Ilkogretim Online - Elementary Education Online, Year; Vol 20 (Issue 5): pp. 1782-1792
http://ilkogretim-online.org
doi: 10.17051/ilkonline.2021.05.196

Intelligent Data Analysis approaches for Knowledge Discovery: Survey and challenges

Maher O Al-Khateeb, Department of Computer Science, Zarqa University, Zarqa, 13132, Jordan
Mohammad A. Hassan, Department of Computer Science, Zarqa University, Zarqa, 13132, Jordan
Ibrahim Al-Shourbaji*, Department of Computer and Network Engineering, Jazan University, 82822-6649 Jazan, Kingdom of Saudi Arabia, alshourbajiibrahim@gmail.com
Muhammad Saidu Aliero, School of Information Technology, Monash University, Subang Jaya, Malaysia

Abstract- With the enormous growth in information and data that are produced by various resources such as organization, companies' phones, health records, social media, and the Internet of Tings (IoT), their analysis becomes a challenge and even more complex due to the increased volume of structured and unstructured data. Knowledge Discovery in Database (KDD) is the process of finding knowledge in data stored by various resources using Intelligent Data Analysis (IDA) techniques which have the ability to analyze and discover knowledge from these data. This paper investigates the main challenges in KDD. Also, it illustrates the IDAs approaches used to address KDD trends in short and finally presents open issues for research and progress in the field of KDD.

Keywords: Knowledge Discovery in Database; Intelligent Data Analysis; Missing values; Data scarcity, Black box; Mathematical model

I. INTRODUCTION

At present, the size of data is becoming huge in terms of the abundance of features and in their dimensionality due to the increased amount of data produced and stored by various resources such as organization, companies' phones, health records, social media and Governments [1]. The main challenge is how to analyze, summarize and discover knowledge from these stored data. Knowledge Discovery in Database (KDD) plays a fundamental role in discovering useful information and identifying hidden patterns in large data warehouses. This huge size of the data can significantly influence the accuracy of most Intelligent Data Analysis (IDA) techniques and their efficiency, especially in the presence of irrelevant, redundant features, missing data or scare data [2].

Data come from different sources are integrated and stored into a single data, called the target data. Then they are transformed into a standard format. IDA methods aim at processing the data to the output in the form of rules or patterns and then they are interpreted to useful information [3]. The process of the KDD is shown in Figure 1.

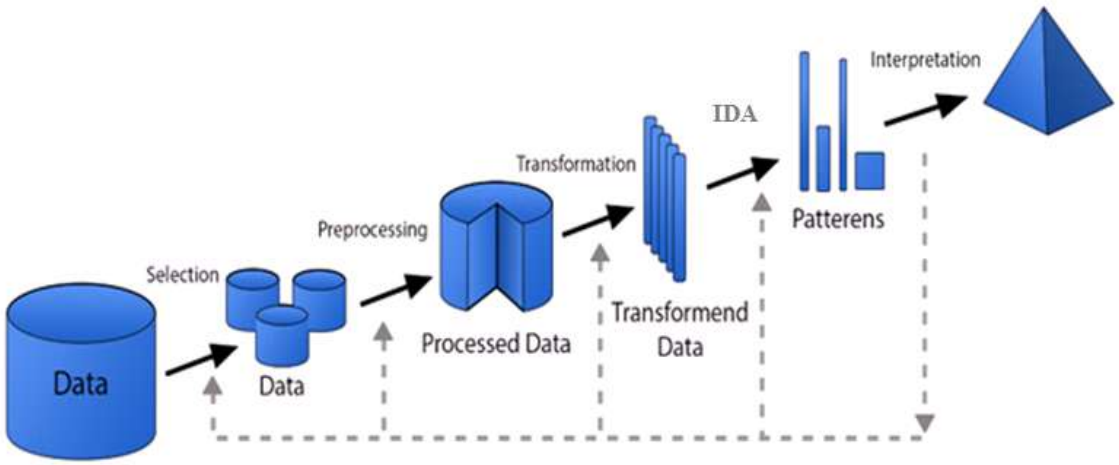

Figure 1 KDD process

In our present paper, we analyze the challenges and issues in KDD. We then discuss some recent approaches that are proposed to address these issues. We highlight several open issues and research directions in KDD. A list of abbreviations used in this paper is provided in Table 1.

This paper is structured as follows: Section 2 offers a brief overview of KDD challenges. Section 3presentsthe approaches that are proposed to overcome KDD challenges. Section 4 discusses open issues. Finally, Section 5 concludes the paper.

Table 1. List of abbreviations definitions

Abbreviations	Description
KDD	Knowledge Discovery in Database
IDA	Intelligent Data Analysis
SVM	Support Vector Machines
PCA	Principal Component Analysis
RBF	Radial Basis Function
MM	Mathematical Model
RF	Random Forest
k-NN	k-Nearest Neighbor
GA	Genetic Algorithm
LAW-LSimpute	Locally Auto Weighted Least Squares Imputation
GP	Genetic Programming
UCI	University of California at Irvine
WPCA	Weighted Principal Component Analysis
LDA	Linear Discriminant Analysis
PSO	Particle Swarm Optimization
JMIM	Joint Mutual Information Maximization
NJMIM	Normalized Joint Mutual Information Maximization
MMCC	Maximum-Minimum Correntropy Criterion
SA	Sensitivity Analysis
AAD	Average Absolute Deviation

VEC	Variable Effect Characteristic
NN	Neural Network
ANN	Artificial Neural Network
FARB	All-Permutation Rule Base
AUC	Area Under the receiver operating Characteristic
PRC	Area under the precision-Recall Curve
GMM	Gaussian Mixture Model
ELM	Extreme Learning Machine
MSE	**Mean square Error**

II. CHALLENGES IN KDD

In this section, the current trends in KDD are discussed. Figure 2 shows list of potential challenges within the KDD. This section is divided into five categories: Section 2.1 missing data, Section 2.2 data scarcity, Section 2.3 data dimensionality reduction, section 2.4 black box model and Section 2.5 mathematical model and in each parta brief overview of the concept is discussed.

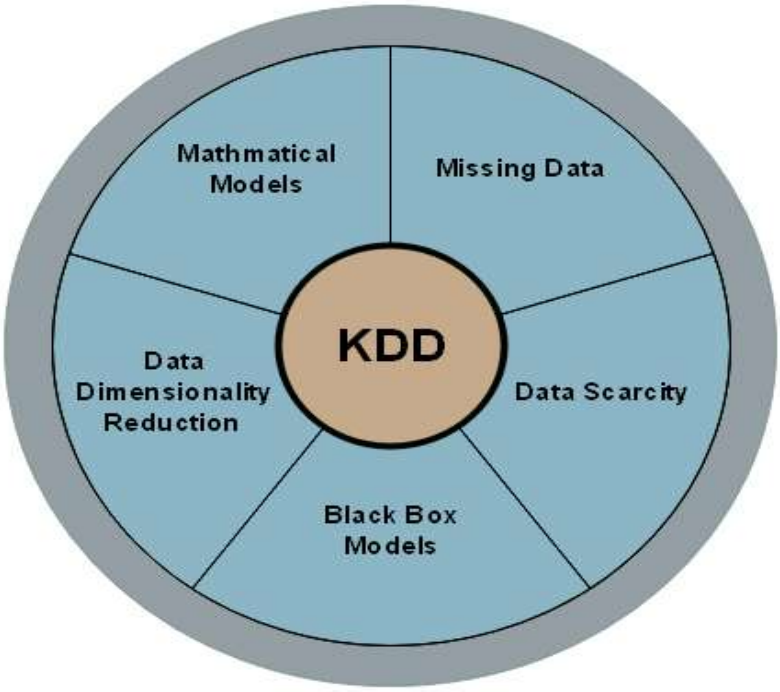

Figure. 2. Main challenges in the KDD

2.1 Missing Data

Rubin first developed a framework for missing values problem in 1976 in an attempt to understand the strengths and limitations of different analytic strategies [4]. Some IDA methods accept missing values whileothers such as SVM and PCA require all values to be available, and therefore, they cannot be applied to data with missing values [5] Therefore, missing value estimation is of paramount importance in KDD. The simplest method to tackle this problemis by removing the entire rows that contain missing values and replacing them with the average, medium or zeros [6]. However, this could lead to biases since the correlation structure between the variables is ignored.

2.2 Data scarcity

Another important problem in KDD is the required amount of data or examples to adequate IDA learning for successful generalization. Data scarcity presents an important challenge for real life studies because collecting large amount of self reported data require time and efforts, especially in healthcare sector [7]. To tackle this challenge, two possible ways can be used: (i) collect more information or (ii) design suitable methods to deal with small amount of data

2.3 Data dimensionality reduction

Over the past decades, several works confirmed that the existence of irrelevant, unimportant or redundant features has a direct effect on reducing the accuracy of IDA learning algorithms and their effectiveness [8, 9, 10]. For example, the authors in [11] found that deleting these features helped in reducing complexity structural and improving classification performance of the RBF neural network. In another work [12], the authors stated that reducing dimensionality of the feature space by eliminating less important features improved neural fuzzy classifier computational speed. Therefore, dimensionality reduction of the feature space is vital in KDD [13].
Dimensional reduction methods use correlation structure between variables to achieve several goals such as reducing number of components, ensuring that these components are independent and providing a model for results interpretation [14]. There are three main dimensions of the preprocessed data and usually represented in the form of flat files. These dimensions include features (i.e. columns), samples (i.e. rows) and the numerical value of features [15, 16]. Figure 3 depicts the association between dimensions.

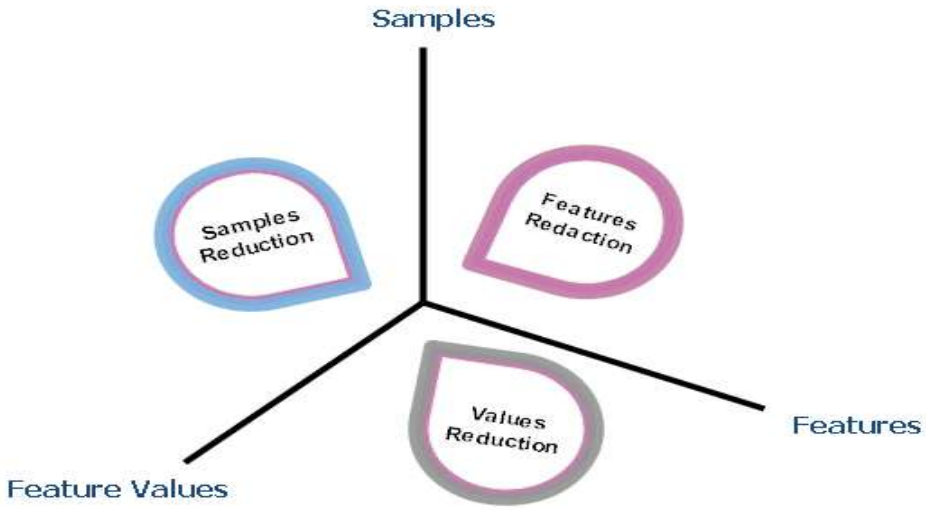

Figure 3. Relationship among Dimensions

2.4 Black box Model

Despite the capability of IDA techniques such as NN, SVM and RF to identify relationships between variables and provide precise results, they do not specify the relative importance of each variable [17]. These methods called black boxes because they are difficult to be understood by humans and their successful requires skilled specialists who can analyze and interpret the outputs [18, 19].
Two main effective strategies can be used to extract useful knowledge from black boxes: Extraction rules and visualization. The first strategy is the most widely adopted method due to its ability to simplify model complexity [20, 21]. However, this approach could extract rules that do not accurately represent the original model. The majority of visualization methods deal with aspects related to the data multidimensionality. The usage of these techniques for black boxes is limited [22] because most of them are specifically designed for a given goal [23].

2.5 Mathematical model

The use of MM has been rapidly increased in different fields within the last few decades due to its ability to provide intuitive understanding of a system and the relationships between its components [24]. Application of MM can be found in various applications such as anticancer therapy [25], ocular hemodynamic and glaucoma [26], biological research [27], physiology and biochemistry of the retina [28], biomechanics of the eye [29] and drug discovery [30].

III. IDA APPROACHES APPLIED TO SOLVE KDD CHALLENGES

In this section, we review efforts that aimed to address KDD problems.

3.1 Missing Data

A least squares formulation based approach was proposed by [31] to estimate missing values in gene expression data using local similarity structures and least squares optimization process. The results showed that a comparative result was achieved by their proposed method when the authors compared its performance with other imputation methods on different datasets.

In another study, a method was presented to find the similarities at task classifying pairs of proteins [32]. They used direct and indirect information about interaction pairs to generate a RF from a training set. The generated RFwas then used to determine similarity between protein pairs. The authors used a modified k-nearest neighbor algorithm to classify pairs. The final results showed that an improvement was attained by their proposed method.

A novel hybrid method to address the problem of missing values was introduced by [33]. The authors combined GMM to handle missing values and ELM to enable multiple imputation method to be executed on a reasonable scale. They confirmed that their proposed method increased the overall accuracy.

PhyloPars web server was suggested to provide a statistically consistent method that combines incomplete set of empirical records with the species phylogeny to produce more effective parameters estimation results for all species [34]. The main objective of the PhyloPars was to build an accurate model for missing data.

In [35], four imputation strategies were investigated to handle missing values in microarray data. They used three local nearest neighbor and global PCA. The authors confirmed that the imputation strategies achieved better results than global PCA based strategy.

A new NSLLSimpute method for missing values estimation in gene expression data was proposed [36]. The results on different datasets showed that their proposed approach obtained higher accuracy results than other methods used in their work.

In [37], the authors proposed an evolutionary k-NN approach for missing data imputation. The authors used GAto optimize k-NNand then they compared between the evolutionary k-NN and k-NN algorithms using three gene expression datasets. The results showed that their optimized algorithm was capable in identifying the value of k and assigning weights to the different attribute in the datasets.

In [38], a LAW-LSimpute framework for missing value estimation was proposed. Their proposed method aimed to automatically weight the neighbors genes based on their importance ranking. An acceleration strategy was added to the proposed LAW-LSimpute method to improve the convergence. Their method was tested on eight benchmark datasets and the final results showed that the suggested approach reduced estimation error efficiently.

3.2 Data scarcity

Several methods are presented to address the problem of data scarcity [39, 40, 41, 42, 43]. A number of research works have confirmed the potential of semi-supervised learning and transfer learning techniques in addressing this problem [44, 45, 46].

The author in [47], applied semi supervised learning and transfer learning methods where the semi supervised learning was used to deduce the amount of unlabeled data and transfer learning was applied to improve classification accuracy. They confirmed that their proposed methods performed well and were found to be effective in dealing with the problem of data scarcity when their proposed approach was evaluated on two datasets.

The authors in [48] combined semi supervised learning, transfer learning and assemble methods to obtain a good model from a scarce data. They used semi supervised learning as a preprocessing method to reduce amount of unlabeled data and then four models based on transfer learning were investigated.

Their idea was to get the distance among models and use the same models to improve predictive accuracy, while transfer learning models are employed to choosethe data from the close models. The proposed method was tested on a data obtained from employees of two different companies. The results showed that the proposed weighted ensemble based model increased the accuracy.
A heuristic model was developed with the purpose of maximizing the benefitsfrom the available data[49]. The data are classified into classes with different errors and the usable data are identified from the available data in order to facilitate data analysis by the decision makers.

3.3 Data dimensionality reduction

Data dimensionality reduction can be divided into feature extraction [50, 51, 52, 53, 54] and feature selection [55, 56, 57] and various approaches were proposed for data dimensionality reduction in recent years [58, 59, 60, 61, 62, 63].
The authors in [64] investigated GP andGAfor data preprocessing to improve the performance of C4.5classifier. GP was applied to construct and discover hidden relationships between features, while they used GA to select the most important features. Ten public datasets obtained from UCI were used to assess the performance of the GP and GA. A significant improvement was attained in the classification accuracy.
In [65],SVM and GA were combined as a hybrid method to develop a classification method named YamiPred.GA was used to identify most crucial features and SVM was employed for parameters optimization. The results showed that their proposed approach outperformed other approaches in terms of accuracy.
A range of studies has applied PCA for dominantly reduction [66, 67, 68]. A WPCA feature selection based method was proposed to analyze and capture relevant features in gene expression datasets [69]. The results revealed that the WPCA performed better than PCA. In another study[70], a novel method for tumor classification was introduced. The method used robust PCA and LDA to identify the most crucial features in seven gene expression datasets. They used identified features as inputs in SVM to classify tumor samples. Their proposed methods produced more effective results for tumor classification.
The work of [71] developed a model to forecast protein structure classes. They used Multi profile Bayes and bi-gram probability to extract features from protein sequences. The PSO was used to select important features from the hybrid space. The performance of their model was compared to other methods using three benchmarking datasets. The results showed that the performance of their method performed better.
RF was applied to identify a set of prognostic genes in a high dimensional data [72]. The performance of their approach was compared with several IDA methods and various split criteria using several real would datasets. Their proposed method outperformed other IDA classifiers.
In [73], two feature selection methods based on information theory were presented to reduce dimensionality. These methods include JMIM and NJMIM. Eleven publicly available datasets along with five feature selection methods were used to evaluate the proposed method and its performance. The JMIM performed better than other methods.
In [74], MMCC was proposed to select informative genes from microarray data. The authors used evolutionary optimization to search for optimal features in each dataset. Their results showed better performance of their method compared to other for 25 used microarray datasets of gene expression. The results also showed that a higher accuracy in classification was attained by SVM in almost all of the datasets in comparison with other results obtained by other well known feature selection methods.

3.4 Black box Models

Sensitivity Analysis (SA) performed well and found to be efficient and effective in extracting valuable knowledge from black box models [75, 76, 77, 78, 79, 80].
The work in [20] used five SA methods and four measures of input importance. The efficiency of their visualization method was evaluated in various classification tasks. The performance capability of the method was also assessed using four real would datasets. They suggested using their method in conjunction with the AAD measure of importance.
The authors in [81] proposed a global SA by combining several visualization techniques such as VEC curve to open black box model. The authors assessed their proposed method capability on different datasets using NN and SVM.

In [82], the authors focused on predicting performance metrics from the available posts in Facebook pages for a company's using SVM. They mainly used SA method to extract useful information from the proposed lifetime post consumer's model.
Numerous approaches have been proposed to determine the relative contribution of every variable in the ANN [83, 84, 85]. In [86], a set of ANN with same architecture were selected to determine the contribution of each input parameters. They used a dataset obtained from "a customer satisfaction survey developed by the transport consortium of the Granada Metropolitan Area (Spain) in 2007" to validate their proposed procedure. The results showed that when each method was used independently, the variable's importance rankings are similar.

3.5 Mathematical model

In [87], the advantages of IDA and FARB were combined to design a MM for classification rules. The proposed approach achieved higher accuracy when it was evaluated on two classifications case studies based on association rules. The results also confirmed that the proposed method can be effectively used for IDA.

IV. Open Issues

This section provides several essential issues that need deserve attention and research in the field of KDD.
After analyzing and discussing the mentioned approaches, it can be observed that there is no independent approach that can address all the issues involved in the KDD. For example, some approaches consider data scarcity, dimensionality reduction and missing data, while some ignore these issues.
The inability to deal with complex interactions between variables prevents IDA techniques to match a set of analysis goals. Therefore, solutions for missing data are essential to meet IDA goals. Recently, one of still problems in estimation missing values methods is how to select the optimal number of nearest neighbors of the missing values [88].
Lacking of sufficient data has a significant problem for the process of IDA techniques. This problem results in inaccurate information on how to extract the appropriate features from a limited amount of available data. The Central Limit Theorem used in statistics [89], supports the idea that lack of sufficient data affects IDA training results because it states that sample size less than thirty is insufficient. To overcome the problems associated with lack of scarce data, more data should be collected and techniques with the capability of handling small amount of data should be designed.
Data dimensionality reduction aims at transforming data variable space from high dimensions to low dimensions space without affecting the correlation structure between variables. Therefore, how to combine between the samples features and values present an open problem which needs focus and attention in the field of KDD.
All kinds of data have different attributes that might pose problems for IDA techniques to extract the most crucial patterns in a given dataset due to the imbalanced classes, and thus, classification results may become unreliable. Various techniques can deal with this problem, for example, undersampling, oversampluing, SMOTE.TomicLinks etc [90, 91, 92]. A comparison of three types of methods developed for class imbalance was carried out by [90],. The authors used different datasets obtained from different application to evaluate seven feature selection measures. They used AUC and PRC to assess the performance of these measures. The results revealed that both single to noise correlation coefficient and feature assessment by sliding thresholds worked well for feature selection in various applications.
Converting black boxes into understandable systems (i.e., white box) has remained a challenge for researchers in the field of KDD and it needs further investigations. Also, MM allows for a more intuitive understanding of the system and its components. However, few works consider it in the field of KDD and how to translate hypothesis into a set of mathematical equations is still an open issue. Therefore, greater attention and development in this area are required from both mathematical and computational modeling communities.

V. Conclusion

In this work, we highlight the main trends in KDD. We discussed solutions that proposed to overcome these trends and pointed out some open research issues of research which can be considered by researchers for future work in the area of KDD.

REFERENCES

[1] Wu, X., Zhu, X., Wu, G. Q., & Ding, W. (2014). Data mining with big data. IEEE transactions on knowledge and data engineering, 26(1), 97-107.

[2] Hans-Peter Kriegel , Karsten M. Borgwardt · Peer Kröger · Alexey Pryakhin · Matthias Schubert and Arthur Zimek., "Future trends in data mining". Data Mining and Knowledge Discovery, Springer. 15:87–97. 2007. DOI 10.1007/s10618-007-0067-9

[3] S. H. Ali, Miner for OACCR: Case of medical data analysis in knowledge discovery. In 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT), (2012), 962-975.

[4] D. B Rubin, Inference and missing data, Biometrika. 63(3) (1976) 581-592.

[5] A. W. C Liew, N. F Law, H. Yan, Missing value imputation for gene expression data: computational techniques to recover missing data from available information. Briefings in bioinformatics, 12(5), (2010), 498-513.

[6] Jiawei Han and Micheline Kamber.," Data Mining: Concepts and Techniques" Second Edition. University of Illinois at Urbana-Champaign. 500 Sansome Street, Suite 400, San Francisco. Elsevier 2006. www.books.elsevier.com.

[7] Maxhuni, A. (2017). Managing the Scarcity of Monitoring Data through Machine Learning in Healthcare Domain (Doctoral dissertation, University of Trento).

[8] Kantardzic, M. (2011). Data mining: concepts, models, methods, and algorithms. John Wiley & Sons.

[9] Mehmed Kantardzic ," Data Mining: Concepts, Models, Methods, and Algorithms" IEEE Computer society, Sponser ,2003.

[10] Langley, P. (1996). Elements of machine learning. Morgan Kaufmann

[11] Fu, X., & Wang, L. (2003). Data dimensionality reduction with application to simplifying RBF network structure and improving classification performance. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 33(3), 399-409.

[12] Azar, A. T., & Hassanien, A. E. (2015). Dimensionality reduction of medical big data using neural-fuzzy classifier. Soft computing, 19(4), 1115-1127.

[13] S. H. Ali, "A novel tool (FP-KC) for handle the three main dimensions reduction and association rule mining," IEEE, 2012 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT), Sousse, 2012, pp. 951-961.

[14] Langley, P., & Sage, S. (1994). Induction of selective Bayesian classifiers. In Uncertainty Proceedings 1994 (pp. 399-406).

[15]. Koller, D., & Sahami, M. (1996). Toward optimal feature selection. Stanford InfoLab

[16]. Barak Chizi and Oded Maimon," Dimension Reduction and Feature Selection", Data Mining and Knowledge Discovery Handbook, 2nd ed., Springer Science and Business Media, LLC 2010.

[17] Cortez, P., & Embrechts, M. J. (2013). Using sensitivity analysis and visualization techniques to open black box data mining models. Information Sciences, 225, 1-17.

[18] Jeffrey W. Seifert, "Data Mining: An Overview," CRS Report for Congress, 2004.

[19] Olden, J. D., & Jackson, D. A. (2002). Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks. Ecological modeling, 154(1-2), 135-150.

[20] Setiono, R. U. D. Y. (2003). Techniques for extracting classification and regression rules from artificial neural networks. Computational intelligence: The experts speak, 99-114.

[21] Martens, D., Baesens, B., Van Gestel, T., & Vanthienen, J. (2007). Comprehensible credit scoring models using rule extraction from support vector machines. European journal of operational research, 183(3), 1466-1476.

[22] Craven, M. W., & Shavlik, J. W. (1992). Visualizing learning and computation in artificial neural networks. International Journal on Artificial Intelligence Tools, 1(03), 399-425.

[23] Tzeng, F. Y., & Ma, K. L. (2005, October). Opening the black box-data driven visualization of neural networks. In Visualization, 2005. VIS 05. IEEE (pp. 383-390). IEEE.

[24] Bradley, P. S., Fayyad, U. M., & Mangasarian, O. L. (1999). Mathematical programming for data mining: Formulations and challenges. INFORMS Journal on Computing, 11(3), 217-238.

[25] Swierniak, A., Kimmel, M., & Smieja, J. (2009). Mathematical modeling as a tool for planning anticancer therapy. European journal of pharmacology, 625(1-3), 108-121.

[26] Harris, A., Guidoboni, G., Arciero, J. C., Amireskandari, A., Tobe, L. A., & Siesky, B. A. (2013). Ocular hemodynamics and glaucoma: the role of mathematical modeling. European journal of ophthalmology, 23(2), 139.

[27] Glynn, P., Unudurthi, S. D., & Hund, T. J. (2014). Mathematical modeling of physiological systems: an essential tool for discovery. Life sciences, 111(1-2), 1-5.

[28] Ganesan, P., He, S., & Xu, H. (2010). Development of an image-based network model of retinal vasculature. Annals of biomedical engineering, 38(4), 1566-1585.
[29] Grytz, R., & Meschke, G. (2010). A computational remodeling approach to predict the physiological architecture of the collagen fibril network in corneo-scleral shells. Biomechanics and modeling in mechanobiology, 9(2), 225-235.
[30] Mirams, G. R., Davies, M. R., Cui, Y., Kohl, P., & Noble, D. (2012). Application of cardiac electrophysiology simulations to pro-arrhythmic safety testing. British journal of pharmacology, 167(5), 932-945.
[31] Kim, H., Golub, G. H., & Park, H. (2004). Missing value estimation for DNA microarray gene expression data: local least squares imputation. Bioinformatics, 21(2), 187-198.
[32] Qi, Y., Klein-Seetharaman, J., & Bar-Joseph, Z. (2005). Random forest similarity for protein-protein interaction prediction from multiple sources. In Biocomputing 2005 (pp. 531-542).
[33] Sovilj, D., Eirola, E., Miche, Y., Björk, K. M., Nian, R., Akusok, A., & Lendasse, A. (2016). Extreme learning machine for missing data using multiple imputations. Neurocomputing, 174, 220-231.
[34] Bruggeman J, Heringa J, Brandt B (2009) PhyloPars: estimation of missing parameter values using phylogeny. Nucleic Acids Research 37(2):W179–W184
[35] Ryan C, Green D, Cagney G, Cunningham P (2010) Missing value imputation for epistatic MAPs. Bioinformatics 11(1):197. doi:10.1186/1471-2105-11-197
[36] Bose, S., Das, C., Gangopadhyay, T., & Chattopadhyay, S. (2013, December). A modified local least squares-based missing value estimation method in microarray gene expression data. In Advanced Computing, Networking and Security (ADCONS), 2013 2nd International Conference on(pp. 18-23). IEEE.
[37] de Silva, H., & Perera, A. S. (2016, September). Missing data imputation using Evolutionary k-Nearest neighbor algorithm for gene expression data. In Advances in ICT for Emerging Regions (ICTer), 2016 Sixteenth International Conference on(pp. 141-146). IEEE.
[38] Yu, Z., Li, T., Horng, S. J., Pan, Y., Wang, H., & Jing, Y. (2017). An Iterative Locally Auto-Weighted Least Squares Method for Microarray Missing Value Estimation. IEEE transactions on nanobioscience, 16(1), 21-33.
[39] Z.-H. Zhou and Y. Jiang, ," Medical diagnosis with C4.5 rule preceded by artificial neural network ensemble" , IEEE Transactions on Information Technology in Biomedicine, vol.7, no.1, pp.37–42. 2003
[40] Z.-H. Zhou and Y. Jiang,., "NeC4.5: Neural ensemble based C4.5". IEEE Transactions on Knowledge and Data Engineering, vol.16, no.6, pp.770–773. 2004.
[41] Li, D.-C., & Lin, Y.-S. "Using virtual sample generation to build up management knowledge in the early manufacturing stages". European Journal of Operational Research, 175, 413-434. 2006.
[42] Yuan J., Ming L. and Zhi H. "Mining Extremely Small Data Sets with Application to Software Reuse". Elsevier Preprint. China. 2008.
[43] Chun-J and Hsiao F. "Virtual Sampling with Data Construction Method". Information Science Reference. Hershey • New York. 2009
[44] Zhu, X. (2005). Semi-supervised learning literature survey, University of Wisconsin – Madison http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.99.9681&rep=rep1&type=pdf
[45] Dai, W, Xue G R, Yang Q, Yu Y, (2007). "Transferring naive bayes classifiers for text classification". In: Proceedings of the national conference on artificial intelligence. Vol. 22. 1. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, p. 540
[46] Shi, X., Fan, W., & Ren, J. (2008, September). Actively transfer domain knowledge. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases(pp. 342-357). Springer, Berlin, Heidelberg.
[47] Maxhuni, A. (2017). Managing the Scarcity of Monitoring Data through Machine Learning in Healthcare Domain (Doctoral dissertation, University of Trento).
[48] [38] Maxhuni, A., Hernandez-Leal, P., Sucar, L. E., Osmani, V., Morales, E. F., & Mayora, O. (2016). Stress modelling and prediction in presence of scarce data. Journal of biomedical informatics, 63, 344-356.
[49] Zakeri, A., Saberi, M., Hussain, O. K., & Chang, E. (2017, August). A Heuristic Machine Learning Based Approach for Utilizing Scarce Data in Estimating Fuel Consumption of Heavy Duty Trucks. In International Conference on Intelligent Networking and Collaborative Systems (pp. 96-107). Springer, Cham.
[50] Chen, L. H., & Chang, S. (1995). An adaptive learning algorithm for principal component analysis. IEEE transactions on neural networks, 6(5), 1255-1263.

[51] Hussein K., "Knowledge Discovery in database by using data mining" Ph.D. Thesis, University of Technology, 2002.
[52] Fu, X., & Wang, L. (2001). Rule extraction by genetic algorithms based on a simplified RBF neural network. In Evolutionary Computation, 2001. Proceedings of the 2001 Congress on (Vol. 2, pp. 753-758). IEEE.
[53] Nian Yan : Classification Using Neural Network Ensemble with Feature Selection., Ph.D thesis Linköpings university , Sweden, 2004.
[54] Zhang, L., Wang, L., Lin, W., & Yan, S. (2014). Geometric optimum experimental design for collaborative image retrieval. IEEE Transactions on Circuits and Systems for Video technology, 24(2), 346-359.
[55] Bins, J., & Draper, B. A. (2001). Feature selection from huge feature sets. In Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on (Vol. 2, pp. 159-165). IEEE.
[56] Zhong, J., Wang, J., Peng, W., Zhang, Z., & Li, M. (2015). A feature selection method for prediction essential protein. Tsinghua Science and Technology, 20(5), 491-499.
[57] Fong, S., Deb, S., Yang, X. S., & Li, J. (2014). Feature selection in life science classification: metaheuristic swarm search. IT Professional, 16(4), 24-29.
[58] Koller, D., & Sahami, M. (1996). Toward optimal feature selection. in: the 13th International Conference on Machine Learning (ML), 1996, pp. 284–292.
[59] Fu, X., & Wang, L. (2003). Data dimensionality reduction with application to simplifying RBF network structure and improving classification performance. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 33(3), 399-409
[60] Halgamuge, S. K., & Wang, L. (Eds.). (2005). Classification and clustering for knowledge discovery (Vol. 4). Springer Science & Business Media..
[61] Langley, P., & Sage, S. (1994). Induction of selective Bayesian classifiers. In Uncertainty Proceedings 1994 (pp. 399-406).
[62] Samaher Hussein " Designing a Software for Knowledge Discovery in Database Using Data Mining and Soft Computing Techniques .IEEE, 2nd International Conference: E-Medical Systems October 29-31, 2008
[63] Burges, C. J. (2009). Geometric methods for feature extraction and dimensional reduction-a guided tour. In Data mining and knowledge discovery handbook (pp. 53-82). Springer, Boston, MA.
[64] Smith, M. G., & Bull, L. (2003, April). Feature construction and selection using genetic programming and a genetic algorithm. In European Conference on Genetic Programming (pp. 229-237). Springer, Berlin, Heidelberg.
[65] Kleftogiannis, D., Theofilatos, K., Likothanassis, S., & Mavroudi, S. (2015). YamiPred: A novel evolutionary method for predicting pre-miRNAs and selecting relevant features. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 12(5), 1183-1192.
[66] Chu, F., & Wang, L. (2005). Applications of support vector machines to cancer classification with microarray data. International journal of neural systems, 15(06), 475-484.
[67] Chu, F., & Wang, L. (2003, July). Gene expression data analysis using support vector machines. In Neural Networks, 2003. Proceedings of the International Joint Conference on(Vol. 3, pp. 2268-2271). IEEE.
[68] Al_Janabi, S., Al_Shourbaji, I., & Salman, M. A. (2017). Assessing the suitability of soft computing approaches for forest fires prediction. Applied Computing and Informatics.
[69] Da Costa, J. F. P., Alonso, H., & Roque, L. (2011). A weighted principal component analysis and its application to gene expression data. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 8(1), 246-252.
[70] Liu, J. X., Xu, Y., Zheng, C. H., Kong, H., & Lai, Z. H. (2015). RPCA-based tumor classification using gene expression data. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 12(4), 964-970.
[71] Hayat, M., Tahir, M., & Khan, S. A. (2014). Prediction of protein structure classes using hybrid space of multi-profile Bayes and bi-gram probability feature spaces. Journal of theoretical biology, 346, 8-15.
[72] Pang, H., George, S. L., Hui, K., & Tong, T. (2012). Gene selection using iterative feature elimination random forests for survival outcomes. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 9(5), 1422-1431.
[73] Bennasar, M., Hicks, Y., & Setchi, R. (2015). Feature selection using joint mutual information maximisation. Expert Systems with Applications, 42(22), 8520-8532.
[74] Mohammadi, M., Noghabi, H. S., Hodtani, G. A., & Mashhadi, H. R. (2016). Robust and stable gene selection via Maximum–Minimum Correntropy Criterion. Genomics, 107(2), 83-87.

[75] Kewley, R. H., Embrechts, M. J., & Breneman, C. (2000). Data strip mining for the virtual design of pharmaceuticals with neural networks. IEEE Transactions on Neural Networks, 11(3), 668-679.
[76] Krause, J., Perer, A., & Ng, K. (2016, May). Interacting with predictions: Visual inspection of black-box machine learning models. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (pp. 5686-5697). ACM.
[77] Cortez, P., Teixeira, J., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009, October). Using data mining for wine quality assessment. In International Conference on Discovery Science (pp. 66-79). Springer, Berlin, Heidelberg.
[78] Moro, S., Cortez, P., & Rita, P. (2015). Using customer lifetime value and neural networks to improve the prediction of bank deposit subscription in telemarketing campaigns. Neural Computing and Applications, 26(1), 131-139.
[79] Embrechts, M. J., Arciniegas, F. A., Ozdemir, M., & Kewley, R. H. (2003). Data mining for molecules with 2-D neural network sensitivity analysis. International Journal of smart engineering system design, 5(4), 225-239..
[80] Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. Decision Support Systems, 62, 22-31.
[81] Cortez, P., & Embrechts, M. J. (2011, April). Opening black box data mining models using sensitivity analysis. In Computational Intelligence and Data Mining (CIDM), 2011 IEEE Symposium on (pp. 341-348). IEEE.
[82] Moro, S., Rita, P., & Vala, B. (2016). Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach. Journal of Business Research, 69(9), 3341-3351.
[83] Gevrey, M., Dimopoulos, I., & Lek, S. (2003). Review and comparison of methods to study the contribution of variables in artificial neural network models. Ecological modelling, 160(3), 249-264.
[84] Paliwal, M., & Kumar, U. A. (2011). Assessing the contribution of variables in feed forward neural network. Applied Soft Computing, 11(4), 3690-3696.
[85] Olden, J. D., Joy, M. K., & Death, R. G. (2004). An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. Ecological Modelling, 178(3-4), 389-397.
[86] De Oña, J., & Garrido, C. (2014). Extracting the contribution of independent variables in neural network models: a new approach to handle instability. Neural Computing and Applications, 25(3-4), 859-869.
[87] Al-Janabi, S., & Alwan, E. (2017, June). Soft Mathematical System to Solve Black Box Problem through Development the FARB Based on Hyperbolic and Polynomial Functions. In Developments in eSystems Engineering (DeSE), 2017 10th International Conference on (pp. 37-42). IEEE.
[88] Colm R, Derek G., Gerard C. and Padraig C.," ŝMissing Value Imputation for Epistatic MAPs" School of Computer Science and Informatics, University College Dublin, Dublin, Ireland. Bioinformatics 2010.
[89] Ross, M. S. ," Introduction to probability and statistics for engineers and scientists".John Wiley & Sons, Inc., 1987.
[90] Fu, X., Wang, L., Chua, K. S., & Chu, F. (2002, November). Training RBF neural networks on unbalanced data. In Neural Information Processing, 2002. ICONIP'02. Proceedings of the 9th International Conference on (Vol. 2, pp. 1016-1020). IEEE.
[91] Patel, A., Al-Janabi, S., AlShourbaji, I., & Pedersen, J. (2015). A novel methodology towards a trusted environment in mashup web applications. computers & security, 49, 107-122.
[92] Yu, L., Han, Y., & Berens, M. E. (2012). Stable gene selection from microarray data via sample weighting. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 9(1), 262-272.
[93] Wasikowski, M., & Chen, X. W. (2010). Combating the small sample class imbalance problem using feature selection. IEEE Transactions on knowledge and data engineering, 22(10), 1388-1400.