Reduction of the Size of Datasets by using Evolutionary Feature Selection: the Case of Noise in a Modern City

Javier Luque, Jamal Toutouh, and Enrique Alba

Departamento de Lenguajes y Ciencias de la Computación,
Universidad de Málaga, Málaga, Spain
javierluque@uma.es,{jamal,eat}@lcc.uma.es

Abstract. Smart city initiatives have emerged to mitigate the negative effects of a very fast growth of urban areas. Most of the population in our cities are exposed to high levels of noise that generate discomfort and different health problems. These issues may be mitigated by applying different smart cities solutions, some of them require high accurate noise information to provide the best quality of serve possible. In this study, we have designed a machine learning approach based on genetic algorithms to analyze noise data captured in the university campus. This method reduces the amount of data required to classify the noise by addressing a feature selection optimization problem. The experimental results have shown that our approach improved the accuracy in 20% (achieving an accuracy of 87% with a reduction of up to 85% on the original dataset).

Keywords: Smart city · Genetic algorithm · Feature selection · Noise

1 Introduction

Smart city has emerged as link of the different stakeholders of the cities to mitigate the negative effects of a very fast growth in urban areas [5]. With the raising of smart cities, countless new applications are appearing to improve our daily life [1, 6, 8, 15, 17]: smart parking, intelligent waste collection, intelligent traffic systems, efficient street lighting, etc. Most of these new applications use intelligent systems able to detect, predict, and efficiently manage different aspects of the city.

The high ambient noise level is an important problem in our present cities, because it causes discomfort, sleep disturbance, reduces cognitive performance, and is responsible of many disease [18]. As road traffic generates about 80% of the noise pollution [14], reducing its noise seems to be an efficient strategy to improve this aspect of the daily life. For this reason, there are many different approaches to measure and evaluate the ambient noise level in the roads [7, 11, 16], e.g,. people carrying a sound-meters to take temporal measures, installing wireless sensor networks (WSN) or, most recently, using smartphones of the inhabitants. Having a better knowledge about the sources of noise is vital to help city managers in taking better decisions.

In this work, we focus on the intelligent analysis of noise data captured by a cyber-physical system (WSN) installed in the University of Málaga's campus [16]. We want to characterize/label the noise level sensed during the day in order to identify the main noise source by using machine learning (i.e., *K-Means Clustering* classification method). In order to do so, we find the periods of time that efficiently classify the days in different groups (clusters) according to the noise levels for these periods. Thus, a feature selection method is applied to select these time periods, since this type of methods provide high competitive results in removing irrelevant data and improving the efficiency and accuracy of the application of machine learning methods [2].

When dividing the day is important to use time periods as small as possible to improve the accuracy of the evaluation of the sensed noise level. However, this critically increases the number of required periods to be evaluated. Therefore, it increases the number of studied features in our feature selection method. In order to deal with feature selection methods with large set of features, classic methods can not be applied. Thus, Genetic Algorithms (GA), as efficient tools to address *hard-to-solve* search and optimization problems, have been successfully applied to address feature selection achieving very competitive results [19, 20].

In this study, we have divided the time line in blocks of 30 minutes. Thus, the size of the set of features is 48, that defines 2^{48} (2.81E14) possible subsets of features. In order to deal with this feature selection problem, we have applied a Genetic Algorithm (GA) to search for the best subset of features to efficiently classify the noise level information. Therefore, the main objective of this paper is to present this machine learning method based on GA and to use it to classify real noise data captured at the university campus.

The rest of this work is structured as follows: Section 2 introduces the noise analysis and presents the noise feature selection problem. Section 3 describes the evolutionary approach designed in this study to address the optimization problem. Section 4 presents the experimental analysis of our proposal. Finally, Section 5 outlines the conclusions and proposes the future work.

2 Smart Campus Efficient Noise Evaluation

In this section, we describe the process to capture noise data and we define the feature selection problem to efficiently classify the data.

2.1 Noise Measurement by the Smart Campus Sensing System

The noise level data analyzed in this study is gathered by the sensing system presented by Touotuh et al. [16]. These sensors measure the ambient noise and detect smart devices with a WiFi or Bluetooth connection. Their principal components are (see Fig. 1): two WiFi and a Bluetooth wireless interfaces, a Real Time Clock (RTC), a noise sound meter, and a Raspberry Pi 3. The global architecture of the system is shown in Fig. 2, where the sensors send information to the data center via Internet by using a client/server model (see Fig. 3).

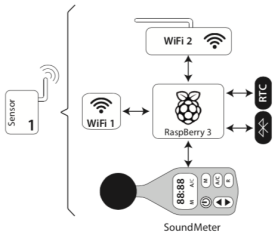

Fig. 1. Hardware scheme.

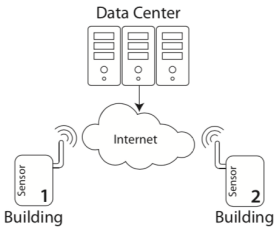

Fig. 2. Architecture.

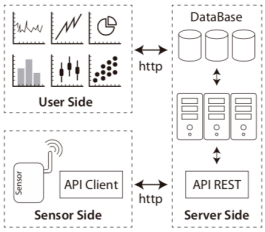

Fig. 3. System model.

The noise sound meter captures the level of noise each second. The sound is evaluated in terms of *equivalent sound pressure level*, i.e. L_{eq}, which expresses the mean of the sound pressure perceived by an individual measured in decibels (dBa) [11] in an interval of time. In our case, L_{eq} is calculated for intervals of one minute. These measures are sent to the server to be stored in the database each hour. Finally, we generate the *daily noise curve* (see Fig. 4) by grouping all the noise level measures of a given day (from 0:00:00H to 23:59:59H).

2.2 The Noise Feature Selection Optimization Problem

The classic feature selection problem over a given dataset consists in selecting a subset of the most relevant data that allows us to characterize the original dataset without losing information [2]. In this work, we want to reduce the data required to accurately classify the daily noise level information.

As it is shown in Fig. 5, the noise levels follow three different patterns corresponding to: the *working days* (from Monday to Friday), *Saturdays*, and *holidays* (including Sundays). *K-Means Clustering* results with K=3 confirmed that these daily noise levels may be grouped in the previously presented three different patterns. This is principally due to the road traffic patterns are mainly dependent on these three types of days, and as it has been stated before, road traffic is the main source of ambient noise in urban areas.

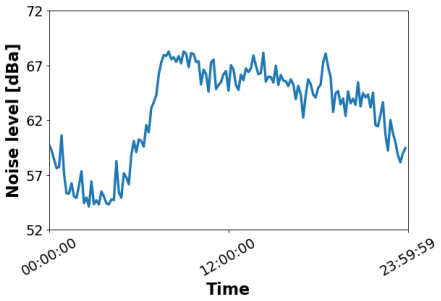

Fig. 4. Example of a daily noise curve of a given day.

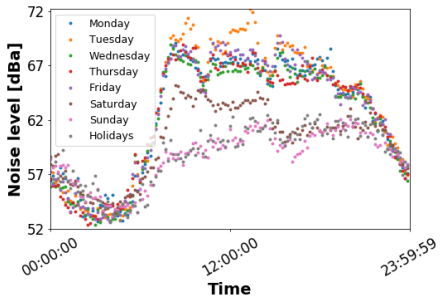

Fig. 5. Level of noise sampled grouped by the weekdays and holidays.

In order to characterize the noise level of a given day, we have calculated the *Area Under the Curve* (AUC) [9]. This method allows to simplify the analysis without loosing the information represented by the daily noise curve. AUC has been employed in different research areas for the same purposes, achieving high accurate results (e.g., medicine [10]).

AUC returns the area between a curve and the *x-axis*. We calculate the AUC by using a number of *trapezoidal approximations* (N), which are delimited by the noise curve, the x-axis and one of the N continuous periods of time. Equation 1 defines this approximation to compute the AUC, where the time line (temporal space) is defined in the range between a=00:00:00H and b=23:59:59H ($x = [a, b]$), $f(x)$ is the daily noise curve, and each time period (*x-step*) is given by $\Delta x = \frac{b-a}{N}$.

$$\int_a^b f(x)\,dx = \frac{\Delta x}{2} \sum_{k=1}^{N} (f(x_{k-1}) + f(x_k)) \tag{1}$$

In the context of the *Noise Feature Selection* (NFS) optimization problem, we deal with the daily noise curve C, divided in an arbitrary daily rate DR (number of time blocks/samples/trapezoids per day). Thus, we define T the set of trapezoids or features that defines the original dataset of C ($\|T\|$=DR), and $ST \subseteq T$ to be a subset of the original set of features. The main target of this work is to find the most efficient (smallest) subset of trapezoids/features (ST) that allows us to maximize the accuracy ($ACC(ST)$) [4], i.e., minimize the error of a given machine learning classification approach (see Equation 2). In this work, we have evaluated our feature selection method over *K-Means Clustering*.

$$Maximize\ ACC(ST)\ subject\ to\ ST \subseteq T \tag{2}$$

3 Noise Feature Selection by Using a Genetic Algorithm

In this section, we describe the approach applied to address NFS optimization problem by using a GA. Thus, we introduce the proposed algorithm, we show the solution representation, we describe the evolutionary operators, and we formulate the objective/fitness function used to evaluate the solutions.

3.1 The Genetic Algorithm

Algorithm 1 shows the pseudocode of the GA. It iteratively applies stochastic operators on a set of solutions (named *individuals*) that define a *population* (P) to improve their quality related to the objective of the problem. Each iteration is called *generation*.

The first steps produce the initial population *P(0)* of $\#p$ individuals (Line 2). An evaluation function associates a fitness value to every individual, indicating its suitability to the problem (Line 4). Then, the search process is guided by a probabilistic technique of selection of the best individuals (parents and generated offspring) according to their quality (Line 5). Iteratively, solutions evolve

Algorithm 1 Generic schema for a GA.

```
1: t ← 0
2: initialization(P(0))
3: while not stopcriterion do
4:         evaluation(P(t))
5:         parents ← selection(P(t))
6:         offspring ← variation operators(parents)
7:         P(t+1) ← replacement(offspring, P(t))
8:         t ← t + 1
9: end while
10: return best solution ever found
```

by the probabilistic application of *variation operators* (lines 6-7). The stopping criterion usually involves a fixed computational effort (e.g., number of generations, number of fitness evaluations or execution time), a threshold on the fitness values or the detection of a stagnation situation.

3.2 Representation

Solutions are encoded as a binary vector $S_o =< i_1, ..., i_N >$, where N is the number of features of the original set. Each index of the vector represents a given feature (period of time to compute the trapezoids), and the corresponding binary value $S_o(i)$ represents the selection *i-th* feature. Thus, if the *i-th* feature is selected, then $S_o(i)$=1, otherwise $S_o(i)$=0.

3.3 Operators

The main evolutionary operators are presented in this section.

Initialization The population is initialized (*P(0)*) by applying a uniform random procedure.

Selection Tournament selection is applied, with tournament size of two solutions (*individuals*). The tournament criteria is based on fitness value.

Evolutionary operators We analyze the application of three different **recombination operators**: the standard one point (1PX) and two points (2PX) crossover [13], and the recombination by *selecting randomly half* (SRH) applied with probability p_C. SRH, which was specifically designed in this study to address NFS, randomly selects a set of features of size $N/2$ and it exchanges them between the two parents to generate two parent offspring solutions. Fig. 6 shows an example of applying SRH to the parents P1 and P2, in order to get the offspring O1 and O2. Finally, we study the application of two different **mutation** operators: the flip bit (FB) and shuffle indexes (SI) mutation [13] used with probability p_M.

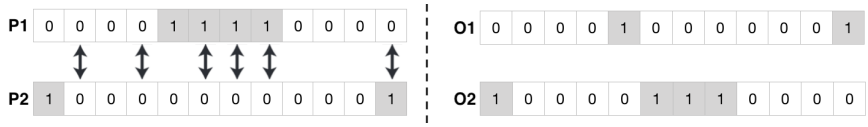

Fig. 6. Example of the SRH crossover operator.

3.4 Fitness Function

The NFS optimization problem is addressed as a minimization problem. Thus, Equation 3 presents the objective/fitness function ($fitness(S)$) used to evaluate the individuals during the evolutionary process, which should be minimized. Each solution S represents a subset of features ST_S and it is used to compute the accuracy measure ($ACC(ST_S) \in \mathbb{R} \subseteq [0,1]$) of using *K-Means* to classify/group the set of AUC daily values in the three day types (see Section 2.2). As it is a minimization problem, the fitness function takes into account the value of 1-$ACC(ST_S)$. Besides, NFS is defined to minimize the size of the ST_s ($\|ST_s\| \leq N$). Thus, we use this metric in the objective function multiplied by an α factor. After some preliminary experiments α was set to 10^{-6}.

$$fitness(S) = (1 - ACC(ST_S)) + \alpha \times \|ST_s\| \quad \in \mathbb{R} \subseteq [0, 1 + \alpha \times N] \quad (3)$$

4 Experimental Analysis

This section presents the experiments carried out to evaluate the proposed optimization problem. The algorithms were implemented in Python, using the evolutionary computation framework DEAP [3] and the Anaconda framework.

The experiments conducted have been run in the cluster belonging to NEO (Networking and Emerging Optimization) group, which consists of 16 nodes (64 cores) equipped with an Intel Core2 Quad CPU (Q9400) @ 2.66 GHz and 4 GB of RAM.

In the following subsections, we define the problem instances, we present the results of configuring the applied GA, and we analyze the final results and discuss the performance of this proposal.

4.1 Problem Instances

In order to evaluate our proposal, we use the noise data sensed during seven weeks (from February 22^{nd}, 2018) by one of the sensors installed in the university campus. Fig. 5 illustrates the sensed data grouping (averaging) them by the type of day. It can be seen that there are three groups: *(a)* working days, that present a very similar pattern, *(b)* holidays and Sundays, that have a clear relation, and *(c)* noise on Saturdays, that has its own particular shape.

We defined two problem instances: the *1-hour data blocks instance* (1-H) and the *30-min data blocks instance* (30-M). The 1-H instance is a low cost instance used to find the best configuration of the operators applied in the proposed GA (see Section 3.3). In this instance, the time line is divided in 24 features, i.e., trapezoids/data blocks of one hour (N=24). The 30-M instance is applied to address the NFS optimization problem. In order to obtain more accurate results than in 1-H instance, the time line is divided in trapezoids/data blocks of 30 minutes, i.e., 48 features (N=48).

4.2 Parameters Calibration

A set of parametric setting experiments were performed over the 1-H instance to determine the best operators and parameter values for the proposed GA. The analysis was carried out by setting the population size ($\#p$) to 50 and the maximum number of generations ($\#g$) to 100.

We analyzed the results of applying three crossover operators (1PX, 2PX, and SRH) with four different probabilities ($p_C \in \{0.2, 0.4, 0.6, 0.8\}$), and two mutation operators (BF and SI) with three different probabilities ($p_M \in \{1/10N, 1/N, 10/N\}$, where N is the number of features). Therefore, we evaluated 72 different parameterizations (three crossover operators, four values of p_C, two mutation operators, and three p_M). All these combinations were studied with the proposed GA, for a total number of 30 independent runs.

We applied Shapiro–Wilk statistical test to check if the results follow a normal distribution. As the test resulted negative, we analyzed them by applying non-parametric statistical tests [12]: first, we used Friedman rank test to rank the configurations, and second, we performed Wilcoxon tests to the three best ranked ones to determine the most competitive configuration.

According to Friedman ranking, the three best configurations were (2PX, 0.6, BF, 10/N), (1PX, 0.8, BS, 10/N), and (SRH, 0.8, BF, 10/N), with p-value$<10^{-10}$. Wilcoxon test did not confirm the differences of comparing these three configurations with each other (p-value$>$0.01). Therefore, we decided to select the parameterization that obtained the best (minimum) median value of these three best ranked ones (2PX, 0.6, BF, 10/N), i.e., the configuration defined by 2PX as crossover operator, p_C=0.6, BF as mutation operator, and p_M=10/N.

4.3 Numerical Results

This subsection reports the numerical results achieved in an exhaustive experimental evaluation of addressing NFS optimization problem over 30-M instance by applying the proposed GA and performing 100 independent executions. The GA configuration is defined by $\#p$=50, $\#g$=200, and the best configuration found in the previous section (2PX, 0.6, BF, 10/N).

Table 1 reports the average, standard deviation, best (minimum), median, and worst (maximum) of the final fitness values, the number of selected features ($\|ST_S\|$), and the accuracy achieved by applying *K-Means* method ($ACC(S)$) over the 100 independent runs. In turn, it shows the result of the accuracy of no applying feature selection (No NFS), i.e., using 48 features.

Table 1. Optimization results: final fitness, number of features, and accuracy.

	Mean±Std	Minimum	Median	Maximum	No NFS
fitness(S)	0.197±13.8%	**0.125**	0.203	0.234	-
$\|ST_S\|$	-	**6**	7	10	48
$ACC(S)$	-	**0.875**	0.797	0.766	0.687

Results in Table 1 clearly state that higher accurate classifications were achieved while the number of features was reduced. All the computed solutions obtained better accuracy and used critically lower number of features than the original dataset, i.e., *No NFS* obtained the worst accuracy (lower than 69%). The best solution found requires six features (data blocks) to achieve a high accuracy in classifying the noise (higher than 87%). The median fitness value was obtained by solutions that use just one feature more (seven). Thus, we can confirm that there are periods of the day in which the ambient noise is definitely different depending on type of the day, as can be infer from Fig. 5.

Finally, we identified the periods of time (features or data blocks) that best characterize the daily noise curve. Thus, we computed how often (the frequency) a given feature has been selected by the best solution found for each independent run, in order to observe the features selected with higher frequency. These results are shown in a histogram (see Fig. 7). This histogram illustrates the frequency of a given feature (*Feature ID*) and the corresponding time. The three most used features for classifying the noise are the ones numbered as 16, 18, and 19, which correspond to the noise level data captured during commercial and school opening hours (from 8:00:00H to 10:00:00H). This time period coincides with the morning peak traffic hours at the campus. Therefore, the main result of our work is that the daily noise level may be accurately classified by activating the noise sensors during just two hours (from 8:00:00H to 10:00:00H). This represents noticeable energy savings.

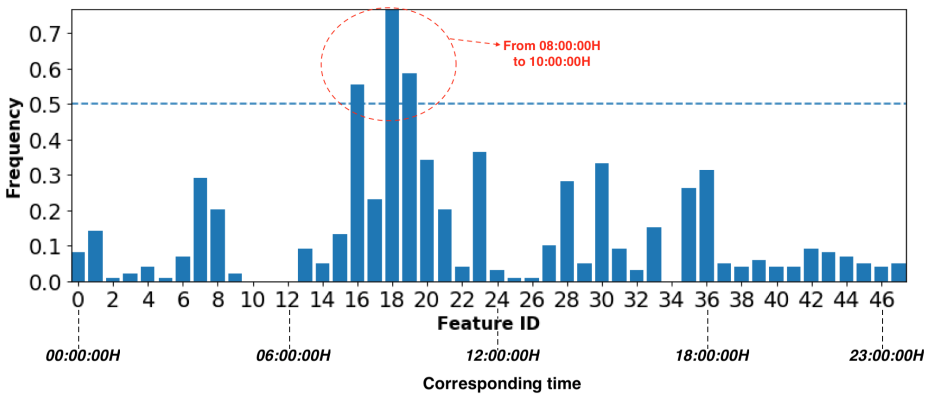

Fig. 7. Histogram of the features selected by the final solutions.

5 Conclusions and Future Work

The ambient noise is an important problem in the present cities, because it principally causes discomfort and it is responsible of many diseases. In this study, we have shown one interesting contribution in the intelligent data analysis of ambient noise level, which can be useful for city managers. The real data used are captured and processed by the cyber-physical system installed at University of Málaga by the authors.

In order to characterize the noise level captured during a day and use such data to classify/group the type of day, we have used the AUC metric as an input to *K-Means clustering* method. We have designed an intelligent and automatic feature selection method based on a GA to improve the accuracy of this classification approach.

As to the numerical findings, we have been able to drastically improve the accuracy of the classification method in about 20% (from less than 69% to higher than 87%), while the number of features were reduced from 48 to 6 (87%). These results have a direct benefit for city managers because we can save lots of time and energy by only measuring at some moments of the day. In this case, we have observed that the most characteristic period of time in the evaluated sensor is during the morning from 8:00:00H to 10:00:00H, which coincide with morning peak traffic hours.

As future work we plan to extend the proposed methodology by extending this approach by adding more noise data from other sensors, analyzing this approach over other datasets (road traffic data), and studying new evolutionary approaches as well as new machine learning methods.

Acknowledgements

This research has been partially funded by the Spanish MINECO and FEDER projects TIN2016-81766-REDT (http://cirti.es), and TIN2017-88213-R (http://6city.lcc.uma.es). University of Malaga. International Campus of Excellence Andalucia TECH.

References

1. Camero, A., Toutouh, J., Stolfi, D.H., Alba, E.: Evolutionary Deep Learning for Car Park Occupancy Prediction in Smart Cities. In: Learning and Intelligent OptimizatioN (LION) 12. pp. 1–15. Springer (2018)
2. Chandrashekar, G., Sahin, F.: A survey on feature selection methods. Computers & Electrical Engineering **40**(1), 16–28 (2014)
3. Fortin, F.A., De Rainville, F.M., Gardner, M.A., Parizeau, M., Gagné, C.: DEAP: Evolutionary algorithms made easy. Journal of Machine Learning Research **13**, 2171–2175 (jul 2012)
4. García, S., Fernández, A., Luengo, J., Herrera, F.: A study of statistical techniques and performance measures for genetics-based machine learning: accuracy and interpretability. Soft Computing **13**(10), 959 (2009)

5. McClellan, S., Jimenez, J.A., Koutitas, G.: Smart Cities: Applications, Technologies, Standards, and Driving Factors. Springer (2017)
6. Mir, Z.H., Toutouh, J., Filali, F., Alba, E.: Qos-aware radio access technology (RAT) selection in hybrid vehicular networks. In: International Workshop on Communication Technologies for Vehicles. pp. 117–128. Springer (2015)
7. Murphy, E., King, E.A.: Testing the accuracy of smartphones and sound level meter applications for measuring environmental noise. Applied Acoustics **106**, 16–22 (2016)
8. Nesmachnow, S., Rossit, D., Toutouth, J.: Comparison of Multiobjective Evolutionary Algorithms for Prioritized Urban Waste Collection in Montevideo, Uruguay. Electronic Notes in Discrete Mathematics (2018), in press
9. Nguyen, D.H., Sanjay Rebello, N.: Students' understanding and application of the area under the curve concept in physics problems. Physical Review Physics Education Research **7**(1) (2011)
10. Pruessner, J.C., Kirschbaum, C., Meinlschmid, G., Hellhammer, D.H.: Two formulas for computation of the area under the curve represent measures of total hormone concentration versus time-dependent change. The official journal of ISPNE **28**(7), 916—-931 (2003)
11. Segura-Garcia, J., Felici-Castell, S., Perez-Solano, J.J., Cobos, M., Navarro, J.M.: Low-cost alternatives for urban noise nuisance monitoring using wireless sensor networks. IEEE Sensors Jour. **15**(2), 836–844 (2015)
12. Sheskin, D.J.: Handbook of parametric and nonparametric statistical procedures. crc Press (2003)
13. Spears, W.M.: Evolutionary Algorithms: the role of mutation and recombination. Springer Berlin Heidelberg (2000)
14. Steele, C.: A critical review of some traffic noise prediction models. Applied acoustics **62**(3), 271–287 (2001)
15. Stolfi, D.H., Alba, E.: Smart mobility policies with evolutionary algorithms: The adapting info panel case. In: Proceedings of the 2015 Annual Conference on Genetic and Evolutionary Computation. pp. 1287–1294. GECCO '15, ACM, New York, NY, USA (2015)
16. Toutouh, J., Arellano-Verdejo, J., Alba, E.: Enabling low cost smart road traffic sensing. In: The 12th edition of the Metaheuristics International Conference (MIC 2017). pp. 13–15 (2017)
17. Toutouh, J., Rossit, D., Nesmachnow, S.: Computational intelligence for locating garbage accumulation points in urban scenarios. In: Learning and Intelligent OptimizatioN (LION) 12. pp. 1–15. Springer (2018)
18. Van Kempen, E.E., Kruize, H., Boshuizen, H.C., Ameling, C.B., Staatsen, B.A., de Hollander, A.E.: The association between noise exposure and blood pressure and ischemic heart disease: a meta-analysis. Environmental health perspectives **110**(3), 307 (2002)
19. Xue, B., Zhang, M., Browne, W.N., Yao, X.: A survey on evolutionary computation approaches to feature selection. IEEE Transactions on Evolutionary Computation **20**(4), 606–626 (2016)
20. Yang, J., Honavar, V.: Feature subset selection using a genetic algorithm, pp. 117–136. Springer (1998)