Machine learning based outlier detection for medical data

R. Vijaya Kumar Reddy[1], Shaik Subhani[2], B. Srinivasa Rao[3], N. Lakshmipathi Anantha[4]
[1,3]Department of IT, Lakireddy Bali Reddy College of Engineering (Autonomus), Mylavaram, India
[2]Department of IT, Sreenidhi Institute of Science and Technology (Autonomus), Telangana, India
[4]Department of CSE, Malla Reddy Engineering College (Autonomus), Telangana, India

Article Info

Article history:

Received Jan 10, 2021
Revised Aug 8, 2021
Accepted Aug 23, 2021

Keywords:

Machine learning algorithm
Medical data analysis
Outlier’s detection
Similarity operations

ABSTRACT

The concept of machine learning generate best results in health care data, it also reduce the work load of health care industry. This algorithm potentially overcome the issues and find out the novel knowledge for development of medical date in health care industry. In this paper propose a new algorithm for finding the outliers using different datasets. Considering that medical data are analytic of mutually health problems and an activity. The proposed algorithm is working based on supervised and unsupervised learning. This algorithm detects the outliers in medical data. The effectiveness of local and global data factor for outlier detection for medical data in real time. Whatever, the model used in this scenario from their training and testing of medical data. The cleaning process based on the complete attributes of dataset of similarity operations. Experiments are conducted in built in various medical datasets. The statistical outcome describe that the machine learning based outlier finding algorithm given that best accurateness.

Corresponding Author:

R. Vijaya Kumar Reddy
Department of IT
Lakireddy Bali Reddy College of Engineering (Autonomus)
Mylavaram, India
E-mail: Vijayakumarr285@gmail.com

1. INTRODUCTION

Outlier recognition is significant themes in data mining, the aim of discovery pattern that happen rarely as opposite to other data mining methods [1]. An outlier is derived significantly from inconsistent of a dataset [2]. The significance of outlier finding is in the sight of the truth so as to outliers can offer raw patterns and precious information about a dataset. Present research cover the fields of outlier discovery along with credit card fraud finding, network intrusion exposure, crime discovery, medical analysis, and detecting unusual parts in image processing [3].

Unsupervised outlier detection, is normally classified into distance-based [4], [5], density-based [6], [7], and distribution-based [3] procedures. This approach finds each data points are produced by a definite arithmetical model, but outliers do not accept this type of model. This method was preliminarily investigated by Knox and Ng [5]. In local information of the dataset differ to the global parameters. Density-based method was at first discussed by Breunig *et al*. [6]. Based on their local point density, local outlier factor is assign to every data point. The data point with a far above the ground for the local (LOF) value is described as an outlier. The clustering-based methods are unsupervised, they d not require any labeled training data, and their appearance in outlier discovery is restricted.

Many real-world applications may come across dissimilar cases for a small set of objects are labeled as outliers to a certain class, but the greater part of data are unlabeled. Significantly improve the efficiency of outlier detection, little bit of proper knowledge is required [8]-[10]. So semi supervised methods to outlier recognition have been urbanized to undertake such type of scenario, which have been consideration of a well-liked track of outlier discovery.

The last three decades, different types of machine learning approaches are proposed like an ant colony optimization, artificial neural networks, particle swarm optimization, evolutionary calculation and support vector machine (SVM). We concentrated on unsupervised methods that have been projected for outlier recognition in this literature. In density based spatial clustering with noise was used to generate numerous data segments by the unique characteristic space X into N dissimilar parts.

Amoli *et al*. [11] define unsupervised inventory for depressive symptomatology (IDS) for quick network, this type of network recognize zero attacks. The major engine of the density-based spatial clustering of applications with noise (DBSCAN) clustering identifies attacks and the second one set up the botnet under dissimilar protocols. The estimation of proposed mock-up, two obtainable datasets was used for checking. These accessible models was evaluated and also compared with multiple approaches like k-means and DBSCAN outlier detection techniques. In these approaches [12] recognize uncharacteristic behaviors in network and system log data also. A survey of machine learning methods for detection techniques discuss by Buczak and Guven variety of approaches [13], and explain the significance of the datasets for training and testing IDS. Nisioti *et al*. [14] provide unsupervised methods for anomaly type of IDS; it was obtainable and compared characteristic selection for intrusion detection.

The study is organized into different sections. Section 2 state the health care system architecture with training and testing. Section 3 deal with outlier detection using machine learning algorithm. Section 4 discusses the real world outlier data detection results. Conclusion of the study is presented in the last section.

2. HEALTH CARE SYSTEM ARCHITECTURE

Medical data profiling is often found in different sources of real time fields. Preprocessing of the data is essential for every real time data for remove noise. In health care industry predictive analytics is one of the significant issues. This paper focus on proper analysis of data using machine learning algorithm [15]. The diagram describe the training and testing phase of health care system. At first level, the medical history of patients and medical check outcomes are collected. Pre-process the data before applying the machine learning algorithms. The attributes related to the knowledge is practical in the training phase. The model is experienced with pre definite dataset and authority using dissimilar metrics and likelihood ratio set up [16].

Figure 1 describes the training and testing phase of health care system. At first level, the medical history of patients and medical check outcomes are collected. Pre-process the data before applying the machine learning algorithms. The attributes related to the knowledge is practical in the training phase. In testing phase check the similarity of new patients features and validates the data. The model is experienced with pre definite dataset and authority using dissimilar metrics and likelihood ratio set up.

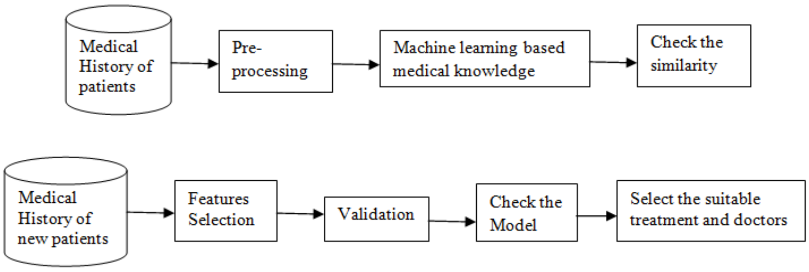
Medical History of patients
Pre-processing
Machine learning based medical knowledge
Check the similarity
Medical History of new patients
Features Selection
Validation
Check the Model
Select the suitable treatment and doctors

Figure 1. Training and testing phase for health care system

3. OUTLIER DETECTION USING MACHINE LEARNING ALGORITHM

3.1. Feature engineering

Feature engineering included the difference between current initial value of patient data and the final value, second one is difference between current medical data value and find the last five values average, and finally machine learning clustering feature based on hundred groups on the aforesaid. If the data is not normalize, decreased the performance of the models. The following Figure 2 represents the discovered outliers shown in red points for the local outlier factor trained on 0.5% dataset [17].

Figure 2 represents the detected outliers based on the random forests algorithm trained on the 0.5% anomaly dataset and Figure 3 shows the detected outliers for the isolation forests models trained on the 0.5% dataset. Figure 4 shows the detected outliers for the isolation forests models trained on the 0.5% dataset. Discovered outlier (x) for the isolation forest (FF) feature form trained on the 0.5% outlier dataset.

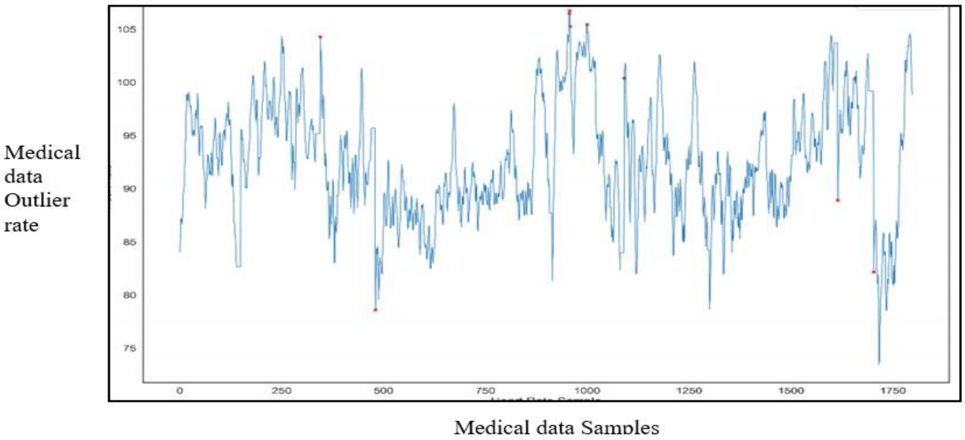

Figure 2. Discovered outlier (x) for the local (LOF) form trained on the 0.5% outlier dataset

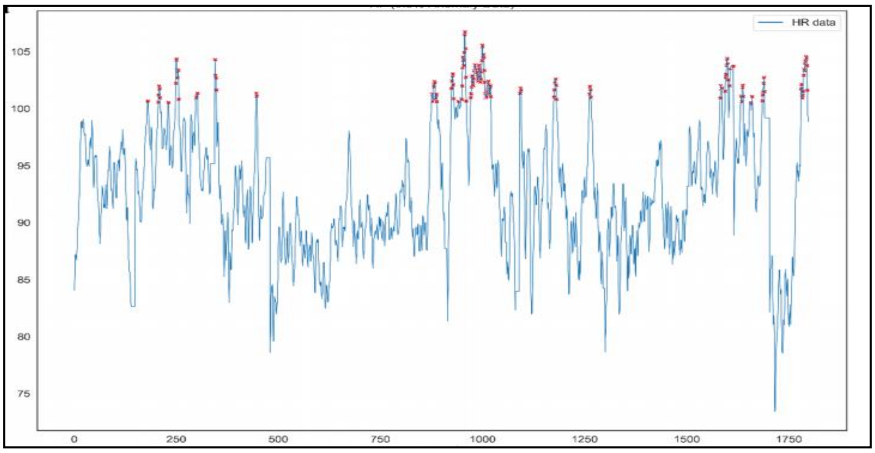

Figure 3. Discovered outlier (x) for the random forest (RF) feature form trained on the 0.5% outlier dataset

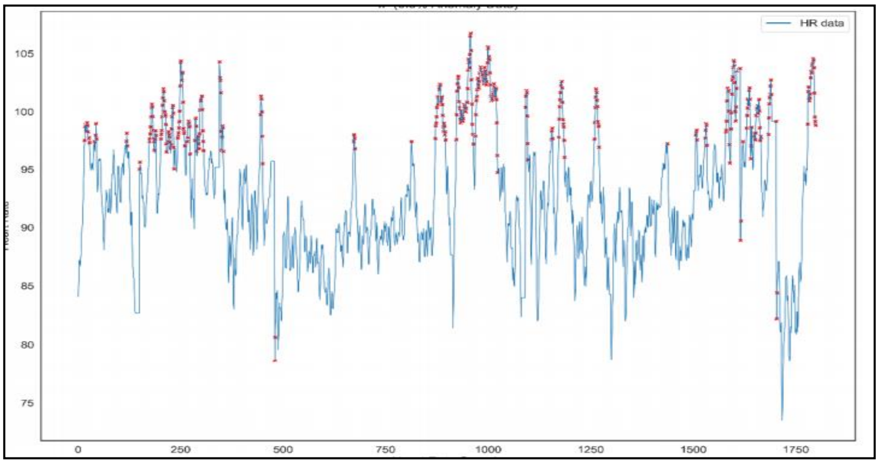

Figure 4. Discovered outlier (x) for the isolation forest (IF) feature form trained on the 0.5% outlier dataset

3.2. Simulated data performance

Every algorithm trained on 0.5% outlier dataset for train and test performances. Create the training and testing dataset in ratio of 80:20 come apart of the data. Sample parameters used a series of loops that connected through modify in every relevant parameter specified the model. This type of changes permitted for us to assess the importance of features to the model and crop any unrelated features in the process.

The proposed algorithm is a plain and well-liked clustering algorithm. It uses distance and a minimum number of points per cluster to classify a point as an outlier. The robust algorithm thusmakes dual predictions: it checks the point is outlier or not. To refine the predictions, check the all clusters other than outlier cluster. Euclidean distance function will be the default used for calculations.

Euclidean distance based method for outlier detection indirectly the neighborhood of an object; it is define by a given radius. A distance based on the threshold can be treated as a neighborhood of object. For each object **o,** we can discover neighbors of an object. Normally, let *r(r>0)* be a threshold distance and ∅ *(0<∅<1)* be a threshold fraction.

$$\text{dist} = \frac{\|o' dist(o,o') \leq r\|}{\|D\|} \leq \emptyset \tag{1}$$

The second approach takes O (n²) time,

- The density of an object and that of its neighbors examine by weight based outlier detection technique.
- So many real world data sets are complex structure. Compare to global data, local neighborhoods are better to measure outliers among objects.

The density-based outlier detection methods concentrate on density values of neighbor points. dist_k(o) is a distance between object o, and KNN. The k-distance neighborhood of o, consider as an entire objects of the distance to o, it is less than dist_k(o).

$$N_k(o) = [o|o' \in D, (o, o') \leq dist_k(o)] \tag{2}$$

Find out the average distance from the objects in $\mathrm{N_k(o)}$ to o. If o has indicated close neighbors o' such that *dist (o, o')* is tiny distance, due the numerical fluctuations of the distance calculate can be very high. So normalization methods applied for overcome the current issues.

$$\text{reachdistk(o,o}') = \max[\mathrm{dist_k(o)}, \mathrm{dist(o, o')}] \tag{3}$$

k is a user-specified parameter and it specify the minimum neighborhood to be check the local density of an object. The local density of an object o is,

$$\text{lrd(o)} = \frac{\|N_k(o)\|}{\sum_{o' \in N_k(o) reachdist_k(o,o')}} \tag{4}$$

we calculate the density for an object of local reachability and compared with its neighbors to spell out the degree to which the object is recognized an outlier. The above algorithm is working procedure as similar to existing ones. But the difference in efficiency and suitable dataset with different data sizes.

$$LOF_K(o) = \frac{\sum_{O' \in N_k(o)} \frac{lrd_k(o')}{lrd_k(o)}}{\|N_k(o)\|} \tag{5}$$

3.3. Proposed algorithm

```
begin
Step1: from cluster1 import ML-A
Step2: outlier1_detection = ML-A ( eps = .2
Step3: measurement=”euclidean”,
Step4: min_samples = 5, n_jobs = -1)
Step5: clusters = outlier_detection.fit_predict (num2)
end
```

4. RESULTS WITH REAL WORLD OUTLIER DATA DETECTION

In real world scenario medical data analysis is most important for generate accurate result. In this research predict the real time results for detecting the outliers. Before predicting the results, check the previous status of the existing data. This algorithm low sensitivity in primary task, after that generate good results. Outlier detection model is implemented suitable real time medical data. The training and testing process is similar to data mining approach for pre processing of data, training of data and testing with training data, finally validate the data and generate accurate results.

The proposed models are suitable for 2.5% outlier dataset. The LOF find out at the top end of the medical data range for outliers. In this research to find radio frequency (RF) and intermediate frequency (IF) of outliers from given data [18]. The model trained on 2.5% of dataset performed in the same way its counterpart that was trained on the 0.5% of outlier dataset. It is important to mention the ideal equilibrium between hit and false alarm rate will based on the task-related penalties of the outlier recognition. The outcome of dissimilar training dataset not changes in the pattern classification. The objective of this job access the potential performance of algorithm based on outliers. Visualized presentation of the model on a novel and real-world medical data samples, noted difference across the multiple models, and to conclude differentiate presentation based on which data the model was trained.

In this research based on real world considerations, we apply this type of approaches for algorithm utilization is too high and it accepts parallel processing also. The throughput of the algorithm supports the processing and distribution of the data in computing resources [19], [20]. The current research skilled to use a classification rule for detection of outliers, and all the way through a validation of medical data. In Table 1 we represent the comparisons of different methods for detecting outliers. Among these methods, proposed novel approach proves the better than traditional approaches.

Figure 5 is a graph represents the comparisons of different methods for detecting outliers. Among these methods, proposed novel approach proves the better than traditional approaches. In my proposed concept generate better results compare to more methods in presence of feature space and efficiency. The computational complexity is very less in proposed machine learning (ML) approach. Here practical applicability is not discussed because dependency factor based on model of the data and size of the data in machine learning. The comparison of different technical result available here. Out of these our proposed results state separately. The dataset of this research paper taken from kaggle website, for our research.

Table 1. Comparison of outlier detection methods

Methods	Computational Complexity	Efficiency	Feature Space	Practical applicability
Statistical Methods [21]	High Complex	Low	Single Variable	Statistical data
Parametric Methods [22]	Low Complex	Moderate	Single Variable	Prior knowledge of data sets
Non Parametric methods [22]	Low Complex	Moderate	Single / Multi Variable	profile of the data required
Distance Methods [23]	Nil	High	Multi Variable	Relation of individual points
Density Methods [24]	Complex	High	Multi Variable	Relation of points and nearest neighbor too
Cluster Approach [25]	Low Complex	High	Single / Multi Variable	clustering of similar data
Neural networks [26]	Nil	Very High	Multi Variable	Simple training data
Proposed ML approach	Nil	Very High	Single / Multi Variable	Based on data size

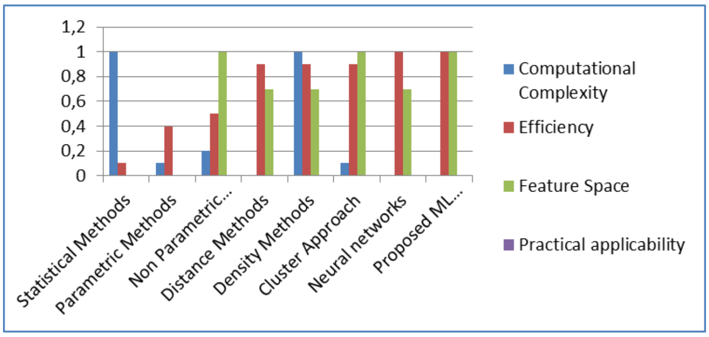

Figure 5. Graph representation of outlier detection methods

5. CONCLUSION

The concept of machine learning generate best results in health care data, it also reduce the work load of health care industry. This algorithm potentially overcome the issues and find out the novel knowledge for development of medical date in health care industry. In this paper recommend a novel method for finding the outliers using different medical datasets. Considering that medical data are analytic of health complications. The proposed approach is working based on supervised and unsupervised learning. This algorithm detects the outliers in medical data. The effectiveness of local and global data factor for outlier detection for medical data in real time. Whatever, the model used in this scenario from their training and testing of medical data. The cleaning process based on the complete dimensional attributes of dataset of similarity operations. Experiments are conducted in built in various medical datasets. The statistical results show that the machine learning based outlier recognition algorithm provided that finest accuracy.

REFERENCES

[1] J. Han, M. Kamber, and J. Pei, "Data Mining: Concepts and Techniques," *Elsevier*, 2012, doi: 10.1016/C2009-0-61819-5.

[2] D. M. Hawkins, "Identification of Outliers", *Chapman & Hall, London*, UK, 1980, doi: 10.1007/978-94-015-3994-4.

[3] O. Alan and C. Catal, "Thresholds based outlier detection approach for mining class outliers: an empirical case study on software measurement datasets," *Expert Systems with Applications*, vol. 38, no. 4, pp. 3440-3445, 2011, doi: 10.1016/j.eswa.2010.08.130.

[4] S. Ramaswamy, R. Rastogi, and K. Shim, "Efficient algorithms for mining outliers from large data set," *SIGMOD Record (ACM Special Interest Group on Management of Data)*, vol. 29, no. 2, pp. 427-438, May. 2000, doi: 10.1145/335191.335437.

[5] E. M. Knox and R. T. Ng, "Algorithms for mining distance based outliers in large dataset," in *Proceedings of the International Conference on Very Large Data Bases*, 1998, pp. 392-403.

[6] M. M. Breunig, H.-P. Kriegel, R. T. Ng and J. Sander, "LOF: identifying density-based local outliers," *ACM SIGMOD Record*, vol. 29, no. 2, pp. 93-104, June. 2000, doi: 10.1145/335191.335388.

[7] J. Ha, S. Seok and J.-S. Lee, "A precise ranking method for outlier detection," *Information Sciences*, vol. 324, pp. 88-107, Dec. 2015, doi: 10.1016/j.ins.2015.06.030.

[8] A. Daneshpazhouh and A. Sami, "Entropy-based outlier detection using semi-supervised approach with few positive examples," *Pattern Recognition Letters*, vol. 49, pp. 77-84, Nov. 2014, doi: 10.1016/j.patrec.2014.06.012.

[9] J. Gao, H. Cheng and P.-N. Tan, "Semi-supervised outlier detection," in *Proceedings of the ACM Symposium on Applied Computing*, pp. 635-636, ACM, Dijon, France, April 2006, doi: 10.1145/1141277.1141421.

[10] Z. Xue, Y. Shang, and A. Feng, "Semi-supervised outlier detection based on fuzzy rough C-means clustering," *Mathematics and Computers in Simulation*, vol. 80, no. 9, pp. 1911-1921, May. 2010, doi: 10.1016/j.matcom.2010.02.007.

[11] P. V. Amoli, T. Hamalainen, G. David, M. Zolotukhin, and M. Mirzamohammad, "Unsupervised network intrusion detection systems for zero-day fast-spreading attacks and botnets," *International Journal of Digital Content Technology and Its Applications*, vol. 10, no. 2, pp. 1-13, 2016.

[12] A. Bohara, U. Thakore, and W. H. Sanders, "Intrusion detection in enterprise systems by combining and clustering diverse monitor data," in *Proceedings of the Symposium and Bootcamp on the Science of Security*, Pittsburgh, PA, USA, 2016, pp. 7-16, doi: 10.1145/2898375.2898400.

[13] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," *IEEE Communications Surveys Tutorials*, vol. 18, no. 2, pp. 1153-1176, 2016, doi: 10.1109/COMST.2015.2494502.

[14] A. Nisioti, A. Mylonas, P. D. Yoo, and V. Katos, "From intrusion detection to attacker attribution: A comprehensive survey of unsupervised methods," *IEEE Communications Surveys Tutorials*, vol. 20, no. 4, pp. 3369-3388, 2018, doi: 10.1109/COMST.2018.2854724.

[15] J. S. Keertan, Y. Nagasai and S. Shaik, "Machine Learning Algorithms for Oil Price Prediction," *International Journal of Innovative Technology and Exploring Engineering*, vol. 8, no.. 8, pp. 958-963, June. 2019.

[16] https://digital.ahrq.gov/key-topics/architecture-health-it

[17] C. Isaksson, M. H. Dunham, "A Comparative Study of Outlier Detection Algorithms," *Machine Learning and Data Mining in Pattern Recognition*, vol. 5632, pp 440-453, 2009, doi: 10.1007/978-3-642-03070-3_33.

[18] S. Shaik and U. Ravibabu, "Classification of EMG Signal Analysis based on Curvelet Transform and Random Forest tree Method," *Journal of Theoretical and Applied Information Technology (JATIT)*, vol. 95, no. 24, pp. 6856-6866, Dec. 2017.

[19] R. Bekkerman, M. Bilenko, and J. Langford, "*Scaling up machine learning: Parallel and distributed approaches*," *Cambridge University Press, Cambridge*, 2012.

[20] R. Vijaya Kumar Reddy and U. Ravi Babu, "A Review on Classification Techniques in Machine Learning," *International Journal of Advance Research in Science And Engineering*, vol. 7, no. 3, March 2018.

[21] M. Alam, "Statistical techniques for anomaly detection", September 2020. [Online]. Available: https://towardsdatascience.com/statistical-techniques-for-anomaly-detection-6ac89e32d17a

[22] J. Joseph, "How to detect outliers using parametric and non-parametric methods: Part I", 2019. [Online]. Available: https://clevertap.com/blog/how-to-detect-outliers-using-parametric-methods-and-non-parametric-methods/

[23] J. Ranjan Sethi, "Study of Distance-Based Outlier Detection Methods", June 2013. [Online]. Available: https://core.ac.uk/download/pdf/53189702.pdf

[24] G. Kumar Jha, N. Kumar, P. Ranjan and K. G. Sharma," Density Based Outlier Detection (DBOD) in Data Mining: A Novel Approach", *Recent Advances in Mathematics, Statistics and Computer Science*, pp. 403-412, 2016. doi: 10.1142/9789814704830_0037

[25] A. Christy, G. Meera Gandhi and S. Vaithyasubramanian, "Cluster Based Outlier Detection Algorithm For Healthcare Data," ISBCC-15, *Procedia Computer Science*, vol. 50, pp. 209-215, 2015, doi: 10.1016/j.procs.2015.04.058.

[26] S. Hawkins, H. He and G. Williams and R. Baxter, "Outlier Detection Using Replicator Neural Networks," *Proceedings of the 4th International Conference on Data Warehousing and Knowledge Discovery*, September 2002, pp. 170-180, doi: 10.1007/3-540-46145-0_17.