of neural networks, SANE applies genetic operators to a population of neurons. Each neuron's task involves establishing connections with other neurons in the population to form a functioning neural network. Since no one neuron can perform well alone, they must specialize or optimize one aspect of the neural network and connect with other neurons that optimize other aspects. SANE thus decomposes the search space, which creates a much more efficient genetic search. Moreover, because of the inherent diversity in the neuron population, SANE can quickly revise its decision policy in response to shifts in the domain.

In this paper, we first use PCA for dimensionality reduction of the breast cancer database. The original database consisting of 30 attributes usually becomes very large and complex hence; the dimensionality is been reduced by PCA before giving to the input of SANE. In this problem SANE is a representative of the class of co-evolutionary neural networks. Then we compare the performance of SANE with modular neural network (MNN), fixed architecture evolutionary neural network (F-ENN) and variable architecture evolutionary neural network (V-ENN).

LITERATURE REVIEW

Fully automatic malignant and benign breast cancer classification is of great importance for research and clinical studies. Recent work has shown that classification of breast cancer is possible via supervised techniques such as artificial neural networks and support vector machine (SVM), and unsupervised classification tech-

niques unsupervised such as self-organization map (SOM) and fuzzy c-means combined with feature extraction techniques. Other supervised classification techniques, such as k-nearest neighbors (k-NN), Support Vector Machine (SVM). Another machine learning methods like LDA, NDA, PCA and ANN has used (Maitra & Chatterjee, 2007; Santos-Andre et al., 1999; Abdalla, Deris, Zaki, & Ghoneim, 2008).

Almost all the current models try to develop a global architecture, which is a very complex problem. Although, some attempts made to develop modular networks (Liu & Yao, 1999), in most cases the modules are combined only after the evolutionary process has finished and not following a cooperative co-evolutionary model. Cho and Shimohara (1998) developed a modular neural network evolved by means of genetic programming.

METHODOLOGY

The paper proposes solution to the problem of Breast Cancer. Database of breast cancer microscopic and clinical tests reports are collected. Based on the tests reports our expert system predicts cancer as either Benign or Malignant. The complete database contains 30 input attributes. The number of attributes is too large for the system to perform. As a result, the first step performed is dimensionality reduction. For this, we use PCA to get a reduced database containing less number of input attributes. The second task is to apply different soft computing techniques over the problem. The classification has done using SANE neuro-evolution algorithm. The SANE uses co-evolutionary approaches to find the optimal network architecture of the neural network along with the weights and bias values.

Further, in the paper numerous soft computing models have used for the prediction of breast cancer. The learning for each model takes place through an iterative process of weight adjustments applied to its initial weight after epoch iteration of the learning process. Each neural network has its own training algorithm, as per the model. The details are shown in Figure 1.

PCA Preprocessing

Principal components analysis is a quantitatively rigorous method for achieving the simplification of data set. The method generates a new set of variables, called principal components. Each principal component is a linear combination of the original variables. All the principal components are orthogonal to each other so there is no redundant information (Sengur, 2007). The principal components as a whole form an orthogonal basis for the space of the data.

Step that involved in PCA process includes:

Step 1: Get the data,
Step 2: Subtract the mean,
Step 3: Calculate the covariance matrix,
Step 4: Calculate the eigenvectors and Eigen values of the covariance matrix,
Step 5: Choosing components and forming a feature vector,
Step 6: Deriving the 'newdata' set, and later getting the old data back.

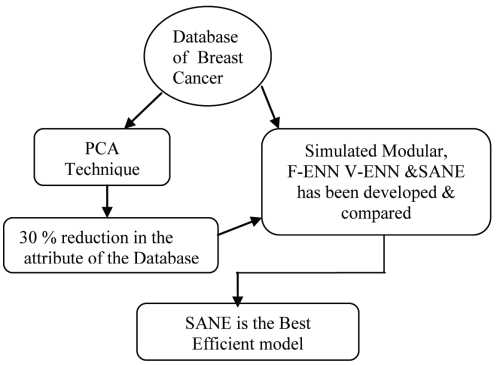
The problem of low-dimensional feature representation can be stated as follows: Let $X=(x_1, x_2,\ldots, x_i,\ldots, x_n)$ represents the $n \times N$ data matrix, where each xi is a feature vector of dimension n. Here n represents the total number of attributes in the breast cancer input and N is the number of input cases in the training set. The PCA can be considered as a linear transformation (1) from the original input vector to a projection feature vector, i.e.

$$Y = W^T X \quad (1)$$

Where Y is the $m \times N$ feature vector matrix, m is the dimension of the feature vector, and transformation matrix W is an $n \times m$ transformation matrix whose columns are the eigenvectors corresponding to the m largest Eigen values computed according to the formula (2):

$$\lambda_{ei} = S_{ei} \quad (2)$$

Figure 1. Methodology of breast cancer prediction



Where e_i, λ are eigenvectors & Eigen values matrix respectively.

Here the total scatter matrix S and the mean of all samples is as follows:

$$S=\sum_{i=1}^{N} (x_i-\mu)(x_i-\mu)^T,$$
$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i . \quad (3)$$

After applying the linear transformation WT, the scatter of the transformed feature vectors $\{y_1,y_2,\ldots, y_N\}$ is W^TSW. In PCA, the projection Wopt is chosen to maximize the determinant of the total scatter matrix of the projected samples, i.e.,

$$W_{opt} = \arg \max_W | W^TSW | = (w_1w_2\ldots w_m) \quad (4)$$

Where $\{w_i \mid i = 1,\ldots. 2, m\}$ is the set of n – dimensional eigenvectors of S corresponding to the m largest eigen values. In other words, the input vector in an n -dimensional space reduced to a feature vector in an m -dimensional subspace. We can see that the dimension of the reduced feature vector m is much less than the dimension of the input faces vector n.

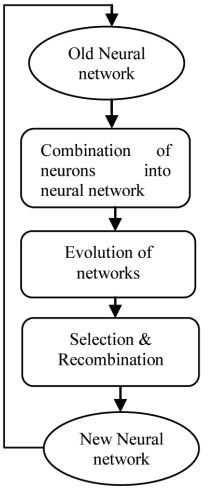
After preprocessing of training database by PCA, feature vectors of reduced dimension are achieved.

Symbiotic Adaptive Neuro-Evolution (SANE)

Symbiotic Adaptive Neuro-Evolution (SANE) evolves three layer networks, where the number of hidden nodes of the network is predefined and fixed, but the network connection topology is not fixed (Moriarty & Miikkulainen, 1997). There are two populations evolved in SANE, a population of neurons and a population of network blueprints. Each individual in the neuron population represents the connection paths and weights of a hidden neuron from the input layer and to the output layer (Smalz & Conrad, 1994). Each gene of an individual contains two parts: one part specifies the connection path (that is which neuron to connect to) and the other specifies the weight of that connection (Whitehead & Choate, 1996).

All the hidden neurons have the same number of connections, but could have different connection paths from the input layer and to the output layer. Networks formed by combining selected individuals from the neuron population. The combined neural information of the network is stored in blue print and individual population represents selected individuals from the neuron population (Hillis, 1990). At the beginning of evolution, combinations created randomly (Paredis, 1994). Effective combination is stored and a new combination created

Figure 2. Conceptual diagram of SANE



by evolving the population (Potter & De Jong, 1995). A well-contributing neuron does not always cooperate well with any other neurons. Therefore, through maintaining a population, well-contributing neurons protected from by elimination due to ineffective cooperation with some other neurons.

SANE is an intra-population neuro-co-evolutionary algorithm that is all the cooperative neurons come from the same population. Each individual of neuron population represents a partial solution instead of a complete network. A complete network formed by a collection of cooperative neuron individuals. The fitness of individual neuron is not evaluated independently but is based on its cooperation. After evaluating all the networks built besides assigning the fitness, each neuron also obtains a fitness value that equals the fitness sum of the best five networks in which the neuron participates. Cooperation only happens when we evaluate individuals, while two populations perform recombination and mutation process independently. The basic block diagram of SANE is as shown in Figure 2.

Algorithm

1. Clear the fitness value of each neuron and blueprint.
2. Select ζ neurons from the population using a blueprint.
3. Create a neural network from the selected neurons.
4. Evaluate the network in the given task.
5. Assign the blueprint the evaluation of the network as its fitness.
6. Repeat steps 2 to 4 for each individual in the blueprint population.
7. Assign each neuron the evaluation of the best five networks in which it participated.
8. Perform crossover and mutation operations on both populations.

During the evaluation stage, each blueprint selects neuron of subpopulations of size ζ to form a neural network. Each blueprint receives the fitness evaluation of the resulting network, and each neuron receives the summed fitness evaluations of the best five networks in which it participated. Calculating fitness from the best

five networks, as opposed to all the neuron's networks, discourages selection against neurons that are crucial in the best networks but ineffective in poor networks.

SIMULATION RESULTS

It is a classificatory problem with two output classes as Malignant and Benign. The features of the database are as follows: radius (mean of distances from center to points on the perimeter), texture (standard deviation of gray-scale values), perimeter, area, smoothness (local variation in radius lengths), compactness (perimeter2 / area - 1.0), concavity (severity of concave portions of the contour), concave points (number of concave portions of the contour), symmetry, fractal dimension ("coastline approximation" - 1). The entire data set consists of 357 benign and 212 malignant cases, totaling to 569 instances in the database.

The SANE algorithm implemented for the Breast Cancer database. The program coded in JAVA. The software uses a data module, which consists of the Breast Cancer data. The data obtained from UCI library as text file is loaded into this module at start. The data divided into training and testing data sets. The SANE algorithm train and tested for different data sets.

The various parameters used for the execution of the algorithm include maximum number of connections per neuron as 24, evolutionary population size of 1000, maximum neurons in hidden layer as 12, SANE elite value of 200, mutation rate of 0.2, and number of generations as 100.

In order to analyze the ability of the system to quickly learn from the training data set and generalize the results over the testing data set, we experiment different percentage breakups of training and testing data sets. The best results have accuracy of 97.88% without PCA and 98.52% with PCA.

The effectiveness of the proposed model in solving the problem, compared with other approaches taken from literature. These include modular neural networks (MNN), fixed architecture evolutionary neural network (F-ANN) and variable architecture evolutionary neural network (V-ANN).

In Modular Neural Networks approach, we tried to divide the entire input space into a set of clusters. Each cluster served as a separate problem domain for diagnosis. Each cluster trained and tested by separate neural network, which forms one of the modules of the complete MNN. Here each neural network used a single hidden layer. The ANN structure used for this problem was (30 x 18 x 1). After training and testing, the system gave accuracy 95.08% without PCA and 96.00% with PCA.

F-ENN (Fixed Architecture Evolutionary Neural Network) we use Genetic Algorithm for the task of training of the neural network. The training involved the setting of weights and biases, to maximize the performance on training data. F-ENN used a single hidden layer structure for this problem (30 x 18 x 1). Genetic Algorithm (GA) applied to structure the parameter optimization. The weight matrix consisted of 30*18 weights between input and hidden layer, 18*1 between the hidden and the output layer and a total of 18 hidden layer biases and 1 output layer bias. This made the total number of variables for the GA as 577. The total number of individuals in the population was 10. A uniform creation function, rank based scaling function and stochastic uniform selection are used. The elite count is two and Single point crossover is used. The program executed until 15 generations. The crossover rate was 0.8. The BPA had a learning rate of 0.05, and momentum of one and the number of training epochs are 30. The accuracy achieved is 95.40% without PCA and 96.56% with PCA.

In variable architecture evolutionary neural network, we tried to optimize the neural network architecture as well as the weights and biases with the help of Genetic Algorithm using connectionist approach. The neural network consists of one hidden layer. In place of an all-connected architecture, few connections are there from input layer to hidden layer and hidden layer to output layer. The information regarding the connections is stored into the

Table 1. Internal architecture of the models

S. No.	Methods	No of Neurons in Input Layer	No of Neurons in Hidden Layer
1	MNN	30	18
2	F-ENN	30	18
3	V-ENN	30	30
4	**SANE**	30	12

Table 2. Comparative analysis of SANE

S.No.	Methods	Accuracy without PCA%	Accuracy with PCA %
1	MNN	95.08%	96.00%
2	F-ENN	95.40%	96.56%
3	V-ENN	95.21%	97.89%
4	**SANE**	**97.88%**	**98.52%**

genetic individual. The parameters of the GA were a maximum number of 30 neurons, 10 as the population size with an elite count of 2. The creation function was uniform and double vector representation done. Rank based fitness selection was used. Stochastic Uniform selection method used and crossover ratio was 0.8. The algorithms run for 15 generations. Extra connections penalized by assigning a penalty of 0.01 per connection. The accuracy in this case has total performance of n 95.21% without PCA and 97.89% with PCA.

Table 1 represents the comparative analyses of all techniques and Table 2 represents the architecture of different soft computing models. It observed that the performance of SANE is the best with 98.52% as compared to the other approaches for the testing data. This naturally proves a good generalizing capability of this co-evolutionary neural network model.

CONCLUSION

A new approach has been attempted and simulated successfully to predict the results of Breast Cancer. Different models like modular neural network, fixed architecture evolutionary neural network (F-ENN) and Variable Architecture evolutionary neural network (V-ENN) and SANE (Symbiotic, Adaptive Neuro-evolution) have been compared SANE gives the maximum efficiency. The whole database divided into training and testing set.

Initially 10% data used for training and 90% for testing and then in following steps of 20/80, 20% training and 80% testing, similarly 30/70, 40/60, 50/50, 60/40, 70/30 and 80/20% for training and testing. The data has been trained and tested for all the models with variation in internal parameters of the simulated model.

The best result obtained using the SANE (Symbiotic, Adaptive Neuro-evolution) approach after the reduction of the attributes of the data set with PCA (principal component analysis) to 98.52% prediction accuracy with PCA and 97.88% without PCA for testing data.

ACKNOWLEDGMENT

The authors sincerely acknowledge the Director ABV-IIITM, Gwalior, India for providing facilities to carry out this research work.

REFERENCES

Abdalla, M. A., Deris, S., Zaki, N., & Ghoneim, D. M. (2008). Breast cancer detection based on statistical textural features classification. In *Proceedings of the 4th International Conference on Innovations in Information Technology* (pp. 728-730).

Cho, S. B., & Shimohara, K. (1998). Evolutionary learning of modular neural networks with genetic programming. *Applied Intelligence*, *9*, 191–200. doi:10.1023/A:1008388118869

Darwin, C. R. (1859). *The origin of species*. London, UK: John Murray.

Hillis, W. D. (1990). Co-evolving parasites improve simulated evolution as an optimization procedure. *Physica D. Nonlinear Phenomena*, *42*, 228. doi:10.1016/0167-2789(90)90076-2

Janghel, R. R., Shukla, A., & Tiwari, R. (2010, June 12-15). Intelligent decision support system for breast cancer. In Y. Tan, Y. Shi, & K. C. Tan (Eds.), *Proceedings of the International Conference on Swarm Intelligence*, Beijing, China (LNCS 6146, pp. 351-358).

Janghel, R. R., Shukla, A., Tiwari, R., & Kala, R. (in press). Diagnosis of breast cancer by modular evolutionary neural networks. *International Journal of Biomedical Engineering and Technology.*

Janghel, R. R., Shukla, A., Tiwari, R., & Tiwari, P. (2009, June 30-July 2). Clinical decision support system for fetal delivery using artificial neural network. In *Proceedings of the IEEE International Conference on New Trends in Information and Service Science*, Beijing, China (pp. 1070-1075).

Liu, Y., & Yao, X. (1999). Ensemble learning via negative correlation. *Neural Networks*, *12*(10), 1399–1404. doi:10.1016/S0893-6080(99)00073-8

Maitra, M., & Chatterjee, A. (2007). Hybrid multiresolution Slantlet transform and fuzzy c-means clustering approach for normal-pathological brain MR image segregation. *Medical Engineering & Physics*, *30*(5), 615–623. doi:10.1016/j.medengphy.2007.06.009

Moriarty, D. E., & Miikkulainen, R. (1997). Forming neural networks through efficient and adaptive coevolution. *Evolutionary Computation*, *5*, 373–399. doi:10.1162/evco.1997.5.4.373

Paredis, J. (1994). Steps towards co-evolutionary classification neural networks. In *Proceedings of the Fourth International Workshop on the Synthesis and Simulation of Living Systems* (pp. 102-108).

Potter, M. A., & De Jong, K. A. (1995). Evolving neural networks with collaborative species. In *Proceedings of the Summer Computer Simulation Conference*.

Rosen, B. E. (1996). Ensemble learning using decorrelated neural networks. *Connection Science*, *8*(3), 373–384. doi:10.1080/095400996116820

Santos-Andre, T. C. S., & da Silva, A. C. R. (1999). A neural network made of a Kohonen's SOM coupled to a MLP trained via backpropagation for the diagnosis of malignant breast cancer from digital mammograms. In *Proceedings of the Joint International Conference on Neural Networks* (pp. 3647-3650).

Sengur, A. (2007). An expert system based on principal component analysis, artificial immune system and fuzzy k-NN for diagnosis of valvular heart diseases. *Computers in Biology and Medicine*, *38*(3), 329–338. doi:10.1016/j.compbiomed.2007.11.004

Smalz, R., & Conrad, M. (1994). Combining evolution with credit apportionment: A new learning algorithm for neural nets. *Neural Networks*, *7*(2), 341–351. doi:10.1016/0893-6080(94)90028-0

Whitehead, B. A., & Choate, T. D. (1996). Cooperative-competitive genetic evolution of radial basis function centers and widths for time series prediction. *IEEE Transactions on Neural Networks*, *7*(4), 869–880. doi:10.1109/72.508930

R. R. Janghel is a PhD student in department of information technology at Indian Institute of Information Technology and Management Gwalior. He has published around 10 papers in various national and international journals/conferences.

Ritu Tiwari is an assistant professor in the IT department of Indian Institute of Information Technology and Management Gwalior. Her field of research includes biometrics, artificial neural networks, signal processing, robotics and soft computing. She has published around 20 papers in various national and international journals/conferences. She has received Young Scientist Award from Chhattisgarh Council of Science and Technology in the year 2006. She also received gold medal in her post graduation from NIT, Raipur.

Rahul Kala is a student of third year integrated post graduate course (BTech + MTech in information communication technology) in Indian Institute of Information Technology and Management Gwalior. His fields of research are robotics, design and analysis of algorithms, artificial intelligence and soft computing. He secured 7th position in the ACM International Collegiate Programming Contest, Kanpur Regionals. He is a student member of ACM. He also secured Al India 8th position in Graduates Aptitude Test in Engineeging-2008 with a percentile of 99.84.

Anupam Shukla is an associate professor in the ICT department of the Indian Institute of Information Technology and Management Gwalior. He has 19 years of teaching experience. His research interest includes speech processing, artificial intelligence, soft computing and bioinformatics. He has published around 62 papers in various national and international journals/conferences. He is referee for 4 international journals and in the editorial board of International Journal of AI and Soft Computing. *He received Young Scientist Award from Madhya Pradesh Government and gold medal from Jadavpur University.*

CALL FOR ARTICLES

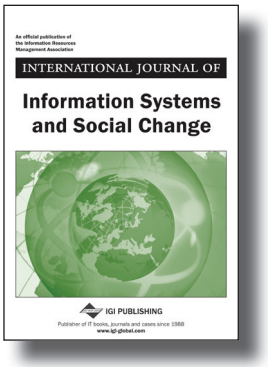
International Journal of Information Systems and Social Change

An official publication of the Information Resources Management Association

The Editor-in-Chief of the *International Journal of Information Systems and Social Change* (IJISSC) would like to invite you to consider submitting a manuscript for inclusion in this scholarly journal.

Mission:

Today's society is fraught with a myriad of constrained opportunities. These challenges are indeed daunting, and the need for change is ever present. The mission of the *International Journal of Information Systems and Social Change* (IJISSC) is to provide an international forum for practitioners and researchers from both the social and natural sciences, along with information systems professionals, software developers, and vendors, to exchange, share, and present useful and innovative ideas and work. It also enables the presentation and distribution of ground-breaking and original ideas and concepts, which can shape future directions of research. These, when applied, can enable policy makers, government officials, business managers, and social workers to spread over various advanced techniques and new applications of information systems. IJISSC also encourages discussions on how information systems can promote social change, and also in turn, how the advances brought about by these social changes can affect the growth in, and future application of, information systems.



ISSN 1941-868X
eISSN 1941-8698
Published quarterly

Coverage:

- Anthropology
- Behavioral science
- Complex adaptive system
- Computer simulation
- Corporate social responsibility
- Decision theory
- Demography
- Disaster management and information systems
- Ecological information systems
- Education information systems
- Environmental information systems
- E-society
- Ethics
- Fuzzy logic
- Geography
- Healthcare information systems
- Infrastructure management system
- Knowledge networks
- Legal information systems
- Natural resource information systems
- Policy development and management
- Political science
- Public-use information systems
- Risk management
- Social informatics
- Social sciences
- Social security information systems
- Socio-technical networks
- Strategic planning
- Urban development information systems
- Web 2.0

All submissions should be e-mailed to:
John Wang, Editor-in-Chief
journalswang@gmail.com

Ideas for Special Theme Issues may be submitted to the Editor-in-Chief.

Please recommend this publication to your librarian. For a convenient easy-to-use library recommendation form, please visit: http://www.igi-global.com/ijissc.