Classification of categorical and numerical data on selected subset of features

Zaher Al Aghbari
Department of Computer Science
University of Sharjah
UAE

1. Abstract

Many Data Mining techniques use the whole features space in the classification process. This feature space might contain irrelevant, or redundant, features that could reduce the accuracy of classification. This paper presents an approach to select a subset of features that are most relevant to the classification application. We use a wrapper approach to search for relevant subset of features, which will be used in the classification of two datasets: categorical teachers' dataset and numerical image dataset. Naïve Bayesian algorithm and *K*-Nearest Neighbor algorithm are used to classify and estimate the accuracy of the categorical data and numerical data, respectively. The experimental results for both categorical and numerical datasets indicate that classification accuracy is improved by removing the irrelevant features and using only the relevant subset of the feature space.

Key-Words: - Data mining, classification, feature selection, wrapper approach, image classification, and categorical data classification.

2. Introduction

Numerous data mining and machine learning applications employ feature selection techniques to improve their performance accuracy and efficiency. Instead of performing the task on the whole input feature set, these applications optimize the solution by selecting only the relevant subset of features, discarding the irrelevant ones, and perform the task on the selected subset of features. As a result, the running time cost of the system is reduced. However, minimizing the set of features may lead to degradation of classification accuracy. Thus, it is essential to include sufficient number features to achieve comparable, or better, classification accuracy as compared to including the whole input set of features.

If we assume that the whole set of features is S, then for some applications it is necessary to develop an algorithm that can reduce the set of features S to a subset of relevant features F, where F < S. Those eliminated features are irrelevant, or redundant, features and can negatively contribute to the classification accuracy. Therefore, before performing the classification task, the relevant subset of features should be searched for. There are two methods to search for the relevant features. In the first method, the search can be performed based on prior knowledge of the feature space and the targeted results; however, this method is subjective and based on the user's intuition and it difficult to apply the same method to different applications (John et. al, 1994). In the second method, a heuristic algorithm is developed to automatically select a subset of features, F, from the whole set of features, S, that will be sufficient to improve accuracy. However, with a moderate size of S, the number of subsets to be considered grows exponentially with the number of features S (Guyon & Elisseeff, 2003). There are two heuristic approaches in the literature to select the relevant subset of features: filter approach and wrapper approach.

The filter approach tries to find a subset of features independently of the inductive algorithm that will use this subset in classification. This is achieved by applying some statistics to select strong relevant features and filter out the weak relevant ones before executing the classification algorithm. In contrast, wrapper approach searches for subsets of features using cross-validation and compares the performance of the classification algorithm with each tested subset in order to select the optimal one. Although the wrapper approach achieves better classification performance compared to filter approach, it requires more time for computations (Guyon & Elisseeff, 2003). The filter approach emphasizes the discovery of relevant features that maximizes the classificaiton accuracy, while the wrapper approach searches for relevant features that minimizes the classification error (Lui & Kender, 2003).

Some scientific applications, such as fusion physics and remote sensing, necessitate the use of feature selection algorithms (Cantu-Paz et al., 2004). In fusion physics, the goal of scientists is not to build a predictor but to identify which features are related with an interesting state of the plasma. In remote sensing, feature selection algorithms are used to automate the identification of human settlements in satellite imagery, which is an essential step in the production of maps of human settlements that are used in studies of urbanization, population movement, etc.

In this paper, we present an approach to select a subset of features that are most relevant to the classification application. We use the Sequential Forward Selection algorithm (SFS) in a wrapper approach to search for relevant subset of features. The selected subset of features will be used in the classification of two datasets: categorical teachers' dataset and numerical image dataset. Naïve Bayesian algorithm and K-Nearest Neighbor algorithm are used to classify and estimate the accuracy of the categorical data and numerical data, respectively.

In Section 2, we survey the related work to classification based on selected subset of features. Then, in Section 3, we present the algorithms for searching a subset of features and the classification algorithms. In Section 4, we present our experimental results. Finally, we conclude the paper in Section 5.

3. Related Work

Developing heuristic algorithms that efficiently searches the space of features and selects the best subset that maintains the same or better performance was a field of research for the past

4 decades. One of the most common feature selection algorithms is genetic algorithms (GA) (Holland, 1975; Laanaya et al., 2005; Vafaie & Imam, 1994; Hao et al., 2003). In GA, a population of candidate solutions of selected subsets of features is always maintained. Candidate solutions are sometimes named as individuals, chromosomes, etc. Each individual is an encoded representation of features of the problems at hand. Each feature in an individual is termed as Gene. The evolution starts from a population of completely random individuals and happens in generations. In each generation, the fitness of the whole population is evaluated; multiple individuals are stochastically selected from the current population based on their fitness, mutated or recombined to form a new population, which becomes current in the next iteration of the algorithm (Laanaya et al., 2005). This generalization process is repeated until a termination condition is achieved such as a solution that satisfies minimum criteria is found, which could be fixed number of generations is reached.

Another feature selection algorithm is called importance score, which is based on greedy-like search (Vafaie & Imam, 1994). The algorithm is based on determining the importance score of each feature using a fitness function and then it performs a greedy-like search to obtain the minimum set of features that maximizes the recognition of some learned rules.

Secquential backwork elimination (SBE) (Marill & Green, 1963) and Sequential forward selection (SFS) (Whitney, 1971) are greedy wrappers used to select the relevant subset of features. SBE start the search with a full set and in each iteration it examines all subsets by removing one feature and retains the subset the gives the highest accuracy as a basis for the next iteration. On the other hand, SFS starts with an empty set and in every iteration it adds one feature to the subset. The search terminates after the accuracy of the current subset cannot be improved by removing (in case of SBE), or adding (in case of SFS), any other feature. However, the drawback of SFS is that once a feature is selected it cannot be removed even if its removal will increase performance accuracy. Similary, in SBE, once a feature is removed it cannot be included even if its inclusion will increase performance accuracy.

A recent algorithm called Basic Sort-Merge Tree (BSMT) (Lui & Kender, 2003) is proposed to choose a very small subset of features. BSMT can be divided into two parts: the creation of a tree of feature subsets, and the manipulation of the tree to create a feature subset of desired cardinality or accuracy. Each part uses a heuristic greedy method. The algorithm reduces the cardinality of the input data by sorting the individual features by their effectiveness in categorization, and then merging these features pairwise into feature sets of cardinality two. Repeating this Sort-Merge process several times results in a subset of features that is efficient and accurate, which is then used in the classification process.

A memory-based algorithm, called leave-one-out cross validation (LOOCV) (Moore & Lee, 1994) employs backward and forward hill-climbing techniques to search for the best subset of features without having to exhaustively evaluate all possible subsets.

In this paper, we present SFS based search algorithm that avoids evaluating all possible subsets of features in a wrapper approach and then the selected subsets of the categorical dataset are classified by a Naïve Bayes algorithm and the selected subsets of the numerical datasets are classified by a K-Nearest Neighbor algorithm.

4. Feature Selection

There are two major approaches, namely the filter approach and wrapper approach, to select the relevant subset of features that will improve system performance in terms of cost and accuracy. As compared to the filter approach, the wrapper approach improves the system performance by reducing the classification error. However, the wrapper approach requires more computations (Guyon & Elisseeff, 2003).

4.1 Feature Subset Selection

Selecting a subset of features has many potential benefits for classification applications:

- Reduces dimensionality to improve classification.
- Reduces compuatational cost and storage requirements.
- Reduces training time.
- Facilitates data understanding.

A simple greedy algorithm called Sequential Forward Selection SFS was proposed by Whitney in 1971 (Whitney, 1971) to search for the best subset of features. SFS (see below) starts with an empty feature subset (see line 1). In each iteration, one feature is added to the feature subset. To determine which feature to add, the algorithm tentatively adds to the candidate feature subset one feature that is not already selected and tests the accuracy of a classifier built on the tentative feature subset. The feature that results in the highest accuracy is added to the feature subset (lines 3-8). If we have added all the features or there is no improvement accrued from adding any further features, the search stops and returns the current set of features (line 9). This algorithm returns a single solution which contains the same selected subset of features on a given problem at every run. As shown in Fig. 1, the SFS algorithm takes as input the whole set of input features and returns the relevant subset of features.

Algorithm: SFS

Input: whole set of input features, S

Output: best subset of features, F

```
1)    Let current subset, $F = \phi$
2)    While size of $F < \tau$, where $\tau$ is the maximum allowed size of $F$.
3)        for each $f \in S$
4)            set $F' \leftarrow f \cup F$
5)            evaluate $F'$ and keep result
6)            set $F \leftarrow F'$ of best result
7)            set $S \leftarrow S - f$ of best result
8)            keep evaluation result of current F
9)    Return $F$
```

4.2 Subset Selection for Categorical Data

Bayes theory is a statistical method that measures the probability of a record in belonging to different classes. A method called Naïve Bayesian classifier (NB-Classifier), which is based on Bayes theory (Tan et al., 2006), is used to measure the accuracy of classification of our categorical teachers dataset into three classes: assistant professor, associate professor and full professor. Each record in the staff dataset consists of six features: name, age, nationality, salary, number of research works, and number of advisees.

The NB-Classifier is particularly suited when the dimensionality of the inputs is high. Despite its simplicity, NB-Classifier can often outperform more sophisticated classification methods. The NB-Classifier requires only one scan of the training data. Furthermore, it can easily handle missing values by simply omitting their probabilities when calculating the likelihoods of membership in each class. This method handles discrete values; however, if an attribute has continuous data, such as salary, these continuous values are divided into ranges. The ranges we used in our experiment are presented in Section 4. Table 1 summarizes the major notation used in this Section and subsequent sections.

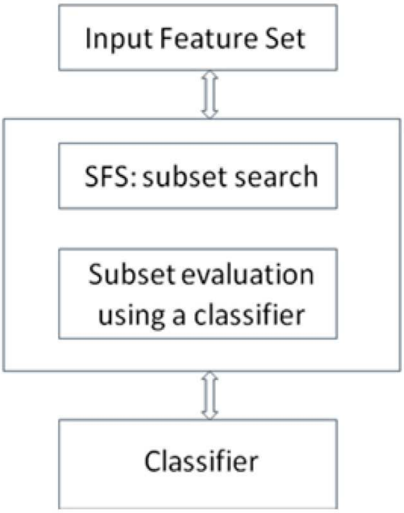


Fig. 1. feature subset search and evaluation using a wrapper approach

The *NB*-Classifier can be viewed as a specialized form of a Bayesian network, termed naïve because it relies on two important simplifying assumptions: independence and normality. That is it assumes that the predictive features x_{ik} of an observed staff record x_i are conditionally independent given the class c_j. These assumptions support very efficient algorithms for both learning and classification. An *NB*-classifier is often represented graphically as shown in Fig. 2, where the direction of the arrows state that the predictive attributes x_{i1}, x_{i2}, …, x_{in} are conditionally independent given the class c_j.

symbol	description
X	A set of s observed records $X = x_1, x_2, \ldots, x_s$
x_{ij}	feature j of the observed record x_i
C	A set of m classes $C = c_1, c_2, \ldots, c_m$
$P(c_i)$	Prior probability associated with class c_i
$P(x_i)$	Probability of occurrence of record x_i
$P(x_i \mid c_j)$	Conditional probability that given class c_j the record x_i satisfies it
$P(c_j \mid x_i)$	Posterior probability that estimates the probability of c_j given x_i

Table 1. summary of notation used in this paper

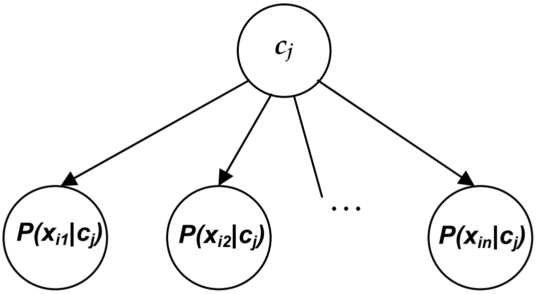


Fig. 2. A Bayesian network that represent the NB-classifier.

Let a set of classes $C = c_1, c_2, \ldots, c_m$ denote the classes of the observed staff records (training set) $X = x_1, x_2, \ldots, x_s$. Consider each observed record x_i as a vector of random variables denoting the predictive feature values $x_{i1}, x_{i2}, \ldots, x_{in}$. Then, given a test instance x to be classified, first, using Bayes rule (Eq. 1) we compute the posterior probabilities of each class and then predict the class with the highest probability as the class of x.

$$P(c_j \mid x_i) = \frac{P(x_i \mid c_j)P(c_j)}{P(x_i)} \tag{1}$$

From the training set, $P(c_j)$ is computed by counting the number of occurrences of c_j. For each feature x_{ik}, the number of occurrences is counted to determine $P(x_i)$. Similarly, assuming categorical features, the probability $P(x_i \mid c_j)$ can be estimated by counting how often each value x_{ik} occurs in the class in the training set.

Since a staff record has *n* independent features, we compute $P(x_{ik} | c_j)$ for every feature and then estimate $P(x_i | c_j)$ by the conjunction of all conditional probabilities of the features as shown in Eq. 2.

$$P(x_i \mid c_j) = \prod_{k=1}^{n} P(x_{ik} \mid c_j) \qquad (2)$$

The posterior probability, Eq. 1, is estimated for every class and then predict the class with the highest probability as the class of the test instance *x*. The *NB*-classifier is simple and efficient approach to classify new staff record instances.

4.3 Subset Selection for Numerical Data

We use the K-Nearest Neighbor (*KNN*) algorithm to classify the numerical dataset (image dataset) using the selected subset of features. Each image is represented by a feature vector of size 64 and an image may belong to one of the following 12 classes: beach, garden, desert, snow, sunset, rose, banana, tomato, copper, tiger, wood, and gorilla.

KNN algorithm measures the classification accuracy of the selected subset of feature based on a distance function, *d(q, p)*, Eq. 3, where $p: p_1, p_2, \ldots, p_d$ and $q: q_1, q_2, \ldots, q_d$ are two vectors representing two images.

$$\text{D(q , p)} = \left(\sum_{i=1}^{d} \left| q_i - p_i \right|^2 \right)^{0.5} \qquad (3)$$

Algorithm: K-Nearest Neighbor

Input: t_{DB}, *K*, I_Q

Output: *Class to which* I_Q *is assigned*

```
L_K = 0
for each t ∈ t_DB do
    compute d(t, I_Q) using Eq. 3
    if L_K contains < K items
        L_K = L_K ∪ t
    else if d(t, I_Q) < d(I_Q , K^th )
        L_K = L_K - K^th
        L_K = L_K ∪ t
Assign t to the majority class in L_K
```

Generally, the *KNN* algorithm works as follows:

- A number of images are prepared to be the training dataset. We performed stratified sampling to build the training dataset, which are representative images from all the pre-defined classes. These representative images in the training set include the class information.

- The algorithm maintains an ascending order list LK that keeps the K nearest neighbors found so far.
- Each image in the database IQ is then compared with each image t in the training set tDB by computing their Euclidean distance (Eq. 3).
- If the list LK contains less than K images from the training dataset, add image t into the list, otherwise, if the distance between IQ and image t is less than the distance between IQ and the Kth neighbor in LK, remove the Kth neighbor from the LK and add t to LK.
- Finally, the query image IQ is classified to the majority class in the retrieved K nearest images in LK.

5. Experimental Results

We performed our experiments on two datasets: categorical teachers' dataset and numerical image dataset. The first part measures the accuracy of classification of our categorical teachers dataset into three classes: assistant professor, associate professor and full professor. Each record in the staff dataset consists of six features: name, age, nationality, salary, number of research works, and number of advisees.

The *NB*-Classifier technique handles discrete values. If a feature value type is continuous, such as salary, these continuous values are discretized by dividing it into ranges. Each of the three classes is given a value that is assistant professor class is assigned 1, associate professor is assigned 2 and full professor is assigned 3. Before the *NB*-Classifier tests features, these features need to be encoded first. In our experiment, we encoded the ages between 20 and 30 as 1, those between 31 and 40 as 2, those between 41 and 50 as 3, and the ages above 51 are encoded as 4. Salary is divided into ranges as follows: 10,000-20,000 as 1, above 20,000-30,000 as 2, above 30,000-40,000 as 3, above 40,000-50,000 as 4, above 50,000-60,000 as 5, above 60,000-70,000 as 6, and above 70,000 is encoded as 7. Number of research is encoded as 1 for 0-25, 2 for 26-50, and 3 for 51 researches and above. Similarly, number of advisees was divided into ranges as follows: 1-5 as 1, 6-10 as 2, 11-15 as 3, and so on. The name and nationality features were not considered in our experiment as they are clearly irrelevant.

A training staff dataset of size 50 records was prepared and used to search for the subset of features and evaluate them. As seen in Table 2, in the first iteration, the number of research works gave the best classification accuracy among all other features and thus it was used as a basis for second iteration. In the second iteration, the subset of number of research works and salary gave the best classification accuracy. In the third iteration, the technique evaluated all possible subsets by adding another unselected to the basis from the second iteration; however, the best subset in this iteration did not give classification accuracy higher than the subset found in the second iteration. Therefore, the search was stopped and the best subset of features found in the second iteration was returned.

Iteration Number	Best Subset of Features	Classification Accuracy
1	{number of research works}	0.67
2	{number of research works, salary}	0.87
3	{number of research works, salary, age}	0.85

Table 2. subset of features of categorical data

In the second part, several experiments were performed using 419 images. The accuracies of all possible combinations of 64 features, which represent the images, are found by SFS and measured by *KNN* classifier.

First, using stratified sampling, a sample of 60 images was selected, five different images from each class were chosen. *KNN* algorithm estimates the class of each image by selecting ten nearset neighbors. All subset of features were evaluated and their accuracies were measured by SFS when training images are included in evaluation and when they are excluded.

Dataset Size	Training Dataset Included	Size of selected subset	Best Classification Acurracy
60	Yes	51	0.35
	No	36	0.67
139	Yes	52	0.67
	No	37	0.79
419	Yes	56	0.74
	No	39	0.80

Table 3. subset of features of image data

Similar experiments were performed on different image dataset sizes; that is using 139 images from the set of 419. The training dataset in each experiment was chosen to be 60% of the dataset size. The number of training images selected from each class is propotional to the number of images in the class and these training images are selected randomly. Table 3 summaries the results of the three experiments. Notice that as we include the training images in the testing phase, the classification accuracy increase, which is expected because the system will be able to correctly classify those images used in the training. Also, note that as the training image dataset is included in the testing phase the size of the subset of features is reduced.

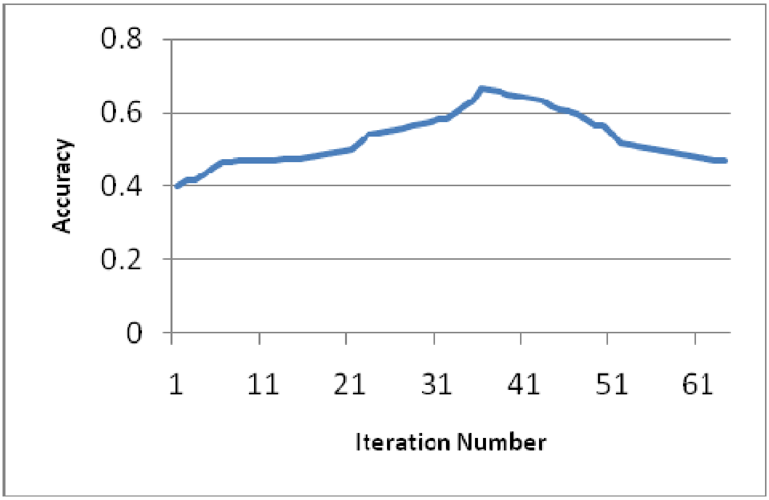


Fig. 3. Classification accuracy of the tested subsets of features.

The highest classification accuracy of the subset of features (with the training image dataset included in the test) at each iteration is shown in Fig. 2 for the image dataset of size 60. The X-axis represents the iteration number and the Y-axis represents the classification accuracy. Note that in the classification accuracy increases as more features are added to the basis of previous iteration till it reaches a peak at which the system had found the best subset of features, then as more feauters are added the accuracy degrades. The other datasets depict the same trend. The highest classification accuracy in Fig. 3 was reached for the following subset of feature numbers: {6, 3, 23, 9, 24, 32, 33, 12, 13, 14, 19, 20, 29, 30, 35, 36, 40, 41, 45, 50, 51, 56, 25, 49, 34, 61, 37, 52, 53, 57, 59, 17, 21, 42, 55}. The order of the features in the subset depicts the order of their inclusion in subset, which is based on their contributions to classification accuracy.

6. Conclusion

In this paper, we presented a wrapper approach to select the best subset of features that result in the highest classification accuracy. We use an SFS approach to search for the best subset of features. The Naïve Bayes algorithm and *K*-Nearest Neighbor algorithm are used to classify and estimate the accuracy of the categorical data and image data, respectively. This approach is evaluated using two datasets: categorical teachers' dataset and image dataset. The experimental results for both categorical and image datasets show the feasibility of the presented techniques in classifying categorical and numerical data. Such techniques are useful in many applications to decrease the performance cost and increase the classification accuracy.

7. References

John, G. H.; Kohavi, R. & Pfleger, K. (1994). Irrelevant Features and the Subset Selection Problem, In *Proceedings of the International Conference on Machine Learning,* pp. 121-129.

Guyon, I. & Elisseeff, A. (2003). Overfitting in Making Comparisons Between Variable Selection Methods, *Journal of Machine Learning Research* , vol. 3, pp. 1371-1382.

Cantu-Paz, E.; Newsam, S. & Kamath, C. (2004). Feature Selection in Scientific Applications, *Proceedings of International Conference on Knowledge Discovery and Data Mining*, Seattle, WA, United States.

Liu, Y. & Kender, J. R. (2003). Sort-Merge Feature Selection for Video Data, *SIAM Data Mining Conference (SDM)*, San Francisco, USA.

Holland, J. (1975). Adaptation in Natural and Artificial Systems, *University of Michigan Press*, Ann Arbor, MI., USA.

Laanaya, H.; Martin, A.; Khenchaf, A. & Aboutajdine, D. (2005). Feature Selection Using Genetic Algorithms For Sonar Images Classification With Support Vector, *ECPS* Conference, 15-18 March, Brest, France.

Vafaie, H. & Imam, I.F. (1994). Feature selection methods: Genetic algorithms vs. greedy-like search, *Proceedings of International Conference on Fuzzy and Intelligent Control Systems*.

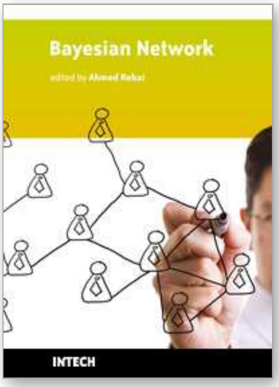
Hao, H.; Liu, C. & Sako, H. (2003). Comparison of Genetic Algorithm and Sequential Search Methods for Classifier Subset Selection, *Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR'03)*, vol. 2, pp. 765-769.

Marill, T. & Green, D.M. (1963). On the Effectiveness of Receptors in Recognition Systems, *IEEE Trans. on Information Theory*, vol. 9, pp.11-17.

Whitney, A.W. (1971). A Direct Method of Nonparametric Measurement Selection, *IEEE Trans. on Computers*, vol. 20, no. 9, pp. 1100-1103.

Moore, A.W. & Lee, M.S. (1994). Efficient Algorithms for Minimizing Cross Validation Error, *in Cohen, W. and Hirsh, H. eds., Machine Learning: Proceedings of the 11th International Conference, Morgan Kaufmann*, 1994.

Tan, P-N.; Steinbach, M. & Kumar, V. (2006). Introductin to Data Mining, *Addison Wesley*.



Bayesian Network
Edited by Ahmed Rebai

ISBN 978-953-307-124-4
Hard cover, 432 pages
Publisher Sciyo
Published online 18, August, 2010
Published in print edition August, 2010

Bayesian networks are a very general and powerful tool that can be used for a large number of problems involving uncertainty: reasoning, learning, planning and perception. They provide a language that supports efficient algorithms for the automatic construction of expert systems in several different contexts. The range of applications of Bayesian networks currently extends over almost all fields including engineering, biology and medicine, information and communication technologies and finance. This book is a collection of original contributions to the methodology and applications of Bayesian networks. It contains recent developments in the field and illustrates, on a sample of applications, the power of Bayesian networks in dealing the modeling of complex systems. Readers that are not familiar with this tool, but have some technical background, will find in this book all necessary theoretical and practical information on how to use and implement Bayesian networks in their own work. There is no doubt that this book constitutes a valuable resource for engineers, researchers, students and all those who are interested in discovering and experiencing the potential of this major tool of the century.

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821