Journal of Critical Reviews

ISSN- 2394-5125 Vol 7, Issue 4, 2020

Review Article

MACHINE LEARNING METHODOLOGY FOR MEDICAL DATA ANALYSIS FOR PREDICTION OF RISK

G.L. Sravanthi[1], N. Harika[2], G. Archana[3], B. Sundara Leela[4]

[1]Asst. Professor, CSE Department, Vignan Nirula Institute of Technology & Science for Women, Peda Palakaluru, Guntur, Andhra Pradesh, India. glsravanthi88@gmail.com
[2]Asst. Professor, CSE Department, Vignan Nirula Institute of Technology & Science for Women, Peda Palakaluru, Guntur, Andhra Pradesh, India. narraharika4@gmail.com
[3]Asst. Professor, CSE Department, Vignan Nirula Institute of Technology & Science for Women, Peda Palakaluru, Guntur, Andhra Pradesh, India. archu.gunakala@gmail.com
[4]Asst.Professor, JNTUK, Andhra Pradesh, India. sundaraleela.b@gmail.com

Received: 19.12.2019 *Revised: 22.01.2020* *Accepted: 24.02.2020*

Abstract
Mining data is a nontrivial procedure of finding information from a large volume of data. Such information can be helpful in settling on significant choices. Medical data show special features including noise coming about because of human just as methodical blunders, missing qualities and even meager conditions. The nature of data has huge ramifications for the nature of the mining results. Medical data classification is important to perform preprocessing steps so as to expel or at least lighten a portion of the issues related with medical data. Clustering is a descriptive-based data mining task. The clustering algorithm is also called as unsupervised learning algorithm that learns the unlabeled dataset and groups or clusters the instances based on their similarity and builds the clustering model. Clustering is same as classification in which data is grouped, but in this, groups are not predefined. In clustering, clusters are not predefined. Classification of different types of clustering is as follows: Hierarchical clustering, Partition clustering, Categorical clustering, Density based clustering and Grid based clustering. The main intension of the research is to classify the medical data with high accuracy value. In order to achieve promising results, a novel data classification methods have been designed that utilize a Improved Cluster Optimal Classifier (ICOC). The proposed method is compared with traditional methods and the results show that the proposed method performance is better and accurate.

Keywords: Medical Data Classification, Machine Learning Methodologies, Data Mining, Data Classification, Cluster Based Classification.

DOI: http://dx.doi.org/10.31838/jcr.07.04.151

INTRODUCTION

Classification is the way toward grouping a data thing into one of the predefined classes. Two steps are to be followed in Classification process [1]. It includes examining the features of a recently displayed object and appointing to it a predefined class. Initial a model is constructed portraying a predefined arrangement of data classes or ideas [2]. Preparing data are utilized to construct the model. Furthermore, the model is utilized for classification. Figure 1.6 represents the Schematic Representation of Classification

Classification is a predictive-based data mining task. In order to accomplish the classification task, the classification algorithm is used to learn the dataset and to build the classifier [15]. The dataset contains a set of features (columns) and instances (rows) with a target-class attribute which contains the class-label associated with each instance of the dataset [16]. The unlabeled data is given to the classifier in order to predict the class-label of the unlabeled instance [17]. Classification can be performed on structured or unstructured data [18]. The goal of classification is to identify the category where new data comes under.

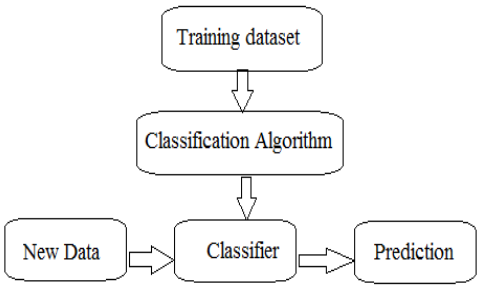

Fig 1: Schematic Classification Representation

A data classifier requires a choice of features that must be custom fitted independently for different issues [19] [20]. Following component determination, classifier improvement requires detachment of the data into preparing and test data and experiences two noteworthy periods of data classifier development as shown in the Figure 2.

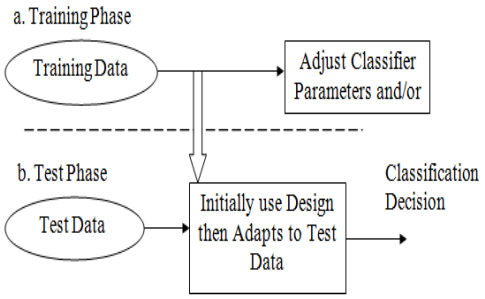

Figure 2: The Two Major Phases of Data Classifier Development

LITERATURE WORK

Adeniyi et al [1] proposed the indicator can be utilized to structure an online application to acknowledge the indicator factors and Automated framework Decision Tree based expectation can be actualized in remote regions like provincial districts or wide open spaces, to mirror like human analytic mastery for forecast of illness.

The Bayesian system is additionally seen as a procedure in medical expectation Particular it has been used for Brest malignancy guess and determination [20]. In future we mean to plan and actualize such framework for online applications [21].

Akhiljabbar et al [2] passed on a local harsh set classification way that deals with arrangement with medical datasets. Five benchmarked medical datasets had been utilized in that exploration work for examining the effect of proposed work in basic leadership. Medical datasets devour tremendous measure of data about the patients, ailments and the doctors. Ailments finding required numerous costly examinations to foresee the maladies. Cost of sickness forecast and finding can be decreased by applying AI and data mining strategies. Sickness forecast and basic leadership assumes a huge job in medical conclusion [22].

AlMuhaideb et al [3] presented data mining task and gave a short synopsis of different data mining calculations utilized for classification, clustering, and affiliation. Exchange was made to empower the sickness finding and guess.

Alsaffar et al [4] displayed a methodology for the usage of total knowledge in remedial examination by applying accord systems. They differentiated the accuracy procured and that system against the diagnostics precision came to through the data of a lone ace. They used the ontological structures of ten diseases. Two data bases were made by setting five contaminations into each learning base. They coordinated two assessments, one with an empty data base and the other with a populated learning base.

For the two examinations, five experts included or conceivably executed signs/reactions and expository tests for each infection [23]. After that system, the individual learning bases were produced reliant on the after effect of the understanding procedures [24]. To play out the evaluation, they dissected the amount of things for each infirmity in the agreed learning bases against the amount of things in the GS (Gold Standard).

Anooj et al [5] showed that so as to evaluate the present general acknowledgment inside the Medical network of Shaken Baby Syndrome (SBS), Abusive Head Trauma (AHT), and a few elective clarifications for discoveries normally observed in manhandled kids. That was a review of doctors much of the time engaged with the assessment of harmed kids at 10 driving youngsters' medical clinics.

Aqueel Ahmed et al [6] proposed a shared data based max relevancy min-excess (MRMR) features. To distinguish the feature importance, the common data is figured between the individual feature and target-class, and to recognize the repetitive feature, the fundamentally unrelated condition is connected. They built up a shared data based feature determination technique (MIFS). In this technique, common data measure is utilized to decide the pertinence between the individual feature and the objective class [24]. The features having comparative data are considered as repetitive features that are to be expelled [25].

Attar Salam et al [7] has introduced an investigation of programmed web use data mining and suggestion framework dependent on current client conduct with snap info. so as to give significant data to the person without unequivocally requesting it. The outcome demonstrates that the KNN classifier was straightforward, reliable, clear, easy to see, high propensity to have alluring characteristics and simple to actualize than most other AI methods explicitly when there is next to zero earlier learning about data appropriation.

AwateSuyash et al [8] have productive therapeutically classification framework dependent on Adaptive Hereditary Fuzzwords (AGFS). In that examination i) creating rules from data just as for the improved standards determination, Adapting of hereditary calculation is done to clarify the investigation issue in hereditary calculation by introducing another administrator called orderly expansion, ii) Proposing a basic method for plotting of enrollment capacity, and iii) Designing a wellness work by permitting the recurrence of event of the principles in the preparation data.

Bai R et al [9] have proposed a coronary illness anticipating framework, which arrange the examination results and give per users a review of the current coronary illness expectation strategies in every class. Neural Systems have been used to make forecasts for therapeutically.

BartoszKrawczyk et al [11] displayed a way to deal with fuzzy delicate sets in basic leadership to abstain from choosing a reasonable level delicate set and to apply that way to deal with tackle medical determination issues. That approach joined Gray social examination with the Dempster Shafer hypothesis of proof.

Basari A et al [12] as per a few perspectives, in any event one PC comprehensible medium encoding guidelines that when executed perform such a strategy as well as a framework for giving such a technique is given.

Benameur L et al [14] utilized to manage this issue. We at that point lead a trial concentrate to explore the nature of various combination techniques for joining classifiers in an outfit. A few combination procedures dependent on discrete and consistent reactions from (neural system) base classifiers are assessed and it is demonstrated that a cautious decision of combination technique can support the acknowledgment pace of the minority class. Specifically, a neural system prepared fuser is appeared to give the best arrangement execution on two separate bosom disease datasets.

Gopi et al [15] utilized a piece of the learning procedure of the regulated learning calculation for feature determination. Inserted based strategies lessen the computational expense than the wrapper strategy. This installed strategy can be generally ordered into three to be specific pruning technique, worked in instrument, and regularization models. In the pruning-based strategy, at first every one of the features are taken into the preparation procedure for structure the arrangement model and the features which have less relationship coefficient worth are evacuated recursively utilizing the help vector machine (SVM).

Bermejo P et al [29] introduce and talk about the examination that was executed with gullible bayes strategy so as to assembled prescient model as a counterfeit analyze for coronary illness dependent on informational collection which contains set of parameters that were estimated for people already. At that point contrast the outcomes and different strategies as per utilizing

similar information that were given from UCI archive information.

PROPOSED WORK

Many data mining techniques exist for Medical data categorization [26]. But, the classification accuracy of these models is restricted frequently when the relationship of input/outcome datasets are composite and/or unsystematic [27]. Analysts attempted to utilize diverse methods for the exact classification of data. MC is essential to perform medical data classification so as to diagnoses the diseases [28].

Feature dimensionality reduction is utilized to decrease the excess features by avoiding irrelevant and repeating features from a dataset. Feature Selection (FS) algorithm can be determined based on two criteria: The first one is the determination of the required memory space and the computational complexity of the feature selection algorithm. The second one is the determination of the worthiness of the features selected by the feature selection algorithm [18] [24]. The worthiness of the selected features is determined in terms of classification accuracy and ability to have good generality, i.e., ability to produce higher classification accuracy with different classifiers.

Here, ICOC is utilized to decrease the traits (features) dimension. At that point when the traits lessening are surrounded, effective classifier is utilized for probability calculation. In this effective classifier, ICOC algorithm is utilized and the Classification strategy is accompanying in Figure 3.

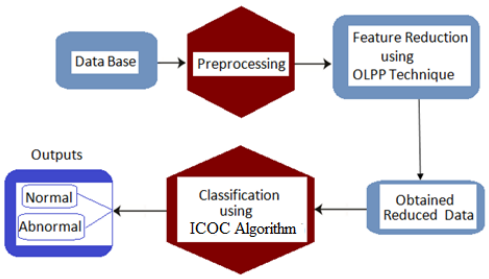

Figure 3: First Hybrid Classification Approach

Data preprocessing techniques discretization is used to unify the presentation of all variables (attributes) and their values in the dataset. Normalization techniques can be used to scale up or scale down the variable-values within a desirable range. The size of a dataset matters when it comes to data analysis and knowledge extraction. Medical dataset can be considered to have width and height. From a medical viewpoint, width refers to the clinical characteristics of a patient, and height refers to patients' records. From a database design viewpoint width refers to attributes and height refers to tuples.

The procedure of feature reduction is generally grouped into four classes to be specific channel, wrapper, installed, and half and half techniques dependent on how the managed learning calculation is utilized in the feature reduction procedure. Data increase, symmetric vulnerability, gain proportion, and so forth and the top positioned features are chosen as huge features by predefined edge esteem.

ICOC is computationally less expensive and space multifaceted nature is less contrasted with subset approach. The reason for this subset approaches is reduction in dimensionality. As initiation of the calculation, Principal Components Analysis is used for this process. A contiguous chart worked by ICOC and the class association within the pattern reflects better outcomes.

This projection incorporates with ongoing advances:

- Obtain a great deal of parameters (features) from the data.
- Mean worth calculation
- Covariance framework is calculated and then learn eigenvector and estimation of covariance lattice
- Concatenating Eigen worth and Eigen vectors mentioned

ICOC have being determined toward streamline the weight in the neural classification. The looking through action of creatures is fundamentally finished with the goal of finding resources that incorporate nourishment and haven. Here the ICOC calculations have being corrected with this assistance in regard to velocity updating without considering performance of the rangers for selecting the resources randomly.

Improved ICOC Algorithm

Step 1: Search solution and head angle

Head angle is represented as,

$$\Psi_i^t = \left(\Psi_{i1}^t \ldots\ldots\ldots \Psi_{i(n-1)}^t\right)$$

The head angle calculation is done using

$$L_i^t(\Psi_i^t) = \left(l_{i1}^t \ldots\ldots\ldots l_{i(n)}^t\right)$$

Polar & Cartesian coordinate transformations are calculated as

$$L_{i1}^t = \prod_{p=1}^{n-1} \cos(\Psi_{ip}^t)$$

$$L_{ij}^t = \sin(\Psi_{i(j-1)}^t \prod_{p=j}^{n-1} \cos(\Psi_{ip}^t); \quad \text{Where(j=2....n-1)}$$

$$L_{in}^t = \sin(\Psi_{i(n-1)}^t)$$

Step 2: Computing the function for Fitness.

Fitness = Min(MSE)

Step 3: Producer is Fixed (Z_p).

Member with best fitness is fixed as producers.

(i) Scanning operation at 0^o

$$Z_z = Z_p^t + \varepsilon_1 d_{max} L_p^t(\Psi^t)$$

(ii) Right hand side scanning operation

$$Z_r = Z_p^t + \varepsilon_1 d_{max} L_p^t\left(\Psi^t + \varepsilon_2 \Phi_{max}/2\right)$$

(iii) left hand side Scanning operation

$$Z_l = Z_p^t + \varepsilon_1 d_{max} L_p^t\left(\Psi^t - \varepsilon_2 \Phi_{max}/2\right)$$

Where, ε_1 focuses to a random member which is normally distributed with zero mean along with solidarity standard deviation. Also, ε_2 represents a random sequence which is uniformly distributed somewhere in the range of zero and one.

RESULTS

This proposed strategy achieves awesome exactness for the restorative information grouping. In this proposed technique fluffy min max neural system with altered gathering scan streamlining agent calculation for characterization is utilized. And furthermore this forecast exactness result by contrasting and different classifiers can be set up.

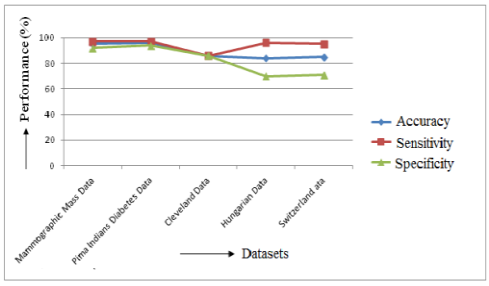

Fig. 4: Performance of classification method.

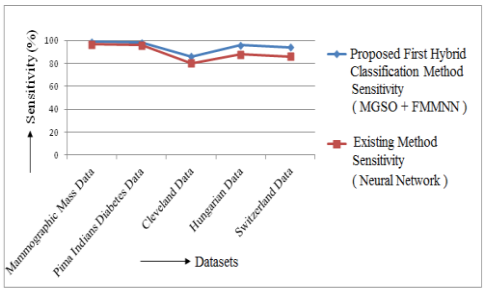

Fig 5: Outcomes of the Sensitivity

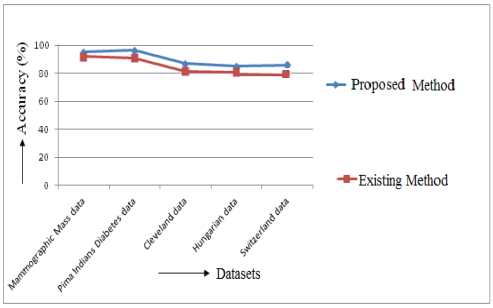

Fig 6: Accuracy Level

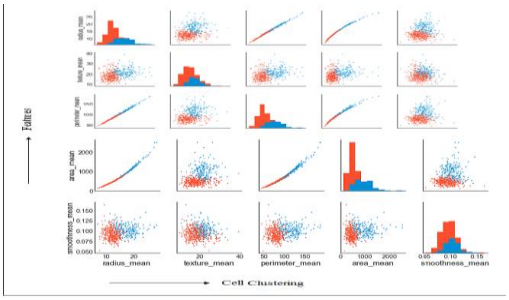

Fig. 7: Cluster Groups

CONCLUSION

Medical data classification is important to perform preprocessing steps so as to expel or at least lighten a portion of the issues related with medical data. Clustering is a descriptive-based data mining task. The implementation of the proposed method was done in ANACONDA SPYDER. For experimentation, the dataset given in the UCI machine learning repository such as, Mammographic Mass Data, Pima Indians Diabetes Data, Cleveland, Hungarian and Switzerland etc. was utilized to break down the presentation for breast cancer, diabetes, heart infections and kidney disease by utilizing the proposed methods using Accuracy, Sensitivity and Specificity. The consequences of proposed techniques are demonstrated that, ICOC calculations accomplishes efficient outcome when contrasted with MGSO alongside FMMNN strategy. The proposed method achieves 94% of accuracy value for therapeutically classification.

REFERENCES

1. Adeniyi D.A, Wai Z, Yongquan Y, (2014), "Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method", Applied Computing and Informatics, Vol. 3.
2. Akhiljabbar M, Deekshatulu B L and Priti Chandra, (2013), "Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm" Elsevier Procedia Technology, Vol. 10, pp. 85-94.
3. AlMuhaideb, Sarab and Mohamed El BachirMenai, (2016), "An Individualized Preprocessing for Medical Data Classification", Elsevier on Procedia Computer Science, Vol.82, pp.35-42.
4. Alsaffar A and Omar N, (2015), "Integrating a Lexicon based approach and K nearest neighbour for Malay sentiment analysis", Journal of Computer Science, 11(4), pp.639-645.
5. Anooj P.K, (2012), "Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules," Elsevier Computer and Information Sciences, Vol. 24, no. 1, pp. 27-40.
6. Aqueel Ahmed and Shaikh Abdul Hannan, (2012), "Data Mining Techniques to Find Out Heart Diseases: An Overview ", International Journal of Innovative Technology and Exploring Engineering, vol.1,no.4,pp.18-23.
7. Attar Salam A Al, Pollex Rbecca L, Robinson John F,Miskie Brooke A, Walcarius Rhonda, Rutt Brian K and Hegele Robert A, (2006), " Research Article, Semi automated segmentation and quantification of adiposet issue in calf and thigh by MRI: a preliminary study in patients with monogenenic metabolic syndrome. *BMC Medical Imaging,* 6-11
8. AwateSuyash P. and Gee James C, (2007), " A Fuzzy, Nonparametric Segmentation Framework for DTI and MRI Anlysis ", *In Proc. Information Processing Medical Imaging (IPMI)*: 296-307

9. Bai R, Wang X and Liao J, (2009), " Folksonomy for theblogosphere: Blog identification and classification", In Computer Science and Information Engineering, WRI World Congresson (Vol. 3, pp. 631-635), IEEE.
10. BallinAyeletAkselrod, GalunMeirav, Gomoriz MosheJohn, Basri Ronen and Brandt Achi, (2006), "Atlas Guided Identification of Brain Structures by Combining 3DSegmentation and SVM Classification", *MICCAI, LNCS*4191, 209-216.
11. BartoszKrawczyk and Gerald Schaefer, (2012), "Ensemble fusion methods for medical data and classification", In proceeding of 11th symposium on neural network application in electrical engineering, pp.143-146.
12. Basari A. S. H., HussinB., Ananta I. G. P and Zeniarja J, (2013), " Opinion mining of movie review using hybrid method of support vector machine and particle swarm optimization", Procedia Engineering, 53, 453-462.
13. Beloufa, Fayssal & Chikh, (2013), " Design of Fuzzy Classifier for Diabetes Disease using Modified Artificial Bee Colony Algorithm", Computer Methods and Programs in Biomedicine, Vol. 112, no. 1, pp. 92-103.
14. Gao, X., Yuan, S.High density lipoproteins-based therapies for cardiovascular disease(2010) Journal of Cardiovascular Disease Research, 1 (3), pp. 99-103.
    DOI: 10.4103/0975-3583.70898
15. Gopi, A.P., Jyothi, R.N.S., Narayana, V.L. et al. (2020), "Classification of tweets data based on polarity using improved RBF kernel of SVM". Int. j. inf. tecnol. (2020). https://doi.org/10.1007/s41870-019-00409-4.
16. A Peda Gopi and Lakshman Narayana Vejendla, (2019)," Certified Node Frequency in Social Network Using Parallel Diffusion Methods", Ingénierie des Systèmes d' Information, Vol. 24, No. 1, 2019,pp.113-117.DOI: 10.18280/isi.240117
17. Lakshman Narayana Vejendla and A Peda Gopi, (2019)," Avoiding Interoperability and Delay in Healthcare Monitoring System Using Block Chain Technology", Revue d'Intelligence Artificielle, Vol. 33, No. 1, 2019,pp.45-48.
18. Lakshman Narayana Vejendla and Bharathi C R,(2018),"Multi-mode Routing Algorithm with Cryptographic Techniques and Reduction of Packet Drop using 2ACK scheme in MANETs", Smart Intelligent Computing and Applications, Vo1.1, pp.649-658. DOI: 10.1007/978-981-13-1921-1_63 DOI: 10.1007/978-981-13-1921-1_63
19. Chauhan SP, Sheth NR, Jivani NP, Rathod IS, Shah PI. "Biological Actions of Opuntia Species." Systematic Reviews in Pharmacy 1.2 (2010), 146-151. Print. doi:10.4103/0975-8453.75064
20. Lakshman Narayana Vejendla, A Peda Gopi and N.Ashok Kumar,(2018)," Different techniques for hiding the text information using text steganography techniques: A survey", Ingénierie des Systèmes d'Information, Vol.23, Issue.6,pp.115-125. DOI: 10.3166/ISI.23.6.115-125
21. A Peda Gopi and Lakshman Narayana Vejendla (2018), "Dynamic load balancing for client server assignment in distributed system using genetic algorithm", Ingénierie des Systèmes d'Information, Vol.23, Issue.6, pp. 87-98. DOI: 10.3166/ISI.23.6.87-98
22. Lakshman Narayana Vejendla and Bharathi C R,(2017),"Using customized Active Resource Routing and Tenable Association using Licentious Method Algorithm for secured mobile ad hoc network Management", Advances in Modeling and Analysis B, Vol.60, Issue.1, pp.270-282. DOI: 10.18280/ama_b.600117
23. Lakshman Narayana Vejendla and Bharathi C R,(2017),"Identity Based Cryptography for Mobile ad hoc Networks", Journal of Theoretical and Applied Information Technology, Vol.95, Issue.5, pp.1173-1181. EID: 2-s2.0-85015373447
24. Lakshman Narayana Vejendla and A Peda Gopi, (2017)," Visual cryptography for gray scale images with enhanced security mechanisms", Traitement du Signal,Vol.35, No.3-4,pp.197-208. DOI: 10.3166/ts.34.197-208
25. A Peda Gopi and Lakshman Narayana Vejendla, (2017)," Protected strength approach for image steganography", Traitement du Signal, Vol.35, No.3-4,pp.175-181.
26. Lakshman Narayana Vejendla and Bharathi C R,(2016),"Secured Key Production and Circulation in Mobile Ad hoc Networks Using Identity Based Cryptography", International conference on Engineering and Technology, Vol.1, pp.202-206.
27. V.Lakshman Narayana, A.Koteswara Rao (2015), "Secured Cloud Data Storage with User Validation", International Journal of Computer Science and Technology, Volume6, Issue 4, pp.78-85.
28. V.Lakshman Narayana, Sk.Mastan, Dr.M.Kishore kumar (2012)," Multiple Routing Configurations For Fast IP Network Recovery", International Journal of Engineering Research & Technology, Vol. 1 Issue 5, July – 2012.
29. Bermejo P., Gámez J. A and Puerta J. M, (2014), "Speeding upincremental wrapper feature subset selection with Naive Bayes classifier", Knowledge-Based Systems, 55, 140-147.
30. Bezdek J.C, Keller J,Krisnapuram R, and Pal N, (1999), " Fuzzy Models and Algorithms for Pattern Recognition and Image Processing", Springer. DOI: 10.1007/b106267.